CLASSIFICATION WITH THE NEAREST NEIGHBOR RULE IN GENERAL FINITE DIMENSIONAL SPACES

By Sébastien Gadat and Thierry Klein and Clément Marteau

Toulouse School of Economics, Université Toulouse I Capitole
Institut Mathématiques de Toulouse, Université Paul Sabatier

Given an n-sample of random vectors $(X_i, Y_i)_{1\leq i\leq n}$ whose joint law is unknown, the long-standing problem of supervised classification aims to *optimally* predict the label Y of a given a new observation X. In this context, the nearest neighbor rule is a popular flexible and intuitive method in non-parametric situations. Even if this algorithm is commonly used in the machine learning and statistics communities, less is known about its prediction ability in general finite dimensional spaces, especially when the support of the density of the observations is $\mathbb{R}^d$. This paper is devoted to the study of the statistical properties of the nearest neighbor rule in various situations. In particular, attention is paid to the marginal law of X, as well as the smoothness and margin properties of the *regression function* $\eta(X) = \mathbb{E}[Y|X]$. We identify two necessary and sufficient conditions to obtain uniform consistency rates of classification and to derive sharp estimates in the case of the nearest neighbor rule. Some numerical experiments are proposed at the end of the paper to help illustrate the discussion.

1. Introduction. The supervised classification model has been at the core of numerous contributions to statistical literature in recent years. It continues to provide interesting problems, both from the theoretical and practical point of views. The classical task in supervised classification is to predict a feature $Y \in \mathcal{M}$ when a variable of interest $X \in \mathbb{R}^d$ is observed, the set $\mathcal{M}$ being finite. In this paper, we focus on the binary classification problem where $\mathcal{M} = \{0, 1\}$.

In order to provide a prediction of the label Y of X, it is assumed that a training set $\mathcal{S}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ is at our disposal, where (X_i, Y_i) are i.i.d. and with a common law $\mathbb{P}_{X,Y}$. This training set $\mathcal{S}_n$ makes it possible to retrieve some information on the joint law of (X, Y) and to provide, depending on some technical conditions, a pertinent prediction. In particular,

AMS 2000 subject classifications: Primary 62G05; secondary 62G20
Keywords and phrases: Supervised classification, nearest neighbor algorithm, plug in rules, minimax classification rates

the regression function η defined as:

$$\eta(x) = \mathbb{E}[Y|X = x], \forall x \in \mathbb{R}^d$$

appears to be of primary interest for the statistician (see Section 2 for a formal description of the model). Indeed, given $x \in \mathbb{R}^d$, the term $\eta(x)$ provides the probability that Y is assigned the label 1, conditionally to the event $\{X = x\}$. Since this function is unknown in practice, prediction rules are based on the training sample $\mathcal{S}_n$.

Several algorithms have been proposed over the years but we do not intend to provide an exhaustive list of the associated papers. For an extended introduction to the supervised classification theory, we refer to [BBL05] or [DGL96]. Among available classification procedures, we can, roughly speaking, divide them into (at least) three families:

- *Approaches based on pure entropy considerations and Empirical Risk Minimization (ERM):* Given a classifier, the miss-classification error can be empirically estimated from the learning sample. The ERM algorithm then selects the classifier that minimizes this empirical risk among a given family of candidates. Several studies such as in [MT99], [BM06], [AT07], [LM14] now provide an almost complete description of their statistical performance. In an almost similar context, some aggregation schemes, first proposed in Boosting procedures by [FS97], have been analyzed in depth in [Lec07] and shown to be adaptive to margin and complexity.
- *Methods derived from geometric interpretation or information theory:* For example, the Support Vector Machine classifier (SVM) aims to maximize the margin of the classification rule. It has been intensively studied in the last two decades because of its low computational cost and excellent statistical performances (see [Vap98], [Ste05] or [BBM08] among others). Classification and Regression Tree is another intuitive standard method that relies on a recursive dyadic partition of the state space, introduced in [BFOS84] and greatly improved by an averaging procedure in [AG97] and [Bre01], which is usually referred to as Random Forest, and later theoretically developed in [BDL08].
- *Plug-in rules:* The main idea is to mimic the Bayes optimal classifier using a plug-in rule after a preliminary estimation of the function η. We refer to [GKKW02] for a general overview, and to [BR03] and [AT07] for some recent statistical results within this framework. The main motivation behind plug-in rules is to transfer properties related

to the classical regression problem (estimation of η from the sample $\mathcal{S}_n$) to a quantitative control of the miss-classification error.

In this general overview, the nearest neighbor rule (see Section 2.2 for a complete description) belongs to the last two classes. It corresponds to a plug-in classifier with a simple geometrical interpretation. It has attracted a great deal of attention for the past few decades, from the seminal works of [FH51] and [CH67]. Given an integer k, the corresponding classifier is based on a feature average of the k-closest observations of X in the training set $\mathcal{S}_n$. We also refer to [Sto77], [Győ78], [Györ81], [DW77] and [DGKL94] for seminal contributions on this prediction rule (both for classification and regression). Recently, this algorithm has received even further attention in mathematical statistics, and is still at the core of several studies: [CG06] examines the situation of general metric space and identifies the importance of the so-called Besicovitch assumption, [HPS08] is concerned with the influence of the integer k on the excess risk of the nearest neighbor rule as well as the two notions of the sample structure while [Sam12] describes an improvement of the standard algorithm.

Most of the results obtained for penalized ERM, SVM or plug-in classifiers are based on complexity considerations (metric entropy or Vapnik dimension). In this paper, we mainly use the asymptotic behavior of the small ball probabilities instead (see [Lia11] and the references therein), which can be seen as a dual quantity of the entropy (see [LS01]). We also deal with the more intricate situation of not bounded away from zero densities (especially for non compactly supported measures). For this purpose, we work with both *smoothness* and *minimal mass* assumptions (see Section 2.3 for more details) that will provide a pertinent estimation of the function η. In particular, it is assumed that we will be able to take advantage of some smoothness properties of the function η in order to improve the prediction of the label Y. According to previous existing studies (see, *e.g.*, [Györ81]), the associated classification rates appear to be comparable to those obtained in an "estimation" framework and, hence, always greater than $\sqrt{n}^{-1}$. However, it has been proven in [MT99] that *fast rates* (*i.e.*, faster than $\sqrt{n}^{-1}$) can be obtained up to some additional *margin* assumption. It is in fact possible to take advantage of the behavior of the law of (X, Y) around the boundary $\{\eta = 1/2\}$ in order to improve the properties of the classification process.

In this paper, we investigate the nearest neighbor rule with margin assumption and marginal distribution μ for the variable X that is not necessarily compactly supported or lower bounded from zero. The contributions pro-

posed below can be broken down into three different categories.

Consistency rate for bounded from below densities. Our first result concerns the optimality of the nearest neighbor classifier Φ_n in the compact case. We prove that this classification rule reaches the minimax rate of convergence for the excess risk obtained by [AT07] (see Theorem 3.2 below). In particular, under some classical assumptions about the distribution F of the couple (X, Y) (which will be illustrated below), we show that:

$$\sup_{F\in\mathcal{F}} [\mathcal{R}(\Phi_n) - \mathcal{R}(\Phi^*)] \leq Cn^{-\frac{1+\alpha}{2+d}},$$

where α denotes the margin parameter, d the dimension of the problem, $\mathcal{R}(\Phi)$ the miss-classification error of a given classifier Φ and Φ^* the Bayes classifier.[1] We obtain this result for both Poisson and Binomial sample-size models. In particular, such a result appears to be a generalization of the ones given in [HPS08] that do not take the margin α into account in their study.

Consistency rate for general densities. In a second step, we investigate the behavior of the nearest neighbor classifier when the marginal density μ (w.r.t. the Lebesgue measure) of X is not bounded from below on its support. Such an improvement is not of secondary importance since it corresponds to the commonly encountered situation of vanishing or non-compactly supported densities. To do this, we use an additional assumption on the *tail* of this distribution and prove that generically:

$$\sup_{F\in\mathcal{F}} [\mathcal{R}(\Phi_n) - \mathcal{R}(\Phi^*)] \leq Cn^{-\frac{1+\alpha}{2+\alpha+d}},$$

as soon as the bandwidth k involved in the classifier is allowed to depend on the spatial position of X. The *tail* assumption on the marginal distribution on X involved in this result, will describe the behavior of the density μ near the set $\{\mu = 0\}$.

Lower bounds. Finally, we derive some lower bounds for the supervised classification problem, which extends the results obtained in [AT07] in a slightly different context. We prove that our Tail Assumption is unavoidable to ensure uniform consistency rates for classification in a non-compact case, regardless of dimension d. We then see how these upper and lower bounds are linked. In particular, we show that a very unfavorable situation of classification occurs when the regression function η oscillates in the tail of the

[1] This result has also been established in the recent work of [Sam12]

distribution μ, *i.e.*, we establish that it is even impossible in these situations to obtain uniform consistency rates and thus elucidate two open questions in [Can13].

The paper is organized as follows. In Section 2, we precisely describe the statistical setting related to the classification problem. Some attention is paid to the nearest neighbor rule. Section 3 is devoted to the bounded from below case where we prove that the nearest neighbor classifier reaches the minimax rate of convergence for the excess risk under mild assumptions. We then extend our study to the general (typically non-compact) case in Section 4. This section is supplemented with some supporting numerical results and a glossary of typical situations of location models. We conclude with a discussion of our results, and potential problems. Proofs and technical results are included in Appendix A. The paper is completed by an adaptation to the smooth discriminant analysis model in Appendix B (see, *e.g.* [MT99] or [HPS08] for another comparison between the so-called Poisson and Binomial models). In particular, although the variables of interest are strongly dependent in this case, we derive (using a Poissonization argument) results similar to those obtained in the classical binary classification model.

We use the following notations throughout the paper. $\mathbb{P}_{X,Y}$ denotes the distribution of the couple (X, Y) and $\mathbb{P}_X$ the marginal distribution of X, which will be assumed to admit a density μ with respect to the Lebesgue measure. Similarly, we set $\mathbb{P}_{\otimes^n} = \prod_{i=1}^n \mathbb{P}_{(X_i,Y_i)}$ and $\mathbb{P} = \mathbb{P}_{(X,Y)} \times \mathbb{P}_{\otimes^n}$. In the same spirit, $\mathbb{E}[.]$, $\mathbb{E}_X[.]$ and $\mathbb{E}_{\otimes^n}[.]$ will hereafter correspond to the expectations w.r.t. the measures $\mathbb{P}$, $\mathbb{P}_X$ and $\mathbb{P}_{\otimes^n}$, respectively. Finally, given two real sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$, we write $a_n \lesssim b_n$ (resp. $a_n \sim b_n$) is a real constant $C \geq 1$ exists such that $a_n \leq Cb_n$ (resp. $\frac{b_n}{C} \leq a_n \leq Cb_n$) for all $n \in \mathbb{N}$.

2. Statistical setting and nearest neighbor classifier.

2.1. *Statistical Classification problem.* In this paper, we study the classical binary supervised classification model (see, *e.g.*, [DGL96] for a complete introduction). An i.i.d. sample $\mathcal{S}_n := (X_i, Y_i)_{i=1\dots n} \in \Omega \times \{0, 1\}$, whose distribution is $\mathbb{P}_{X,Y}$ and where $\Omega = \mathrm{Supp}(\mu)$ is an open set of $\mathbb{R}^d$, is at our disposal. Given a new incoming observation X, our goal is to predict its corresponding label Y. To do this, we use a classifier that provides a decision rule for this problem. Formally, a classifier is a measurable mapping Φ from $\mathbb{R}^d$ to $\{0, 1\}$. Given a classifier Φ, its corresponding miss-classification error

is then defined as:

$$\mathcal{R}(\Phi) = \mathbb{P}\left(\Phi(X) \neq Y\right).$$

In practice, the most interesting classifiers are those associated with the smallest possible error. In this context it is well known (see, *e.g.*, [BBL05]) that the Bayes classifier Φ^* defined as:

$$\Phi^*(X) = \mathbf{1}_{\left\{\eta(X)>\frac{1}{2}\right\}}, \text{ where } \eta(x) := \mathbb{E}\left[Y|X=x\right] \ \forall x \in \Omega, \tag{2.1}$$

minimizes the miss-classification error, *i.e.*,

$$\mathcal{R}(\Phi^*) \leq \mathcal{R}(\Phi), \ \forall \Phi : \mathbb{R}^d \longrightarrow \{0,1\}.$$

The classifier $\Phi^\star$ provides the best decision rule in the sense that it leads to the lowest possible miss-classification error. Unfortunately, Φ^* is not available since the regression function η explicitly depends on the underlying distribution of (X,Y). In some sense, the Bayes classifier can be considered as an *oracle* that provides a benchmark error. Hence, the main challenge in this supervised classification setting is to construct a classifier Φ whose miss-classification error will be as close as possible to the smallest possible one. In particular, the excess risk (also referred to as the *regret*) defined as

$$\mathcal{R}(\Phi) - \mathcal{R}(\Phi^*),$$

appears to be of primary importance. We are interested here in the statistical properties of the nearest neighbor classifier (see Section 2.2 below for more details) based on the sample $\mathcal{S}_n$. In particular, we investigate the asymptotic properties of the excess risk through the **minimax** paradigm. Given a set $\mathcal{F}$ of possible distributions F for (X,Y), the minimax risk is defined as:

$$\delta_n(\mathcal{F}) := \inf_{\Phi} \sup_{F\in\mathcal{F}} \left[\mathcal{R}(\Phi) - \mathcal{R}(\Phi^*)\right],$$

where the infimum in the above formula is taken over all $\mathcal{S}_n$ measurable classifiers. A classifier Φ_n is then said to be minimax over the set $\mathcal{F}$ if:

$$\sup_{F\in\mathcal{F}} \left[\mathcal{R}(\Phi_n) - \mathcal{R}(\Phi^*)\right] \leq C\delta_n(\mathcal{F}),$$

for some constant C. The considered set $\mathcal{F}$ will be detailed later on and will depend on the behavior of (μ,η) over $\mathbb{R}^d$ through some smoothness, margin and minimal mass hypotheses.

2.2. *The nearest neighbor rule .* In this paper, we focus on the nearest neighbor classifier, which is perhaps one of the most widespread and simplest classification procedures. Suppose that the state space is $(\mathbb{R}^d, \|.\|)$ where $\|.\|$ is a reference distance. Given any sample $\mathcal{S}_n$ and for any $x \in \mathbb{R}^d$, we first build the reordered sample $\big(X_{(j)}(x), Y_{(j)}(x)\big)_{1\leq j\leq n}$ with respect to the distances $\|X_i - x\|$, namely:

$$\|X_{(1)}(x) - x\| \leq \|X_{(2)}(x) - x\| \leq \ldots \leq \|X_{(n)}(x) - x\|.$$

In this context $X_{(m)}(x)$ is the m-nearest neighbor of x w.r.t. the distance $\|.\|$ and $Y_{(m)}(x)$ its corresponding label. Given any integer k in $\mathbb{N}$, the principle of the nearest neighbor algorithm is to construct a decision rule based on the k-nearest neighbor of the input X: the $\mathcal{S}_n$-measurable classifier $\Phi_{n,k}$ is:

$$\Phi_{n,k}(X) = \begin{cases} 1 & \text{if} \quad \dfrac{1}{k}\displaystyle\sum_{j=1}^{k} Y_{(j)}(X) > \dfrac{1}{2}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.2}$$

For all $x \in \Omega$, the term $\frac{1}{k}\sum_{j=1}^{k} Y_{(j)}(x)$ appears to be an estimator of the regression function $\eta(x)$. In particular, we can write the classifier $\Phi_{n,k}$ as

$$\Phi_{n,k}(X) = \mathbf{1}_{\{\hat{\eta}_n(X)>1/2\}} \qquad \text{where} \qquad \hat{\eta}_n(x) = \frac{1}{k}\sum_{j=1}^{k} Y_{(j)}(x) \quad \forall x \in \Omega. \tag{2.3}$$

Hence, the nearest neighbor procedure can be considered as a plug-in classifier, *i.e.*, a preliminary estimator of the regression function η is plugged in our decision rule. It is worth noting that the integer k is a *regularization parameter*. Indeed, if k is too small, the classifier $\Phi_{n,k}$ will only use a small amount of the neighbors of X, leading to a large variance during the classification process. On the other hand, large values of k will introduce some bias into the decision rule since we use observations that may be far away from the input X. In other words, the statistical performances of $\Phi_{n,k}$ will depend on a careful choice of the integer k. In particular, the number of neighbors $k = k_n$ considered should carefully grow to $+\infty$ with respect to n.

For this purpose, we introduce some baselines assumptions into the following section that will make it possible to characterize an optimal value for this regularization parameter.

2.3. *Baseline assumptions.* It is well known that no reliable prediction can be made in a distribution-free setting (see [DGL96]). We restrict the class of possible distributions of (X, Y) below.

Since the nearest neighbor rule is a plug-in classification rule, we expect to take advantage of some smoothness properties of η in order to improve the classification process. In fact, when η is smooth, the respective values of $\eta(x_1)$ and $\eta(x_2)$ are comparable for close enough x_1, x_2. In other words, we can infer the sign of $\eta(x) - \frac{1}{2}$ from those of the neighbors of x.

Assumption ***A1****.* (Smoothness) *The regression function η belongs to the Hölder class of parameter 1 with a radius L, which is denoted $\mathcal{C}^{1,0}(\Omega, L)$ and corresponds to the set of functions such that*

$$\forall (x_1, x_2) \in \Omega^2 \qquad |\eta(x_1) - \eta(x_2)| \leq L|x_1 - x_2|.$$

REMARK 2.1. *It would be tempting to consider some more general smoothness classes for the regression function η. Nevertheless, the standard nearest neighbor algorithm does not make it possible to use smoothness indexes greater than 1. An alternative procedure has been proposed in [Sam12]: the idea is then to balance the $(Y_{(j)})_{j=1..k}$ with a suitable monotonous weighting sequence. However, this modification complicates the statistical analysis and may alter the ideas developed below. We therefore chose to fix the smoothness of η to 1 (i.e. restrict our study to $\mathcal{C}^{1,0}(\Omega, L)$).*

Our second assumption was introduced by [Tsy04] in the binary supervised classification model (see [MT99] in a *smooth discriminant analysis* setting).

Assumption ***A2****.* (Margin assumption) *For any $\alpha > 0$, a constant $C > 0$ exists such that:*

$$\mathbb{P}_X \left(0 < \left| \eta(X) - \frac{1}{2} \right| < \epsilon \right) \leq C\epsilon^{\alpha}, \quad \forall \epsilon > 0.$$

In such a case, we write $(\mu, \eta) \in \mathbf{M}_{\alpha}$.

