# Multi-class classification in nonparametric active learning

**Boris Ndjia Njike**
University of Mons

**Xavier Siebert**
University of Mons

## Abstract

Several works have recently focused on nonparametric active learning, especially in the binary classification setting under Hölder smoothness assumptions on the regression function. These works have highlighted the benefit of active learning by providing better rates of convergence compared to the passive counterpart. In this paper, we extend these results to multiclass classification under a more general smoothness assumption, which takes into account a broader class of underlying distributions. We present a new algorithm called `MKAL` for multiclass $k$-nearest neighbors active learning, and prove its theoretical benefits. Additionally, we empirically study `MKAL` on several datasets and discuss its merits and potential improvements.

## 1 Introduction

Active learning is a paradigm of machine learning in which the learner does not have access to a fully labeled dataset. Starting with unlabeled data, the labels are obtained interactively and at some cost from a so-called oracle, and then used to progressively construct a classifier. Active learning algorithms aim at simultaneously reducing the labeling cost, and at achieving better performances than standard passive learning (Dasgupta, 2011; Settles, 2010). Their performance is typically measured in terms of the number of labels required to achieve a given classification error. The theory of active learning is now fairly well developed, both in parametric (Hanneke, 2011; Balcan et al., 2010, 2007, 2009; Hanneke and Yang, 2015) and in nonparametric settings (Castro and Nowak, 2008; Minsker, 2012; Locatelli et al., 2017, 2018; Hanneke, 2018), and provides a framework to study conditions under which active learning is favorable.

In this paper, we are particularly interested in the nonparametric setting, for which seminal work has been achieved by Castro and Nowak (Castro and Nowak, 2008) in the binary classification setting. Their work relies on assumptions on the smoothness of boundary decision between classes, and on the noise distribution to provide an active learning algorithm that provably achieves a better rate of convergence than in passive learning. Subsequent works, for example (Minsker, 2012; Locatelli et al., 2018; Hanneke, 2018) rely on a slightly modified version of the assumptions used in (Castro and Nowak, 2008) to provide algorithms with a (minimax) rate of convergence on the order $n^{-\beta(\alpha+1)/(2\alpha+d-\alpha\beta)}$, where $\alpha$ and $\beta$ are smoothness and noise parameters, $d$ is the dimension of the data space, and $n$ the number of label requests. These parameters will be discussed in details in Section 3 but we can already notice that the active learning rate compares favorably to the passive learning one, which is of the order of $n^{-\beta(\alpha+1)/(2\alpha+d)}$.

Unfortunately, most of these algorithms suffer a major limitation: the smoothness assumption (e.g., Hölder, see Audibert and Tsybakov (2007)) requires the use of a strong density assumption. This implies the existence of the density function of the marginal probability of the unlabeled data, and it also requires the support of this marginal probability to be bounded. Additionally, these preceding studies only considered the binary classification setting. In this paper, we consider a more universal smoothness assumption which takes into account a broader class of probabilities while avoiding the use of the strong density assumption. We propose a new active learning algorithm based on $k$-nearest neighbors in a multi-class setting and prove theoretically that it performs better than its passive counterpart. The paper is organized as follows: Section 2 presents related work on active learning, nearest neighbors and multiclass classification. In Section 3 we introduce the main notations and assumptions used in this work. Section 4 contains an overall description of our algorithm, while Section 5 presents its theoretical

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. 

properties. Section 6 provides an experimental study on several datasets and a discussion of its results, and we conclude with Section 7. Some of the proofs being quite long, they are relegated to the Supplementary Material, which contains additional experimental results as well.

## 2 Related works

**Active learning.** Previous works (Dasgupta et al., 2007; Balcan et al., 2009; Hanneke, 2011; Balcan et al., 2010; Hanneke and Yang, 2015; Beygelzimer et al., 2009; Castro and Nowak, 2008; Minsker, 2012; Locatelli et al., 2017) in active learning have shown that we can obtain a good performance by using a much smaller number of labeled sample than in passive learning.

As stated in Section 1, our work is most related to that of Minsker (2012), Locatelli et al. (2017) in a nonparametric setting, which extended the seminal work from Castro and Nowak (2008) by assuming that the regression function belongs to the Hölder class of functions. In this case, under some additional assumptions (related to Tsybakov's noise and the strong density assumption, see Section 3.2) and some range of values of distributional parameters, they provided algorithms with faster rates of convergence than those obtained in passive learning by Audibert and Tsybakov (2007). Furthermore, these algorithms are adaptive to their distributional parameters unlike that of Castro and Nowak (2008). Although these active learning algorithms are very interesting regarding their advantage with respect to their passive counterpart, the use of the Hölder smoothness is a limiting factor. Indeed, it requires the use of a strong density assumption and thus the existence of the density function of the marginal probability $P_X$ on the unlabeled data as well as the boundedness of the support of $P_X$. These limitations motivated the development of our algorithm with a more general smoothness assumption, valid for both discrete and continuous distributions, and which does not require the support of $P_X$ to be bounded.

**Nearest neighbors.** Nearest neighbors (NN) classification has been widely characterized in passive learning (see for example Cover and Hart (1967); Kulkarni and Posner (1995); Devroye et al. (1994); Biau and Devroye (2015)). In particular, its theoretical performance has been considered for Hölder regression function. Moreover, using also Tsybakov's noise and strong density assumptions, Chaudhuri and Dasgupta (2014) showed that the convergence rate behaves as $n^{\alpha(\beta+1)/(2\alpha+d)}$, the same as that obtained earlier in Audibert and Tsybakov (2007) (where $\alpha$, $\beta$, $d$ represent the smoothness parameter, the noise parameter and the dimension of the data space, respectively, see Section 3.2). The nearest neighbors method has also been used recently in active learning. For instance, Kontorovich et al. (2016) considered an active learning method which derives a subsample of the data on which the 1-NN method is applied. They showed that this approach is statistically consistent. However, their assumptions differ from ours in terms of smoothness and noise. Similarly, the algorithm proposed in (Hanneke, 2018) outputs a 1-NN classifier based on a subsample of the data, such that the label of each instance of this subsample is determined with high probability by the labels of its neighbors within a large pool of data. The number of neighbors is adaptively chosen for each instance in the subsample, leading to the minimax rate $n^{-\alpha(\beta+1)/(2\alpha+d-\alpha\beta)}$ under the same assumptions as in (Locatelli et al., 2017). In this work, we follow the same procedure as Hanneke (2018) under a more general smoothness assumption, which will be defined in Section 3.1.

**Multiclass-classification** Most of the previous results in active learning in the same line of research as ours (such as Minsker (2012); Locatelli et al. (2017); Hanneke (2018)) are limited to binary classification. Our work extends those results to multiclass classification, assuming that the label set contains more than two classes. In passive learning, Reeve and Brown (2017) derived procedures for cost-sensitive multiclass classification, where different misclassification errors incur different costs. They considered the same noise and smoothness assumptions as ours, and assumed that the data belong to a manifold which is characterized by an intrinsic dimension $D$ which could be much smaller than the dimension of the ambient space, leading to a rate of convergence which behaves as $n^{-\alpha(\beta+1)/(2\alpha+D)}$. Some recent works have also considered multiclass classificatication with nearest neighbors method under similar assumptions as ours (e.g., Puchkin and Spokoiny (2020); Györfi and Weiss (2021)). For instance, Puchkin and Spokoiny (2020) used a nearest neighbors method to provide an aggregated estimate. Their procedure is adaptive both to the noise and smoothness parameters and leads to a non-asymptotic analysis which generalizes that of Chaudhuri and Dasgupta (2014).

## 3 Notations and assumptions

### 3.1 General notations

Let $(\mathcal{X}, \rho)$ be a metric space, where $\mathcal{X} \subset \mathbb{R}^d$ is called the instance space. Let the number of classes $M \geq 2$ and the label space $\mathcal{Y} = \{1, \dots, M\}$. Let $\Omega(\mathcal{Y})$ be defined as the $(M-1)$-simplex consisting of probability

vectors over $\mathcal{Y}$. Let $P$ be a probability defined on $\mathcal{X}\times\mathcal{Y}$. We assume that the probability $P$ is determined by the marginal probability $P_X$ defined on $\mathcal{X}$ and by the regression function $\eta$ defined by:

$$\begin{array}{rccl}\eta & : & \mathcal{X} & \to & \Omega(\mathcal{Y}) \\ & & x & \mapsto & \eta(x) = (\eta_1(x), \ldots, \eta_M(x)),\end{array} \quad (1)$$

where $\eta_i(x) = P(Y = i | X = x)$. Let $f : \mathcal{X} \longrightarrow \mathcal{Y}$ a classifier whose risk is $R(f) = P(f(X) \neq Y)$. It is easy to show that the function defined by

$$f^*(x) = \underset{i\in\{1,\ldots,M\}}{\arg\max}\ \eta_i(x) \quad (2)$$

achieves the minimum risk. In practice $P$ is unknown and the function $f^*$ is unreachable. In standard statistical learning, the algorithm has access to an i.i.d sample $(X_1, Y_1), \ldots, (X_w, Y_w)$ based on which it constructs a classifier $\widehat{f}_w$ with minimum excess risk $\mathcal{R}(\widehat{f}_w) - \mathcal{R}(f^*)$ with high probability, where the latter can be rewritten as :

$$\mathcal{R}(\widehat{f}_w) - \mathcal{R}(f^*) = E_X(\eta_{f^*(X)}(X) - \eta_{\hat{f}_w(X)}(X)). \quad (3)$$

For multiclass classification purposes, let us introduce $\zeta(v)$ such that, given a vector $v \in [0,1]^M$,

$$\zeta(v) = v_{i_v} - \max_{j\neq i_v} v_j, \quad \text{where} \quad i_v = \underset{j\in\{1,\ldots,M\}}{\arg\max}\ v_j. \quad (4)$$

For $x \in \mathcal{X}$, and $r > 0$, let $B(x,r) = \{z,\ \rho(x,z) \leq r\}$ and $\text{supp}(P_X) = \{x \in \mathcal{X},\ \ P_X(B(x,r)) > 0,\ \ \forall r > 0\}$.

### 3.2 Assumptions

We will make the following three assumptions on $P$.

**Assumption 1:** (Tsybakov's noise assumption)
The probability distribution $P$ satisfies the **margin noise** assumption with parameter $\beta \geq 0$ if for all $0 < \epsilon \leq 1$, there is $C = C(\beta) \in [1, +\infty[$ such that

$$P_X(x \in \mathcal{X},\ \zeta(\eta(x)) \leq \epsilon) < C\epsilon^{\beta}. \quad \text{(H1)}$$

This assumption generalizes the standard noise condition of Audibert and Tsybakov (2007) to the multiclass setting.

**Assumption 2:** ($(\alpha, L)$-smoothness assumption)
Let $0 < \alpha \leq 1$ and $L > 1$. The regression function is $(\alpha, L)$**-smooth** if for all $x, z \in \text{supp}(P_X)$ we have:

$$\| \eta(x) - \eta(z) \|_\infty \leq L.P_X(B(x, \rho(x,z)))^{\alpha/d}, \quad \text{(H2)}$$

where $d$ is the dimension of the instance space. This assumption was introduced in Chaudhuri and Dasgupta (2014) and is particularly well suited for $k$-NN classification. Indeed, it allows to nicely control the misclassification in the low-density regions. Previous works on nonparametric active learning (such as Locatelli et al. (2017); Minsker (2012)) that assume Hölder smoothness instead of $(\alpha, L)$-smoothness also need a strong density assumption. As stated in Theorem 3.1, the $(\alpha, L)$-smoothness allows to generalize the Hölder smoothness, thereby avoiding a strong density assumption.

**Theorem 3.1** ($(\alpha, L)$-smoothness assumption generalizes $(\alpha_H, L_H)$-Hölder and strong density assumptions, Chaudhuri and Dasgupta (2014))**.**
*Suppose that the regression function $\eta$ is $(\alpha_H, L_H)$-Hölder continuous:*

$$\| \eta(x) - \eta(z) \|_\infty \leq L_H \rho(x,z)^{\alpha_H},$$

*and that $P_X$ satisfies the strong density assumption i.e., $P_X$ has a density $p_X$ and there exist $r_0 > 0$, $c_0 > 0$ and $p_0 > 0$ such that:*

$$Vol(B(x,r) \cap \text{supp}(P_X)) \geq c_0\, Vol(B(x,r)),$$

*for all $r \leq r_0$ and $p_X(x) > p_0, x \in \text{supp}(P_X)$.*
*Then there is a constant $L > 1$ such that for any $x, z \in \text{supp}(P_X)$, we have:*

$$\| \eta(x) - \eta(z) \|_\infty \leq L.P_X(B(x, \rho(x,z)))^{\alpha_H/d}.$$

**Assumption 3:** (Doubling probability)
The marginal distribution $P_X$ is a **doubling-probability** if there exists a constant $C_{db} > 0$ such that for any $x \in \mathcal{X}$, and $r > 0$, we have:

$$P_X(B(x,r)) \leq C_{db} P_X(B(x, r/2)). \quad \text{(H3)}$$

This notion was first used in the context of metric space dimension by Edgar (2000) and was recently adapted to the area of machine learning where it helps to formally define the notion of intrinsic dimension (e.g., Kpotufe (2011)), that is the actual dimension of the region in which the data is located. This assumption allows to considerably reduce the complexity of the classification problem and to bypass the so-called curse of dimension. Contrary to the strong density assumption, Assumption 3 does not require the existence of the density of $P_X$, and is thus more universal. In this work, we will use Assumption 3 for geometrical reasons, and also in a weaker form, so that it is sufficient to deal with balls $B(x,r)$ such that $P_X(B(x,r))$ is sufficiently large. Some procedures to construct a doubling probability are described in Kpotufe (2011).

### 3.3 Active learning procedure

Let us consider $\mathcal{K} = \{(X_1, Y_1), \ldots, (X_w, Y_w)\}$ a pool of labeled data, with $w > 1$ an integer. In active learning, the labels are hidden, and only the unlabeled

part $\mathcal{K}_{\mathcal{X}} = \{X_1, \ldots, X_w\}$ is accessible. The labels of some carefully selected points $(X_{i_1}, \ldots, X_{i_j} \ldots)$ are iteratively requested until some stopping condition related to the number $n$ of label requests allowed, called the *budget*. Based on the resulting labeled set, a reasonably good estimator of the Bayes classifier (2) is provided.

In our procedure described in Section 4, instead of requesting the label of some points $X_{i_j}$, the labels of their $k$ nearest neighbors in $\mathcal{K}_{\mathcal{X}}$ are requested, where $k$ is chosen adaptively by our algorithm. Therefore, the labels of points $X_{i_j}$ are provided by using the $k-$nearest neighbors procedure. The definition of $k$-nearest neighbors procedure in multiclass classification is recalled below for convenience.

**Definition 3.1** ($k$-nearest neighbors procedure)**.** *Let $w > 0$ and $k < w$ two integers. Let $\mathcal{K} = \{(X_1, Y_1), \ldots, (X_w, Y_w)\}$ an i.i.d. labeled sample. Let $I_k(x)$ the indices of the $k$ nearest neighbors of $x$ in the sample $\mathcal{K}$. The $k$-NN procedure can be defined through the estimator:*

$$\begin{array}{rccl} \hat{\eta}_k & : & \mathcal{X} & \to & \Omega(\mathcal{Y}) \\ & & x & \mapsto & \hat{\eta}_k(x) = (\hat{\eta}_{k,1}(x), \ldots, \hat{\eta}_{k,M}(x)), \end{array}$$

*where $\hat{\eta}_{k,i}(x) = \frac{1}{k} \sum_{j \in I_k(x)} \mathbb{1}_{Y_j = i}$. Then the label of $x$ obtained by the $k$-NN procedure is $\arg\max_{i \leq M} \hat{\eta}_{k,i}(x)$.*

### 3.4 Specific notations

Before describing in details our algorithm, let us introduce some specific variables and notations that will be used throughout the remainder of the work.
For $\epsilon, \delta \in (0,1)$, $k \geq 1$, $c \geq 7.10^6$ and $C$ from (H1):

$$b_{\delta,k} = \sqrt{\frac{2}{k}\left(\log\left(\frac{1}{\delta}\right) + \log\log\left(\frac{1}{\delta}\right) + \log\log(e.k)\right)},$$

$$k(\epsilon, \delta) = \frac{c}{\Delta^2}\left[\log(\frac{1}{\delta}) + \log\log(\frac{1}{\delta}) + \log\log\left(\frac{512\sqrt{e}}{\Delta}\right)\right],$$

$$\Delta = \max(\frac{\epsilon}{2}, \left(\frac{\epsilon}{C}\right)^{\frac{1}{\beta+1}}), \tag{5}$$

$$\phi_n = \sqrt{\frac{1}{n}\left(\log\left(\frac{1}{\delta}\right) + \log\log\left(\frac{1}{\delta}\right)\right)}. \tag{6}$$

For $X_s \in \mathcal{K}_{\mathcal{X}} = \{X_1, \ldots, X_w\}$, we denote henceforth by $X_s^{(k)}$ its $k$-th nearest neighbor in $\mathcal{K}_{\mathcal{X}}$, and $Y_s^{(k)}$ the corresponding label. For an integer $k \geq 1$, let

$$\bar{\eta}_k(X_s) = (\bar{\eta}_{k,1}(X_s), \ldots, \bar{\eta}_{k,M}(X_s)),$$

where for $i \in \mathcal{Y}$, $\bar{\eta}_{k,i}(X_s) = \frac{1}{k} \sum_{j=1}^{k} \eta_i(X_s^{(j)})$.

## 4 The MKAL algorithm

Given a pool $\mathcal{K}_{\mathcal{X}}$ of unlabeled data (with $|\mathcal{K}_{\mathcal{X}}| = w$), the label budget $n$, a precision parameter $\epsilon$, a confidence parameter $\delta \in (0, \frac{1}{2})$, and parameters related to the assumptions which were introduced in Section 3.1, our algorithm aims at providing a classifier $\hat{f}_{n,w}$ based on a set of points $\mathcal{S}_{ac}$ chosen from $\mathcal{K}_{\mathcal{X}}$ to be most *informative*. The first element of the set is $X_{t_1} = X_1$ arbitrarily chosen in $\mathcal{K}$. A point $X_i$ is then considered *informative* if its label cannot be inferred from the previous observations $X_{i'}$ (with $i' < i$). Furthermore, the set $\mathcal{S}_{ac}$ is provided in such a way that, with high confidence, the classifier $\widehat{f}_{n,w}$ has a zero pointwise excess error $\eta_{f^*(x)}(x) - \eta_{\hat{f}_{n,w}(x)}(x)$ at points $x$ which satisfy $\zeta(\eta(x)) > \Delta_0$ for some $\Delta_0 > 0$ suitably chosen. The label of an informative point is then inferred by the $k$-nearest neighbors procedure defined above, where $k$ is chosen adaptively by our algorithm for each point. This is reasonable for practical situations where the uncertainty about the label of $X_t$ has to be overcome, and it is related to the assumption (H2). A similar procedure was already used in Hanneke (2018) and Kontorovich et al. (2016).

Our algorithm `MKAL` (Multiclass $k$-NN Active Learning, Algorithm 1) works iteratively until the budget $n$ is consumed or if $X_w$ is reached. It mainly uses two subroutines: `Reliable` to determine the informativeness of a point and `ConfidentLabel` to infer its label. These subroutines are detailed in Sections 4.1 and 4.2, respectively.

### 4.1 Reliable subroutine

Given the current point $X \in \mathcal{K}_{\mathcal{X}}$, the `Reliable` subroutine aims at determining if $X$ is informative. Intuitively, `Reliable` attempts to find out if the label of $X$ can be inferred using the previously labeled points. More formally, let $X'$ a point that was examined before $X$, whose label has been inferred as $\hat{Y}'$. In this case, if $\zeta(\eta(X'))$ is sufficiently large, we can easily see that labeling $X'$ by $\hat{Y}'$ gives a lower bound guarantee $\ell > 0$ such that $\zeta(\eta(X')) \geq O(\ell)$. If

$$\min(P_X(B(X, \rho(X', X)), P_X(B(X', \rho(X', X))) \leq O(\ell^{d/\alpha}). \tag{7}$$

then, by using assumption (H2), we can obtain a lower bound guarantee on $\zeta(\eta(X))$ of the order of $O(\ell)$, as for $\zeta(\eta(X'))$. In this case, we can easily infer the label of $X$ by that of $X'$.
Given a current point $X$, the `Reliable` subroutine thus determines with high probability if there exists a previous informative point $X'$, which satisfies equation (7).

