# On the Rate of Convergence of the Functional $k$-NN Estimates

Gérard Biau, Frédéric Cérou, Arnaud Guyader

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

# *On the Rate of Convergence of the Functional $k$-NN Estimates*

G. Biau — F. Cérou — A. Guyader

**N° 6861**

Février 2009

Thème NUM


ISSN 0249-6399 ISRN INRIA/RR--6861--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
RENNES - BRETAGNE ATLANTIQUE

# On the Rate of Convergence of the Functional $k$-NN Estimates

G. Biau* , F. Cérou† , A. Guyader‡

Thème NUM — Systèmes numériques
Équipe-Projet ASPI

Rapport de recherche n° 6861 — Février 2009 — 13 pages

**Abstract:** Let $\mathcal{F}$ be a general separable metric space and denote by $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ independent and identically distributed $\mathcal{F} \times \mathbb{R}$-valued random variables with the same distribution as a generic pair $(\mathbf{X}, Y)$. In the regression function estimation problem, the goal is to estimate, for fixed $\mathbf{x} \in \mathcal{F}$, the regression function $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ using the data $\mathcal{D}_n$. Motivated by a broad range of potential applications, we propose, in the present contribution, to investigate the properties of the so-called $k_n$-nearest neighbor regression estimate. We present explicit general finite sample upper bounds, and particularize our results to important function spaces, such as reproducing kernel Hilbert spaces, Sobolev spaces or Besov spaces.

**Key-words:** Regression, Nearest neighbors, Rates of convergence.

The author names appear in alphabetical order.

* LSTA & LPMA - Université Pierre et Marie Curie - Paris VI, Boîte 158, 175 rue du Chevaleret, 75013 Paris, France
† INRIA Rennes - Bretagne Atlantique, Campus de Beaulieu, 35042 Rennes Cedex, France
‡ Equipe de Statistique IRMAR, Université de Haute Bretagne, Place du Recteur H. Le Moal, CS 24307, 35043 Rennes Cedex, France

Centre de recherche INRIA Rennes – Bretagne Atlantique
IRISA, Campus universitaire de Beaulieu, 35042 Rennes Cedex
Téléphone : +33 2 99 84 71 00 — Télécopie : +33 2 99 84 71 71

# Sur la vitesse de convergence de l'estimateur fonctionnel des $k$ plus proches voisins

**Résumé :** Soit $\mathcal{F}$ un espace métrique séparable. On s'intéresse à l'estimation de la fonction de régression $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ associée à un couple aléatoire $(\mathbf{X}, Y)$ à valeurs dans $\mathcal{F}\times\mathbb{R}$, à partir d'un échantillon i.i.d. $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i)_{1\leq i\leq n}\}$ de même loi que $(\mathbf{X}, Y)$. Dans ce contexte, l'estimateur des $k$ plus proches voisins consiste à sélectionner dans $\mathcal{D}_n$ les $k$ plus proches voisins de $\mathbf{X}$ et à calculer la moyenne des $Y_i$ associés. Un nombre croissant d'applications portant sur des données $X_i$ fonctionnelles, on s'intéresse aux vitesses de convergence de la méthode des $k$ plus proches voisins dans ce contexte. On présente en particulier des bornes explicites pour des espaces fonctionnels $\mathcal{F}$ classiques : espaces de Hilbert à noyau reproduisant, espaces de Sobolev et espaces de Besov.

**Mots-clés :** Régression, Plus proches voisins, Vitesses de convergence.

# 1 Introduction

Let $\mathcal{F}$ be a general separable metric space equipped with metric $d$. Denote by $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ independent and identically distributed $\mathcal{F} \times \mathbb{R}$-valued random variables with the same distribution as a generic pair $(\mathbf{X}, Y)$ satisfying $\mathbb{E}|Y| < \infty$. In the regression function estimation problem, the goal is to estimate, for fixed $\mathbf{x} \in \mathcal{F}$, the regression function $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ using the data $\mathcal{D}_n$. With this respect, we say that a regression function 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, most of the results to date 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 [8] for a comprehensive introduction to the subject and an overview of most standard methods and developments.

