Nonparametric Active learning under local smoothness assumption

Boris Ndjia Njike *[1] et Xavier Siebert †[1]

[1]Department of Mathematics and Operational Research, Faculté Polytechnique, Université de Mons, Rue de Houdain 9, 7000 Mons, Belgium.

September 6, 2019

Abstract

There is a large body of work on convergence rates either in passive or active learning. Here we first outline some of the main results that have been obtained, more specifically in a nonparametric setting under assumptions about the smoothness and the margin noise. We discuss the relative merits of these underlying assumptions by putting active learning in perspective with recent work on passive learning. We provide a novel active learning algorithm with a rate of convergence better than in passive learning, using a particular smoothness assumption customized for k-nearest neighbors. This smoothness assumption provides a dependence on the marginal distribution of the instance space unlike those are commonly used in the recent litterature. Our algorithm thus avoids the strong density assumption that supposes the existence of the density function of the marginal distribution of the instance space and is therefore more generally applicable.

Keywords: Nonparametric learning, active learning, nearest-neighbors, smoothness condition.

1 Introduction

Active learning is a machine learning approach for reducing the data labelling effort. Given an instance space $\mathcal{X}$ or a pool of unlabelled data $\{X_1, \ldots, X_w\}$ provided by a distribution P_X, the learner focuses its labeling effort only on the most "informative" points so that a model built from them can achieve the best possible guarantees [6]. Such guarantees are particularly interesting when they are significantly better than those obtained in passive learning [10]. In the context of this work, we consider binary classification (where the label Y of X takes its value in $\{0, 1\}$) in a nonparametric setting. Extensions to multiclass classification and adaptive algorithms are discussed in the last section.

The nonparametric setting has the advantage of providing guarantees with many informations such as the dependence on the dimensional and distributional parameters by using some hypotheses on the regularity of the decision boundary [4], on the regression function [20, 15], and on the geometry of instance space (called strong density assumption) [1, 15, 20]. One of the initial works on nonparametric active learning [4] assumed that the decision boundary is the graph of a smooth function, that a margin assumption very similar to Tsybakov's noise assumption [18] holds, and that distribution P_X is uniform. This led to a better guarantee than in passive learning. Instead of the assumption on the decision boundary, other works [20, 15] supposed rather that the regression function is smooth (in some sense). This assumption, along with Tsybakov's noise assumption and strong density assumption also gave a better guarantee than in passive learning. Moreover, unlike in [4], they provided algorithms that are adaptive with respect to the margin's noise and to the smoothness parameters.

However, recent work [5] pointed out some disadvantages of the preceding smoothness assumption, and

*borisedgar.ndjianjike@umons.ac.be

†xavier.siebert@umons.ac.be

extended it in the context of passive learning with k-nearest neighbors (k-nn) by using another smoothness assumption that is able to sharply characterize the rate of convergence for all probability distributions that satisfy it.

In this paper, we thus extend the work of [5] to the active learning setting, and provide a novel algorithm that outputs a classifier with the same rate of convergence as other recent algorithms that were using more restrictive hypotheses, as for example [20, 15].

Section 2 introduces general definitions, Section 3 presents previous related work on convergence rates in active and passive non-parametric learning, with a special emphasis on the assumptions related to our work. Section 4 describes our algorithm, Section 5 provides the theoretical motivations behind our algorithm and proofs of some results. Finally, Section 6 concludes this paper with a discussion of possible extensions of this work.

2 Preliminaries

We begin with some general definitions and notations about active learning in binary classification, then summarize the main assumptions that are typically used to study the rate of convergence of active learning algorithms in the framework of statistical learning theory.

2.1 Active learning setting

Let $(\mathcal{X}, \rho)$ a metric space. In this paper we set $\mathcal{X} = \mathbb{R}^d$ and refer to it as the instance space, and ρ the Euclidean metric. Let $\mathcal{Y} = \{0, 1\}$ the label space. We assume that the couples (X, Y) are random variables distributed according to an unknown probability P over $\mathcal{X} \times \mathcal{Y}$. Let us denote P_X the marginal distribution of P over $\mathcal{X}$.

Given $w \in \mathbb{N}$ and an i.i.d sample $(X_1, Y_1), \ldots, (X_w, Y_w)$ drawn according to probability P, the learning problem consists in minimizing the risk $\mathcal{R}(f) = P(Y \neq f(X))$ over all measurable functions, called classifiers $f : \mathcal{X} \to \mathcal{Y}$.

In active learning, the labels are not available from the beginning but we can request iteratively at a certain cost (to a so-called oracle) a given number n of samples, called the budget ($n \leq w$). In passive learning, all labels are available from the beginning, and $n = w$. At any time, we choose to request the label of a point X according to the previous observations. The point X is chosen to be most "informative", which amounts to belonging to a region where classification is difficult and requires more labeled data to be collected. Therefore, the goal of active learning is to design a sampling strategy that outputs a classifier $\widehat{f}_n$ whose excess risk is as small as possible with high probability over the requested samples, as reviewed in [6, 10, 7].

Given x in $\mathcal{X}$, let us introduce $\eta(x) = \mathbb{E}(Y|X = x) = P(Y = 1 | X = x)$ the regression function. As done in [17], it is easy to show that the function $f^*(x) = \mathbb{1}_{\eta(x) \geq 1/2}$ achieves the minimum risk and that $\mathcal{R}(f^*) = \mathbb{E}_X(\min(\eta(X), 1 - \eta(X)))$. Because P is unknown, the function f^* is unreachable and thus the aim of a learning algorithm is to return a classifier $\widehat{f}_n$ with minimum excess risk $\mathcal{R}(\widehat{f}_n) - \mathcal{R}(f^*)$ with high probability over the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$.

2.2 k nearest neighbors (k-nn) classifier

Given two integers k, n such that $k < n$, and a test point $X \in \mathcal{X}$, the k-nn classifier predicts the label of X by giving the majority vote of its k nearest neighbors amongst the sample $X_1, \ldots, X_n$. For $k = 1$, the k-nn classifier returns the label of the nearest neighbor of X amongst the sample $X_1, \ldots, X_n$. Often k grows with n, in which case the method is called k_n-nn. For a complete discussion of nearest neighbors classification, see for example [3, 5].

2.3 Regularity, noise and strong density assumptions

Let $B(x, r) = \{x' \in \mathcal{X},\ \rho(x, x') \leq r\}$ and $B^o(x, r) = \{x' \in \mathcal{X},\ \rho(x, x') < r\}$ the closed and open balls (with respect to the Euclidean metric ρ), respectively, 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.

Definition 1 (Hölder-continuity)**.**
*Let $\eta : \mathcal{X} \to [0,1]$ the regression function. We say that η is (α, L)-**Hölder continuous** $(0 < \alpha \leq 1$, and $L > 0)$ if $\forall\ x, x' \in \mathcal{X}$,*

$$|\eta(x) - \eta(x')| \leq L\rho(x,x')^{\alpha}. \tag{H1}$$

The notion of Hölder continuity ensures that the proximity between two closest (according to the metric ρ) points is reflected in a similar value for the conditional probability $\eta(x)$.
This definition remains true for a general metric space, but for the case where ρ is the Euclidean metric, we should always have $0 < \alpha \leq 1$, otherwise η becomes constant.

Definition 2 (Strong density)**.**
*Let P the distribution probability defined over $\mathcal{X} \times \mathcal{Y}$ and P_X the marginal distribution of P over $\mathcal{X}$. We say that P satisfies the **strong density** assumption if there exists some constants $r_0 > 0$, $c_0 > 0$, $p_{min} > 0$ such that for all $x \in supp(P_X)$:*

$$\begin{aligned} &\lambda(B(x,r) \cap supp(P_X)) \geq c_0 \lambda(B(x,r)),\ \forall r \leq r_0 \\ &and\ p_X(x) > p_{min}. \end{aligned} \tag{H2}$$

where p_X is the density function of the marginal distribution P_X and λ is the Lebesgue measure.

The strong density assumption ensures that, given a realisation $X = x$ according to P_X, there exists an infinite number of realisations $X_1 = x_1, \ldots, X_m = x_m, \ldots$ in a neighborhood of x.

Definition 3 (Margin noise)**.**
*We say that P satisfies **margin noise** or **Tsybakov's noise** assumption with parameter $\beta \geq 0$ if for all $0 < \epsilon \leq 1$*

$$P_X(x \in \mathcal{X},\ |\eta(x) - 1/2| < \epsilon) < C\epsilon^{\beta}, \tag{H3}$$

for $C := C(\beta) \in [1, +\infty[$.

The margin noise assumption gives a bound on the probability that the label of the training points in the neigborhood of a test point x differs from the label of x given by the conditional probability $\eta(x)$. It also describes the behavior of the regression function in the vicinity of the decision boundary $\eta(x) = \frac{1}{2}$. When β goes to infinity, we observe a "jump" of η around to the decision boundary, and then we obtain Massart's noise condition [19]. Small values of β allow for η to "cuddle" $\frac{1}{2}$ when we approach the decision boundary.

Definition 4 ((α, L)-smooth)**.**
*Let $0 < \alpha \leq 1$ and $L > 1$. The regression function is (α, L)-**smooth** if for all $x, z \in supp(P_X)$ we have:*

$$|\eta(x) - \eta(z)| \leq L.P_X(B^o(x, \rho(x,z)))^{\alpha/d}. \tag{H4}$$

Theorem 1 states that the (α, L)-smooth assumption (H4) is more general than the Hölder continuity assumption (H1).

