Rates of Convergence of the Functional k-Nearest Neighbor Estimate

Gérard BIAU [a,1], Frédéric CÉROU [b] and Arnaud GUYADER [b,c]

[a] LSTA & LPMA
Université Pierre et Marie Curie – Paris VI
Boîte 158, 175 rue du Chevaleret
75013 Paris, France
gerard.biau@upmc.fr

[b] INRIA Rennes Bretagne Atlantique
Aspi project-team
Campus de Beaulieu, 35042 Rennes Cedex, France
Frederic.Cerou@irisa.fr

[c] Université Rennes 2 – Haute Bretagne
Campus Villejean
Place du Recteur Henri Le Moal, CS 24307
35043 Rennes Cedex, France
arnaud.guyader@uhb.fr

Abstract

Let $\mathcal{F}$ be a separable Banach space, and let $(\mathbf{X}, Y)$ be a random pair taking values in $\mathcal{F} \times \mathbb{R}$. Motivated by a broad range of potential applications, we investigate rates of convergence of the k-nearest neighbor estimate $r_n(\mathbf{x})$ of the regression function $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$, based on n independent copies of the pair $(\mathbf{X}, Y)$. Using compact embedding theory, we present explicit and general finite sample bounds on the expected squared difference $\mathbb{E}[r_n(\mathbf{X}) - r(\mathbf{X})]^2$, and particularize our results to classical function spaces such as Sobolev spaces, Besov spaces and reproducing kernel Hilbert spaces.

Index Terms — Regression estimation, Nearest neighbor estimate, Rates of convergence, Compact embedding, Reproducing kernel Hilbert space, Sobolev space.

AMS 2000 Classification: 62G05, 62G08.

[1] Corresponding author.

1 Introduction

Let $(\mathcal{F}, \|.\|)$ be a separable Banach space (possibly infinite-dimensional), and let $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ be independent $\mathcal{F} \times \mathbb{R}$-valued random variables with the same distribution as a generic pair $(\mathbf{X}, Y)$ such that $\mathbb{E}Y^2 < \infty$. In the regression function estimation problem, the goal is to estimate the regression function $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ using the data $\mathcal{D}_n$. With this respect, we will say that a regression estimate $r_n(\mathbf{x})$ is consistent if $\mathbb{E}[r_n(\mathbf{X}) - r(\mathbf{X})]^2 \to 0$ as $n \to \infty$.

In the classical statistical setting, each observation $\mathbf{X}_i$ is supposed to be a collection of numerical measurements represented by a d-dimensional vector. Thus, to date, most of the results pertaining to regression estimation have been reported in the finite-dimensional case, where it is assumed that $\mathcal{F}$ is the standard Euclidean space $\mathbb{R}^d$. We refer the reader to the monograph of Györfi, Kohler, Krzyżak and Walk [12] for a comprehensive introduction to the subject and an overview of most standard methods and developments in $\mathbb{R}^d$.

However, in an increasing number of practical applications, input data items are in the form of random functions (speech recordings, multiple time series, images...) rather than standard vectors, and this casts the regression problem into the general class of functional data analysis. Here, “random functions” means that the variable $\mathbf{X}$ takes values in a space $\mathcal{F}$ of functions on a subset of $\mathbb{R}^d$, equipped with an appropriate norm. For example, $\mathcal{F}$ could be the Banach space of continuous real functions on $\mathcal{X} = [0,1]^d$ with the norm

$$\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|,$$

but many other choices are possible. The challenge in this context is to infer the regression structure by exploiting the infinite-dimensional nature of the observations. The last few years have witnessed important developments in both the theory and practice of functional data analysis, and many traditional statistical tools have been adapted to handle functional inputs. The book of Ramsay and Silverman [14] provides a presentation of the area.

Interestingly, functional observations also arise naturally in the so-called kernel methods for general pattern analysis. These methods are based on the choice of a proper similarity measure, given by a positive definite kernel defined between pairs of objects of interest, to be used for inferring general types of relations. The key idea is to embed the observations at hand into a (typically infinite-dimensional) Hilbert space, called the feature space, and to compute inner products efficiently directly from the original data items using the kernel function. For an exhaustive presentation of kernel methodologies and related algorithms, we refer the reader to Schölkopf and Smola [15], and Shawe-Taylor and Cristianini [16].

Motivated by this broad range of potential applications, we propose, in the present contribution, to investigate rates of convergence properties of the k_n-nearest neighbor (k_n-NN) regression estimate, assuming that the $\mathbf{X}_i$'s take values in a general separable Banach space $(\mathcal{F}, \|.\|)$, typically infinite-dimensional. Recall that, for $\mathbf{x}$ in $\mathcal{F}$, the k_n-NN estimate is defined by

$$r_n(\mathbf{x}) = \frac{1}{k_n} \sum_{i=1}^{k_n} Y_{(i,n)}(\mathbf{x}),$$

where $(\mathbf{X}_{(1,n)}(\mathbf{x}), Y_{(1,n)}(\mathbf{x})), \ldots, (\mathbf{X}_{(n,n)}(\mathbf{x}), Y_{(n,n)}(\mathbf{x}))$ denotes a reordering of the data according to the increasing values of $\|\mathbf{X}_i - \mathbf{x}\|$ (ties are broken in favor of smallest indices). This procedure is one of the oldest approaches to regression analysis, dating back to Fix and Hodges [7, 8]. It is among the most popular nonparametric methods, with over 900 research articles published on the method since 1981 alone. For implementation, it requires only a measure of distance in the sample space, hence its popularity as a starting-point for refinement, improvement and adaptation to new settings (see for example Devroye, Györfi and Lugosi [5], Chapter 19).

Stone [17] proved the striking result that the estimate r_n is universally consistent if $\mathcal{F} = \mathbb{R}^d$, provided $k_n \to \infty$ and $k_n/n \to 0$. Here, "universally consistent" means that the method is consistent for all distributions of $(\mathbf{X}, Y)$ with $\mathbb{E}Y^2 < \infty$ (universally consistent regression estimates can also be obtained by other local averaging methods as long as $\mathcal{F} = \mathbb{R}^d$, see e.g. [12]). It turns out that the story is radically different in general spaces $\mathcal{F}$. In this respect, Cérou and Guyader [2] present counterexamples indicating that the estimate r_n is not universally consistent for general $\mathcal{F}$, and they argue that restrictions on $\mathcal{F}$ and the distribution of $(\mathbf{X}, Y)$ cannot be dispensed with. In short, $\mathcal{F}$ must be separable for the norm $\|.\|$, as already noticed by Cover and Hart [3], page 23. To see this, take for $(\mathcal{F}, \|.\|)$ the non-separable space of continuous functions from $]0, 1]$ to $[0, 1]$ equipped with the supremum norm $\|.\| = \|.\|_\infty$, and define the random function $\mathbf{X}$ as follows: let α be a $[0, 1]$-valued random variable with distribution $\mu = \frac{1}{2}\delta_0 + \frac{1}{2}\mathcal{U}_{[0,1]}$ (i.e., a mixture of a Dirac at 0 and the uniform distribution over $[0, 1]$) and let $\mathbf{X} : t \mapsto \sin(\alpha/t)$. Letting $Y = 0$ if $\mathbf{X} = 0$ and $Y = 1$ otherwise, the regression function has the form