**Algorithm 1:** Multiclass $k$-NN Active Learning (`MKAL`)

**Input:** a pool $\mathcal{K}_{\mathcal{X}} = \{X_1, \ldots, X_w\}$, label budget $n$, smoothness parameters $(\alpha, L)$, margin noise parameters $(\beta, C)$, doubling constant parameter $C_{db}$, confidence parameter $\delta$, accuracy parameter $\epsilon$.

**Output:** 1-NN classifier $\widehat{f}_{n,w}$

1 $s = 1$ ▷ index of point currently examined
2 $\widehat{\mathcal{S}}^{(1)}, \widehat{\mathcal{S}}^{(2)} = \emptyset$ ▷ current active sets
3 $t = n$ ▷ current label budget
4 $I = \emptyset$ ▷ Set of informative points indices (used for theoretical proofs)
5 **while** $t > 0$ *and* $s < w$ **do**
6  Let $\delta_s = \frac{\delta}{32Ms^2}$
7  **if** $k(\epsilon, \delta_s) \leq t$ **then**
8  $T$=`Reliable`$(X_s, \delta_s, \alpha, L, C_{db}, \widehat{\mathcal{S}}^{(1)})$
9  **if** $T$=*False* **then**
10  $[\widehat{Y}_s, Q_s]$=`confidentLabel`$(X_s, k(\epsilon, \delta_s), \delta_s)$
11  $\widehat{\ell}_s = \zeta(\hat{\eta}(X_s)) - 2b_{\delta_s, |Q_s|}$
12  $t = t - |Q_s|$
13  $I = I \cup \{s\}$
14  **if** $\widehat{\ell}_s \geq 0.1 b_{\delta_s, |Q_s|}$ **then**
15  $\widehat{\mathcal{S}}^{(1)} = \widehat{\mathcal{S}}^{(1)} \cup \{(X_s, \widehat{Y}_s, \widehat{\ell}_s)\}$
16  **else**
17  $\widehat{\mathcal{S}}^{(2)} = \widehat{\mathcal{S}}^{(2)} \cup \{(X_s, \widehat{Y}_s, \widehat{\ell}_s)\}$
18  s=s+1
19 $\mathcal{S}^{(1)}_{ac} = \{(X_s, \widehat{Y}_s),\ (X_s, \widehat{Y}_s, \widehat{\ell}_s) \in \widehat{\mathcal{S}}^{(1)}\}$
20 $\mathcal{S}^{(2)}_{ac} = \{(X_s, \widehat{Y}_s),\ (X_s, \widehat{Y}_s, \widehat{\ell}_s) \in \widehat{\mathcal{S}}^{(2)}\}$
21 $\widehat{f}_{n,w} \leftarrow$ `1-NN` $(\mathcal{S}^{(1)}_{ac} \cup \mathcal{S}^{(2)}_{ac})$

**Input and output variables.** The `Reliable` subroutine takes as input the current point $X$, a confidence parameter $\delta$, the parameters related respectively to the assumptions (H2) and (H3), a set $\widehat{\mathcal{S}}$, which contains some points considered as informative before reaching $X$. Each element of $\widehat{\mathcal{S}}$ is a triplet, $(X, \hat{Y}, \ell)$, where $\hat{Y}$ is the inferred label of $X$ and $\ell$ is a lower bound guarantee on $\zeta(\eta(X))$. If `Reliable`$(X, \delta, \alpha, L, C_{db}, \widehat{\mathcal{S}})$ outputs `True`, the point $X$ is not considered to be informative. By convention, `Reliable`$(X, \delta, \alpha, L, C_{db}, \emptyset)$ always returns `False`. Note that the input $\hat{\mathcal{S}}$ corresponds to a subset of the current active set $\mathcal{S}_{ac}$. More precisely, it corresponds to the set of points $\mathcal{S}^{(1)}_{ac}$ defined in `MKAL` which contains the labeled points where we have obtained a higher guarantee on their inferred labels.

It is important to note that, outside the subroutine `Reliable`, The set $\mathcal{S}^{(1)}_{ac}$ is used with another one: $\mathcal{S}^{(2)}_{ac}$, together with which the represent the active set $\mathcal{S}_{ac}$. This is for reasons related to the proofs, as details in the supplementary Materials but, in practical situations, it is enough to only consider $\mathcal{S}_{ac}$ as $\mathcal{S}^{(1)}_{ac}$ in `MKAL`.

**Auxiliary subroutines.** Because $P_X$ appearing in (7) is unknown, it is impossible to use this expression directly in the `Reliable` subroutine. It can nevertheless be estimated with arbitrary precision and confidence using only unlabeled data from $\mathcal{K}$. For that purpose, we combine the `Reliable` subroutine with two auxiliary subroutines named `Estprob` and `BerEst`, inspired from Kontorovich et al. (2016).

The `Reliable` subroutine uses `EstProb`$(x, r, \epsilon_o, 50, \delta)$ which in turn uses the subroutine `BerEst`$(\epsilon_o, \delta, 50, p)$. The subroutine `BerEst` consists in adaptively estimating with high probability the expectation of a Bernoulli variable $Z \sim p$. The variables at the beginning of `BerEst` subroutine $(p_1, \ldots, p_4)$ are sampled at the beginning for theoretical analysis where we want a concentration inequality to hold for a number of samples greater than 4 (see Kontorovich et al. (2016); Maurer and Pontil (2009) for more details). Besides, the `EstProb`$(x, r, \epsilon_o, 50, \delta)$ subroutine uses a particular version of the `BerEst` subroutine where the Bernoulli variable corresponds to $p_i = 1_{X_i \in B(x,r)}$, with $X_i \in \mathcal{K}_{\mathcal{X}}$. It begins by setting the Bernoulli variable $p = 1_{X \in B(x,r)}$, with $X \in \mathcal{K}$, and then outputs an estimation of the probability-ball $P_X(B(x, r))$ which corresponds to the output of `BerEst`$(\epsilon_o, \delta, 50, p)$.
For the estimation of $P_X(B(x, r))$, instead of using the unlabeled points from $\mathcal{K}_{\mathcal{X}}$, we can also use a (large) separate set of unlabeled points independent of $\mathcal{K}_{\mathcal{X}}$.

### 4.2 ConfidentLabel subroutine

The `ConfidentLabel` subroutine aims at inferring the label of an informative point $X$ with some confidence. It is the main source of the advantage over passive learning, indeed, it adaptively finds the appropriate number of label requests for each informative point $X$, taking into account that :

- points with large margin $\zeta(\eta(X))$ require fewer label requests, as controlled by the cut-off condition in line 9.
- points with a small margin $\zeta(\eta(X))$ are too noisy and we should not waste too many label requests on those. The maximum number of label requests is controlled by the parameter $k'$ defined below.

**Algorithm 2:** `Reliable` subroutine

**Input:** an instance $X$, the confidence parameter $\delta$, the smoothness parameters $L$, $\alpha$, the doubling constant $C_{db}$, a set $\widehat{\mathcal{S}} \subset \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^+$

**Output:** $T$

**for** $(X', Y', \ell) \in \widehat{\mathcal{S}}$ **do**

  $\widehat{p}_{X'} =$ `EstProb` $\left(X', \rho(X, X'), \left(\frac{\ell}{64LC_{db}^3}\right)^{d/\alpha}, 50, \delta\right)$

  $\widehat{p}_{X} =$ `EstProb` $\left(X, \rho(X, X'), \left(\frac{\ell}{64LC_{db}^3}\right)^{d/\alpha}, 50, \delta\right)$

  **if** $\exists\ (X', Y, \ell) \in \widehat{\mathcal{S}}$ *with* ( $\widehat{p}_{X'} \leq \frac{75}{94}\left(\frac{\ell}{64LC_{db}^3}\right)^{d/\alpha}$ *OR* $\widehat{p}_X \leq \frac{75}{94}\left(\frac{\ell}{64LC_{db}^3}\right)^{d/\alpha}$ ) **then**

    $T = True$

  **else**

    $T = False$

**Algorithm 3:** `EstProb` subroutine

**Input:** an instance $x \in \mathcal{X}$, a positive number $r > 0$, an accuracy parameter $\epsilon_o$, an integer parameter $u$, a confidence parameter $\delta$

**Output:** $\widehat{p}_X$ ▷ An estimate of $P_X(B(x, r))$

Set $p \sim \mathbb{1}_{X \in B(x,r)}$ a Bernoulli variable, with $X \sim \mathcal{K}_{\mathcal{X}}$

$\widehat{p}_x =$ `BerEst`$(\epsilon_o, \delta, u, p)$ ▷ `BerEst` subroutine

**Input and output variables.** The input variables of the `confidentLabel` subroutine are the current point $X$, an integer $k'$, and a confidence parameter $\delta$. The parameter $k'$ represents the maximum number of label requests we are allowed to do (if the budget $t$ is large enough), and it is independent of $X$. Furthermore, it is computed such that all the $k'$-NN are at most at some distance from $X$. In that case the majority in expectation of the empirical majority of the $k'$-NN labels differs from $\eta(X)$ by less than some margin.

## 5 Theoretical properties of `MKAL`

Our main theoretical result is Theorem 5.1 below which provides guarantees about the statistical performance of `MKAL` and requires some additional notations. Let $\mathcal{A}_{a,w}$ be the set of active learning algorithms on $\mathcal{K}_{\mathcal{X}}$, and $\mathcal{P}(\alpha, \beta)$ the set of probabilities such that each element satisfies assumptions (H2) and (H1), and its marginal probability is a doubling-probability. For $A \in \mathcal{A}_{a,w}$, and $P \in \mathcal{P}(\alpha, \beta)$, we denote by $\widehat{f}_{A,n,w,P} := \widehat{f}_{n,w}$ the classifier that is provided by $A$, under an environment governed by the probability $P$.

**Algorithm 4:** `BerEst` subroutine (Bernoulli Estimation)

**Input:** accuracy parameter $\epsilon_o$, confidence parameter $\delta'$, budget parameter $u$. ▷ $u$ does not depend on the label budget $n$

$Z \sim p$ a Bernoulli variable from which we can sample.

**Output:** $\widehat{p}$

Sample $p_1, \ldots, p_4$ ▷ with respect to $\sim p$

$S = \{p_1, \ldots, p_4\}$

$K = \frac{4u}{\epsilon_o}\log(\frac{8u}{\delta'\epsilon_o})$

**for** $i = 3 : \log_2(u\log(2K/\delta')/\epsilon_o)$ **do**

  $m = 2^i$

  $S = S \cup \{p_{m/2+1}, \ldots, p_m\}$

  $\widehat{p} = \frac{1}{m}\sum_{j=1}^{m} p_j$

  **if** $\widehat{p} > u\log(2m/\delta')/m$ **then**

    Break

Output $\widehat{p}$

**Algorithm 5:** `confidentLabel` subroutine

**Input:** an instance $X$, integer $k'$, confidence parameter $\delta$.

**Output:** $\widehat{Y}$, $Q$

$Q = \emptyset$

$k = 1$

**while** $k \leq k'$ **do**

  Request the label $Y^{(k)}$ of $X^{(k)}$

  $Q = Q \cup \{(X^{(k)}, Y^{(k)})\}$

  **for** *i=1 to M* **do**

    $\hat{\eta}_i = \frac{1}{|Q|}\sum_{(X,Y)\in Q} \mathbb{1}_{Y=i}$

  $\hat{\eta} = (\hat{\eta}_1, \ldots, \hat{\eta}_M)$

  **if** $\zeta(\hat{\eta}) > 4b_{\delta,k}$ **then**

    Break ▷ cut-off condition with $b_{\delta,k}$ from (3.4)

  $k = k + 1$

$\widehat{Y} = \arg\max_{i \in \mathcal{Y}} \hat{\eta}_i$

### 5.1 Main idea

The main technicality of our result involves a suitable decomposition of the excess risk as follows: let $\hat{f}_{n,w}$ be the classifier provided by `MKAL`, and a parameter

$\Delta_0 > 0$ which will be chosen suitably in equation (9). We have:

$$R(\hat{f}_{n,w}) - R(f^*) = E_X(\eta_{f^*(X)}(X) - \eta_{\hat{f}_{n,w}(X)}(X))$$
$$= E_X((\eta_{f^*(X)}(X) - \eta_{\hat{f}_{n,w}(X)}(X))\mathbb{1}_{\zeta(\eta(X))>\Delta_0/2})$$
$$+ E_X((\eta_{f^*(X)}(X) - \eta_{\hat{f}_{n,w}(X))}(X)\mathbb{1}_{\zeta(\eta(X))\leq\Delta_0/2}).$$

Let us assume that, for $n$ large enough, we have with high probability:

- For all $x \in supp(P_X)$, the pointwise error satisfies:

$$\eta_{f^*(x)}(x) - \eta_{\hat{f}_{n,w}(x)}(x) \leq \Delta_0. \tag{8}$$

- For all $x \in supp(P_X)$ with large margin, that is $\zeta(\eta(X)) > \Delta_0/2$, we have $\hat{f}_{n,w}(x) = f^*(x)$.

In that case,

$$R(\hat{f}_{n,w}) - R(f^*) \leq \Delta_0 P_X(x,\ \zeta(\eta(X)) \leq \Delta_0/2)$$
$$\leq C\Delta_0^{\beta+1} \quad \text{by assumption (H1)}$$

By choosing

$$\Delta_0 = \max\left(\epsilon, \left(\frac{\epsilon}{C}\right)^{\frac{1}{\beta+1}}\right), \tag{9}$$

we obtain:

$$R(\hat{f}_{n,w}) - R(f^*) \leq \epsilon.$$

### 5.2 Preliminary result

The following Proposition gives a sufficient condition to obtain (8).

**Proposition 5.1.**
*Let $\hat{\eta}$ be the estimator provided by* `MKAL` *and $\hat{f}$ the corresponding classifier. Let us assume that there exists an event $A$ such that for a large budget $n$, we have in $A$: for all $x \in supp(P_X)$,*

$$\| \eta(x) - \hat{\eta}(x) \|_\infty \leq \frac{\Delta}{2}. \tag{10}$$

*for some $\Delta > 0$.*
*Then, in the same event $A$,*

$$\eta_{f^*(x)}(x) - \eta_{\hat{f}(x)}(x) \leq \Delta. \tag{11}$$

*Proof.*
Let us assume that, for $x \in supp(P_X)$,

$$\| \eta(x) - \hat{\eta}(x) \|_\infty \leq \frac{\Delta}{2} \tag{12}$$

and

$$\eta_{f^*(x)}(x) - \eta_{\hat{f}(x)}(x) > \Delta. \tag{13}$$

Because (12) holds, we have:

$$\begin{cases} |\eta_{f^*(x)}(x) - \hat{\eta}_{f^*(x)}(x)| \leq \frac{\Delta}{2} \\ |\eta_{\hat{f}(x)}(x) - \hat{\eta}_{\hat{f}(x)}(x)| \leq \frac{\Delta}{2} \end{cases}$$

Consequently, by (13), we have:

$$\Delta < \eta_{f^*(x)}(x) - \eta_{\hat{f}(x)}(x)$$
$$\leq \hat{\eta}_{f^*(x)}(x) + \frac{\Delta}{2} - (\hat{\eta}_{\hat{f}(x)}(x) - \frac{\Delta}{2})$$
$$\leq \underbrace{\hat{\eta}_{f^*(x)}(x) - \hat{\eta}_{\hat{f}(x)}(x)}_{\leq 0 \text{ by definition}} + \Delta$$
$$\leq \Delta,$$

which is a contradiction. And then, necessarily, equation (10) is sufficient for having (11). □

### 5.3 Main result

The following Theorem is the theoretical core of our algorithm and will be fully proven in the Appendix (see Supplementary Material).

**Theorem 5.1** (Label complexity for the `MKAL` algorithm)**.**
*Let us consider the set $\mathcal{P}(\alpha, \beta)$ defined above in Section 5.1 and such that $\alpha\beta < d$. Let $\epsilon$, $\delta \in (0, \frac{1}{2})$. Let $\Delta$ be the parameter defined by (17). Let $n$, $w \in \mathbb{N}$ the label budget and the number of unlabeled points, respectively.*

***If*** $n \geq \tau(\epsilon, \delta) \times \left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d-\alpha\beta}{\alpha(\beta+1)}}$
$$\times \left[2\log\left(\frac{32MT^2_{\epsilon,\delta}}{\delta}\right) + \log\log\left(\frac{512\sqrt{e}}{\Delta}\right)\right],$$
***and*** $w \geq \tilde{O}\left(\left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d}{\alpha(\beta+1)}}\right)$,
***and*** $w \geq \frac{400\log\left(\frac{12800w^2}{\delta(\frac{1}{64L}\bar{c}\phi_n)^{d/\alpha}}\right)}{(\frac{1}{64L}\bar{c}\phi_n)^{d/\alpha}}$,

*where $\tilde{O}$ hides logarithmic factors, $L$ appears in (H2), $\bar{c} = 0.1$, $\phi_n$ is defined by (18), $\tau(\epsilon, \delta)$, $T_{\epsilon,\delta}$ are polynomial on $\log(\frac{1}{\epsilon})$, $\log(\frac{1}{\delta})$,*

***then*** *with probability at least $1 - \delta$ we have for all $x \in supp(P_X)$,*

$$\| \eta(x) - \hat{\eta}(x) \|_\infty \leq \frac{\Delta}{2},$$

***and***

$$\sup_{P \in \mathcal{P}(\alpha,\beta)} \left[R(\widehat{f}_{n,w}) - R(f^*)\right] \leq \epsilon. \tag{14}$$

## 6 Numerical experiments

In this Section we present numerical experiments to illustrate the behavior of our MKAL algorithm. We repeated each experiment 10 times, and present additional results in the Supplementary Material. The implementation of the code of MKAL can be found on the following webpage: https://github.com/xsiebert/MKAL

### 6.1 Datasets

We start first with a couple of synthetic binary datasets and then consider multiclass ones. For each dataset, we generate 100000 points as a training set for the algorithm, and 30000 points as a test set. The points are equally distributed between classes. The confidence and accuracy parameters ($\delta$ and $\epsilon$, respectively) have both been set to 0.1 for the experiments presented here.

**Binary datasets** We generated several synthetic datasets such that the margin and regularity assumptions described in Section 3.2 are satisfied. Moreover, the parameters are chosen such that the classification problem is challenging.

The first dataset is generated from a two-dimensional uniform distribution on $(x_1, x_2) \in [-1, 1]^2$ with a regression function $\eta(x_1, x_2) = 0.5(1 + \sin(\frac{\pi}{2} x_2))$. This regression function is independent of $x_1$ and the optimal classification boundary is $x_2 = 0$.

The second dataset is similar to that used in Berlind and Urner (2015), and corresponds to $\eta(x_1, x_2) = \frac{1}{2}(1 + \sin(\frac{\pi}{2} x_1) \sin(\frac{\pi}{2} x_2))$. The optimal classification boundary is given by $x_1 = 0$ and $x_2 = 0$.

**Multiclass datasets** We generate points from a mixture model with $M > 2$ classes corresponding to isotropic gaussian distributions. The centers of the gaussians are chosen randomly and the data rescaled to fit in $[-1, 1]^2$. The standard deviation is set to 0.2 to create overlap between classes. Results with different settings are provided in the Supplementary Material.

### 6.2 Results and discussion

#### 6.2.1 Informative points selection

Figure 1 illustrates the selection of the informative points by our algorithm, both for the binary and the multiclass cases. The black dots indicate the points that have been considered as non-informative by the Reliable subroutine, because their label could be inferred by the points already available in the active set, as described in Section 4.1. Conversely, the yellow crosses indicate the points that have been considered informative by the same subroutine. The latter are for the most part located close to the class boundaries, where classification is indeed expected to be more difficult. In the case $M = 5$, the informative points are concentrated at the boundaries between more than two classes.

