On the Rate of Convergence of the Bagged Nearest Neighbor Estimate

Gérard Biau, Frédéric Cérou, Arnaud Guyader

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

On the Rate of Convergence of the Bagged Nearest Neighbor Estimate

G. Biau — F. Cérou — A. Guyader

N° 6860

Février 2009

Thème NUM

apport de recherche

ISSN 0249-6399 ISRN INRIA/RR--6860--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 Bagged Nearest Neighbor Estimate

G. Biau* , F. Cérou† , A. Guyader‡

Thème NUM — Systèmes numériques
Équipe-Projet ASPI

Rapport de recherche n° 6860 — Février 2009 — 29 pages

Abstract: Bagging is a simple way to combine estimates in order to improve their performance. This method, suggested by Breiman in 1996, proceeds by resampling from the original data set, constructing a predictor from each bootstrap sample, and decide by combining. By bagging an n-sample, the crude nearest neighbor regression estimate is turned out into a consistent weighted nearest neighbor regression estimate, which is amenable to statistical analysis. Letting the resampling size k_n grows with n in such a manner that $k_n \to \infty$ and $k_n/n \to 0$, it is shown that this estimate achieves optimal rates of convergence, independently from the fact that resampling is done with or without replacement.

Key-words: Bagging, Resampling, 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 du plus proche voisin baggé

Résumé : 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 $\mathbb{R}^d \times \mathbb{R}$, 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 dit du plus proche voisin baggé consiste à tirer dans $\mathcal{D}_n$ un très grand nombre de sous-échantillons bootstrap indépendants de taille $k_n < n$, à considérer pour chacun d'eux l'estimateur du plus proche voisin et enfin à calculer la moyenne de tous ces estimateurs pour prédire. Il a récemment été prouvé par Biau et Devroye (2008) que l'estimateur ainsi obtenu est universellement convergent à condition que k_n tende vers l'infini et k_n/n tende vers 0 lorsque $n \to \infty$. Nous montrons dans ce travail que, mieux encore, la vitesse de convergence de cet estimateur baggé est optimale.

Mots-clés : Bootstrap, Rééchantillonnage, Plus proches voisins, Vitesse de convergence.

1 Introduction

1.1 Bagging

Ensemble methods are popular machine learning algorithms which train multiple learners and combine their predictions. The success of ensemble algorithms on many benchmark data sets has raised considerable interest in understanding why such methods succeed and identifying circumstances in which they can be expected to produce good results. It is now well-known that the generalization ability of an ensemble can be significantly better than that of a single predictor, and ensemble learning has therefore been a hot topic during the past years. For a comprehensive review of the domain, we refer the reader to Dietterich [6] and the references therein.

One of the first and simplest ways to combine predictors in order to improve their performance is bagging (**b**ootstrap **agg**regat**ing**), suggested by Breiman [2]. This ensemble method proceeds by generating bootstrap samples from the original data set, constructing a predictor from each bootstrap sample, and decide by combining. It is one of the most effective computationally intensive procedures to improve on unstable estimators or classifiers, especially for large, high dimensional data set problems where finding a good model or classifier in one step is impossible because of the complexity and scale of the problem. Bagging has attracted much attention and is frequently applied, although its statistical mechanisms are not yet fully understood and are still under active investigation. Recent theoretical contributions to bagging and related methodologies include those of Friedman and Hall [9], Bühlmann and Yu [3], Hall and Samworth [11], Buja and Stuetzle [4], and Biau and Devroye [1].

It turns out that Breiman's bagging principle has a simple application in the context of nearest neighbor methods. Nearest neighbor predictors are one of the oldest approach to regression and classification, dating back to Fix and Hodges [7, 8]. A major attraction of nearest neighbor procedures is their simplicity. For implementation, they require only a measure of distance in the sample space, along with samples of training data, hence their popularity as a starting point for refinement, improvement and adaptation to new settings (see for example Devroye, Györfi and Lugosi [5], Chapter 19). Before we formalize the link between bagging and nearest neighbors, some definitions are in order. Throughout the paper, we suppose that we are given a sample $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ of i.i.d. $\mathbb{R}^d \times \mathbb{R}$-valued random variables with the same distribution as a generic pair $(\mathbf{X}, Y)$ satisfying $\mathbb{E}|Y| < \infty$. The space $\mathbb{R}^d$ is equipped with the standard Euclidean metric. For fixed $\mathbf{x} \in \mathbb{R}^d$, our mission 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 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 $n \to \infty$. It is universally consistent if this property is true for all distributions of $(\mathbf{X}, Y)$ with $\mathbb{E}[Y^2] < \infty$.

1.2 Bagging and nearest neighbors

Recall that the 1-nearest neighbor (1-NN) regression estimate sets $r_n(\mathbf{x}) = Y_{(1)}(\mathbf{x})$ where $Y_{(1)}(\mathbf{x})$ is the observation of the feature vector $\mathbf{X}_{(1)}(\mathbf{x})$ whose

Euclidean distance to $\mathbf{x}$ is minimal among all $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Ties are broken in favor of smallest indices. Contrary to some beliefs, it is clearly not, in general, a consistent estimate (Devroye, Gÿorfi and Lugosi [5], Chapter 5). However, by bagging, one may turn the 1-NN estimate into a consistent one, provided that the size of the bootstrap sample is sufficiently small.

We proceed as follows, via a randomized basic regression estimate r_{k_n} in which $1 \leq k_n \leq n$ is a parameter. The elementary predictor r_{k_n} is the 1-NN rule for a random subsample $\mathcal{S}_n$ drawn with (or without) replacement from $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$, with $\mathrm{Card}(\mathcal{S}_n) = k_n$. We apply bagging, that is, we repeat the random sampling an infinite number of times, and take the average of the individual outcomes. Thus, theoretically, the bagged regression estimate $r_n^\star$ is defined by

$$r_n^\star(\mathbf{x}) = \mathbb{E}^\star \left[r_{k_n}(\mathbf{x}) \right],$$

where $\mathbb{E}^\star$ denotes expectation with respect to the resampling distribution, conditionally on the data set $\mathcal{D}_n$. In practice, the resampling distribution is implemented by Monte-Carlo: we repeat the random sampling m times, and take the average of the individual outcomes. Formally, if $Z_t = r_{k_n}(\mathbf{x})$ is the prediction in the t-th round of bagging, we let the bagged regression estimate be approximately

$$r_n^\star(\mathbf{x}) \approx \frac{1}{T} \sum_{t=1}^{T} Z_t,$$

where $Z_1, \ldots, Z_T$ are the outcomes in the individual rounds. However, for definiteness, we shall always treat in this paper the version of the bagged nearest neighbor predictor which uses an infinite number of simulations in the aggregation step.

The next result, proved in [1], shows that for an appropriate choice of k_n, the bagged version of the 1-NN regression estimate is universally consistent:

Theorem 1.1. *If $k_n \to \infty$ and $k_n/n \to 0$, then $r_n^\star$ is universally consistent.*

In this theorem, the fact that resampling is done with or without replacement is irrelevant. Thus, by bagging, one may turn the crude 1-NN procedure into a consistent one, provided the size of the resamples is sufficiently small. To understand the statistical forces driving Theorem 1.1, recall that if we let $V_1 \geq V_2 \geq \ldots \geq V_n \geq 0$ denote weights (depending on n) that sum to one, and $V_1 \to 0$, $\sum_{i>\varepsilon n} V_i \to 0$ for all $\varepsilon > 0$ as $n \to \infty$, then the regression estimate

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

with $(\mathbf{X}_{(1)}(\mathbf{x}), Y_{(1)}(\mathbf{x})), \ldots, (\mathbf{X}_{(n)}(\mathbf{x}), Y_{(n)}(\mathbf{x}))$ the reordering of the data such that

$$\|\mathbf{x} - \mathbf{X}_{(1)}(\mathbf{x})\| \leq \ldots \leq \|\mathbf{x} - \mathbf{X}_{(n)}(\mathbf{x})\|$$

is called the weighted nearest neighbor regression estimate. It is known to be universally consistent (Stone [15], and Problems 11.7, 11.8 of Devroye, Györfi

and Lugosi [5]). The crux to prove Theorem 1.1 is to observe that $r_n^\star$ is in fact a weighted nearest neighbor estimate with

$$V_i = \mathbb{P}(i\text{-th nearest neighbor of } \mathbf{x} \text{ is the 1-NN in a random selection}).$$

Then, a moment's thought shows that for the "without replacement" sampling, V_i is hypergeometric:

$$V_i = \begin{cases} \dfrac{\binom{n-i}{k_n-1}}{\binom{n}{k_n}}, & i \le n-k_n+1 \\ 0, & i > n-k_n+1, \end{cases}$$

whereas for sampling "with replacement",

$$V_i = \left(1-\frac{i-1}{n}\right)^{k_n} - \left(1-\frac{i}{n}\right)^{k_n}.$$

The core of the proof of Theorem 1.1 is then to show that, in both cases, the weights V_i satisfy the conditions $V_1 \to 0$ and $\sum_{i>\varepsilon n} V_i \to 0$ for all $\varepsilon > 0$ as $n \to \infty$. These bagging weights have been independently exhibited by Steele [14], who shows on practical examples that substantial reductions in prediction error are possible under bootstrap sub-sampling.