$r(\mathbf{x}) = \mathbf{1}_{[\mathbf{x}\neq 0]}$. It is then easy to see that the k_n-NN regression estimate is not consistent for this distribution of $(\mathbf{X}, Y)$. And there is worse: even if the space $(\mathcal{F}, \|.\|)$ is separable, the estimate r_n may still have bad asymptotic behavior. Indeed, by working out arguments in Preiss [13], Cérou and Guyader [2] exhibit a random $\mathbf{X}$ with Gaussian distribution in a separable Hilbert space $\mathcal{F}$ for which the estimate r_n fails to be consistent. On the positive side, these authors provide a general condition, called the μ-continuity condition, which ensures the consistency of the estimate.

In this note, we go one step further in the analysis and study the rates of convergence of $\mathbb{E}[r_n(\mathbf{X}) - r(\mathbf{X})]^2$ as $n \to \infty$, when $\mathbf{X}$ is allowed to take values in the separable Banach space $\mathcal{F}$. This important question has been first addressed by Kulkarni and Posner [11], who put forward the essential role played by the covering numbers of the support of the distribution of $\mathbf{X}$. Building upon the ideas in [11] and exploiting recent advances on compact embeddings of functional Banach spaces, we present explicit and general finite sample upper bounds on $\mathbb{E}[r_n(\mathbf{X}) - r(\mathbf{X})]^2$, and particularize our results to classical function spaces such as Sobolev spaces, Besov spaces and reproducing kernel Hilbert spaces.

2 Rates of convergence

2.1 Bias-variance tradeoff

Setting

$$\tilde{r}_n(\mathbf{x}) = \frac{1}{k_n}\sum_{i=1}^{k_n} r\left(\mathbf{X}_{(i,n)}(\mathbf{x})\right),$$

we start the analysis with the standard variance/bias decomposition (Györfi, Kohler, Krzyżak and Walk [12])

$$\mathbb{E}\left[r_n(\mathbf{X}) - r(\mathbf{X})\right]^2 = \mathbb{E}\left[r_n(\mathbf{X}) - \tilde{r}_n(\mathbf{X})\right]^2 + \mathbb{E}\left[\tilde{r}_n(\mathbf{X}) - r(\mathbf{X})\right]^2. \qquad (2.1)$$

The first term is a variance term, which can be upper bounded independently of the topological structure of the space $\mathcal{F}$. Proof of the next proposition can be found for example in [12], Chapter 6 (here and throughout the document, the symbol $\mathbb{V}$ denotes variance):

Proposition 2.1 *Suppose that, for all* $\mathbf{x} \in \mathcal{F}$,

$$\sigma^2(\mathbf{x}) = \mathbb{V}[Y \mid \mathbf{X} = \mathbf{x}] \leq \sigma^2.$$

Then

$$\mathbb{E}\left[r_n(\mathbf{X}) - \tilde{r}_n(\mathbf{X})\right]^2 \leq \frac{\sigma^2}{k_n}.$$

The right-hand term in (2.1), which is a bias term, needs more careful attention. Let the symbol $\lfloor . \rfloor$ denote the integer part function. A quick inspection of the finite-dimensional proof (see [12], page 95) reveals the following result:

Proposition 2.2 *Suppose that, for all* $\mathbf{x}$ *and* $\mathbf{x}' \in (\mathcal{F}, \|.\|)$,

$$|r(\mathbf{x}) - r(\mathbf{x}')| \leq L\|\mathbf{x} - \mathbf{x}'\|, \tag{2.2}$$

for some positive constant L. *Then*

$$\mathbb{E}\left[\tilde{r}_n(\mathbf{X}) - r(\mathbf{X})\right]^2 \leq L^2\,\mathbb{E}\|\mathbf{X}_{(1,\lfloor \frac{n}{k_n} \rfloor)}(\mathbf{X}) - \mathbf{X}\|^2.$$

Putting Proposition 2.1 and Proposition 2.2 together, we obtain

$$\mathbb{E}\left[r_n(\mathbf{X}) - r(\mathbf{X})\right]^2 \leq \frac{\sigma^2}{k_n} + L^2\,\mathbb{E}\|\mathbf{X}_{(1,\lfloor \frac{n}{k_n} \rfloor)}(\mathbf{X}) - \mathbf{X}\|^2. \tag{2.3}$$

Thus, in order to bound the rate of convergence of $\mathbb{E}[r_n(\mathbf{X}) - r(\mathbf{X})]^2$, we need to analyze the rate of convergence of *the* nearest neighbor distance in the Banach space $\mathcal{F}$. As noticed in Kulkarni and Posner [11], this task can be achieved via the use of covering numbers of totally bounded sets (Kolmogorov and Tihomirov [10]). Some recalls are in order. Let $\mathcal{B}_{\mathcal{F}}(\mathbf{x}, \varepsilon)$ denote the open ball in $\mathcal{F}$ centered at $\mathbf{x}$ of radius ε.

Definition 2.1 *Let* $\mathcal{A}$ *be a subset of* $\mathcal{F}$. *The* ε*-covering number* $\mathcal{N}(\varepsilon)$ $[= \mathcal{N}(\varepsilon, \mathcal{A})]$ *is defined as the smallest number of open balls of radius* ε *that cover the set* $\mathcal{A}$. *That is*

$$\mathcal{N}(\varepsilon) = \inf\left\{ r \geq 1 \,:\, \exists\, \mathbf{x}_1, \ldots, \mathbf{x}_r \in \mathcal{F} \text{ such that } \mathcal{A} \subset \bigcup_{i=1}^{r} \mathcal{B}_{\mathcal{F}}(\mathbf{x}_i, \varepsilon) \right\}.$$

A set $\mathcal{A} \subset \mathcal{F}$ is said to be totally bounded if $\mathcal{N}(\varepsilon) < \infty$ for all $\varepsilon > 0$. In particular, any relatively compact set is totally bounded, and the converse assertion is true if the space $\mathcal{F}$ is complete. All totally bounded sets are bounded, and the converse assertion is satisfied when $\mathcal{F}$ is finite-dimensional. Figure 1 below illustrates this important concept in the finite-dimensional setting, with $(\mathcal{F}, \|.\|) = (\mathbb{R}^2, \|.\|_\infty)$ and $\mathcal{A} = (-1, 1)^2$.

As a function of ε, $\mathcal{N}(\varepsilon)$ is nonincreasing, piecewise-constant and right-continuous. The following discrete function, called the metric covering radius, can be interpreted as a pseudo-inverse of the function $\mathcal{N}(\varepsilon)$.