However, in an increasing number of practical applications, input data items are in the form of random functions (speech recordings, times 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 compact subset of $\mathbb{R}^d$ with an appropriate norm. For example, $\mathcal{F}$ could be the Banach space of continuous 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 data analysis tools have been adapted to handle functional inputs. The book of Ramsay and Silverman [9] 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 (possibly 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 [10], and Shawe-Taylor and Cristianini [11].

Motivated by this broad range of potential applications, we propose, in the present contribution, to investigate the properties of the so-called $k_n$-nearest neighbor ($k_n$-NN) regression estimate, assuming that the $\mathbf{X}_i$'s take values in a general separable metric space $(\mathcal{F}, d)$, possibly 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 $(\mathbf{X}_{(1,n)}(\mathbf{x}), Y_{(1,n)}(\mathbf{x})), \ldots, (\mathbf{X}_{(1,n)}(\mathbf{x}), Y_{(1,n)}(\mathbf{x}))$ denotes a reordering of the data according to the increasing values of $d(\mathbf{x}, \mathbf{X}_i)$ (ties are broken in favor of smallest indices). This procedure is one of the oldest approaches to regression, dating back to Fix and Hodges [4, 5]. It is among the most popular nonparametric methods used in regression analysis, 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 [2], Chapter 19).

Stone [13] 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, regardless of the underlying distribution of $(\mathbf{X}, Y)$ (universally consistent regression estimates can also be obtained by other local averaging methods as long as $\mathcal{F} = \mathbb{R}^d$, see e.g. [8]). It turns out that the story is radically different in general metric spaces $\mathcal{F}$. In this respect, Cérou and Guyader [1] 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 regression function $r$ cannot be dispensed with.

In this paper, 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 a general separable space $\mathcal{F}$. This important question has been first addressed by Kulkarni and Posner [7], who put forward the essential role played by the covering numbers of the support of the distribution. Building upon the ideas in [7] and exploiting recent results of Smale [12] on metric entropy, we present explicit general finite sample upper bounds, and particularize our results to important function spaces, such as reproducing kernel Hilbert spaces, Sobolev spaces or Besov spaces.

## 2 Rates of convergence

Setting

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

we start the analysis with the standard variance/bias decomposition (Györfi, Kohler, Krzyżak and Walk [8])

$$\mathbb{E}\,|r_n(\mathbf{X}) - r(\mathbf{X})|^2 = \mathbb{E}\,|r_n(\mathbf{X}) - \tilde{r}_n(\mathbf{X})|^2 + \mathbb{E}\,|\tilde{r}_n(\mathbf{X}) - r(\mathbf{X})|^2\,. \qquad (1)$$

The first term is a variance term, which can be upper-bounded independently of the space $\mathcal{F}$. Proof of the next proposition can be found in [8], page 94:

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

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

*Then, for all* $\mathbf{x} \in \mathbb{R}^d$ *and all* $n \geq 1$,

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

The right-hand term in (1), which is a bias term, needs more careful attention. Let the symbol $\lfloor . \rfloor$ denote the integer part function. First, a quick inspection of the finite-dimensional proof ([8], page 95) reveals that

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

$$|r(\mathbf{x}) - r(\mathbf{x}')| \leq C\|\mathbf{x} - \mathbf{x}'\|,$$

*for some nonnegative constant* $C$. *Then*

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

Putting Proposition 2.1 and 2.2 together, we obtain

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

Thus, in order to bound the rate of convergence of $\mathbb{E}|r_n(\mathbf{X}) - r(\mathbf{X})|^2$, we need to analyse the rate of convergence of *the* nearest neighbor distance in a general separable metric setting. As noticed in Kulkarni and Posner [7], this task can be achieved via the use of covering numbers of totally bounded sets (Kolmogorov and Tihomirov [6]). Let $\mathcal{B}(\mathbf{x}, \varepsilon)$ denote the open ball in $\mathcal{F}$ centered at $\mathbf{x}$ of radius $\varepsilon$.

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

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

A set $\mathcal{A} \subset \mathcal{F}$ is said to be totally bounded if $\mathcal{N}(\varepsilon) < \infty$ for all $\varepsilon > 0$. In particular, every relatively compact set is totally bounded and all totally bounded sets are bounded. The converse assertions are not true in general. Figure 1 below illustrates this important concept in the finite-dimensional setting $(\mathcal{F}, d) = (\mathbb{R}^2, \|.\|_\infty)$ and $\mathcal{A} = (0,1)^2 = \mathcal{B}(0,1)$.