In the present paper, we go one step further in the analysis and study the rates of convergence of $\mathbb{E}\left[r_n^\star(\mathbf{x}) - r(\mathbf{x})\right]^2$ as $n \to \infty$. Building upon [1], we will distinguish between the "with replacement" (Section 2) and "without replacement" cases (Section 3). In both cases, we show that, for $d \ge 3$, the estimate $r^\star$ is of optimum rate for the class of distributions of $(\mathbf{X}, Y)$ such that $\mathbf{X}$ has compact support and the regression function r is Lipschitz. We wish to emphasize that all the results are obtained by letting the resampling size k_n grows with n in such a manner that $k_n \to \infty$ and $k_n/n \to 0$. These results are of interest because the majority of bagging experiments employ relatively large resample sizes. In fact, much of the evidence *against* the performance of bagged nearest nearest neighbor methods is for full sample size resamples (see the discussion in Breiman [2], Paragraph 6.4), except the notable results of Hall and Samworth [11] and Steele [14], who also report encouraging numerical calculations in the context of regression and classification.

2 Bagging "with replacement"

This bagging-type is sometimes called moon-bagging, standing for **m o**ut **o**f **n** **b**oostrap **agg**regat**ing**. In this case, the bagged 1-NN regression estimate takes the form

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

where

$$V_i = \left(1-\frac{i-1}{n}\right)^{k_n} - \left(1-\frac{i}{n}\right)^{k_n}.$$

Here and throughout the document, the symbol $\mathbb{V}$ denotes variance and $\Gamma(t)$ is the Gamma function

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} \mathrm{d}x, \quad t > 0.$$

Remind that if p is a positive integer, then $\Gamma(p) = (p-1)!$, with the convention $0! = 1$.

Let us now recall various topological definitions that will be used in the paper. We first define the well-known notion of covering numbers which characterize the massiveness of a set (Kolmogorov and Tihomirov [13]). As put forward in Kulkarni and Posner [12], these quantities play a key role is the context of nearest neighbor analysis. Let $\mathcal{B}(\mathbf{x}, \varepsilon)$ denote the open Euclidean ball in $\mathbb{R}^d$ centered at $\mathbf{x}$ of radius ε.

Definition 2.1. *Let $\mathcal{A}$ be a subset of $\mathbb{R}^d$. 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*

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

A set $\mathcal{A} \subset \mathbb{R}^d$ is bounded if and only if $\mathcal{N}(\varepsilon) < \infty$ for all $\varepsilon > 0$. Note that as a function of ε, $\mathcal{N}(\varepsilon)$ 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 $\mathcal{N}^{-1}(r)$ $[= \mathcal{N}^{-1}(r, A)]$ is defined as the smallest radius such that there exist r balls of this radius which cover the set $\mathcal{A}$. That is*

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

We note that $\mathcal{N}^{-1}(r)$ is a non increasing discrete function of r. Finally, we let the support $\mathcal{S}(\mu)$ of a probability measure μ be defined as the collection of all $\mathbf{x}$ with $\mu(\mathcal{B}(\mathbf{x}, \varepsilon)) > 0$ for all $\varepsilon > 0$.

From now on, we denote by μ the distribution of $\mathbf{X}$. Throughout the paper, it will be asumed that $\mathbf{X}$ is bounded. Letting $\rho = \mathcal{N}^{-1}(1, \mathcal{S}(\mu))$, we observe that 2ρ is an upper bound of the diameter of the support of μ. We are now in a position to state the main result of this section.

Theorem 2.1. *Suppose that $\mathbf{X}$ is bounded, and set $\rho = \mathcal{N}^{-1}(1, \mathcal{S}(\mu))$. Suppose in addition that, for all $\mathbf{x}$ and $\mathbf{x}' \in \mathbb{R}^d$,*

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

and

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

for some nonnegative constants σ^2 and C. Then

(*i*) *If* $d = 1$,

$$\mathbb{E}\left[r_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 \leq \frac{2\sigma^2 k_n}{n}\left(1+\frac{1}{n}\right)^{2k_n} + \frac{32C^2\rho^2}{k_n}\left(1+\frac{1}{n}\right)^{k_n}.$$

(*ii*) *If* $d = 2$,

$$\mathbb{E}\left[r_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 \leq \frac{2\sigma^2 k_n}{n}\left(1+\frac{1}{n}\right)^{2k_n} + \frac{16C^2\rho^2}{k_n}\left(1+\frac{1}{n}\right)^{k_n}\left[1+\ln(k_n+1)\right].$$

(*iii*) *If* $d \geq 3$,

$$\mathbb{E}\left[r_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 \leq \frac{2\sigma^2 k_n}{n}\left(1+\frac{1}{n}\right)^{2k_n} + \frac{8C^2\rho^2}{1-2/d}\left[\frac{1}{n^{k_n}} + \alpha_d\left(1+\frac{1}{n}\right)^{k_n} {k_n}^{-2/d}\right],$$

where

$$\alpha_d = 2\Gamma\left(\frac{d-2}{d}\right)\Gamma\left(\frac{d+2}{d}\right).$$

By balancing the variance and bias terms in the theorem above, we are led to the following corollary.

Corollary 2.1. *Under the assumptions of Theorem 2.1, we have:*

(*i*) *If* $d = 1$, *for* $k_n = (4C\rho/\sqrt{\sigma^2 e})n^{1/2}$,

$$\mathbb{E}\left[r_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 \leq \frac{\Lambda}{\sqrt{n}},$$

where $\Lambda = 16e^{3/2}\sigma C\rho$.

(*ii*) *If* $d = 2$, *for* $k_n = (2\sqrt{2}C\rho/\sqrt{\sigma^2 e})\sqrt{n(1+\ln n)}$,

$$\mathbb{E}\left[r_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 \leq \Lambda\sqrt{\frac{1+\ln n}{n}},$$

where $\Lambda = 8\sqrt{2}e^{3/2}\sigma C\rho$.

(*iii*) *If* $d \geq 3$, *for* $k_n = [4(1+e\alpha_d C^2\rho^2)/(\sigma^2 e^2(1-2/d))]^{d/(d+2)}\, n^{\frac{d}{d+2}}$,

$$\mathbb{E}\left[r_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 \leq \Lambda C^{\frac{2d}{d+2}}\, n^{-\frac{2}{d+2}},$$

where

$$\Lambda = 4\sigma^2 e^2\left[\frac{4(1+e\alpha_d)\rho^2}{\sigma^2 e^2(1-2/d)}\right]^{d/(d+2)}.$$

Two important remarks are in order.

1. First, we note that, for $d \geq 3$ and a suitable choice of k_n, the bagged nearest neighbor estimate achieves both the minimax $n^{-2/(d+2)}$ rate and the optimal constant $C^{2d/(d+2)}$ on the class of distributions of $(\mathbf{X}, Y)$ such that $\mathbf{X}$ has compact support and the regression function r is Lipschitz with Lipschitz constant C (Györfi, Kohler, Krzyżak and Walk [10], Chapter 3). Seconds, the bounds are valid for finite sample sizes, so that we are in fact able to approach the minimax lower bound not only asymptotically but even for finite sample sizes. On the other hand, the estimate with the optimal rate of convergence depends on the unknown distribution of $(\mathbf{X}, Y)$, and especially on the covering radius ρ and the smoothness of the regression function measured by the constant C. It is an open question whether there exists an optimal adaptive (i.e., data-dependent) choice of k_n which still achieves the optimal rate of convergence over traditional classes of regression functions. Preliminary empirical results for cross-validation are reported in Hall and Samworth [11].