Theorem 1. *[5]*
Suppose that $\mathcal{X} \subset \mathbb{R}^d$, that the regression function η is (α_h, L_h)-Hölder continuous, and that P_X satisfies H2. Then there is a constant $L > 1$ such that for any $x, z \in supp(P_X)$, we have:

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

3 Convergence rates in nonparametric active learning

3.1 Previous works

Active learning theory has been mostly studied during the last decades in a parametric setting, see for example [2, 11, 7] and references therein. One of the pioneering works studying the achievable limits in active learning in a nonparametric setting [4] required that the decision boundary is the graph of a Hölder continuous function with parameter α (H1). Using a notion of margin's noise (with parameter β) very similar to (H3), the following minimax rate was obtained:

$$O\left(n^{-\frac{\beta}{2\beta+\gamma-2}}\right), \tag{1}$$

where $\gamma = \frac{d-1}{\alpha}$ and d is the dimension of instance space ($\mathcal{X} = \mathbb{R}^d$).

Note that this result assumes the knowledge of smoothness and margin's noise parameters, whereas an algorithm that achieves the same rate, but that adapts to these parameters was proposed recently in [16].

In this paper, we consider the case where the smoothness assumption refers to the regression function both in passive and in active learning.

In passive learning, by assuming that the regression function is Hölder continuous (H1), along with (H3) and (H2), the minimax rate was established by [1] :

$$O\left(n^{-\frac{\alpha(\beta+1)}{2\alpha+d}}\right). \tag{2}$$

In active learning, using the same assumptions (H1), (H3) and (H2), with the additional condition $\alpha\beta < d$, the following minimax rate was obtained [15] :

$$\tilde{O}\left(n^{-\frac{\alpha(\beta+1)}{2\alpha+d-\alpha\beta}}\right), \tag{3}$$

where $\tilde{O}$ indicates that there may be additional logarithmic factors. This active learning rate given by (3) thus represents an improvement over the passive learning rate (2) that uses the same hypotheses.

With another assumption on the regression function relating the L_2 and L_∞ approximation losses of certain piecewise constant or polynomial approximations of η in the vicinity of the decision boundary, the same rate (3) was also obtained by [20].

3.2 Link with k-nn classifiers

For practicals applications, an interesting question is if k-nn classifiers attain the rate given by (2) in passive learning and by (3) in active learning.

In passive learning, under assumptions (H1), (H3) and (H2), and for suitable k_n, it was shown in [5] that k_n-nn indeed achieves the rate (2).

In active learning a pool-based algorithm that outputs a k-nn classifier has been proposed in [14], but its assumptions differ from ours in terms of smoothness and noise, and the number of queries is constant. Similarly, the algorithm proposed in [9] outputs a 1-nn classifier based on a subsample of the initial pool, such that the label of each instance of this subsample is determined with high probability by the labels of its neighbors. The number of neighbors is adaptively chosen for each instance in the subsample, leading to the minimax rate (3) under the same assumptions as in [15].

To obtain more general results on the rate of convergence for k-nn classifiers in metric spaces under minimal assumptions, the more general smoothness assumption given by (H4) was used in [5]. By using a k-nn algorithm, and under assumptions (H3) and (H4), the rate of convergence obtained in [5] is also on the order of (2).

Additionally, this rate avoids the strong density assumption (H2) and therefore allows more classes of probability. In addition, the α-smooth assumption is more universal than Hölder continuity assumption. It just holds for any pair of distributions P_X and η. In Hölder continuity, strong density assumption implicitly assumes the existence of the density p_X, and according to (H2), it also implies that the support of P_X has finite Lebesgue measure; this is very restrictive and excludes important densities like Gaussian densities as noticed in [8].

3.3 Contributions of the current work

In this paper, we provide an active learning algorithm under the assumptions (H4) and (H3) that were used in passive learning in [5]. The α-smooth assumption (H4) involves a dependence on the marginal distribution P_X unlike the Hölder continuity assumption (H1).

In the following, we will show that the rate of convergence of our algorithm remains the same as (3), despite the use of more general hypotheses.

4 KALLS algorithm

4.1 Setting

As explained in Section 2.1, we consider a pool of i.i.d unlabeled examples $\mathcal{K} = \{X_1, X_2, \dots, X_w\}$. Let $n \leq w$ the budget, that is the maximum number of points whose label we are allowed to query to the oracle. The objective of the algorithm is to build a subsample $\{X_{t_i},\ i \geq 1\}$ whose labels are considered most "informative", and which we call the *active set*. More precisely, a point X_{t_i} is considered "informative" if its label cannot be inferred from the previous observations X_{t_j}(with $t_j < t_i$). The sequence $(t_i)_{i\geq 1}$ of indices is an increasing sequence of integers, starting arbitrarily with $X_{t_1} = X_1$ and stopping when the budget n is attained or when $X_{t_i} = X_w$ for some t_i.

When a point $\{X_{t_i}\}$ is considered informative, instead of requesting its label, we request the labels of its nearest neighbors, as was done in [9]. This differs from the setting of [16], where the label of X_{t_i} is requested several times. This is reasonable for practical situations where the uncertainty about the label of X_{t_i} has to be overcome, and it is related to the (α, L)-smooth assumption (H4). The number of neighbors k_{t_i} is adaptively determined such that with high confidence while respecting the budget, we can predict the true label as $f^*(X_{t_i})$ of X_{t_i} by empirical mean of the labels of his k_{t_i} nearest neighbors.

The final active set output by the algorithm will thus be $\widehat{S} = \{(X_{t_1}, \widehat{Y}_{t_1}), \dots, (X_{\widehat{t}_l}, \widehat{Y}_{\widehat{t}_l})\}$ with $\widehat{t}_l \leq w$. This set $\widehat{S}$ is the set of points considered to be "informative" by removing the points that are too noisy and thus that require many more labels. We show that the active set $\widehat{S}$ is sufficient to predict the label of any new point by a 1-nn classification rule $\widehat{f}_{n,w}$.

Before beginning the description of our algorithm, let us introduce some variables and notations: For $\epsilon, \delta \in (0,1)$, $k \geq 1$, set:

$$b_{\delta,k} = \sqrt{\frac{2}{k}\left(\log\left(\frac{1}{\delta}\right) + \log\log\left(\frac{1}{\delta}\right) + \log\log(ek)\right)}. \tag{4}$$

For $\epsilon, \delta \in (0,1)$, $s \geq 1$

$$k(\epsilon, \delta) = \frac{c}{\Delta^2}\left[\log(\frac{1}{\delta}) + \log\log(\frac{1}{\delta}) + \log\log\left(\frac{512\sqrt{e}}{\Delta}\right)\right] \tag{5}$$

where

$$\Delta = \max(\frac{\epsilon}{2}, \left(\frac{\epsilon}{2C}\right)^{\frac{1}{\beta+1}}), \quad c \geq 7.10^6. \tag{6}$$

Let

$$\phi_n = \sqrt{\frac{1}{n}\left(\log\left(\frac{1}{\delta}\right) + \log\log\left(\frac{1}{\delta}\right)\right)}. \tag{7}$$

For $X_s \in \mathcal{K} = \{X_1, \dots, X_w\}$, we denote henceforth 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

$$\widehat{\eta}_k(X_s) = \frac{1}{k}\sum_{i=1}^{k} Y_s^{(i)}, \quad \bar{\eta}_k(X_s) = \frac{1}{k}\sum_{i=1}^{k} \eta(X_s^{(i)}). \tag{8}$$

4.2 Algorithm

Below we provide a description of the KALLS algorithm (Algorithm 1), that aims at determining the active set defined in Section 4.1 and the related 1-nn classifier $\widehat{f}_{n,w}$ under the assumptions (H4) and (H3). The complete proofs of the convergence of the algorithm are in Section 5.

For the algorithm KALLS, the inputs are a pool $\mathcal{K}$ of unlabelled data of size w, the budget n, the smoothness parameters α and L from (H4), the margin noise parameters β, C from (H3), a confidence parameter $\delta \in (0,1)$ and an accuracy parameter $\epsilon \in (0,1)$. For the moment, these parameters are fixed from the beginning. Adaptive algorithms such as [15] could be exploited, in particular for the α and β parameters.

The final active set is obtained such that, with high confidence, the 1-nn classifier $\widehat{f}_{n,w}$ based on it agrees with the Bayes classifier at points that lie beyond some margin $\Delta_o > 0$ of the decision boundary.

Formally, given $x \in \mathcal{X}$ such that $|\eta(x) - 1/2| > \Delta_0$, we have $\widehat{f}_{n,w}(x) = \mathbb{1}_{\eta(x) \geq 1/2}$ with high confidence. We will show that, with a suitable choice of Δ_o, the hypothesis (H3) leads to the desired rate of convergence (3).

Algorithm 1: k-nn Active Learning under Local Smoothness (KALLS)

Input: a pool $\mathcal{K} = \{X_1, \ldots, X_w\}$, label budget n, smoothness parameters α, L, margin noise parameters β, C, confidence parameter δ, accuracy parameter ϵ.

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

$s = 1$ ▷ index of point currently examined

$\widehat{\mathcal{S}} = \emptyset$ ▷ current active set

$t = n$ ▷ current label budget

$I = \emptyset$ ▷ set of "informative points"; used for providing the label complexity

for $s \leq w$ **do**

  Let $\delta_s = \frac{\delta}{32 s^2}$

while $t > 0$ *and* $s < w$ **do**

  T=`Reliable`$(X_s, \delta_s, \alpha, L, I)$

  if T=*True* **then**

    $s = s + 1$

  else

    $[\widehat{Y}, Q_s]$=`confidentLabel`$(X_s, k(\epsilon, \delta_s), t, \delta)$

    $\widehat{LB}_s = \left| \frac{1}{|Q_s|} \sum_{(X,Y) \in Q_s} Y - \frac{1}{2} \right| - b_{\delta_s, |Q_s|}$ ▷ Lower bound guarantee on $|\eta(X_s) - \frac{1}{2}|$

    $I = I \cup \{(X_s, \widehat{LB}_s, |Q_s|)\}$

    $t = t - |Q_s|$

    if $\widehat{LB}_s \geq 0$ **then**

      $\widehat{\mathcal{S}} = \widehat{\mathcal{S}} \cup \{(X_s, \widehat{Y})\}$

$\widehat{f}_{n,w} \leftarrow$ `Learn` $(\widehat{\mathcal{S}})$

KALLS (Algorithm 1) uses two main subroutines : `Reliable` and `ConfidentLabel`, which are detailed below (Sections 4.4 and 4.3, respectively). It finally outputs a 1-nn classifier $\widehat{f}_{n,w}$ using a subroutine called `Learn`.

