On the Kernel Rule for Function Classification

Christophe ABRAHAM [a], Gérard BIAU [b,*] and Benoît CADRE [b]

[a] ENSAM-INRA, UMR Biométrie et Analyse des Systèmes,
2 place Pierre Viala, 34060 Montpellier Cedex 1, France;
[b] Institut de Mathématiques et de Modélisation de Montpellier,
UMR CNRS 5149, Equipe de Probabilités et Statistique,
Université Montpellier II, CC 051,
Place Eugène Bataillon, 34095 Montpellier Cedex 5, France

Abstract

Let X be a random variable taking values in a function space $\mathcal{F}$, and let Y be a discrete random label with values 0 and 1. We investigate asymptotic properties of the moving window classification rule based on independent copies of the pair (X, Y). Contrary to the finite dimensional case, it is shown that the moving window classifier is not universally consistent in the sense that its probability of error may not converge to the Bayes risk for some distributions of (X, Y). Sufficient conditions both on the space $\mathcal{F}$ and the distribution of X are then given to ensure consistency.

Index Terms **— Classification, kernel rule, consistency, universal consistency, metric entropy.**

***AMS 2000 Classification*: 62G08.**

1 Introduction

In many experiments, scientists and practitioners often collect samples of curves and other functional observations. For instance, curves arise naturally as observations in the investigation of growth, in climate analysis, in food industry or in speech recognition; Ramsay and Silverman [25] discuss other examples. The aim of the present paper is to investigate whether the classical nonparametric classification rule based on kernel (as discussed, for

*Corresponding author. Email: biau@math.univ-montp2.fr .

example, in Devroye, Györfi and Lugosi [9]) can be extended to classify functions.

Classical classification deals with predicting the unknown nature Y, called a *label*, of an observation X with values in $\mathbb{R}^d$ (see Boucheron, Bousquet and Lugosi [6] for a recent survey). Both X and Y are assumed to be random, and the distribution of (X, Y) just describes the frequency of encountering particular pairs in practice. We require for simplicity that the label only takes two values, say 0 and 1. If we denote by F_0 (*resp.* F_1) the conditional distribution of X given $Y = 0$ (*resp.* $Y = 1$), then the distribution of X is given by $\pi F_0 + (1 - \pi)F_1$, where $\pi = \mathbf{P}\{Y = 0\}$. Note that, in this framework, the label Y is random, and this casts the classification problem into a bounded regression problem.

The statistician creates a *classifier* $g : \mathbb{R}^d \to \{0, 1\}$ which represents her guess of the label of X. An error occurs if $g(X) \neq Y$, and the probability of error for a particular classifier g is

$$L(g) = \mathbf{P}\{g(X) \neq Y\}.$$

It is easily seen that the *Bayes rule*

$$g^*(x) = \begin{cases} 0 & \text{if } \mathbf{P}\{Y = 0|X = x\} \geq \mathbf{P}\{Y = 1|X = x\} \\ 1 & \text{otherwise,} \end{cases} \tag{1.1}$$

is the optimal decision, in the sense that, for any decision function $g : \mathbb{R}^d \to \{0, 1\}$,

$$\mathbf{P}\{g^*(X) \neq Y\} \leq \mathbf{P}\{g(X) \neq Y\}.$$

Unfortunately, the Bayes rule depends on the distribution of (X, Y), which is unknown to the statistician. The problem is thus to construct a reasonable classifier g_n based on independent observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ with the same distribution as (X, Y).

Among the various ways to define such classifiers, one of the most simple and popular is probably the *moving window rule* given by

$$g_n(x) = \begin{cases} 0 & \text{if } \sum_{i=1}^n \mathbf{1}_{\{Y_i=0, X_i \in B_{x,h_n}\}} \geq \sum_{i=1}^n \mathbf{1}_{\{Y_i=1, X_i \in B_{x,h_n}\}} \\ 1 & \text{otherwise,} \end{cases} \tag{1.2}$$

where h_n is a (strictly) positive real number, depending only on n and called the *smoothing factor*, and B_{x,h_n} denotes the closed ball of radius h_n centered at x. It is possible to make the decision even smoother using a kernel K (that is, a nonnegative and monotone function decreasing along rays starting from the origin) by giving more weights to closer points than to more distant ones, deciding 0 if $\sum_{i=1}^n \mathbf{1}_{\{Y_i=0\}} K\big((x-X_i)/h_n\big) \geq \sum_{i=1}^n \mathbf{1}_{\{Y_i=1\}} K\big((x-X_i)/h_n\big)$ and 1 otherwise, but that will not concern us here. Kernel-based rules are derived from the kernel estimate in density estimation originally studied by Akaike [2], Rosenblatt [27], and Parzen [23]; and in regression estimation, introduced by Nadaraya [21], [22], and Watson [31]. For particular choices of K, statistical analyses of rules of this sort and/or the corresponding regression function estimates have been studied by many authors. For a complete and updated list of references, we refer the reader to the monograph by Devroye, Györfi and Lugosi [9], Chapter 10. Now, if we are given any classification rule g_n based on the training data $(X_1, Y_1), \ldots, (X_n, Y_n)$, the best we can expect from the classification function g_n is to achieve the Bayes error probability $L^* = L(g^*)$. Generally, we cannot hope to obtain a function that exactly achieves the Bayes error probability, and we rather require that the error probability

$$L_n = \mathbf{P}\big\{g_n(X) \neq Y | (X_1, Y_1), \ldots, (X_n, Y_n)\big\}$$

gets arbitrarily close to L^* with large probability. More precisely, a classification rule g_n is called *consistent* if

$$\mathbf{E}L_n = \mathbf{P}\{g_n(X) \neq Y\} \to L^* \quad \text{as } n \to \infty\,,$$

and *strongly consistent* if

$$\lim_{n\to\infty} L_n = L^* \quad \text{with probability one.}$$

A decision rule can be consistent for a certain class of distributions of (X, Y), but may not be consistent for others. On the other hand, it is clearly desirable to have a rule that gives good performance for all distributions. With this respect, a decision rule is called *universally* (strongly) consistent if it is (strongly) consistent for any distribution of the pair (X, Y). When X is $\mathbb{R}^d$-valued, it is known from Devroye and Krzyżak [10] that the classical conditions $h_n \to 0$ and $nh_n^d \to \infty$ as $n \to \infty$ ensure that the moving window rule (1.2) is universally strongly consistent.