2. For $d = 1$, the obtained rate is not optimal, whereas it is optimal up to a *log* term for $d = 2$. This low-dimensional phenomenon is also true for the standard k_n-nearest neighbor regression estimate, which does not achieve the optimal rates in dimensions 1 and 2 (see Problem 6.1 and Problem 6.7 in [10], Chapter 3).

Setting

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

the proof of Theorem 2.1 will rely on the variance/bias decomposition

$$\mathbb{E}\left[r_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 = \mathbb{E}\left[r_n^\star(\mathbf{X}) - \tilde{r}_n^\star(\mathbf{X})\right]^2 + \mathbb{E}\left[\tilde{r}_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2. \tag{1}$$

The first term is treated in Proposition 2.1 and the second one in Proposition 2.3.

Proposition 2.1. *Suppose that, for all* $\mathbf{x} \in \mathbb{R}^d$,

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

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

$$\mathbb{E}\left[r_n^\star(\mathbf{x}) - \tilde{r}_n^\star(\mathbf{x})\right]^2 \leq \frac{2\sigma^2 k_n}{n}\left(1 + \frac{1}{n}\right)^{2k_n}.$$

The message of Proposition 1 is that, when resampling is done with replacement, the variance term of the bagged NN estimate is $\mathrm{O}(k_n/n)$.

To analyse the bias term in (1), we will need the following result, which bounds the convergence rate of the expected i-th nearest neighbor distance in terms of the metric covering radii of the support of the distribution μ of $\mathbf{X}$. Proposition 2.2 is a generalization of Theorem 1, page 1032 in Kulkarni and Posner [12], which only report results for the rate of convergence of *the* nearest neighbor convergence rate. Therefore, this result is interesting by itself. As for now, we let the symbol $\lfloor . \rfloor$ denote the integer part function.

Proposition 2.2. *Suppose that* $\mathbf{X}$ *is bounded. Then*

$$\mathbb{E}\|\mathbf{X}_{(i)}(\mathbf{X}) - \mathbf{X}\|^2 \leq \frac{8i}{n} \sum_{j=1}^{\lfloor n/i \rfloor} \left[\mathcal{N}^{-1}\left(j, \mathcal{S}(\mu)\right)\right]^2.$$

For any bounded set A in the Euclidean d-space, the covering radius satisfies $\mathcal{N}^{-1}(r, A) \leq \mathcal{N}^{-1}(1, A) r^{-1/d}$ (see [13]). Hence the following corollary:

Corollary 2.2. *Suppose that* $\mathbf{X}$ *is bounded, and set* $\rho = \mathcal{N}^{-1}(1, \mathcal{S}(\mu))$. *Then*

(*i*) *If* $d = 1$,

$$\mathbb{E}\|\mathbf{X}_{(i)}(\mathbf{X}) - \mathbf{X}\|^2 \leq \frac{16\rho^2 i}{n}.$$

(*ii*) *If* $d = 2$,

$$\mathbb{E}\|\mathbf{X}_{(i)}(\mathbf{X}) - \mathbf{X}\|^2 \leq \frac{8\rho^2 i}{n}\left[1 + \ln\left(\frac{n}{i}\right)\right].$$

(*iii*) *If* $d \geq 3$,

$$\mathbb{E}\|\mathbf{X}_{(i)}(\mathbf{X}) - \mathbf{X}\|^2 \leq \frac{8\rho^2 \lfloor n/i \rfloor^{-2/d}}{1 - 2/d}.$$

We are now in a position to upper bound the bias term in (1).

Proposition 2.3. *Suppose that* $\mathbf{X}$ *is bounded, and set* $\rho = \mathcal{N}^{-1}(1, \mathcal{S}(\mu))$. *Suppose in addition that, for all* $\mathbf{x}$ *and* $\mathbf{x}' \in \mathbb{R}^d$,

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

for some nonnegative constant C. *Then*

(*i*) *If* $d = 1$,

$$\mathbb{E}\left[\tilde{r}_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 \leq \frac{32C^2\rho^2}{k_n}\left(1 + \frac{1}{n}\right)^{k_n}.$$

(*ii*) *If* $d = 2$,

$$\mathbb{E}\left[\tilde{r}_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 \leq \frac{16C^2\rho^2}{k_n}\left(1 + \frac{1}{n}\right)^{k_n}\left[1 + \ln(k_n + 1)\right]$$

(*iii*) *If* $d \geq 3$,

$$\mathbb{E}\left[\tilde{r}_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 \leq \frac{8C^2\rho^2}{1 - 2/d}\left[\frac{1}{n^{k_n}} + \alpha_d\left(1 + \frac{1}{n}\right)^{k_n} {k_n}^{-2/d}\right],$$

where

$$\alpha_d = \Gamma\left(\frac{d-2}{d}\right)\Gamma\left(\frac{d+2}{d}\right).$$

The take-home message here is that, for $d \geq 3$, $\mathbb{E}\left[\tilde{r}_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 = \mathrm{O}(k_n^{-2/d})$.

3 Bagging "without replacement"

We analyse in this section the rate of convergence of the bagged nearest neighbor estimate, assuming this time that, at each step, the k_n observations are distinctly chosen at random within the sample set $\mathcal{D}_n$. This alternative aggregation scheme is called subagging (for **sub**sample **agg**regat**ing**) in Bühlmann and Yu [3]. In this case, the bagged 1-NN regression estimate takes the form

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

where

$$V_i = \begin{cases} \dfrac{\binom{n-i}{k_n-1}}{\binom{n}{k_n}}, & i \leq n-k_n+1 \\ 0, & i > n-k_n+1. \end{cases}$$

Due to the fact that there is no repetition in the sampling process, the analysis turns out to be simpler. Our result is as follows.

Theorem 3.1. *Suppose that* $\mathbf{X}$ *is bounded, and set* $\rho = \mathcal{N}^{-1}(1, \mathcal{S}(\mu))$. *Suppose in addition that, for all* $\mathbf{x}$ *and* $\mathbf{x}' \in \mathbb{R}^d$,

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

and

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

for some nonnegative constants σ^2 *and* C. *Then*

(*i*) *If* $d = 1$,

$$\mathbb{E}\left[r_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 \leq \frac{k_n}{n} \frac{\sigma^2}{(1 - k_n/n + 1/n)^2} + \frac{8C^2\rho^2}{k_n}.$$

(*ii*) *If* $d = 2$,

$$\mathbb{E}\left[r_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 \leq \frac{k_n}{n} \frac{\sigma^2}{(1 - k_n/n + 1/n)^2} + \frac{4C^2\rho^2}{k_n}(1 + \ln k_n).$$

(*iii*) *If* $d \geq 3$,

$$\mathbb{E}\left[r_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 \leq \frac{k_n}{n} \frac{\sigma^2}{(1 - k_n/n + 1/n)^2} + \frac{4C^2\rho^2}{1 - 2/d}\, k_n^{-2/d}.$$

By balancing the variance and bias terms in the theorem above, we are led to the following corollary.

Corollary 3.1. *Under the assumptions of Theorem 3.1, we have:*

(*i*) *If* $d = 1$, *for* $k_n = n^{1/2}$,

$$\mathbb{E}\left[r_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 \leq \frac{\Lambda}{\sqrt{n}},$$

where $\Lambda = \sigma^2 + 8C^2\rho^2$.

(*ii*) *If* $d = 2$, *for* $k_n = \sqrt{n(1+\ln n)}$,

$$\mathbb{E}\left[r_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 \leq \Lambda \sqrt{\frac{1+\ln n}{n}},$$

where

$$\Lambda = 5\sigma^2 + 4C^2\rho^2.$$

(*iii*) *If* $d \geq 3$, *for* $k_n = n^{\frac{d}{d+2}}$,

$$\mathbb{E}\left[r_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 \leq \left(\sigma^2 C_d^2 + \Lambda\right) n^{-\frac{2}{d+2}},$$

where

$$C_d = \max_{n\geq 1} \frac{1}{\left(1 - n^{-2/(d+2)} + \frac{1}{n}\right)} \quad \textit{and} \quad \Lambda = \frac{4C^2\rho^2}{1-2/d}.$$

As in bagging with replacement,Theorem 3.1 expresses the fact that the without replacement bagged NN estimate achieves the standard minimax rate of convergence for $d \geq 3$, with bounds which are valid for finite sample sizes. However, and contrary to bagging with replacement, the obtained constant in front of $n^{-2/(d+2)}$ is not optimal, and attentions shows that $C_d \to \infty$ as $d \to \infty$. We do not know whether this constant can be sharpened or not.