4.3 `ConfidentLabel` subroutine

If the point X_s is considered informative, the `confidentLabel` subroutine is used to determine with a given level of confidence, the label of the current point X_s. This is done by using the labels of its $k(\epsilon, \delta_s)$ nearest neighbors, where $k(\epsilon, \delta_s)$ is chosen such that, with high probability, the empirical majority of

$k(\epsilon, \delta_s)s$ labels differs from the majority in expectation by less than some margin, and all the $k(\epsilon, \delta_s)$ nearest neighbors are at most at some distance from X_s.

Algorithm 2: `confidentLabel` subroutine

Input: an instance X, integer k', budget parameter $t \geq 1$, confidence parameter δ.
Output: $\widehat{Y}$, Q

$Q = \emptyset$
$k = 1$
while $k \leq \min(k', t)$ **do**
  Request the label $Y^{(k)}$ of $X^{(k)}$
  $Q = Q \cup \{(X^{(k)}, Y^{(k)})\}$
  $k = k + 1$
  if $\left|\frac{1}{k}\sum_{i=1}^{k} Y^{(i)} - \frac{1}{2}\right| > 2b_{\delta,k}$ **then**
    exit ▷ cut-off condition
  else
    Return in step 4
$\widehat{\eta} \leftarrow \frac{1}{|Q|}\sum_{(X,Y)\in Q} Y$
$\widehat{Y} = \mathbb{1}_{\widehat{\eta}\geq 1/2}$

4.4 `Reliable` subroutine

The `Reliable` subroutine is a binary test about the point X_s currently considered, to verify if the label of X_s can be inferred with high confidence using the labels of the points currently in the active set. If it is the case, the point X_s is not considered to be informative, its label is not requested and it is not added to the active set.
When a point X_s is relatively far away from the decision boundary, the subroutine `ConfidentLabel` provides a lower confidence bound $O(\widehat{LB}_s) \leq |\eta(X_s) - \frac{1}{2}|$; for a new point X_t, we have a low degree of uncertainty (and then uninformative) if $|\eta(X_t) - \frac{1}{2}|$ entails the same confidence lower bound $O(\widehat{LB}_s)$ (for some previous informative point X_s). We can see that by smoothness assumption, it suffices to have $P_X(B(X_t, \rho(X_t, X_s)) \leq O((\widehat{LB}_s)^{d/\alpha})$. Because P_X is unknown, we use the subroutine `BerEst` to adaptively estimate with hight probability (over the data) $P_X(B(X_t, \rho(X_t, X_s))$ up to $O((\widehat{LB}_s)^{d/\alpha})$.

Algorithm 3: `Reliable` subroutine

Input: an instance X, a confidence parameter δ, smoothness parameters α, L, a set $I \subset \mathcal{X} \times \mathbb{R} \times \mathbb{N}$
Output: T

for $(X', c, k) \in I$ **do**
  if $c \geq 0$ **then**
    $\widehat{p}_{X'} = \texttt{Estprob}\left(\rho(X, X'), \left(\frac{c}{64L}\right)^{d/\alpha}, 50, \delta\right)$
if $\exists\ (X', c, k) \in I$ *such that* $c > 0.1b_{\delta,k}$ *and* $\widehat{p}_{X'} \leq \frac{75}{94}\left(\frac{c}{64L}\right)^{d/\alpha}$ **then**
  $T = True$
else
  $T = False$

The `Reliable` subroutine uses $\texttt{EstProb}(r, \epsilon_o, 50, \delta)$ as follows:

1. Call the subroutine `BerEst`$(\epsilon_o, \delta, 50)$.

2. To draw a single p_i in `BerEst`$(\epsilon_o, \delta, 50)$, sample randomly an example X_i from $\mathcal{K}$, and set $p_i = \mathbb{1}_{X_i \in B(X,r)}$.

Algorithm 4: `BerEst` subroutine (Bernoulli Estimation)

Input: accuracy parameter ϵ_o, confidence parameter δ', budget parameter u. ▷ u does not depend on the label budget n

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 $\widehat{p}$

4.5 `Learn` subroutine

The `Learn` subroutine takes as input the set of points that were considered informative and relatively less noisy, we apply the passive learning on this subset by using the 1-nn classifier.

Algorithm 5: `Learn` subroutine

Input: $\widehat{\mathcal{S}}$

Output: $\widehat{f}_{n,w}$

$\widehat{f}_{n,w} \leftarrow$ the 1-nn classifier on $\widehat{\mathcal{S}}$

5 Theoretical motivations

This Section provides the main theoretical motivations behind the `KALLS` algorithm, and is organized as follows:

Section 5.2.1 outlines the main ideas of the proof, in Section 5.2.2 we adaptively determine the number of label requests needed to accurately predict the label of an informative point that is relatively far from the boundary decision. In Section 5.2.3, we provide some lemmas that illustrate a sufficient condition for a point to be informative, in Section 5.2.4, we give theorems that allow us to classify each instance relatively far from the decision boundary. Finally in Section 5.2.5, we provide the label complexity and establish Theorem 3.

5.1 Notations

Some notations will be used throughout the proofs are listed here for convenience.

As defined in Section 2.3, let $B(x,r) = \{x' \in \mathcal{X},\ \rho(x,x') \leq r\}$ and $B^o(x,r) = \{x' \in \mathcal{X},\ \rho(x,x') < r\}$ the closed and open balls (with respect to the Euclidean metric ρ), respectively, 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.

Given $\mathcal{K} = \{X_1, \ldots, X_w\}$, Let us denote by $\mathcal{A}_{a,w}$ the set of active learning algorithms on $\mathcal{K}$, and $\mathcal{P}(\alpha, \beta) :=$ the set of probabilities that satisfy the hypotheses (H4) and (H3), where α is the parameter in (H4) and β in (H3). For $A \in \mathcal{A}_{a,w}$, and n the budget, we denote by $\widehat{f}_{A,n,w} := \widehat{f}_{n,w}$ the classifier that is provided by A.

For $p \in (0, 1]$, and $x \in supp(P_X)$, let us define $r_p(x) = \inf\{r > 0, P_X(B(x,r)) \geq p\}$.

5.2 Main ideas of the proof

Theorem 2 is the main result of this paper, which provides bounds on the excess risk for the `KALLS` algorithm.

Theorem 2 (Excess risk for the `KALLS` algorithm.)**.**
Let the set $\mathcal{P}(\alpha, \beta)$ such that $\alpha\beta < d$ where d is the dimension of the input space $\mathcal{X} = \mathbb{R}^d$. Then, we have:

$$\inf_{A \in \mathcal{A}_{a,w}} \sup_{P \in \mathcal{P}(\alpha,\beta)} \mathbb{E}_n \left[R(\widehat{f}_{n,w}) - R(f^*) \right] \leq \tilde{O}\left(n^{\frac{\alpha(\beta+1)}{2\alpha+d-\alpha\beta}} \right). \tag{9}$$

Where $\mathbb{E}_n$ is with respect to the randomness of the algorithm $A \in \mathcal{A}_{a,w}$.

The result (9) is also be stated below (Theorem 3) in a more practical form using label complexity (10). This latter form will be used in the proof.

Theorem 3 (Label complexity for the `KALLS` algorithm.)**.**
Let the set $\mathcal{P}(\alpha, \beta)$ such that $\alpha\beta < d$. Let ϵ, $\delta \in (0, 1)$. for all n, $w \in \mathbb{N}$ such that: if

$$n \geq \tilde{O}\left(\left(\frac{1}{\epsilon} \right)^{\frac{2\alpha+d-\alpha\beta}{\alpha(\beta+1)}} \right), \tag{10}$$

$$w \geq \tilde{O}\left(\left(\frac{1}{\epsilon} \right)^{\frac{2\alpha+d}{\alpha(\beta+1)}} \right) \tag{11}$$

and

$$w \geq \frac{400 \log\left(\frac{12800 w^2}{\delta (\frac{1}{64L}\bar{c}\phi_n)^{d/\alpha}} \right)}{(\frac{1}{64L}\bar{c}\phi_n)^{d/\alpha}}, \tag{12}$$

(where L is defined in (H4)*, $\bar{c} = 0.1$ and ϕ_n is defined by* (7)*)*
then with probability at least $1 - \delta$, we have:

$$\inf_{A \in \mathcal{A}_{a,w}} \sup_{P \in \mathcal{P}(\alpha,\beta)} \left[R(\widehat{f}_{n,w}) - R(f^*) \right] \leq \epsilon. \tag{13}$$

5.2.1 Main idea of the proof

For a classifier $\widehat{f}_{n,w}$, it is well known[17] that the excess of risk is:

$$R(\widehat{f}_{n,w}) - R(f^*) = \int_{\{x,\, \widehat{f}_{n,w}(x) \neq f^*(x)\}} |2\eta(x) - 1| dP_X(x). \tag{14}$$

We thus aim to proof that (10) is a sufficient condition to have with probability $\geq 1 - \delta$, $\widehat{f}_{n,w}$ agrees with f^* on $\{x,\ |\eta(x) - 1/2| > \Delta_o\}$, for $\Delta_o > 0$. Introducing Δ_o in (14) leads to:

$$\begin{aligned} R(\widehat{f}_{n,w}) - R(f^*) &= \int_{\{x,\, \widehat{f}_{n,w}(x) \neq f^*(x)\}} |2\eta(x) - 1| dP_X(x) \\ &\leq 2\Delta_o P_X(|\eta(x) - 1/2| < \Delta_o). \end{aligned}$$

Therefore, if $\Delta_o \leq \frac{\epsilon}{2}$ then, $R(\widehat{f}_{n,w}) - R(f^*) \leq \epsilon$. Otherwise, if $\Delta_o > \frac{\epsilon}{2}$, by the hypothesis (H3), we have $R(\widehat{f}_{n,w}) - R(f^*) \leq 2C\Delta_o^{\beta+1}$. In the latter case, setting $\Delta_o = \left(\frac{\epsilon}{2C}\right)^{\frac{1}{\beta+1}}$ guarantees $R(\widehat{f}_{n,w}) - R(f^*) \leq \epsilon$. Altogether, the value $\Delta_o = \max(\frac{\epsilon}{2}, \left(\frac{\epsilon}{2C}\right)^{\frac{1}{\beta+1}}) = \Delta$ (see (6)) guarantees $R(\widehat{f}_{n,w}) - R(f^*) \leq \epsilon$.