In this paper, we wish to investigate consistency properties of the moving window rule (1.2) in the setting of random functions, that is when X takes values in a metric space $(\mathcal{F}, \rho)$ instead of $\mathbb{R}^d$. The scope of applications is vast, including disciplines such as medicine (discriminate electrocardiograms from two different groups of patients), finance (sell or buy stocks regarding their evolution in time), biometrics (discern pertinent fingerprint patterns), or image recognition on the Internet.

In this general framework, the infinite dimensional version of the classification rule (1.2) under study reads

$$g_n(x) = \begin{cases} 0 & \text{if } \sum_{i=1}^n \mathbf{1}_{\{Y_i=0, X_i \in B_{x,h_n}\}} \geq \sum_{i=1}^n \mathbf{1}_{\{Y_i=1, X_i \in B_{x,h_n}\}} \\ 1 & \text{otherwise,} \end{cases} \tag{1.3}$$

where B_{x,h_n} is now a ball in $(\mathcal{F}, \rho)$ – the optimal decision remains the Bayes one $g^* : \mathcal{F} \to \{0, 1\}$ as in (1.1). Probably due to the difficulty of the problem, and despite nearly unlimited applications, the theoretical literature on regression and classification in infinite dimensional spaces is relatively recent. Key references on this topic are Rice and Silverman [26], Kneip and Gasser [16], Kulkarni and Posner [18], Ramsay and Silverman [25], Bosq [5], Ferraty and Vieu [12], [13], [14], Diabo-Niang and Rhomari [11], Hall, Poskitt and Presnell [15], Abraham, Cornillon, Matzner-Løber and Molinari [1], Antoniadis and Sapatinas [3], and Biau, Bunea and Wegkamp [4]. We also mention that Cover and Hart [8] consider classification of Banach space valued elements as well, but they do not establish consistency.

As pointed out by an Associate Editor, the classification rule (1.3) is feeded with infinite dimensional observations as inputs. In particular, it does not require any preliminary dimension reduction or model selection step. On the other hand, in the so-called "filtering approach", one first reduces the infinite dimension of the observations by considering only the first d coefficients of the data on an appropriate basis, and then perform finite dimensional classification. For more on this alternative approach, we refer the reader to Hall, Poskitt and Presnell [15], Abraham, Cornillon, Matzner-Løber and Molinari [1], Biau, Bunea and Wegkamp [4] and the references therein.

As a first important contribution, we show in Section 2 that the universal consistency result valid for the rule (1.2) in the finite dimensional case breaks down as soon as X is allowed to take values in a space of functions. More precisely, we are able to exhibit a normed function space and a distribution of (X, Y) such that the moving window rule (1.3) fails to be consistent. This negative finding makes it legitimate to put some restrictions both on the functional space and the distribution of X in order to obtain the desired consistency properties. Sufficient conditions of this sort are given in Section 3 (Theorem 3.1 deals with consistency whereas Theorem 3.2 with strong consistency) along with examples of applications. These conditions involve both the support of the distribution of X and the way this distribution locally spreads out. For the sake of clarity, proofs are gathered in Section 4.

2 Non-universal consistency of the moving window rule

Let $(h_n)_{n\geq 1}$ be a *given* sequence of smoothing factors such that $h_n \to 0$ as $n \to \infty$. Our purpose in this section is to show that there exists a normed function space $\big(\mathcal{F}, \|.\|\big)$, a random variable X taking values in this space and a distribution of (X, Y) such that the moving window rule (1.3) fails to be consistent. For any pair (X, Y), we denote by $\eta(x)$ the conditional probability that Y is 1 given $X = x$, *i.e*,

$$\eta(x) = \mathbf{P}\{Y = 1 | X = x\} = \mathbf{E}[Y | X = x] \,.$$

Good candidates may be designed as follows.

Preliminaries Define the space $\big(\mathcal{F}, \|.\|\big)$ as the space of functions from $]0, 1]$ to $[0, 1]$ endowed with the supremum norm $\|.\| = \|.\|_\infty$, and let X be a random variable (to be specified later) taking values in $\mathcal{F}$. Choose finally a label Y which is 1 with probability one, and thus $\eta(x) = 1$ and $L^* = 0$. Following the lines of the proof of Theorem 2.2 in Devroye, Györfi and Lugosi [9] (Chapter 2, page 16) it is easily seen that

$$\begin{aligned}
\mathbf{P}\{g_n(X) \neq Y\} - L^* &= \mathbf{E}\big[|2\eta(X) - 1| \mathbf{1}_{\{g_n(X) \neq g^*(X)\}}\big] \\
&= \mathbf{E}\big[\mathbf{1}_{\{g_n(X) \neq g^*(X)\}}\big] \,,
\end{aligned}$$

where the last equality follows from our choice of η. We emphasize that g_n predicts the label 0 when there are no data falling around x, *i.e.*, setting $N(x) = \sum_{i=1}^n \mathbf{1}_{\{X_i \in B_{x,h_n}\}}$, when $N(x) = 0$. When x belongs to $\mathbb{R}^d$, the conditions $h_n \to 0$ and $nh_n^d \to \infty$ as $n \to \infty$ ensure that the misspecification when $N(x) = 0$ is not crucial for consistency (see Devroye and Krzyżak [10]). The remainder of the paragraph shows that things are different when x is a function. Observe first that