To prove Theorem 3.1, we start again by the variance/bias decomposition

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

with

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

and we analyse each term separately.

Proposition 3.1. *Suppose that, for all* $\mathbf{x} \in \mathbb{R}^d$,

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

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

$$\mathbb{E}\left[r_n^\star(\mathbf{x}) - \tilde{r}_n^\star(\mathbf{x})\right]^2 \leq \frac{k_n}{n}\, \frac{\sigma^2}{(1 - k_n/n + 1/n)^2}.$$

As in the with replacement case, Proposition 3.1 above shows that the variance term of the without repetition-bagged NN estimate is $\mathrm{O}(k_n/n)$. In order to analyse the bias term, we need the following proposition, which bounds the convergence rate of the expected nearest neighbor distance in terms of the metric covering radii of the support of the distribution μ of $\mathbf{X}$. This result sharpens the constant of Theorem 1, page 1032 in Kulkarni and Posner [12] and of Proposition 2.2 in the case of *the* nearest neighbor.

Proposition 3.2. *Suppose that* $\mathbf{X}$ *is bounded. Then*

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

Corollary 3.2. *Suppose that* $\mathbf{X}$ *is bounded, and set* $\rho = \mathcal{N}^{-1}(1, \mathcal{S}(\mu))$. *Then*

(*i*) *If* $d = 1$,

$$\mathbb{E}\|\mathbf{X}_{(1)}(\mathbf{X}) - \mathbf{X}\|^2 \leq \frac{8\rho^2}{n}.$$

(*ii*) *If* $d = 2$,

$$\mathbb{E}\|\mathbf{X}_{(1)}(\mathbf{X}) - \mathbf{X}\|^2 \leq \frac{4\rho^2}{n}(1 + \ln n).$$

(*iii*) *If* $d \geq 3$,

$$\mathbb{E}\|\mathbf{X}_{(1)}(\mathbf{X}) - \mathbf{X}\|^2 \leq \frac{4\rho^2 n^{-2/d}}{1 - 2/d}.$$

Recall that for any bounded set A in the Euclidean d-space, the covering radius satisfies $\mathcal{N}^{-1}(r, A) \leq \mathcal{N}^{-1}(1, A) r^{-1/d}$. Hence the following corollary:

Proposition 3.3. *Suppose that* $\mathbf{X}$ *is bounded, and set* $\rho = \mathcal{N}^{-1}(1, \mathcal{S}(\mu))$. *Suppose in addition that, for all* $\mathbf{x}$ *and* $\mathbf{x}' \in \mathbb{R}^d$, *and*

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

for some nonnegative constant C. *Then*

(*i*) *If* $d = 1$,

$$\mathbb{E}\left[\tilde{r}_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 \leq \frac{8C^2\rho^2}{k_n}.$$

(*ii*) *If* $d = 2$,

$$\mathbb{E}\left[\tilde{r}_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 \leq \frac{4C^2\rho^2}{k_n}(1 + \ln k_n).$$

(*iii*) *If* $d \geq 3$,

$$\mathbb{E}\left[\tilde{r}_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 \leq \frac{4C^2\rho^2 {k_n}^{-2/d}}{1 - 2/d}.$$

We are now in a position to upper bound the bias term.

Thus, for $d \geq 3$, we have as in the with replacement case :

$$\mathbb{E}\left[\tilde{r}_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 = \mathrm{O}(k_n^{-2/d}).$$

4 Proofs

4.1 Proof of Proposition 2.1

First, we observe that, for all $\mathbf{x} \in \mathbb{R}^d$,

$$\begin{aligned}
&\mathbb{E}\left[r_n^\star(\mathbf{x}) - \tilde{r}_n^\star(\mathbf{x})\right]^2 \\
&\quad = \mathbb{E}\left[\sum_{i=1}^n V_i \left(Y_{(i)}(\mathbf{x}) - r(\mathbf{X}_{(i)}(\mathbf{x}))\right)\right]^2 \\
&\quad = \mathbb{E}\left[\sum_{i=1}^n V_i^2 \left(Y_{(i)}(\mathbf{x}) - r(\mathbf{X}_{(i)}(\mathbf{x}))\right)^2\right] \\
&\quad \text{(by conditioning on the } \mathbf{X}_i\text{'s and since } r(\mathbf{X}_{(i)}(\mathbf{x})) = \mathbb{E}\left[Y_{(i)}(\mathbf{x})\right]\text{)} \\
&\quad = \mathbb{E}\left[\sum_{i=1}^n V_i^2 \sigma^2 \left(\mathbf{X}_{(i)}(\mathbf{x})\right)\right] \\
&\leq \sigma^2 \sum_{i=1}^n V_i^2. \qquad (2)
\end{aligned}$$

Next, an easy calculation shows that

$$\begin{aligned}
\sum_{i=1}^n V_i^2 &= \sum_{i=1}^n \left[\left(1 - \frac{i-1}{n}\right)^{k_n} - \left(1 - \frac{i}{n}\right)^{k_n}\right]^2 \\
&= 2\sum_{i=0}^{n-1} \left(1 - \frac{i}{n}\right)^{k_n} \left[\left(1 - \frac{i}{n}\right)^{k_n} - \left(1 - \frac{i+1}{n}\right)^{k_n}\right] - 1.
\end{aligned}$$

Let the map $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = (1-x)^{k_n}$. Then, by the mean value theorem,

$$0 \leq \left(1 - \frac{i}{n}\right)^{k_n} - \left(1 - \frac{i+1}{n}\right)^{k_n} \leq -\frac{1}{n} f'\left(\frac{i}{n}\right) = \frac{k_n}{n}\left(1 - \frac{i}{n}\right)^{k_n - 1}.$$

Thus,

$$\sum_{i=1}^n V_i^2 \leq \frac{2k_n}{n} \sum_{i=0}^{n-1} \left(1 - \frac{i}{n}\right)^{2k_n - 1} - 1.$$

In addition, let the map $g : \mathbb{R} \to \mathbb{R}$ be defined by $g(x) = (1-x)^{2k_n - 1}$. Observing that

$$\int_0^1 g(x)\mathrm{d}x = \frac{1}{2k_n},$$

we obtain

$$\begin{aligned}
\sum_{i=1}^n V_i^2 &\leq 2k_n \left[\frac{1}{n}\sum_{i=0}^{n-1} g\left(\frac{i}{n}\right) - \int_0^1 g(x)\mathrm{d}x\right] \\
&= 2k_n \sum_{i=0}^{n-1} \int_{i/n}^{(i+1)/n} \left[g\left(\frac{i}{n}\right) - g(x)\right] \mathrm{d}x.
\end{aligned}$$

Invoking again the mean value theorem, we may write, for all $x \in [i/n, (i+1)/n]$,

$$0 \leq g\left(\frac{i}{n}\right) - g(x) \leq -\frac{1}{n} g'\left(\frac{i}{n}\right).$$

Therefore,

$$\sum_{i=1}^{n} V_i^2 \leq \frac{2k_n}{n^2} \sum_{i=0}^{n-1} \left[-g'\left(\frac{i}{n}\right)\right].$$

A moment's thought shows that

$$\frac{1}{n}\sum_{i=0}^{n-1} \left[-g'\left(\frac{i}{n}\right)\right] \leq -\int_{-1/n}^{1-1/n} g'(x)\mathrm{d}x = g\left(-\frac{1}{n}\right) - g\left(1-\frac{1}{n}\right).$$

Putting all pieces together, we finally obtain

$$\begin{aligned}\sum_{i=1}^{n} V_i^2 &\leq \frac{2k_n}{n}\left[\left(1+\frac{1}{n}\right)^{2k_n-1} - \left(\frac{1}{n}\right)^{2k_n-1}\right] \\ &\leq \frac{2k_n}{n}\left(1+\frac{1}{n}\right)^{2k_n}.\end{aligned}$$

This, together with inequality (2), concludes the proof of the proposition.