The Bayes classifier depends on the sign of $\eta(X) - 1/2$. Intuitively, it would be easier to mimic the behavior of this classifier when the mass around the set $\{\eta = 1/2\}$ is small. On the other hand, the decision process may be more complicated when $\eta(X)$ is close to $1/2$ with a large probability. Quantifying this closeness is the purpose of this margin assumption.

For the sake of convenience, we use the set $\mathcal{F}_{L,\alpha}$ throughout the paper, which contains distributions that satisfy both Assumptions **A1** and **A2**, namely:

$$\mathcal{F}_{L,\alpha} := \big\{\mathbb{P}_{(X,Y)} \,:\, \mathbb{P}_X(dx) = \mu(x)dx \text{ and } \mathcal{L}(Y|X) \sim \mathcal{B}(\eta(X)) \\ \text{with } \eta \in \mathcal{C}^{1,0}(\Omega, L) \text{ and } (\mu, \eta) \in \mathbf{M}_\alpha\big\}$$

We now turn to our last assumption that involves the marginal distribution of the variable X.

2.4. *Minimal Mass Assumption.* In the sequel, this type of hypothesis will play a very important role.

*Assumption **A3**.* (Strong Minimal Mass Assumption) *There exists $\kappa > 0$ such that the marginal density μ of X satisfies $\mu \in \mathfrak{M}_{mma}(\Omega, \kappa)$ where*

$$\mathfrak{M}_{mma}(\Omega,\kappa) := \Big\{\mathbb{P}_X \,:\, \mathbb{P}_X(dx) = \mu(x)dx \,| \\ \exists \delta_0 > 0,\, \forall \delta \le \delta_0,\, \forall x \in \Omega : \mathbb{P}_X(X \in B(x,\delta)) \ge \kappa\mu(x)\delta^d\Big\}.$$

This assumption guarantees that $\mathbb{P}_X$ possesses a minimal amount of mass on each ball $B(x,\delta)$, this lower bound being balanced by the level of the density on x. In some sense, distributions in $\mathfrak{M}_{mma}(\Omega,\kappa)$ will make it possible to obtain reliable predictions of the regression function η according to its Lipschitz property. The Strong Minimal Mass Assumption **A3** may be seen as a refinement of the so-called Besicovitch assumption that is quite popular in the statistical literature (see, *e.g.*, [Dev81] for a version of the Besicovitch assumption used for pointwise consistency or [CG06] for a general discussion on this hypothesis in finite or infinite dimension). It is worth pointing out that the Besicovitch assumption introduced in [CG06] states that η satisfies the following μ-continuity property:

$$(2.4)\quad \forall \epsilon > 0 \quad \lim_{\delta \to 0} \mathbb{P}_X \left\{ x : \frac{1}{\mu(B(x,\delta))} \int_{B(x,\delta)} |\eta(z) - \eta(x)| d\mu(z) > \epsilon \right\} = 0$$

In our setting, since η is L-Lipschitz (Assumption **A1**), we can check that for all $x \in \Omega$

$$\int_{B(x,\delta)} |\eta(z) - \eta(x)| \mu(z) dz \le L \int_{B(x,\delta)} |x - z| \mu(z) dz \le L\delta\mu(B(x,\delta)),$$

which implies that the right hand side of (2.4) vanishes as soon as $\delta \leq \epsilon/L$. We will see that Assumption **A3** is necessary to obtain quantitative estimates for any finite dimensional classification problem in a general setting. In a slightly different framework, our Assumption **A3** is similar to the *Strong Density Assumption* used in the paper of [AT07] when the density μ is lower bounded on its (compact) support, which is assumed to possess some geometrical properties ((c_0, r_0) regularity). This setting is at the core of the study presented in Section 3 below. Assumption **A3** also recalls the notion of standard sets used in [Cas07] for the estimation of compact support sets. More generally, the following examples present some standard distributions that satisfy Assumption **A3**.

Example 2.1.

- *In $\mathbb{R}^d$, it is not difficult to check that Gaussian measures with non-degenerated covariance matrices satisfy $\mathfrak{M}_{mma}(\Omega, \kappa)$. As a simple example, consider a standard Gaussian law $\mu \sim \mathcal{N}(0,1)$. For any $x \in \mathbb{R}$ and $\delta > 0$, if x belongs to a compact set K, then a constant C_K exists such that $(2\pi)^{-1/2} \int_{x-\delta}^{x+\delta} e^{-t^2/2} dt \geq C_K e^{-x^2/2}\delta$. Now, if $x \longrightarrow +\infty$, we can check that:*

$$(2\pi)^{-1/2} \int_{x-\delta}^{x+\delta} (2\pi)^{-1/2} e^{-t^2/2} dt \sim (2\pi)^{-1/2} e^{-x^2/2} \left[\frac{e^{x\delta}}{x-\delta} - \frac{e^{-x\delta}}{x+\delta} \right] e^{-\delta^2/2}.$$

  The bracket above is always greater than δ when $(x\delta)^{-1} = O(1)$. Now, if $\delta = o(1/x)$, a simple Taylor expansion yields

$$(2\pi)^{-1/2} \int_{x-\delta}^{x+\delta} (2\pi)^{-1/2} e^{-t^2/2} dt \sim \mu(x) \frac{1+2x\delta}{x} \gtrsim \mu(x)\delta.$$

- *The same computations are still possible for symmetric Laplace distributions ($e^t \int_{t-\delta}^{t+\delta} e^{-x} dx = [e^{\delta} - e^{-\delta}] \sim 2\delta$ when δ is small. Thus, any Laplace distribution belongs to $\mathfrak{M}_{mma}(\Omega, \kappa)$. In a same way, when μ is a standard Cauchy distribution, we can check that:*

$$\begin{aligned} \int_{x-\delta}^{x+\delta} \frac{dt}{1+t^2} &= \frac{1}{1+x^2} \int_{\delta}^{\delta} \frac{1}{1+h\frac{2x+h}{1+x^2}} dh \\ &\sim \frac{1}{1+x^2} \left[2\delta - \frac{2}{3} \frac{\delta^3}{1+x^2} + +8 \frac{\delta^3 x^2}{(1+x^2)^2} o(\delta^3) \right] \\ &\gtrsim \frac{\delta}{1+x^2} \end{aligned}$$

Typically, distributions that do not satisfy the Strong Minimal Assumption (A3) possess some important oscillations in their tails (when the density μ is close to 0). In such a setting, the alternative set $\widetilde{\mathfrak{M}}_{mma}(\Omega,\kappa)$, defined as follows, may be considered:

$$\widetilde{\mathfrak{M}}_{mma}(\Omega,\kappa) := \Big\{ \mathbb{P}_X \ : \ \mathbb{P}_X(dx) = \mu(x)dx | \, \exists(\rho, C) \in]0;+\infty[^2$$
$$\exists \delta_0 > 0, \, \forall \delta \leq \delta_0 : \forall x \in \Omega : \mu(x) \geq e^{-C\delta^{-\rho}} \implies \mathbb{P}_X(B(x,\delta)) \geq \kappa \mu(x) \delta^d \Big\}.$$

The interest of the weaker $\widetilde{\mathfrak{M}}_{mma}(\Omega,\kappa)$ compared to $\mathfrak{M}_{mma}(\Omega,\kappa)$ is that the statistical abilities of the nearest neighbor rule are still the same with $\mathfrak{M}_{mma}(\Omega,\kappa)$ or $\widetilde{\mathfrak{M}}_{mma}(\Omega,\kappa)$. Moreover, an analytic criterion that ensures $\widetilde{\mathfrak{M}}_{mma}(\Omega,\kappa)$ can be found (see Proposition 4.1. This is not the case for the *uniform* assumption $\mathfrak{M}_{mma}(\Omega,\kappa)$ (it is indeed more difficult to ensure the lower bound on the global set Ω).

Although all the subsequent results may be established for a weaker version of the minimal mass assumption (based on the set $\widetilde{\mathfrak{M}}_{mma}(\Omega,\kappa)$), we will restrict ourselves to its strong formulation (Assumption **A3**). In Section 3, we prove that the nearest neighbor rule is optimal in the minimax sense provided that the margin and smoothness assumptions hold, with a marginal density of the variable X bounded away from 0 and a suitable choice of k. In Section 4, we will see that $\mathfrak{M}_{mma}(\Omega,\kappa)$ is not yet sufficient to derive consistent classifiers for non compactly supported densities, and a last additional hypothesis is needed.

3. Bounded away from zero densities.

3.1. *Minimax consistency of the nearest neighbor rule.* In this section, we are interested in the special case of a marginal density μ bounded from below by a strictly positive constant μ_-. In this context, we can state an upper bound on the consistency rate of the nearest neighbor rule.

THEOREM 3.1. *Assume that Assumptions **A1-A3** hold. The nearest neighbor classifier Φ_{n,k_n} with $k_n = \lfloor n^{\frac{2}{2+d}} \rfloor$ satisfies*

$$\sup_{\mathbb{P}_{X,Y} \in \mathcal{F}_{L,\alpha} \cap \mathfrak{M}_{mma}(\Omega,\kappa)^{\mu_-}} [\mathcal{R}(\Phi_{n,k_n}) - \mathcal{R}(\Phi^*)] \lesssim n^{-\frac{1+\alpha}{2+d}},$$

where $\mathfrak{M}_{mma}(\Omega,\kappa)^{\mu_-}$ denotes the subset of densities of $\mathfrak{M}_{mma}(\Omega,\kappa)$ that are bounded from below by μ_-.

Theorem 3.1 establishes a consistency rate of the nearest neighbor rule over $\mathcal{F}_{L,\alpha} \cap \mathfrak{M}_{mma}(\Omega, \kappa)^{\mu_-}$. A detailed proof of is presented in Section A.2. Implicitly, we restrict our analysis to compactly supported observations, this assumption being at the core of several statistical analyses (see, *e.g.*, [GKKW02], [BBL05], [MT99] or [HPS08] among others). It is worth pointing out that this setting falls into the framework considered in [AT07].

DEFINITION 3.1 (Strong Density Assumption (SDA), [AT07]). *The marginal distribution of the variable X satisfies the Strong Density Assumption if*

- *it admits a density μ w.r.t. the Lebesgue measure of $\mathbb{R}^d$,*
- *the density μ satisfies:*

$$\mu_- \leq \mu(x) \leq \mu_+, \quad \forall x \in \mathrm{Supp}(\mu)$$

  for some constants $(\mu_-, \mu_+) \in]0, +\infty[^2$.
- *The support of μ is (c_0, r_0)-regular, namely:*

$$\lambda\left[\mathrm{Supp}(\mu) \cap B(x,r)\right] \geq c_0 \lambda[B(x,r)], \forall r \leq r_0,$$

  for some positive constants c_0 and r_0.

As soon as the marginal density is bounded from below by a strictly positive constant, then both SDA and Strong Minimal Mass Assumption (**A3**) are equivalent, as stated in the following proposition.

PROPOSITION 3.1. *For bounded away from zero density, the SDA is equivalent to the Strong Minimal Mass Assumption.*

PROOF. As soon as the support of μ is (c_0, r_0)-regular and the density is lower bounded by $\mu_- > 0$, then SDA implies a minimal mass type assumption since $\forall \delta \leq r_0$:

$$\mathbb{P}_X(B(x,\delta)) = \int_{B(x,\delta)} \mu(z)dz \geq \mu_- \times \lambda[B(x,\delta) \cap \mathrm{Supp}(\mu)] \geq c_0 \gamma_d \mu_- \delta^d.$$

Conversely, we can also check the fact that the Strong Minimal Mass Assumption (A3) implies the SDA (including the (c_0, r_0)-regularity of μ). Indeed, since for any x and $\delta \leq \delta_0$:

$$1 \geq \int_{B(x,\delta)} \mu(x)dx \geq C\mu(x)\delta^d,$$

then the density μ is upper bounded and we obtain that:

$$\int_{B(x,\delta)} \mu(x)dx \leq \|\mu\|_\infty \lambda \left[\mathrm{Supp}(\mu) \cap B(x,r)\right].$$

We therefore obtain:

$$\lambda \left[\mathrm{Supp}(\mu) \cap B(x,r)\right] \geq C \frac{\mu(x)}{\|\mu\|_\infty} \delta^d \geq C \frac{\mu_-}{\|\mu\|_\infty} \delta^d.$$

This concludes the proof of this proposition. □

It is possible to link the constants (c_0, r_0) involved in SDA with κ involved in $\mathfrak{M}_{mma}(\Omega, \kappa)^{\mu_-}$, but we have omitted their relationships here for the sake of simplicity. Minimax rates of excess risk under the SDA are established in [AT07]. A consequence of Proposition 3.1 is that the same lower bound is still valid with $\mathfrak{M}_{mma}(\Omega, \kappa)^{\mu_-}$.

THEOREM 3.2 (Theorem 3.3, [AT07]). *Assume that Assumptions **A1-A3** hold and a $\mu_- > 0$ exists such that $\mu(x) > \mu_-$ for all $x \in \Omega$. Then, the minimax classification rate is lower bounded as follows:*

$$\inf_{\Phi} \sup_{\mathbb{P}_{X,Y} \in \mathcal{F}_{L,\alpha} \cap \mathfrak{M}_{mma}(\Omega,\kappa)^{\mu_-}} \left[\mathcal{R}(\Phi) - \mathcal{R}(\Phi^*)\right] \gtrsim n^{-\frac{1+\alpha}{2+d}}.$$

Thanks to the previous lower bound, we can conclude that the nearest neighbor rule achieve the minimax rate of convergence in the particular case where the density μ is lower bounded on its (compact) support. As already discussed in [MT99] or [AT07], the higher the margin index α is, the smaller the excess risk will be. On the other hand, the performance deteriorates as the dimension of the considered problem increases. This corresponds to the classical curse of the dimensionality. The lower bound obtained by [AT07] is based on an adaptation of standard tools from nonparametric statistics (Assouad's Lemma). This proof is of primary importance for next lower bound results. It is recalled in Section A for the sake of convenience.

3.2. *The Smooth discriminant analysis model (Binomial sample-size).* While the supervised classification model (also referred to as the Poisson sample-size model) has been intensively studied in the last decades, the smooth discriminant analysis model has been considered as an alternative approach. This model is presented in [MT99] and is referred to as a binomial model in [HPS08]. It assumes that we have two independent samples $\mathcal{S}_1 = (X_1, \ldots, X_n)$ and $\mathcal{S}_2 = (\tilde{X}_1, \ldots, \tilde{X}_n)$ of i.i.d. random variables at our disposal, with densities f and g respectively. Given a new incoming observation, the goal is then

to predict its corresponding label, namely to determine whether X comes from the density f or g.

In the classification setting, the positions are drawn according to μ and the labels are then sampled using $\mathcal{B}(\eta(X))$, which makes the values of the labels $(Y_{(i)})_{1\leq i\leq n}$ completely independent each other, conditionally to their positions $(X_{(i)})_{1\leq i\leq n}$. This key observation is no longer true in the smooth discriminant analysis: conditionally to ordered spatial inputs induced in the nearest neighbor rule, the random variables $(Y_{(1)},\ldots,Y_{(k_n)})$ **are not independent**. This significantly complicates the analysis of the nearest neighbor rule and is a major difference with the standard classification task.

We briefly provide our main result on the nearest neighbor rule with the smooth discriminant analysis below. More complete details can be found in Appendix B.

THEOREM 3.3. *The nearest neighbor classifier* Φ_{n,k_n} *with* $k_n = \lfloor n^{\frac{2}{2+d}} \rfloor$ *satisfies*

$$\sup_{\mathbb{P}_{X,Y}\in\mathcal{F}_{L,\alpha}\cap\mathfrak{M}_{mma}(\Omega,\kappa)^{\mu-}} \left[\mathcal{R}^{Binom}(\Phi_{n,k_n}) - \mathcal{R}^{Binom}(\Phi^*)\right] \lesssim \log(n) n^{-\frac{1+\alpha}{2+d}},$$

where $\mathcal{R}^{Binom}$ *denotes the risk in the smooth discriminant analysis setting.*

To the best of our knowledge, the performance of the nearest neighbor classifier in the binomial sample-size model has only been studied in [HPS08]. In their paper, the difference between the Poisson and the binomial model is studied through Reny's representation of order statistics. In contrast, we directly compute an upper bound of the binomial model. Our main argument relies on a Poissonization of the sample size (see, *e.g.*, [Kac49]). Even if it is a standard alternative to cope with dependencies in probability, such a method has not yet been applied for smooth discriminant analysis.

Regarding the obtained consistency rates now, our result misses a log term in the smooth discriminant analysis setting. In [HPS08], the authors show that the difference of the excess risk between the classification and the smooth discriminant analysis is on the order of $o\left(k^{-1} + \left(\frac{k}{n}\right)^{4/d}\right)$ for *twice* differentiable functions η (instead of only the Lipschitz situation in our case) and their resulting rate is $n^{-2/(4+d)}$ for the optimal choice $k_n = n^{4/(4+d)}$. Following their argument with a *Lipschitz* regression function η, their excess risk becomes $n^{-1/(2+d)}$ for the binomial model. Hence, for a margin $\alpha = 0$, our result in Theorem 3.3 is weaker than the one in [HPS08] (because of

our log term). This is not yet the case as soon as the margin $\alpha > 0$ since the result of [HPS08] does not take this parameter, which may be central to obtain fast rates, into account. Moreover, the approach of [HPS08] does not seem to simply manage the margin information of the classification.

Finally, our Poissonization method also applies for general densities that are not necessarily bounded from below (see Appendix B). This is a major difference with the results of [HPS08] that are valid with a compactly supported and bounded away from zero density μ.

4. General finite dimensional case.

4.1. *The Tail Assumption.* Results of the previous section are designed for the problem of supervised binary classification with compactly supported inputs and lower bounded densities. Such an assumption is an important prior on the problem that may be improper in several practical settings. Various situations involve Gaussian, Laplace, Cauchy or Pareto distributions on the observations, and both the compactness and the boundedness away from zero assumptions may seem to be very unrealistic. This is even more problematic when dealing with functional classification with a Gaussian White Noise model (GWN). In such a case, observations are described through an infinite sequence of Gaussian random variables and the SDA or $\mathfrak{M}_{mma}(\Omega, \kappa)^{\mu-}$ are far from being well-tailored for this situation (see [Lia11] for a discussion and further references).

This section is dedicated to a more general case of binary supervised classification problems where the marginal density μ of X is no longer assumed to be lower bounded on its support. The main problem related to such a setting is that we have to predict labels in places where few (or even no) observations are available in the training set. In order to address this problem, we take the following assumption.