$$\begin{aligned}
\mathbf{1}_{\{g_n(X) \neq g^*(X)\}} &\geq \mathbf{1}_{\{g^*(X)=1, g_n(X)=0\}} \\
&\geq \mathbf{1}_{\{\eta(X)>1/2, N(X)=0\}} \\
&= \mathbf{1}_{\{N(X)=0\}}
\end{aligned}$$

since $\eta(X) = 1$. Therefore, we are led to

$$\begin{aligned}
\mathbf{P}\{g_n(X) \neq Y\} - L^* &\geq \mathbf{E}\big[\mathbf{1}_{\{N(X)=0\}}\big] \\
&= \mathbf{E}\Big[\mathbf{E}\big[\mathbf{1}_{\{N(X)=0\}} | X\big]\Big] \,.
\end{aligned}$$

Clearly, the distribution of $N(X)$ given X is binomial $\mathrm{Bin}(n, P_X)$, with

$$P_X = \mathbf{P}\big\{\|X' - X\| \le h_n | X\big\}\,,$$

where X' is an independent copy of X. It follows that

$$\begin{aligned}\mathbf{P}\{g_n(X) \neq Y\} - L^* &\geq \mathbf{E}\big[(1 - P_X)^n\big]\\ &\geq \mathbf{E}[1 - nP_X]\\ &= 1 - n\mathbf{P}\big\{\|X' - X\| \le h_n\big\}\,.\end{aligned}$$

Having disposed of this preliminary step, we propose now to prove the existence of a $\mathcal{F}$-valued random variable X such that $n\mathbf{P}\big\{\|X - X'\| \le h_n\big\}$ goes to zero as n grows.

Example Take $U_0, U_1, U_2, \ldots$ to be an infinite sequence of independent random variables uniformly distributed on $[0,1]$ and let X be the random function from $]0,1]$ to $[0,1]$ constructed as follows: for $t = 2^{-i}$, $i = 0, 1, 2, \ldots$, set $X(t) = X(2^{-i}) = U_i$, and for $t \in]2^{-(i+1)}, 2^{-i}[$, define $X(t)$ by linear interpolation. We thus obtain a continuous random function X which is linear on each interval $[2^{-(i+1)}, 2^{-i}]$. Denote by X' an independent copy of X derived from $U'_0, U'_1, U'_2, \ldots$ Attention shows that, with probability one, the following equality holds:

$$\|X - X'\| = \sup_{i \ge 0} |U_i - U'_i| = 1\,.$$

Therefore, for all n large enough,

$$\mathbf{P}\{g_n(X) \neq Y\} - L^* \geq 1\,,$$

what shows that the moving window rule cannot be consistent for the considered distribution of (X, Y).

Note that the same result holds if $U_0, U_1, \ldots$ are chosen independently with a standard Gaussian distribution. In this case, X is a continuous Gaussian process.

One can argue that our example is rather pathological, as the distance between two random functions X and X' is almost surely equal to one. Things can be slightly modified to avoid this inconvenience. To this aim, construct first, for each integer $k \geq 1$, a random function X_k as above with the U_i's uniformly distributed on $[0, k^{-1}]$, and denote by $(X'_k)_{k\geq 1}$ an independent copy of the sequence $(X_k)_{k\geq 1}$. A trivial verification shows that, with probability one, for $k, k' \geq 1$,

$$\|X_k - X'_{k'}\| = \max\{k^{-1}, k'^{\,-1}\}\,.$$

Second, denote by K a discrete random variable satisfying $\mathbf{P}\{K=k\}=w_k$, where $(w_k)_{k\geq 1}$ is a sequence of positive weights adding to one. Define the conditional distribution of X given $\{K=k\}$ as the distribution of X_k and denote by X' an independent copy of X associated with K' (independent of K). Then it is a simple exercise to prove that for any sequence of smoothing factors $(h_n)_{n\geq 1}$ verifying $h_n \to 0$ as $n \to \infty$, one can find a sequence of weights $(w_k)_{k\geq 1}$ and thus a random variable X such that

$$\liminf_{n\to\infty} \mathbf{P}\{g_n(X) \neq Y\} - L^* \geq 1\,.$$

Thus, the moving window rule is not universally consistent, whatever the choice of the sequence $(h_n)_{n\geq 1}$.

3 Consistent classification in function spaces

3.1 Notation and assumptions

In view of the results of Section 2, we are now interested in finding sufficient conditions ensuring consistency of the moving window rule (1.3).

Let us first introduce the abstract mathematical model. Let X be a random variable taking values in a metric space $(\mathcal{F}, \rho)$ and let Y be a random label with values 0 and 1. The distribution of the pair (X, Y) is completely specified by μ, the probability measure of X and by η, the regression function of Y on X. That is, for any Borel-measurable set $A \subset \mathcal{F}$,

$$\mu(A) = \mathbf{P}\{X \in A\}$$

and, for any $x \in \mathcal{F}$, $\eta(x) = \mathbf{P}\{Y = 1 | X = x\}$. Given independent copies $(X_1, Y_1), \ldots, (X_n, Y_n)$ of (X, Y), the goal is to classify a new random element from the same distribution μ, independent of the training data, using the moving window rule. Let us now recall the important and well-known notions of covering numbers and metric entropy which characterize the massiveness of a set. Following Kolmogorov and Tihomirov [17], these quantities have been extensively studied and used in various applications. Denote by $S_{x,\varepsilon}$ the open ball of radius ε about a point $x \in \mathcal{F}$.