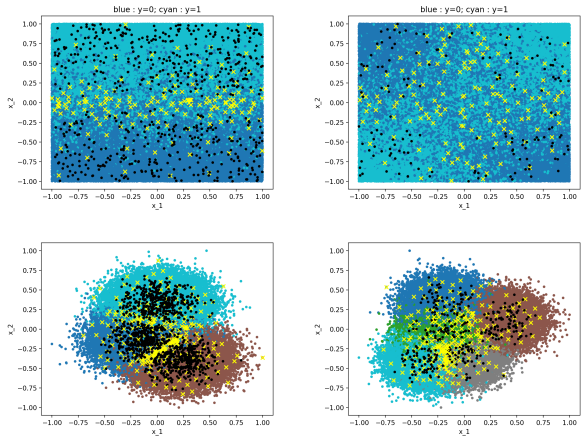


Figure 1: Selection of informative points by MKAL for two binary datasets (top) and two multiclass datasets (bottom), with the points colored according to their class label. The yellow crosses indicate the points that have been considered informative, and the black dots the points that have been considered as non-informative by the Reliable subroutine.

#### 6.2.2 Performance of MKAL

Figure 2 shows the performance of our active MKAL algorithm on the datasets of Figure 1, compared to the passive 1-NN counterpart, as well as a 5-NN passive classifier for comparison. As the number of labeled points increases, the MKAL algorithm reaches a point where it surpasses both passive nearest neighbors classifiers. This is due to the fact that the points used to construct the 1-NN classifier with MKAL are carefully chosen, as shown on Figure 1, so that they are all guaranteed to be informative by the Reliable subroutine. On the opposite, in passive learning the points are chosen randomly, and some of them are likely to be poor choices. The second binary dataset is more difficult to classify than the first one, and active learning with MKAL appears to be even more helpful in this case, as indicated by Figure 2. In the multiclass case, the advantage of MKAL, although significative, is less pronounced in the setting considered, because some classes are well separated. It becomes more pronounced as the noise level increases (see Supplementary Material).

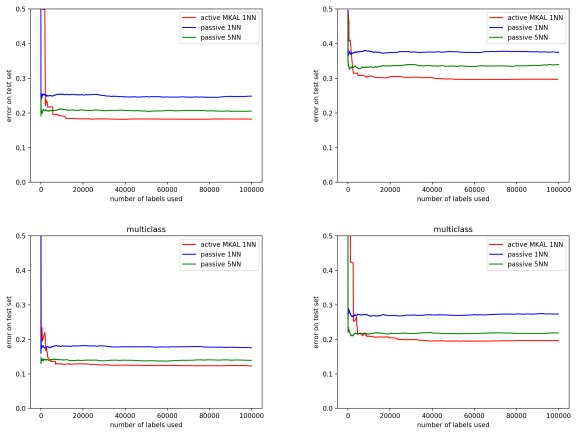


Figure 2: Comparison of the errors for the datasets of Figure 1 for the MKAL algorithm and the passive 1-NN counterpart, as well as a 5-NN passive learning. The error is computed on the test set and is plotted with respect to the number of points whose labels have been obtained.

## 7 Conclusions and perspectives

In this paper we present a novel algorithm for nonparametric active learning, which addresses two limitations of existing methods. First, we consider a more general smoothness assumption which takes into account a broader class of underlying distributions. Second, we extend the results from binary classification to a multiclass context. We prove the theoretical benefits of our algorithm and empirically illustrate it on several datasets.

Our results show that the careful selection of informative points by our algorithm allows the construction of a 1-NN classifier which is statistically consistent and provides better rate of convergence than passive counterparts. We believe that this rate is minimax according to the lower bound obtained in (Minsker, 2012) in the binary setting. A following step will consist in providing a similar rate of convergence in the case of cost-sensitive learning where a related work is due to (Krishnamurthy et al., 2017).

### Acknowledgements

The authors are deeply grateful to Christophe Denis and Mohamed Hebiri for their inspiring discussions. We also thank the anonymous reviewers for useful comments during the review process. We acknowledge the University of Mons for providing support of this work.

## References

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

Maria-Florina Balcan, Andrei Broder, and Tong Zhang. Margin based active learning. In *International Conference on Computational Learning Theory*, pages 35–50. Springer, 2007.

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

Maria-Florina Balcan, Steve Hanneke, and Jennifer Wortman Vaughan. The true sample complexity of active learning. *Machine learning*, 80(2-3): 111–139, 2010.

Christopher Berlind and Ruth Urner. Active nearest neighbors in changing environments. In *International Conference on Machine Learning*, pages 1870–1879, 2015.

Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. Importance weighted active learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 49–56, 2009.

Gérard Biau and Luc Devroye. *Lectures on the nearest neighbor method*. Springer, 2015.

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

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pages 3437–3445, 2014.

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

Sanjoy Dasgupta. Two faces of active learning. *Theoretical computer science*, 412(19):1767–1781, 2011.

Sanjoy Dasgupta, Daniel J Hsu, and Claire Monteleoni. *A general agnostic active learning algorithm*. Citeseer, 2007.

Luc Devroye, Laszlo Gyorfi, Adam Krzyzak, and Gábor Lugosi. On the strong universal consistency of nearest neighbor regression function estimates. *The Annals of Statistics*, pages 1371–1385, 1994.

Gerald A Edgar. Packing measure in general metric space. *Real Analysis Exchange*, 26(2):831–852, 2000.

László Györfi and Roi Weiss. Universal consistency and rates of convergence of multiclass prototype algorithms in metric spaces. *Journal of Machine Learning Research*, 22(151):1–25, 2021.

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

Steve Hanneke. Nonparametric active learning, part 1: Smooth regression functions. http://www.stevehanneke.com/, 12 2018.

Steve Hanneke and Liu Yang. Minimax analysis of active learning. *The Journal of Machine Learning Research*, 16(1):3487–3602, 2015.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

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. k-nn regression adapts to local intrinsic dimension. In *Advances in Neural Information Processing Systems*, pages 729–737, 2011.

Akshay Krishnamurthy, Alekh Agarwal, Tzu-Kuo Huang, Hal Daumé III, and John Langford. Active learning for cost-sensitive classification. In *International Conference on Machine Learning*, pages 1915–1924. PMLR, 2017.

Sanjeev R Kulkarni and Steven E Posner. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Transactions on Information Theory*, 41(4):1028–1039, 1995.

Andrea Locatelli, Alexandra Carpentier, and Samory Kpotufe. Adaptivity to noise parameters in nonparametric active learning. *Proceedings of Machine Learning Research vol*, 65:1–34, 2017.

Andrea Locatelli, Alexandra Carpentier, and Samory Kpotufe. An adaptive strategy for active learning with smooth decision boundary. In *Algorithmic Learning Theory*, pages 547–571, 2018.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

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

Wolfgang Mulzer. Five proofs of Chernoff's bound with applications. *arXiv preprint arXiv:1801.03365*, 2018.

Nikita Puchkin and Vladimir Spokoiny. An adaptive multiclass nearest neighbor classifier. *ESAIM: Probability and Statistics*, 24:69–99, 2020.

Henry WJ Reeve and Gavin Brown. Minimax rates for cost-sensitive learning on manifolds with approximate nearest neighbours. In *International Conference on Algorithmic Learning Theory*, pages 11–56, 2017.

Burr Settles. Active learning literature survey. *University of Wisconsin, Madison*, page 11, 2010.

Mathukumalli Vidyasagar. *Learning and generalisation: with applications to neural networks*. Springer Science & Business Media, 2013.

# Supplementary Materials

In this document we provide proofs that are too long to include in the main manuscript, as well as some additional results from computer simulations.

## A Missing proofs

This Section is organized as follows: in Section A.1, we introduce some additional notations. In Section A.2 we formally prove Theorem 5.1 from the main manuscript.

### A.1 Notations

Some notations that will be used throughout the proofs are listed below. Most of them were already introduced in the main document but are repeated here for convenience.

As defined in Section 3 of the main manuscript, let $B(x,r) = \{x' \in \mathcal{X},\ \rho(x,x') < r\}$ the open ball with respect to the Euclidean metric $\rho$, centered at $x \in \mathcal{X}$ with radius $r > 0$. Let $\text{supp}(P_X) = \{x \in \mathcal{X},\ \forall r > 0,\ P_X(B(x,r)) > 0\}$ the support of the marginal distribution $P_X$.

For $p \in (0,1]$, and $x \in supp(P_X)$, let us define

$$r_p(x) = \inf\{r > 0, P_X(B(x,r)) \geq p\}. \tag{15}$$

As in the main part of the manuscript, for a point $X_s \in \mathcal{K} = \{X_1, \ldots, X_w\}$, we denote by $X_s^{(k)}$ its $k$-th nearest neighbor in $\mathcal{K}$, and $Y_s^{(k)}$ the corresponding label.

For an integer $k \geq 1$, let

$$\widehat{\eta}_k(X_s) = (\widehat{\eta}_{k,1}(X_s), \ldots, \widehat{\eta}_{k,M}(X_s)) \quad , \quad \bar{\eta}_k(X_s) = (\bar{\eta}_{k,1}(X_s), \ldots, \bar{\eta}_{k,M}(X_s)),$$

where for $i \in \mathcal{Y}$,

$$\widehat{\eta}_{k,i}(X_s) = \frac{1}{k}\sum_{j=1}^{k} \mathbb{1}_{Y_s^{(j)}=i}\,, \quad \bar{\eta}_{k,i}(X_s) = \frac{1}{k}\sum_{j=1}^{k} \eta_i(X_s^{(j)}). \tag{16}$$

Let us also introduce some specific variables and notations that will be used throughout the remainder of this document.

For $\epsilon, \delta \in (0,1)$, $k \geq 1$, $c \geq 7.10^6$ and $C$ is defined in (H1) from the main manuscript:

$$b_{\delta,k} = \sqrt{\frac{2}{k}\left(\log\left(\frac{1}{\delta}\right) + \log\log\left(\frac{1}{\delta}\right) + \log\log(e.k)\right)},$$

$$k(\epsilon,\delta) = \frac{c}{\Delta^2}\left[\log(\frac{1}{\delta}) + \log\log(\frac{1}{\delta}) + \log\log\left(\frac{512\sqrt{e}}{\Delta}\right)\right],$$

$$\Delta = \max(\frac{\epsilon}{2}, \left(\frac{\epsilon}{C}\right)^{\frac{1}{\beta+1}}), \tag{17}$$

$$\phi_n = \sqrt{\frac{1}{n}\left(\log\left(\frac{1}{\delta}\right) + \log\log\left(\frac{1}{\delta}\right)\right)}. \tag{18}$$

### A.2 Detailed proof of Theorem 5.1

Let us restate Theorem 5.1 from the main manuscript:

**Theorem A.1** (Label complexity for the MKAL algorithm). *Let us consider the set $\mathcal{P}(\alpha, \beta)$ defined above in Section 5 and such that $\alpha\beta < d$. Let $\epsilon$, $\delta \in (0, \frac{1}{2})$. Let $\Delta$ be the parameter defined by (17). Let $n$, $w \in \mathbb{N}$ used in MKAL respectively as the label budget and the number on unlabeled points.*

$$\textbf{\textit{If}} \quad n \geq \tau(\epsilon, \delta) \times \left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d-\alpha\beta}{\alpha(\beta+1)}} \times \left[2\log\left(\frac{32MT_{\epsilon,\delta}^2}{\delta}\right) + \log\log\left(\frac{512\sqrt{e}}{\Delta}\right)\right], \tag{19}$$

$$\textbf{\textit{and}}\ w \geq \tilde{O}\left(\left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d}{\alpha(\beta+1)}}\right), \tag{20}$$

$$\textbf{\textit{and}} \quad w \geq \frac{400\log\left(\frac{12800w^2}{\delta(\frac{1}{64L}\bar{c}\phi_n)^{d/\alpha}}\right)}{(\frac{1}{64L}\bar{c}\phi_n)^{d/\alpha}}, \tag{21}$$

*where $\tilde{O}$ hides logarithmic factors, $L$ appears in (H2), $\bar{c} = 0.1$, $\phi_n$ is defined by (18), $\tau(\epsilon, \delta)$, $T_{\epsilon,\delta}$ are polynomial on $\log(\frac{1}{\epsilon})$, $\log(\frac{1}{\delta})$,*

***then*** *with probability at least $1 - \delta$ we have for all $x \in supp(P_X)$,*

$$\| \eta(x) - \hat{\eta}(x) \|_\infty \leq \frac{\Delta}{2} \tag{22}$$

***and***

$$\sup_{P \in \mathcal{P}(\alpha,\beta)} \left[R(\widehat{f}_{n,w}) - R(f^*)\right] \leq \epsilon. \tag{23}$$

The proof of the above theorem is organized as follows:

- In Section A.2.1, we introduce some technical lemmas mostly related to concentration inequalities, algebra results, etc... which will be very helpful.
- In Section A.2.2, we determine the number of label requests made in the neighbourhood of an informative point. Remarkably, when an informative point is relatively far from the decision boundary, its label can be determined (with high probability) by using only a small number of label requests.
- In Section A.2.4, by using some conditions on $w$ ((27), (21)), a particular condition on the last informative point (50), and the Lemma A.8 related to the estimation of the probability-ball and the informativeness of a point, we prove that with high probability, we have:
  - Each instance $x$ in the support of $P_X$ with large $\zeta(\eta(x))$ (greater than a specific threshold) is correctly labeled (with high probability) by the classifier provided by MKAL, thus the prediction error is only made on points with small $\zeta(\eta(x))$ (Theorem A.4).
- In Section A.2.5, we prove that with high probability, the Equation (19) is sufficient to have (50).
- In Section A.2.6, by combining results from Section A.2.4, Section (A.2.5) and Proposition 5.1, we prove that with high probability that, when the condition (19), (20), (21), hold, then the equations (22) and (23) also hold.

#### A.2.1 Technical Lemmas

**Lemma A.1** (Chernoff bounds). *(Mulzer, 2018)*
*Suppose $X_1, \ldots, X_m$ are independent random variables taking value in $\{0, 1\}$. Let $X$ denote their sum and $\mu = E(X)$ its expected value. Then,*

- *For any $\delta \in (0, 1)$,*
$$P_m(X \leq (1 - \delta)\mu) \leq \exp(-\delta^2\mu/2), \tag{24}$$
*where $P_m$ is the probability with respect to the sample $X_1, \ldots, X_m$.*

- *Additionally, for any $\delta' \geq 1$, we have:*

$$P_m(X \geq (1+\delta')\mu) \leq \exp(-\delta'\mu/4). \tag{25}$$

**Lemma A.2** (Hoeffding's inequality). *(Hoeffding, 1963)*

- ***First version**:*
  *Let $X$ a random variable with $E(X) = 0$, $a \leq X \leq b$, then for $v > 0$,*

$$E(e^{vX}) \leq e^{v^2(b-a)^2/8}.$$

- ***Second version**:*
  *Let $X_1, \ldots, X_m$ be independent random variables such that $-1 \leq X_i \leq 1$, $(i = 0, \ldots, m)$. The empirical mean of these variables is defined as*

$$\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i.$$

  *Then we have:*

$$P(|\bar{X} - E(\bar{X})| \geq t) \leq \exp(-mt^2/2).$$

**Lemma A.3** (Kontorovich *et al.*). *(Kontorovich et al., 2016)*
*Let $\delta' \in (0,1)$, $\epsilon_o > 0$, $t \geq 7$ and set $g(t) = 1 + \frac{8}{3t} + \sqrt{\frac{2}{t}}$. Let $p_1, p_2, \ldots \in \{0,1\}$ be i.i.d Bernoulli random variables with expectation $p$. Let $\widehat{p}$ be the output of `BerEst`$(\epsilon_o, \delta', t)$. There exists an event $A'$, such that $P(A') \geq 1 - \delta'$, and on $A'$, we have:*

1. *If $\widehat{p} \leq \frac{\epsilon_o}{g(t)}$ then $p \leq \epsilon_o$, otherwise, we have $p \geq \frac{2-g(t)}{g(t)}\epsilon_0$.*
2. *The number of random draws in the `BerEst` subroutine (Algorithm 4 in the main manuscript) is at most $\frac{8t\log(\frac{8t}{\delta'\psi})}{\psi}$, where $\psi := \max(\epsilon_o, \frac{p}{g(t)})$.*

**Lemma A.4** (Logarithmic relationship). *(Vidyasagar, 2013)*
*Suppose $a, b, c > 0$, $abe^{c/a} > 4\log_2(e)$, and $u \geq 1$. Then:*

$$u \geq 2c + 2a\log(ab) \Rightarrow\ u > c + a\log(bu).$$

**Lemma A.5** (Chaudhuri and Dasgupta). *(Chaudhuri and Dasgupta, 2014)*
*For $p \in (0,1]$, and $x \in supp(P_X)$, given $p \in (0,1]$, let $r_p(x)$ be defined in (15) For all , and $x \in supp(P_X)$, we have:*

$$P_X(B(x, r_p(x)) \geq p.$$

**Lemma A.6** (Kaufmann *et al.*). *(Kaufmann et al., 2016)*
*Let $\zeta(u) = \sum_{k\geq 1} k^{-u}$. Let $X_1, X_2, \ldots$ be independent random variables, identically distributed, such that, for all $v > 0$, $E(e^{vX_1}) \leq e^{v^2\sigma^2/2}$. For every positive integer $t$, let $S_t = X_1 + \ldots + X_t$. Then, for all $\gamma > 1$ and $r \geq \frac{8}{(e-1)^2}$:*

$$P\left(\bigcup_{t\in\mathbb{N}^*}\left\{|S_t| > \sqrt{2\sigma^2 t(r + \gamma\log\log(et))}\right\}\right) \leq \sqrt{e}\zeta(\gamma(1-\frac{1}{2r}))(\frac{\sqrt{r}}{2\sqrt{2}}+1)^\gamma\exp(-r).$$

**Lemma A.7.**
*Let $m \geq 1$ and $u \geq 20$. Then we have:*

$$m \geq 2u\log(\log(u)) \Longrightarrow m \geq u\log(\log(m)).$$

*Proof.*
Define $\phi(m) = m - u\log(\log(m))$, and let $m_0 = 2u\log(\log(u))$. We have:

$$\begin{aligned}\phi(m_0) &= 2u\log(\log(u)) - u(\log(\log(2u\log(\log(u))))) \\ &= 2u\log(\log(u)) - u\log(\log(2u) + \log(\log(\log(u))))\end{aligned}$$

It can be shown numerically that $\phi(m_0) \geq 0$ for $u \geq 20$.
Also, we have: $\phi'(m) = \frac{m\log(m)-u}{m\log(m)} \geq 0$ for all $m \geq m_0$ (notice that $m_0 \geq u$ for $u \geq 20$). Then it is easy to see that $\phi(m) \geq \phi(m_0)$ for all $m \geq m_0$. This establishes the lemma. □

### A.2.2 Label complexity on informative points

**Theorem A.2.**
*Let $\epsilon, \delta \in (0,1)$. Set $\Delta = \max(\epsilon, \left(\frac{\epsilon}{C}\right)^{\frac{1}{\beta+1}})$, and $p_\epsilon = \left(\frac{31\Delta}{1024L}\right)^{d/\alpha}$, where $\alpha$, $L$, $\beta$, $C$ are parameters used in (H1) and (H2) from the main manuscript.*
*For $p \in (0,1]$, and $x \in supp(P_X)$, let us introduce $r_p(x) = \inf\{r > 0, P_X(B(x,r)) \geq p\}$ and $k_s := k(\epsilon, \delta_s)$ defined in (A.1) (where $\delta_s = \frac{\delta}{32Ms^2}$).*
*For $k, s \geq 1$, set $\alpha_{k,s} = \sqrt{\frac{2}{k}\log(\frac{32s^2}{\delta})}$. There exists an event $A_1$ with probability at least $1 - \frac{\delta}{16}$, such that on $A_1$, for all $1 \leq s \leq w$, if*

$$k_s \leq (1 - \alpha_{k_s,s})p_\epsilon(w-1), \tag{26}$$

*then the $k_s$ nearest neighbors of $X_s$ (in the pool $\mathcal{K}_\mathcal{X}$) belong to the ball $B(X_s, r_{p_\epsilon}(X_s))$. Additionally, the condition*

$$w \geq 4\bar{C}\left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d}{\alpha(\beta+1)}}\left(\log\left(\frac{16384M\sqrt{e}}{\delta\epsilon}\right) + \log\left(4\bar{C}\left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d}{\alpha(\beta+1)}}\right)\right), \tag{27}$$

*where $\bar{C}$ is an absolute constant, is sufficient to have (26).*

*Proof.*
Fix $x \in supp(P_X)$. For $k \in \mathbb{N}$, let us denote $X_x^{(k)}$, the $k^{th}$ nearest neighbor of $x$ in the pool. We have

$$P(\rho(x, X_x^{(k_s+1)}) > r_{p_\epsilon}(x)) \leq P(\sum_{i=1}^{w} \mathbb{1}_{X_i \in B(x, r_{p_\epsilon}(x))} \leq k_s).$$

Then by using Lemma A.1 and Lemma A.5, and if $k_s$ satisfies (26), we have:

$$\begin{aligned}P(\rho(x, X_x^{(k_s+1)}) > r_{p_\epsilon}(x)) &\leq P(\sum_{i=1}^{w} \mathbb{1}_{X_i \in B(x, r_{p_\epsilon}(x))} \leq (1 - \alpha_{k_s,s})p_\epsilon(w-1)) \\ &\leq P\left(\sum_{i=1}^{w} \mathbb{1}_{X_i \in B(x, r_{p_\epsilon}(x))} \leq (1 - \alpha_{k_s,s})P_X(B(x, r_{p_\epsilon}(x)))(w-1)\right) \\ &\leq \exp(-\alpha_{k_s,s}^2(w-1)P_X(B(x, r_{p_\epsilon}(x))/2) \\ &\leq \exp(-\alpha_{k_s,s}^2(w-1)p_\epsilon/2) \\ &\leq \exp(-\alpha_{k_s,s}^2 k_s/2) \\ &\leq \exp(-\log(32s^2/\delta)) \\ &= \frac{\delta}{32s^2}.\end{aligned}$$

Fix $x = X_s$. Given $X_s$, there exists an event $A_{1,s}$, such that $P(A_{1,s}) \geq 1 - \delta/(32s^2)$, and on $A_{1,s}$, if

$$k_s \leq (1 - \alpha_{k_s,s})p_\epsilon(w-1), \tag{28}$$

we have $B(X_s, r_{p_\epsilon}(X_s)) \cap \{X_1, \ldots, X_w\} \geq k_s$. By setting $A_1 = \cap_{s\geq 1} A_{1,s}$, we have $P(A_1) \geq 1 - \delta/16$, and on $A_1$, for all $1 \leq s \leq w$, if $k_s \leq (1 - \alpha_{k_s,s})p_\epsilon(w-1)$, then $B(X_s, r_{p_\epsilon}(X_s)) \cap \{X_1, \ldots, X_w\} \geq k_s$.