5.2.2 First Part: adaptive label requests on informative points

Lemma 1 (Chernoff [21])**.**
Suppose $X_1, \dots, X_m$ are independent random variables taking value in $\{0,1\}$. Let X denote their sum and $\mu = E(X)$ its expected value. Then, for any $\delta > 0$,

$$P_m(X \leq (1-\delta)\mu) \leq \exp(-\delta^2\mu/2),$$

where P_m is the probability with respect to the sample $X_1, \dots, X_m$.

Lemma 2 (Logarithmic relationship, [23])**.**
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 3 (Chaudhuri and Dasgupta, [5])**.**
For $p \in (0,1]$, and $x \in supp(P_X)$, let us define $r_p(x) = \inf\{r > 0, P_X(B(x,r)) \geq p\}$. For all $p \in (0,1]$, and $x \in supp(P_X)$, we have:

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

Theorem 4.
Let $\epsilon, \delta \in (0,1)$. Set $\Delta = \max(\epsilon, \left(\frac{\epsilon}{2C}\right)^{\frac{1}{\beta+1}})$, and $p_\epsilon = \left(\frac{31\Delta}{1024L}\right)^{d/\alpha}$, where α, L, β, C are parameters used in (H3) *and* (H4)*.*
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 (5) *(where $\delta_s = \frac{\delta}{32s^2}$).*
For $k, s \geq 1$, set $\tau_{k,s} = \sqrt{\frac{2}{k}\log(\frac{32s^2}{\delta})}$. 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

$$k_s \leq (1 - \tau_{k_s,s})p_\epsilon(w-1) \tag{15}$$

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

$$w \geq \tilde{O}\left(\left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d}{\alpha(\beta+1)}}\right) \tag{16}$$

is sufficient to have (15)*.*

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 1 and Lemma 3, and if k_s satisfies (15), we have:

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

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

$$k_s \leq (1 - \tau_{k_s,s})p_\epsilon(w-1),$$

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 $k_s \leq (1 - \tau_{k_s,s})p_\epsilon(w-1)$, 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 (16) is sufficient to guarantee (15): the relation (15) implies $w \geq \frac{k_s}{(1-\tau_{k_s,s})p_\epsilon} + 1$. We can see by a bit calculus, that $\tau_{k_s,s} \leq \frac{1}{2}$, then

$$\begin{aligned}
\frac{k_s}{(1-\tau_{k_s,s})p_\epsilon} + 1 &\leq \frac{2k_s}{p_\epsilon} + 1 \\
&\leq 4\frac{k_s}{p_\epsilon} \qquad \left(\text{because } \frac{k_s}{p_\epsilon} \geq 1\right) \\
&= \frac{4c}{p_\epsilon \Delta^2}\left[\log(\frac{32s^2}{\delta}) + \log\log(\frac{32s^2}{\delta}) + \log\log\left(\frac{512\sqrt{e}}{\Delta}\right)\right] \\
&= \frac{b}{\Delta^{2+\frac{d}{\alpha}}}\left[\log(\frac{32s^2}{\delta}) + \log\log(\frac{32s^2}{\delta}) + \log\log\left(\frac{512\sqrt{e}}{\Delta}\right)\right] \\
&\qquad \text{where } b = \left(\frac{1024L}{31}\right)^{d/\alpha}.4c \\
&\leq \bar{C}\left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d}{\alpha(\beta+1)}}\left[\log(\frac{32s^2}{\delta}) + \log\log(\frac{32s^2}{\delta}) + \log\log\left(\frac{512\sqrt{e}}{\Delta}\right)\right] \\
&\qquad \text{as } \Delta = \max(\epsilon, \left(\frac{\epsilon}{2C}\right)^{\frac{1}{\beta+1}}), \text{ where } \bar{C} = b(2C)^{\frac{2\alpha+d}{\alpha(\beta+1)}} \\
&\leq \bar{C}\left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d}{\alpha(\beta+1)}}\left[2\log(\frac{32s^2}{\delta}) + \log\left(\frac{512\sqrt{e}}{\epsilon}\right)\right] \\
&\qquad \text{as} \log(x) \leq x, \text{ and } \Delta \geq \epsilon \\
&\leq 2\bar{C}\left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d}{\alpha(\beta+1)}}\left[\log(s^2) + \log\left(\frac{16384\sqrt{e}}{\delta\epsilon}\right)\right] \\
&\leq 4\bar{C}\left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d}{\alpha(\beta+1)}}\left[\log(s) + \log\left(\frac{16384\sqrt{e}}{\delta\epsilon}\right)\right] \\
&\leq 4\bar{C}\left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d}{\alpha(\beta+1)}}\left[\log(w) + \log\left(\frac{16384\sqrt{e}}{\delta\epsilon}\right)\right]
\end{aligned} \tag{17}$$

Now, we are going to apply the lemma 2. If we set in lemma 2

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

we can easily see that $c \geq a$, $a \geq 4$ and then

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

then, the relation

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

is sufficient to guarantee (17). □

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

Motivation for choosing k_s for X_s

In The following, we give, a justification for the choice of $k_s := k(\epsilon, \delta_s)$ (5). We also prove that, if the budget allows us, under a margin condition, the cut-off condition used in Algorithm 2

$$\left| \frac{1}{k} \sum_{i=1}^{k} Y_s^{(i)} - \frac{1}{2} \right| \leq c_5 b_{\delta_s, k}$$

will be violated with a number of label requests lower than $k(\epsilon, \delta_s)$. The intuition behind this, is with theorem 5, to adapt (with respect to the noise) the number of label requested; that is to say, fewer label requests on a less-noisy point, and most label requests on a noisy point. This provide a significant savings in the number of request needed to predict with high probability the right label.

Lemma 4. *[12]*
Let X be a random variable with $E(X) = 0$, $a \leq X \leq b$, then for $v > 0$,

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

Lemma 5. *[13]*
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 6.
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. □

Theorem 5.
Let $\delta \in (0,1)$, and $\epsilon \in (0,1)$. Let us assume that w satisfies (11). For X_s, set $\tilde{k}(\epsilon, \delta_s)$ (with $\delta_s = \frac{\delta}{32s^2}$) as

$$\tilde{k}(\epsilon, \delta_s) = \frac{c}{4|\eta(X_s) - \frac{1}{2}|^2} \left[\log(\frac{32s^2}{\delta}) + \log\log(\frac{32s^2}{\delta}) + \log\log\left(\frac{256\sqrt{e}}{|\eta(X_s) - \frac{1}{2}|} \right) \right],$$

where $c \geq 7.10^6$. For $k \geq 1$, $s \leq w$, let $\Delta = \max(\frac{\epsilon}{2}, \left(\frac{\epsilon}{2C}\right)^{\frac{1}{\beta+1}})$ and $b_{\delta_s,k}$ defined in (4).
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 (8), for all $s \in \{1, \ldots, w\}$,*

$$|\widehat{\eta}_k(X_s) - \bar{\eta}_k(X_s)| \ \leq b_{\delta_s,k}. \tag{18}$$

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

$$|\frac{1}{\bar{k}_s}\sum_{i=1}^{\bar{k}_s} Y_s^{(i)} - \frac{1}{2}| \geq 2b_{\delta_s, \bar{k}_s} \tag{19}$$

and

$$f^*(X_s) = \mathbb{1}_{\widehat{\eta}_{\bar{k}_s}(X_s) \geq \frac{1}{2}}, \tag{20}$$

Where $\bar{k}_s$ is the number of requests made in $\texttt{ConfidentLabel}(X_s)$.

Proof.