Definition 3.1 *Let $\mathcal{G}$ be a subset of the metric space $(\mathcal{F}, \rho)$. The ε-*covering number *$\mathcal{N}(\varepsilon)$ $\big(= \mathcal{N}(\varepsilon, \mathcal{G}, \rho)\big)$ is defined as the smallest number of open balls of radius ε that cover the set $\mathcal{G}$. That is*

$$\mathcal{N}(\varepsilon) = \inf\left\{k \geq 1 : \exists x_1, \ldots, x_k \in \mathcal{F} \text{ with } \mathcal{G} \subset \bigcup_{i=1}^{k} S_{x_i,\varepsilon}\right\}.$$

The logarithm of the ε-covering number is often referred to as the *metric entropy* or *ε-entropy.* A set $\mathcal{G} \subset \mathcal{F}$ is said to be *totally bounded* if $\mathcal{N}(\varepsilon) < \infty$ for all $\varepsilon > 0$. In particular, every relatively compact set is totally bounded and all totally bounded sets are bounded.

Our first basic assumption in the present paper is that there exists a sequence $(\mathcal{F}_k)_{k\geq 1}$ of *totally bounded subsets* of $\mathcal{F}$ such that

$$\mathcal{F}_k \subset \mathcal{F}_{k+1} \quad \text{for all } k \geq 1 \quad \text{and} \quad \mu\Big(\bigcup_{k\geq 1} \mathcal{F}_k\Big) = 1 \quad (\mathbf{H1}).$$

Various examples will be discussed below. It is worth pointing out that this condition is mild. It is for example satisfied whenever $(\mathcal{F}, \rho)$ is a separable metric space. Note also that a similar requirement is imposed by Kulkarni and Posner [18] who study the problem of nearest neighbor estimation under arbitrary sampling in a general separable metric space.

Our second assumption asks that the following differentiation result holds:

$$\lim_{h\to 0} \frac{1}{\mu(B_{x,h})} \int_{B_{x,h}} \eta \, \mathrm{d}\mu = \eta(x) \quad \text{in } \mu\text{-probability}, \quad (\mathbf{H2})$$

which means that for every $\varepsilon > 0$,

$$\lim_{h\to 0} \mu\Big\{x \in \mathcal{F} : \Big|\frac{1}{\mu(B_{x,h})} \int_{B_{x,h}} \eta \, \mathrm{d}\mu - \eta(x)\Big| > \varepsilon\Big\} = 0 .$$

If $\mathcal{F}$ is a finite dimensional vector space, this differentiation theorem turns to be true (Rudin [28], Chapter 8). There have been several attempts to generalize this kind of results to general metric spaces (see Mattila [19], Preiss and Tiser [24], Tiser [29] and the references therein for examples, counterexamples and discussions). The general finding here is that equality (**H2**) holds in typically infinite dimensional spaces if we ask conditions both on the structure of the space $\mathcal{F}$ (such as to be an Hilbert) and the measure μ (such as to be Gaussian) – see the examples below.

We draw attention on the fact that condition (**H2**) holds as soon as the regression function η is μ-a.e. continuous. In particular, when μ is nonatomic, (**H2**) holds for a piecewise continuous function η. Piecewise continuous η's in $\mathbb{R}^d$ are associated with the so-called "change-set problem" – an interesting problem of spatial statistics, which is undoubtedly no less interesting in functional spaces.

Before we present our consistency results, we illustrate the generality of the approach by working out several examples for different classes.

3.2 Examples

• As a first example, just take $\mathcal{F} = \mathbb{R}^d$ endowed with any norm $\|.\|$. In this case, condition (**H1**) is obviously true and (**H2**) holds according to the classical differentiation theorem (Rudin [28], Chapter 8).

• Consider now the less trivial situation where the regression function η is μ-a.e. continuous – so that (**H2**) is superfluous – and where the random elements to be classified are known to be bounded and Hölder functions of some order $\alpha > 0$ defined on a bounded, convex subset Ξ of $\mathbb{R}^d$ with nonempty interior. Note that the standard Brownian paths on $[0, 1]$ satisfy this condition with $\alpha > 1/2$, and that in the important case where X is a Gaussian process, the Hölder parameter α may be estimated using an Hölder property of the covariance function of X, see Ciesielski [7]. The natural balls

$$\mathcal{F}_k = \{\text{all continuous functions } f : \Xi \to \mathbb{R} \text{ with } \|f\|_\infty \leq k\}$$

are not totally bounded in $\mathcal{F}$ endowed with the supremum norm $\|.\|_\infty$. However, a slight change in the definition of the balls leads to a tractable model. That is, take

$$\mathcal{F} = \{\text{all bounded continuous functions } f : \Xi \to \mathbb{R}\}$$

and, for each $k \geq 1$,

$$\mathcal{F}_k = \{\text{all continuous functions } f : \Xi \to \mathbb{R} \text{ with } \|f\|_\alpha \leq k\}$$

with

$$\|f\|_\alpha = \sup_t \big|f(t)\big| + \sup_{s \neq t} \frac{\big|f(s) - f(t)\big|}{\|s - t\|^\alpha},$$

where the suprema are taken over all points in the interior of Ξ and $\|.\|$ denotes the norm on $\mathbb{R}^d$. Bounds on the metric entropy of the classes $\mathcal{F}_k$ with respect to the supremum norm were among the first known after the introduction of covering numbers. In the present context, it can be shown (see, for example, van der Vaart and Wellner [30], Chapter 2.7) that there exists a constant A depending only on α, d, k and Ξ such that

$$\log \mathcal{N}\big(\varepsilon, \mathcal{F}_k, \|.\|_\infty\big) \leq A \left(\frac{1}{\varepsilon}\right)^{d/\alpha}$$

for every $\varepsilon > 0$.