Now, let us proof that the condition (27) is sufficient to guarantee (26).
The relation (26) implies

$$w \geq \frac{k_s}{(1-\alpha_{k_s,s})p_\epsilon} + 1. \tag{29}$$

We can see by a bit of calculus, that $\alpha_{k_s,s} \leq \frac{1}{2}$, and then

$$\begin{aligned}
\frac{k_s}{(1-\alpha_{k_s,s})p_\epsilon} + 1 &\leq \frac{2k_s}{p_\epsilon} + 1 \\
&\leq 4\frac{k_s}{p_\epsilon} \qquad \left(\text{because } \frac{k_s}{p_\epsilon} \geq 1\right) \\
&= \frac{4c}{p_\epsilon \Delta^2}\left[\log(\frac{32Ms^2}{\delta}) + \log\log(\frac{32Ms^2}{\delta}) + \log\log\left(\frac{512\sqrt{e}}{\Delta}\right)\right] \\
&= \frac{b}{\Delta^{2+\frac{d}{\alpha}}}\left[\log(\frac{32Ms^2}{\delta}) + \log\log(\frac{32Ms^2}{\delta}) + \log\log\left(\frac{512\sqrt{e}}{\Delta}\right)\right],
\end{aligned}$$

where $b = 4c\left(\frac{1024L}{31}\right)^{d/\alpha}$.

$$\begin{aligned}
\frac{k_s}{(1-\alpha_{k_s,s})p_\epsilon} + 1 &\leq \bar{C}\left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d}{\alpha(\beta+1)}}\left[\log(\frac{32Ms^2}{\delta}) + \log\log(\frac{32Ms^2}{\delta}) + \log\log\left(\frac{512\sqrt{e}}{\Delta}\right)\right] \\
&\qquad \text{as } \Delta = \max(\epsilon, \left(\frac{\epsilon}{C}\right)^{\frac{1}{\beta+1}}), \text{ where } \bar{C} = b(C)^{\frac{2\alpha+d}{\alpha(\beta+1)}} \\
&\leq \bar{C}\left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d}{\alpha(\beta+1)}}\left[2\log(\frac{32Ms^2}{\delta}) + \log\left(\frac{512\sqrt{e}}{\epsilon}\right)\right] \\
&\qquad \text{as} \log(x) \leq x, \text{ and } \Delta \geq \epsilon \\
&\leq 2\bar{C}\left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d}{\alpha(\beta+1)}}\left[\log(s^2) + \log\left(\frac{16384M\sqrt{e}}{\delta\epsilon}\right)\right] \\
&\leq 4\bar{C}\left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d}{\alpha(\beta+1)}}\left[\log(s) + \log\left(\frac{16384M\sqrt{e}}{\delta\epsilon}\right)\right] \\
&\leq 4\bar{C}\left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d}{\alpha(\beta+1)}}\left[\log(w) + \log\left(\frac{16384M\sqrt{e}}{\delta\epsilon}\right)\right].
\end{aligned}$$

We can now apply Lemma A.4, where we set

$$a = 4\bar{C}\left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d}{\alpha(\beta+1)}}, \quad c = 4\bar{C}\left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d}{\alpha(\beta+1)}}\log\left(\frac{16384M\sqrt{e}}{\delta\epsilon}\right), \quad b = 1.$$

We can easily notice that $c \geq a$, $a \geq 4$ and thus

$$abe^{c/a} \geq 4e > \log_2(e).$$

Then, the relation

$$w \geq 4\bar{C}\left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d}{\alpha(\beta+1)}}\left(\log\left(\frac{16384M\sqrt{e}}{\delta\epsilon}\right) + \log\left(4\bar{C}\left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d}{\alpha(\beta+1)}}\right)\right)$$

is sufficient to guarantee (28). □

Let us note that the guarantee obtained in the preceding theorem corresponds to that obtained in passive setting ($w = n$)Reeve and Brown (2017).

**Theorem A.3.**
*Let $\delta \in (0,1)$, and $\epsilon \in (0,1)$. Let us assume that $w$ satisfies (27). For $X_s$, set $\tilde{k}(\epsilon,\delta_s)$ (with $\delta_s = \frac{\delta}{32Ms^2}$) as*

$$\tilde{k}(\epsilon,\delta_s) = \frac{c}{16(\zeta(\eta(X_s))^2}\left[\log(\frac{32Ms^2}{\delta}) + \log\log(\frac{32Ms^2}{\delta}) + \log\log\left(\frac{256\sqrt{e}}{\eta_{f^*(X_s)}(X_s) - \max\limits_{j\neq f^*(X_s)}\eta_j(X_s)}\right)\right],$$

*where $c \geq 7.10^6$. For $k \geq 1$, $s \leq w$, let $\Delta = \max(\frac{\epsilon}{2}, (\frac{\epsilon}{C})^{\frac{1}{\beta+1}})$ and $b_{\delta_s,k}$ defined in (16).*
*Then, there exists an event $A_2$, such that $P(A_2) \geq 1-\delta/8$, and on $A_1 \cap A_2$, we have:*

1. *For $k \geq 1$, $\widehat{\eta}_k(X_s)$ and $\bar{\eta}_k(X_s)$ defined in (16), for all $s \in \{1,\dots,w\}$,*

$$\| \widehat{\eta}_k(X_s) - \bar{\eta}_k(X_s) \|_\infty \leq b_{\delta_s,k}. \tag{30}$$

2. *For all $s \leq w$, if $\zeta(\eta(X_s)) \geq \frac{1}{4}\Delta$, then, $\tilde{k}(\epsilon,\delta_s) \leq k(\epsilon,\delta_s)$, and the subroutine* `ConfidentLabel`$(X_s)$:=`ConfidentLabel`$(X_s, k(\epsilon,\delta_s),\delta_s)$ *uses at most $\tilde{k}(\epsilon,\delta_s)$ label requests. We also have*

$$\zeta(\hat{\eta}_{\bar{k}_s}(X_s)) := \hat{\eta}_{\bar{k}_s,\hat{Y}_s}(X_s) - \max_{j\neq\hat{Y}_s}\hat{\eta}_{\bar{k}_s,j}(X_s) \geq 4b_{\delta_s,\bar{k}_s} \tag{31}$$

*and*

$$\hat{Y}_s = f^*(X_s), \tag{32}$$

*where $\bar{k}_s$ is the number of requests made in* `ConfidentLabel`$(X_s)$, *$\hat{Y}_s$ is the output of the subroutine* `ConfidentLabel`$(X_s)$.

*Proof.*

1. Let us begin with the proof of the first part.
Here, we follow the proof of Theorem 8 in (Kaufmann et al., 2016), with few modifications.
Let $s \in \{1,\dots,w\}$. and $l \in \{1,\dots,M\}$. Set $S_{k,l} = \sum_{i=1}^{k}\left(\mathbb{1}_{Y_s^{(i)}=l} - \eta_l(X_s^{(i)})\right)$. Given $\{X_1,\dots,X_w\}$, $E\left(\mathbb{1}_{Y_s^{(i)}=l} - \eta_l(X_s^{(i)})\right) = 0$, and the random variables $\left\{\mathbb{1}_{Y_s^{(i)}=l} - \eta_l(X_s^{(i)}),\ i=1,\dots,k\right\}$ are independent. Then by Lemma A.2, given $\{X_1,\dots,X_w\}$, as $\mathbb{1}_{Y_s^{(i)}=l} - \eta_l(X_s^{(i)})$ takes values in $[-1,1]$, we have $E\left(\exp\left(v(\mathbb{1}_{Y_s^{(i)}=l} - \eta_l(X_s^{(i)}))\right)\right) \leq \exp(v^2/2)$ for all $v > 0$. Furthermore, set $z = \log(\frac{32Ms^2}{\delta})$, and $r = z + 3\log(z)$. We have $r \geq \frac{8}{(e-1)^2}$, and by Lemma A.6, with $\gamma = 3/2$, we have:

$$\begin{aligned} P\left(\bigcup_{k\in\mathbb{N}^*}\left\{|S_{k,l}| > \sqrt{2k(r+\gamma\log\log(ek))}\right\}\right) &\leq \sqrt{e}\zeta(3/2(1-\frac{1}{2r}))(\frac{\sqrt{r}}{2\sqrt{2}}+1)^{3/2}\exp(-r) \\ &= \frac{\sqrt{e}}{8}\zeta\left(\frac{3}{2} - \frac{3}{4(z+3\log(z))}\right)\frac{(\sqrt{z+3\log(z)}+\sqrt{8})^{3/2}}{z^3}\frac{\delta}{32Ms^2}. \end{aligned}$$

It can be shown numerically that for $z \geq 2.03$, which holds for all $\delta \in (0,1)$, $s \geq 1$,

$$\frac{\sqrt{e}}{8}\zeta\left(\frac{3}{2} - \frac{3}{4(z+3\log(z))}\right)\frac{(\sqrt{z+3\log(z)}+\sqrt{8})^{3/2}}{z^3} \leq 1.$$

Then, we have, given $s \in \{1,\dots,w\}$, there exists an event $A'_{2,s,l}$ such that $P(A'_{2,s,l}) \geq 1-\delta/32Ms^2$, and simultaneously for all $k \geq 1$, we have:

$$|S_{k,l}| \leq \sqrt{2k\left(\log\left(\frac{32Ms^2}{\delta}\right) + \log\log\left(\frac{32Ms^2}{\delta}\right) + \log\log(ek)\right)}.$$

By setting $A_2' = \cap_{l=1}^{M} \cap_{s\geq 1} A_{2,s,l}'$, we have $P(A_2') \geq 1 - \delta/16$, and on $A_2'$, we have for all $s \in \{1, \ldots, w\}$, for all $k \geq 1$, and $l \in \{1, \ldots, M\}$,

$$|\widehat{\eta}_{k,l}(X_s) - \bar{\eta}_{k,l}(X_s)| \ \leq b_{\delta_s,k}.$$

And then

$$\| \ \widehat{\eta}_k(X_s) - \bar{\eta}_k(X_s) \ \|_\infty \leq b_{\delta_s,k}.$$

2. For the proof of the second part of Theorem A.3, we are going to show that there exists an event $A_2''$ such that (31) and (32) hold on $A_2' \cap A_2'' \cap A_1$.
Given $\{X_1, \ldots, X_w\}$, and $X_s \in \{X_1, \ldots, X_w\}$, $l \in \{1, \ldots, M\}$, by Lemma A.2, there exists an event $A_{2,s,l}''$, with $P(A_{2,s,l}'') \geq 1 - \delta/32Ms^2$, and on $A_{2,s,l}''$, we have for all $k \geq 1$,

$$|\widehat{\eta}_{k,l}(X_s) - \bar{\eta}_{k,l}(X_s)| \leq \sqrt{\frac{2\log(\frac{32Ms^2}{\delta})}{k}}.$$

By setting $A_2'' = \cap_{l=1}^{M} \cap_{s\geq 1} A_{2,s,l}''$, we have $P(A_2'') \geq 1 - \delta/16$ and on $A_2''$, we have for all $k \geq 1$, $s \geq 1$, $l \in \{1, \ldots, M\}$,

$$|\widehat{\eta}_{k,l}(X_s) - \bar{\eta}_{k,l}(X_s)| \leq \sqrt{\frac{2\log(\frac{32Ms^2}{\delta})}{k}}. \tag{33}$$

On the event $A_1$, we have, for all $k \leq k_s$, $l \in \{1, \ldots, M\}$, by the $\alpha$-smoothness assumption (H2) from the main manuscript,

$$|\eta_l(X_s) - \bar{\eta}_{k,l}(X_s)| \leq \frac{31}{1024}\Delta. \tag{34}$$

On the event $A_2'' \cap A_1$, by (33), (34) we have simultaneously for all $k \geq 1$, $l \leq M$:

$$\begin{cases} \eta_l(X_s) \leq \bar{\eta}_{k,l}(X_s) + \frac{31}{1024}\Delta \\ \eta_l(X_s) \geq \bar{\eta}_{k,l}(X_s) - \frac{31}{1024}\Delta, \end{cases} \tag{35}$$

and

$$\begin{cases} \widehat{\eta}_{k,l}(X_s) - \bar{\eta}_{k,l}(X_s) \leq \sqrt{\frac{2\log(\frac{32Ms^2}{\delta})}{k}} \\ \widehat{\eta}_{k,l}(X_s) \geq \bar{\eta}_{k,l}(X_s) - \sqrt{\frac{2\log(\frac{32Ms^2}{\delta})}{k}}. \end{cases} \tag{36}$$

For $k \geq 1$, let $\hat{l}_{s,k}$ be defined as the index of the largest component of $\hat{\eta}_k(X_s)$, and $j \in \{1, \ldots, M\} \setminus \{\hat{l}_{s,k}\}$. We have:

$$\begin{aligned} \hat{\eta}_{k,\hat{l}_{s,k}}(X_s) - \hat{\eta}_{k,j}(X_s) &\overset{(1)}{\geq} \hat{\eta}_{k,f^*(X_s)}(X_s) - \hat{\eta}_{k,j}(X_s) \\ &\overset{(2)}{\geq} \bar{\eta}_{k,f^*(X_s)}(X_s) - \bar{\eta}_{k,j}(X_s) - 2\sqrt{\frac{2\log(\frac{32Ms^2}{\delta})}{k}} \\ &\overset{(3)}{\geq} \eta_{f^*(X_s)}(X_s) - \max_{i\neq f^*(X_s)} \eta_i(X_s) - 2\sqrt{\frac{2\log(\frac{32Ms^2}{\delta})}{k}} - \frac{62}{1024}\Delta + \max_{i\neq f^*(X_s)} \eta_i(X_s) - \eta_j(X_s) \end{aligned} \tag{37}$$

(1) is obtained by definition of $\hat{l}_{s,k}$, (2), (3) are obtained by respectively using (36), (35).
Let us introduce the following conditions on $k$:

$$k \geq \frac{1024}{(\eta_{f^*(X_s)}(X_s) - \max_{i\neq f^*(X_s)} \eta_i(X_s))^2} 162\log(\frac{32s^2M}{\delta}) \tag{38}$$

$$k \geq \frac{1024}{(\eta_{f^*(X_s)}(X_s) - \max_{i\neq f^*(X_s)} \eta_i(X_s))^2} 72\log\log(\frac{32s^2M}{\delta}) \tag{39}$$

$$k \geq 4\frac{73728e}{(\eta_{f^*(X_s)}(X_s) - \max\limits_{i\neq f^*(X_s)}\eta_i(X_s))^2}\log\log\left(\frac{\sqrt{73728e}}{\eta_{f^*(X_s)}(X_s) - \max\limits_{i\neq f^*(X_s)}\eta_i(X_s)}\right) \tag{40}$$

We will prove the following claim: if $k$ satisfies (38), (39), (40), and $\eta_{f^*(X_s)}(X_s) - \max\limits_{i\neq f^*(X_s)}\eta_i(X_s) \geq \frac{1}{4}\Delta$ then the right-hand side term in (37) is $\geq$ than $4b_{k,s}$. Additionally, it is easy to see that $\tilde{k}_{\epsilon,\delta_s}$ satisfies (38), (39), (40).

- Firstly, let us prove that $\max\limits_{i\neq f^*(X_s)}\eta_i(X_s) - \eta_j(X_s) \geq 0$ when $k$ satisfies (38), (39), (40). Let us assume that $k$ satisfies (38), (39), (40) and $\max\limits_{i\neq f^*(X_s)}\eta_i(X_s) - \eta_j(X_s) < 0$. In this case, by Lemma A.7, (by taking $m = ek$ and $u = \frac{73728e}{(\eta_{f^*(X_s)}(X_s) - \max\limits_{i\neq f^*(X_s)}\eta_i(X_s))^2}$ which imply $m \geq 1$ and $u \geq 20$) (40) leads to

$$k \geq \frac{1024}{(\eta_{f^*(X_s)}(X_s) - \max\limits_{i\neq f^*(X_s)}\eta_i(X_s))^2}\,72\log\log(ek). \tag{41}$$

  As $\max\limits_{i\neq f^*(X_s)}\eta_i(X_s) - \eta_j(X_s) < 0$, we necessarily have $j = f^*(X_s)$, and $\hat{l}_{s,k} \neq f^*(X_s)$ (by definition of $j$). Then:

$$\begin{aligned}
0 \leq \hat{\eta}_{k,\hat{l}_{s,k}}(X_s) - \hat{\eta}_{k,f^*(X_s)}(X_s) &\leq \bar{\eta}_{k,\hat{l}_{s,k}}(X_s) - \bar{\eta}_{k,f^*(X_s)}(X_s) + 2b_{k,\delta_s} \\
&\leq \eta_{\hat{l}_{s,k}}(X_s) - \eta_{f^*(X_s)}(X_s) + \frac{62}{1024}\Delta + 2b_{k,\delta_s} \\
&\leq \frac{62}{1024}\Delta - (\eta_{f^*(X_s)}(X_s) - \eta_{\hat{l}_{s,k}}(X_s)) + \frac{1}{2}(\eta_{f^*(X_s)}(X_s) - \eta_{\hat{l}_{s,k}}(X_s)) \\
&= \frac{62}{1024}\Delta - \frac{1}{2}(\eta_{f^*(X_s)}(X_s) - \eta_{\hat{l}_{s,k}}(X_s)) \\
&\leq \frac{62}{1024}\Delta - \frac{1}{8}\Delta \\
&< 0,
\end{aligned}$$

  which leads to a contradiction, then $\max\limits_{i\neq f^*(X_s)}\eta_i(X_s) - \eta_j(X_s) \geq 0$.