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

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

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

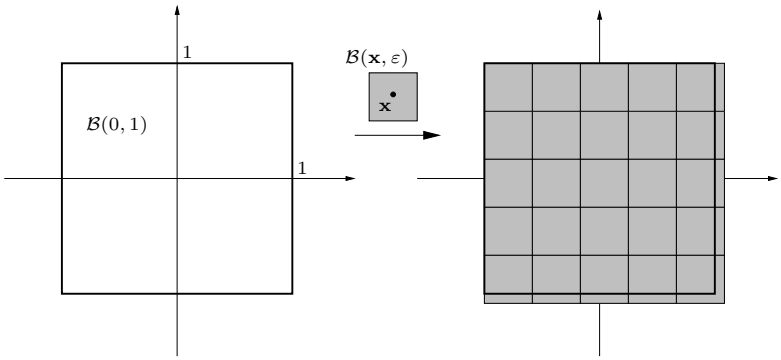


Figure 1: An $\varepsilon$-covering of $\mathcal{B}(0,1)$ in dimension 2.

We note that $\mathcal{N}^{-1}(r)$ is a non increasing function of $r$ (see Figure 2).

**Exemple 2.1.** *If* $(\mathcal{F},d)=(\mathbb{R}^d,\|.\|_\infty)$ *and* $\mathcal{A}=\mathcal{B}(0,1)=(-1,+1)^d$, *then*

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

*In addition*

$$\mathcal{N}^{-1}(i,\mathcal{A})=i^{-\frac{1}{d}}.$$

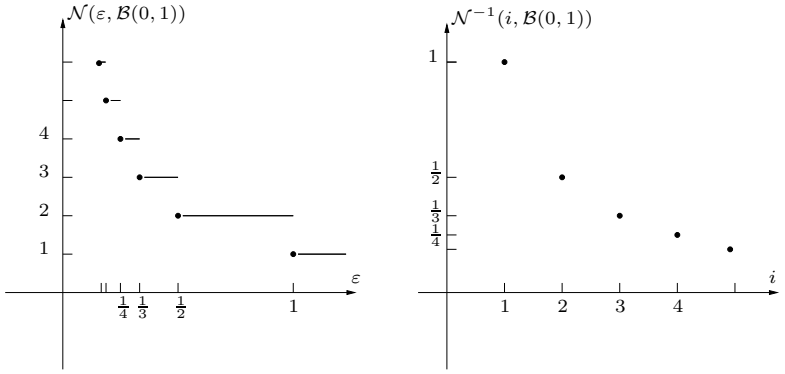


Figure 2: Covering numbers and covering radii for the unit ball $(-1,+1)$.

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

Proposition 2.3 below bounds the convergence rate of the expected nearest neighbor distance in terms of the metric covering radii of the support of the distribution $\mu$. It sharpens the constant of Theorem 1, page 1032 in Kulkarni and Posner [7].

**Proposition 2.3.** *Let $\mathbf{X}_1, \ldots, \mathbf{X}_N$ be i.i.d. according to a probability measure $\mu$ with $\mathcal{S}(\mu)$ a totally bounded subset of $(\mathcal{F}, d)$. Then*

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

**Proof.** All the covering and metric numbers we use in this proof are pertaining to the bounded 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 cumulative distribution function of the Euclidean distance between $\mathbf{X}$ and $\mathbf{X}'$ conditionally on $X$. Set finally

$$D_{(1)}(\mathbf{X}) = d\left(\mathbf{X}, \mathbf{X}_{(1,N)}(\mathbf{X})\right).$$

Clearly,

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

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,

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

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

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

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

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

which leads to

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

Since $p(1-p)^N \leq \frac{e^{-1}}{N+1} \leq \frac{1}{2N}$ for all $p \in [0,1]$, we can deduce

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

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

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

Since $\mathcal{N}(\sqrt{\varepsilon}) = i$ for $\mathcal{N}^{-1}(i) \leq \sqrt{\varepsilon} < \mathcal{N}^{-1}(i-1)$, we can upper bound this last quantity as follows

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

For the last inequality, recall that the sequence $(\mathcal{N}^{-1}(i))$ is non increasing, so that