• Now, if we do not suppose that the regression function η is μ-a.e. continuous, then we have to ask a bit more both on the underlying space $\mathcal{F}$ and the measure μ to ensure that assumption (**H2**) holds. Assume for example that $\mathcal{F}$ is a Hilbert space and that μ is a centered Gaussian measure with the following spectral representation of its covariance operator:

$$Rx = \sum_{i \geq 1} c_i (x, e_i) e_i \,,$$

where $(.,.)$ is the scalar product and $(e_i)_{i\geq 1}$ is an orthonormal system in $\mathcal{F}$. If the sequence $(c_i)_{i\geq 1}$ satisfies

$$0 < \frac{c_{i+1}}{c_i} \leq q, \quad i \geq 1, \tag{3.1}$$

where $q < 1$, then (**H2**) holds (Preiss and Tiser [24]). As an illustration, keep $\mathcal{F}$ and the $\mathcal{F}_k$'s defined as in the previous example, and still assume that $\mu(\cup_{k\geq 1}\mathcal{F}_k) = 1$. Let Q be a probability measure on Ξ. Consider the $L_2(Q)$ norm defined by

$$\|f\|_{2,Q}^2 = \int |f|^2 \,\mathrm{d}Q$$

and the Hilbert space $\big(\mathcal{F}, \|.\|_{2,Q}\big)$. Then it can be shown (van der Vaart and Wellner [30], Chapter 2.7) that there exists a constant B, depending only on α, d, k and Ξ such that

$$\log \mathcal{N}\big(\varepsilon, \mathcal{F}_k, \|.\|_{2,Q}\big) \leq B\Big(\frac{1}{\varepsilon}\Big)^{d/\alpha}$$

for every $\varepsilon > 0$. Thus, any Gaussian measure – or any mixture of two Gaussian measures – whose covariance operator satisfies requirement (3.1) above and meeting the condition $\mu(\cup_{k\geq 1}\mathcal{F}_k) = 1$ can be dealt with the tools presented in the present paper.

3.3 Results

Following the notation of the introduction for the finite dimensional case, we let L^* and L_n be the probability of error for the Bayes rule and the moving window rule, respectively. In this paragraph, we establish consistency (Theorem 3.1) and strong consistency (Theorem 3.2) of the moving window rule g_n under assumptions (**H1**), (**H2**) and general conditions on the smoothing factor h_n. The notation $\mathcal{G}^c$ stands for the complement of any subset $\mathcal{G}$ in $\mathcal{F}$. For simplicity, the dependence of h_n on n is always understood and we write $\mathcal{N}_k(\varepsilon)$ instead of $\mathcal{N}(\varepsilon, \mathcal{F}_k, \rho)$.

Theorem 3.1 [CONSISTENCY] *Assume that* (**H1**) *and* (**H2**) *hold. If* $h \to 0$ *and, for every* $k \geq 1$, $\mathcal{N}_k(h/2)/n \to 0$ *as* $n \to \infty$, *then*

$$\lim_{n\to\infty} \mathbf{E}L_n = L^* .$$

Theorem 3.2 [STRONG CONSISTENCY] *Assume that* (**H1**) *and* (**H2**) *hold. Let* $(k_n)_{n\geq 1}$ *be an increasing sequence of positive integers such that*

$$\sum_{n\geq 1} \mu(\mathcal{F}^c_{k_n}) < \infty .$$

If $h \to 0$ *and*

$$\frac{n}{(\log n)\mathcal{N}^2_{k_n}(h/2)} \to \infty \quad \text{as } n \to \infty ,$$

then

$$\lim_{n\to\infty} L_n = L^* \quad \text{with probability one.}$$

Remark 1 Practical applications exceed the scope of this paper. However, the applied statistician should be aware of the following two points.

First, for a particular n, asymptotic results provide little guidance in the selection of h. On the other hand, selecting the wrong value of h may lead to catastrophic error rates – in fact, the crux of every nonparametric estimation problem is the choice of an appropriate smoothing factor. The question of how to select automatically and optimally a data-dependent smoothing factor h will be addressed in a future work. Note however that one can always find a sequence of smoothing factors satisfying the conditions of Theorem 3.1 and Theorem 3.2.

Second, in practice, the random elements are always observed at discrete sampling times only (deterministic or random) and are possibly contaminated with measurement errors. The challenge then is to explore properties of classifiers based on estimated functions rather than on true (but unobserved) functions.

4 Proofs

4.1 Preliminary results

Define

$$\eta_n(x) = \frac{\sum_{i=1}^n Y_i \mathbf{1}_{\{X_i \in B_{x,h}\}}}{n\mu(B_{x,h})} ,$$

and observe that the decision rule can be written as

$$g_n(x) = \begin{cases} 0 & \text{if } \frac{\sum_{i=1}^n Y_i \mathbf{1}_{\{X_i \in B_{x,h}\}}}{n\mu(B_{x,h})} \le \frac{\sum_{i=1}^n (1-Y_i)\mathbf{1}_{\{X_i \in B_{x,h}\}}}{n\mu(B_{x,h})} \\ 1 & \text{otherwise.} \end{cases}$$

Thus, by Theorem 2.3 in Devroye, Györfi and Lugosi [9] (Chapter 2, page 17) – whose extension to the infinite dimensional setting is straightforward – Theorem 3.1 will be demonstrated if we show that

$$\mathbf{E}\left[\int \left|\eta(x) - \eta_n(x)\right|\mu(\mathrm{d}x)\right] \to 0 \quad \text{as } n \to \infty$$

and Theorem 3.2 if we prove that

$$\int \left|\eta(x) - \eta_n(x)\right|\mu(\mathrm{d}x) \to 0 \quad \text{with probability one as } n \to \infty\,.$$

Proofs of Theorem 3.1 and Theorem 3.2 will strongly rely on the following three lemmas. Proof of Lemma 4.1 is a straightforward consequence of assumption (**H2**) and the Lebesgue dominated convergence theorem.