- Secondly, let us prove that

$$\eta_{f^*(X_s)}(X_s) - \max_{i\neq f^*(X_s)}\eta_i(X_s) - 2\sqrt{\frac{2\log(\frac{32Ms^2}{\delta})}{k}} - \frac{62}{1024}\Delta \geq 2b_{\delta_s,k},$$

  when $k$ satisfies (38), (39), (40). As previously, we can easily see (by using Lemma A.7) that if $k$ satisfies (38), (39), (40), then

$$k \geq \frac{1024}{(\eta_{f^*(X_s)}(X_s) - \max\limits_{i\neq f^*(X_s)}\eta_i(X_s))^2}\left[54\log(\frac{32s^2M}{\delta}) + 24.\log\log(\frac{32s^2M}{\delta}) + 24.\log\log(ek)\right]. \tag{42}$$

As $\eta_{f^*(X_s)}(X_s) - \max_{i \neq f^*(X_s)} \eta_i(X_s) \geq \frac{1}{4}\Delta$, we obtain:

$$\begin{aligned}(\eta_{f^*(X_s)}(X_s) - \max_{i \neq f^*(X_s)} \eta_i(X_s)) - 2\sqrt{\frac{2\log(\frac{32Ms^2}{\delta})}{k}} - \frac{62}{1024}\Delta &\geq (1 - \frac{248}{1024})(\eta_{f^*(X_s)}(X_s) - \max_{i \neq f^*(X_s)} \eta_i(X_s)) \\ &\quad - 2\sqrt{\frac{2\log(\frac{32Ms^2}{\delta})}{k}} \\ &\geq \frac{776}{1024}(\eta_{f^*(X_s)}(X_s) - \max_{i \neq f^*(X_s)} \eta_i(X_s)) \\ &\quad - 2\sqrt{\frac{2\log(\frac{32Ms^2}{\delta})}{k}} \\ &\geq 4b_{\delta_s,k} \quad \text{by using (42)}\end{aligned}$$

- Finally, we can easily see that $\tilde{k}(\epsilon, \delta_s)$ satisfies (38), (39), (40). Additionally, when $\zeta(\eta(X_s)) = \eta_{f^*(X_s)}(X_s) - \max_{i \neq f^*(X_s)} \eta_i(X_s) \geq \frac{1}{4}\Delta$, we have $\tilde{k}(\epsilon, \delta_s) \leq k(\epsilon, \delta_s)$.

Consequently, when $\zeta(\eta(X_s)) \geq \frac{1}{4}\Delta$, , the subroutine `ConfidentLabel`$(X_s)$ ends after at most $\tilde{k}(\epsilon, \delta_s)$ label requests. Let us denote by $\bar{k}_s = k(\epsilon, \delta_s)$ the minimum value of $k$ that satisfies:

$$\zeta(\hat{\eta}_k(X_s)) = \hat{\eta}_{k,\hat{l}_{s,k}}(X_s) - \max_{i \neq \hat{l}_{s,k}} \hat{\eta}_{k,i}(X_s) \geq 4b_{\delta_s,k}. \tag{43}$$

Let us prove that on $A_2' \cap A_2'' \cap A_1$, $f^*(X_s) = \hat{l}_{s,\bar{k}_s}$.
If $f^*(X_s) \neq \hat{l}_{s,\bar{k}_s}$, then we have:

$$\begin{aligned}\eta_{\hat{l}_{s,\bar{k}_s}}(X_s) - \max_{i \neq f^*(X_s)} \eta_i(X_s) &\geq \eta_{\hat{l}_{s,\bar{k}_s}}(X_s) - \eta_{f^*(X_s)}(X_s) + \frac{1}{4}\Delta \\ &\geq \bar{\eta}_{\bar{k}_s,\hat{l}_{s,\bar{k}_s}}(X_s) - \bar{\eta}_{\bar{k}_s,f^*(X_s)}(X_s) - \frac{62}{1024}\Delta + \frac{1}{4}\Delta \quad \text{by the smoothness assumption} \\ &\geq \hat{\eta}_{\bar{k}_s,\hat{l}_{s,\bar{k}_s}}(X_s) - \hat{\eta}_{\bar{k}_s,f^*(X_s)}(X_s) + \frac{97}{512}\Delta - 2b_{\delta_s,\bar{k}_s} \quad \text{by (30)} \\ &\geq \hat{\eta}_{\bar{k}_s,\hat{l}_{s,\bar{k}_s}}(X_s) - \max_{i \neq \hat{l}_{s,\bar{k}_s}} \hat{\eta}_{\bar{k}_s,i} + \frac{97}{512}\Delta - 2b_{\delta_s,\bar{k}_s} \qquad \text{because } f^*(X_s) \neq \hat{l}_{s,\bar{k}_s} \\ &\geq \frac{97}{512}\Delta > 0 \quad \textit{by (43)}.\end{aligned}$$

This contradicts the fact that $f^*(X_s) \neq \hat{l}_{s,\bar{k}_s}$, then we necessarily have $f^*(X_s) = \hat{l}_{s,\bar{k}_s}$. By setting $A_2 = A_2' \cap A_2''$, we have $P(A_2) \geq 1 - \delta/8$ and on $A_1 \cap A_2$, the item 1 and item 2 hold simultaneously.

□

### A.2.3 Sufficient condition to be a non-informative point

**Lemma A.8.**
*Let $\epsilon$, $\delta \in (0,1)$, $r > 0$. Let us assume that $w$ satisfies (21).*

*There exists an event $A_3$, such that $P(A_3) \geq 1 - \delta/16$, we have, on $A_3$, for all $s \leq w$:*
*If there exists $1 \leq s' < s$, such that $X_{s'}$ is an informative point, and $(X_{s'}, \widehat{Y}_{s'}, \widehat{\ell}_{s'}) \in \widehat{S}^{(1)}$ (the current set $\widehat{S}^{(1)}$ just before attaining $X_s$ defined in `MKAL` (Algorithm(1))), and that satisfies:*

$$\left(\widehat{p}_{X_{s'}} \leq \frac{75}{94}\left(\frac{1}{64LC_{db}^3}\widehat{\ell}_{s'}\right)^{d/\alpha} \text{ or } \widehat{p}_{X_s} \leq \frac{75}{94}\left(\frac{1}{64LC_{db}^3}\widehat{\ell}_{s'}\right)^{d/\alpha}\right), \tag{44}$$

*where*

$$\widehat{p}_{X_{s'}} := \texttt{Estprob}(X_{s'}, \rho(X_s, X_{s'}), \left(\frac{1}{64LC_{db}^3}\widehat{\ell}_{s'}\right)^{d/\alpha}, 50, \delta_s)$$

*and*

$$\widehat{p}_{X_s} := \texttt{Estprob}(X_s, \rho(X_s, X_{s'}), \left(\frac{1}{64LC_{db}^3}\widehat{\ell}_{s'}\right)^{d/\alpha}, 50, \delta_s);$$

*then*

$$\min(P_X(B(X_s, \rho(X_{s'}, X_s))), P_X(B(X_{s'}, \rho(X_{s'}, X_s)))) \leq \left(\frac{1}{64LC_{db}^3}\widehat{\ell}_{s'}\right)^{d/\alpha}. \tag{45}$$

*Otherwise, if (44) does not hold, i.e:*

$$\min(\widehat{p}_{X_{s'}}, \widehat{p}_{X_s}) > \frac{75}{94}\left(\frac{1}{64LC_{db}^3}\widehat{\ell}_{s'}\right)^{d/\alpha},$$

*then*

$$\min(P_X(B(X_s, \rho(X_{s'}, X_s))), P_X(B(X_{s'}, \rho(X_{s'}, X_s)))) \geq \frac{28}{47}\left(\frac{1}{64LC_{db}^3}\widehat{\ell}_{s'}\right)^{d/\alpha}. \tag{46}$$

*Proof.*
By following the scheme of subroutine `Estprob`, this Lemma is a direct application of Lemma A.3 by taking for all $s \leq w$, $t = 50$, $\epsilon_o = \left(\frac{1}{64LC_{db}^3}\widehat{\ell}_{s'}\right)^{d/\alpha}$, $\delta' = \delta/32s^2$, $r = \rho(X_s, X_{s'})$, $A_{3,s} := A'$. And then, if we set $A_3 = \cap_{s\geq 1} A_{3,s}$, we have $P(A_3) \geq 1 - \delta/16$, and on the event $A_3$, we can easily deduce (45) and (46) in each case.
On the other hand, for all $s \leq w$, the number of draws in $\texttt{Estprob}(X_s, \rho(X_s, X_{s'}), \left(\frac{1}{64LC_{db}^3}\widehat{\ell}_{s'}\right)^{d/\alpha}, 50, \delta_s)$ (respectively $\texttt{Estprob}(X_{s'}, \rho(X_s, X_{s'}), \left(\frac{1}{64LC_{db}^3}\widehat{\ell}_{s'}\right)^{d/\alpha}, 50, \delta_s)$) is always lower than $w$. Indeed, by Lemma A.3, the number of draws is at most:

$$N := \frac{400\log(\frac{12800s^2}{\delta\psi})}{\psi} \quad \text{where} \quad \psi = \max((\frac{1}{64LC_{db}^3}\widehat{\ell}_{s'})^{d/\alpha}, \frac{75}{94}P_X(B(X_s, \rho(X_s, X_{s'})))).$$

Then we have:

$$\begin{aligned}
N &\leq \frac{400\log\left(\frac{12800s^2}{\delta(\frac{1}{64LC_{db}^3}\widehat{\ell}_{s'})^{d/\alpha}}\right)}{(\frac{1}{64LC_{db}^3}\widehat{\ell}_{s'})^{\frac{d}{\alpha}}} \\
&\leq \frac{400\log\left(\frac{12800s^2}{\delta(\frac{1}{64LC_{db}^3}\bar{c}b_{\delta_{s'},|Q_{s'}|})^{d/\alpha}}\right)}{(\frac{1}{64LC_{db}^3}\bar{c}b_{\delta_{s'},|Q_{s'}|})^{d/\alpha}} \qquad (\text{as } \widehat{\ell}_{s'} \geq \bar{c}b_{\delta_{s'},|Q_{s'}|}, \text{ with } \bar{c} = 0.1) \\
&\leq \frac{400\log\left(\frac{12800w^2}{\delta(\frac{1}{64LC_{db}^3}\bar{c}\phi_n)^{d/\alpha}}\right)}{(\frac{1}{64LC_{db}^3}\bar{c}\phi_n)^{d/\alpha}} \qquad (\text{we can easily see that } b_{\delta_{s'},|Q_{s'}|} \geq \phi_n) \\
&\leq w \qquad (\text{by (21)}).
\end{aligned} \tag{47}$$

In equation (47), $b_{\delta_{s'},|Q_{s'}|}$ is defined by (A.1), and $|Q_{s'}|$ represents the number of label requests used in the subroutine `ConfidentLabel` (Algorithm 5) at the stage $s'$. □

### A.2.4 Label the instance space

**Theorem A.4.**
*Let $\epsilon$, $\delta \in (0,1)$. Let*

$$T_{\epsilon,\delta} = \frac{1}{\tilde{p}_\epsilon}\log(\frac{8}{\delta}),\ \textit{and}\ \tilde{p}_\epsilon = \left(\frac{\Delta}{128LC_{db}^3}\right)^{d/\alpha},\ \textit{with}\ \Delta = \max(\frac{\epsilon}{2}, \left(\frac{\epsilon}{C}\right)^{\frac{1}{\beta+1}}). \tag{48}$$

*Let $I$ the set of indexes of informative points used in `MKAL` (Algorithm 1 in the main document). Let us consider its last update in `MKAL` and also denoted it by $I$.*
*Then, set $s_I = \max I$ the index of the last informative point. Let $\widehat{S}_{ac} = \widehat{S}_{ac}^{(1)} \cup \widehat{S}_{ac}^{(2)}$ be the active set obtained in `MKAL` and denote by $\widehat{f}_{n,w}$ the output `1-NN`$(\widehat{S}_{ac})$. There exists an event $A_4$ such that $P(A_4) \geq 1-\delta/8$, and on $A_1 \cap A_2 \cap A_3 \cap A_4$, we have*

1. $$\sup_{x \in supp(P_X)} \min_{\bar{X} \in \{X_1,\ldots,X_{T_{\epsilon,\delta}}\}} P_X(B(x, \rho(\bar{X}, x))) \leq \tilde{p}_\epsilon. \tag{49}$$

2. *If $w$ satisfies (27) and (21) and the following condition holds*

$$s_I \geq T_{\epsilon,\delta}, \tag{50}$$

*then, for all $x \in supp(P_X)$ such that $\zeta(\eta(x)) > \frac{1}{2}\Delta$, there exists $s := s(x) \in I$ such that:*

$$\zeta(\eta(X_s)) \geq \frac{1}{4}\Delta \tag{51}$$

*and*

$$f^*(x) = f^*(X_s). \tag{52}$$

*In addition, we have*

$$\widehat{f}_{n,w}(x) = f^*(x). \tag{53}$$

*Proof.*
This proof is based on results from (Hanneke, 2018), (Reeve and Brown, 2017) with some additional modifications.

1. Let us begin by proving the first part of Theorem A.4.
For $x \in \text{supp}(P_X)$, let us introduce

$$r_{\tilde{p}_\epsilon}(x) = \inf\{r > 0,\ P_X(B(x,r)) \geq \tilde{p}_\epsilon\}.$$

By Lemma A.5, we have $P_X(B(x, r_{\tilde{p}_\epsilon}(x)) \geq \tilde{p}_\epsilon$. Then each $\bar{X} \in \{X_1,\ldots,X_{T_{\epsilon,\delta}}\}$ belongs to $B(x, r_{\tilde{p}_\epsilon}(x))$ with probability at least $\tilde{p}_\epsilon$. If we denote $\widehat{P}$ the probability over the data, we have:

$$\begin{aligned}
&\widehat{P}(\exists \bar{X} \in \{X_1,\ldots,X_{T_{\epsilon,\delta}}\},\ P_X(B(x,\rho(x,\bar{X})) \leq \tilde{p}_\epsilon) \\
&= 1 - \widehat{P}(\forall \bar{X} \in \{X_1,\ldots,X_{T_{\epsilon,\delta}}\},\ P_X(B(x,\rho(x,\bar{X})) > \tilde{p}_\epsilon) \\
&= 1 - \prod_{i=1}^{T_{\epsilon,\delta}} \widehat{P}(P_X(B(x,\rho(x,X_i)) > \tilde{p}_\epsilon) \\
&\geq 1 - \prod_{i=1}^{T_{\epsilon,\delta}} \widehat{P}(\rho(x,X_i) > r_{\tilde{p}_\epsilon}(x)) \\
&= 1 - \prod_{i=1}^{T_{\epsilon,\delta}} (1 - \widehat{P}(\rho(x,X_i) \leq r_{\tilde{p}_\epsilon}(x))) \\
&\geq 1 - (1-\tilde{p}_\epsilon)^{T_{\epsilon,\delta}} \\
&\geq 1 - \exp(-T_{\epsilon,\delta}\tilde{p}_\epsilon) \\
&= 1 - \delta/8.
\end{aligned}$$

Then, there exists an event $A_4$, such that $P(A_4) \geq 1 - \delta/8$ and (49) holds on $A_4$. Thus we can easily conclude the first part.

2. For the second part of Theorem A.4, let $x \in \text{supp}(P_X)$. By (49), on $A_4$ there exists $X_x \in \{X_1, \ldots, X_{T_{\epsilon,\delta}}\}$ such that:

$$P_X(B(x, \rho(X_x, x))) \leq \tilde{p}_\epsilon. \tag{54}$$

Let us consider the following inequality, which will be very useful: let $y_1, y_2 \in \mathcal{Y}$, for all $t, z \in \mathcal{X}$, we have:

$$|\eta_{y_1}(t) - \eta_{y_2}(t) + \eta_{y_2}(z) - \eta_{y_1}(z)| \leq 2 \parallel \eta(t) - \eta(z) \parallel_\infty . \tag{55}$$

By assumption (H2) from the main manuscript and equation (54), we have:

$$\parallel \eta(x) - \eta(X_x) \parallel_\infty \leq \frac{1}{128}\Delta. \tag{56}$$

By applying (55) with $\eta(x)$ and $\eta(X_x)$, altogether, using (56) and the fact that $\zeta(\eta(x)) \geq \frac{1}{2}\Delta$ we have:

$$\forall\, y \neq f^*(x), \begin{cases} \eta_{f^*(x)}(X_x) - \eta_y(X_x) \leq (\frac{1}{2} + \frac{1}{32})(\eta_{f^*(x)}(x) - \eta_y(x)) \\ (\frac{1}{2} - \frac{1}{32})\Delta \leq \eta_{f^*(x)}(X_x) - \eta_y(X_x). \end{cases} \tag{57}$$

Because $s_I \geq T_{\epsilon,\delta}$, there exists $s'$ such that $X_x := X_{s'}$ and $X_{s'}$ passes through the subroutine `Reliable`.

We have to consider two cases:

a) $X_{s'}$ **is uninformative**. Then there exists $s < s'$, such that $X_s \in \widehat{S}_{ac}^{(1)}$, and

$$\min(\widehat{p}_{X_s}, \widehat{p}_{X_{s'}}) \leq \frac{75}{94}\left(\frac{1}{64LC_{db}^3}\widehat{\ell}_s\right)^{d/\alpha},$$

where $\widehat{p}_{X_s} := \texttt{Estprob}(X_s, \rho(X_s, X_{s'}), \left(\frac{1}{64LC_{db}^3}\widehat{\ell}_s\right)^{d/\alpha}, 50, \delta_s\ )$, and
$\widehat{p}_{X_{s'}} := \texttt{Estprob}(X_{s'}, \rho(X_s, X_{s'}), \left(\frac{1}{64LC_{db}^3}\widehat{\ell}_s\right)^{d/\alpha}, 50, \delta_s)$. Then by Lemma A.8,

$$\min(P_X(B(X_s, \rho(X_s, X_{s'}))), P_X(B(X_{s'}, \rho(X_s, X_{s'})))) \leq \left(\frac{1}{64LC_{db}^3}\widehat{\ell}_s\right)^{d/\alpha}. \tag{58}$$

Necessary, we have $\eta_{f^*(x)}(X_s) - \max\limits_{j \neq f^*(x)} \eta_j(X_s) \geq \frac{1}{4}\Delta$.

Indeed, if

$$\eta_{f^*(x)}(X_s) - \max_{j \neq f^*(x)} \eta_j(X_s) < \frac{1}{4}\Delta, \tag{59}$$

then on $A_1 \cap A_2 \cap A_3$, by denoting $\bar{k}_s$ the number of request labels in $\texttt{ConfidentLabel}(X_s) := \texttt{ConfidentLabel}(X_s, k(\epsilon, \delta_s), \delta_s)$, we have:

$$\begin{aligned} \parallel \eta(X_x) - \eta(X_s) \parallel_\infty &\leq \frac{1}{64}(\widehat{\ell}_s) \\ &= \frac{1}{64}(\zeta(\hat{\eta}_{\bar{k}}(X_s)) - 2b_{\bar{k}_s,\delta_s}). \end{aligned}$$

Let $\hat{i}_s = \arg\max\limits_{j \in \mathcal{Y}} \hat{\eta}_{\bar{k}_s,j}(X_s)$, then:

$$\begin{aligned} \parallel \eta(X_x) - \eta(X_s) \parallel_\infty &\leq \frac{1}{64}(\hat{\eta}_{\bar{k}_s,\hat{i}_s}(X_s) - \max_{j \neq \hat{i}_s} \hat{\eta}_{\bar{k}_s,j}(X_s) - 2b_{\bar{k}_s}) \\ &\leq \frac{1}{64}(\bar{\eta}_{\bar{k}_s,\hat{i}_s}(X_s) - \max_{j \neq \hat{i}_s} \bar{\eta}_{\bar{k}_s,j}(X_s)) \quad \text{By (30)} \\ &\leq \frac{1}{64}(\eta_{\hat{i}_s}(X_s) - \max_{j \neq \hat{i}_s} \eta_j(X_s)) + \frac{62}{64 * 1024}\Delta \quad \text{by assumption (H2).} \end{aligned}$$

- If $\eta_{\hat{i}_s}(X_s) - \max_{j \neq \hat{i}_s} \eta_j(X_s) \leq 1/4\Delta$, then by using (57) and applying (55) with $\eta(X_s)$ and $\eta(X_x)$, we have for all $y \neq f^*(x)$,
$$\eta_{f^*(x)}(X_s) - \eta_y(X_s) \geq 0.459\Delta > \frac{1}{4}\Delta,$$
which contradicts (59).
- If $\eta_{\hat{i}_s}(X_s) - \max_{j \neq \hat{i}_s} \eta_j(X_s) > 1/4\Delta$, then by applying (55) with $\eta(X_x)$ and $\eta(X_s)$, we have
$$\eta_{\hat{i}_s}(X_x) - \max_{j \neq \hat{i}_s} \eta_j(X_x) > \frac{1}{4}(1 - \frac{1}{32} - \frac{62}{8 * 1024})\Delta.$$
Then altogether with (57), we have $\hat{i}_s = f^*(X_x) = f^*(x)$ and consequently $\eta_{f^*(x)}(X_s) - \max_{j \neq f^*(x)} \eta_j(X_s) > 1/4\Delta$ which contradicts (59).