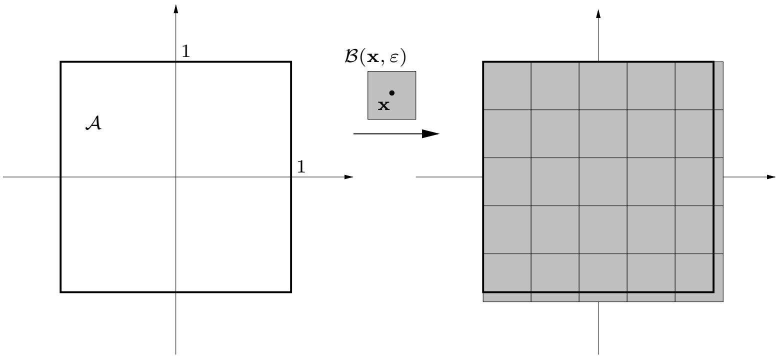

Figure 1: An ε-covering of $\mathcal{A} = (-1,1)^2$ in $(\mathbb{R}^2, \|.\|_\infty)$.

Definition 2.2 *Let $\mathcal{A}$ be a subset of $\mathcal{F}$. The metric covering radius $\mathcal{N}^{-1}(r)$ $[= \mathcal{N}^{-1}(r, \mathcal{A})]$ is defined as the smallest radius such that there exist r open balls of this radius which cover the set $\mathcal{A}$. That is*

$$\mathcal{N}^{-1}(r) = \inf \left\{ \varepsilon > 0 \, : \exists \, \mathbf{x}_1, \dots, \mathbf{x}_r \in \mathcal{F} \text{ such that } \mathcal{A} \subset \bigcup_{i=1}^{r} \mathcal{B}_{\mathcal{F}}(\mathbf{x}_i, \varepsilon) \right\} .$$

We note that $\mathcal{N}^{-1}(r)$ is a nonincreasing function of r (see Figure 2 for an illustration). Observe also that both $\mathcal{N}$ and $\mathcal{N}^{-1}$ are increasing with respect to the inclusion, that is $\mathcal{N}(\varepsilon, \mathcal{A}) \leq \mathcal{N}(\varepsilon, \mathcal{B})$ and $\mathcal{N}^{-1}(r, \mathcal{A}) \leq \mathcal{N}^{-1}(r, \mathcal{B})$ for $\mathcal{A} \subset \mathcal{B}$.

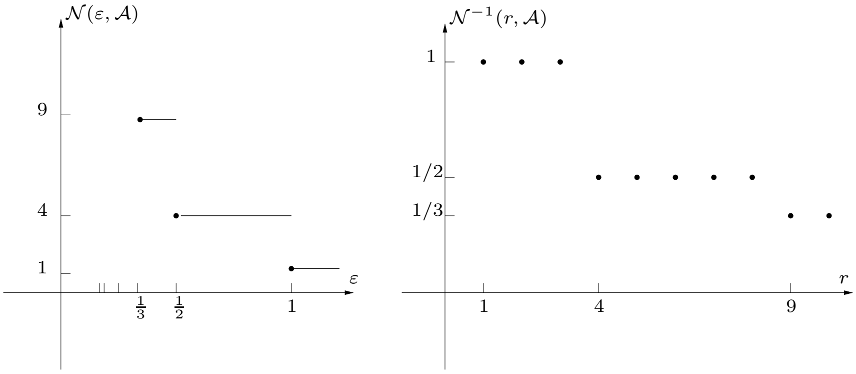

Figure 2: Covering numbers and covering radii of the set $\mathcal{A} = (-1,1)^2$ in $(\mathbb{R}^2, \|.\|_\infty)$.

Finally, we let the support $\mathcal{S}(\mu)$ of the probability measure μ of $\mathbf{X}$ be defined as the collection of all $\mathbf{x}$ with $\mu(\mathcal{B}_{\mathcal{F}}(\mathbf{x},\varepsilon)) > 0$ for all $\varepsilon > 0$. Throughout the paper, it will be assumed that $\mathcal{S}(\mu)$ is totally bounded. Observe then that $2\mathcal{N}^{-1}(1, \mathcal{S}(\mu))$ is an upper bound on the diameter of $\mathcal{S}(\mu)$.

Proposition 2.3 below bounds the convergence rate of the expected squared nearest neighbor distance in terms of the metric covering radii of $\mathcal{S}(\mu)$. This result sharpens the constant of Theorem 1, page 1032 in Kulkarni and Posner [11].

Proposition 2.3 *Let $\mathbf{X}_1, \ldots, \mathbf{X}_p$ be independent $\mathcal{F}$-valued random variables, distributed according to a common probability measure μ. Suppose that $\mathcal{S}(\mu)$ is a totally bounded subset of $(\mathcal{F}, \|.\|)$. Then*

$$\mathbb{E}\|\mathbf{X}_{(1,p)} - \mathbf{X}\|^2 \leq \frac{4}{p} \sum_{i=1}^{p} \left[\mathcal{N}^{-1}\left(i, \mathcal{S}(\mu)\right)\right]^2.$$

Proof of Proposition 2.3 All the covering and metric numbers we use in this proof are pertaining to the set $\mathcal{S}(\mu)$. Therefore, to lighten notation a bit, we set $\mathcal{N}(\varepsilon) = \mathcal{N}(\varepsilon, \mathcal{S}(\mu))$ and $\mathcal{N}^{-1}(r) = \mathcal{N}^{-1}(r, \mathcal{S}(\mu))$.

Let $\mathbf{X}'$ be a random variable distributed as and independent of $\mathbf{X}$ and let, for $\varepsilon > 0$,

$$F_{\mathbf{X}}(\varepsilon) = \mathbb{P}\left(\|\mathbf{X} - \mathbf{X}'\| \leq \varepsilon \,|\, \mathbf{X}\right)$$

be the conditional cumulative distribution function of the distance between $\mathbf{X}$ and $\mathbf{X}'$. Set finally

$$D_{(1)}(\mathbf{X}) = \|\mathbf{X}_{(1,p)}(\mathbf{X}) - \mathbf{X}\|.$$

Clearly,

$$\mathbb{P}\left(D^2_{(1)}(\mathbf{X}) > \varepsilon\right) = \mathbb{E}\left[\mathbb{P}(D_{(1)}(\mathbf{X}) > \sqrt{\varepsilon} \,|\, \mathbf{X})\right] = \mathbb{E}\left[\left(1 - F_{\mathbf{X}}(\sqrt{\varepsilon})\right)^p\right].$$