Lemma 4.1 *Assume that* (**H2**) *holds. If $h \to 0$, then*

$$\int \left|\eta(x) - \mathbf{E}\eta_n(x)\right|\mu(\mathrm{d}x) = \int \left|\eta(x) - \frac{\int_{B_{x,h}} \eta(t)\mu(\mathrm{d}t)}{\mu(B_{x,h})}\right|\mu(\mathrm{d}x) \to 0$$

as $n \to \infty$.

Lemma 4.2 *Let k be a fixed positive integer. Then, for every $h > 0$,*

$$\int_{\mathcal{F}_k} \frac{1}{\mu(B_{x,h})}\mu(\mathrm{d}x) \le \mathcal{N}_k\left(\frac{h}{2}\right).$$

Proof Since, by assumption, $\mathcal{F}_k$ is totally bounded, there exist $a_1, \ldots, a_{\mathcal{N}_k(h/2)}$ elements of $\mathcal{F}$ such that

$$\mathcal{F}_k \subset \bigcup_{j=1}^{\mathcal{N}_k(h/2)} B_{a_j,h/2}\,.$$

Therefore

$$\int_{\mathcal{F}_k} \frac{1}{\mu(B_{x,h})}\mu(\mathrm{d}x) \le \sum_{j=1}^{\mathcal{N}_k(h/2)} \int_{B_{a_j,h/2}} \frac{1}{\mu(B_{x,h})}\mu(\mathrm{d}x)\,.$$

Then $x \in B_{a_j,h/2}$ implies $B_{a_j,h/2} \subset B_{x,h}$ and thus

$$\int_{\mathcal{F}_k} \frac{1}{\mu(B_{x,h})}\mu(\mathrm{d}x) \le \sum_{j=1}^{\mathcal{N}_k(h/2)} \int_{B_{a_j,h/2}} \frac{1}{\mu(B_{a_j,h/2})}\mu(\mathrm{d}x) = \mathcal{N}_k\left(\frac{h}{2}\right).$$

■

Lemma 4.3 *Let k be a fixed positive integer. Then, for all $n \geq 1$,*

$$\mathbf{E}\Big[\int_{\mathcal{F}_k} \big|\eta_n(x) - \mathbf{E}\eta_n(x)\big|\mu(\mathrm{d}x)\Big] \leq \Big(\frac{1}{n}\mathcal{N}_k\Big(\frac{h}{2}\Big)\Big)^{1/2}.$$

Proof According to Devroye, Györfi and Lugosi [9] (Chapter 10, page 157) one has, for every $x \in \mathcal{F}$ and $n \geq 1$:

$$\mathbf{E}\Big[\big|\eta_n(x) - \mathbf{E}\eta_n(x)\big|\Big] \leq \frac{1}{\sqrt{n\mu(B_{x,h})}}.$$

Consequently,

$$\begin{aligned}
\mathbf{E}\Big[\int_{\mathcal{F}_k} \big|\eta_n(x) - \mathbf{E}\eta_n(x)\big|\mu(\mathrm{d}x)\Big] &\leq \int_{\mathcal{F}_k} \frac{1}{\sqrt{n\mu(B_{x,h})}}\mu(\mathrm{d}x) \\
&\leq \Big(\int_{\mathcal{F}_k} \frac{1}{n\mu(B_{x,h})}\mu(\mathrm{d}x)\Big)^{1/2} \\
&\quad \text{(by Jensen's inequality)} \\
&\leq \Big(\frac{1}{n}\mathcal{N}_k\Big(\frac{h}{2}\Big)\Big)^{1/2} \\
&\quad \text{(by Lemma 4.2).}
\end{aligned}$$

■

4.2 Proof of Theorem 3.1

We have, for every $k \geq 1$,

$$\begin{aligned}
&\mathbf{E}\Big[\int \big|\eta(x) - \eta_n(x)\big|\mu(\mathrm{d}x)\Big] \\
&\quad = \mathbf{E}\Big[\int_{\mathcal{F}_k} \big|\eta(x) - \eta_n(x)\big|\mu(\mathrm{d}x)\Big] + \mathbf{E}\Big[\int_{\mathcal{F}_k^c} \big|\eta(x) - \eta_n(x)\big|\mu(\mathrm{d}x)\Big] \\
&\quad \leq \int_{\mathcal{F}_k} \big|\eta(x) - \mathbf{E}\eta_n(x)\big|\mu(\mathrm{d}x) + \mathbf{E}\Big[\int_{\mathcal{F}_k} \big|\eta_n(x) - \mathbf{E}\eta_n(x)\big|\mu(\mathrm{d}x)\Big] + 2\mu(\mathcal{F}_k^c),
\end{aligned}$$

where in the last inequality we used the fact that $\eta(x) \leq 1$ and $\mathbf{E}\eta_n(x) \leq 1$ for every $x \in \mathcal{F}$ and $n \geq 1$. As a consequence, using Lemma 4.3,

$$\mathbf{E}\Big[\int \big|\eta(x) - \eta_n(x)\big|\mu(\mathrm{d}x)\Big] \leq \int \big|\eta(x) - \mathbf{E}\eta_n(x)\big|\mu(\mathrm{d}x) + \Big(\frac{1}{n}\mathcal{N}_k\Big(\frac{h}{2}\Big)\Big)^{1/2} + 2\mu(\mathcal{F}_k^c).$$

Therefore, according to Lemma 4.1 and the assumptions on h, for every $k \geq 1$,

$$\limsup_{n\to\infty} \mathbf{E}\left[\int \left|\eta(x) - \eta_n(x)\right|\mu(\mathrm{d}x)\right] \leq 2\mu(\mathcal{F}_k^c)\,.$$

The conclusion follows under (**H1**) if we let k converge to infinity. ■

4.3 Proof of Theorem 3.2

Let $(k_n)_{n\geq 1}$ be the sequence defined in Theorem 3.2. We first proceed to show that

$$\int_{\mathcal{F}_{k_n}} \left|\eta(x) - \eta_n(x)\right|\mu(\mathrm{d}x) \to 0 \quad \text{with probability one as } n \to \infty\,. \tag{4.1}$$