4.2 Proof of Proposition 2.2

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

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

Clearly,

$$\begin{aligned}\mathbb{P}\left(D_{(i)}^2(\mathbf{X}) > \varepsilon\right) &= \mathbb{E}\left[\mathbb{P}\left(D_{(i)}(\mathbf{X}) > \sqrt{\varepsilon}|\mathbf{X}\right)\right] \\ &= \mathbb{E}\left[\sum_{j=0}^{i-1} \binom{n}{j} \left[F_{\mathbf{X}}\left(\sqrt{\varepsilon}\right)\right]^j \left[1 - F_{\mathbf{X}}\left(\sqrt{\varepsilon}\right)\right]^{n-j}\right]. \qquad (3)\end{aligned}$$

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 $\mathcal{P}_i \cap \mathcal{P}_m = \emptyset$. 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

$$\mathbb{E}\left[\frac{1}{F_{\mathbf{X}}\left(\sqrt{\varepsilon}\right)}\right] \leq \mathcal{N}(\sqrt{\varepsilon}/2). \tag{4}$$

Consequently, combining inequalities (3), (4) and technical Lemma 4.1, we obtain

$$\begin{aligned} \mathbb{P}\left(D^2_{(i)}(\mathbf{X}) > \varepsilon\right) &= \mathbb{E}\left[\frac{1}{F_{\mathbf{X}}\left(\sqrt{\varepsilon}\right)} \sum_{j=0}^{i-1} \binom{n}{j} \left[F_{\mathbf{X}}\left(\sqrt{\varepsilon}\right)\right]^{j+1} \left[1 - F_{\mathbf{X}}\left(\sqrt{\varepsilon}\right)\right]^{n-j}\right] \\ &\leq \frac{i}{n+1} E\left[\frac{1}{F_{\mathbf{X}}\left(\sqrt{\varepsilon}\right)}\right] \\ &\leq \frac{i}{n} \mathcal{N}\left(\sqrt{\varepsilon}/2\right). \end{aligned}$$

Thus, since $\mathbb{P}(D^2_{(i)}(\mathbf{X}) > \varepsilon) = 0$ for $\varepsilon > 4[\mathcal{N}^{-1}(1)]^2$, we obtain

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

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

$$\begin{aligned}
\mathbb{E}\left[D^2_{(i)}(\mathbf{X})\right] &\leq 4\left[\mathcal{N}^{-1}\left(\left\lfloor\frac{n}{i}\right\rfloor\right)\right]^2 + \frac{4i}{n}\int_{[\mathcal{N}^{-1}(\lfloor n/i\rfloor)]^2}^{[\mathcal{N}^{-1}(1)]^2} \mathcal{N}(\sqrt{\varepsilon})\mathrm{d}\varepsilon \\
&\leq 4\left[\mathcal{N}^{-1}\left(\left\lfloor\frac{n}{i}\right\rfloor\right)\right]^2 + \frac{4i}{n}\sum_{j=2}^{\lfloor n/i\rfloor}\int_{[\mathcal{N}^{-1}(j)]^2}^{[\mathcal{N}^{-1}(j-1)]^2} j\ \mathrm{d}\varepsilon \\
&= 4\left[\mathcal{N}^{-1}\left(\left\lfloor\frac{n}{i}\right\rfloor\right)\right]^2 \\
&\quad + \frac{4i}{n}\left[2\left[\mathcal{N}^{-1}(1)\right]^2 - \left\lfloor\frac{n}{i}\right\rfloor\left[\mathcal{N}^{-1}\left(\left\lfloor\frac{n}{i}\right\rfloor\right)\right]^2 + \sum_{j=2}^{\lfloor n/i\rfloor-1}\left[\mathcal{N}^{-1}(j)\right]^2\right] \\
&\leq \frac{8i}{n}\left[\mathcal{N}^{-1}(1)\right]^2 + \frac{4i}{n}\left[\mathcal{N}^{-1}\left(\left\lfloor\frac{n}{i}\right\rfloor\right)\right]^2 + \frac{4i}{n}\sum_{j=2}^{\lfloor n/i\rfloor-1}\left[\mathcal{N}^{-1}(j)\right]^2,
\end{aligned}$$

where the last statement follows from the inequality

$$-\frac{4i}{n}\left\lfloor\frac{n}{i}\right\rfloor + 4 \leq \frac{4i}{n}.$$

In conclusion, we are led to

$$\mathbb{E}\left[D^2_{(i)}(\mathbf{X})\right] \leq \frac{8i}{n}\sum_{j=1}^{\lfloor n/i\rfloor}\left[\mathcal{N}^{-1}(j)\right]^2,$$

as desired.

4.3 Proof of Corollary 2.2

For any bounded set A in the Euclidean d-space, the covering radius satisfies $\mathcal{N}^{-1}(r, A) \leq \mathcal{N}^{-1}(1, A)r^{-1/d}$ (see [13]). Consequently, using Proposition 2.2, we obtain

(*i*) For $d = 1$,

$$\begin{aligned}
\mathbb{E}\|\mathbf{X}_{(i)}(\mathbf{X}) - \mathbf{X}\|^2 &\leq \frac{8\rho^2 i}{n}\sum_{j=1}^{\lfloor n/i\rfloor} j^{-2} \\
&\leq \frac{8\rho^2 i}{n}\left[1 + \int_1^{\lfloor n/i\rfloor} x^{-2}\mathrm{d}x\right] \\
&\leq \frac{16\rho^2 i}{n}.
\end{aligned}$$

(*ii*) For $d = 2$,

$$\begin{aligned}\mathbb{E}\|\mathbf{X}_{(i)}(\mathbf{X}) - \mathbf{X}\|^2 &\leq \frac{8\rho^2 i}{n} \sum_{j=1}^{\lfloor n/i \rfloor} j^{-1} \\ &\leq \frac{8\rho^2 i}{n} \left[1 + \int_1^{\lfloor n/i \rfloor} x^{-1} \mathrm{d}x\right] \\ &\leq \frac{8\rho^2 i}{n} \left[1 + \ln\left(\frac{n}{i}\right)\right].\end{aligned}$$

(*iii*) For $d \geq 3$,

$$\begin{aligned}\mathbb{E}\|\mathbf{X}_{(i)}(\mathbf{X}) - \mathbf{X}\|^2 &\leq \frac{8\rho^2 i}{n} \sum_{j=1}^{\lfloor n/k_n \rfloor} j^{-2/d} \\ &\leq \frac{8\rho^2 i}{n} \int_0^{\lfloor n/i \rfloor} x^{-2/d} \mathrm{d}x \\ &= \frac{8\rho^2 \lfloor n/i \rfloor^{-2/d}}{1 - 2/d}.\end{aligned}$$

In the last statement, we used the inequality $i/n \leq 1/\lfloor n/i \rfloor$.

4.4 Proof of Proposition 2.3

We write first

$$\begin{aligned}\mathbb{E}\left[\tilde{r}_n^\star(\mathbf{x}) - r(\mathbf{x})\right]^2 &= \mathbb{E}\left[\sum_{i=1}^n V_i \left(r(\mathbf{X}_{(i)}(\mathbf{x})) - r(\mathbf{x})\right)\right]^2 \\ &\leq \mathbb{E}\left[\sum_{i=1}^n V_i \left|r(\mathbf{X}_{(i)}(\mathbf{x})) - r(\mathbf{x})\right|\right]^2 \\ &\leq C^2\, \mathbb{E}\left[\sum_{i=1}^n V_i \left\|\mathbf{X}_{(i)}(\mathbf{x}) - \mathbf{x}\right\|\right]^2 \\ &\leq C^2 \left[\sum_{i=1}^n V_i\, \mathbb{E}\|\mathbf{X}_{(i)}(\mathbf{x}) - \mathbf{x}\|^2\right] \\ &\qquad \text{(by Cauchy-Schwarz inequality and since} \sum_{i=1}^n V_i = 1).\end{aligned}$$

Thus, integrating with respect to the distribution of $\mathbf{X}$,

$$\mathbb{E}\left[\tilde{r}_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 \leq C^2 \left[\sum_{i=1}^n V_i\, \mathbb{E}\|\mathbf{X}_{(i)}(\mathbf{X}) - \mathbf{X}\|^2\right].$$

To bound the term $\mathbb{E}\|\mathbf{X}_{(i)}(\mathbf{X}) - \mathbf{X}\|^2$, we apply Corollary 2.2 and distinguish the cases $d = 1$, $d = 2$ and $d \geq 3$.