Next, take $\mathcal{B}_1, \ldots, \mathcal{B}_{\mathcal{N}(\sqrt{\varepsilon}/2)}$ a $\sqrt{\varepsilon}/2$-covering of $\mathcal{S}(\mu)$, and define an $\mathcal{N}(\sqrt{\varepsilon}/2)$-partition of $\mathcal{S}(\mu)$ as follows. For each $\ell = 1, \ldots, \mathcal{N}(\sqrt{\varepsilon}/2)$, let

$$\mathcal{P}_\ell = \mathcal{B}_\ell - \bigcup_{j=1}^{\ell-1} \mathcal{B}_j.$$

Then $\mathcal{P}_\ell \subset \mathcal{B}_\ell$ and

$$\bigcup_{\ell=1}^{\mathcal{N}(\sqrt{\varepsilon}/2)} \mathcal{B}_\ell = \bigcup_{\ell=1}^{\mathcal{N}(\sqrt{\varepsilon}/2)} \mathcal{P}_\ell,$$

with $\mathcal{P}_\ell \cap \mathcal{P}_{\ell'} = \emptyset$. Also,

$$\sum_{\ell=1}^{\mathcal{N}(\sqrt{\varepsilon}/2)} \mu(\mathcal{P}_\ell) = 1.$$

Thus, letting $p_\ell = \mu(\mathcal{P}_\ell)$, we may write

$$\begin{aligned} F_{\mathbf{X}}\left(\sqrt{\varepsilon}\right) &\geq \mathbb{P}(\exists\, \ell = 1, \ldots, \mathcal{N}(\sqrt{\varepsilon}/2) : \mathbf{X} \in \mathcal{P}_\ell \text{ and } \mathbf{X}' \in \mathcal{P}_\ell \,|\, \mathbf{X}) \\ &= \mathbb{E}\left[\sum_{\ell=1}^{\mathcal{N}(\sqrt{\varepsilon}/2)} \mathbf{1}_{[\mathbf{X}\in\mathcal{P}_\ell]}\mathbf{1}_{[\mathbf{X}'\in\mathcal{P}_\ell]} \,\middle|\, \mathbf{X}\right] \\ &= \sum_{\ell=1}^{\mathcal{N}(\sqrt{\varepsilon}/2)} p_\ell \mathbf{1}_{[\mathbf{X}\in\mathcal{P}_\ell]}. \end{aligned}$$

As a by-product, we remark that, for all $\varepsilon > 0$, $F_{\mathbf{X}}\left(\sqrt{\varepsilon}\right) > 0$ almost surely. Moreover

$$\mathbb{E}\left[\frac{1}{F_{\mathbf{X}}\left(\sqrt{\varepsilon}\right)}\right] \leq \mathbb{E}\left[\frac{1}{\sum_{\ell=1}^{\mathcal{N}(\sqrt{\varepsilon}/2)} p_\ell \mathbf{1}_{[\mathbf{X}\in\mathcal{P}_\ell]}}\right] = \mathbb{E}\left[\sum_{\ell=1}^{\mathcal{N}(\sqrt{\varepsilon}/2)} \frac{1}{p_\ell}\mathbf{1}_{[\mathbf{X}\in\mathcal{P}_\ell]}\right],$$

which leads to

$$\mathbb{E}\left[\frac{1}{F_{\mathbf{X}}\left(\sqrt{\varepsilon}\right)}\right] \leq \mathcal{N}\left(\frac{\sqrt{\varepsilon}}{2}\right).$$

Since $t(1-t)^p \leq t\exp(-pt) \leq \frac{1}{2p}$ for all $t \in [0,1]$, we deduce

$$\begin{aligned} \mathbb{P}(D^2_{(1)}(\mathbf{X}) > \varepsilon) &= \mathbb{E}\left[\left(1 - F_{\mathbf{X}}(\sqrt{\varepsilon})\right)^p\right] \\ &= \mathbb{E}\left[\frac{1}{F_{\mathbf{X}}\left(\sqrt{\varepsilon}\right)} F_{\mathbf{X}}(\sqrt{\varepsilon})\left(1 - F_{\mathbf{X}}(\sqrt{\varepsilon})\right)^p\right] \\ &\leq \frac{1}{2p}\,\mathbb{E}\left[\frac{1}{F_{\mathbf{X}}\left(\sqrt{\varepsilon}\right)}\right] \\ &\leq \frac{\mathcal{N}(\sqrt{\varepsilon}/2)}{2p}. \end{aligned}$$

Next, using the fact that $\mathbb{P}(D_{(1)}(\mathbf{X}) > \varepsilon) = 0$ for $\varepsilon \geq 2\mathcal{N}^{-1}(1)$, we may write

$$\begin{aligned} \mathbb{E}\left[D^2_{(1)}(\mathbf{X})\right] &= \int_0^\infty \mathbb{P}\left(D^2_{(1)}(\mathbf{X}) > \varepsilon\right) \mathrm{d}\varepsilon \\ &= \int_0^{4[\mathcal{N}^{-1}(1)]^2} \mathbb{P}\left(D^2_{(1)}(\mathbf{X}) > \varepsilon\right) \mathrm{d}\varepsilon \\ &= \int_0^{4[\mathcal{N}^{-1}(p)]^2} \mathbb{P}\left(D^2_{(1)}(\mathbf{X}) > \varepsilon\right) \mathrm{d}\varepsilon + \int_{4[\mathcal{N}^{-1}(p)]^2}^{4[\mathcal{N}^{-1}(1)]^2} \mathbb{P}\left(D^2_{(1)}(\mathbf{X}) > \varepsilon\right) \mathrm{d}\varepsilon. \end{aligned}$$

Thus

$$\begin{aligned}\mathbb{E}\left[D_{(1)}^2(\mathbf{X})\right] &\leq 4\left[\mathcal{N}^{-1}(p)\right]^2 + \frac{1}{2p}\int_{4[\mathcal{N}^{-1}(p)]^2}^{4[\mathcal{N}^{-1}(1)]^2} \mathcal{N}(\sqrt{\varepsilon}/2)\mathrm{d}\varepsilon \\ &= 4\left[\mathcal{N}^{-1}(p)\right]^2 + \frac{2}{p}\int_{[\mathcal{N}^{-1}(p)]^2}^{[\mathcal{N}^{-1}(1)]^2} \mathcal{N}(\sqrt{\varepsilon})\mathrm{d}\varepsilon \\ &= 4\left[\mathcal{N}^{-1}(p)\right]^2 + \frac{2}{p}\sum_{i=2}^{p}\int_{[\mathcal{N}^{-1}(i)]^2}^{[\mathcal{N}^{-1}(i-1)]^2} \mathcal{N}(\sqrt{\varepsilon})\mathrm{d}\varepsilon.\end{aligned}$$