According to Lemma 4.3, we have

$$\begin{aligned}
&\mathbf{E}\left[\int_{\mathcal{F}_{k_n}} \left|\eta(x) - \eta_n(x)\right|\mu(\mathrm{d}x)\right] \\
&\quad \leq \int \left|\eta(x) - \mathbf{E}\eta_n(x)\right|\mu(\mathrm{d}x) + \mathbf{E}\left[\int_{\mathcal{F}_{k_n}} \left|\eta_n(x) - \mathbf{E}\eta_n(x)\right|\mu(\mathrm{d}x)\right] \\
&\quad \leq \int \left|\eta(x) - \mathbf{E}\eta_n(x)\right|\mu(\mathrm{d}x) + \left(\frac{1}{n}\mathcal{N}_{k_n}\left(\frac{h}{2}\right)\right)^{1/2}.
\end{aligned}$$

Therefore, applying Lemma 4.1 and the assumptions on h, we obtain

$$\mathbf{E}\left[\int_{\mathcal{F}_{k_n}} \left|\eta(x) - \eta_n(x)\right|\mu(\mathrm{d}x)\right] \to 0 \quad \text{as } n \to \infty\,.$$

Consequently, (4.1) will be proved if we show that

$$\int_{\mathcal{F}_{k_n}} \left|\eta(x) - \eta_n(x)\right|\mu(\mathrm{d}x) - \mathbf{E}\left[\int_{\mathcal{F}_{k_n}} \left|\eta(x) - \eta_n(x)\right|\mu(\mathrm{d}x)\right] \to 0$$

with probability one as $n \to \infty$. To do this, we use McDiarmid's inequality [20] for

$$\int_{\mathcal{F}_{k_n}} \left|\eta(x) - \eta_n(x)\right|\mu(\mathrm{d}x) - \mathbf{E}\left[\int_{\mathcal{F}_{k_n}} \left|\eta(x) - \eta_n(x)\right|\mu(\mathrm{d}x)\right].$$

Fix the training data at $(x_1, y_1), \ldots, (x_n, y_n)$ and replace the i-th pair (x_i, y_i) by $(\hat{x_i}, \hat{y_i})$, changing the value of $\eta_n(x)$ to $\eta^*_{ni}(x)$. Clearly,

$$\begin{aligned}
&\left| \int_{\mathcal{F}_{k_n}} \big|\eta(x) - \eta_n(x)\big| \mu(\mathrm{d}x) - \int_{\mathcal{F}_{k_n}} \big|\eta(x) - \eta^*_{ni}(x)\big| \mu(\mathrm{d}x) \right| \\
&\quad \le \int_{\mathcal{F}_{k_n}} \big|\eta_n(x) - \eta^*_{ni}(x)\big| \mu(\mathrm{d}x) \\
&\quad \le \frac{2}{n} \int_{\mathcal{F}_{k_n}} \frac{1}{\mu(B_{x,h})} \mu(\mathrm{d}x) \\
&\quad \le \frac{2}{n} \mathcal{N}_{k_n}\left(\frac{h}{2}\right),
\end{aligned}$$

where the last inequality arises from Lemma 4.2. So, by McDiarmid's inequality [20], for every $\alpha > 0$,

$$\begin{aligned}
&\mathbf{P}\left\{ \left| \int_{\mathcal{F}_{k_n}} \big|\eta(x) - \eta_n(x)\big| \mu(\mathrm{d}x) - \mathbf{E}\left[\int_{\mathcal{F}_{k_n}} \big|\eta(x) - \eta_n(x)\big| \mu(\mathrm{d}x) \right] \right| \ge \alpha \right\} \\
&\quad \le 2 \exp\left(- \frac{\rho\, n}{\mathcal{N}^2_{k_n}(h/2)} \right),
\end{aligned}$$

for some positive constant ρ depending only on α. Thus, using the assumption on h and the Borel-Cantelli lemma, we conclude that

$$\int_{\mathcal{F}_{k_n}} \big|\eta(x) - \eta_n(x)\big| \mu(\mathrm{d}x) - \mathbf{E}\left[\int_{\mathcal{F}_{k_n}} \big|\eta(x) - \eta_n(x)\big| \mu(\mathrm{d}x) \right] \to 0$$

with probability one as $n \to \infty$. This proves (4.1).

To finish the proof, let us denote for all $n \ge 1$ and $i = 1, \ldots, n$,

$$Z_i^n = \int_{\mathcal{F}^c_{k_n}} \frac{\mathbf{1}_{\{X_i \in B_{x,h}\}}}{\mu(B_{x,h})} \mu(\mathrm{d}x) \,.$$

Observe that

$$\mathbf{E}\left[\frac{1}{n} \sum_{i=1}^n Z_i^n \right] = \mu(\mathcal{F}^c_{k_n}) \,.$$

Applying the Borel-Cantelli Lemma together with the condition $\sum_{n \ge 1} \mu(\mathcal{F}^c_{k_n}) < \infty$ yields

$$\frac{1}{n} \sum_{i=1}^n Z_i^n \to 0 \quad \text{with probability one as } n \to \infty \,. \tag{4.2}$$

Write finally

$$\begin{aligned}\int \big|\eta(x) - \eta_n(x)\big|\mu(\mathrm{d}x) &= \int_{\mathcal{F}_{k_n}} \big|\eta(x) - \eta_n(x)\big|\mu(\mathrm{d}x) + \int_{\mathcal{F}_{k_n}^c} \big|\eta(x) - \eta_n(x)\big|\mu(\mathrm{d}x)\\ &\le \int_{\mathcal{F}_{k_n}} \big|\eta(x) - \eta_n(x)\big|\mu(\mathrm{d}x) + \mu(\mathcal{F}_{k_n}^c) + \frac{1}{n}\sum_{i=1}^{n} Z_i^n\,,\end{aligned}$$

and this last term goes to 0 according to (**H1**), (4.1) and (4.2). This completes the proof of Theorem 3.2. ■

Acknowledgments. The authors greatly thank an Associate Editor for a careful reading of the paper and many helpful comments.