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

Then the decomposition

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

leads to the desired result. ■

**Exemple 2.2.** *: if* $(\mathcal{F}, d) = (\mathbb{R}^d, \|.\|_\infty)$*,* $\mathcal{S}(X) \subset \mathcal{B}(0,1)$ *and* $d \geq 3$*, then:*

$$\mathbb{E}\|X_{(1,N)} - X\|_\infty^2 \leq \frac{4}{N}\sum_{i=1}^{N} i^{-\frac{2}{d}} \sim \frac{4d}{d-2}\ N^{-\frac{2}{d}}.$$

**Remark 2.1.** *: In the same way as for the proof of Theorem 2.3, one can easily show that*

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

*which sharpens also the constant of Theorem 1, page 1032 in Kulkarni and Posner [7].*

In our context, the goal is to estimate $\mathcal{N}^{-1}(i, \mathcal{S}(\mu))$ when $X$ takes values in a general metric space. Unfortunately, in infinite dimension, even if the support is bounded (e.g. $\mathcal{S}(\mu)$ contained in the unit ball), it will not be totally bounded, so that $\mathcal{N}^{-1}(i, \mathcal{S}(\mu)) = +\infty$ for all $i$ and the previous theorem is useless. The trick is then to use an auxiliary distance to measure the distances between $X$ and the training data. This is the key idea of compact embedding.

In the sequel, we assume that $(\mathcal{F}, \|.\|_\mathcal{F})$ and $(\mathcal{G}, \|.\|_\mathcal{G})$ are two general normed vector spaces. Moreover, $\mathcal{B}_\mathcal{F}(0,1)$ is the open unit ball with radius 1 in $(\mathcal{F}, \|.\|_\mathcal{F})$ :

$$\mathcal{B}_\mathcal{F}(0,1) = \{\mathbf{x} \in \mathcal{F} :\ \|\mathbf{x}\|_\mathcal{F} < 1\}.$$

**Definition 2.3.** *An inclusion* $I : (\mathcal{F}, \|.\|_\mathcal{F}) \hookrightarrow (\mathcal{G}, \|.\|_\mathcal{G})$ *is called a compact embedding if* $I(\mathcal{B}_\mathcal{F}(0,1))$ *is totally bounded in* $(\mathcal{G}, \|.\|_\mathcal{G})$*.*

We give now several examples of compact embeddings.

**1. Reproducing Kernel Hilbert Spaces.**
Let $\mathcal{X}$ be a compact domain in $\mathbb{R}^d$ with smooth boundary and $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a Mercer kernel, i.e. $K$ is continuous, symmetric and positive definite. This last property means that for all finite sets $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \subset \mathcal{X}$, the matrix $[K(\mathbf{x}_i, \mathbf{x}_j)]_{1 \leq i,j \leq m}$ is positive definite. The reproducing kernel Hilbert space $\mathcal{H}_K$ associated with the kernel $K$ is defined as the closure of the linear span of the set of functions $\{K_\mathbf{x} = K(\mathbf{x}, .),\ \mathbf{x} \in \mathcal{X}\}$ with the inner product satisfying

$$\forall \mathbf{x} \in \mathcal{X}, \forall f \in \mathcal{H}_K \qquad \langle K_\mathbf{x}, f\rangle_K = f(\mathbf{x}).$$

Let us also denote $(C(\mathcal{X}), \|.\|_\infty)$ the space of continuous functions on $\mathcal{X}$ equipped with the supremum norm. If $K$ is $\mathcal{C}^\infty$ then it has been proved by Cucker and Smale [12] that the inclusion

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

is a compact embedding. If $\mathcal{B}_R$ stands for the closed ball with radius $R$ in $(\mathcal{H}_K, \|.\|_K)$, it implies that the closure $\overline{I_K(\mathcal{B}_R)}$ is a compact subset of $(C(\mathcal{X}), \|.\|_\infty)$ and we can even obtain some upper-bounds on the covering numbers $\mathcal{N}(\varepsilon, \overline{I_K(\mathcal{B}_R)})$.