Since $\mathcal{N}(\sqrt{\varepsilon}) = i$ for $\mathcal{N}^{-1}(i) \leq \sqrt{\varepsilon} < \mathcal{N}^{-1}(i-1)$, we obtain

$$\begin{aligned}\mathbb{E}\left[D_{(1)}^2(\mathbf{X})\right] &\leq 4\left[\mathcal{N}^{-1}(p)\right]^2 + \frac{2}{p}\sum_{i=2}^{p} i\left(\left[\mathcal{N}^{-1}(i-1)\right]^2 - \left[\mathcal{N}^{-1}(i)\right]^2\right) \\ &= \frac{4}{p}\left[\mathcal{N}^{-1}(1)\right]^2 + \frac{2}{p}\sum_{i=2}^{p-1}\left[\mathcal{N}^{-1}(i)\right]^2 + 2\left[\mathcal{N}^{-1}(p)\right]^2 \\ &\leq \frac{4}{p}\sum_{i=1}^{p}\left[\mathcal{N}^{-1}(i)\right]^2.\end{aligned}$$

To state the last inequality, recall that the sequence $(\mathcal{N}^{-1}(i))_{i\geq 1}$ is nonincreasing, so that

$$\left[\mathcal{N}^{-1}(p)\right]^2 \leq \frac{\sum_{i=2}^{p}\left[\mathcal{N}^{-1}(i)\right]^2}{p-1}.$$

The decomposition

$$\left[\mathcal{N}^{-1}(p)\right]^2 = \frac{p-1}{p}\left[\mathcal{N}^{-1}(p)\right]^2 + \frac{1}{p}\left[\mathcal{N}^{-1}(p)\right]^2$$

leads to the desired result. ■

Example 2.1 *Take* $(\mathcal{F}, \|.\|) = (\mathbb{R}^d, \|.\|_\infty)$ *and suppose that* $\mathcal{S}(\mu) \subset \mathcal{A} = (-1,1)^d$. *Then a moment's thought shows that*

$$\mathcal{N}(\varepsilon, \mathcal{A}) = \left(\frac{1}{\varepsilon}\right)^d \mathbf{1}_{[\varepsilon^{-1}\in\mathbb{N}]} + \left(\left\lfloor\frac{1}{\varepsilon}\right\rfloor + 1\right)^d \mathbf{1}_{[\varepsilon^{-1}\notin\mathbb{N}]}. \qquad (2.4)$$

In addition

$$\mathcal{N}^{-1}(i, \mathcal{A}) = i^{-\frac{1}{d}}\mathbf{1}_{[i^{1/d}\in\mathbb{N}]} + \left\lfloor i^{1/d}\right\rfloor^{-1}\mathbf{1}_{[i^{1/d}\notin\mathbb{N}]}.$$

Consequently, for $d \geq 3$, *by Proposition 2.3,*

$$\mathbb{E}\|\mathbf{X}_{(1,p)} - \mathbf{X}\|^2 \lesssim p^{-\frac{2}{d}},$$

where the notation $x \lesssim y$ means $x \leq Ay$ for some positive constant A. Combining this result with inequality (2.3), we conclude that

$$\mathbb{E}\left[r_n(\mathbf{X}) - r(\mathbf{X})\right]^2 \lesssim \frac{\sigma^2}{k_n} + L^2 \left\lfloor \frac{n}{k_n} \right\rfloor^{-\frac{2}{d}}.$$

Thus, for the choice $k_n \propto n^{\frac{2}{d+2}}$,

$$\mathbb{E}\left[r_n(\mathbf{X}) - r(\mathbf{X})\right]^2 \lesssim n^{-\frac{2}{d+2}}.$$

This shows that the nearest neighbor estimate is of optimal rate for the class of smooth distributions $(\mathbf{X}, Y)$ such that $\mathbf{X}$ has compact support, the regression function r is Lipschitz with constant L and, for all $\mathbf{x} \in \mathbb{R}^d$, $\mathbb{V}[Y \mid \mathbf{X} = \mathbf{x}] \leq \sigma^2$ (Ibragimov and Khasminskii [9] and Györfi, Kohler, Krzyżak and Walk [12], Chapter 3 and Theorem 6.2).

Example 2.1 strongly relies on the fact that bounded subsets of $(\mathbb{R}^d, \|.\|_\infty)$ are in fact totally bounded, as expressed by identity (2.4). Indeed, as shown in Proposition 2.3, a key step in obtaining rates of convergence for the nearest neighbor regression estimate is the derivation of covering numbers for the support of the distribution μ of $\mathbf{X}$. Unfortunately, in infinite-dimensional spaces, closed balls are bounded but not totally bounded, so that $\mathcal{N}^{-1}(i, \mathcal{S}(\mu)) = \infty$ most of the time and Proposition 2.3 is useless.

To correct this situation, a possible route is to assume that the observations we are dealing with behave in fact more regularly than a generic element of the ambient space $\mathcal{F}$, thereby reducing the general complexity of $\mathcal{S}(\mu)$. To illustrate this idea, suppose for example that $\mathcal{F}$ is the space $\mathcal{C}([0,1])$ of continuous real functions on $[0,1]$ equipped with the supremum norm $\|.\|_\infty$. Then, guided by the experience and practical considerations, it may be fair to suppose that the random curves $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are smooth enough, so that the support of their common distribution μ is in fact included and bounded in $\mathcal{D}^m([0,1])$, the space of m times differentiable functions with bounded derivatives, endowed with its canonical norm. Next, in this context, it can be proved that $\mathcal{N}^{-1}(i, \mathcal{S}(\mu)) < \infty$, and the show may go on. This example will be thoroughly discussed in the next section, together with other illustrations.

Thus, taking a general point of view, we will now suppose that the support of μ is bounded and included in a subspace $(\mathcal{G}, \|.\|_\mathcal{G})$ of $(\mathcal{F}, \|.\|)$, and that the embedding $(\mathcal{G}, \|.\|_\mathcal{G}) \hookrightarrow (\mathcal{F}, \|.\|)$ is compact. Here, "compact embedding" means that the unit ball (and thus, any ball) in $(\mathcal{G}, \|.\|_\mathcal{G})$ is totally bounded in $(\mathcal{F}, \|.\|)$. Put differently, balls in $\mathcal{G}$ (with respect to $\|.\|_{\mathcal{G}}$) become totally bounded as we see them as subsets of $\mathcal{F}$, endowed with the original metric $\|.\|$. The crux then is to identify covering numbers of balls in $\mathcal{G}$ with respect to the norm $\|.\|$. This will be the topic of the next section.

3 Compact embeddings

As we are now working with two different spaces, to avoid notational confusion we will rather denote by $\|.\|_{\mathcal{F}}$ the original norm of $\mathcal{F}$. Thus, in our context, $(\mathcal{G}, \|.\|_{\mathcal{G}})$ is a separable Banach subspace of $(\mathcal{F}, \|.\|_{\mathcal{F}})$ and, to simplify notation a bit, we let in the sequel $\mathcal{B}_{\mathcal{G}}(R)$ be the open ball in $(\mathcal{G}, \|.\|_{\mathcal{G}})$ of radius $R > 0$ centered at the origin, that is

$$\mathcal{B}_{\mathcal{G}}(R) = \{\mathbf{x} \in \mathcal{G} : \ \|\mathbf{x}\|_{\mathcal{G}} < R\}.$$

Definition 3.1 *The embedding $I : (\mathcal{G}, \|.\|_{\mathcal{G}}) \hookrightarrow (\mathcal{F}, \|.\|_{\mathcal{F}})$ is called compact if $I(\mathcal{B}_{\mathcal{G}}(1))$ is totally bounded in $(\mathcal{F}, \|.\|_{\mathcal{F}})$.*

Note that this definition is equivalent to require that the closure $\overline{I(\mathcal{B})}$ is compact for any bounded set $\mathcal{B} \subset \mathcal{G}$. It turns out that many interesting Banach spaces can be embedded into a larger functional space. To convince the reader, four examples are discussed below.