1. Let us begin with the proof of the first item.
Here, we follow the proof of Theorem 8 in [13], with few additional modifications.
Let $s \in \{1, \ldots, w\}$. Set $S_k = \sum_{i=1}^{k} \left(Y_s^{(i)} - \eta(X_s^{(i)})\right)$. Given $\{X_1, \ldots, X_w\}$, $E(Y_s^{(k)} - \eta(X_s^{(k)})) = 0$, and the random variables $\left\{Y_s^{(i)} - \eta(X_s^{(i)}),\ i = 1, \ldots, k\right\}$ are independent. Then by lemma 4, given $\{X_1, \ldots, X_w\}$, as $Y_s^{(1)} - \eta(X_s^{(1)})$ takes values in $[-1, 1]$, we have $E(e^{v(Y_s^{(1)} - \eta(X_s^{(1)}))}) \leq e^{v^2/2}$ for all $v > 0$. Furthermore, set $z = \log(\frac{32s^2}{\delta})$, and $r = z + 3\log(z)$. We have $r \geq \frac{8}{(e-1)^2}$, and by lemma 5, with $\gamma = 3/2$, we have:

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

It can be shown numerically that for $z \geq 2.03$, (it 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 $s \in \{1, \ldots, w\}$, there exists an event $A'_{2,s}$ such that $P(A'_{2,s}) \geq 1 - \delta/32s^2$, and simultaneously for all $k \geq 1$, we have:

$$|S_k| \leq \sqrt{2k\left(\log\left(\frac{32s^2}{\delta}\right) + \log\log\left(\frac{32s^2}{\delta}\right) + \log\log(ek)\right)}.$$

By setting $A'_2 = \cap_{s \geq 1} A'_{2,s}$, we have $P(A'_2) \geq 1 - \delta/16$, and on A'_2, we have for all $s \in \{1, \ldots, w\}$, for all $k \geq 1$,

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

2. Finally, we are going to show that there exists and event A''_2 such that the item 2 holds on $A'_2 \cap A''_2 \cap A_1$.
Given $\{X_1, \ldots, X_w\}$, and $X_s \in \{X_1, \ldots, X_w\}$, by Hoeffding's inequality, there exists an event $A''_{2,s}$, with $P(A''_{2,s}) \geq 1 - \delta/32s^2$, and on $A''_{2,s}$, we have:

$$|\widehat{\eta}_k(X_s) - \bar{\eta}_k(X_s)| \leq \sqrt{\frac{2\log(\frac{32s^2}{\delta})}{k}}.$$

This implies that:

$$|\widehat{\eta}_k(X_s) - \frac{1}{2}| \geq |\bar{\eta}_k(X_s) - \frac{1}{2}| - \sqrt{\frac{2\log(\frac{32s^2}{\delta})}{k}}. \tag{21}$$

On the event A_1, we have, for all $k \leq k_s$, by α-smoothness assumption (H4),

$$|\eta(X_s) - \eta(X_s^{(k)})| \leq \frac{31}{1024}\Delta. \tag{22}$$

And then, if $|\eta(X_s) - \frac{1}{2}| \geq \frac{1}{2}\Delta$, then $|\eta(X_s) - \frac{1}{2}| \geq \frac{1}{32}\Delta$. (22) becomes

$$|\eta(X_s^{(k)}) - \frac{1}{2}| \geq \frac{1}{1024}|\eta(X_s) - \frac{1}{2}|.$$

Then (21) becomes:

$$|\widehat{\eta}_k(X_s) - \frac{1}{2}| \geq \frac{1}{1024}|\eta(X_s) - \frac{1}{2}| - \sqrt{\frac{2\log(\frac{32s^2}{\delta})}{k}}. \tag{23}$$

A sufficient condition for k to satisfy (19), is

$$\frac{1}{1024}|\eta(X_s) - \frac{1}{2}| - \sqrt{\frac{2\log(\frac{32s^2}{\delta})}{k}} \geq 2b_{\delta_s,k}$$

and then:

$$\frac{1}{1024}|\eta(X_s) - \frac{1}{2}| - \sqrt{\frac{2\log(\frac{32s^2}{\delta})}{k}} \geq 2\sqrt{\frac{2}{k}\left(\log\left(\frac{32s^2}{\delta}\right) + \log\log\left(\frac{32s^2}{\delta}\right) + \log\log(ek)\right)}$$

this implies:

$$k \geq \frac{1024}{|\eta(X_s) - \frac{1}{2}|^2}\left(\sqrt{2\log(\frac{32s^2}{\delta})} + 2\sqrt{2\left(\log\left(\frac{32s^2}{\delta}\right) + \log\log\left(\frac{32s^2}{\delta}\right) + \log\log(ek)\right)}\right)^2. \tag{24}$$

On the other hand, the right hand side is smaller than:

$$\frac{1024}{|\eta(X_s) - \frac{1}{2}|^2}\left(\sqrt{2\log(\frac{32s^2}{\delta})} + 2\sqrt{2\log\left(\frac{32s^2}{\delta}\right)} + 2\sqrt{2\log\log\left(\frac{32s^2}{\delta}\right)} + 2\sqrt{2\log\log(ek)}\right)^2.$$

To deduce (24), it suffices to have the expression into the brackets is lower than:

$$\frac{\sqrt{k}}{32}|\eta(X_s) - \frac{1}{2}|.$$

Then, it suffices to have simultaneously:

$$\sqrt{2\log(\frac{32s^2}{\delta})} \leq \frac{1}{9}\frac{\sqrt{k}}{32}|\eta(X_s) - \frac{1}{2}|$$

$$\sqrt{2\log\log(\frac{32s^2}{\delta})} \leq \frac{1}{6}\frac{\sqrt{k}}{32}|\eta(X_s) - \frac{1}{2}|$$

$$\sqrt{2\log\log(ek)} \leq \frac{1}{6}\frac{\sqrt{k}}{32}|\eta(X_s) - \frac{1}{2}|$$

Equivalently, we have:

$$k \geq \frac{1024}{|\eta(X_s) - \frac{1}{2}|^2}\,162\log(\frac{32s^2}{\delta}) \tag{25}$$

$$k \geq \frac{1024}{|\eta(X_s) - \frac{1}{2}|^2} 72 \log\log(\frac{32 s^2}{\delta}) \tag{26}$$

$$k \geq \frac{1024}{|\eta(X_s) - \frac{1}{2}|^2} 72 \log\log(ek) \tag{27}$$

We can apply the lemma 6 in (27) by taking: $m = ek$ and $u = \frac{73728e}{|\eta(X_s)-\frac{1}{2}|^2}$. we have $m \geq 1$ and $u \geq 20$ and then, a sufficient condition to have (27) is:

$$k \geq 2\frac{73728e}{|\eta(X_s) - \frac{1}{2}|^2} \log\log\left(\frac{73728e}{|\eta(X_s) - \frac{1}{2}|^2}\right)$$

or

$$k \geq 4\frac{73728e}{|\eta(X_s) - \frac{1}{2}|^2} \log\log\left(\frac{\sqrt{73728e}}{|\eta(X_s) - \frac{1}{2}|}\right) \tag{28}$$

We can easily see that $\tilde{k}_s := \tilde{k}(\epsilon, \delta_s)$ satisfies (25), (26), (28). then

$$|\frac{1}{\tilde{k}_s} \sum_{i=1}^{\tilde{k}_s} Y_s^{(i)} - \frac{1}{2}| \geq 2 b_{\delta_s, \tilde{k}_s}. \tag{29}$$

As $|\eta(X_s) - \frac{1}{2}| \geq \frac{1}{2}\Delta$, we can easily see that $\tilde{k}(\epsilon, \delta_s) \leq k(\epsilon, \delta_s)$. By taking the minimum value $\bar{k}_s = \bar{k}(\epsilon, \delta_s)$ that satisfies (29), we can see that when the budget allows us, the subroutine `ConfidentLabel` requests $\bar{k}_s$ labels, and we have:

$$|\frac{1}{\bar{k}_s} \sum_{i=1}^{\bar{k}_s} Y_s^i - \frac{1}{2}| \geq 2 b_{\delta_s, \bar{k}_s}. \tag{30}$$

By setting $A_2'' = \cap_{s \geq 1} A_{2,s}''$, we have $P(A_2'') \geq 1 - \delta/16$, and we can deduce (19).

We have on A_2', for all $s \leq w$, $k \leq k(\epsilon, \delta_s)$,

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

And then, on $A_1 \cap A_2'$, we have for all $s \leq w$, $k \leq k(\epsilon, \delta_s)$:

$$\begin{aligned} |\eta(X_s) - \widehat{\eta}_k(X_s)| &\leq |\eta(X_s) - \bar{\eta}_k(X_s)| + |\bar{\eta}(X_s) - \widehat{\eta}_k(X_s)| \\ &\leq \frac{31}{1024}\Delta + b_{\delta_s, k} \end{aligned} \tag{31}$$

Assume without loss of generality that $\eta(X_s) \geq \frac{1}{2}$, which leads to:

$$\begin{aligned} \widehat{\eta}_{\bar{k}_s}(X_s) - \frac{1}{2} &= \widehat{\eta}_{\bar{k}_s}(X_s) - \eta(X_s) + \eta(X_s) - \frac{1}{2} \\ &\geq -|\widehat{\eta}_{\bar{k}_s}(X_s) - \eta(X_s)| + \eta(X_s) - \frac{1}{2}. \end{aligned} \tag{32}$$

If $\eta(X_s) - \frac{1}{2} \geq \frac{1}{2}\Delta$, with (31), the expression (32) becomes:

$$\begin{aligned}\widehat{\eta}_{\bar{k}_s}(X_s) - \frac{1}{2} &\geq -\frac{31}{1024}\Delta - b_{\delta_s,\bar{k}_s} + \frac{1}{2}\Delta \\ &= \frac{481}{1024}\Delta - b_{\delta_s,\bar{k}_s} \\ &\geq -b_{\delta_s,\bar{k}_s} \end{aligned} \tag{33}$$

On the other hand, we have by (19),

$$|\widehat{\eta}_{\bar{k}_s}(X_s) - \frac{1}{2}| \geq 2b_{\delta_s,\bar{k}_s},$$

that is to say:

$$\widehat{\eta}_{\bar{k}_s}(X_s) - \frac{1}{2} \geq 2b_{\delta_s,\bar{k}_s} \quad \text{or} \quad \widehat{\eta}_{\bar{k}_s}(X_s) - \frac{1}{2} \leq -2b_{\delta_s,\bar{k}_s}.$$

By (33), we have necessarily $\widehat{\eta}_{\bar{k}_s}(X_s) - \frac{1}{2} \geq 2b_{\delta_s,\bar{k}_s}$, and then:

$$\widehat{\eta}_{\bar{k}_s} - \frac{1}{2} \geq \max(-b_{\delta_s,\bar{k}_s}, 2b_{\delta_s,\bar{k}_s}) = 2b_{\delta_s,\bar{k}_s} \geq 0,$$

Thus we can easily deduce (20).
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.