Assumption ***A4****.* (Tail Assumption) *Afunction ψ that satisfies $\psi(\epsilon) \to 0$ as $\epsilon \to 0$ and that increases in a neighborhood of 0 exists such that*

$$\mathbb{P}_{(X,Y)} \in \mathcal{P}_{\mathcal{T},\psi} := \left\{ \mathbb{P}_X \,:\, \exists\, \epsilon_0 \in \mathbb{R}_+^* : \forall \epsilon < \epsilon_0,\, \mathbb{P}_X\left(\{\mu < \epsilon\}\right) \leq \psi(\epsilon) \right\},$$

where $\mathcal{P}_{\mathcal{T},Id}$ corresponds to the particular case where $\psi = Id$.

The aim of this Tail Assumption is to ensure that the set where μ is small has a small mass. We use the notation $\mathcal{T}$ because of the interpretation on the *tail* of μ, but $\mathcal{P}_{\mathcal{T},\psi}$ is not just an assumption on the tail of the μ. It is, in fact, an assumption on the behavior of μ near the set $\{\mu = 0\}$. We

provide some examples of marginal distribution below that satisfy this tail requirement. In Section 4.2 below, we prove that the Tail Assumption (**A4**) is unavoidable in this setting. In Section 4.3, we investigate the performances of the nearest neighbor rule in this setting.

EXAMPLE 4.1. *Following are several families of densities in* $\mathcal{P}_{\mathcal{T},\psi}$.

- *Laplace distributions obviously satisfy* $\mathcal{P}_{\mathcal{T},Id}$, *and a straightforward integration by parts shows that Gamma distributions* $\Gamma(k,\theta)$ *satisfy* $\mathcal{P}_{\mathcal{T},\psi}$ *with* $\psi(\epsilon) = \epsilon \log(\epsilon^{-1})^{k-1}$ *(the term around* $x = 0$ *is on the order of* $\epsilon^{k/(k-1)}$ *and thus negligible compared to the term around* $+\infty$*).*
- *An immediate computation shows that the family of Pareto distributions of parameters* (x_0, k) *satisfies* $\mathcal{P}_{\mathcal{T},\psi}$ *where* $\psi(\epsilon) = \epsilon^{k/(k+1)}$, *regardless of the value of* x_0.
- *The family of Cauchy distributions satisfies* $\mathcal{P}_{\mathcal{T},\psi}$ *with* $\psi(\epsilon) = \sqrt{\epsilon}$.
- *Univariate Gaussian laws* γ_{m,σ^2} *with mean* m *and variance* σ^2 *satisfy*

$$\gamma_{m,\sigma^2}(x) \leq \epsilon \Longleftrightarrow |x-m| \geq t_{\sigma,\epsilon} := \sqrt{2}\sigma\sqrt{\log\left(\frac{1}{\epsilon}\right) + \log(\frac{1}{\sigma\sqrt{2\pi}})},$$

and a standard result on the size of Gaussian tails (see [BNC89]) yields

$$\gamma_{m,\sigma^2}\left(\gamma_{m,\sigma^2} \leq \epsilon\right) = \frac{\epsilon}{t_{\sigma,\epsilon}}\left[1 - \frac{1}{t_{\sigma,\epsilon}^2} + \frac{1.3}{t_{\sigma,\epsilon}^4}\ldots\right] \lesssim \frac{\epsilon}{\sqrt{\log\left(\frac{1}{\epsilon}\right)}}.$$

Hence, univariate Gaussian laws satisfy $\mathcal{P}_{\mathcal{T},\psi}$ *with* $\psi(\epsilon) = \epsilon \log(\epsilon^{-1})^{-1/2}$.
- *If* $\mathbf{m}$ *is any real vector of* $\mathbb{R}^d$ *and* Σ^2 *a covariance matrix whose spectrum is* $\lambda_1 \geq \ldots \lambda_d \geq 0$*:*

$$\gamma_{\mathbf{m},\Sigma^2}\left(\gamma_{\mathbf{m},\Sigma^2} \leq \epsilon\right) = \gamma_{\mathbf{0},\Sigma^2}\left(\gamma_{\mathbf{0},\Sigma^2} \leq \epsilon\right) \lesssim \gamma_{\mathbf{0},\Sigma^2}\left(\|X\| \geq \sqrt{2\lambda_1 \log\left(\frac{1}{\epsilon}\right)}\right).$$

Careful inspection of Theorem 1 of [HLS02] now yields

$$\gamma_{\mathbf{0},\Sigma^2}\left(\|X\| \geq \sqrt{2\lambda_1 \log\left(\frac{1}{\epsilon}\right)}\right) \sim C_{\Sigma^2} \log\left(\frac{1}{\epsilon}\right)^{r/2-1} \epsilon,$$

where C_{Σ^2} *is a constant that only depends on the spectrum of* Σ^2 *and* r *is the multiplicity of the eigenvalue* λ_1. *In particular,* $\gamma_{\mathbf{m},\Sigma^2}$ *satisfy* $\mathcal{P}_{\mathcal{T},\psi}$ *where* $\psi(\epsilon) = C_{\Sigma^2}\epsilon \log(\epsilon^{-1})^{r/2-1}$.

4.2. *Non-consistency results.* We first justify the introduction of the sets $\mathfrak{M}_{mma}(\Omega, \kappa)$ and $\mathcal{P}_{\mathcal{T},\psi}$ and discuss their influences regarding uniform lower bounds and even consistency of any estimator. To do this, we first state that the Minimal Mass Assumption (**A3**) is necessary to obtain *uniformly* consistent classification rules. Second, we assert that the Tail Assumption (**A4**) is also unavoidable.

THEOREM 4.1. *Assume that the law $\mathbb{P}_{X,Y}$ belongs to $\mathcal{F}_{L,\alpha}$, then:*

i) No classification rule can be universally consistent if Assumptions **A1-A3** hold and not **A4**. *For any discrimination rule Φ_n and for any $\epsilon < 4^{-\alpha}$, a distribution $\mathbb{P}_{(X,Y)}$ in $\mathcal{F}_{L,\alpha} \cap \mathfrak{M}_{mma}(\Omega, \kappa)$ exists such that:*

$$\mathcal{R}(\Phi_n) - \mathcal{R}(\Phi^*) \geq \epsilon.$$

ii) No classification rule can be universally consistent if Assumption **A1, A2, A4** hold and not **A3**. *For any discrimination rule Φ_n and for any $\epsilon < 4^{-\alpha}$, a distribution $\mathbb{P}_{(X,Y)}$ in $\mathcal{F}_{L,\alpha} \cap \mathcal{P}_{\mathcal{T},Id}$ exists such that:*

$$\mathcal{R}(\Phi_n) - \mathcal{R}(\Phi^*) \geq \epsilon.$$

The first result *i*) asserts that even if the Minimal Mass Assumption **A3** holds for the underlying density on X, it is not possible to expect a uniform consistency result over the entire class of non-compactly considered densities. In some sense, the support of the variable X seems to be too large to obtain reliable predictions with any classifiers without additional assumptions. As discussed above, the Tail Assumption **A4** may make it possible to counterbalance this curse of support effect (see next section). Such statistical damage has also been observed for the estimation of densities that are supported on the real line instead of being compactly supported, even though such dramatic consequences are not shown here. We refer to [RBRTM11] and the references therein for a more detailed description.

The second result *ii*) states that the Strong Minimal Mass Assumption **A3** cannot be skipped for uniform consistency rates and no compactly supported densities. This is in line with the former studies of [Győ78] and [DGKL94]. In particular, Lemma 2.2 of [DGKL94] takes advantage of some of the positive consequences of this type of assumption. Our proof relies on the construction of a sample size dependent law on (X, Y) that violates our Minimal Mass Assumption **A3** but *that keeps the regression function η in our smoothness class $\mathcal{F}_{L,\alpha}$*. This is a major difference with former counter examples built in [DGL96] where the non uniform consistency is obtained with a family of *non-smooth* regression functions η. In our study, we also

obtained a family of *smooth* regression functions for which such phenomena occur. Even in this case, it is still possible to keep the excess risk strictly positive for any classifier Φ_n (and no longer for only nearest neighbor rules).

Finally, it should be noted that our inconsistency results always occur when building a network of regression functions η that oscillate around the value $1/2$ at the neighborhood of the set $\{\mu = 0\}$. In a sense, Theorem 4.1 contributes to the understanding of one of the opens question put forth in [Can13] on the behavior of the nearest neighbor rule when η is oscillating about $1/2$ in the tail.

4.3. *Minimax rates of convergence.* In the meantime, when both **A2**, **A3** and **A4** hold, we are able to precisely describe the corresponding minimax rate of convergence.

4.3.1. *Minimax lower bound.*

THEOREM 4.2. *Assume that Assumptions **A1-A4** hold. Then*

$$\inf_{\Phi_n} \sup_{\mathbb{P}_{(X,Y)} \in \mathcal{F}_{L,\alpha} \cap \mathfrak{M}_{mma}(\Omega,\kappa) \cap \mathcal{P}_{\mathcal{T},Id}} [\mathcal{R}(\Phi_n) - \mathcal{R}(\Phi^*)] \gtrsim n^{-\frac{1+\alpha}{2+\alpha+d}}.$$

For the sake of convenience, we briefly outline the proof of Theorem 3.2 borrowed from [AT07] in Section A.1. It is then adapted to our new set of assumptions.

Theorem 4.5 below provides some lower bounds for different tails of distributions (through the function ψ). It should be noted that we recover the known rate of compactly supported densities with the so-called *Mild Density Assumption* of [AT07] in the particular case $\psi = Id$. This implies that in the non-compact case, the rate cannot be improved compared to the compact setting, even with an Additional Tail assumption.

4.3.2. *An upper bound for the nearest neighbor rule.* When the density is no longer bounded away from 0, the integer k_n will be chosen in order to counterbalance the vanishing probability of the small balls in the tail of the distributions. For example, when $\psi = Id$, we show that a suitable choice of the integer k_n is:

$$k_n := \lfloor n^{\frac{2}{3+\alpha+d}} \rfloor,$$

which appears to be quite different from the one in the previous section.

THEOREM 4.3. *Assume that* ***A1-A3*** *hold and if the Tail Assumption* ***A4*** *is driven by* $\psi = Id$*, the choice* $k_n := \lfloor n^{\frac{2}{3+\alpha+d}} \rfloor$ *yields:*

$$\sup_{\mathbb{P}_{(X,Y)} \in \mathcal{F}_{L,\alpha} \cap \mathcal{P}_{\mathcal{T},Id} \cap \mathfrak{M}_{mma}(\Omega,\kappa)} [\mathcal{R}(\Phi_{n,k_n}) - \mathcal{R}(\Phi^*)] \lesssim n^{-\frac{(1+\alpha)}{(3+\alpha+d)}}.$$

The proof of Theorem 4.3 is provided in Section A.3. The above results indicate that the price to pay for the classification from entries in compact sets to arbitrary large sets of $\mathbb{R}^d$ is translated by the degradation from $n^{-(1+\alpha)/(2+d)}$ to at least $n^{-(1+\alpha)/(2+\alpha+d)}$ (see, *e.g.*, Theorem 4.2 when $\psi(\epsilon) \sim \epsilon$). Our upper bound for the nearest neighbor rule does not exactly match this lower bound since we obtain $n^{-(1+\alpha)/(3+\alpha+d)}$ in a similar situation . At this step, obtaining the appropriate minimax rate requires slight changes inside the construction of the nearest neighbor rule. This is the purpose of the next paragraph.

4.3.3. *Minimax upper bound for an optimal nearest neighbor rule.* The upper bound proposed in the theorem can be improved if we change the way in which the regularization parameter k_n is constructed. We use a nearest neighbor algorithm with a number of neighbors that depends on the position of the observation x according to the value of the density $\mu(x)$. More formally, we define for all $j \in \mathbb{N}$

$$\Omega_{n,0} := \left\{ x \in \mathbb{R}^d : \ \mu(x) \geq n^{\frac{-\alpha}{2+\alpha+d}} \right\},$$

and

$$\Omega_{n,j} = \left\{ x \in \mathbb{R}^d : \ \frac{n^{\frac{-\alpha}{2+\alpha+d}}}{2^j} \leq \mu(x) < \frac{n^{\frac{-\alpha}{2+\alpha+d}}}{2^{j+1}} \right\}.$$

Setting $k_{n,0} = \lfloor n^{\frac{2}{2+\alpha+d}} \log(n) \rfloor$, we then use for all $j \in \mathbb{N}$

$$k_n(x) = \lfloor k_{n,0} 2^{-2j/(2+d)} \rfloor \vee 1 \quad \text{when} \quad x \in \Omega_{n,j}. \tag{4.1}$$

According to (4.1), the number of neighbors involved in the decision process depends on the spatial position of the input X. In some sense, this position is linked to the tail. The statistical performances of the corresponding nearest neighbor classifier is displayed below. Such a construction of this sequence of "slices" may be interpreted as a spatial adaptive bandwidth selection. This bandwidth is smaller at points $x \in \mathbb{R}^d$ such that $\mu(x)$ is small. In a sense, this idea is close to the one introduced in [GL14] that provides a similar slicing procedure to obtain an adaptive minimax density estimation on $\mathbb{R}^d$.

THEOREM 4.4. *Assume that* ***A1-A3*** *hold and that the Tail Assumption* ***A4*** *is driven by* $\psi = Id$. *Then, if* Φ^*_{n,k_n} *is the classifier associated with (4.1), we have:*

$$\sup_{\mathbb{P}_{(X,Y)} \in \mathcal{F}_{L,\alpha} \cap \mathcal{P}_{\mathcal{T},Id} \cap \mathfrak{M}_{mma}(\Omega,\kappa)} \left[\mathcal{R}(\Phi^*_{n,k_n}) - \mathcal{R}(\Phi^*)\right] \lesssim n^{-\frac{(1+\alpha)}{(2+\alpha+d)}} (\log n)^{\frac{1}{2}+\frac{1}{d}}.$$

We stress that the upper bound obtained in Theorem 4.4 nearly matches the lower bound proposed in Theorem 4.2, up to a log-term. This log-term can be removed by the use of additional technicalities that are omitted in our proof. Hence, Theorems 4.4 and 4.2 make it possible to identify the exact minimax rate of classification when the Tail Assumption is driven by $\psi = Id$, that is:

$$\inf_{\Phi} \sup_{\mathbb{P}_{(X,Y)} \in \mathcal{F}_{L,\alpha} \cap \mathcal{P}_{\mathcal{T},Id} \cap \mathfrak{M}_{mma}(\Omega,\kappa)} \left[\mathcal{R}(\Phi^*_{n,k_n}) - \mathcal{R}(\Phi^*)\right] \sim n^{-\frac{1+\alpha}{2+\alpha+d}}.$$

4.3.4. *Generalizations.* We propose several extensions of our previous results (lower and upper bounds) for more general tails of distribution. We also propose to enlighten the *Minimal Mass Assumption* $\mathfrak{M}_{mma}(\Omega, \kappa)$.

Effect of the tail: from $\mathcal{P}_{\mathcal{T},Id}$ *to* $\mathcal{P}_{\mathcal{T},\psi}$.

THEOREM 4.5. *Assume that Assumptions* ***A1-A4*** *hold. For any tail* $\mathcal{T}$ *parameterized by a function* ψ, *we obtain the following results:*

i) *Lower bound: the minimax classification rate satisfies:*

$$\inf_{\Phi_n} \sup_{\mathbb{P}_{(X,Y)} \in \mathcal{F}_{L,\alpha} \cap \mathcal{P}_{\mathcal{T},\psi} \cap \mathfrak{M}_{mma}(\Omega,\kappa)} [\mathcal{R}(\Phi_n) - \mathcal{R}(\Phi^*)] \gtrsim \epsilon^{1+\alpha}_{n,\alpha,d},$$

where $\epsilon_{n,\alpha,d}$ *satisfies the balance*

$$n^{-1} = \{\epsilon_{n,\alpha,d}\}^{2+d} \times \psi^{-1}\left(\{\epsilon_{n,\alpha,d}\}^{\alpha}\right). \tag{4.2}$$

ii) *Upper bound: the nearest neighbor rule satisfies*

$$\sup_{\mathbb{P}_{(X,Y)} \in \mathcal{F}_{L,\alpha} \cap \mathcal{P}_{\mathcal{T},\psi} \cap \mathfrak{M}_{mma}(\Omega,\kappa)} [\mathcal{R}(\Phi_{n,k_n}) - \mathcal{R}(\Phi^*)] \leq C\nu^{1+\alpha}_{n,\alpha,d}$$

with $k_n = \nu^{-2}_{n,\alpha,d}$ *where* $\nu_{n,\alpha,d}$ *fulfills the balance:*

$$n^{-1} = \psi^{-1}(\{\nu_{n,\alpha,d}\}^{1+\alpha})\{\nu_{n,\alpha,d}\}^{2+d}. \tag{4.3}$$

It would also be possible to propose some generalizations using the sliced nearest neighbor rule presented in Sections 4.3.2 and 4.3.3 for tails driven by a general function ψ, even if we do not include this additional result for the purpose of clarity.

Meeting the Minimal Mass Assumption $\widetilde{\mathfrak{M}}_{mma}(\Omega,\kappa)$. We now obtain similar rates when using the weaker assumption $\widetilde{\mathfrak{M}}_{mma}(\Omega,\kappa)$ instead of $\mathfrak{M}_{mma}(\Omega,\kappa)$: the lower bounds of $\mu(B(x,\delta))$ are only useful for some points x such that $\mu(x)$ is large enough. We can state the next corollary.

COROLLARY 4.1. *Assume that* ***A1,A2,A4*** *hold and* $P_{(X,Y)} \in \widetilde{\mathfrak{M}}_{mma}(\Omega,\kappa)$, *then*

$$\sup_{\mathbb{P}_{(X,Y)} \in \mathcal{F}_{L,\alpha} \cap \mathcal{P}_{\mathcal{T},\psi} \cap \widetilde{\mathfrak{M}}_{mma}(\Omega,\kappa)} [\mathcal{R}(\Phi_{n,k_n}) - \mathcal{R}(\Phi^*)] \lesssim \nu_{n,\alpha,d}^{1+\alpha},$$

with $k_n = \nu_{n,\alpha,d}^{-2}$ *where* $\nu_{n,\alpha,d}$ *satisfies the balance*

$$n^{-1} = \psi^{-1}(\{\nu_{n,\alpha,d}\}^{1+\alpha})\{\nu_{n,\alpha,d}\}^{2+d}.$$

The condition $\mathfrak{M}_{mma}(\Omega,\kappa)$ cannot be easily described through an analytical condition because of its uniform nature over Ω. In contrast, $\widetilde{\mathfrak{M}}_{mma}(\Omega,\kappa)$ is more tractable in view of the criterion given by the next result (Proposition 4.1). Using a log-density model, we write the density μ as

$$\mu(x) = e^{-\varphi(x)}, \forall x \in \mathbb{R}^d.$$

PROPOSITION 4.1. *Let* $\varphi \in \mathcal{C}^1(\Omega)$ *and assume that a real number* $a > 0$ *exists such that:*

$$\lim_{x:\mu(x)\longrightarrow 0} \frac{\|\nabla\varphi(x)\|}{\varphi(x)^a} = 0,$$

then a suitable κ *can be found such that* $\mu = e^{-\varphi} \in \widetilde{\mathfrak{M}}_{mma}(\Omega,\kappa)$.