Example 3.1 (Differentiable functions) *Let $\mathcal{X}$ be a compact domain in $\mathbb{R}^d$ with smooth boundary. For every $m \in \mathbb{N}$, let $\mathcal{D}^m(\mathcal{X})$ be the Banach space of m times differentiable functions with bounded partial derivatives, that is*

$$\mathcal{D}^m(\mathcal{X}) = \left\{ f : \mathcal{X} \to \mathbb{R}, \ \|f\|_{\mathcal{D}^m} = \sum_{|\alpha| \leq m} \|D^\alpha f\|_\infty < \infty \right\},$$

where the sum is taken over all multi-indices $\alpha = (\alpha_1, \ldots, \alpha_d)$ such that $|\alpha| = \alpha_1 + \cdots + \alpha_d \leq m$. Then the inclusion

$$I_m : (\mathcal{G}, \|.\|_{\mathcal{G}}) = (\mathcal{D}^m(\mathcal{X}), \|.\|_{\mathcal{D}^m}) \hookrightarrow (\mathcal{F}, \|.\|_{\mathcal{F}}) = (C(\mathcal{X}), \|.\|_\infty)$$

is a compact embedding. Moreover, for every $\varepsilon > 0$ and $R > 0$,

$$\ln \mathcal{N}\left(\varepsilon, \overline{I_m\left(\mathcal{B}_{\mathcal{G}}(R)\right)}\right) \leq \left(\frac{RC}{\varepsilon}\right)^{\frac{d}{m}},$$

for some positive constant C independent of ε and R (Kolmogorov and Tihomirov [10]). This implies, for $i \in \mathbb{N}^\star$ and $R > 0$,

$$\mathcal{N}^{-1}\left(i, \overline{I_m\left(\mathcal{B}_{\mathcal{G}}(R)\right)}\right) \leq RC\left(\ln(i+1)\right)^{-\frac{m}{d}}.$$

Example 3.2 (Sobolev spaces) *Let again $\mathcal{X}$ be a compact domain in $\mathbb{R}^d$ with smooth boundary. For every $s \in \mathbb{N}$ and $p \geq 1$, let $W^{s,p}(\mathcal{X})$ be the usual Sobolev space equipped with the norm*

$$\|f\|_{W^{s,p}} = \sum_{|\alpha| \leq m} \|D^\alpha f\|_p.$$

The Rellich-Kondrakov Theorem asserts that, for $s_1 > s_2$, the inclusion

$$I_{s_1,s_2} : (\mathcal{G}, \|.\|_{\mathcal{G}}) = (W^{s_1,p}(\mathcal{X}), \|.\|_{W^{s_1,p}}) \hookrightarrow (\mathcal{F}, \|.\|_{\mathcal{F}}) = (W^{s_2,p}(\mathcal{X}), \|.\|_{W^{s_2,p}})$$

is compact. It can be proved (see for example Edmunds and Triebel [6], page 105) that for every $\varepsilon > 0$ and $R > 0$,

$$\ln \mathcal{N}\left(\varepsilon, \overline{I_{s_1,s_2}\left(\mathcal{B}_{\mathcal{G}}(R)\right)}\right) \leq \left(\frac{RC}{\varepsilon}\right)^{\frac{d}{s_1 - s_2}},$$

for some positive constant C independent of ε and R. This implies, for $s_1 > s_2$, $i \in \mathbb{N}^\star$ and $R > 0$,

$$\mathcal{N}^{-1}\left(i, \overline{I_{s_1,s_2}\left(\mathcal{B}_{\mathcal{G}}(R)\right)}\right) \leq RC \left(\ln(i+1)\right)^{-\frac{s_1 - s_2}{d}}.$$

This result can be extended to the more general context of Sobolev-type function spaces (Edmunds and Triebel [6]).

Example 3.3 (Besov spaces) *Let $\mathcal{X}$ be a compact domain in $\mathbb{R}^d$ with smooth boundary, and let $(B^s_{pq}(\mathcal{X}), \|.\|_{spq})$ be the Besov space on $\mathcal{X}$ (Edmunds and Triebel [6]). If $1 \leq p, q \leq \infty$ and $s > d/p$, then the inclusion*

$$I_s : (\mathcal{G}, \|.\|_{\mathcal{G}}) = \left(B^s_{pq}(\mathcal{X}), \|.\|_{spq}\right) \hookrightarrow (\mathcal{F}, \|.\|_{\mathcal{F}}) = (C(\mathcal{X}), \|.\|_\infty)$$

is compact. Besides, using a general result in [6], page 105, we have, for every $\varepsilon > 0$ and $R > 0$,

$$\ln \mathcal{N}\left(\varepsilon, \overline{I_s\left(\mathcal{B}_{\mathcal{G}}(R)\right)}\right) \leq \left(\frac{RC}{\varepsilon}\right)^{\frac{d}{s}},$$

and this gives raise to the bound

$$\mathcal{N}^{-1}\left(i, \overline{I_s\left(\mathcal{B}_{\mathcal{G}}(R)\right)}\right) \leq RC \left(\ln(i+1)\right)^{-\frac{s}{d}}.$$

As mentioned in [6], this inequality can be extended to compact embeddings of Besov-type function spaces.

Example 3.4 (Reproducing kernel Hilbert spaces) *Let $\mathcal{X}$ be a compact domain in $\mathbb{R}^d$, and let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Mercer kernel, i.e., K is continuous, symmetric and positive definite. Recall that we say that K is positive definite if for all finite sets $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \subset \mathcal{X}$, the $m \times m$ matrix $[K(\mathbf{x}_i, \mathbf{x}_j)]_{1 \leq i,j \leq m}$ is positive definite. Typical examples of Mercer kernels are the Gaussian kernel $K(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2)$ and the kernel $K(\mathbf{x}, \mathbf{x}') = (c^2 + \|\mathbf{x} - \mathbf{x}'\|^2)^{-\alpha}$ with $\alpha > 0$.*

For $\mathbf{x} \in \mathcal{X}$, let $K_{\mathbf{x}} = K(\mathbf{x}, .)$. According to Moore-Aronszajn's Theorem (Aronszajn [1]), there exists a unique Hilbert space $\mathcal{H}_K$ of functions on $\mathcal{X}$ satisfying the following conditions:

(i) For all $\mathbf{x} \in \mathcal{X}$, $K_{\mathbf{x}} \in \mathcal{H}_K$;

(ii) The span of the set $\{K_{\mathbf{x}} = K(\mathbf{x}, .),\ \mathbf{x} \in \mathcal{X}\}$ is dense in $\mathcal{H}_K$;

(iii) For all $f \in \mathcal{H}_K$, $f(\mathbf{x}) = \langle K_{\mathbf{x}}, f \rangle$.