Indeed, Cucker and Smale have shown (see [12], Theorem D) that for $h > d$, $\varepsilon > 0$ and $R > 0$:

$$\ln \mathcal{N}(\varepsilon, \overline{I_K(\mathcal{B}_R)}) \leq \left(\frac{RC_h}{\varepsilon}\right)^{\frac{2d}{h}},$$

where $C_h$ is a constant independent of $\varepsilon$ and $R$. This readily implies that for $h > d$, $i \in \mathbb{N}^*$ and $R > 0$:

$$\mathcal{N}^{-1}(i, \overline{I_K(\mathcal{B}_R)}) \leq RC_h (\ln i)^{-\frac{h}{2d}}.$$

More recently, with stronger assumptions, Zhou [14] has even improved this result. For convolution type kernels $K(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}' - \mathbf{x})$ on $[0,1]^d$, he has indeed provided estimates depending on the decay of $\hat{k}$, the Fourier transform of $k$. For example, when $\hat{k}$ decays exponentially, we have:

$$\ln \mathcal{N}(\varepsilon, \overline{I_K(\mathcal{B}_R)}) \leq C_{k,d} \left(\ln \frac{R}{\varepsilon}\right)^{d+1},$$

where $C_{k,d}$ depends only on the kernel and on the dimension. This implies that:

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

In particular, this last result can be applied to the classical Gaussian kernel:

$$K(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}' - \mathbf{x}) = \exp\left\{-\frac{\|\mathbf{x}' - \mathbf{x}\|^2}{\sigma^2}\right\}.$$

**2. Sobolev Spaces.**

Here again $\mathcal{X}$ is a compact domain in $\mathbb{R}^d$ with smooth boundary. For all $m \in \mathbb{N}$ and $p \geq 1$, define the usual Sobolev space $W^{m,p}(\mathcal{X})$ equipped with the norm:

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

where the sum is over the multi-indexes $\alpha = (\alpha_1, \ldots, \alpha_d)$ such that $\alpha_1 + \cdots + \alpha_d \leq m$. The first results date back to Sobolev in 1938 and it is now well-known that for all $m > d/2$, the inclusion

$$I_m : (W^{m,p}(\mathcal{X}), \|.\|_{W^{m,p}}) \hookrightarrow (C(\mathcal{X}), \|.\|_\infty)$$

is a compact embedding. Moreover, with the same notations as above, it can be proved (see [3] p.105 and [12] Proposition 6) that for $\varepsilon > 0$ and $R > 0$:

$$\ln \mathcal{N}(\varepsilon, \overline{I_m(\mathcal{B}_R)}) \leq \left(\frac{RC_m}{\varepsilon}\right)^{\frac{d}{m}},$$

where $C_m$ is a constant independent of $\varepsilon$ and $R$. This implies that for $m > d/2$, $i \in \mathbb{N}^*$ and $R > 0$:

$$\mathcal{N}^{-1}(i, \overline{I_m(B_R)}) \leq RC_m (\ln i)^{-\frac{m}{d}}.$$

The result mentioned in [3] can in fact be applied in many more situations since they also give covering number estimates for a lot of classical "Sobolev type" function spaces.

**3. Besov Spaces.**
Here again $\mathcal{X}$ is a compact domain in $\mathbb{R}^d$ with smooth boundary. Without delving into the details of construction, let us just say that $(B^s_{pq}(\mathcal{X}), \|.\|_{spq})$ is the classical Besov space on $\mathcal{X}$. If $1 \leq p, q \leq \infty$ and $s > d/p$, then the inclusion

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

is a compact embedding. It can be deduced from a general result by Edmunds and Triebel (see [3] p.105) that for $\varepsilon > 0$ and $R > 0$:

$$\ln \mathcal{N}(\varepsilon, \overline{I_s(\mathcal{B}_R)}) \leq \left(\frac{RC_m}{\varepsilon}\right)^{\frac{d}{s}},$$

which gives raise to the same upper-bound as in Sobolev spaces :

$$\mathcal{N}^{-1}(i, \overline{I_s(\mathcal{B}_R)}) \leq RC_s(\ln i)^{-\frac{s}{d}}.$$

Here again, the result mentioned in [3] can be applied in many more situations since they give compact embeddings for a lot of classical "Besov type" function spaces.