PROOF. For any $\delta > 0$, we compute a lower bound of

$$\mathbb{P}_X\left(B(x,\delta)\right) = \int_{B(x,\delta)} e^{-\varphi(z)} dz.$$

The Jensen Inequality applied to the normalized Lebesgue measure over $B(x,t)$, which is denoted $\bar{dz}$, yields

(4.4)
$$\int_{B(x,\delta)} e^{-\varphi(z)} dz \geq \frac{\pi^{d/2}\delta^d}{\Gamma(d/2+1)} \exp\left(-\varphi(x) + \int_{B(x,\delta)} [\varphi(z) - \varphi(x)]\bar{dz}\right).$$

A first order Taylor expansion leads to

$$\int_{B(x,\delta)} [\varphi(z) - \varphi(x)]\bar{dz} \leq \sup_{z\in B(x,\delta)} \|\nabla\varphi(z)\| \int_{B(x,\delta)} \|z-x\|\bar{dz} \leq \delta \sup_{z\in B(x,\delta)} \|\nabla\varphi(z)\|.$$

Now, our assumption on φ implies that a large enough C_a exists such that:

$$\|\nabla\varphi(z)\| \leq C_a(1+\varphi(z)^a).$$

Thus, the lower bound (4.4) becomes:

$$\int_{B(x,\delta)} e^{-\varphi(z)}dz \geq \frac{\pi^{d/2}\delta^d}{\Gamma(d/2+1)} e^{-\varphi(x)} e^{-C_a\delta(1+\sup_{z\in B(x,\delta)}\varphi^a(z))}.$$

It is now sufficient to consider points x such that $\varphi \leq \delta^{-1/a}$ (equivalent to $\mu \geq e^{-\delta^{-1/a}}$) to obtain a meaningful lower bound Hence, $\widetilde{\mathfrak{M}}_{mma}(\Omega,\kappa)$ is satisfied choosing

$$\rho = 1/a \qquad \text{and} \qquad \kappa = \frac{\pi^{d/2}}{2\Gamma(d/2+1)} e^{-C_a}.$$

□

4.4. *Practical settings on typical examples .* The aim of this section is to illustrate the results obtained above. We first describe a location model for which we can derive explicit upper and lower bounds in several different cases. We then propose a small numerical study in order to enhance the discussion regarding the importance of the Tail Assumption and we conclude by drawing a comparison between the standard nearest neighbor and sliced nearest neighbor rules.

Explicit rates for specific location models. We investigate here the influence of the function ψ in $\mathcal{P}_{\mathcal{T},\psi}$ as well as the one of the margin parameter on the convergence rates through several specific location models. These models are defined as follows: given any positive random variable Z (whose cumulative distribution function is denoted as F) and two real location values a and b, the random variable X is given by:

$$X = \epsilon Z + Yb + (1-Y)a, \tag{4.5}$$

where ϵ is a Rademacher random variable (whose values is ± 1) independent of Z, and Y is the label of the observation, sampled independently of ϵ and Z with a Bernoulli law $\mathcal{B}(1/2)$. Using a translation invariance argument, it is enough in the next study to consider $a = 0$ and $b > 0$. Table 1 illustrates the rate reached by the nearest neighbor procedure in each situation.

Law of Z	Tail ψ	Margin	$k_n \sim n^\beta$	Upper bound
Gauss	$\psi(\epsilon) \propto \epsilon \log(1/\epsilon)^{r/2-1}$	$\alpha = 1$	$\beta = 2/(4+d)$	$n^{-2/(4+d)} \log(n)^{\beta(r)}$
Laplace	$\psi(\epsilon) \propto \epsilon$	$\alpha = 1$	$\beta = 2/(4+d)$	$n^{-2/(4+d)}$
Gamma	$\psi(\epsilon) \propto \epsilon \log(1/\epsilon)^{k-1}$	$\alpha = 1$	$\beta = 2/(4+d)$	$n^{-2/(4+d)} \log(n)^{\beta(k)}$
Cauchy	$\psi(\epsilon) \propto \sqrt{\epsilon}$	$\alpha = 1$	$\beta = 1/(3+d)$	$n^{-2/(3+d)}$
Power-Pareto	$\psi(\epsilon) \propto \epsilon^{p/(p+1)}$	$\alpha = 1 \wedge p$	$\beta = \frac{2(p+1)}{p(3+\alpha+d)+2+d}$	$n^{\frac{-4(p+1)}{p(3+\alpha+d)+2+d}}$

TABLE 1
Convergence rates for location models with several tail sizes.

A numerical study for 'power laws'. In order to illustrate Equations (4.2) and (4.3), we consider some specific cases of "power laws" such that:

$$\mathbb{P}_X(\mu(X) < \epsilon) = \psi(\epsilon) \sim \epsilon^g \quad \text{when} \quad \epsilon \longrightarrow 0^+,$$

for some $g > 0$. In this case, the upper bound on the Nearest Neighbor classifier is given by

$$\mathcal{R}(\Phi_n) - \mathcal{R}(\Phi^*) \lesssim n^{-\frac{(1+\alpha)}{1+\alpha+\frac{2+d}{g}}}$$

although the lower bound derived from (4.2) is:

$$\inf_{\Phi_n} \sup_{\mathbb{P}_{(X,Y)} \in \mathcal{F}_{L,\alpha} \cap \mathcal{P}_{\mathcal{T},\psi} \cap \mathfrak{M}_{mma}(\Omega,\kappa)} [\mathcal{R}(\Phi_n) - \mathcal{R}(\Phi^*)] \gtrsim n^{-\frac{(1+\alpha)}{\alpha+\frac{2+d}{g}}}$$

We immediately observe that the classification rates are seriously damaged when g is small. In contrast, for very thin tails, the rate can be arbitrarily close to n^{-1}. For this purpose, we illustrate this phenomenon with a family of distributions $\mathcal{P}_g$, where the parameter $g > 0$ influences the tail size. We define the cumulative distribution function of the positive random variable Z:

$$\forall t \geq 0 \qquad F_g(t) = 1 - \frac{1}{(t+1)^g}.$$

Then, for two real values (a, b), we sample n observations (X_i, Y_i) according to the previous model and the Bayes classifier is given by:

$$\Phi^*(X) = \mathbf{1}_{\{X > (a+b)/2\}}.$$

In this example, the margin α is equal to 1 and η is L-Lipschitz. We then consider $k_n = \lfloor n^{2/5} \rfloor + 1$ to assess the statistical performance of the Nearest Neighbor classifier. Figure 1 represents the excess risk obtained by the

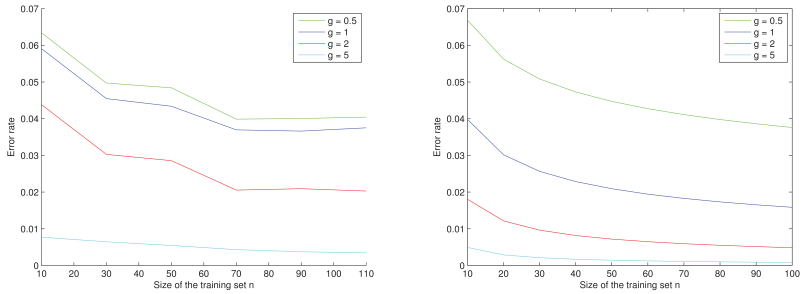

FIG 1. *Example of observed empirical rates and upper bound theoretical rates given by* (4.3) *for several power law distributions of parameter* g.

Nearest Neighbor classifier and the successive degradation of the convergence rate when g decreases to 0 (on the left, the empirical performance of the Nearest Neighbor rule with the underlying distributions and on the right for the upper bound theoretically derived from Theorem (4.3)). These numerical experiments are consistent with the theoretical result obtained in Theorem 4.5 .

Comparison between the standard nearest neighbor and its sliced counterpart. We provide here a short numerical study that aims to compare the results reached by the standard nearest neighbor rule described in Theorem 4.3 and the ones obtained by its sliced counterpart described in Section 4.3.3 and in Theorem 4.4. To measure such an improvement, we have chosen to once again use some non-compactly supported distributions and several different location models.

On the one hand, as pointed out in Theorem 4.3, the standard nearest neighbor will be tuned with a number of neighbor $k_n := \lfloor n^{\frac{2}{3+\alpha+d}} \rfloor + 1$.

On the other hand, the sliced nearest neighbor rule described in Theorem 4.4 requires a preliminary estimation of the law of observation $\mathbb{P}_X$. To do this, we used the recent kernel density estimation package[2] provided by [BGK10], which is an adaptive estimator based on linear diffusion processes. Given any training set $(X_i, Y_i)_{1 \leq i \leq n}$, we first built the preliminary estimator $\hat{\mu}_n$ of the unknown density μ. This estimator is interesting because of its adaptive smoothing properties and because it includes a very fast automatic bandwidth selection algorithm.

The sliced nearest neighbor rule then uses a number of neighbors that depends on the design point X. If the density estimate is large enough, that

[2]kde.m is available on the author's Website of [BGK10].

is, if $\hat{\mu}_n(X) \geq n^{-\frac{\alpha}{2+\alpha+d}}$:

$$k_n(X) := \lfloor n^{\frac{2}{2+\alpha+d}} \rfloor + 1.$$

Otherwise, when $2^{-(j+1)} \leq \hat{\mu}_n(X) n^{\frac{\alpha}{2+\alpha+d}} \leq 2^{-(j)}$, the number $k_n(X)$ is:

$$k_n(X) := \lfloor n^{\frac{2}{2+\alpha+d}} 2^{-\frac{2j}{2+d}} \rfloor + 1.$$

To draw some reliable comparisons, we also used some various laws for the random variable Z involved in the definition of the location model (4.5) (Normal distributions, Cauchy distributions, and Power laws) whose parameters are described in Table 1. The two location parameters are still denoted a and b and fixed such that $a = -b$.

In each situation, we used a Monte-Carlo strategy with 1000 replications to compute the mean excess risk of each nearest neighbor rule. We used a training set of cardinal n, as well as a test set of size 200. Results are given in Table 2.

Law of Z	$n = 100$			$n = 500$			$n = 1000$		
Gauss, $a = 1, \sigma = 2$	$19.2_{.6}$	$18.1_{.6}$	6%	$16.4_{.5}$	$13.9_{.5}$	15%	$15.4_{.5}$	$12_{.5}$	22%
Cauchy, $a = \frac{1}{2}, \gamma = \frac{1}{2}$	$2.6_{.2}$	$1.9_{.2}$	26%	$1.4_{.1}$	$1.2_{.1}$	14%	$0.9_{.05}$	$0.8_{.05}$	6%
Cauchy, $a = \frac{1}{2}, \gamma = 1$	$4.4_{.3}$	$3.6_{.2}$	18%	$3.1_{.3}$	$2.2_{.2}$	28%	$2.3_{.2}$	$1.4_{.2}$	37%
Power, $a = \frac{1}{2}, \gamma = 1$	$3.8_{.3}$	$3_{.3}$	20%	$2.7_{.2}$	$2.1_{.2}$	22%	$1.9_{.2}$	$1.5_{.1}$	19%
Power, $a = \frac{1}{2}, \gamma = 2$	$2_{.2}$	$1.7_{.2}$	13%	$1.2_{.2}$	$1.0_{.1}$	15%	$0.7_{.1}$	$0.6_{.1}$	14%

TABLE 2

Mean excess risk multiplied by 100 *(left: standard nearest neighbor; middle: sliced nearest neighbor; right: percentage of improvement). Standard errors are given in small script.*

We may observe in Table 2 that the sliced version of the nearest neighbor always outperforms the standard one. Such a numerical result is consistent with the theoretical ones of Theorem 4.3 and 4.4. Note also that the relative improvement of the sliced nearest neighbor rule seems to increase when the number of observations n growth, meaning that each excess risk of the two procedures varies with a different power of n.

Finally, it should be mentioned that we have not tried to modify the dimension of the observations X. Indeed, the difference of the upper bounds given by Theorems 4.3 and 4.4 becomes more and more negligible when the dimension is increasing. This should also be the case in the empirical study that will be in the subject of a future work. Likewise, the statistical study of the empirical sliced nearest neighbor rule should also be addressed in a future study, since a balance between the estimation $\hat{\mu}_n$ of the density μ and the excess risk of classification with the sliced rule may exist. We have left this problem open for a future study.

SUPPLEMENTARY MATERIAL

Supplement A: Main proofs for this paper : Classification with the nearest neighbor rule in general finite dimensional spaces: necessary and sufficient conditions.
(doi: COMPLETED BY THE TYPESETTER; .pdf). See in the temporary Appendix section after references.

References.

[AG97] Yali Amit and Donald Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9:1545–1588, 1997.

[Ass83] P Assouad. Deux remarques sur l'estimation. *Comptes Rendus de l'Académie des Sciences*, 296:1021–1024, 1983.

[AT07] Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *Ann. Statist.*, 35(2):608–633, 2007.

[BBL05] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat.*, 9:323–375, 2005.

[BBM08] Gilles Blanchard, Olivier Bousquet, and Pascal Massart. Statistical performance of support vector machines. *Ann. Statist.*, 36(2):489–531, 2008.

[BDL08] Gérard Biau, Luc Devroye, and Gábor Lugosi. Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.*, 9:2015–2033, 2008.

[BFOS84] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and regression trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA, 1984.

[BGK10] Z.I. Botev, J.F. Grotowski, and D.P. Kroese. Kernel density estimation via diffusion. *The Annals of Statistics*, 38:2916–2957, 2010.

[BH79] J Bretagnole and C. Huber. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete.*, 47:119–137, 1979.

[BM06] P.L. Bartlett and S. Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135 (3):311–334, 2006.

[BNC89] O. E. Barndorff-Nielsen and D. R. Cox. *Asymptotic techniques for use in statistics*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1989.

[BR03] Peter J. Bickel and Ya'acov Ritov. Nonparametric estimators which can be "plugged-in". *Ann. Statist.*, 31(4):1033–1053, 2003.

[Bre01] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[Can13] Timothy Cannings. Nearest neighbour classification in the tails of a distribution. *Preprint*, 2013.

[Cas07] Alberto Rodriguez Casal. Set estimation under convexity type assumptions. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 43(6):763 – 774, 2007.

[CG06] Frédéric Cérou and Arnaud Guyader. Nearest neighbor classification in infinite dimension. *ESAIM Probab. Stat.*, 10:340–355 (electronic), 2006.

[CH67] T. Covert and P.E. Hart. Nearest neighbour pattern classification. *IEEE, Transactions on Information Theory*, 13:21–27, 1967.

[Dev81] Luc Devroye. On the almost everywhere convergence of nonparametric regression function estimates. *Ann. Statist.*, 9(6):1310–1319, 1981.

[DGKL94] Luc Devroye, László Györfi, Adam Krzyżak, and Gábor Lugosi. On the strong universal consistency of nearest neighbor regression function estimates. *Ann. Statist.*, 22(3):1371–1385, 1994.

[DGL96] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.

[DW77] Luc P. Devroye and T. J. Wagner. The strong uniform consistency of nearest neighbor density estimates. *Ann. Statist.*, 5(3):536–540, 1977.

[FH51] E Fix and J.L. Hodges. Discriminatory analysis, nonparametric discrimination, consistency properties. *Randolph Field, Texas, Project 21-49-004, Report 4*, 1951.

[FS97] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.*, 55(1, part 2):119–139, 1997. Second Annual European Conference on Computational Learning Theory (EuroCOLT '95) (Barcelona, 1995).

[GKKW02] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002.

[GL14] A. Goldenshluger and O. Lepski. On adaptive minimax density estimation on R^d. *Probab. Theory Related Fields*, 159(3-4):479–543, 2014.

[Győ78] László Győrfi. On the rate of convergence of nearest neighbor rules. *IEEE Trans. Inform. Theory*, 24(4):509–512, 1978.

[Gyö81] László Györfi. The rate of convergence of k_n-NN regression estimates and classification rules. *IEEE Trans. Inform. Theory*, 27(3):362–364, 1981.

[HLS02] Jürg Hüsler, Regina Y. Liu, and Kesar Singh. A formula for the tail probability of a multivariate normal distribution and its applications. *J. Multivariate Anal.*, 82(2):422–430, 2002.

[HPS08] Peter Hall, Byeong U. Park, and Richard J. Samworth. Choice of neighbor order in nearest-neighbor classification. *Ann. Statist.*, 36(5):2135–2152, 2008.

[Kac49] M. Kac. On deviations between theoretical and empirical distributions. *Proc. Nat. Acad. Sci. U. S. A.*, 35:252–257, 1949.

[Kle03] Thierry Klein. *Inégalités de concentration,martingales et arbres aléatoires*. PhD thesis, Université de Versailles-Saint-Quentin (France), 2003.

[Lec07] Guillaume Lecué. Simultaneous adaptation to the margin and to complexity in classification. *Ann. Statist.*, 35(4):1698–1721, 2007.

[Lia11] Heng Lian. Convergence of functional k-nearest neighbor regression estimate with functional responses. *Electron. J. Stat.*, 5:31–40, 2011.

[LM14] S. Loustau and C. Marteau. Minimax fast rates for discriminant analysis with errors in variables. 2014. To appear in Bernoulli.

[LS01] W. V. Li and Q.-M. Shao. Gaussian processes: inequalities, small ball probabilities and applications. In *Stochastic processes: theory and methods*, volume 19 of *Handbook of Statist.*, pages 533–597. North-Holland, Amsterdam, 2001.

[MT99] Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829, 1999.

[RBRTM11] Patricia Reynaud-Bouret, Vincent Rivoirard, and Christine Tuleau-Malot. Adaptive density estimation: a curse of support? *J. Statist. Plann. Inference*, 141(1):115–139, 2011.

[Sam12] R. Samworth. Optimal weighted nearest neighbour classifiers. *Annals of Statistics*, 40:2733–2763, 2012.

[Sha00] Qi-Man Shao. A comparison theorem on moment inequalities between negatively associated and independent random variables. *J. Theoret. Probab.*, 13(2):343–356, 2000.