The Hilbert space $\mathcal{H}_K$ is said to be the reproducing kernel Hilbert space (for short, RKHS) associated with the kernel K. It can be shown that $\mathcal{H}_K$ consists of continuous functions and, provided K is a $\mathcal{C}^\infty$ Mercer kernel, that the inclusion

$$I_K : (\mathcal{G}, \|.\|_{\mathcal{G}}) = (\mathcal{H}_K, \|.\|_K) \hookrightarrow (\mathcal{F}, \|.\|_{\mathcal{F}}) = (C(\mathcal{X}), \|.\|_\infty)$$

is a compact embedding (Cucker and Smale [4], Theorem D). Moreover, as proved in [4], for all $h > d$, $\varepsilon > 0$ and $R > 0$,

$$\ln \mathcal{N}\left(\varepsilon, \overline{I_K\left(\mathcal{B}_{\mathcal{G}}(R)\right)}\right) \leq \left(\frac{RC}{\varepsilon}\right)^{\frac{2d}{h}},$$

where C is a positive constant independent of ε and R. This readily implies that for $h > d$, $i \in \mathbb{N}^\star$ and $R > 0$,

$$\mathcal{N}^{-1}\left(i, \overline{I_K\left(\mathcal{B}_{\mathcal{G}}(R)\right)}\right) \leq RC\left(\ln(i+1)\right)^{-\frac{h}{2d}}.$$

This result has been improved by Zhou [18], who studies convolution-type kernels on $[0,1]^d$, i.e., kernels of form $K(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}' - \mathbf{x})$. Zhou provides estimates of $\ln \mathcal{N}(\varepsilon, \overline{I_K\left(\mathcal{B}_{\mathcal{G}}(R)\right)})$ depending on the decay of $\hat{k}$, the Fourier transform of k. For example, when $\hat{k}$ decays exponentially, one has

$$\ln \mathcal{N}\left(\varepsilon, \overline{I_K\left(\mathcal{B}_{\mathcal{G}}(R)\right)}\right) \leq C\left(\ln \frac{R}{\varepsilon}\right)^{d+1},$$

where C depends only on the kernel and the dimension. This implies

$$\mathcal{N}^{-1}\left(i, \overline{I_K\left(\mathcal{B}_{\mathcal{G}}(R)\right)}\right) \leq R \exp\left\{-\left(\frac{\ln(i+1)}{C}\right)^{\frac{1}{d+1}}\right\}.$$

This result can typically be applied to the Gaussian kernel.

Motivated by Examples 3.1-3.4 above, we shall impose the following set of assumptions on the distribution μ of $\mathbf{X}$:

$A1$ There exists a subspace $(\mathcal{G}, \|.\|_{\mathcal{G}})$ of $(\mathcal{F}, \|.\|_{\mathcal{F}})$ such that the support $\mathcal{S}(\mu)$ is bounded in $(\mathcal{G}, \|.\|_{\mathcal{G}})$, that is $\mathcal{S}(\mu) \subset \mathcal{B}_{\mathcal{G}}(R)$ for some positive constant R.

$A2$ There exists a compact embedding

$$I : (\mathcal{G}, \|.\|_{\mathcal{G}}) \hookrightarrow (\mathcal{F}, \|.\|_{\mathcal{F}}).$$

$A3$ There exists a function $\phi :]0, \infty[\to]0, \infty[$ such that

$$\left[\mathcal{N}^{-1}\left(i, \overline{I\left(\mathcal{B}_{\mathcal{G}}(R)\right)}\right)\right]^2 \leq \phi\left(\ln(i+1)\right), \quad i \in \mathbb{N}^\star,$$

where the covering number is taken with respect to $\|.\|_{\mathcal{F}}$.

The boundedness condition in assumption $A1$ is standard when establishing rates of convergence of nonparametric estimates, see e.g. Györfi, Kohler, Krzyżak and Walk [12]. As noticed in Theorem 7 of Kulkarni and Posner [11], this condition can be slightly relaxed, at the price of obtaining rates of convergence in probability.

Assumption $A2$ means that the balls in $\mathcal{G}$ (with respect to $\|.\|_{\mathcal{G}}$) are totally bounded as subsets of the space $(\mathcal{F}, \|.\|_{\mathcal{F}})$. This condition is not restrictive, and it is in particular satisfied by our leading Examples 3.1-3.4. From a practical perspective, we wish to emphasize that one usually has some latitude in choosing the space $\mathcal{G}$. This choice will typically be based on the regularity of the data (curves) to be processed. Roughly speaking, the smoother they are, the "smaller" the support of μ, and therefore the faster the convergence. On the other hand, we note that the Lipschitz condition in (2.2) needs to be valid in $(\mathcal{F}, \|.\|_{\mathcal{F}})$ — typically in $(\mathcal{C}(\mathcal{X}), \|.\|_\infty)$ — which is a stronger requirement than a Lipschitz condition in $(\mathcal{G}, \|.\|_{\mathcal{G}})$. To overcome this difficulty, we may decide to choose a "smaller" space $\mathcal{F}$, where the Lipschitz condition will be easier fulfilled. However, this operation may lead to slower rates of convergence, since they essentially depend on the difference of regularity (on the difference of "size" in some sense) between $(\mathcal{F}, \|.\|_{\mathcal{F}})$ and $(\mathcal{G}, \|.\|_{\mathcal{G}})$, as enlightened by Example 3.2.

Finally, and in view of the presented examples, the requirement $A3$ should be understood as a general notation, which will be crucial in the statement of Lemma 3.1 and Theorem 3.1 below.

We will need the following lemma.