(i) If $d = 1$, for $i = 1, \ldots, n$,

$$\mathbb{E}\|\mathbf{X}_{(i)}(\mathbf{X}) - \mathbf{X}\|^2 \leq \frac{16\rho^2 i}{n}.$$

Consequently,

$$\mathbb{E}\left[\tilde{r}_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 \leq 16C^2\rho^2 \sum_{i=1}^{n} V_i \frac{i}{n},$$

and by definition of the V_i's

$$\mathbb{E}\left[\tilde{r}_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 \leq 16C^2\rho^2 \sum_{i=1}^{n} \left[\left(1 - \frac{i-1}{n}\right)^{k_n} - \left(1 - \frac{i}{n}\right)^{k_n}\right] \frac{i}{n}.$$

Thus

$$\begin{aligned}
\mathbb{E}\left[\tilde{r}_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 &\leq 16C^2\rho^2 \sum_{i=1}^{n} \left[\left(1 - \frac{i}{n} + \frac{1}{n}\right)^{k_n} - \left(1 - \frac{i}{n}\right)^{k_n}\right] \frac{i}{n} \\
&= 16C^2\rho^2 \sum_{i=1}^{n} \left[\sum_{j=1}^{k_n} \binom{k_n}{j} \frac{1}{n^j} \left(1 - \frac{i}{n}\right)^{k_n - j}\right] \frac{i}{n} \\
&= 16C^2\rho^2 \sum_{j=1}^{k_n} \binom{k_n}{j} \frac{1}{n^{j-1}} \left[\frac{1}{n} \sum_{i=1}^{n} \left(1 - \frac{i}{n}\right)^{k_n - j} \frac{i}{n}\right].
\end{aligned}$$

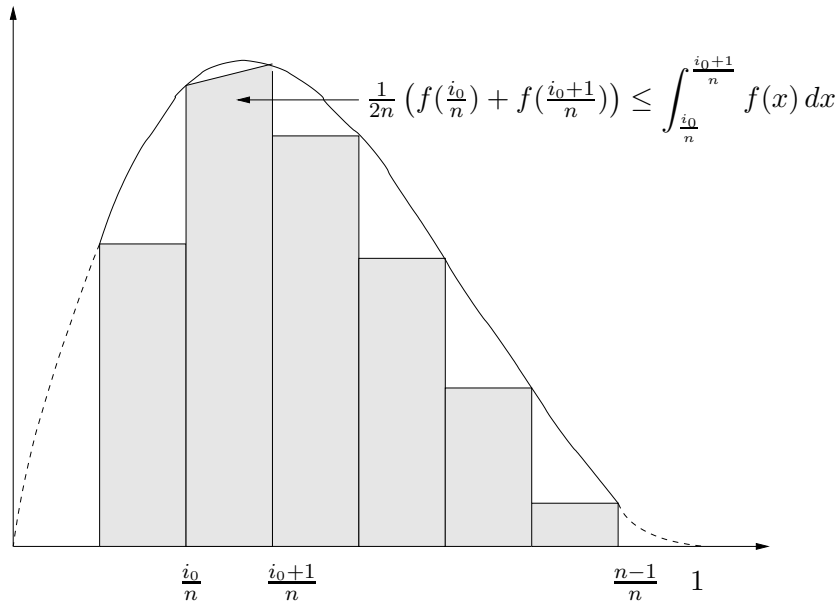

Figure 1: Illustration of $\frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{i}{n}\right)^{k_n - j} \frac{i}{n} \leq 2\int_0^1 x(1-x)^{k_n - j}\mathrm{d}x$.

For all $j = 1, \ldots, k_n$, we now use the inequality

$$\frac{1}{n}\sum_{i=1}^{n} \frac{i}{n}\left(1-\frac{i}{n}\right)^{k_n-j} \leq 2\int_0^1 x(1-x)^{k_n-j}\mathrm{d}x$$

This inequality is clearly true for $j = k_n$, even without the factor 2 in front of the integral. For $j < k_n$, it is illustrated on Figure 1 where $f(x) = x(1-x)^{k_n-j}$. The factor 2 is this time necessary because f is not monotonous on $[0, 1]$.

Consequently,

$$\mathbb{E}\left[\tilde{r}_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 \leq 32C^2\rho^2 \sum_{j=1}^{k_n} \binom{k_n}{j} \frac{1}{n^{j-1}} \int_0^1 x(1-x)^{k_n-j}\mathrm{d}x.$$

Recalling the general formula

$$\int_0^1 x^{p-1}(1-x)^{q-1}\mathrm{d}x = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}, \quad p, q > 0, \tag{5}$$

we obtain

$$\begin{aligned}
&\mathbb{E}\left[\tilde{r}_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 \\
&\leq 32C^2\rho^2 \sum_{j=1}^{k_n} \binom{k_n}{j} \frac{1}{n^{j-1}} \frac{\Gamma(2)\Gamma(k_n-j+1)}{\Gamma(k_n-j+3)} \\
&\leq 32C^2\rho^2 \sum_{j=1}^{k_n} \binom{k_n}{j} \frac{1}{n^{j-1}} \frac{1}{(k_n-j+1)(k_n-j+2)} \\
&= 32C^2\rho^2 \sum_{j=1}^{k_n} \binom{k_n}{j-1} \frac{1}{n^{j-1}} \frac{1}{j(k_n-j+2)} \\
&= 32C^2\rho^2 \sum_{j=0}^{k_n-1} \binom{k_n}{j} \frac{1}{n^{j}} \frac{1}{(j+1)(k_n-j+1)}.
\end{aligned}$$

Observing finally that $(j+1)(k_n-j+1) \geq k_n$ for all $j = 0, \ldots, k_n-1$, we conclude

$$\mathbb{E}\left[\tilde{r}_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 \leq \frac{32C^2\rho^2}{k_n} \sum_{j=0}^{k_n-1} \binom{k_n}{j} \frac{1}{n^j} \leq \frac{32C^2\rho^2}{k_n}\left(1+\frac{1}{n}\right)^{k_n}.$$

(*ii*) For $d = 2$, a reasoning similar to the one reported in statement (*i*) above can be followed, to show that

$$\begin{aligned}
&\mathbb{E}\left[\tilde{r}_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 \\
&\leq 16C^2\rho^2 \left[\frac{1}{k_n}\left(1+\frac{1}{n}\right)^{k_n}\right. \\
&\qquad \left. - \sum_{j=1}^{k_n} \binom{k_n}{j} \frac{1}{n^{j-1}} \int_0^1 x(1-x)^{k_n-j}\ln x \mathrm{d}x\right].
\end{aligned} \tag{6}$$

Denoting by H_n the n-th harmonic number, i.e.,

$$H_n = 1 + \frac{1}{2} + \ldots + \frac{1}{n},$$

we may write, using technical Lemma 4.2,

$$\begin{aligned}
&- \sum_{j=1}^{k_n} \binom{k_n}{j} \frac{1}{n^{j-1}} \int_0^1 x(1-x)^{k_n-j} \ln x \mathrm{d}x \\
&= \sum_{j=1}^{k_n} \binom{k_n}{j} \frac{1}{n^{j-1}} \frac{H_{k_n-j+2} - 1}{(k_n-j+1)(k_n-j+2)} \\
&= \sum_{j=0}^{k_n-1} \binom{k_n}{j} \frac{1}{n^j} \frac{H_{k_n-j+1} - 1}{(j+1)(k_n-j+1)}.
\end{aligned}$$

For all $j = 0, \ldots, k_n - 1$, we have $(j+1)(k_n - j + 1) \geq k_n$ as well as

$$\begin{aligned}
H_{k_n-j+1} - 1 &= \frac{1}{2} + \ldots + \frac{1}{k_n - j + 1} \\
&\leq \int_1^{k_n-j+1} \frac{\mathrm{d}x}{x} \\
&= \ln(k_n - j + 1) \\
&\leq \ln(k_n + 1).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&- \sum_{j=1}^{k_n} \binom{k_n}{j} \frac{1}{n^{j-1}} \int_0^1 x(1-x)^{k_n-j} \ln x \mathrm{d}x \\
&\leq \frac{\ln(k_n+1)}{k_n} \sum_{j=0}^{k_n-1} \binom{k_n}{j} \frac{1}{n^j} \\
&\leq \frac{\ln(k_n+1)}{k_n} \left(1 + \frac{1}{n}\right)^{k_n}. \qquad (7)
\end{aligned}$$

Combining inequalities (6) and (7) leads to the desired result.