[Ste05] Ingo Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Trans. Inform. Theory*, 51(1):128–142, 2005.

[Sto77] Charles J. Stone. Consistent nonparametric regression. *Ann. Statist.*, 5(4):595–645, 1977. With discussion and a reply by the author.

[Tsy04] Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166, 2004.

[Tsy09] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[Vap98] Vladimir N. Vapnik. *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, Inc., New York, 1998. A Wiley-Interscience Publication.

APPENDIX A: PROOFS

Recall that $\mathbb{E}$ (resp. $\mathbb{E}_X$, $\mathbb{E}_{\otimes^n}$) denote the expectation with respect to the measure $\mathbb{P}$ (resp. $\mathbb{P}_X$, $\mathbb{P}_{\otimes^n}$).

A.1. Proofs of the lower bounds. The proofs of the lower bounds presented in both Theorem 4.1 and Theorem 4.2 are inspired from the construction proposed in [AT07]. It is based on Assouad's cube method (see [Ass83], and [BH79]). This approach reduces the problem of obtaining a lower bound on the minimax risk to the problem of testing several couples of hypotheses. We refer to [Tsy09] for a comprehensive introduction to this useful method for deriving lower bounds on minimax risk.

A.1.1. *Baseline structure of the network.* We present here the common structure of the network of laws on (X, Y), that is, the definition of the underlying measure $\mathbb{P}_{(X,Y)}$ on $\mathbb{R}^d \times \{0, 1\}$ (through the density μ and the regression function η).

Definition of η. Let $(q, m) \in (\mathbb{N}^*)^2$ and $(x_1, \ldots, x_m) \in \mathbb{R}^d$. We denote by B_i the Euclidean ball of center x_i and of radius $2/q$, such that for any i and j we have $B_i \bigcap B_j = \emptyset$ (we choose $|x_i - x_j| \geq 5/q$). Now consider a C^∞ function φ such that $\|\varphi\|_\infty = 1$, φ is compactly supported in $[0, 2]$ such that $\varphi(x) = 1$ when $|x| \leq 1$, and $\varphi(x) = 0$ for any $x > 3/2$. Now let $\Phi_j(x) = c_\varphi q^{-1} \varphi(q|x - x_j|)$ so that $\Phi_j(x)$ is also C^∞ and supported in $B_j := B(x_j, 2/q)$. Denote by $A_0 = \bigcup_{j=1}^m B_j$ and let $A_1 = [0, 1]^d \bigcap A_0^c$ and $A = A_0 \bigcup A_1$ be the support of the density μ.

Definition of the Assouad Hypercube of regression functions. We define $\Sigma_m = \{-1, 1\}^m$, and for any $\sigma \in \Sigma_m$:

$$\forall 1 \leq j \leq m,\ \forall x \in B_j :\ \eta_\sigma(x) = \frac{1 + \sigma_j \Phi_j(x)}{2}, \quad \text{and} \quad \eta_\sigma(x) = \frac{1}{2} \,\text{if}\, x \in A_1.$$

Figure 2 shows the regression function η_σ for two opposite values of σ_j and for a particular ball B_j.

The density μ. We use in the sequel a measure μ in the sequel that does not depend on σ. Indeed, we even consider only some constant densities on each B_j. In particular, the measure μ of each ball B_j is ω (that will be chosen later) and the density μ is then given by

$$\mu(x) = \frac{\omega}{\lambda(B_j)} = \frac{\omega q^d}{\gamma_d 2^d},\ \forall x : in B_j$$

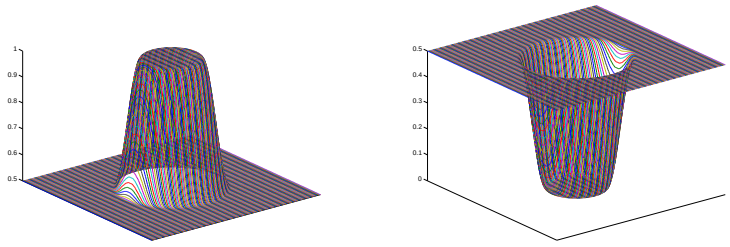

FIG 2. *Example of function η_σ on a particular ball B_j of size $1/q$. The value of η_σ oscillates either between $1/2$ and 1 if $\sigma_j = 1$ or 0 and $1/2$ if $\sigma_j = -1$.*

where γ_d is the Lebesgue measure of the unit Euclidean ball of $\mathbb{R}^d$. We now define μ on A_1 as

$$\mu(x) = \frac{1 - m\omega}{\lambda(A_1)}.$$

A schematic representation of this measure can be seen on the left of Figure 3.

Margin condition. For the sake of convenience, for any $\sigma \in \Sigma_m$, we denote $\mathbb{P}_\sigma := \mathbb{P}_{(X,Y),\sigma}$ the law of the couple (X, Y). Following the arguments of [AT07], consider any $\sigma \in \Sigma_m$:

$$\begin{aligned}\mathbb{P}_\sigma\left(0 < |\eta_\sigma(X) - \frac{1}{2}| \leq t\right) &= m\mathbb{P}_\sigma\left(0 < c_\varphi\varphi(q[|X - x_1|]) \leq 2tq\right) \\ &= m\int_{B(x_1,2/q)} \mathbf{1}_{\{0\leq c_\varphi\varphi(q[|x-x_1|])\leq 2tq\}}\mu(x)dx\end{aligned}$$

Since φ is equal to 1 on $[0, 1]$, we then obtain that:

$$\begin{aligned}\mathbb{P}_\sigma\left(0 < |\eta_\sigma(X) - \frac{1}{2}| \leq t\right) &\leq m\int_{B(x_1,2/q)} \mathbf{1}_{\{c_\varphi\leq 2tq\}}\mu(x)dx \\ &= \mathbf{1}_{\{c_\varphi\leq 2tq\}}m\omega \lesssim t^\alpha.\end{aligned}$$

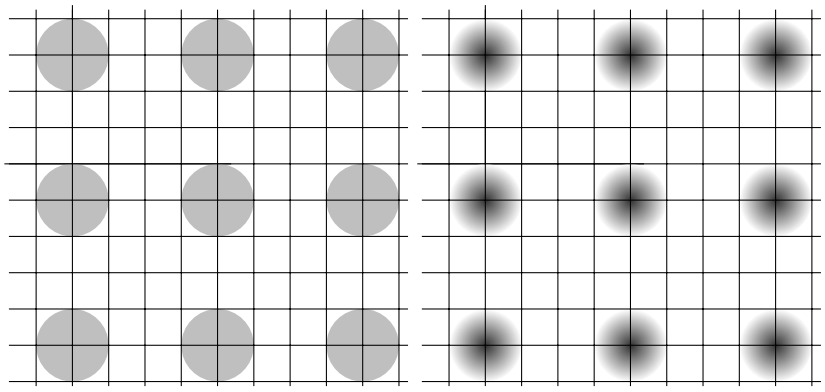

FIG 3. *Simplified representation of the measure $\mathbb{P}_X$, the gray level is proportional to the value of the density μ. Left: measure used in Section A.1.1 or A.1.2 (compactly supported measure or not) and in Section A.1.3 ($\mathfrak{M}_{mma}(\Omega,\kappa)$ is fulfilled and not the Tail Assumption). Right: measure used in Section A.1.4 when the tail is fulfilled and not $\mathfrak{M}_{mma}(\Omega,\kappa)$.*

as soon as:

$$m\omega = O(q^{-\alpha}).$$

Smoothness of η_σ. We briefly check that the regression functions are Lipschitz, uniformly with respect to any choice of q. First, it should be observed that:

$$\forall (x,\tilde{x}) \in B_j \qquad |\eta_\sigma(x) - \eta_\sigma(\tilde{x})| = \frac{|\Phi_j(x) - \Phi_j(\tilde{x})|}{2} \leq \frac{c_\varphi \|\varphi'\|_\infty}{2} \|x - \tilde{x}\|$$

On the contrary, when $(x,\tilde{x}) \in A_1$, $\eta_\sigma(x) = \eta_\sigma(\tilde{x}) = 1/2$. It now remains to study the situation when $x \in A_1$ and $\tilde{x} \in B_j$ for one j. When $\tilde{x}$ is in the exterior ring of size $3/(2q)$ (the set $B_j \cap B(x_j, 3/(2q))^c$), we have:

$$\eta_\sigma(x) = \eta_\sigma(\tilde{x}) = 1/2.$$

Now, if $\tilde{x}$ belongs to $B(x_j, 3/(2q))$:

$$|\eta_\sigma(x) - \eta_\sigma(\tilde{x})| = \left|\frac{\Phi_j(\tilde{x})}{2}\right| \leq \frac{c_\varphi \|\varphi\|_\infty}{2q} \leq \frac{c_\varphi \|\varphi\|_\infty}{} \|x - \tilde{x}\|$$

Hence, we can deduce the uniform Lipschitz bound (note that the case $x \in B_j$ and $\tilde{x} \in B_k$ can be treated in the same way):

$$\forall (x,\tilde{x}) \in (\mathbb{R}^d)^2, \forall \sigma \in \Sigma_n \qquad |\eta_\sigma(x) - \eta_\sigma(\tilde{x})| \leq c_\varphi \frac{\|\varphi'\|_\infty + \|\varphi\|_\infty}{} \|x - \tilde{x}\|.$$

Minoration of the risk. Following the arguments of Theorem 3.5 in [AT07], we have that:

$$\mathcal{R}_n \geq m\frac{\omega}{q}\left(1 - q^{-1}\sqrt{n\omega}\right).$$

A.1.2. *Proof of Theorem 4.2 and of Theorem 4.5, i).* We first study the situation of rates when $\mathfrak{M}_{mma}(\Omega,\kappa)$ and $\mathcal{P}_{\mathcal{T},\psi}$ are in force and use a measure similar to the one represented on the left of Figure 3.

Besicovitch-like condition $\mathfrak{M}_{mma}(\Omega,\kappa)$:. We aim to show that our network satisfies the lower bound involved in $\mathfrak{M}_{mma}(\Omega,\kappa)$. Consider $\delta \to 0^+$ and $q \to +\infty$. If $x \in A_0$ and $\delta = o(1/q)$ then one ball B_j intersects at the least half of $B(x,\delta)$ and since μ is stepwise constant:

$$\mathbb{P}_X\left(B(x,\delta)\right) \geq \frac{\mu(x)\lambda\left(B(x,\delta)\right)}{2} \geq \frac{\gamma_d}{2}\mu(x)\delta^d$$

If δ is now proportional to $1/q$, the last inequality is still true up to a constant (which is not illustrated here for the sake of simplicity). Now if $q^{-1} = o(\delta)$, $B(x,\delta)$ contains a number $N_{\delta,q}$ of balls $(B_j)_{1\leq j\leq m}$ such that $N_{\delta,q} \geq C_d\frac{\delta^d}{q-d}$. In this case, we still have

$$\mathbb{P}_X\left(B(x,\delta)\right) \geq \mathbb{P}_X\left(B(x,\delta)\cap \cup_{j=1}^m B_j\right) \geq N_{\delta,q}\times\omega \geq C_d\delta^d q^d\omega = \frac{C_d}{\gamma_d}\mu(x)\delta^d.$$

Hence, the measure $\mathbb{P}_X$ belongs to $\mathfrak{M}_{mma}(\Omega,\kappa)$ with a constant κ independent of q.

Tail Assumption $\mathcal{P}_{\mathcal{T},Id}$ or $\mathcal{P}_{\mathcal{T},\psi}$. First, note that $\mathbb{P}_X$ is built such that if $x \in A_0$

$$\mathbb{P}_X(\mu < \epsilon) = 0 \text{ if } \epsilon < \omega q^d/(\gamma_d 2^d) \text{ and } \mathbb{P}_X(\mu < \epsilon) = m\omega \text{ if } \epsilon > \omega q^d/(\gamma_d 2^d).$$

Note that the density on A_1 is bounded from below and, as a result, we will not take the tail property on this set into account.
Since ψ is increasing in a neighborhood of 0, the tail property $\mathbb{P}_X(\mu < \epsilon) \lesssim \psi(\epsilon)$ is fulfilled as soon as:

$$m\omega \lesssim \psi\left(\frac{\omega q^d}{\gamma_d 2^d}\right).$$

Calibration for the minoration. Recall that $\mathcal{R}_n \geq m\frac{\omega}{q}\left(1 - q^{-1}\sqrt{n\omega}\right)$ and that we must satisfy the following constraints

$$m\omega = O(q^{-\alpha}) \text{ and } m\omega \lesssim \psi\left(\frac{\omega q^d}{\gamma_d 2^d}\right).$$

The lower bound above is meaningful as soon as we choose $\omega \leq \frac{q^2}{n}$. If we denote $\epsilon_{n,\alpha,d} = q^{-1}$, the values of m, q, ω that provide a tradeoff between all these constraints are obtained with

$$m\omega = q^{-\alpha}, \ \frac{\omega q^d}{\gamma_d 2^d} = \psi^{-1}(m\omega), \ \omega = \frac{q^2}{2n}.$$

In particular, the constraints are optimized when $\epsilon_{n,\alpha,d}$ solves $2^{-d}\gamma_d^{-1}\frac{\epsilon_{n,\alpha,d}^{-2}}{2n}\epsilon_{n,\alpha,d}^{-d} = \psi^{-1}(\epsilon_{n,\alpha,d}^{\alpha})$, which leads to the lower bound

$$\mathcal{R}_n \gtrsim \epsilon_{n,\alpha,d}^{1+\alpha} \quad \text{with} \quad n^{-1} = \epsilon_{n,\alpha,d}^{d+2}\psi^{-1}(\epsilon_{n,\alpha,d}^{\alpha}).$$

In the above calibration, we obtain that:

$$\omega = q^{-d}\psi^{-1}(q^{-\alpha}) \quad \text{and} \quad m = q^d \frac{q^{-\alpha}}{\psi^{-1}(q^{-\alpha})}.$$

This ends the proof of Theorem 4.2 and Theorem 4.5, i). □

Looking carefully at the proof of the theorem above, we can see that the influence of ψ is as follows:

- If $\epsilon = o(\psi(\epsilon))$, then the construction of the network yields a non compactly supported distribution since:

$$\lambda(Supp(\mu)) \gtrsim mq^{-d} = \frac{q^{-\alpha}}{\psi^{-1}(q^{-\alpha})} \longrightarrow +\infty \qquad \text{as} \qquad q \longrightarrow +\infty.$$

  As pointed out in paragraph 4.4, a polynomial decay of the density when x grows to ∞ yields such a tail size.
- In the opposite situation, when $\psi(\epsilon) = O(\epsilon)$, the corresponding density has a compact support. In particular, when $\psi(\epsilon) \sim \epsilon$, our network is exactly the same as the one used in [AT07] and we naturally recover the lower bound $n^{-(1+\alpha)/(2+\alpha+d)}$.

A.1.3. *Proof of Theorem 4.1, item i).* We study the specific case where the Besicovitch has to be fulfilled although the Tail Assumption is no longer necessary. In such a case, we still use the construction shown on the left of Figure 3 and provided in Section A.1.2 but m *can be chosen much greater than* q^d. For example, for a parameter $\tau > 0$ chosen in the sequel, we assume that $m = q^{d+\tau} >> q^d$ as $q \longrightarrow +\infty$. In such a case, the underlying measure $\mathbb{P}_X$ is no longer compactly supported.

Using the same argument as above, Assumption **A3** is still satisfied since the number m of balls B_i does not influence the minoration of $\mathbb{P}_X(B(x,\delta))$.

We have to satisfy the following constraints:

$$m\omega = O(q^{-\alpha}), \omega \leq \frac{q^2}{n}.$$

We keep the value of ω as:

$$\omega = q^2/(2n),$$

and the calibration of q with respect to m yields

$$q = n^{\frac{1}{2+d+\alpha+\tau}}.$$

We then obtain the lower bound

$$\mathcal{R}_n \geq c_\phi n^{-\frac{1+\alpha}{2+\alpha+d+\tau}}.$$

By increasing the size of τ ($\tau_n = n$ for example), it can then be observed that it is possible to obtain any arbitrary value between 0 and c_ϕ. Hence, for any classifier Φ_n, a distribution on (X, Y) exists such that Assumptions **A1-A3** hold and that the classifier Φ_n cannot be consistent.

A.1.4. *Proof of Theorem 4.1, item ii).* We then study what could happen when the Tail Assumption is satisfied but Assumption **A3** can be violated. The idea is to pick the density of observations to ensure the validity of the Tail Assumption. To do this, we consider the new marginal on X whose μ defined as:

$$\forall x \in B_j \qquad \mu(x) = \omega \frac{q^{\gamma d}\left(1 - |x - x_j| q^{\gamma}\right)_+}{\int_{B(0,1)} \left(1 - |x|\right)_+ dx}$$

so that:

$$\int_{B_j} \mu(x) dx = \omega.$$

The obtained measure is represented on the right of Figure 3. We proceed in the same way as in paragraph A.1.1: φ is still lower bounded by a strictly positive constant (as soon as $\gamma \geq 1$) and the Margin Assumption is satisfied as soon as $m\omega = O(q^{-\alpha})$.
It should also be observed that Assumption **A3** is not satisfied here. In fact, when we choose $\gamma > 1$ and the reference radius δ as $\delta = q^{-a}$ for $a \in [1, \gamma[$:

$$\mathbb{P}_X(B(x_j, q^{-a})) = \omega \quad \text{and} \quad \delta^d \mu(x_j) = c q^{-ad} \omega q^{\gamma d} = c\omega q^{d(\gamma - a)},$$

where

$$c = \frac{1}{\int_{B(0,1)} \left(1 - |x|\right)_+ dx}$$

The left hand side becomes negligible with respect to the right hand side as soon as $q \longrightarrow +\infty$.
We now check that such a definition of density μ satisfies the Tail Assumption. Consider any $\epsilon > 0$. We then have:

$$\begin{aligned}
&\mathbb{P}_X\left(\{\mu < \epsilon\}\right) \\
&= m\omega \int_{B(x_1,1/q)} cq^{\gamma d}\left(1 - |x - x_1|q^{\gamma}\right)_+ \mathbf{1}_{\{c\omega q^{\gamma d}(1-|x-x_j|q^{\gamma})_+ \leq \epsilon\}} dx \\
&= m\omega \int_{B(x_1,1/q)} cq^{\gamma d}\left(1 - |x - x_1|q^{\gamma}\right)_+ \mathbf{1}_{\{(1-|x-x_j|q^{\gamma})_+ \leq c^{-1}\omega^{-1}q^{-\gamma d}\epsilon\}} dx.
\end{aligned}$$

Consider the variable $y = q^{\gamma}(x - x_1)$. We then obtain

$$\mathbb{P}_X\left(\{\mu < \epsilon\}\right) = m\omega \int_{B(0,1)} c\left(1 - |y|\right)_+ \mathbf{1}_{\{(1-|y|)_+ \leq c^{-1}\omega^{-1}q^{-\gamma d}\epsilon\}} dy \leq \gamma_d m q^{-\gamma d}\epsilon.$$

As a consequence, the Tail Assumption is true as soon as $m = O(q^{\gamma d})$. We point out that since we chose $\gamma > 1$ in the sequel, m is then greater than q^d and the support of μ is no longer compact since $q \longrightarrow +\infty$.
Following the roadmap of paragraph A.1.2, we then obtain the lower bound calibrations of q and ω such that:

$$\mathcal{R}_n \geq n^{-\frac{1+\alpha}{2+\alpha+\gamma d}}.$$

Again, a sufficiently large value of γ makes it possible to obtain arbitrarily slow rates (and even non-consistent classifiers).