In the general case, we suppose that the support $\mathcal{S}(\mu)$ is bounded in $(\mathcal{F}, d)$ and denote $R > 0$ such that $\mathcal{S}(\mu) \subset \mathcal{B}_R$. In all above mentioned cases, there exists a compact embedding :

$$I : (\mathcal{F}, d) \hookrightarrow (C(\mathcal{X}), \|.\|_\infty)$$

and we can upper-bound the covering radius :

$$\mathcal{N}^{-1}(i, \mathcal{S}(\mu)) \leq \mathcal{N}^{-1}(i, \overline{I(\mathcal{B}_R)}) \leq \phi(\ln i),$$

where $\phi$ is a function depending on the considered compact embedding. Anyway, in all these cases $\phi$ has some common features which lead to the following result.

**Lemma 2.1.** *If $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ is a function satisfying the assumptions :*

*(i) $\phi$ is non increasing and $\lim_{t\to+\infty} \phi(t) = 0$;*

*(ii) $\frac{\phi'(t)}{\phi(t)} \xrightarrow[t\to+\infty]{} 0$;*

*(iii) $\int_1^{+\infty} \phi(\ln u)\, du = +\infty$.*

*Then we have the following equivalence when $N$ goes to infinity :*

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

**Proof.** Since $\phi$ is a non increasing function and $\int_1^{+\infty} \phi(\ln u)\, du = +\infty$, we have the classical equivalence between sum and integral when $N$ goes to $+\infty$:

$$\sum_1^N \phi(\ln i) \sim \int_1^N \phi(\ln u)\, \mathrm{d}u$$

By the assumption $\lim_{u\to+\infty} \frac{\phi'(\ln u)}{\phi(\ln u)} = 0$, we have also the following equivalence when $u$ goes to $+\infty$

$$\phi(\ln u) \sim \phi(\ln u) \left(1 + \frac{\phi'(\ln u)}{\phi(\ln u)}\right),$$

and since $\int_1^{+\infty} \phi(\ln u)\, \mathrm{d}u = +\infty$, we can deduce that

$$\int_1^N \phi(\ln u)\, \mathrm{d}u \sim \int_1^N \phi(\ln u) \left(1 + \frac{\phi'(\ln u)}{\phi(\ln u)}\right) \mathrm{d}u,$$

but this last term is simply:

$$\int_1^N \phi(\ln u) \left(1 + \frac{\phi'(\ln u)}{\phi(\ln u)}\right) \mathrm{d}u = [u\phi(\ln u)]_1^N \sim N\phi(\ln N).$$

Putting all things together, we can conclude that:

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

■

We summarize now the assumptions on the model.

**Hypothesis 2.1** ($\mathcal{H}$)**.** *We make the following assumptions:*

*(i) The support $\mathcal{S}(\mu)$ is bounded in $(\mathcal{F}, \|.\|_{\mathcal{F}})$ : there exists an $R > 0$ such that $\mathcal{S}(\mu) \subset \mathcal{B}_{\mathcal{F}}(0, R)$;*

*(ii) There exists a compact embedding*

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

*and a function $\phi$ satisfying the assumptions of Lemma 2.1 so that for all $i \in \mathbb{N}^\star$:*

$$\mathcal{N}^{-1}(i, \overline{I(\mathcal{B}_R)}) \leq \phi(\ln i);$$

*(iii) There exists an $\alpha \in \mathbb{R}$ such that:*

$$\forall (\mathbf{x}, \mathbf{x}') \in \mathcal{F} \times \mathcal{F} \qquad |r(\mathbf{x}') - r(\mathbf{x})| \leq C \|\mathbf{x}' - \mathbf{x}\|_{\mathcal{G}}.$$

Finally, we can give a complete overview of the preceding results.

**Theorem 2.1.** *Under assumption ($\mathcal{H}$), we have :*

$$\mathbb{E}\, |r_n(\mathbf{X}) - r(\mathbf{X})|^2 \leq \frac{\sigma^2}{k_n} + C^2\ \phi(\ln \lfloor \frac{n}{k_n} \rfloor),$$

# References

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

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

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

[4] 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.

[5] 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.

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

[7] 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.

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

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

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

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

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

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

[14] Zhou, D.-X (2002). The Covering Number in Learning Theory, *Journal of Complexity*, **18**, 739-767.

Centre de recherche INRIA Rennes – Bretagne Atlantique
IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399