(*iii*) For $d \geq 3$, starting again as in (*i*), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\tilde{r}_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 \\
&\leq \frac{8C^2\rho^2}{1-2/d} \sum_{i=1}^{n} V_i \left\lfloor \frac{n}{i} \right\rfloor^{-2/d}. \qquad (8)
\end{aligned}$$

To upper bound the right-hand side, we note that for all $i = 1, \ldots, n-1$,

$$\left\lfloor \frac{n}{i} \right\rfloor^{-2/d} \leq \left(\frac{i/n}{1 - i/n}\right)^{2/d},$$

and set consequently

$$S_n = \frac{1}{n^{k_n}} + \sum_{i=1}^{n-1} \left[\left(1 - \frac{i-1}{n}\right)^{k_n} - \left(1 - \frac{i}{n}\right)^{k_n}\right] \left(\frac{i/n}{1-i/n}\right)^{2/d}.$$

We obtain

$$\begin{aligned} S_n &= \frac{1}{n^{k_n}} + \sum_{i=1}^{n-1} \left[\sum_{j=1}^{k_n} \binom{k_n}{j} \frac{1}{n^j} \left(1 - \frac{i}{n}\right)^{k_n - j} \right] \left(\frac{i/n}{1 - i/n} \right)^{2/d} \\ &= \frac{1}{n^{k_n}} + \sum_{j=1}^{k_n} \binom{k_n}{j} \frac{1}{n^{j-1}} \left[\frac{1}{n} \sum_{i=1}^{n-1} \left(1 - \frac{i}{n}\right)^{k_n - j - 2/d} \left(\frac{i}{n}\right)^{2/d} \right] \\ &\leq \frac{1}{n^{k_n}} + 2 \sum_{j=1}^{k_n} \binom{k_n}{j} \frac{1}{n^{j-1}} \int_0^1 x^{2/d} (1-x)^{k_n - j - 2/d} \mathrm{d}x. \end{aligned}$$

Applying formula (5) again, together with the identity

$$\Gamma\left(p + \frac{d-2}{d}\right) = \Gamma\left(\frac{d-2}{d}\right) \prod_{\ell=1}^{p} \left(\ell - \frac{2}{d}\right),$$

we obtain

$$\begin{aligned} S_n &\leq \frac{1}{n^{k_n}} + \alpha_d \sum_{j=1}^{k_n} \binom{k_n}{j} \frac{1}{n^{j-1}} \frac{1}{(k_n - j + 1)} \prod_{\ell=1}^{k_n - j} \left(1 - \frac{2}{d\ell}\right) \\ &\quad (\text{with } \alpha_d = 2\Gamma((d-2)/d)\,\Gamma((d+2)/d) \\ &= \frac{1}{n^{k_n}} + \alpha_d \sum_{j=1}^{k_n} \frac{k_n!}{j!(k_n - j + 1)!} \frac{1}{n^{j-1}} \prod_{\ell=1}^{k_n - j} \left(1 - \frac{2}{d\ell}\right) \\ &= \frac{1}{n^{k_n}} + \alpha_d \sum_{j=1}^{k_n} \frac{1}{n^{j-1}} \binom{k_n}{j-1} \frac{1}{j} \prod_{\ell=1}^{k_n - j} \left(1 - \frac{2}{d\ell}\right) \\ &= \frac{1}{n^{k_n}} + \alpha_d \sum_{j=0}^{k_n - 1} \frac{1}{n^j} \binom{k_n}{j} \frac{1}{j+1} \prod_{\ell=1}^{k_n - j - 1} \left(1 - \frac{2}{d\ell}\right) \end{aligned}$$

and by technical Lemma 4.3:

$$\begin{aligned} S_n &\leq \frac{1}{n^{k_n}} + \alpha_d \sum_{j=0}^{k_n - 1} \binom{k_n}{j} \frac{1}{n^j} k_n^{-2/d} \\ &\leq \frac{1}{n^{k_n}} + \alpha_d \left(1 + \frac{1}{n}\right)^{k_n} k_n^{-2/d}. \end{aligned} \tag{9}$$

Finally, combining inequalities (8) and (9) concludes the proof of Proposition 2.3.

4.5 Proof of Proposition 3.1

Starting as in the proof of Proposition 2.1, we write

$$\mathbb{E}\left[r_n^\star(\mathbf{x}) - \tilde{r}_n^\star(\mathbf{x})\right]^2 \leq \sigma^2 \sum_{i=1}^{n} V_i^2,$$

where, for $i = 1, \ldots, n - k_n + 1$,

$$\begin{aligned} V_i &= \frac{\binom{n-i}{k_n-1}}{\binom{n}{k_n}} \\ &= \frac{k_n}{n-k_n+1} \prod_{j=0}^{k_n-2} \left(1 - \frac{i}{n-j}\right) \\ &\leq \frac{k_n}{n-k_n+1} \prod_{j=0}^{k_n-2} \left(1 - \frac{i}{n}\right) \\ &= \frac{k_n}{n-k_n+1} \left(1 - \frac{i}{n}\right)^{k_n-1}. \end{aligned}$$

This yields

$$\begin{aligned} \sum_{i=1}^{n} V_i^2 &\leq \frac{k_n^2}{(n-k_n+1)^2} \sum_{i=1}^{n-k_n+1} \left(1 - \frac{i}{n}\right)^{2(k_n-1)} \\ &\leq \frac{k_n^2\, n}{(n-k_n+1)^2} \frac{1}{n} \sum_{i=1}^{n} \left(1 - \frac{i}{n}\right)^{2(k_n-1)}. \end{aligned}$$

Observing finally that

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \left(1 - \frac{i}{n}\right)^{2(k_n-1)} &\leq \int_0^1 (1-x)^{2(k_n-1)} \mathrm{d}x \\ &= \frac{1}{2k_n - 1}, \end{aligned}$$

we conclude that

$$\sum_{i=1}^{n} V_i^2 \leq \frac{k_n^2\, n}{(2k_n-1)(n-k_n+1)^2} \leq \frac{k_n}{n} \frac{1}{(1 - k_n/n + 1/n)^2}.$$

4.6 Proof of Proposition 3.2

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,

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

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

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

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

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

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

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

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.

4.7 Proof of Corollary 3.2

Recall that the covering radius satisfies the property

$$\mathcal{N}^{-1}(r, \mathcal{S}(\mu)) \leq \mathcal{N}^{-1}(1, \mathcal{S}(\mu)) r^{-1/d}$$

(see [13]). Thus, using Proposition 3.2, we obtain

(*i*) For $d = 1$,

$$\mathbb{E}\|\mathbf{X}_{(1)}(\mathbf{X}) - \mathbf{X}\|^2 \leq \frac{4\rho^2}{n} \sum_{i=1}^{n} i^{-2} \leq \frac{8\rho^2}{n}.$$

(*ii*) For $d = 2$,

$$\mathbb{E}\|\mathbf{X}_{(1)}(\mathbf{X}) - \mathbf{X}\|^2 \leq \frac{4\rho^2}{n}\sum_{i=1}^{n} i^{-1} \leq \frac{4\rho^2}{n}(1 + \ln n).$$

(*iii*) For $d \geq 3$,

$$\begin{aligned}\mathbb{E}\|\mathbf{X}_{(1)}(\mathbf{X}) - \mathbf{X}\|^2 &\leq \frac{4\rho^2}{n}\sum_{i=1}^{n} i^{-2/d} \\ &\leq \frac{4\rho^2}{n}\int_0^n x^{-2/d}\mathrm{d}x \\ &\leq \frac{4\rho^2 n^{-2/d}}{1-2/d}.\end{aligned}$$

4.8 Proof of Proposition 3.3

Recall that

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

and observe that

$$\tilde{r}_n^\star(\mathbf{x}) = \mathbb{E}^\star\left[r(\mathbf{X}_{(1)}^\star(\mathbf{x}))\right],$$

where $\mathbf{X}_{(1)}^\star(\mathbf{x})$ is the nearest neighbor of $\mathbf{x}$ in a random subsample $\mathcal{S}_n$ drawn without replacement from $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ with $\mathrm{Card}(\mathcal{S}_n) = k_n$, and $\mathbb{E}^\star$ denotes expectation with respect to the resampling distribution, conditionally on the data set $\mathcal{D}_n$. Consequently, by Jensen's inequality,

$$\begin{aligned}\mathbb{E}\left[\tilde{r}_n^\star(\mathbf{x}) - r(\mathbf{x})\right]^2 &= \mathbb{E}\left[\mathbb{E}^\star\left[r\left(\mathbf{X}_{(1)}^\star(\mathbf{x})\right) \mid \mathcal{D}_n\right] - r(\mathbf{x})\right]^2 \\ &= \mathbb{E}\left[\mathbb{E}^\star\left[r\left(\mathbf{X}_{(1)}^\star(\mathbf{x})\right) - r(\mathbf{x}) \mid \mathcal{D}_n\right]\right]^2 \\ &\leq \mathbb{E}\left[\mathbb{E}^\star\left[\left(r\left(\mathbf{X}_{(1)}^\star(\mathbf{x})\right) - r(\mathbf{x})\right)^2 \mid \mathcal{D}_n\right]\right] \\ &= \mathbb{E}\left[r\left(\mathbf{X}_{(1)}^\star(\mathbf{x})\right) - r(\mathbf{x})\right]^2 \\ &\leq C^2\,\mathbb{E}\|\mathbf{X}_{(1)}^\star(\mathbf{x}) - r(\mathbf{x})\|^2\end{aligned}$$

Since $\mathrm{Card}(\mathcal{S}_n) = k_n$, we can now apply Corollary 3.2, replacing n with k_n.