Finally, we have $\eta_{f^*(x)}(X_s) - \max_{j \neq f^*(x)} \eta_j(X_s) \geq \frac{1}{4}\Delta$ and then $f^*(x) = f^*(X_s)$.

b) $X_x$ **is informative**. In this case, $s = s'$ and then by using (57) we always obtain (51) and (52).

Now, let us prove (53).
Let $X_x^{(1)}$ be the nearest neighbor of $x$ in $\widehat{S}_{ac} = \widehat{S}_{ac}^{(1)} \cup \widehat{S}_{ac}^{(2)}$. Additionally, let $X_x$ which satisfies (54). Without loss of generality, let us suppose that $X_x$ is uninformative (does not pass through `Reliable` subroutine), otherwise it is easy to obtain (53). In this case, we have previously seen that there exist $s < s' \leq s_{\epsilon,\delta}$ such that $X_x = X_{s'}$ and (58) holds. Without loss of generality, we can assume that:

$$P_X(B(X_s, \rho(X_s, X_{s'}))) = \min(P_X(B(X_s, \rho(X_s, X_{s'}))), P_X(B(X_{s'}, \rho(X_s, X_{s'})))). \tag{60}$$

Otherwise, we can use the relation:

$$P_X(B(X_{s'}, \rho(X_s, X_{s'}))) \leq P_X(B(X_s, 2\rho(X_s, X_{s'}))) \leq C_{db} P_X(B(X_s, \rho(X_s, X_{s'}))).$$

By the smoothness assumption, we have:

$$\begin{aligned} \| \eta(x) - \eta(X_x^{(1)}) \|_\infty &\leq L.P_X(B(x, \rho(x, X_s)))^{\alpha/d} \\ &\leq L.P_X(B(x, \rho(x, X_s)))^{\alpha/d} \\ &\leq C_{db}^3 L.P_X(B(x, \frac{1}{8}\rho(x, X_s)))^{\alpha/d}. \end{aligned} \tag{61}$$

Without loss of generality, we assume that $x \neq X_s$, otherwise (53) obviously holds. By (49), we have:

$$\rho(x, X_{s'}) \leq r_{\tilde{p}_\epsilon}(x). \tag{62}$$

Additionally, using (58),(60), (51), (52), (30), we have:

$$\begin{aligned} L.P_X(B(X_s, \rho(X_s, X_{s'})))^{\alpha/d} &\leq \frac{1}{64C_{db}^3}\widehat{\ell}_s \\ &\leq \frac{1}{64C_{db}^3}(\zeta(\eta(X_s)) + \frac{62}{1024}\Delta) \\ &\leq \frac{1}{64C_{db}^3}(\eta_{f^*(x)}(X_s) - \max_{j \neq f^*(x)} \eta_j(X_s) + \frac{62}{1024}\Delta). \end{aligned} \tag{63}$$

However, using (55), we have for all $j \neq f^*(x)$,

$$\begin{aligned} \eta_{f^*(x)}(X_s) - \eta_j(X_s) &\leq 2L.P_X(B(X_s, \rho(X_s, X_{s'})))^{\alpha/d} + \eta_{f^*(x)}(X_{s'}) - \eta_j(X_{s'}) \\ &\leq \frac{1}{32}(\eta_{f^*(x)}(X_s) - \eta_j(X_s) + \frac{62}{1024}\Delta) + \eta_{f^*(x)}(X_{s'}) - \eta_j(X_{s'}) \\ &\leq \frac{1}{32}(\eta_{f^*(x)}(X_s) - \eta_j(X_s) + \frac{62}{1024}\Delta) + (\frac{1}{2} + \frac{1}{64})(\eta_{f^*(x)}(x) - \eta_y(x)) \quad \textit{by (57)} \end{aligned}$$

Then,

$$\eta_{f^*(x)}(X_s) - \eta_j(X_s) \leq \frac{32}{31}\frac{4255}{8192}(\eta_{f^*(x)}(x) - \eta_j(x)),$$

and (63) becomes

$$L.P_X(B(X_s, \rho(X_s, X_{s'})))^{\alpha/d} \leq \frac{3}{128C_{db}^3}\zeta(\eta(x)),$$

which implies

$$\rho(X_s, X_{s'}) \leq r_{\wedge(x)}(X_s), \tag{64}$$

where

$$\wedge(x) = \left(\frac{3}{128LC_{db}^3}\zeta(\eta(x))\right)^{d/\alpha}.$$

By the triangular inequality, we have:

$$\frac{1}{4}\rho(x, X_s) < \frac{1}{2}\rho(x, X_s) \leq \frac{1}{2}\left(\rho(x, X_{s'}) + \rho(X_{x'}, X_s)\right), \tag{65}$$

which implies (by using (62), (64)).

$$\frac{1}{4}\rho(x, X_s) < \max(r_{\tilde{p}_\epsilon}(x), r_{\wedge(x)}(X_s)) \tag{66}$$

We will consider two situations:

- $r_{\tilde{p}_\epsilon}(x) \geq r_{\wedge(x)}(X_s)$: in this case, (61) becomes

$$\| \eta(x) - \eta(X_x^{(1)}) \|_\infty \leq \frac{1}{128}\Delta. \tag{67}$$

- $r_{\tilde{p}_\epsilon}(x) < r_{\wedge(x)}(X_s)$: in this case, (61) becomes

$$\begin{aligned}\| \eta(x) - \eta(X_x^{(1)}) \|_\infty &\leq C_{db}^3 L.P_X(B(x, \frac{1}{8}\rho(x, X_s)))^{\alpha/d} \\ &\leq C_{db}^3 L.P_X(B(X_s, \frac{1}{4}\rho(x, X_s)))^{\alpha/d} \\ &\leq \frac{3}{128}\zeta(\eta(x)).\end{aligned}$$

Because $\zeta(\eta(x)) > \frac{1}{2}\Delta$, we obtain in both cases that

$$\| \eta(x) - \eta(X_x^{(1)}) \|_\infty \leq \frac{3}{128}\zeta(\eta(x))$$

Furthermore, by using (55), we have for all $j \neq f^*(x)$,

$$\begin{aligned}\eta_{f^*(x)}(X_x^{(1)}) - \eta_j(X_x^{(1)}) &\geq (\eta_{f^*(x)}(x) - \eta_j(x)) - \frac{6}{128}\zeta(\eta(x)) \\ &\geq \frac{1}{4}\Delta.\end{aligned}$$

Then $f^*(X_x^{(1)}) = f^*(x)$ and by Theorem A.3, the subroutine `ConfidentLabel` $(X_x^{(1)})$ outputs

$$\widehat{Y}_x^{(1)} = f^*(X_x^{(1)}). \tag{68}$$

We easily deduce that:

$$f_{n,w}(x) = \widehat{Y}_x^{(1)} = f^*(X_x^{(1)}) = f^*(x).$$

□

### A.2.5 Label complexity of MKAL

**Lemma A.9.**
*Let us assume that $w$ satisfies (27), (21), and $w \geq T_{\epsilon,\delta}$. Then, there exists an event $A_5$ such that $P(A_5) \geq 1-\delta/8$, and on $A_1 \cap A_2 \cap A_3 \cap A_5$. The condition (19) is sufficient to guarantee (50).*

Before beginning the proof, let us define a notion that will be used through the proof.

**Definition A.1** ($p$-probability-packing)**.**
*Let a set $\mathcal{F} \subset supp(P_X)$. Let $\{x_1, \ldots, x_m\} \subset \mathcal{F}$ and $p \in (0,1]$. We say that the set $\{x_1, \ldots, x_m\}$ is a $p$-probability-packing set of $\mathcal{F}$ if:*

$$\forall s, s' \leq m,\ s \neq s' \Longrightarrow \rho(x_s, x_{s'}) > r_p(x_s) \vee r_p(x_{s'}), \tag{69}$$

*where $r_p$ is defined by (15), and $a \vee b = \max(a,b)$ for $a, b \in \mathbb{R}$.*

This notion of $p$-probability-packing comes from the Definition 1.4 in (Edgar, 2000). It will be used on a particular set of the form $\{x \in supp(P_X),\ \gamma \leq \zeta(\eta(x)) \leq \gamma'\}$, where $0 < \gamma < \gamma'$. This allows us to upper bound the number of informative points whose labels are inferred with very high confidence.

*Proof.*
Let us consider the last update of $I$, the set of indexes of informative points used in MKAL (Algorithm 1 in the main document).
Set $s_I = \max I$, the index of the last informative point. We consider two cases:

1. **First case:** $s_I = w$: we can easily see that (50) is satisfied, and we trivially have that the condition (19) is sufficient to guarantee (50).

2. **Second case:** $s_I < w$: then the total number of label requests up to $s_I$ is:

$$\sum_{s \in I} |Q_s|, \tag{70}$$

where $|Q_s|$ is the number of label requests used in the subroutine ConfidentLabel (Algorithm(5)) with input $X_s$. Let $s \in I$. For brevity, let us denote ConfidentLabel$(X_s)$:=ConfidentLabel$(X_s, k(\epsilon,\delta_s), \delta_s)$. If $s < s_I$, the subroutine ConfidentLabel$(X_s)$ implicitly assumes that the process of label request does not takes into account the constraint related to the budget $n$ (very large budget with respect to $k(\epsilon,\delta_s)$). Then we have:

$$n > \sum_{\substack{s \in I \\ s < s_I}} |Q_s|. \tag{71}$$

On the other hand, we want to guarantee the condition (50). For this, necessarily for all $s \in I$ such that $s \leq T_{\epsilon,\delta}$ and $s < s_I$ at the end of the subroutine ConfidentLabel$(X_s)$, the budget $n$ is not yet reached and then we can replace the relation (71) by

$$n > \sum_{\substack{s \in I \\ s < s_I \\ s \leq T_{\epsilon,\delta}}} |Q_s|. \tag{72}$$

Then, necessarily, (50) holds when (72) holds.
Also, for $s \in I$, by Theorem A.3, if we assume that $\zeta(\eta(X_s)) \geq \frac{1}{4}\Delta$, we have that $|Q_s| \leq \tilde{k}(\epsilon,\delta_s)$, and the subroutine ConfidentLabel$(X_s)$ terminates when the cut-off condition (31) is satisfied. The right-hand side of (72) is equal to:

$$\sum_{\substack{s \in I \\ s < s_I \\ s \leq T_{\epsilon,\delta} \\ \zeta(\eta(X_s)) \geq \frac{1}{4}\Delta}} |Q_s| + \sum_{\substack{s \in I \\ s < s_I \\ s \leq T_{\epsilon,\delta} \\ \zeta(\eta(X_s)) \leq \frac{1}{4}\Delta}} |Q_s|. \tag{73}$$

Firstly, let us consider the first term in (73) and denote it by $T_1$. Let us denote by $B_s$ the event:

$$B_s = \{\zeta(\eta(X_s)) \geq \frac{1}{4}\Delta\}.$$

We have

$$\mathbb{1}_{B_s} = \sum_{j=1}^{m_\epsilon} \mathbb{1}_{B_{s,j}}, \tag{74}$$

where

$$B_{s,j} = \{2^{j-1}\frac{1}{4}\Delta \leq \zeta(\eta(X_s)) \leq 2^j\frac{1}{4}\Delta\} \quad \text{and } m_\epsilon = \left\lceil \log_2\left(\frac{1}{\frac{1}{4}\Delta}\right)\right\rceil.$$

Then,

$$\begin{aligned} T_1 &\leq \sum_{\substack{s\in I \\ s<s_I \\ s\leq T_{\epsilon,\delta} \\ \zeta(\eta(X_s))\geq\frac{1}{4}\Delta}} \tilde{k}(\epsilon,\delta_s) \quad \text{by Theorem A.3} \\ &= \sum_{\substack{s\in I \\ s<s_I \\ s\leq T_{\epsilon,\delta}}} \sum_{j=1}^{m_\epsilon} \tilde{k}(\epsilon,\delta_s)\mathbb{1}_{B_{s,j}}. \end{aligned} \tag{75}$$

On $B_{s,j}$,

$$\begin{aligned} \tilde{k}(\epsilon,\delta_s) &\leq \frac{4c}{2^{2j}\Delta^2}\left[\log(\frac{32Ms^2}{\delta}) + \log\log(\frac{32Ms^2}{\delta}) + \log\log\left(\frac{512\sqrt{e}}{2^j\Delta}\right)\right] \\ &\leq \frac{4c}{2^{2j}\Delta^2}\left[2\log(\frac{32Ms^2}{\delta}) + \log\log\left(\frac{512\sqrt{e}}{\Delta}\right)\right]. \end{aligned} \tag{76}$$

Then (75) becomes:

$$T_1 \leq \frac{4c}{\Delta^2}\left[2\log(\frac{32MT^2_{\epsilon,\delta}}{\delta}) + \log\log\left(\frac{512\sqrt{e}}{\Delta}\right)\right]\sum_{j=1}^{m_\epsilon} 2^{-2j} \sum_{\substack{s\in I \\ s\leq s_I \\ s\leq T_{\epsilon,\delta}}} \mathbb{1}_{B_{s,j}}. \tag{77}$$

In (77), the term $N_j = \sum_{\substack{s\in I \\ s\leq s_I \\ s\leq T_{\epsilon,\delta}}} \mathbb{1}_{B_{s,j}}$ represents the numbers of informative points that belong to the set

$$I_j = \{x,\ \gamma_{j-1} \leq \zeta(\eta(X_s)) \leq \gamma_j\}, \tag{78}$$

where $\gamma_j = 2^j.\frac{\Delta}{4}$, $j = 1\ldots, m_\epsilon$. We will prove that

$$N_j \leq O\left((\gamma_j)^{\beta-\frac{d}{\alpha}}\right), \tag{79}$$

and proceed in two steps:

- The set of informative points that belong to $I_j$ forms a $p_j$-probability-packing set (for $p_j$ well chosen) of $I_j$.
- The cardinal of any $p_j$-probability-packing set satisfies (79).

(a) Let us begin with first step:
Let $X_s, X_{s'}$ any two informative points that belong to $I_j$. Without loss of generality, let us assume that $s < s'$. As $X_s \in I_j$, we have $\zeta(\eta(X_s)) \geq \frac{\Delta}{4}$ and by Theorem A.3, the number of label requests $\bar{k}_s$ used in `ConfidentLabel`$(X_s)$ satisfies:

$$\zeta(\widehat{\eta}_{\bar{k}_s}(X_s)) \geq 4b_{\delta_s,\bar{k}_s}, \tag{80}$$

where $\widehat{\eta}_{\bar{k}_s}(X_s)$ and $b_{\delta_s,\bar{k}_s}$ are respectively defined by (16) and (A.1).
Then $X_s \in \widehat{S}_{ac}^{(1)}$. Additionally, as $X_s$ and $X_{s'}$ are both informative points, by Lemma A.8, we necessary have on event $A_3$ (see Lemma A.8), that

$$\min(\widehat{p}_X, \widehat{p}_{X'}) \geq \frac{75}{94}\left(\frac{1}{64LC_{db}^3}\widehat{\ell}_s\right)^{d/\alpha}. \tag{81}$$

On the event $A_3$, equation (81) necessary implies:

$$\min(P_X(B(X_s, \rho(X_{s'}, X_s))), P_X(B(X_{s'}, \rho(X_{s'}, X_s)))) \geq \frac{28}{47}\left(\frac{1}{64LC_{db}^3}\widehat{\ell}_s\right)^{d/\alpha}. \tag{82}$$

Using the quantity $\bar{\eta}_{\bar{k}_s}(X_s)$ defined by (16), we have by Theorem A.3, on the event $A_2$,

$$\| \bar{\eta}_{\bar{k}_s}(X_s) - \widehat{\eta}_{\bar{k}_s}(X_s) \|_\infty \leq b_{\delta_s,\bar{k}_s}.$$

Then using also the smoothness assumption, on the event $A_1 \cap A_2$, we have simultaneously for all $l \leq M$:

$$\begin{cases} \eta_l(X_s) \leq \bar{\eta}_{\bar{k}_s,l}(X_s) + \frac{31}{1024}\Delta \\ \eta_l(X_s) \geq \bar{\eta}_{\bar{k}_s,l}(X_s) - \frac{31}{1024}\Delta, \end{cases} \tag{83}$$

and

$$\begin{cases} \widehat{\eta}_{\bar{k}_s,l}(X_s) - \bar{\eta}_{\bar{k}_s,l}(X_s) \leq b_{\delta_s,\bar{k}_s} \\ \widehat{\eta}_{\bar{k}_s,l}(X_s) \geq \bar{\eta}_{\bar{k}_s,l}(X_s) - b_{\delta_s,\bar{k}_s}. \end{cases} \tag{84}$$

Additionally, by Theorem A.3, we have $f^*(X_s) = \underset{l\in\mathcal{Y}}{\arg\max}\, \widehat{\eta}_{\bar{k}_s,l}(X_s)$ and then, by using (83), (84), we have:

$$\zeta(\widehat{\eta}_{\bar{k}_s}(X_s)) \geq \frac{225}{256}\zeta(\eta(X_s)) - 2b_{\delta_s,\bar{k}_s}. \tag{85}$$

Therefore, we have on $A_1 \cap A_2$:

$$\begin{aligned}
\ell_s &= \zeta(\widehat{\eta}_{\bar{k}_s}(X_s)) - 2b_{\delta_s,\bar{k}_s} \\
&= \zeta(\widehat{\eta}_{\bar{k}_s}(X_s)) - \frac{8}{3}b_{\delta_s,\bar{k}_s} + \frac{2}{3}b_{\delta_s,\bar{k}_s} \\
&\geq \frac{1}{3}\zeta(\widehat{\eta}_{\bar{k}_s}(X_s)) + \frac{2}{3}b_{\delta_s,\bar{k}_s} \quad \text{by using (80)} \\
&\geq \frac{1}{3}\left(\frac{225}{256}\zeta(\eta(X_s)) - 2b_{\delta_s,\bar{k}_s}\right) + \frac{2}{3}b_{\delta_s,\bar{k}_s} \quad \text{by using (85)} \\
&= \frac{225}{768}\zeta(\eta(X_s)) \\
&\geq \frac{225}{768}\gamma_{j-1} \text{ as } X_s \in I_j \\
&= \frac{225}{1536}\gamma_j.
\end{aligned} \tag{86}$$

Then, the equation (86) becomes:

$$\min(P_X(B(X_s, \rho(X_{s'}, X_s))), P_X(B(X_{s'}, \rho(X_{s'}, X_s)))) \geq \frac{28}{47}\left(\frac{1}{L}\frac{225}{52224C_{db}^3}\gamma_j\right)^{d/\alpha}. \tag{87}$$

In the same way, we also obtain (87) if $s' < s$.
Then, if we set

$$p_j = \frac{28}{47}\left(\frac{1}{L}\frac{225}{52224C_{db}^3}\gamma_j\right)^{d/\alpha}, \tag{88}$$

we have that, by (15) and (87):

$$\rho(X_{s'}, X_s) \geq r_{p_j}(X_s) \vee r_{p_j}(X_{s'}). \tag{89}$$

Therefore, the set of informative points that belong to $I_j = \{x,\ \gamma_{j-1} \leq \zeta(\eta(x)) \leq \gamma_j\}$ forms an $p_j$-probability-packing set.