□

5.2.3 Second part: sufficient condition to be an informative point

As noticed in Section 4.4, a sufficient condition for a point X_t to be considered as an uninformative point is:

$$P_X(B(X_t, \rho(X_t, X_s))) \leq O((\widehat{LB}_s)^{d/\alpha}). \tag{34}$$

for some previous informative point X_s (with $\widehat{LB}_s > 0$ defined in Lemma 8). As P_X is unknown, we provide a computational scheme sufficient to obtain (34).
Firstly we follow the general procedure used in [14], which can be able to estimate adaptively the expectation of a Bernoulli random variable. And secondly, we apply it to the Bernoulli variable $\mathbb{1}_A$ where $A = \{x,\ x \in B(X_t, r)\}$

Lemma 7. *[14]*
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`*(ϵ_o, δ', 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 $p \leq \epsilon_o$.*
2. *Let $\psi := \max(\epsilon_o, \frac{p}{g(t)})$. The number of random draws in the* `BerEst` *subroutine (Algorithm 4) is at most $\frac{8t\log(\frac{8t}{\delta'\psi})}{\psi}$.*

Lemma 8.
Let $\epsilon,\ \delta \in (0,1)$, $r > 0$. As previously, for $k \geq 1$, let be defined

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

For $(X_s, \widehat{LB}_s, |Q_s|) \in I$, (where I is the set defined in `KALLS`*), where*

$$\widehat{LB}_s = \left| \frac{1}{|Q_s|} \sum_{(X,Y)\in Q_s} Y - \frac{1}{2} \right| - b_{\delta_s, |Q_s|}$$

and Q_s is defined in subroutine `ConfidentLabel` *(Algorithm 2). Let us assume that w satisfies* (12) *There exists and 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 $1 \leq s' \leq s$ and $(X_{s'}, \widehat{LB}_{s'}, |Q_{s'}|) \in I$, such that:

$$\widehat{LB}_{s'} \geq \bar{c} b_{\delta_{s'}, |Q_{s'}|} \text{ and } \widehat{p}_{X_{s'}} \leq \frac{75}{94} \left(\frac{1}{64L} \widehat{LB}_{s'} \right)^{d/\alpha} \quad (\text{with } \bar{c} = 0.1)$$

(where $\widehat{p}_{X_{s'}} :=$ `Estprob`$(\rho(X_s, X_{s'}), \left(\frac{1}{64L}\widehat{LB}_{s'}\right)^{d/\alpha}, 50, \delta_s)$*, then*

$$P_X(B(X_s, \rho(X_s, X_{s'}))) \leq \left(\frac{1}{64L} \widehat{LB}_{s'} \right)^{d/\alpha}. \tag{35}$$

Proof.
By following the scheme of subroutine `Estprob`, it is a direct application of lemma 7 by taking for all $s \leq w$, $t = 50$, $\epsilon_o = \left(\frac{1}{64L}\widehat{LB}_s\right)^{d/\alpha}$, $\delta' = \delta_s$, $r = \rho(X_s, X_{s'})$, $A_{3,s} := A'$. And then, if we set $A_3 = \cap_{s\geq 1} A_{3,s}$, we have $P(A_3) \geq 1 - \delta/16$, and (43) follows immediately.

On the other hand, for all $s \leq w$, the number of draws in `Estprob`$(\rho(X_s, X_{s'}), \left(\frac{1}{64L}\widehat{LB}_{s'}\right)^{d/\alpha}, 50, \delta_s)$ is always lower than w. Indeed, by lemma 7, the number of draws is at most:

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

Then we have:

$$\begin{aligned}
N &\leq \frac{400 \log\left(\frac{12800 s^2}{\delta(\frac{1}{64L}\widehat{LB}_{s'})^{d/\alpha}}\right)}{(\frac{1}{64L}\widehat{LB}_{s'})^{\frac{d}{\alpha}}} \\
&\leq \frac{400 \log\left(\frac{12800 s^2}{\delta(\frac{1}{64L}\bar{c} b_{\delta_{s'}, |Q_{s'}|})^{d/\alpha}}\right)}{(\frac{1}{64L}\bar{c} b_{\delta_{s'}, |Q_{s'}|})^{d/\alpha}} \qquad (\text{as } \widehat{LB}_{s'} \geq \bar{c} b_{\delta_{s'}, |Q_{s'}|}) \\
&\leq \frac{400 \log\left(\frac{12800 w^2}{\delta(\frac{1}{64L}\bar{c}\phi_n)^{d/\alpha}}\right)}{(\frac{1}{64L}\bar{c}\phi_n)^{d/\alpha}} \qquad (\text{we can easily see that } b_{\delta_{s'}, |Q_{s'}|} \geq \phi_n) \\
&\leq w \qquad (\text{by (12)}).
\end{aligned}$$

□

5.2.4 Label the instance space

Theorem 6.
Let ϵ, $\delta \in (0,1)$. Let $T_{\epsilon,\delta} = \frac{1}{\tilde{p}_\epsilon} \ln(\frac{8}{\delta})$, and $\tilde{p}_\epsilon = \left(\frac{\Delta}{128L}\right)^{d/\alpha}$, with $\Delta = \max(\frac{\epsilon}{2}, \left(\frac{\epsilon}{2C}\right)^{\frac{1}{\beta+1}})$. Let $I \subset \mathcal{X} \times \mathbb{R} \times \mathbb{N}$ the set used in `KALLS`*.*
Set $s_I = \max I'$ with $I' = \{s, (X_s, \widehat{LB}_s, |Q_s|) \in I\}$ (index of the last informative point). 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.

$$\sup_{x \in supp(P_X)} \min_{\bar{X} \in \{X_1, \ldots, X_{T_{\epsilon,\delta}}\}} P_X(B(x, \rho(\bar{X}, x))) \leq \tilde{p}_\epsilon. \tag{36}$$

2. If w satisfies (11) *and* (12) *and the following condition holds*

$$s_I \geq T_{\epsilon,\delta}, \tag{37}$$

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

$$|\eta(X_s) - \frac{1}{2}| \geq \frac{1}{2}\Delta \tag{38}$$

and

$$f^*(x) = f^*(X_s) \tag{39}$$

Proof.

1. Let us first prove the item 1.
   As in Section 5.1, 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 3, 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 (36) holds. And then, we can easily conclude the first item.

2. Let $x \in \text{supp}(P_X)$, by (36), on A_4, there exists $X_x \in \{X_1, \ldots, X_{T_{\epsilon,\delta}}\}$ such that:

$$P_X(B(x, \rho(X_x, x))) \leq \tilde{p}_\epsilon. \tag{40}$$

   By (H4), we have:

$$|\eta(x) - \eta(X_x)| \leq \frac{1}{128}\Delta < \frac{1}{32}\Delta \tag{41}$$

   Then if $|\eta(x) - \frac{1}{2}| > \Delta$, we have:

$$(1 - \frac{1}{32})\Delta < |\eta(X_x) - \frac{1}{2}| < (1 + \frac{1}{32})\Delta \tag{42}$$

   As $s_I \geq T_{\epsilon,\delta}$, then there exists s' such that $X_x := X_{s'}$ and $X_{s'}$ passes through the subroutine `Reliable`; we have two cases:

a) Firstly, $X_{s'}$ is uninformative, and then there exists $s < s'$, such that

$$\widehat{LB}_s \geq \bar{c} b_{\delta_s,|Q_s|} \text{ and } \widehat{p}_{X_s} \leq \frac{75}{94}\left(\frac{1}{64L}\widehat{LB}_s\right)^{d/\alpha}$$

(where $\widehat{p}_{X_s} := \texttt{Estprob}(\rho(X_s, X_{s'}), \left(\frac{1}{64L}\widehat{LB}_s\right)^{d/\alpha}, 50, \delta_s)$, then

$$P_X(B(X_{s'}, \rho(X_s, X_{s'}))) \leq \left(\frac{1}{64L}\widehat{LB}_s\right)^{d/\alpha}. \tag{43}$$

Necessary, we have $|\eta(X_s) - \frac{1}{2}| \geq \frac{1}{32}\Delta$. Indeed, if $|\eta(X_s) - \frac{1}{2}| < \frac{1}{32}\Delta$, then on $A_1 \cap A_2$, by denoting $\bar{k}_s$ the number of request labels in $\texttt{ConfidentLabel}(X_s) := \texttt{ConfidentLabel}(X_s, k(\epsilon, \delta_s), t, \delta_s)$, (where $t = n - \sum_{s_i \in I', s_i < s} |Q_{s_i}|$) we have:

$$\begin{aligned}
\widehat{LB}_s &= |\widehat{\eta}_{\bar{k}_s}(X_s) - \frac{1}{2}| - b_{\delta_s,\bar{k}_s} \\
&\leq |\widehat{\eta}_{\bar{k}_s}(X_s) - \bar{\eta}_{\bar{k}_s}(X_s)| + |\bar{\eta}_{\bar{k}_s}(X_s) - \frac{1}{2}| - b_{\delta_s,\bar{k}_s} \\
&\leq |\bar{\eta}_{\bar{k}_s}(X_s) - \frac{1}{2}| \quad \text{(by (18))} \\
&\leq |\eta(X_s) - \frac{1}{2}| + \frac{1}{32}(1 - \frac{1}{32})\Delta \quad \text{(by (H4) and Theorem 4)} \qquad (44) \\
&< \frac{1}{32}\Delta + \frac{1}{32}(1 - \frac{1}{32})\Delta \\
&= \frac{63}{1024}\Delta \qquad (45)
\end{aligned}$$