4.9 Some technical lemmas

Lemma 4.1. *For $j = 0, \ldots, n-1$, let the map $\varphi_{n,j}(p)$ be defined by*

$$\varphi_{n,j}(p) = \binom{n}{j} p^{j+1}(1-p)^{n-j}, \quad 0 \leq p \leq 1.$$

Then for all $i \leq n$

$$\sup_{0\leq p\leq 1}\sum_{j=0}^{i-1}\varphi_{n,j}(p) \leq \frac{i}{n+1}.$$

Proof of Lemma 4.1 Each map $\varphi_{n,j}$ is nonnegative, continuously increasing on the interval $[0,(j+1)/(n+1)]$ and decreasing on $[(j+1)/(n+1),1]$. Consequently, the supremum of the continuous function $\sum_{j=0}^{i-1}\varphi_{n,j}(p)$ is achieved at some point $p_\star$ of the interval $[1/(n+1),i/(n+1)]$. That is,

$$\begin{aligned}\sup_{0\leq p\leq 1}\sum_{j=0}^{i-1}\varphi_{n,j}(p) &= \sum_{j=0}^{i-1}\varphi_{n,j}(p_\star)\\ &= p_\star\sum_{j=0}^{i-1}\binom{n}{j}p_\star^j(1-p_\star)^{n-j}\\ &\leq p_\star\sum_{j=0}^{n}\binom{n}{j}p_\star^j(1-p_\star)^{n-j}\\ &= p_\star \leq \frac{i}{n+1}.\end{aligned}$$

■

Lemma 4.2. *Let, for each integer $m\geq 0$,*

$$I_m = -\int_0^1 x(1-x)^m\ln x\mathrm{d}x.$$

Then

$$I_m = \frac{H_{m+2}-1}{(m+1)(m+2)},$$

where H_n is the n-th harmonic number, i.e.,

$$H_n = 1+\frac{1}{2}+\ldots+\frac{1}{n}.$$

Proof of Lemma 4.2 Two successive integrations by parts of I_m show that

$$\begin{aligned}&(m+2)(m+1)I_m\\ &\quad= -(m+2)\int_0^1(m+1)(1-x)^m x\ln x\mathrm{d}x\\ &\quad= -\int_0^1(m+2)(1-x)^{m+1}(1+\ln x)\mathrm{d}x.\\ &\quad= \left[\left((1-x)^{m+2}-1\right)(1+\ln x)\right]_0^1+\int_0^1\frac{1-(1-x)^{m+2}}{x}\mathrm{d}x\end{aligned}$$

Thus

$$\begin{aligned}(m+2)(m+1)I_m &= -1+\int_0^1\frac{1-(1-x)^{m+2}}{x}\mathrm{d}x\\ &= -1+\int_0^1\frac{1-u^{m+2}}{1-u}\mathrm{d}u\\ &= -1+\sum_{k=0}^{m+1}\int_0^1 u^k\mathrm{d}u\\ &= -1+H_{m+2}.\end{aligned}$$

■

Lemma 4.3. *For each $d \geq 3$, each $k_n \geq 1$, and $j = 0, \ldots, k_n - 1$, we have*

$$\frac{1}{j+1} \prod_{\ell=1}^{k_n-j-1} \left(1 - \frac{2}{d\ell}\right) \leq k_n^{-2/d}.$$

Proof of Lemma 4.3 First, since $0 \leq 1 - x \leq e^{-x}$ for all $x \in [0, 1]$,

$$\prod_{\ell=1}^{k_n-j-1} \left(1 - \frac{2}{d\ell}\right) \leq \exp\left(-\frac{2}{d} \sum_{\ell=1}^{k_n-j-1} \frac{1}{\ell}\right).$$

Thus, using $1 + 1/2 + \ldots + 1/p \geq \ln(p + 1)$, we deduce

$$\prod_{\ell=1}^{k_n-j-1} \left(1 - \frac{2}{d\ell}\right) \leq (k_n - j)^{-2/d}$$

To conclude, we use the fact that, for $j = 0, \ldots, k_n - 1$,

$$\frac{1}{j+1} (k_n - j)^{-2/d} \leq k_n^{-2/d}.$$

To see this, note that the inequality above may be written under the equivalent form

$$(1 - j/k_n)^{-2/d} \leq 1 + j = 1 + k_n \cdot \frac{j}{k_n}.$$

The result can be deduced from the comparison between the map $\varphi : x \mapsto (1 - x)^{-2/d}$ and $\psi : x \mapsto 1 + k_n x$ on the interval $[0, 1 - 1/k_n]$. Just note that $\varphi(0) = \psi(0)$, $\varphi(1 - 1/k_n) = k_n^{-2/d} < 1/k_n = \psi(1 - 1/k_n)$ since $d \geq 3$, and φ is convex while ψ is affine (see Figure 2). ■

References

[1] Biau, G. and Devroye, L. (2008). On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification, *Technical Report, Université Pierre et Marie Curie*, Paris.

[2] Breiman, L. (1996). Bagging predictors, *Machine Learning*, **24**, 123-140.

[3] Bühlmann, P. and Yu, B. (2002). Analyzing bagging, *The Annals of Statistics*, **30**, 927-961.

[4] Buja, A. and Stuetzle, W. (2006). Observations on bagging, *Statistica Sinica*, **16** 323-352.

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

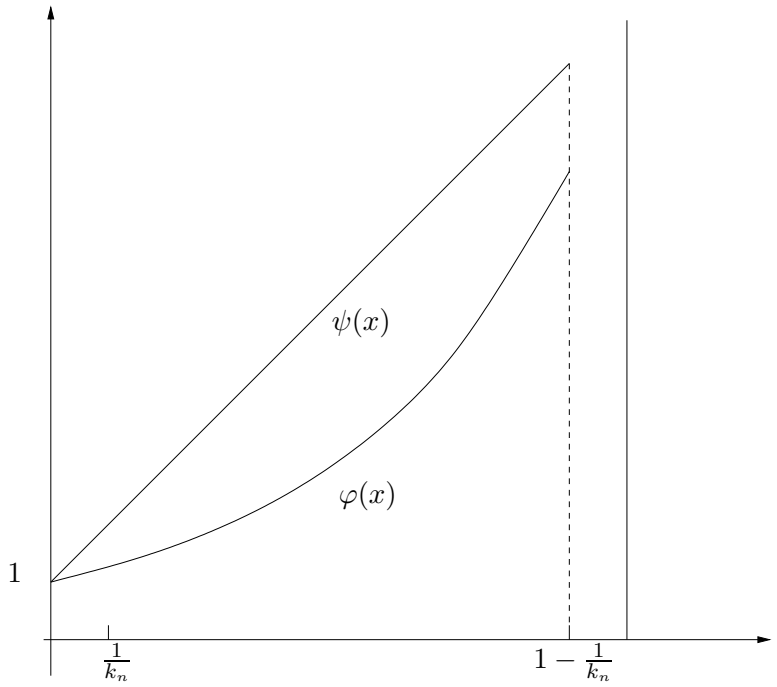

Figure 2: Illustration of $\varphi(x) \leq \psi(x)\ \forall x \in [0, 1-1/k_n]$, see text.

[6] Dietterich, T.G. (2000). Ensemble methods in machine learning, in J. Kittler and F. Roli (Eds.), *First International Workshop on Multiple Classifier Systems*, Lecture Notes in Computer Science, 1-15, Springer-Verlag, New York.

[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] Friedman, J.H. and Hall, P. (2000). On bagging and nonlinear estimation, *Technical Report, Stanford University*, Stanford.

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

[11] Hall, P. and Samworth, R.J. (2005). Properties of bagged nearest neighbour classifiers, *Journal of the Royal Statistical Society B*, **67**, 363-379.

[12] 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.

[13] 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.

[14] Steele, B.M. (2009) Exact bootstrap k-nearest neighbor learners, *Machine Learning*, **74**, 235-255.

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

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