Lemma 3.1 *Suppose that the function ϕ satisfies the following properties:*

- *(i) ϕ is nonincreasing and $\lim_{t\to\infty} t\phi(\ln t) = \infty$;*
- *(ii) ϕ is differentiable on $]0, \infty[$ and $\dfrac{\phi'(u)}{\phi(u)} \to 0$ as $u \to \infty$;*
- *(iii) One has $\displaystyle\int_1^\infty \phi(\ln t)\mathrm{d}t = \infty$.*

Then, as $p \to \infty$,

$$\frac{1}{p}\sum_{i=1}^{p} \phi(\ln i) \sim \phi(\ln p).$$

As a consequence, we have

$$\frac{1}{p}\sum_{i=1}^{p} \phi(\ln i) \lesssim \phi(\ln p).$$

Proof of Lemma 3.1 Since ϕ is a nonincreasing function satisfying (iii), we have, as $p \to \infty$,

$$\sum_{i=1}^{p} \phi(\ln i) \sim \int_1^p \phi(\ln t)\mathrm{d}t.$$

Moreover, by assumption (ii), as $t \to \infty$,

$$\phi(\ln t) \sim \phi(\ln t)\left(1 + \frac{\phi'(\ln t)}{\phi(\ln t)}\right).$$

Consequently, using (iii), we deduce that

$$\int_1^p \phi(\ln t)\,\mathrm{d}t \sim \int_1^p \phi(\ln t)\left(1 + \frac{\phi'(\ln t)}{\phi(\ln t)}\right)\mathrm{d}t.$$

The right term above may be simply evaluated as

$$\int_1^p \phi(\ln t)\left(1+\frac{\phi'(\ln t)}{\phi(\ln t)}\right)\mathrm{d}t = [t\phi(\ln t)]_1^p \sim p\phi(\ln p),$$

by assumption (i). Putting all pieces together, we conclude that

$$\frac{1}{p}\sum_{i=1}^{p}\phi(\ln i) \sim \phi(\ln p) \quad \text{as } p \to \infty.$$

■

We are now in a position to state the main result of the paper, which is a straightforward consequence of inequality (2.3), Proposition 2.3 and Lemma 3.1.

Theorem 3.1 *Suppose that assumptions A1-A3 are satisfied and that the function ϕ satisfies the conditions of Lemma 3.1. Suppose in addition that, for all $\mathbf{x}$ and $\mathbf{x}' \in \mathcal{F}$,*

$$\sigma^2(\mathbf{x}) = \mathbb{V}[Y \mid \mathbf{X} = \mathbf{x}] \le \sigma^2$$

and

$$|r(\mathbf{x}) - r(\mathbf{x}')| \le L\|\mathbf{x} - \mathbf{x}'\|_{\mathcal{F}},$$

for some positive constants σ^2 and L. Then

$$\mathbb{E}\left[r_n(\mathbf{X}) - r(\mathbf{X})\right]^2 \lesssim \frac{\sigma^2}{k_n} + L^2\phi\left(\ln\left\lfloor\frac{n}{k_n}\right\rfloor\right).$$

Theorem 3.1 can be illustrated in light of Examples 3.1-3.4. For differentiable functions (Example 3.1), we have $\phi(t) \propto t^{-2m/d}$, and the result reads

$$\mathbb{E}\left[r_n(\mathbf{X}) - r(\mathbf{X})\right]^2 \lesssim \frac{\sigma^2}{k_n} + L^2\left(\ln\left\lfloor\frac{n}{k_n}\right\rfloor\right)^{-\frac{2m}{d}}.$$

Therefore, with the choice $k_n \propto (\ln n)^{\frac{2m}{d}}$,

$$\mathbb{E}\left[r_n(\mathbf{X}) - r(\mathbf{X})\right]^2 \lesssim (\ln n)^{-\frac{2m}{d}}.$$

Similarly, in Sobolev spaces (Example 3.2), the choice $k_n \propto (\ln n)^{\frac{2(s_1-s_2)}{d}}$ leads to

$$\mathbb{E}\left[r_n(\mathbf{X}) - r(\mathbf{X})\right]^2 \lesssim (\ln n)^{-\frac{2(s_1-s_2)}{d}}.$$

In Besov spaces (Example 3.3), $\phi(t) \propto t^{-2s/d}$ and, with $k_n \propto (\ln n)^{\frac{2s}{d}}$, we obtain

$$\mathbb{E}\left[r_n(\mathbf{X}) - r(\mathbf{X})\right]^2 \lesssim (\ln n)^{-\frac{2s}{d}}.$$

Finally, in reproducing kernel Hilbert spaces (Example 3.4), attention shows that the choice $k_n \propto (\ln n)^{-\frac{h}{d}}$ results in

$$\mathbb{E}\left[r_n(\mathbf{X}) - r(\mathbf{X})\right]^2 \lesssim (\ln n)^{-\frac{h}{d}}.$$

For convolution-type kernels (Zhou [18]), the choice $k_n \propto \exp\left\{2\left(\frac{\ln n}{C}\right)^{\frac{1}{d+1}}\right\}$ implies

$$\mathbb{E}\left[r_n(\mathbf{X}) - r(\mathbf{X})\right]^2 \lesssim \exp\left\{-2\left(\frac{\ln n}{C}\right)^{\frac{1}{d+1}}\right\}.$$

The general finding here is that these rates of convergence are much slower than the traditional finite-dimensional rates (see Example 3.1). On the other hand, to the best of our knowledge, they are the first explicit available rates for the functional k_n-NN estimate. It is an open problem to know whether these rates are optimal over the smoothness classes we consider.

Acknowledgments. We thank the Associate Editor and two referees for valuable comments and insightful suggestions.

References

[1] Aronszajn, N. (1950). Theory of reproducing kernels, *Transactions of the American Mathematical Society*, **68**, 337-404.

[2] Cérou, F. and Guyader, A. (2006). Nearest neighbor classification in infinite dimension, *ESAIM: Probability and Statistics*, **10**, 340-355.

[3] Cover, T.M. and Hart, P.E. (1967). Nearest neighbor pattern classification, *IEEE Transactions on Information Theory*, **13**, 21-27.

[4] Cucker, F. and Smale, S. (2002). On the mathematical foundations of learning, *Bulletin of the American Mathematical Society*, **39**, 1-49.

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

[6] Edmunds, L.E. and Triebel, H. (1996). *Function Spaces, Entropy Numbers and Differential Operators*, Cambridge University Press, Cambridge.

[7] Fix, E. and Hodges, J.L. (1951). Discriminatory analysis. Nonparametric discrimination: Consistency properties, *Technical Report 4, Project Number 21-49-004, USAF School of Aviation Medicine*, Randolph Field, Texas.

[8] Fix, E. and Hodges, J.L. (1952). Discriminatory analysis: Small sample performance, *Technical Report 11, Project Number 21-49-004, USAF School of Aviation Medicine*, Randolph Field, Texas.

[9] Ibragimov, I.A. and Khasminskii, R.Z. (1981). *Statistical Estimation: Asymptotic Theory*, Springer-Verlag, New York.

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

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

[12] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*, Springer-Verlag, New York.

[13] Preiss, D. (1981). Gaussian measures and the density theorem, *Commentationes Mathematicae Universitatis Carolinae*, **1**, 181-193.

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

[15] Schölkopf, B. and Smola, A.J. (2002). *Learning with Kernels*, The MIT Press, Cambridge.

[16] Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*, Cambridge University Press, Cambridge.

[17] Stone, C.J. (1977). Consistent nonparametric regression, *The Annals of Statistics*, **5**, 595-645.

[18] Zhou, D.-X. (2002). The covering number in learning theory, *Journal of Complexity*, **18**, 739-767.