By (H4) and (43), we have:

$$\begin{aligned}
|\eta(X_{s'}) - \frac{1}{2}| &\leq |\eta(X_s) - \frac{1}{2}| + \frac{1}{64}\widehat{LB}_s \\
&< \frac{1}{32}\Delta + \frac{1}{64}.\frac{63}{1024}\Delta \quad \text{(by (45))} \\
&= (\frac{1}{32} + \frac{1}{64}.\frac{63}{1024})\Delta \\
&\leq (1 - \frac{1}{32})\Delta
\end{aligned}$$

that contradicts (42), then we have $|\eta(X_s) - \frac{1}{2}| \geq \frac{1}{32}\Delta$. Therefore, by (43), (44), we have:

$$\begin{aligned}
P_X(B(X_{s'}, \rho(X_s, X_{s'}))) &\leq \left(\frac{1}{64L}\widehat{LB}_s\right)^{d/\alpha} \\
&\leq \left(\frac{1}{64L}\left(|\eta(X_s) - \frac{1}{2}| + \frac{1}{32}(1 - \frac{1}{32})\Delta\right)\right)^{d/\alpha} \\
&\leq \left(\frac{1}{64L}\left(|\eta(X_s) - \frac{1}{2}| + (1 - \frac{1}{32})|\eta(X_s) - \frac{1}{2}|\right)\right)^{d/\alpha} \\
&= \left(\frac{1}{64L}(2 - \frac{1}{32})|\eta(X_s) - \frac{1}{2}|\right)^{d/\alpha} \\
&= \left(\frac{63}{2048L}|\eta(X_s) - \frac{1}{2}|\right)^{d/\alpha} \qquad (46)
\end{aligned}$$

On the other hand, by (40), we have:

$$\begin{aligned} P_X(B(x,\rho(X_{s'},x))) &\leq \tilde{p}_\epsilon \\ &= \left(\frac{1}{128L}\Delta\right)^{d/\alpha} \\ &\leq \left(\frac{1}{128L}|\eta(x)-\frac{1}{2}|\right)^{d/\alpha} \end{aligned} \tag{47}$$

We have:

$$\begin{aligned} |\eta(x)-\eta(X_s)| &\leq |\eta(x)-\eta(X_{s'})| + |\eta(X_{s'})-\eta(X_s)| \\ &\leq L.P_X(B(x,\rho(X_{s'},x)))^{\alpha/d} + L.P_X(B(X_{s'},\rho(X_{s'},X_s)))^{\alpha/d} \quad \text{(by (H4))} \\ &\leq \frac{1}{128}|\eta(x)-\frac{1}{2}| + \frac{63}{2048}|\eta(X_s)-\frac{1}{2}| \quad \text{(by (46) and (47))} \qquad (48) \\ &\leq \frac{1}{128}|\eta(x)-\frac{1}{2}| + \frac{\frac{63}{2048}}{1-\frac{63}{2048}}|\eta(X_{s'})-\frac{1}{2}| \quad \text{(by (H4) and (46))} \\ &\leq \frac{1}{128}|\eta(x)-\frac{1}{2}| + \frac{63}{1985}(1+\frac{1}{128})|\eta(x)-\frac{1}{2}| \quad \text{(by (41))} \\ &= \frac{79}{1985}|\eta(x)-\frac{1}{2}| \qquad (49) \end{aligned}$$

b) Secondly, $X_{s'}$ is informative, in this case, $s = s'$ (in the first case) and then we always obtains the equation (49).

(49) becomes:

$$\begin{aligned} |\eta(X_s)-\frac{1}{2}| &\geq \left(1-\frac{79}{1985}\right)|\eta(x)-\frac{1}{2}| \qquad (50) \\ &\geq \left(1-\frac{79}{1985}\right)\Delta \\ &\geq \frac{1}{2}\Delta \qquad (51) \end{aligned}$$

Then

$$|\eta(X_s)-\frac{1}{2}| \geq \frac{1}{2}\Delta \tag{52}$$

On $A_1 \cap A_2$, by theorem 5, the subroutine `ConfidentLabel`(X_s) uses at most $\tilde{k}(\epsilon,\delta_s)$ request labels, and returns the right label (with respect to the Bayes classifier) of X_s.
Let us proof that $f^*(x) = f^*(X_s)$. Let us assume without loss of generality that $\eta(X_s) - \frac{1}{2} \geq 0$. We will show that $\eta(x) - \frac{1}{2} \geq 0$. We have:

$$\begin{aligned} \eta(x)-\frac{1}{2} &= \eta(x)-\eta(X_s)+\eta(X_s)-\frac{1}{2} \\ &\geq \eta(X_s)-\frac{1}{2}-\frac{79}{1985}|\eta(x)-\frac{1}{2}| \quad \text{(by (49))} \\ &\geq (1-\frac{79}{1985})|\eta(x)-\frac{1}{2}| - \frac{79}{1985}|\eta(x)-\frac{1}{2}| \quad \text{(by (49))} \\ &= \frac{1827}{1985}|\eta(x)-\frac{1}{2}| \\ &\geq 0 \end{aligned}$$

Then $f^*(x) = f^*(X_s)$.

□

Lemma 9.
Let $x \in supp(P_X)$ such that $|\eta(x) - \frac{1}{2}| > \widehat{\Delta}$. Let $I \subset \mathcal{X} \times \mathbb{R} \times \mathbb{N}$ the set used in `KALLS`.
Set $s_I = \max I'$, with $I' = \{s, (X_s, \widehat{LB}_s, |Q_s|) \in I\}$ (index of the last informative point). Let us assume that $s_I \geq T_{\epsilon,\delta}$. Let $\widehat{\mathcal{S}}$ the final active set use in subroutine `Learn` (1). *Let $\widehat{f}_{n,w}$ the output of the subroutine* `Learn`. *Let us assume that w satisfies* (11) *and* (12). *We have on $A_1 \cap A_2 \cap A_3 \cap A_4$*

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

5.2.5 Label complexity

Lemma 10.
Let us assume that w satisfies (11), (12), *and $w \geq T_{\epsilon,\delta}$. Then, there exists an even A_5 such that $P(A_5) \geq 1 - \delta/8$, and on $A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5$, the condition* (10) *is sufficient to guarantee* (37)

Proof.
Let $I \subset \mathcal{X} \times \mathbb{R} \times \mathbb{N}$ the set used in `KALLS`.
Set $s_I = \max I'$, with $I' = \{s, (X_s, \widehat{LB}_s, |Q_s|) \in I\}$ (index of the last informative point). We consider two cases:

1. **First case:** $s_I = w$: we can easily see that (37) is satisfied. And we have trivially that the condition (10) is sufficient to guarantee (37).

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

$$\sum_{s \in I'} |Q_s| \tag{53}$$

where Q_s is the output in the subroutine `ConfidentLabel2`. Let $s \in I'$. For brevity, let us denote `ConfidentLabel`(X_s, t):=`ConfidentLabel`$(X_s, k(\epsilon, \delta_s), t, \delta_s)$, (where $t = n - \sum_{s_i \in I', s_i < s} |Q_{s_i}|$ the budget parameter). If $s \neq s_I$, the subroutine `ConfidentLabel`(t, X_s) implicitly assumes that the process of label request do not takes into account the constraint related to the budget n (very large budget with respect to $k(\epsilon, \delta_s)$) such that `ConfidentLabel`(X_s, t)=`ConfidentLabel`$(X_s, t = \infty)$. Then we have:

$$n > \sum_{\substack{s \in I' \\ s < s_I}} |Q_s| \tag{54}$$

On the other hand, we want to guarantee the condition (37), for this, necessary 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, t), the budget n is not yet reached and then we can replace the relation (54) by

$$n > \sum_{\substack{s \in I' \\ s < s_I \\ s \leq T_{\epsilon,\delta}}} |Q_s| \tag{55}$$

Then, necessary, (37) holds when (55) holds.
Also, for $s \in I'$, by theorem5, if we assume that $|\eta(X_s) - \frac{1}{2}| \geq \frac{1}{2}\Delta$, we have that $|Q_s| \leq \tilde{k}(\epsilon, \delta_s)$, and the subroutine `ConfidentLabel`$(X_s, t)$, (with $t = n - \sum_{s_i \in I', s_i < s} |Q_{s_i}|$) terminates when the cut-off condition (19) is satisfied. The left hand side of (55) is equal to:

$$\sum_{\substack{s \in I' \\ s < s_I \\ s \leq T_{\epsilon,\delta} \\ |\eta(X_s) - \frac{1}{2}| \geq \frac{1}{2}\Delta}} |Q_s| + \sum_{\substack{s \in I' \\ s < s_I \\ s \leq T_{\epsilon,\delta} \\ |\eta(X_s) - \frac{1}{2}| \leq \frac{1}{2}\Delta}} |Q_s| \tag{56}$$

Firstly, let us consider the first term in (56) and denote it by T_1. Let us denote by B_s the event:

$$B_s = \{|\eta(X_s) - \frac{1}{2}| \geq \frac{1}{2}\Delta\}.$$

We have

$$\mathbb{1}_{B_s} = \sum_{j=1}^{m_\epsilon} \mathbb{1}_{B_{s,j}} \tag{57}$$

where

$$B_{s,j} = \{2^{j-1}\frac{1}{2}\Delta \leq |\eta(X_s) - \frac{1}{2}| \leq 2^j\frac{1}{2}\Delta\} \quad \text{and } m_\epsilon = \max\left(0, \left\lceil \log_2\left(\frac{1}{\frac{1}{2}\Delta}\right)\right\rceil\right).$$

Then,

$$\begin{aligned} T_1 &\leq \sum_{\substack{s\in I' \\ s<s_I \\ s\leq T_{\epsilon,\delta} \\ |\eta(X_s)-\frac{1}{2}|\geq\frac{1}{2}\Delta}} \tilde{k}(\epsilon,\delta_s) \quad \text{by theorem 5} \\ &= \sum_{\substack{s\in I' \\ s<s_I \\ s\leq T_{\epsilon,\delta}}} \sum_{j=1}^{m_\epsilon} \tilde{k}(\epsilon,\delta_s)\mathbb{1}_{B_{s,j}} \end{aligned} \tag{58}$$