(b) As second step, let us determine an upper bound of the cardinal of any $p_j$-probability-packing set of $I_j$. Let $\Lambda_j = \{x_1, \ldots, x_{F_j}\}$ any $p_j$-probability-packing set of $I_j$.
For all $s, s' \leq F_j$, we obviously have:

$$s \neq s' \Longrightarrow B(x_s, \frac{r_{p_j}(x_s)}{2}) \cap B(x_{s'}, \frac{r_{p_j}(x_{s'})}{2}) = \emptyset. \tag{90}$$

Then, we have:

$$\begin{aligned} P_X(\bigcup_{s=1}^{F_j} B(x_s, r_{p_j}(x_s)/2)) &= \sum_{s=1}^{F_j} P_X(B(x_s, r_{p_j}(x_s)/2)) \\ &\geq C_{db} \sum_{s=1}^{F_j} P_X(B(x_s, r_{p_j}(x_s))) \\ &\qquad \text{by assumption (H3) from the main manuscript} \\ &\geq C_{db} F_j p_j \\ &\qquad \text{by (15).} \end{aligned} \tag{91}$$

On the other hand, if $z \in B(x_s, r_{p_j}(x_s)/2)$ for some $s \leq F_j$, by assumption (H2) from the main manuscript and equation (15), we have:

$$\| \eta(z) - \eta(x_s) \|_\infty \leq c_0 \gamma_j, \tag{92}$$

where $c_0 = \frac{28}{47}^{\alpha/d} \frac{225}{52224 C_{db}^3}$.
Furthermore, as $\zeta(\eta(x_s)) \geq \frac{1}{2}\gamma_j$, by using (55), we can easily see that $f^*(z) = f^*(x_s)$. Let $j_s$ be defined as:

$$j_s = \arg\max_{j \neq f^*(x_s)} \eta_j(x_s).$$

Thus

$$\begin{aligned} \zeta(\eta(z)) &\leq \eta_{f^*(x_s)}(z) - \eta_{j_s}(z) \\ &\leq \eta_{f^*(x_s)}(x_s) - \eta_{j_s}(x_s) + 2 \| \eta(z) - \eta(x_s) \|_\infty \quad \text{by (55)} \\ &\leq \gamma_j + 2c_0\gamma_j = \tilde{c}\gamma_j \quad \text{by (92)}, \end{aligned}$$

where $\tilde{c} = 1 + 2c_0$.
Now we can upper bound $F_j$ by using assumption (H1) from the main manuscript,

$$\begin{aligned} C_{db} F_j p_j \leq P_X(\bigcup_{s=1}^{F_j} B(x_s, r_{p_j}(x_s)/2)) &\leq P_X(z;\ \zeta(\eta(z)) \leq \tilde{c}\gamma_j) \\ &\leq C(\tilde{c}\gamma_j)^\beta. \end{aligned} \tag{93}$$

Then,

$$\begin{aligned} F_j &\leq \frac{C}{C_{db}} \frac{(\tilde{c}\gamma_j)^\beta}{p_j} \\ &= \tilde{b}(\gamma_j)^{\beta - \frac{d}{\alpha}}. \end{aligned} \tag{94}$$

Then the cardinal of any $p_j$-probability-packing set of $I_j$ is upper bounded by $O\left((\gamma_j)^{\beta - \frac{d}{\alpha}}\right)$, consequently, equation (79) holds.

Equation (77) becomes:

$$\begin{aligned} T_1 &\leq \frac{16c\tilde{b}}{\Delta^2}\left[2\log(\frac{32MT_{\epsilon,\delta}^2}{\delta}) + \log\log\left(\frac{512\sqrt{e}}{\Delta}\right)\right]\sum_{j=1}^{m_\epsilon} 2^{-2j}(\gamma_j)^{\beta-\frac{d}{\alpha}} \\ &= 4^{\frac{d}{\alpha}-\beta+2}c\tilde{b}\Delta^{\beta-\frac{d}{\alpha}-2}\left[2\log(\frac{32MT_{\epsilon,\delta}^2}{\delta}) + \log\log\left(\frac{512\sqrt{e}}{\Delta}\right)\right]\sum_{j=1}^{m_\epsilon} 2^{(-2+\beta-\frac{d}{\alpha})j} \\ &\leq b_0\left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d-\alpha\beta}{\alpha(\beta+1)}}\left[2\log(\frac{32MT_{\epsilon,\delta}^2}{\delta}) + \log\log\left(\frac{512\sqrt{e}}{\Delta}\right)\right]m_\epsilon, \end{aligned} \tag{95}$$

where $b_0 = 4^{\frac{d}{\alpha}-\beta+2}c\tilde{b}(C)^{\frac{2\alpha+d-\alpha\beta}{\alpha(\beta+1)}}$. Equation (95) holds because we have $\alpha\beta \leq d$, $\Delta = \max\left(\frac{\epsilon}{2}, \left(\frac{\epsilon}{C}\right)^{\frac{1}{\beta+1}}\right)$.

Now, it remains to upper bound the second term in (73), which is denoted by $T_2$. By Lemma A.1, equation (25), there exists an event $A_5$ such that $P(A_5) \geq 1-\delta/8$, and on $A_5$, we have:

$$T_2 \leq \sum_{s\leq T_{\epsilon,\delta}} |Q_s|\mathbb{1}_{\tilde{B}_s} \leq k(\epsilon,\delta')\left(1 + \frac{4}{P_X(\tilde{B})T_{\epsilon,\delta}}\log\left(\frac{8}{\delta}\right)\right)P_X(\tilde{B})T_{\epsilon,\delta},$$

because $|Q_s| \leq k(\epsilon,\delta')$ (according to the subroutine `ConfidentLabel`) for all $s \leq T_{\epsilon,\delta}$ and where $\tilde{B} = \{x,\ \zeta(\eta(x)) \leq \frac{\Delta}{4}\}$, $\delta' = \frac{\delta}{32MT_{\epsilon,\delta}^2}$ and $k(\epsilon,\delta')$ is defined in (A.1).
Consequently, we have:

$$\begin{aligned} T_2 &\leq k(\epsilon,\delta')\left(P_X(\tilde{B})T_{\epsilon,\delta} + 4\log\left(\frac{8}{\delta}\right)\right) \\ &\leq k(\epsilon,\delta')\left(T_{\epsilon,\delta}\frac{1}{4^\beta}C\Delta^\beta + 4\log\left(\frac{8}{\delta}\right)\right) \quad \text{by assumption (H1) from the main manuscript} \\ &= k(\epsilon,\delta')\left(\left(\frac{128LC_{db}^3}{\Delta}\right)^{d/\alpha}\log\left(\frac{8}{\delta}\right)\frac{1}{4^\beta}C\Delta^\beta + 4\log\left(\frac{8}{\delta}\right)\right) \quad \text{by (48)} \\ &= k(\epsilon,\delta')\log\left(\frac{8}{\delta}\right)\left(C_{db}^{3d/\alpha}(128)^{d/\alpha-\beta}32^\beta C\left(\frac{1}{\Delta}\right)^{d/\alpha-\beta} + 4\right). \end{aligned} \tag{96}$$

As $\alpha\beta \leq d$, $\Delta \leq 1$, $C \geq 1$, the term $C_{db}^{3d/\alpha}(128)^{d/\alpha-\beta}32^\beta C\left(\frac{1}{\Delta}\right)^{d/\alpha-\beta}$ in (96) is greater than 1. Thus, (96) becomes:

$$\begin{aligned} T_2 &\leq 5k(\epsilon,\delta')\log\left(\frac{8}{\delta}\right)C_{db}^{3d/\alpha}(128)^{d/\alpha-\beta}32^\beta C\left(\frac{1}{\Delta}\right)^{d/\alpha-\beta} \\ &= 5c\left[\log(\frac{1}{\delta'}) + \log\log(\frac{1}{\delta'}) + \log\log\left(\frac{512\sqrt{e}}{\Delta}\right)\right]\log\left(\frac{8}{\delta}\right)C_{db}^{3d/\alpha}(128)^{d/\alpha-\beta}32^\beta C\left(\frac{1}{\Delta}\right)^{d/\alpha-\beta+2} \\ &\leq \left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d-\alpha\beta}{\alpha(\beta+1)}}\left[2\log(\frac{1}{\delta'}) + \log\log\left(\frac{512\sqrt{e}}{\Delta}\right)\right]\log\left(\frac{8}{\delta}\right)\tilde{u}, \end{aligned} \tag{97}$$

where $\tilde{u} = 5c\,(C)^{\frac{2\alpha+d-\alpha\beta}{\alpha(\beta+1)}}C_{db}^{3d/\alpha}(128)^{d/\alpha-\beta}32^\beta C$. Equation (97) holds by using the definition of $\Delta$ (17).

By combining (97) and (95), the term obtained in (73) is less than:

$$\begin{aligned} &b_0\left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d-\alpha\beta}{\alpha(\beta+1)}}\left[2\log(\frac{32MT_{\epsilon,\delta}^2}{\delta}) + \log\log\left(\frac{512\sqrt{e}}{\Delta}\right)\right]m_\epsilon + \\ &\left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d-\alpha\beta}{\alpha(\beta+1)}}\left[2\log(\frac{1}{\delta'}) + \log\log\left(\frac{512\sqrt{e}}{\Delta}\right)\right]\log\left(\frac{8}{\delta}\right)\tilde{u}. \end{aligned}$$

Thus, if the label budget $n$ satisfies

$$n \geq 2b_0 \left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d-\alpha\beta}{\alpha(\beta+1)}} \left[2\log\left(\frac{32MT^2_{\epsilon,\delta}}{\delta}\right) + \log\log\left(\frac{512\sqrt{e}}{\Delta}\right)\right] \max\left(m_\epsilon, \log\left(\frac{8}{\delta}\right)\tilde{u}\right), \tag{98}$$

we have that $n$ satisfies (72), and (50) is necessary satisfied.

□

### A.2.6 Proof of Theorem 5.1

**Lemma A.10** (Proof of Equation (22))**.**
*Let $\hat{\eta}$ be the estimator of the regression function $\eta$ provided by our algorithm* `MKAL` *(Algorithm 1 from the main document). Let us assume that the condition (19) holds. Then, on the event $A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5$, we have:*

$$\| \eta(x) - \hat{\eta}(x) \|_\infty \leq \frac{\Delta}{2},$$

*for all $x \in supp(P_X)$ such that $\zeta(\eta(x)) \leq \frac{1}{2}\Delta$.*

*Proof.*
The idea consists in a bias-variance decomposition. Let $\hat{x}$ be the nearest neighbor of $x$ in the final active set $\widehat{S}_{ac} = \widehat{S}^{(1)}_{ac} \cup \widehat{S}^{(2)}_{ac}$. Thus, we have:

$$\begin{aligned} \| \eta(x) - \hat{\eta}(x) \|_\infty &= \| \eta(x) - \hat{\eta}_{k_{\hat{x}}}(\hat{x}) \|_\infty \\ &\leq \underbrace{\| \eta(x) - \eta(\hat{x}) \|_\infty}_{B_1} + \underbrace{\| \eta(\hat{x}) - \bar{\eta}_{k_{\hat{x}}}(\hat{x}) \|_\infty}_{B_2} + \underbrace{\| \bar{\eta}_{k_{\hat{x}}}(\hat{x}) - \hat{\eta}_{k_{\hat{x}}}(\hat{x}) \|_\infty}_{V}, \end{aligned} \tag{99}$$

where $k_{\hat{x}}$ is the number of label requests used in the subroutine `ConfidentLabel`$(\hat{x})$ (Algorithm 5 from the main document).

On the event $A_1$, the quantity $B_2$ can be bounded as:

$$\| \eta(\hat{x}) - \bar{\eta}_{k_{\hat{x}}}(\hat{x}) \|_\infty \leq \frac{31}{1024}\Delta. \tag{100}$$

Now, let us give a bound on the quantity $B_1$. By Theorem A.4, on the event $A_4$, there exists $X_x \in \{X_1, \ldots, X_{T_{\epsilon,\delta}}\}$ such that:

$$P_X(B(x, \rho(x, X_x))) \leq \tilde{p}_\epsilon, \tag{101}$$

where $\tilde{p}_\epsilon = \left(\frac{\Delta}{128LC^3_{db}}\right)^{d/\alpha}$. We consider two cases depending on whether $X_x$ passes through the Reliable subroutine or not.

1. If $X_x$ passes through `Reliable`.
In this case,

$$\begin{aligned} \| \eta(x) - \eta(\hat{x}) \|_\infty &\leq L.P_X(B(x, \rho(x, \hat{x})))^{\alpha/d} \\ &\leq L.P_X(B(x, \rho(x, X_x)))^{\alpha/d} \\ &\leq \frac{\Delta}{128}. \end{aligned} \tag{102}$$

2. If $X_x$ does not pass through `Reliable`.
The point $X_x$ can be seen as $X_x := X_{s'}$, with $s' \leq w$. By definition of `MKAL` (Algorithm 1 in the main document), there exists $X_s$ $(s < s')$ with $\widehat{\ell}_s > 0$ such that on the event $A_3$ we have:

$$P_X(B(X_s, \rho(X_s, X_{s'}))) \leq \left(\frac{\widehat{\ell}_s}{64LC^3_{db}}\right)^{d/\alpha} \tag{103}$$

Let $k_s$ be the number of label requests used in ConfidentLabel($X_s$) subroutine (Algorithm 5 in the main document) and $\hat{i}_{k_s} = \arg\max_{l \leq M} \hat{\eta}_{k_s,l}$. We have on $A_1 \cap A_2$:

$$\begin{aligned}
\widehat{\ell}_s &= \zeta(\hat{\eta}_{k_s}(X_s)) - 2b_{\delta_s,k_s} \\
&= \hat{\eta}_{k_s,\hat{i}_{k_s}}(X_s) - \max_{j \neq \hat{i}_{k_s}} \hat{\eta}_{k_s,j}(X_s) - 2b_{\delta_s,k_s} \\
&\leq \bar{\eta}_{k_s,\hat{i}_{k_s}}(X_s) - \bar{\eta}_{k_s,j}(X_s) \quad \text{for all } j \neq \hat{i}_{k_s} \\
&\leq \eta_{\hat{i}_{k_s}}(X_s) - \eta_j(X_s) + \frac{62}{1024}\Delta \quad \text{for all } j \neq \hat{i}_{k_s}. \qquad (104)
\end{aligned}$$

Using (55), we have for all $j \neq \hat{i}_{k_s}$:

$$\eta_{\hat{i}_{k_s}}(X_s) - \eta_j(X_s) \leq \frac{32}{31}\left(\eta_{\hat{i}_{k_s}}(X_x) - \eta_j(X_x) + \frac{62}{32 * 1024}\Delta\right). \qquad (105)$$

If $\eta_{\hat{i}_{k_s}}(X_x) - \eta_j(X_x) \geq \frac{35}{64}\Delta$ for all $j \neq \hat{i}_{k_s}$, then $f^*(X_x) = \hat{i}_{k_s}$. By using (55) and (101), we have for all $j \neq f^*(X_x)$:

$$\begin{aligned}
\eta_{f^*(X_x)}(x) - \eta_j(x) &\geq \eta_{f^*(X_x)}(X_x) - \eta_j(X_x) - \frac{1}{64}\Delta \qquad (106) \\
&\geq \frac{17}{32}\Delta, \qquad (107)
\end{aligned}$$

which implies $f^*(X_x) = f^*(x)$ and then $\zeta(\eta(x)) \geq \frac{17}{32}\Delta$, which contradicts the fact that $\zeta(\eta(x)) \leq \frac{1}{2}\Delta$. Then there exists $j_o \neq \hat{i}_{k_s}$ such that $\eta_{\hat{i}_{k_s}}(X_x) - \eta_{j_o}(X_x) \leq \frac{35}{64}\Delta$. In this case, (104) becomes

$$\widehat{\ell}_s \leq \frac{311}{512}\Delta, \qquad (108)$$

and

$$P_X(B(X_s, \rho(X_s, X_x)) \leq \left(\frac{311}{512 * 64LC_{db}^3}\Delta\right)^{d/\alpha}. \qquad (109)$$

Then,

$$\rho(X_s, X_x) \leq r_{\hat{p}_\epsilon}(X_s), \qquad (110)$$

where $\hat{p}_\epsilon = \left(\frac{311}{512*64LC_{db}^3}\Delta\right)^{d/\alpha}$. Likewise, by (101), we have:

$$\rho(x, X_x) \leq r_{\tilde{p}_\epsilon}(x). \qquad (111)$$

Furthermore, we have:

$$P_X(B(x, \rho(x, X_s)) \leq C_{db}^3 P_X(B(x, \frac{1}{8}\rho(x, X_s)), \qquad (112)$$

and $\frac{1}{8}\rho(x, X_s) \leq \frac{1}{8}\left(\rho(x, X_x) + \rho(X_x, X_s)\right) \leq \frac{1}{4}\max(r_{\tilde{p}_\epsilon}(x), r_{\hat{p}_\epsilon}(X_s))$. We consider two cases:

- **If** $r_{\tilde{p}_\epsilon}(x) \geq r_{\hat{p}_\epsilon}(X_s)$: then we have

$$\begin{aligned}
P_X(B(x, \rho(x, X_s) &\leq C_{db}^3 P_X(B(x, \frac{1}{4}r_{\tilde{p}_\epsilon}(x))) \\
&\leq C_{db}^{3d/\alpha}\tilde{p}_\epsilon.
\end{aligned}$$

  Then,

$$\begin{aligned}
\| \eta(x) - \eta(\hat{x}) \|_\infty &\leq L.P_X(B(x, \rho(x, \hat{x})))^{\alpha/d} \\
&\leq L.P_X(B(x, \rho(x, X_s)))^{\alpha/d} \\
&\leq \frac{1}{128}\Delta.
\end{aligned}$$

- **If** $r_{\tilde{p}_\epsilon}(x) \leq r_{\hat{p}_\epsilon}(X_s)$: then we have

$$\begin{aligned} P_X(B(x,\rho(x,X_s)) &\leq C_{db}^3 P_X(B(x,\frac{1}{8}\rho(x,X_s))) \\ &\leq C_{db}^3 P_X(B(X_s,\frac{1}{4}\rho(x,X_s))) \\ &\leq C_{db}^3 P_X(B(X_s,\frac{1}{2}r_{\hat{p}_\epsilon}(X_s))) \\ &\leq C_{db}^{3d/\alpha}\hat{p}_\epsilon. \end{aligned}$$

Then,

$$\begin{aligned} \| \eta(x)-\eta(\hat{x}) \|_\infty &\leq L.P_X(B(x,\rho(x,\hat{x})))^{\alpha/d} \\ &\leq L.P_X(B(x,\rho(x,X_s)))^{\alpha/d} \\ &\leq \frac{311}{512*64}\Delta. \end{aligned}$$

Finally, we have that the term $B_1$ is bounded by $\frac{1}{128}\Delta$.

Now, let us find a bound on V (see (99)). The point $\hat{x}$ can be seen as $\hat{x}=X_{s'}$ with $s'\leq w$. By Theorem A.3, we have:

$$\| \widehat{\eta}_{k_{\hat{x}}}(\hat{x}) - \bar{\eta}_{k_{\hat{x}}}(\hat{x}) \|_\infty \leq b_{\delta_{s'},k_{\hat{x}}}. \tag{113}$$

We consider two cases depending on whether $k_{\hat{x}}$ reaches $k(\epsilon,\delta_{s'})$ or not in the `ConfidentLabel`$(\hat{x})$ subroutine (Algorithm 5 in the main document).

(a) If $k_{\hat{x}}=k(\epsilon,\delta_{s'})$: then we can easily (by also using Lemma A.7) see that $b_{\delta_{s'},k_{\hat{x}}}\leq\frac{1}{512\sqrt{e}}\Delta$.