A.2. Proof of Theorem 3.1. Let $\epsilon > 0$ be a given real number (whose value will be specified later), and define:

$$\mathcal{B}_\epsilon := \left\{ x \in \mathbb{R}^d \quad | \quad |\eta(x) - 1/2| \leq \epsilon \right\}.$$

Applying Proposition A.1 in Section A.5, the excess risk can be decomposed as follows:

$$\begin{aligned}
\mathcal{R}(\Phi_n) - \mathcal{R}(\Phi^*) &= \mathbb{E}\left[|2\eta(X) - 1|\mathbf{1}_{\{\Phi_n(X) \neq \Phi^*(X)\}}\right], \\
&= \underbrace{\mathbb{E}\left[|2\eta(X) - 1|\mathbf{1}_{\{\Phi_n(X) \neq \Phi^*(X)\}}\mathbf{1}_{X \in \mathcal{B}_\epsilon}\right]}_{:=T_{1,\epsilon}} \\
&\quad + \underbrace{\mathbb{E}\left[|2\eta(X) - 1|\mathbf{1}_{\{\Phi_n(X) \neq \Phi^*(X)\}}\mathbf{1}_{X \in \mathcal{B}_\epsilon^c}\right]}_{:=T_{2,\epsilon}}.
\end{aligned}$$

Now, the Margin Assumption **A2** yields:

$$T_{1,\epsilon} \leq 2\mathbb{E}\left[|\eta(X) - 1/2|\mathbf{1}_{X\in\mathcal{B}_\epsilon}\right] \leq 2\epsilon\mathbb{P}_X(X \in \mathcal{B}_\epsilon) \leq 2C\epsilon^{1+\alpha}. \tag{A.1}$$

In order to control $T_{2,\epsilon}$, define:

$$\forall j \geq 1 \qquad \mathcal{B}_{\epsilon,j} := \left\{ x \in \mathbb{R}^d \quad | \quad 2^{j-1}\epsilon \leq |\eta(x) - 1/2| \leq 2^j\epsilon \right\}.$$

Now,

$$\begin{aligned} T_{2,\epsilon} &= 2\sum_{j\geq 1} \mathbb{E}\left[|\eta(X) - 1/2|\mathbf{1}_{\{\Phi_n(X)\neq\Phi^*(X)\}}\mathbf{1}_{\{X\in\mathcal{B}_{\epsilon,j}\}}\right] \\ &\leq 2\epsilon\sum_{j\geq 1} 2^j\mathbb{E}_X\left[\mathbf{1}_{\{X\in\mathcal{B}_{\epsilon,j}\}}\mathbb{E}_{\otimes^n}\left(\mathbf{1}_{\{\Phi_n(X)\neq\Phi^*(X)\}}\right)\right]. \end{aligned}$$

We can apply Proposition A.2 (see Section A.5 below) to obtain:

$$T_{2,\epsilon} \leq 4\epsilon\sum_{j\geq 1} 2^j\mathbb{E}_X\left[\mathbf{1}_{\{X\in\mathcal{B}_{\epsilon,j}\}}\exp\left(-2k_n\lfloor 2^{j-1}\epsilon - \Delta_n(X)\rfloor_+^2\right)\right]. \tag{A.2}$$

Since μ is lower bounded by $a > 0$ on Ω, we can apply Proposition A.3 with $a = \mu_-$ to obtain:

$$\Delta_n(X) \leq C\left(\left(\frac{k_n}{n}\mu_-^{-1}\right)^{1/d} + \exp\left(-3k_n/14\right)\right).$$

Now, we consider $\epsilon = \epsilon_n \geq 2\Delta_n(X)$, for example by choosing:

$$\epsilon_n := 2C\left(\left(\frac{k_n}{n}a^{-1}\right)^{1/d} + \exp\left(-3k_n/14\right)\right). \tag{A.3}$$

With ϵ_n defined as in (A.3), we deduce that $2^{j-1}\epsilon_n - \Delta_n(X) \geq 2^{j-1}\epsilon_n - \frac{\epsilon_n}{2} \geq \epsilon_n\left(2^{j-1} - \frac{1}{2}\right) > 0$. Thus, (A.2) becomes:

$$T_{2,\epsilon_n} \leq 4\epsilon_n\sum_{j\geq 1} 2^j\mathbb{E}_X\left[\mathbf{1}_{\{0<|\eta(X)-1/2|<2^j\epsilon_n\}}\exp\left(-2k_n\epsilon_n^2\left(2^{j-1} - 1/2\right)^2\right)\right].$$

Now, in order to control the previous bound, we choose k_n such that:

$$k_n = \epsilon_n^{-2}. \tag{A.4}$$

Thanks to (A.3), the constraint (A.4) then yields:

$$\epsilon_n \sim n^{\frac{-1}{2+d}} \quad \text{and} \quad k_n \sim n^{\frac{2}{2+d}}. \tag{A.5}$$

We then obtain that:

$$\begin{aligned} T_{2,\epsilon_n} &\leq 4\epsilon_n \sum_{j\geq 1} 2^j \mathbb{E}_X \left[\mathbf{1}_{\{0<|\eta(X)-1/2|<2^j\epsilon_n\}} \exp\left(-\frac{2^{2j}}{8}\right)\right], \\ &\leq \epsilon_n \sum_{j\geq 1} 2^{j+2} \exp\left(-\frac{2^{2j}}{8}\right) \mathbb{P}_X\left(|\eta(X)-1/2|<2^j\epsilon_n\right). \end{aligned}$$

The Margin Assumption applied to $\mathbb{P}_X\left(|\eta(X)-1/2|<2^j\epsilon_n\right)$ leads to:

$$T_{2,\epsilon_n} \leq \epsilon_n^{1+\alpha} \sum_{j\geq 1} 2^{j(1+\alpha)+2} \exp\left(-\frac{2^{2j}}{8}\right).$$

The series on the right hand side converges. This last bound associated with (A.1) leads to:

$$\sup_{F\in\mathcal{F}} \left[\mathcal{R}(\Phi_n) - \mathcal{R}(\Phi^*)\right] \leq C n^{-\frac{1+\alpha}{2+d}}.$$

A.3. Proof of the upper bounds: Theorem 4.3 and Theorem 4.5 ii).

Proof of Theorem 4.3. We consider a constant γ and use the following decomposition of $\mathbb{R}^d$ for a suitable $\gamma > 0$ (that will be chosen later on):

$$\mathbb{R}^d = \underbrace{\{x : 0 \leq \mu(x) \leq n^{-\gamma}\}}_{R_n} \cup \underbrace{\{x : \mu > n^{-\gamma}\}}_{Q_n}.$$

We follow the roadmap of the proof of Theorem 3.1 and keep the notation $\mathcal{B}_\epsilon$, which refers to $\mathcal{B}_\epsilon := \left\{x \in \mathbb{R}^d : |\eta(x)-1/2| \leq \epsilon\right\}$. Thanks to Proposition A.1, we obtain:

$$\begin{aligned} \mathcal{R}(\Phi_n) - \mathcal{R}(\Phi^*) &= \mathbb{E}\left[|2\eta(X)-1|\mathbf{1}_{\{\Phi_n(X)\neq\Phi^*(X)\}}\right] \\ &= \underbrace{\mathbb{E}\left[|2\eta(X)-1|\mathbf{1}_{\{\Phi_n(X)\neq\Phi^*(X)\}}\mathbf{1}_{X\in R_n}\right]}_{:=T_{R_n}} \\ &\quad + \underbrace{\mathbb{E}\left[|2\eta(X)-1|\mathbf{1}_{\{\Phi_n(X)\neq\Phi^*(X)\}}\mathbf{1}_{X\in Q_n}\right]}_{:=T_{Q_n}}. \end{aligned}$$

Study of R_n. The Tail Assumption **A4** in the particular case where $\psi = Id$ leads to:

$$T_{R_n} \leq \mathbb{P}_X\left(X \in R_n\right) = \mathbb{P}_X(\mu(X) \leq n^{-\gamma}) \lesssim n^{-\gamma}.$$

Study of Q_n. Following the proof of Theorem 3.1 with $a = n^{-\gamma}$, Equations (A.2)-(A.4) yield:

$$T_{Q_n} \leq C\epsilon_n^{1+\alpha}, \tag{A.6}$$

where ϵ_n and k_n satisfy the balance equations

$$\epsilon_n \sim 2C\left(\frac{k_n}{n}a^{-1}\right)^{1/d} = 2C\left(\frac{k_n}{n^{1-\gamma}}\right)^{1/d} \qquad \text{and} \qquad k_n = \epsilon_n^{-2}.$$

The equilibria are met in the two terms above with

$$k_n \sim Cn^{\frac{2(1-\gamma)}{2+d}}, \qquad \text{and} \qquad \epsilon_n \lesssim n^{-\frac{(1-\gamma)}{2+d}}. \tag{A.7}$$

Final control of the risk.. From the previous bounds, we obtain that:

$$\mathcal{R}(\Phi_n) - \mathcal{R}(\Phi^*) \lesssim n^{-\frac{(1-\gamma)(1+\alpha)}{2+d}} + n^{-\gamma}. \tag{A.8}$$

We optimize the last expression with respect to γ by setting

$$(1-\gamma)(1+\alpha) = \gamma(2+d) \Leftrightarrow \gamma = \frac{1+\alpha}{3+\alpha+d}.$$

The above choices allow us to conclude that:

$$\sup_{F\in\mathcal{F}}\left[\mathcal{R}(\Phi_n) - \mathcal{R}(\Phi^*)\right] \leq Cn^{-\frac{1+\alpha}{3+\alpha+d}}.$$

□

Proof of Theorem 4.5. ii) We follow the roadmap of the previous proof and replace the threshold $n^{-\gamma}$ with a_n, which should be carefully chosen. The key balance is still $k_n = \nu_n^{-2}$ on the set $\{\mu \geq a_n\}$ with the optimal setting:

$$\frac{k_n}{na_n} \lesssim \nu_n^d$$

Since we want to obtain a minimal value for ν_n, this last equation leads to the choice:

$$a_n = \frac{1}{n\nu_n^{2+d}}, \tag{A.9}$$

and the upper bound of the excess risk we obtained is then

$$\sup_{F\in\mathcal{F}}\left[\mathcal{R}(\Phi_n) - \mathcal{R}(\Phi^*)\right] \lesssim \nu_n^{1+\alpha} + \psi(a_n).$$

The natural equilibrium is found when plug-in (A.9) in this last upper bound and ν_n are be fixed so that $\psi^{-1}\left(\nu_n^{1+\alpha}\right) = n^{-1}\nu_n^{-(2+d)}$. We then obtain the rate $\nu_n^{1+\alpha}$ with the balance equation:

$$\nu_n^{2+d}\psi^{-1}(\nu_n^{1+\alpha}) \sim n^{-1}.$$

□

A.4. Proof of Theorem 4.4 (sliced nearest neighbor).

PROOF. We use the partition of Ω naturally derived from the slices $\Omega_{n,0}$ and $(\Omega_{n,j})_{j\geq 1}$:

$$\Omega_{n,0} := \left\{x|\mu(x) \geq n^{-\gamma}\right\} \quad \text{and} \quad \Omega_{n,j} := \left\{x|n^{-\gamma}2^{-(j+1)} \leq \mu(x) \leq n^{-\gamma}2^{-j}\right\}.$$

For this purpose, let $\gamma \in (0,1)$ (that will be specified later) and:

$$k_{n,j} = k_{n,0}2^{-2j/(2+d)} \quad \text{with} \quad k_{n,0} = n^{\frac{2(1-\gamma)}{2+d}}\log(n).$$

We then use the following decomposition of the excess risk:

$$\begin{aligned}
&\mathcal{R}(\Phi_n) - \mathcal{R}(\Phi^*) \\
&= \mathbb{E}\left[|2\eta(X)-1|\mathbf{1}_{\{\Phi_n(X)\neq\Phi^*(X)\}}\left(\mathbf{1}_{\Omega_{n,0}} + \sum_{j=1}^{+\infty}\mathbf{1}_{\Omega_{n,j}}\right)\right], \\
&= \mathbb{E}[|2\eta(X)-1|\mathbf{1}_{\{\Phi_n(X)\neq\Phi^*(X)\}}\mathbf{1}_{\Omega_{n,0}}] \\
&\quad + \sum_{j=1}^{+\infty}\mathbb{E}[|2\eta(X)-1|\mathbf{1}_{\{\Phi_n(X)\neq\Phi^*(X)\}}\mathbf{1}_{\Omega_{n,j}}, \\
&:= T_n + \sum_{j=1}^{+\infty} R_{n,j}.
\end{aligned}$$

Study of T_n. The density is lower bounded by $n^{-\gamma}$ on $\Omega_{n,0}$. The proof of Theorem 4.3 yields (see (A.6) and (A.7)):

$$T_n \leq n^{-(1+\alpha)\frac{1-\gamma}{2+d}}.$$

Study of $R_{n,j}$. For any $j > J_0(n) := (1-\gamma)\frac{\log(n)}{\log(2)}$ and for any $x \in \Omega_{n,j}$ with $j > J_0$, we have:

$$\mu(x) < n^{-\gamma}2^{-(1-\gamma)\frac{\log(n)}{\log(2)}} = 1/n.$$

The Tail Assumption with $\psi = Id$ leads to:

$$\sum_{j>J_0} R_{n,j} \leq \mathbb{P}_X\left[\mu(X) \leq n^{-1}\right] \lesssim n^{-1}.$$

Study of $R_{n,j}, j \leq J_0(n)$. We consider the intermediary slices, and for $1 \leq j \leq J_0(n)$:

$$R_{n,j} = \overbrace{\mathbb{E}\left[|2\eta(X)-1|\mathbf{1}_{\{\Phi_n(X)\neq\Phi^*(X)\}}\left[\mathbf{1}_{|\eta(X)-1/2|\leq\epsilon_{n,j}}\right]\mathbf{1}_{X\in\Omega_{n,j}}\right]}^{:=R_{n,j,1}} + \underbrace{\mathbb{E}\left[|2\eta(X)-1|\mathbf{1}_{\{\Phi_n(X)\neq\Phi^*(X)\}}\left[\mathbf{1}_{|\eta(X)-1/2|>\epsilon_{n,j}}\right]\mathbf{1}_{X\in\Omega_{n,j}}\right]}_{:=R_{n,j,2}}$$

where $\epsilon_{n,j}$ will be chosen later. To bound $R_{n,j,1}$, we use the fact that $|\eta - 1/2| \leq \epsilon_{n,j}$ as well as the Tail Assumption on the set $\Omega_{n,j} \subset \left\{\mu(X) \leq n^{-\gamma}2^{-j}\right\}$ to obtain:

$$R_{n,j,1} \leq 2\epsilon_{n,j} n^{-\gamma} 2^{-j}. \tag{A.10}$$

Thanks to Proposition A.2, we can bound the term $R_{n,j,2}$ as follows:

$$R_{n,j,2} \leq 4\mathbb{E}\left[\mathbf{1}_{X\in\Omega_{n,j}} \exp\left(-2k_{n,j}\lfloor\epsilon_{n,j} - \Delta_n(X)\rfloor_+^2\right)\right].$$

The term $\epsilon_{n,j}$ is then chosen such that $\epsilon_{n,j} - \Delta_n(X) \leq \epsilon_{n,j}/2$. According to Proposition A.3, we obtain:

$$\epsilon_{n,j} = c\left(k_{n,j}\frac{n^\gamma 2^{j+1}}{n}\right)^{1/d},$$

where c is chosen large enough. With this value, we obtain the following simplifications:

$$\epsilon_{n,j}^2 k_{n,j} = c^2 \frac{k_{n,j}^{1+\frac{2}{d}} 2^{\frac{j+1}{d}}}{n^{\frac{1-\gamma}{d}}} = c^2 \frac{k_{n,0}^{1+\frac{2}{d}}}{n^{\frac{1-\gamma}{d}}} 2^{-\frac{2j}{2+d}\left(1+\frac{2}{d}\right)} 2^{(2j+2)/d} = c^2 2^{2/d} \log(n)^{1+2/d}. \tag{A.11}$$

Taken together, (A.10) and (A.11) lead to:

$$R_{n,j} \leq 2\epsilon_{n,j} n^{-\gamma} 2^{-j} + 4\exp(-c^2 2^{2/d}\log(n)^{1+2/d})\mathbb{P}(X \in \Omega_{n,j}).$$

We then sum up all these terms for $j \geq 1$:

$$\begin{aligned} R_n &\lesssim \sum_{j=1}^{J_0} \frac{\log(n)^{1/2+1/d}}{\sqrt{k_{n,j}}} n^{-\gamma} 2^{-j} + \frac{1}{n} \sum_{j=1}^{J_0} \mathbb{P}(X \in \Omega_{n,j}), \\ &\lesssim n^{-\gamma} \log(n)^{1/2+1/d} \sum_{j=1}^{J_0} 2^{-j} k_{n,j}^{-1/2} + \frac{1}{n}, \\ &\lesssim n^{-\gamma} n^{-\frac{1-\gamma}{2+d}} \log(n)^{1/2+1/d} \sum_{j=1}^{J_0} 2^{-j+\frac{j+1}{2+d}} + \frac{1}{n}, \\ &\lesssim n^{-\gamma} n^{-\frac{1-\gamma}{2+d}} \log(n)^{1/2+1/d} + \frac{1}{n}. \end{aligned}$$

We can see in this last upper bound that we obtain an improvement between the standard rule and the one fixed here since the term $n^{-\gamma}$ that appears in the tail of μ on the right hand side of (A.8) is transformed into $n^{-\gamma} \times n^{-\frac{1-\gamma}{2+d}}$ up to a log term.