References

[1] Abraham, C., Cornillon, P.A., Matzner-Løber, E. and Molinari, N. (2003). Unsupervised curve clustering using B-splines, *Scandinavian Journal of Statistics*, Vol. 30, pp. 581–595.

[2] Akaike, H. (1954). An approximation to the density function, *Annals of the Institute of Statistical Mathematics*, Vol. 6, pp. 127–132.

[3] Antoniadis, A. and Sapatinas, T. (2003). Wavelet methods for continuous-time prediction using representations of autoregressive processes in Hilbert spaces, *Journal of Multivariate Analysis*, Vol. 87, pp. 133–158.

[4] Biau, G., Bunea, F. and Wegkamp, M.H. (2005). Functional classification in Hilbert spaces, *IEEE Transactions on Information Theory*, Vol. 51, pp. 2163–2172.

[5] Bosq, D. (2000). *Linear Processes in Function Spaces – Theory and Applications*, Lecture Notes in Mathematics, Springer–Verlag, New York.

[6] Boucheron, S., Bousquet, O. and Lugosi, G. (2005). Theory of classification: A survey of some recent advances, *ESAIM: Probability and Statistics*, in press.

[7] Ciesielski, Z. (1961). Hölder conditions for the realizations of Gaussian processes, *Transactions of the American Mathematical Society*, Vol. 99, pp. 403–413.

[8] Cover, T.M. and Hart, P.E. (1965). Nearest neighbor pattern classification, *The Annals of Mathematical Statistics*, Vol. 36, pp. 1049–1051.

[9] Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*, Springer–Verlag, New York.

[10] Devroye, L. and Krzyżak, A. (1989). An equivalence theorem for L_1 convergence of the kernel regression estimate, *Journal of Statistical Planning and Inference*, Vol. 23, pp. 71–82.

[11] Diabo-Niang, S. and Rhomari, N. (2001). Nonparametric regression estimation when the regressor takes its values in a metric space, *University Paris VI, Technical Report*, http://www.ccr.jussieu.fr/lsta .

[12] Ferraty, F. and Vieu, P. (2000). Dimension fractale et estimation de la régression dans des espaces vectoriels semi-normés, *Comptes Rendus de l'Académie des Sciences de Paris*, Vol. 330, pp. 139–142.

[13] Ferraty, F. and Vieu, P. (2002). The functional nonparametric model and application to spectrometric data, *Computational Statistics*, Vol. 17, pp. 545–564.

[14] Ferraty, F. and Vieu, P. (2003). Curves discrimination: a nonparametric functional approach, *Computational Statistics and Data Analysis*, Vol. 44, pp. 161–173.

[15] Hall, P., Poskitt, D. and Presnell, B. (2001). A functional data-analytic approach to signal discrimination, *Technometrics*, Vol. 43, pp. 1–9.

[16] Kneip, A. and Gasser, T. (1992). Statistical tools to analyse data representing a sample of curves, *The Annals of Statistics*, Vol. 20, pp. 1266–1305.

[17] Kolmogorov, A.N. and Tihomirov, V.M. (1961). ε-entropy and ε-capacity of sets in functional spaces, *American Mathematical Society Translations, Series 2*, Vol. 17, pp. 277–364.

[18] Kulkarni, S.R. and Posner, S.E. (1995). Rates of convergence of nearest neighbor estimation under arbitrary sampling, *IEEE Transactions on Information Theory*, Vol. 41, pp. 1028–1039.

[19] Mattila, P. (1980). Differentiation of measures on uniform spaces, *Lecture Notes in Mathematics*, Vol. 794, pp. 261–283, Springer–Verlag, Berlin.

[20] McDiarmid, C. (1989). On the method of bounded differences, in *Surveys in Combinatorics 1989*, pp. 148–188, Cambridge University Press, Cambridge.

[21] Nadaraya, E.A. (1964). On estimating regression, *Theory of Probability and its Applications*, Vol. 9, pp. 141–142.

[22] Nadaraya, E.A. (1970). Remarks on nonparametric estimates for density functions and regression curves, *Theory of Probability and its Applications*, Vol. 15, pp. 134–137.

[23] Parzen, E. (1962). On estimation of a probability density function and mode, *The Annals of Mathematical Statistics*, Vol. 33, pp. 1065–1076.

[24] Preiss, D. and Tiser, J. (1982). Differentiation of measures on Hilbert spaces, *Lecture Notes in Mathematics*, Vol. 945, pp. 194–207, Springer–Verlag, Berlin.

[25] Ramsay, J.O. and Silverman, B.W. (1997). *Functional Data Analysis*, Springer–Verlag, New York.

[26] Rice, J.A. and Silverman, B.W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves, *Journal of the Royal Statistical Society, Series B*, Vol. 53, pp. 233–243.

[27] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function, *The Annals of Mathematical Statistics*, Vol. 27, pp. 832–837.

[28] Rudin, W. (1987). *Real and Complex Analysis – 3rd Edition*, McGraw-Hill, New York.

[29] Tiser, J. (1988). Differentiation theorem for Gaussian measures on Hilbert space, *Transactions of the American Mathematical Society*, Vol. 308, pp. 655–666.

[30] Van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes – With Applications to Statistics*, Springer–Verlag, New York.

[31] Watson, G.S. (1964). Smooth regression analysis, *Sankhyā, Series A*, Vol. 26, pp. 359–372.