(b) If $k_{\hat{x}}<k(\epsilon,\delta_{s'})$, then necessarily (by definition of our algorithm `MKAL`), we have $\zeta(\hat{\eta}_{k_{\hat{x}}}(\hat{x}))\geq 4b_{\delta_{s'},k_{\hat{x}}}$. In this case, we have on the event $A_1\cap A_2$,

$$\begin{aligned} b_{\delta_{s'},k_{\hat{x}}} &\leq \frac{1}{2}(\zeta(\hat{\eta}_{k_{\hat{x}}}(\hat{x}))-2b_{\delta_{s'},k_{\hat{x}}}) \\ &\leq \frac{1}{2}(\eta_{\hat{i},k_{\hat{x}}}(\hat{x})-\eta_{j,k_{\hat{x}}}(\hat{x})+\frac{62}{1024}\Delta) \quad \text{for all } j\neq\hat{i}, \end{aligned} \tag{114}$$

where $\hat{i}=\underset{j\in\{1\ldots,M\}}{\arg\max}\,\hat{\eta}_{k_{\hat{x}},j}(\hat{x})$. Previously, we have proven that:

$$\| \eta(x)-\eta(\hat{x}) \|_\infty \leq \frac{1}{128}\Delta. \tag{115}$$

Then, if $\eta_{\hat{i}}(\hat{x})-\eta_j(\hat{x})>\frac{33}{64}\Delta$ for all $j\neq\hat{i}$, we have $\hat{i}=f^*(\hat{x})$ and $f^*(x)=f^*(\hat{x})$ by using (55) and (115).
Let $j_x$ be defined as $\underset{j\neq f^*(x)}{\arg\max}\,\eta_j$. We have:

$$\begin{aligned} \frac{33}{64}\Delta<\zeta(\eta(\hat{x})) &\leq \eta_{f^*(x)}(\hat{x})-\eta_{j_x}(\hat{x}) \\ &\leq \eta_{f^*(x)}(x)-\eta_{j_x}(x)+\frac{1}{64}\Delta \quad \text{by (55)} \\ &\leq \frac{1}{2}\Delta+\frac{1}{64}\Delta=\frac{33}{64}\Delta, \end{aligned}$$

which leads to a contradiction. Thus, there exists $j_o\neq\hat{i}$ such that $\eta_{\hat{i}}(\hat{x})-\eta_{j_o}(\hat{x})\leq\frac{33}{64}\Delta$. In this case, (114) becomes:

$$b_{\delta_{s'},k_{\hat{x}}}\leq\frac{295}{1024}\Delta.$$

By considering the two previous cases, V can be bounded by $\frac{295}{1024}\Delta$. Finally, (99) becomes:

$$\| \eta(x) - \hat{\eta}(x) \|_\infty \leq \frac{1}{128}\Delta + \frac{31}{1024}\Delta + \frac{295}{1024}\Delta \leq \frac{1}{2}\Delta.$$

□

**Proof of Equation (23):**
Let $\hat{f}_{n,w}$ be the classifier provided by MKAL and $\hat{\eta}$ the corresponding estimator of the regression function. Let us assume the conditions (19), (27), (21) hold. We have on the event $A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5$:

$$\begin{aligned}
R(\hat{f}_{n,w}) - R(f^*) &= E_X(\eta_{f^*(X)}(X) - \eta_{\hat{f}_{n,w}(X)}(X)) \\
&= E_X((\eta_{f^*(X)}(X) - \eta_{\hat{f}_{n,w}(X)}(X))\mathbb{1}_{\zeta(\eta(X))>\Delta/2}) + E_X((\eta_{f^*(X)}(X) - \eta_{\hat{f}_{n,w}(X))}(X)\mathbb{1}_{\zeta(\eta(X))\leq\Delta/2}) \\
&= E_X((\eta_{f^*(X)}(X) - \eta_{\hat{f}_{n,w}(X))}(X)\mathbb{1}_{\zeta(\eta(X))\leq\Delta/2}) \quad \text{by Theorem A.4} \\
&\leq \epsilon \quad \text{by using Proposition 5.1, Lemma A.10, Assumption (H1) .}
\end{aligned}$$

# B ADDITIONAL EXPERIMENTS

In this Section we present some additional computer simulations that were conducted to test our algorithm, and that were not included in the main manuscript to avoid overloading it.

As in the main manuscript, we start first with binary datasets (Section B.1) and then consider multiclass ones (Section B.2). For each dataset, we generate 100000 points as a training set for the algorithm, and 30000 points as a test set. The points are equally distributed between classes. The confidence and accuracy parameters ($\delta$ and $\epsilon$, respectively) have both been set to 0.1 for the experiments presented here.

Compared to the main manuscript, experiments are systematically repeated 10 times and the average results are shown, along with their standard deviation.

For the curves showing the test error as a function of the number of labels used (Figures 3, 5, 7 and 9) the inserts show the envelope of the mean and standard deviation for the passive 1-NN and 5-NN algorithms (blue and green curves, respectively). The corresponding curves for the MKAL algorithm are superimposed as such, because the number of labels used at each step of the algorithm is different for each of the 10 rounds, making averages more cumbersome.

The results are briefly summarized and discussed in Section B.3.

## B.1 Binary datasets

We used the same binary datasets as in the main manuscript, and reported the results of the 10 repetitions in Figure 3.

## B.2 Multiclass datasets

We generate points from a mixture model with $M > 2$ classes corresponding to isotropic gaussian distributions.

### B.2.1 $M = 3$ classes

In the main manuscript, we presented results with the centers of the gaussian distributions chosen randomly and a standard deviation is set to 0.2.

Here to make the results more easily reproducible and to repeat the same experiments 10 times, the centers are initially set to (-0.5, 0.5), (0, 0.5) and (0.5, -0.5), and the generated data are then rescaled to fit in $[-1, 1]^2$. The standard deviation $\sigma$ is set to 0.25 to create a substantial overlap between classes.

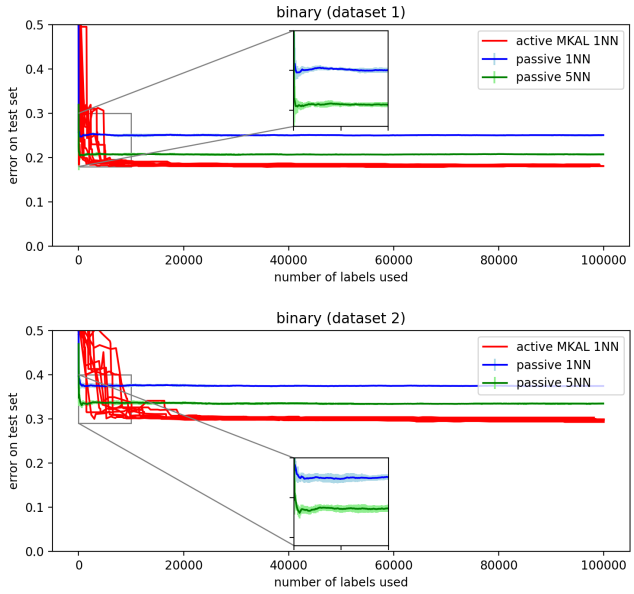


Figure 3: Comparison of the test errors for the binary datasets of for the `MKAL` algorithm (red) and the passive 1-NN counterpart (blue), as well as a 5-NN passive learning (green).

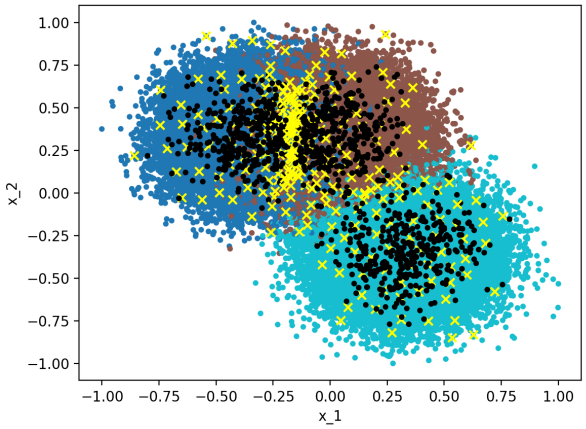


Figure 4: Selection of informative points by `MKAL` for a dataset composed of 3 gaussian distributions ($\sigma = 0.25$), with the points colored according to their class label. The yellow crosses indicate the points that have been considered informative, and the black dots the points that have been considered as non-informative by the `Reliable` subroutine.

### B.2.2 $M = 5$ classes

As in the $M = 3$ case, the results in the main manuscript correspond to randomly chosen centers of the gaussian distributions with a standard deviation is set to 0.2.

Here to make the results more easily reproducible and to repeat the same experiments 10 times, the centers are initially set to (0.5, 0), (-0.5, 0), (0, 0.5), (0.25, -0.5) and (-0.25, -0.5), and the generated data are then rescaled to fit in $[-1, 1]^2$. The standard deviation $\sigma$ is first set to 0.25 to create overlap between classes, then to 0.1 to reduce it considerably.

To study the influence of the noise on the results, we repeated the experiments with $\sigma = 0.1$, which considerably reduces the overlap between classes, thus making the classification problem easier. Figure 6 shows that many

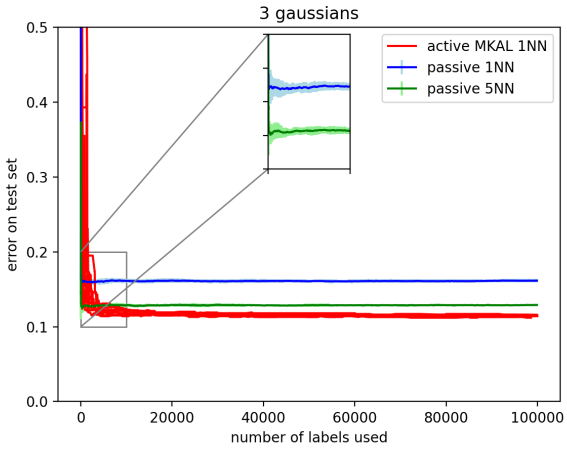


Figure 5: Comparison of the test errors on a dataset composed of 3 gaussian distributions for the `MKAL` algorithm (red) and the passive 1-NN counterpart (blue), as well as a 5-NN passive learning (green).

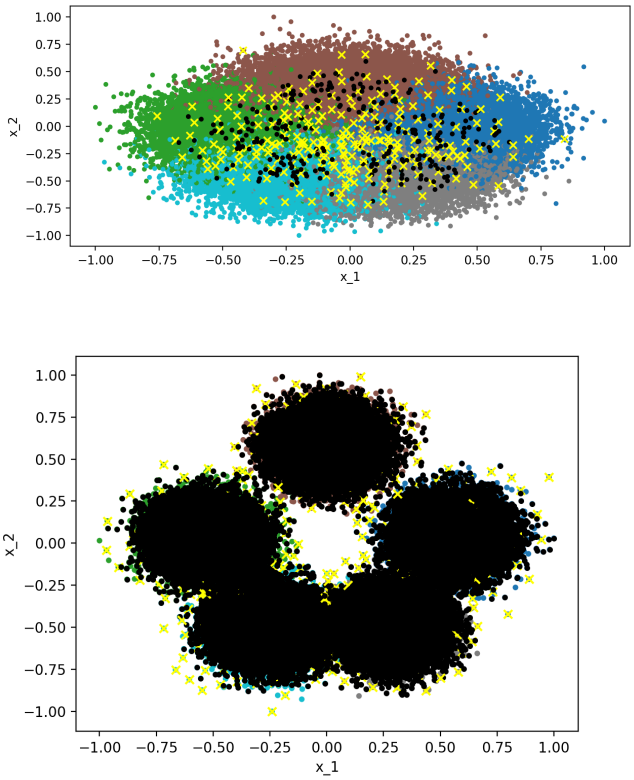


Figure 6: Selection of informative points by `MKAL` for a dataset composed of 5 gaussian distributions ($\sigma = 0.25$, top and $\sigma = 0.1$, bottom), with the points colored according to their class label. The yellow crosses indicate the points that have been considered informative, and the black dots the points that have been considered as non-informative by the `Reliable` subroutine.

points have been identified as non-informative by our algorithm in this case, but Figure 7 (bottom) indicates nevertheless that the advantage of our `MKAL` algorithm is less pronounced. Indeed, choosing appropriately the informative points is less important in such case, where the classification is easier, and all classifiers considered reach an error close to 0.

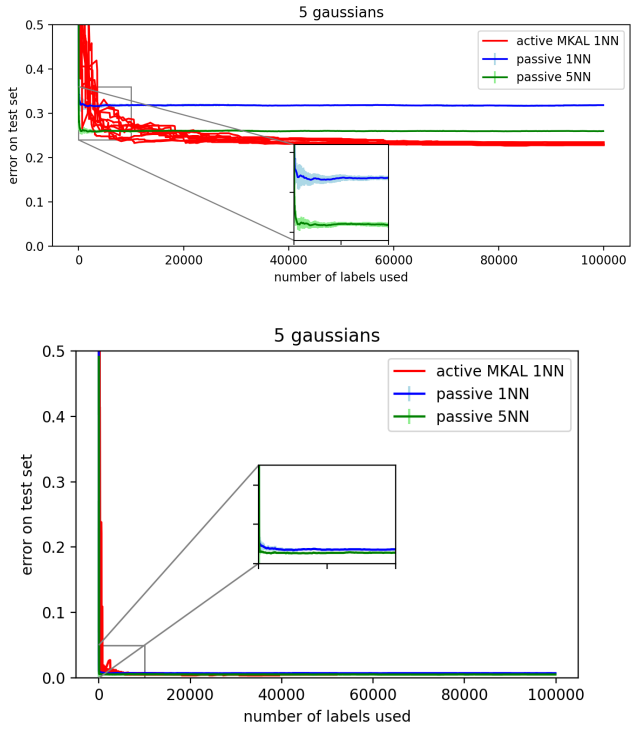


Figure 7: Comparison of the test errors on a dataset composed of 5 gaussian distributions for the `MKAL` algorithm (red) and the passive 1-NN counterpart (blue), as well as a 5-NN passive learning (green). The top graph corresponds to $\sigma = 0.25$ and the bottom one to $\sigma = 0.1$.

### B.2.3 $M = 10$ classes

We present here additional results with 10 gaussian distributions whose centers have been equally spread around a circle of radius 0.5.

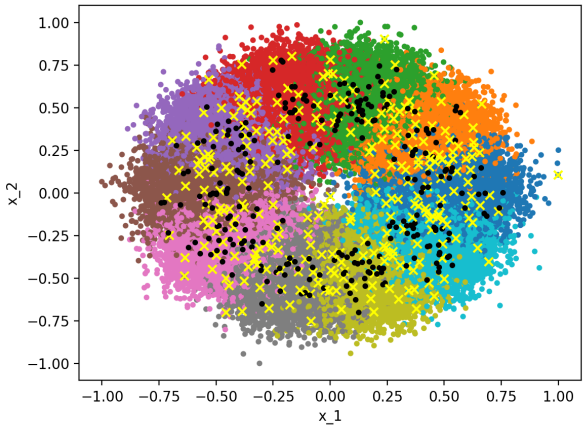


Figure 8: Selection of informative points by `MKAL` for a dataset composed of 10 gaussian distributions ($\sigma = 0.25$), with the points colored according to their class label. The yellow crosses indicate the points that have been considered informative, and the black dots the points that have been considered as non-informative by the `Reliable` subroutine.

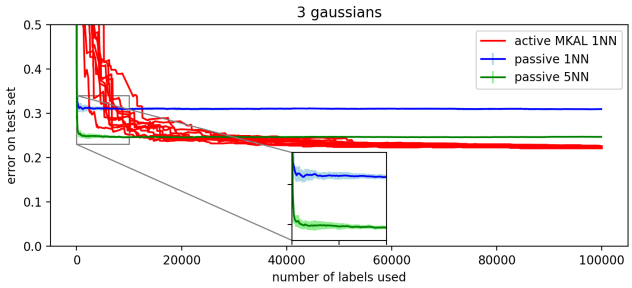


Figure 9: Comparison of the test errors on a dataset composed of 10 gaussian distributions for the `MKAL` algorithm (red) and the passive 1-NN counterpart (blue), as well as a 5-NN passive learning (green).

### B.3 Summary and discussion

The results in Section B, in particular Figures 3, 5, 7 and 9 confirm that our algorithm performs better than its passive 1-NN and 5-NN counterparts in a statistically consistent manner. However, there is clearly more variability in the `MKAL` algorithm (red curves in the aforementioned Figures), which reflects a sensitivity to the (random) choice of the initial point. If this point is close to a boundary between classes, it increases the number of points whose labels must be queried to find its label (parameter $k(\epsilon, \delta_s)$ in the `MKAL` algorithm, which is used by the `confidentLabel` subroutine). The variability in the passive 1-NN and 5-NN classifiers is much lower, as shown in the inserts.

A summary of the final test errors after convergence of the algorithm is presented in Table 1. This shows that our `MKAL` algorithm outperforms the passive 1-NN and 5-NN classifiers on most datasets, except the dataset with 5 gaussian distributions and $\sigma = 0.1$, which is the less noisy dataset. This shows that our algorithm is particularly well suited for difficult classification problems.

| dataset | active `MKAL` | passive 1-NN | passive 5-NN |
|---|---|---|---|
| first binary | **0.182** $\pm$ 0.002 | 0.251 $\pm$ 0.002 | 0.207 $\pm$ 0.002 |
| second binary | **0.298** $\pm$ 0.002 | 0.374 $\pm$ 0.002 | 0.335 $\pm$ 0.002 |
| 3 gaussians ($\sigma = 0.25$) | **0.114** $\pm$ 0.001 | 0.162 $\pm$ 0.002 | 0.129 $\pm$ 0.001 |
| 5 gaussians ($\sigma = 0.25$) | **0.230** $\pm$ 0.002 | 0.318 $\pm$ 0.002 | 0.260 $\pm$ 0.001 |
| 5 gaussians ($\sigma = 0.1$) | **0.005** $\pm$ 0.001 | 0.007 $\pm$ 0.001 | **0.005** $\pm$ 0.001 |
| 10 gaussians ($\sigma = 0.25$) | **0.224** $\pm$ 0.002 | 0.309 $\pm$ 0.002 | 0.247 $\pm$ 0.001 |

Table 1: Summary of the test errors (mean $\pm$ standard deviation computed on 10 repetitions of each experiment) reached after convergence of the algorithms on several datasets.

## References


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

Maria-Florina Balcan, Andrei Broder, and Tong Zhang. Margin based active learning. In *International Conference on Computational Learning Theory*, pages 35–50. Springer, 2007.

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

Maria-Florina Balcan, Steve Hanneke, and Jennifer Wortman Vaughan. The true sample complexity of active learning. *Machine learning*, 80(2-3):111–139, 2010.

Christopher Berlind and Ruth Urner. Active nearest neighbors in changing environments. In *International Conference on Machine Learning*, pages 1870–1879, 2015.

Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. Importance weighted active learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 49–56, 2009.

Gérard Biau and Luc Devroye. *Lectures on the nearest neighbor method*. Springer, 2015.

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

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pages 3437–3445, 2014.

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

Sanjoy Dasgupta. Two faces of active learning. *Theoretical computer science*, 412(19):1767–1781, 2011.

Sanjoy Dasgupta, Daniel J Hsu, and Claire Monteleoni. *A general agnostic active learning algorithm*. Citeseer, 2007.

Luc Devroye, Laszlo Gyorfi, Adam Krzyzak, and Gábor Lugosi. On the strong universal consistency of nearest neighbor regression function estimates. *The Annals of Statistics*, pages 1371–1385, 1994.

Gerald A Edgar. Packing measure in general metric space. *Real Analysis Exchange*, 26(2):831–852, 2000.

László Györfi and Roi Weiss. Universal consistency and rates of convergence of multiclass prototype algorithms in metric spaces. *Journal of Machine Learning Research*, 22(151):1–25, 2021.

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

Steve Hanneke. Nonparametric active learning, part 1: Smooth regression functions. http://www.stevehanneke.com/, 12 2018.

Steve Hanneke and Liu Yang. Minimax analysis of active learning. *The Journal of Machine Learning Research*, 16(1):3487–3602, 2015.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

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. k-nn regression adapts to local intrinsic dimension. In *Advances in Neural Information Processing Systems*, pages 729–737, 2011.

Akshay Krishnamurthy, Alekh Agarwal, Tzu-Kuo Huang, Hal Daumé III, and John Langford. Active learning for cost-sensitive classification. In *International Conference on Machine Learning*, pages 1915–1924. PMLR, 2017.

Sanjeev R Kulkarni and Steven E Posner. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Transactions on Information Theory*, 41(4):1028–1039, 1995.

Andrea Locatelli, Alexandra Carpentier, and Samory Kpotufe. Adaptivity to noise parameters in nonparametric active learning. *Proceedings of Machine Learning Research vol*, 65:1–34, 2017.

Andrea Locatelli, Alexandra Carpentier, and Samory Kpotufe. An adaptive strategy for active learning with smooth decision boundary. In *Algorithmic Learning Theory*, pages 547–571, 2018.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

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

Wolfgang Mulzer. Five proofs of Chernoff's bound with applications. *arXiv preprint arXiv:1801.03365*, 2018.

Nikita Puchkin and Vladimir Spokoiny. An adaptive multiclass nearest neighbor classifier. *ESAIM: Probability and Statistics*, 24:69–99, 2020.

Henry WJ Reeve and Gavin Brown. Minimax rates for cost-sensitive learning on manifolds with approximate nearest neighbours. In *International Conference on Algorithmic Learning Theory*, pages 11–56, 2017.

Burr Settles. Active learning literature survey. *University of Wisconsin, Madison*, page 11, 2010.

Mathukumalli Vidyasagar. *Learning and generalisation: with applications to neural networks*. Springer Science & Business Media, 2013.