Final equilibrium. We now fix the optimal value of γ with the conjunction of the upper bounds for R_n and T_n:

$$\mathcal{R}(\Phi_n) - \mathcal{R}(\Phi^*) \lesssim n^{-(1+\alpha)\frac{1-\gamma}{2+d}} + n^{-\gamma-\frac{1-\gamma}{2+d}} \log(n)^{1/2+1/d} + O(1/n).$$

The balance equilibrium is reached with $(1+\alpha)\frac{1-\gamma}{2+d} = \gamma + \frac{1-\gamma}{2+d}$, meaning that $\gamma = \frac{1}{2+\alpha+d}$. This concludes the proof. □

A.5. Technical results. In the following, we use the result reported in [Győ78] that compares the excess risk of any classifier with the Bayes procedure.

PROPOSITION A.1 ([Győ78]). *For any classifier Ψ, we have:*

$$\mathcal{R}(\Psi) - \mathcal{R}(\Phi^*) = \mathbb{E}\left[|2\eta(X) - 1| \mathbf{1}_{\{\Psi(X) \neq \Phi^*(X)\}}\right].$$

The following lemma is concerned with the concentration of the plug-in estimator $\hat{\eta}_n$ (see (2.3)).

LEMMA A.1 (Concentration of $\hat{\eta}_n$). *In the classification model,*

$$\mathbb{P}_{\otimes^n}\left(|\hat{\eta}_n(X) - \mathbb{E}_{\otimes^n}\left(\eta_n(X)\right)| > s\right) \leq 2\exp\left(-2k_n s^2\right).$$

PROOF. Once again, we can observe that conditionally to the $(X_{(i)})_{1\leq i\leq n}$, the corresponding labels $(Y_{(i)}(X))_{1\leq i\leq n}$ are *independent* Bernoulli random variables with respective parameters $\eta(X_{(i)})$. We can now use the Hoeffding inequality as follows:

$$\begin{aligned}
&\mathbb{P}_{\otimes^n}\left(|\hat{\eta}_n - \mathbb{E}_{\otimes^n}\left(\eta_n(X)\right)| > s\right)\\
&\quad = \mathbb{E}_{\otimes^n}\left(\mathbb{P}_{\otimes^n}\left(|\eta_n(X) - \mathbb{E}_{\otimes^n}\left(\eta_n(X)\right)| > s|(X_1,\ldots,X_n)\right)\right).\\
&\quad \leq \mathbb{E}_{\otimes^n}\left(\mathbb{P}_{\otimes^n}\left(\left|\frac{1}{k_n}\left[\sum_{i=1}^{k_n} Y_{(i)}(X) - \eta(X_{(i)})\right]\right| > s|(X_1,\ldots,X_n)\right)\right).\\
&\quad \leq \mathbb{E}_{\otimes^n}\left(2\exp(-2k_n s^2)|(X_1,\ldots,X_n)\right) \leq 2\exp(-2k_n s^2).
\end{aligned}$$

□

We first state an important upper bound of the error rate when the design point X is fixed.

PROPOSITION A.2. *For any $\epsilon > 0$ and any $X \in \Omega$, if $\Delta_n(X) := |\mathbb{E}_{\otimes^n}\hat{\eta}_n(X) - \eta(X)|$, we have:*

$$\mathbf{1}_{\{|\eta(X)-\frac{1}{2}|\geq\epsilon\}}\,\mathbb{E}_{\otimes^n}\left[\mathbf{1}_{\{\Phi_n(X)\neq\Phi^*(X)\}}\right] \leq 2\mathbf{1}_{\{|\eta(X)-\frac{1}{2}|\geq\epsilon\}}\,e^{-2k_n\lfloor\epsilon-\Delta_n(X)\rfloor_+^2},$$

where $\lfloor a\rfloor_+$ refers to the positive part of any real number a.

PROOF. In the event $\{\Phi_n(X) \neq \Phi^*(X)\}$, $\hat{\eta}_n(X) - 1/2$ and $\eta(X) - 1/2$ do not have the same sign. Therefore:

$$\mathbf{1}_{\{|\eta(X)-\frac{1}{2}|\geq\epsilon\}}\mathbf{1}_{\{\Phi_n(X)\neq\Phi^*(X)\}} \leq \mathbf{1}_{\{|\eta(X)-1/2|\geq\epsilon\}}\mathbf{1}_{\{|\hat{\eta}_n(X)-\eta(X)|\geq\epsilon\}}.$$

Then:

$$\begin{aligned}
&\mathbb{E}_{\otimes^n}\left[\mathbf{1}_{\{\Phi_n(X)\neq\Phi^*(X)\}}\mathbf{1}_{\{|\eta(X)-\frac{1}{2}|\geq\epsilon\}}\right]\\
&\quad \leq \mathbb{E}_{\otimes^n}\left[\mathbf{1}_{\{|\eta(X)-1/2|\geq\epsilon\}}\mathbf{1}_{\{|\hat{\eta}_n(X)-\eta(X)|\geq\epsilon\}}\right]\\
&\quad = \mathbf{1}_{\{|\eta(X)-\frac{1}{2}|\geq\epsilon\}}\,\mathbb{P}_{\otimes^n}\left(|\hat{\eta}_n(X) - \eta(X)| \geq \epsilon\right)\\
&\quad \leq \mathbf{1}_{\{|\eta(X)-\frac{1}{2}|\geq\epsilon\}}\,\mathbb{P}_{\otimes^n}\left(|\hat{\eta}_n(X) - \mathbb{E}_{\otimes^n}\left(\hat{\eta}_n(X)\right)| \geq \epsilon - \Delta_n(X)\right)
\end{aligned}$$

where the last line follows from the triangular inequality and the definition of $\Delta_n(X)$. Lemma A.1 applied with $s = \lfloor\epsilon - \Delta_n(X)\rfloor_+$ now leads to the conclusion of the proof. □

This last proposition clearly underlines the effect of the bias term $\Delta_n(X)$ in the misclassification error rate. This bias term is then upper bounded by the next result.

PROPOSITION A.3. *Assume that μ belongs to $\mathfrak{M}_{mma}(\Omega, \kappa)$ and η belongs to $\mathcal{C}^{1,0}(\Omega, L)$. Then, a constant $C > 0$ exists such that for any $a > 0$:*

$$|\mathbb{E}_{\otimes^n}\hat{\eta}_n(x) - \eta(x)| \leq L\left(\frac{2}{\kappa}\right)^{1/d}\left(\frac{k_n}{na}\right)^{1/d} + 2\exp\left(-\frac{3k_n}{14}\right),$$

for all $x \in K_a$ where

$$K_a := \left\{x \in \mathbb{R}^d \,|\, \mu(x) \geq a\right\}.$$

PROOF. We first propose a control of the bias and then use a concentration inequality in order to obtain the bound.

Decomposition of the bias. Let $x \in K_a$ be fixed. According to the definition of $\hat{\eta}_n(x)$ (see (2.3)):

$$\mathbb{E}_{\otimes^n}[\hat{\eta}_n(x)] = \mathbb{E}_{\otimes^n}\left[\frac{1}{k_n}\sum_{j=1}^{k_n} Y_{(j)}(x)\right] = \mathbb{E}_{\otimes^n}\left[\frac{1}{k_n}\sum_{j=1}^{k_n}\eta(X_{(j)})\right].$$

Hence:

$$\Delta_n(x) = |\mathbb{E}_{\otimes^n}[\hat{\eta}_n(x)] - \eta(x)| = \left|\mathbb{E}_{\otimes^n}\left(\frac{1}{k_n}\sum_{i=1}^{k_n}\eta(X_{(i)}) - \eta(X)\right)\right|.$$

For any $t \geq 0$, we write:

$$\Delta_n(x) = \left|\mathbb{E}_{\otimes^n}\left(\frac{1}{k_n}\sum_{i=1}^{k_n}\eta(X_{(i)}) - \eta(X)\right)\left(\mathbf{1}_{\|X_{k_n}-X\|<t} + \mathbf{1}_{\|X_{k_n}-X\|\geq t}\right)\right|.$$

Using the fact that η and η_n belong to $[0, 1]$, we then have:

$$\Delta_n(x) \leq \left|\mathbb{E}_{\otimes^n}\left(\frac{\mathbf{1}_{\|X_{(k_n)}-x\|<t}}{k_n}\sum_{i=1}^{k_n}(\eta(X_{(i)}) - \eta(x))\right)\right| + \mathbb{P}_{\otimes^n}\left[\|X_{(k_n)} - x\| \geq t\right].$$

Now, since $\eta \in \mathcal{C}^{1,0}(\Omega, L)$:

$$\begin{aligned}
\Delta_n(n) &\leq \frac{1}{k_n}\sum_{i=1}^{k_n}\mathbb{E}_{\otimes^n}\left[\mathbf{1}_{\|X_{(k_n)}-x\|<t}\left|\eta(X_{(i)})-\eta(x)\right|\right] \\
&\quad +\mathbb{P}_{\otimes^n}\left[\|X_{(k_n)}-x\|\geq t\right], \\
&\leq \frac{L}{k_n}\mathbb{E}_{\otimes^n}\left[\mathbf{1}_{\|X_{(k_n)}-x\|<t}\sum_{i=1}^{k_n}\left\|X_{(i)}-x\right\|\right]+\mathbb{P}_{\otimes^n}\left(\|X_{(k_n)}-x\|\geq t\right), \\
&\leq tL+\mathbb{P}_{\otimes^n}\left(\|X_{(k_n)}-x\|\geq t\right). \qquad \text{(A.12)}
\end{aligned}$$

Concentration inequality. We now turn our attention to the control of the last term in the r.h.s. of (A.12). In the following, for all $x \in \Omega$ and for all $t > 0$, $\mu(B(x,t))$ will denote the mass of the ball $B(x,t)$ w.r.t the measure $\mathbb{P}_X$, i.e.

$$\mu(B(x,t)) := \int_{B(x,t)}\mu(z)dz.$$

Within this context, we sometimes omit the dependency of this quantity w.r.t. the point x and write $\mu_t = \mu(B(x,t))$. Then:

$$\begin{aligned}
&\mathbb{P}_{\otimes^n}\left(\left\|X_{(k_n)}-x\right\|\geq t\right) \\
&\quad = \mathbb{P}_{\otimes^n}\left(\sum_{i=1}^{n}\mathbf{1}_{\{X_i\in B(x,t)\}}\leq k_n\right), \\
&\quad = \mathbb{P}_{\otimes^n}\left(\frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{1}_{\{X_i\in B(x,t)\}}-\mu\left(B(x,t)\right)\right]\leq \frac{k_n}{n}-\mu\left(B(x,t)\right)\right), \\
&\quad \leq \mathbb{P}_{\otimes^n}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{1}_{\{X_i\in B(x,t)\}}-\mu\left(B(x,t)\right)\right]\right|\geq \mu(B(x,t))-\frac{k_n}{n}\right),
\end{aligned}$$

as soon as $\mu(B(x,t)) > \frac{k_n}{n}$. Since $\mathbb{P}_X \in \mathfrak{M}_{mma}(\Omega,\kappa)$ and $x \in K_a$,

$$\mu(B(x,t)) \geq \kappa\mu(x)t^d \geq \kappa a t^d.$$

We therefore choose t such that:

$$\kappa a t^d \geq 2\frac{k_n}{n} \Leftrightarrow t \geq \left(2\frac{k_n}{n}\frac{1}{\kappa a}\right)^{1/d}. \qquad \text{(A.13)}$$

In particular, with the choice of t given by (A.13), we obtain:

$$\mu(B(x,t))-\frac{k_n}{n} \geq \frac{\mu(B(x,t))}{2}. \qquad \text{(A.14)}$$

We then use the following version of the Bennett inequality. If W_i are random variables such that $W_i \leq b$, $v = \sum_{i=1}^n \mathbb{E}(W_i^2)$, let $S = \sum_{i=1}^n W_i - \mathbb{E}(W_i)$ then for any $x > 0$:

$$\mathbb{P}\left(S \geq x\right) \leq \exp\left(-\frac{x^2}{2(v + bx/3)}\right).$$

We apply this version to the random variables $Z_i = \frac{\mathbf{1}_{\{X_i \in B(x,t)\}}}{\sqrt{\mu_t(1-\mu_t)}}$. We then have $b = \frac{1}{\sqrt{\mu_t(1-\mu_t)}}$, $v = \frac{n}{1-\mu_t}$ and $x = \frac{n\sqrt{\mu_t}}{2\sqrt{(1-\mu_t)}}$. The exponential bound obtained is then

$$\exp\left(-\frac{\frac{n\mu_t}{4(1-\mu_t)}}{2(\frac{1}{1-\mu_t} + \frac{1}{\sqrt{\mu_t(1-\mu_t)}}\frac{\sqrt{\mu_t}}{2\sqrt{(1-\mu_t)}}/3)}\right) = \exp\left(-\frac{3n\mu_t}{28}\right).$$

Hence:

$$\begin{aligned}
&\mathbb{P}_{\otimes^n}\left(\left\|X_{(k_n)} - x\right\| \geq t\right) \\
&\quad\leq \mathbb{P}_{\otimes^n}\left(\left|\frac{1}{n}\sum_{i=1}^n \left[\mathbf{1}_{\{X_i \in B(x,t)\}} - \mu\left(B(x,t)\right)\right]\right| \geq \frac{\mu(B(x,t))}{2}\right), \\
&\quad\leq 2\exp\left(-\frac{3n\mu_t}{28}\right).
\end{aligned}$$

Now, using (A.14) $\mu_t \geq 2k_n/n$, we obtain:

$$\mathbb{P}_{\otimes^n}\left(\left\|X_{(k_n)} - x\right\| \geq t\right) \leq 2\exp\left(-\frac{3k_n\mu_t}{14}\right). \tag{A.15}$$

Final Bound. According to (A.12) and (A.15), we obtain:

$$\begin{aligned}
\Delta_n(X) &\leq Lt + 2\exp\left(-\frac{3k_n}{14}\right), \\
&\leq L\left(\frac{2}{\kappa}\right)^{1/d}\left(\frac{k_n}{na}\right)^{1/d} + 2\exp\left(-\frac{3k_n}{14}\right).
\end{aligned}$$

This concludes the proof of Proposition A.3. □

APPENDIX B: THE SMOOTH DISCRIMINANT ANALYSIS MODEL

B.1. Statistical setting. In this subsection, we focus our attention on an alternative binary classification model. We assume that we have two independent samples $\mathcal{S}_1 = (X_1, \ldots, X_n)$ and $\mathcal{S}_2 = (\tilde{X}_1, \ldots, \tilde{X}_n)$ of i.i.d. random variables at our disposal, with respective densities f and g. We assume that the support of f and g are included in a set K. In the following, we can label each element of the sample $\mathcal{S}_1$ (resp. $\mathcal{S}_2$) by 0 (resp. 1).
In this context, given a new incoming observation, the goal is to predict its corresponding label, namely to determine whether $X \sim f$ or $X \sim g$. This setting is known as a smooth discriminant analysis model and has been popularized, in particular, by Mammen and Tsybakov (1999).

As in the classical binary classification model, a classifier is defined as a measurable function of the samples $\mathcal{S}_1$ and $\mathcal{S}_2$, having values in $\{0, 1\}$. The risk of each classifier Φ_n is defined as:

$$\mathcal{R}(\Phi_n) := \frac{1}{2}\left[\int_{\{x:\Phi_n(x)=1\}} f(x)dx + \int_{\{x:\Phi_n(x)=0\}} g(x)dx\right].$$

It can then be proved that the Bayes Classifier defined as:

$$\Phi^\star(x) = \mathbf{1}_{\{f(x)\geq g(x)\}},$$

provides the smallest possible risk w.r.t. all possible classifiers. At this step, if we rewrite the Bayes classifier as:

$$\Phi^\star(x) = \mathbf{1}_{\{\eta(x)\geq \frac{1}{2}\}}, \text{ where } \eta(x) = \frac{f(x)}{f(x)+g(x)},$$

A strong analogy with the classical binary classification problem defined in Section 2 (see in particular (2.1)) can be observed. The next assumption is equivalent to Assumption $A1$.

Assumption $\widetilde{\mathbf{A1}}$. *An L exists such that $\eta = \frac{f}{f+g}$ belongs to $\mathcal{C}^{1,0}(\Omega, L)$.*
In keeping with the previous study, we work with both a Margin and a Smoothness Assumption on the regression function η. Once again, the Margin Assumption is quite close to the one introduced for the classical binary classification problem.

Assumption $\widetilde{\mathbf{A2}}$. *An $\alpha > 0$ and a constant $C > 0$ exist such that*

$$\int_{|\eta-\frac{1}{2}|<\epsilon} (f+g) \leq C\epsilon^\alpha$$

.
We also consider the case where the marginal density on X satisfies a Minimal Mass Assumption.

Assumption $\widetilde{\mathbf{A3}}$. *A* κ *exists such that* $\mu = \frac{f+g}{2}$ *satisfies the condition introduced by* $\mathfrak{M}_{mma}(\Omega, \kappa)$.

In this context, the principle of the Nearest Neighbor Algorithm introduced in Section 2.2 remains the same. We first aggregate the two samples $\mathcal{S}_1$ and $\mathcal{S}_2$ to obtain $\mathcal{S} = \left\{(X_1, 0), \dots (X_n, 0), (\tilde{X}_1, 1), \dots, (\tilde{X}_n, 1)\right\} = (\mathcal{X}_i, \mathcal{Y}_i)_{1 \le i \le 2n}$. Then, if $\mathcal{X}_{(m)}(x)$ is the m-nearest neighbor of x and $\mathcal{Y}_{(m)}(x)$ is its label, the K nearest neighbor decision rule is

$$\Phi_{n,k}(x) = \begin{cases} 1 & if \quad \frac{1}{k}\sum_{j=1}^{k} \mathcal{Y}_{(j)}(x) > \frac{1}{2}, \\ 0 & otherwise. \end{cases} \tag{B.1}$$

B.2. Case of bounded from below densities. The following theorem provides a control on its corresponding minimax excess risk when the underlying density on X is bounded from below by a strictly positive constant.