On $B_{s,j}$,

$$\begin{aligned} \tilde{k}(\epsilon,\delta_s) &\leq \frac{c}{2^{2j}\Delta^2}\left[\log(\frac{32s^2}{\delta}) + \log\log(\frac{32s^2}{\delta}) + \log\log\left(\frac{512\sqrt{e}}{2^j\Delta}\right)\right] \\ &\leq \frac{c}{2^{2j}\Delta^2}\left[2\log(\frac{32s^2}{\delta}) + \log\log\left(\frac{512\sqrt{e}}{\Delta}\right)\right] \end{aligned} \tag{59}$$

Then (58) becomes:

$$\begin{aligned} T_1 &\leq \frac{c}{\Delta^2}\left[2\log(\frac{32T_{\epsilon,\delta}^2}{\delta}) + \log\log\left(\frac{512\sqrt{e}}{\Delta}\right)\right]\sum_{j=1}^{m_\epsilon} 2^{-2j} \sum_{\substack{s\in I' \\ s\leq s_I \\ s\leq T_{\epsilon,\delta}}} \mathbb{1}_{B_{s,j}} \\ &\leq \frac{c}{\Delta^2}\left[2\log(\frac{32T_{\epsilon,\delta}^2}{\delta}) + \log\log\left(\frac{512\sqrt{e}}{\Delta}\right)\right]\sum_{j=1}^{m_\epsilon} 2^{-2j} \sum_{s\leq T_{\epsilon,\delta}} \mathbb{1}_{B_{s,j}} \end{aligned} \tag{60}$$

By Hoeffding's inequality, there exists an even A_5 such that $P(A_5) \geq 1 - \delta/8$, and on A_5, we have for all $j \leq m_\epsilon$,

$$\sum_{s\leq T_{\epsilon,\delta}} \mathbb{1}_{B_{s,j}} \leq T_{\epsilon,\delta} P_X(x,\ |\eta(x) - \frac{1}{2}| \leq 2^j\frac{1}{2}\Delta) + T_{\epsilon,\delta}\sqrt{\frac{1}{2T_{\epsilon,\delta}}\log\left(\frac{8}{\delta}\right)}$$

$$\begin{aligned} \frac{c}{\Delta^2}\sum_{s\leq T_{\epsilon,\delta}} \mathbb{1}_{B_{s,j}} &\leq \frac{c}{\Delta^2}\left(T_{\epsilon,\delta}P_X(x,\ |\eta(x) - \frac{1}{2}| \leq 2^j\frac{1}{2}\Delta) + T_{\epsilon,\delta}\sqrt{\frac{1}{2T_{\epsilon,\delta}}\log\left(\frac{8}{\delta}\right)}\right) \\ &\leq \frac{c}{\Delta^2}\left(T_{\epsilon,\delta}P_X(x,\ |\eta(x) - \frac{1}{2}| \leq 2^j\frac{1}{2}\Delta) + \sqrt{\frac{T_{\epsilon,\delta}}{2}\log\left(\frac{8}{\delta}\right)}\right) \\ &\leq \frac{c}{\Delta^2}\left(T_{\epsilon,\delta}2^{\beta j}\frac{1}{2^\beta}C\Delta^\beta + \frac{1}{\sqrt{\tilde{p}_\epsilon}}\log\left(\frac{8}{\delta}\right)\right) \quad \text{by (H3)} \\ &= 2^{\beta(j-1)}O\left(\left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d-\alpha\beta}{\alpha(\beta+1)}}\right) + O\left(\left(\frac{1}{\epsilon}\right)^{\frac{4\alpha+d}{2\alpha(\beta+1)}}\right) \end{aligned}$$

Then, (60) becomes:

$$T_1 \leq \tilde{O}\left(\left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d-\alpha\beta}{\alpha(\beta+1)}}\right)\sum_{j=1}^{m_\epsilon} 2^{(\beta-2)j} + \tilde{O}\left(\left(\frac{1}{\epsilon}\right)^{\frac{4\alpha+d}{2\alpha(\beta+1)}}\right)\sum_{j=1}^{m_\epsilon} 2^{-2j}$$
$$\leq \max\left(\tilde{O}\left(\left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d-\alpha\beta}{\alpha(\beta+1)}}\right), \tilde{O}\left(\left(\frac{1}{\epsilon}\right)^{\frac{4\alpha+d}{2\alpha(\beta+1)}}\right)\right) \tag{61}$$

where $\tilde{O}$ includes the logarithmic terms.
Secondly, by using the same argument as with the term T_1, the second term T_2 in (56) also satisfies the same relation (61). Then the term in (56) is least than:

$$\max\left(\tilde{O}\left(\left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d-\alpha\beta}{\alpha(\beta+1)}}\right), \tilde{O}\left(\left(\frac{1}{\epsilon}\right)^{\frac{4\alpha+d}{2\alpha(\beta+1)}}\right)\right) \tag{62}$$

Then if

$$n \geq \max\left(\tilde{O}\left(\left(\frac{1}{\epsilon}\right)^{\frac{2\alpha+d-\alpha\beta}{\alpha(\beta+1)}}\right), \tilde{O}\left(\left(\frac{1}{\epsilon}\right)^{\frac{4\alpha+d}{2\alpha(\beta+1)}}\right)\right)$$

we have that n satisfies (55) , and (37) is necessary satisfied.

□

Proof of Theorem 2:
Let us assume that (10) holds. Then, by Lemma 10, on $A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5$, (37) also holds, this implies that by Lemma 9, the final classifier $\widehat{f}_{n,w}$ agrees with the Bayes classifier f^* on $\{x,\ |\eta(x) - 1/2| > \Delta\}$. Thus, (13) holds with probability at least $1 - \left(\frac{\delta}{16} + \frac{\delta}{8} + \frac{\delta}{16} + \frac{\delta}{8} + \frac{\delta}{8}\right) = 1 - \delta/2 > 1 - \delta$.

6 Conclusion

In this paper we have reviewed the main results for convergence rates in a nonparametric setting, with a special emphasis on the relative merits of the assumptions about the smoothness and the margin noise. By putting active learning in perspective with recent work on passive learning that used a particular smoothness assumption customized for k-nn, we provided a novel active learning algorithm with a rate of convergence comparable to stat-of-the art active learning algorithms, but with less restrictive assumptions. Interesting future directions include an extension to multi-class instead of binary classification. For example, [22] provides a step in this direction, since it extends the work of [5] to the context of multiclass. Adaptive algorithms, i.e. where the parameters α, β describing the smoothness and margin noise are unknown should also be explored in our setting. Previous work in this direction was done in [15]. Practical implementations of the `KALLS` algorithm are underway.

References

[1] Audibert, J.Y., Tsybakov, A.B., et al.: Fast learning rates for plug-in classifiers. The Annals of statistics **35**(2), 608–633 (2007)

[2] Balcan, M.F., Hanneke, S., Vaughan, J.W.: The true sample complexity of active learning. Machine learning **80**(2-3), 111–139 (2010)

[3] Biau, G., Devroye, L.: Lectures on the nearest neighbor method. Springer (2015)

[4] Castro, R.M., Nowak, R.D.: Minimax bounds for active learning. IEEE Transactions on Information Theory **54**(5), 2339–2353 (2008)

[5] Chaudhuri, K., Dasgupta, S.: Rates of convergence for nearest neighbor classification. In: Advances in Neural Information Processing Systems, pp. 3437–3445 (2014)

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

[7] Dasgupta, S.: Active learning theory. Encyclopedia of Machine Learning and Data Mining pp. 14–19 (2017)

[8] Döring, M., Györfi, L., Walk, H.: Rate of convergence of k-nearest-neighbor classification rule. The Journal of Machine Learning Research **18**(1), 8485–8500 (2017)

[9] Hanneke, S.: Nonparametric active learning, part 1: Smooth regression functions (2018). Http://www.stevehanneke.com/

[10] Hanneke, S., Yang, L.: Minimax analysis of active learning. The Journal of Machine Learning Research **16**(1), 3487–3602 (2015)

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

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

[13] Kaufmann, E., Cappé, O., Garivier, A.: On the complexity of best-arm identification in multi-armed bandit models. The Journal of Machine Learning Research **17**(1), 1–42 (2016)

[14] Kontorovich, A., Sabato, S., Urner, R.: Active nearest-neighbor learning in metric spaces. In: Advances in Neural Information Processing Systems, pp. 856–864 (2016)

[15] Locatelli, A., Carpentier, A., Kpotufe, S.: Adaptivity to noise parameters in nonparametric active learning. Proceedings of Machine Learning Research vol **65**, 1–34 (2017)

[16] Locatelli, A., Carpentier, A., Kpotufe, S.: An adaptive strategy for active learning with smooth decision boundary. In: Algorithmic Learning Theory, pp. 547–571 (2018)

[17] Lugosi, G.: Pattern classification and learning theory. In: Principles of nonparametric learning, pp. 1–56. Springer (2002)

[18] Mammen, E., Tsybakov, A.B., et al.: Smooth discrimination analysis. The Annals of Statistics **27**(6), 1808–1829 (1999)

[19] Massart, P., Nédélec, É., et al.: Risk bounds for statistical learning. The Annals of Statistics **34**(5), 2326–2366 (2006)

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

[21] Mulzer, W.: Five proofs of chernoff's bound with applications. arXiv preprint arXiv:1801.03365 (2018)

[22] Reeve, H.W., Brown, G.: Minimax rates for cost-sensitive learning on manifolds with approximate nearest neighbours. In: International Conference on Algorithmic Learning Theory, pp. 11–56 (2017)

[23] Vidyasagar, M.: Learning and generalisation: with applications to neural networks. Springer Science & Business Media (2013)