THEOREM B.1. *Assume that Assumptions* $\tilde{\mathbf{A1}} - \tilde{\mathbf{A3}}$ *hold and that the density of* X *is lower bounded by* $\mu_- > 0$. *If* $k_n = \lfloor n^{\frac{2}{2+d}} \rfloor$ *, then*

$$[\mathcal{R}(\Phi_{n,k_n}) - \mathcal{R}(\Phi^*)] \lesssim \left(\frac{\log n}{n}\right)^{\frac{1+\alpha}{2+d}}.$$

We can immediately observe that we lose a log term compared to the minimax rate obtained in Section 2.3 for the classification model.
According to our knowledge, the optimal rates in the present model (that is, the smooth discriminant analysis) have never been investigated under Assumptions $\tilde{\mathbf{A1}} - \tilde{\mathbf{A3}}$. Nevertheless, it appears that under some slightly different assumptions (see [Tsy04] and [AT07]), the minimax risk of the smooth discriminant analysis problem is the same as the one of the classification model. It is therefore reasonable to think that our rate is near-optimal (optimal up to a log term).
From a technical point of view, the main counterpart when using the smooth discriminant analysis setting is that conditionally to the spatial positions of the ordered aggregate sample $(X_{(j)}(x))_{1 \le j \le 2n}$, the corresponding labels

$(Y_{(j)}(x))_{1\leq j\leq 2n}$ are no longer independent of each other. This is a significant difference with the classification model considered in Section 2 and makes it quite difficult to obtain a concentration inequality for the empirical "regression" function:

$$\eta_n(x) = \frac{1}{k_n}\sum_{j=1}^{k_n} Y_{(j)}(x).$$

In order to get around this problem, we adopt a Poisson embedding (see Subsection B.4 for further details), which makes it possible to satisfy a control of the excess risk, up to some additional logarithmic terms. We assume that such a logarithmic term may be removed using some concentration inequalities on negatively associated random variables. In fact, we hypothesize that for asymptotically large n, the random sequence $(\sum_{j=1}^{p} Y_{(j)}(x))_{1\leq p\leq k_n}$ is negatively associated as soon as $k_n/n \longmapsto 0$, which may make it possible to remove the logarithmic term (see *e.g.* [Kle03],[Sha00]).

B.3. Case of general densities. We provide a short paragraph here about the case of general densities and introduce the Tail Assumption necessary to derive a uniform rate of consistency for the nearest neighbor rule.

THEOREM B.2. *Assume that Assumptions* $\widetilde{\mathbf{A1}} - \widetilde{\mathbf{A3}}$ *hold and that a function* ψ *exists such that* $\mathbb{P}_X \in \mathcal{P}_{\mathcal{T},\psi}$. *Then:*

$$[\mathcal{R}(\Phi_n) - \mathcal{R}(\Phi^*)] \lesssim \left(\frac{\log n}{n}\right)^{\frac{1+\alpha}{3+\alpha+d}} \qquad \text{if} \qquad \psi = Id.$$

Otherwise:

$$[\mathcal{R}(\Phi_n) - \mathcal{R}(\Phi^*)] \lesssim \nu_{n,\alpha,d}^{1+\alpha} \quad \text{where} \quad \frac{\log n}{n} = \psi^{-1}(\{\nu_{n,\alpha,d}\}^{1+\alpha})\{\nu_{n,\alpha,d}\}^{2+d}.$$

We only provide the proof of Theorem B.1: the one of Theorem B.2 relies on a mixed association of some arguments in Theorem 4.3 and Theorem B.1.

B.4. Proof of Theorem B.1. A straightforward decomposition yields the following proposition.

PROPOSITION B.1. *For any classifier* Ψ, *we have:*

$$\mathcal{R}(\Psi) - \mathcal{R}(\Phi^*) = \int_{\Psi\neq\Phi^*} |2\eta(x) - 1|\,\frac{f(x)+g(x)}{2}dx.$$

PROOF OF THEOREM B.1. The beginning of the proof is similar to the one of Theorem 3.1. We use Proposition B.1 and Assumption $\widetilde{\mathbf{A3}}$ to obtain:

$$\begin{aligned}\mathcal{R}(\Phi_n) - \mathcal{R}(\Phi^*) &\leq 2Cc_2\epsilon^{1+\alpha} \\ &+c_2\epsilon \sum_{j\geq 1} 2^j \int_{K\cap\{0<|\eta(x)-1/2|<2^j\epsilon\}} \mathbb{P}\left(|\eta_n(x)-\eta(x)| > 2^{j-1}\epsilon\right) dx. \qquad \text{(B.2)}\end{aligned}$$

At this step, the major difference between the discriminant analysis and the classification model is the control of the deviation inequality on the right hand side of the equation above. Indeed, conditionally to $((\mathcal{X}_{(i)}(x))_{1\leq i\leq 2n}$, the corresponding labels $((\mathcal{Y}_{(i)}(x))_{1\leq i\leq 2n}$ are no longer independent.
We use Proposition B.2 with:

$$\epsilon_n = 2C\left(\frac{k_n}{n}\right)^{1/d}, \qquad \text{(B.3)}$$

where C is the constant that appears in Proposition B.2 *ii*). We then obtain:

$$\begin{aligned}&\mathbb{P}\left(|\eta_n(x)-\eta(x)| > 2^{j-1}\epsilon_n\right) \\ &\leq 2\pi n\left[\exp\left(-2k_n\left((2^{j-1}-1/2)\epsilon_n\right)^2\right) + \mathbf{1}_{\left\{2^j\left(\frac{k_n}{n}\right)^{1/d}\leq C^{-1}\right\}} e^{-n}\right].\end{aligned}$$

The suitable choice of k_n and ϵ_n then becomes:

$$k_n\epsilon_n^2 \sim a\log(n) \qquad \text{(B.4)}$$

for a sufficiently large universal constant $a > 0$. Hence, (B.3) and (B.4) yields:

$$\epsilon_n \sim \left(\frac{\log(n)}{n}\right)^{\frac{1}{1+d}} \qquad \text{and} \qquad k_n \sim \log(n)^{d/(2+d)} n^{2/(2+d)}.$$

This last choice implies:

$$\mathbb{P}\left(|\eta_n(x)-\eta(x)| > 2^{j-1}\epsilon_n\right) \lesssim \exp(-2a2^{2j}) + 2\pi n e^{-n}\mathbf{1}_{\left\{2^j\leq C^{-1}\left(\frac{n}{k_n}\right)^{1/d}\right\}} \qquad \text{(B.5)}$$

Plug in (B.5) in (B.2) allows us to conclude:

$$\begin{aligned}&\mathcal{R}(\Phi_n) - \mathcal{R}(\Phi^*) \\ &\leq 2Cc_2\epsilon_n^{1+\alpha} + c_2\epsilon_n^{1+\alpha}\sum_{j\geq 1} 2^{j(1+\alpha)+1}\exp(-\tilde{C}2^{2j}) \\ &\quad +\epsilon_n^{1+\alpha}2\pi n e^{-n} \sum_{j:2^j\leq C^{-1}\left(\frac{n}{k_n}\right)^{1/d}} 2^{j(1+\alpha)} \\ &\lesssim \epsilon_n^{1+\alpha} + 2\pi n e^{-n}\left(\frac{n}{k_n}\right)^{(1+\alpha)/d},\end{aligned}$$

where the last inequality follows from standard upper bounds on geometrical sums. We then deduce:

$$\sup_{F\in\mathcal{F}}\left[\mathcal{R}(\Phi_n)-\mathcal{R}(\Phi^*)\right]\leq C\left(\frac{\log(n)}{n}\right)^{\frac{1+\alpha}{2+d}}.$$

□

PROPOSITION B.2 (Concentration of η_n with Poisson approximation). *In the smooth discriminant analysis model, assume that $k_n < n$. We then have:*

i) For any $t \geq 0$:

$$\mathbb{P}\left(|\eta_n(x)-\mathbb{E}\eta_n(x)|>t\right)\leq 2\pi n\left[2\exp(-2k_nt^2)+\mathbf{1}_{\{|t|\leq 1\}}e^{-n}\right].$$

ii) For any $t \geq 0$ and if $k_n >> \log(n)$, there exists a constant C such that:

$$\mathbb{P}\left(|\eta_n(x)-\eta(x)|>t\right)$$
$$\leq 2\pi n\left[2\exp\left(-2k_n\left[t-C\left(\frac{k_n}{n}\right)^{1/d}\right]^2\right)+\mathbf{1}_{\{|t|\leq 1\}}e^{-n}\right].$$

PROOF. Poissonization:
In order to eliminate the dependency between the ordered statistics $(\mathcal{X}_{(i)})_{1\leq i\leq 2n}$, we use an idea introduced by [Kac49] and randomize the size of the sample by using some Poisson random variables.
First consider (N_1, N_2) two independent random variables following a Poisson distribution $\mathcal{P}(n)$ as well as a $N \sim \mathcal{P}(2n)$ independent of (N_1, N_2). We now build an artificial sample of size $N_1 + N_2$ divided into two parts:

$$\mathcal{S}_{0-1}^{\mathcal{P}}=(X_i,Y_i)_{1\leq i\leq N_1+N_2}\sim(gd\lambda\otimes\delta_0)^{\otimes N_1}\bigotimes(fd\lambda\otimes\delta_1)^{\otimes N_2}.$$

Then, if σ denotes a random permutation picked uniformly in $\mathfrak{S}_{N_1+N_2}$ and independent of the previous realizations, the permutated sample:

$$\mathcal{S}_{0-1}^{\mathcal{P},\sigma}=(X_{\sigma(i)},Y_{\sigma(i)})_{1\leq i\leq N_1+N_2},$$

follows the same distribution as the sample of i.i.d. realizations:

$$\mathcal{S}_{\eta}^{\mathcal{P}}=(U_i,V_i)_{1\leq i\leq N},$$

where $U\sim\frac{f+g}{2}d\lambda$ and $V|U\sim\mathcal{B}(\eta(U))$ with $\eta(U)=\frac{f(U)}{f(U)+g(U)}$.
Recall that η_n is built from our original sample $\mathcal{S}$ according to:

$$\eta_n(x)=\frac{1}{k_n}\sum_{j=1}^{k_n}\mathcal{Y}_{(j)}(x).$$

Our aim is to study the deviation of η_n from its mean. For this purpose, we introduce its Poisson counterpart built from $\mathcal{S}_{0-1}^{\mathcal{P},\sigma}$:

$$\eta_n^{N_1,N_2}(x) = \frac{1}{k_n} \sum_{j=1}^{k_n \wedge (N_1+N_2)} Y_{(j)}(x),$$

which is independent of the random choice of σ. At last, we define η_n^N by

$$\eta_n^N = \frac{1}{k_n} \sum_{j=1}^{k_n \wedge N} V_{(j)}(x).$$

where we have used the convention $\sum_{j\geq 1}^{0} a_j = 0$ for any sequence $(a_j)_{j\geq 1}$.
Proof of *i)*. Since $\mathcal{L}\left(\eta_n(x)\right) = \mathcal{L}\left(\eta_n(x)^{N_1,N_2} | N_1 = N_2 = n\right)$, we deduce that

$$\begin{aligned}
\mathbb{P}\left(|\eta_n(x) - \mathbb{E}\eta_n(x)| > t\right) &= \mathbb{P}\left(\left|\eta_n^{N_1,N_2}(x) - \mathbb{E}\eta_n^{N_1,N_2}(x)\right| > t | N_1 = N_2 = n\right) \\
&= \frac{\mathbb{P}\left(\left|\eta_n^{N_1,N_2}(x) - \mathbb{E}\eta_n^{N_1,N_2}(x)\right| > t, N_1 = N_2 = n\right)}{\mathbb{P}(N_1 = N_2 = n)} \\
&\leq \frac{\mathbb{P}\left(\left|\eta_n^{N_1,N_2}(x) - \mathbb{E}\eta_n^{N_1,N_2}(x)\right| > t\right)}{\mathbb{P}(N_1 = N_2 = n)}. \qquad \text{(B.6)}
\end{aligned}$$

Again, we can observe that $\mathcal{L}\left(\eta_n^{N_1,N_2}(x)\right) = \mathcal{L}\left(\eta_n^N(x)\right)$ and

$$\begin{aligned}
\mathbb{P}\left(\left|\eta_n^{N_1,N_2}(x) - \mathbb{E}\eta_n^{N_1,N_2}(x)\right| > t\right) &= \mathbb{P}\left(\left|\eta_n^N(x) - \mathbb{E}\eta_n^N(x)\right| > t\right) \\
&= \mathbb{P}\left(\left|\eta_n^N(x) - \mathbb{E}\eta_n^N(x)\right| > t, N > k_n\right) \\
&\quad + \mathbb{P}\left(\left|\eta_n^N(x) - \mathbb{E}\eta_n^N(x)\right| > t, N < k_n\right).
\end{aligned}$$

We now study the two terms of the upper bound separately.

Upper bound of $\mathbb{P}\left(\left|\eta_n^N(x) - \mathbb{E}\eta_n^N(x)\right| > t, N > k_n\right)$. We can use the standard Hoeffding inequality:

$$\mathbb{P}\left(\left|\eta_n^N(x) - \mathbb{E}\eta_n^N(x)\right| > t | N > k_n\right) \leq 2\exp(-2k_n t^2),$$

and trivially bounding $\mathbb{P}(N > k_n)$ by 1 yields:

$$\mathbb{P}\left(\left|\eta_n^N(x) - \mathbb{E}\eta_n^N(x)\right| > t, N > k_n\right) \leq 2\exp(-2k_n t^2). \qquad \text{(B.7)}$$

Upper bound of $\mathbb{P}\left(\left|\eta_n^N(x) - \mathbb{E}\eta_n^N(x)\right| > t, N < k_n\right)$. From the Chernoff bound of the left Poisson tail, we obtain:

$$\mathbb{P}(N < k_n) \leq \frac{e^{-2n}(e2n)^{k_n}}{k_n^{k_n}} = e^{-2n\left(1-\frac{k_n}{2n}\log\left(\frac{ek_n}{n}\right)\right)} \leq e^{-n},$$

as soon as $k_n < n$. Moreover, we have:

$$\begin{aligned}
&\mathbb{P}\left(\left|\eta_n^N(x) - \mathbb{E}\eta_n^N(x)\right| > t | N < k_n\right) \\
&= \mathbb{P}\left(\frac{N}{k_n}\frac{\sum_{j=1}^N \left[V_{(j)}(x) - \eta(U_{(j)}(x)\right]}{N} > t | k_n > N\right) \\
&\leq \mathbf{1}_{\{|t|\leq 1\}}.
\end{aligned}$$

It follows that:

$$\mathbb{P}\left(\left|\eta_n^N(x) - \mathbb{E}\eta_n^N(x)\right| > t, N < k_n\right) \leq \mathbf{1}_{\{|t|\leq 1\}} e^{-n}. \tag{B.8}$$

We then deduce from (B.6), (B.7) and (B.8) that:

$$\mathbb{P}\left(|\eta_n(x) - \mathbb{E}\eta_n(x)| > t\right) \leq \frac{2\exp(-2k_n t^2) + \mathbf{1}_{\{|t|\leq 1\}} e^{-n}}{\mathbb{P}(N_1 = N_2 = n)}.$$

The Stirling formula concludes the proof of i):

$$\mathbb{P}\left(|\eta_n(x) - \mathbb{E}\eta_n(x)| > t\right) \leq 2\pi n \left(2\exp(-2k_n t^2) + \mathbf{1}_{\{|t|\leq 1\}} e^{-n}\right)$$

Proof of ii). We now build the decomposition already used in the proof of Theorem 3.1.

$$\begin{aligned}
&\mathbb{P}\left(|\eta_n(x) - \eta(x)| > t\right) \\
&= \mathbb{P}\left(\left|\eta_n^{N_1,N_2}(x) - \eta(x)\right| > t | N_1 = N_2 = n\right) \\
&\leq \frac{\mathbb{P}\left(\left|\eta_n^{N_1,N_2}(x) - \eta(x)\right| > t\right)}{\mathbb{P}(N_1 = N_2 = n)} \\
&\leq \frac{\mathbb{P}\left(\left|\eta_n^{N_1,N_2}(x) - \eta(x)\right| > t, N_1 + N_2 > k_n\right)}{\mathbb{P}(N_1 = N_2 = n)} \\
&\quad + \frac{\mathbb{P}\left(\left|\eta_n^{N_1,N_2}(x) - \eta(x)\right| > t, N_1 + N_2 \leq k_n\right)}{\mathbb{P}(N_1 = N_2 = n)} \\
&\leq \frac{\mathbb{P}\left(\left|\eta_n^N(x) - \mathbb{E}\eta_n^N(x)\right| > t - \left|\mathbb{E}\eta_n^N(x) - \eta(x)\right| | N > k_n\right)\mathbb{P}(N > k_n)}{\mathbb{P}(N_1 = N_2 = n)} \\
&\quad + \frac{\mathbf{1}_{\{|t|\leq 1\}} e^{-n}}{\mathbb{P}(N_1 = N_2 = n)}.
\end{aligned}$$

We can slightly modify Proposition A.3 to obtain that for any $m > k_n$:

$$\left|\mathbb{E}\left[\eta_n^N(x) - \eta(x) | N = m\right]\right| \leq C\left(\left(\frac{k_n}{m}\right)^{1/d} + e^{-3k_n/14}\right).$$

Now if we set $m_n = 2n - (2n)^\beta$ for $\beta < 1$, we can write:

$$\begin{aligned}
&\left|\mathbb{E}\left[\eta_n^N(x) - \eta(x)\right]\right| \\
&\quad\leq \mathbb{P}(N < m_n) + \left|\sum_{m=m_n}^{+\infty} \mathbb{E}\left[\eta_n^N(x) - \eta(x) | N = m\right]\mathbb{P}(N = m)\right| \\
&\quad\leq \mathbb{P}(N < m_n) + C\mathbb{P}(N > m_n)\left(\left(\frac{k_n}{m_n}\right)^{1/d} + e^{-3k_n/14}\right) \\
&\quad\leq \mathbb{P}(N < m_n) + C\left(\left(\frac{k_n}{n}\right)^{1/d} + e^{-3k_n/14}\right).
\end{aligned}$$

where we have used $m_n > n$ for n large enough. Again, the Chernoff bound on the Poisson tail yields

$$\mathbb{P}(N < m_n) \leq e^{-n^{2\beta-1}}.$$

Any choice of $\beta \in (\frac{1}{2}, 1)$ implies

$$\left|\mathbb{E}\left[\eta_n^N(x) - \eta(x)\right]\right| \lesssim \left(\frac{k_n}{n}\right)^{1/d}.$$

The end of the proof is now straightforward using the bounds already given in the proof of i). □

Toulouse School of Economics
Université Toulouse 1 - Capitole.
21 allée de Brienne, 31000 Toulouse, France.
E-mail: sebastien.gadat@math.univ-toulouse.fr

Institut Mathématiques de Toulouse
Université Toulouse 3 - Paul Sabatier
118 route de Narbonne, 31400 Toulouse, France.
E-mail: thierry.klein@math.univ-toulouse.fr

Institut Mathématiques de Toulouse
Institut National des Sciences Appliquées
135, avenue de Rangueil 31 077 Toulouse Cedex 4, France.
E-mail: clement.marteau@math.univ-toulouse.fr