# On the Rate of Convergence of the Bagged Nearest Neighbor Estimate

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

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

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

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

**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 subsample, and decide by combining. By bagging an $n$-sample, the crude nearest neighbor regression estimate is turned into a consistent weighted nearest neighbor regression estimate, which is amenable to statistical analysis. Letting the resampling size $k_n$ grows appropriately with $n$, it is shown that this estimate may achieve optimal rate of convergence, independently from the fact that resampling is done with or without replacement. Since the estimate with the optimal rate of convergence depends on the unknown distribution of the observations, adaptation results by data-splitting are presented.

*Index Terms* — Bagging, Resampling, Nearest neighbor estimate, Rates of convergence.

*AMS 2000 Classification*: 62G05, 62G20.

[1] Corresponding author.

# 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 [11] 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 subsamples from the original data set, constructing a predictor from each resample, and decide by combining. It is one of the most effective computationally intensive procedures to improve on unstable estimates or classifiers, especially for large, high dimensional data set problems where finding a good model 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 [14], Bühlmann and Yu [4], Hall and Samworth [18], Buja and Stuetzle [5], 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 approaches to regression and classification (Fix and Hodges [12, 13], Cover and Hart [8], Cover [6, 7], Györfi [16], Venkatesh, Snapp and Psaltis [28], Psaltis, Snapp and Venkatesh [25]). 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 [10], 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 $\mathbb{E}Y^2 < \infty$. 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. It is clearly not, in general, a consistent estimate (Devroye, Györfi and Lugosi [10], Chapter 5). However, by bagging, one may turn the 1-NN estimate into a consistent one, provided that the size of the resamples 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 $\{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$, 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, 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 $\mathcal{D}_n$.

The following 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 that 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 deterministic weights that sum to one, 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 a weighted nearest neighbor regression estimate. It is known to be universally consistent provided $V_1 \to 0$ and $\sum_{i>\varepsilon n} V_i \to 0$ for all $\varepsilon > 0$ as $n \to \infty$ (Stone [27], Devroye [9], and Problems 11.7, 11.8 of Devroye, Györfi and Lugosi [10]). 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 "with replacement" sampling

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

whereas for sampling "without replacement", $V_i$ is hypergeometric:

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

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 weights have been independently exhibited by Steele [26], who also shows on practical examples that substantial reductions in prediction error are possible by bagging the 1-NN estimate. Note also that this new expression for the 1-NN bagged estimate makes any Monte-Carlo approach unnecessary to evaluate the estimate. Indeed, up to now, this predictor was implemented by Monte-Carlo, i.e., by repeating the random sampling $T$ times, and taking 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, the bagged regression estimate was approximately evaluated as

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

where $Z_1, \ldots, Z_T$ are the outcomes in the individual rounds. Clearly, writing the 1-NN bagged estimate as an (exact) weighted nearest neighbor predictor makes such calculations useless.

On the other hand, the fact that the bagged 1-NN estimate reduces to a weighted nearest neighbor estimate may seem at first sight somehow disappointing. Indeed, we get the ordinary $k_n$-NN rule back by the choice

$$V_i = \begin{cases} 1/k_n & \text{if } i \leq k_n \\ 0 & \text{otherwise,} \end{cases}$$

and, with an appropriate choice of the sequence $(k_n)$, this regression estimate is known to have optimal asymptotic properties (see Chapter 6 in Györfi, Kohler, Krzyżak and Walk [17] and the references therein). Thus, the question is: Why would one care about the bagged nearest neighbor rule then? The answer is twofold. First, bagging the 1-NN is a very popular technique for regression and classification in the machine learning community, and most — if not all — empirical studies report practical improvements over the traditional $k_n$-NN method. Secondly (and most importantly), analysing 1-NN bagging is part of a larger project trying to understand the driving forces behind the random forests estimates, which were defined by Breiman in [3]. In short, random forests are some of the most successful ensemble methods that exhibit performance on the level of boosting and support vector machines. These learning procedures typically involve a resampling step, which may be interpreted as a particular 1-NN bagged procedure based on the so-called "layered nearest neighbor" proximities (Lin and Jeon [24], Biau and Devroye [1]).

Thus, in the present paper, we go one step further in bagging investigation and study the rate of convergence of $\mathbb{E}\left[r_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2$ to 0 as $n \to \infty$. We will start our analysis by stating a comprehensive theorem on the rate of convergence of general weighted nearest neighbor estimates (subsection 2.1). Then, this result will be particularized to 1-NN bagging, by distinguishing the "with replacement" (subsection 2.2) and the "without replacement" (subsection 2.3) cases. For the sake of clarity, technical proofs are postponed to section 3.

Throughout the document, we will be interested in rate of convergence results for the class $\mathcal{F}$ of $(1, C, \rho, \sigma^2)$-smooth distributions $(\mathbf{X}, Y)$ such that $\mathbf{X}$ has compact support with diameter $2\rho$, the regression function $r$ is Lipschitz with constant $C$ and, for all $\mathbf{x} \in \mathbb{R}^d$, $\sigma^2(\mathbf{x}) = \mathbb{V}[Y \,|\, \mathbf{X} = \mathbf{x}] \leq \sigma^2$ (the symbol $\mathbb{V}$ denotes variance). It is known (see for example Ibragimov and Khasminskii [19, 20, 21]) that for the class $\mathcal{F}$, the sequence $(n^{-\frac{2}{d+2}})$ is the optimal minimax rate of convergence. In particular,

$$\liminf_{n\to\infty}\inf_{r_n}\sup_{(\mathbf{X},Y)\in\mathcal{F}}\frac{\mathbb{E}[r_n(\mathbf{X})-r(\mathbf{X})]^2}{\left((\rho C)^d\sigma^2\right)^{\frac{2}{d+2}}\,n^{-\frac{2}{d+2}}}\geq\Delta$$

for some positive constant $\Delta$ independent of $C$, $\rho$ and $\sigma^2$. Here the infimum is taken over all estimates $r_n$, i.e., over all square integrable measurable functions of the data. As a striking result, we prove in subsections 2.2 and 2.3 that, irrespectively of the resampling type, for $d\geq 3$ and a suitable choice of the sequence $(k_n)$, the estimate $r_n^\star$ is of optimum rate for the class $\mathcal{F}$, that is

$$\limsup_{n\to\infty}\sup_{(\mathbf{X},Y)\in\mathcal{F}}\frac{\mathbb{E}[r_n^\star(\mathbf{X})-r(\mathbf{X})]^2}{\left((\rho C)^d\sigma^2\right)^{\frac{2}{d+2}}\,n^{-\frac{2}{d+2}}}\leq\Lambda$$

for some positive $\Lambda$ independent of $C$, $\rho$ and $\sigma^2$. Since the parameter $k_n$ of the estimate with the optimal rate of convergence depends on the unknown distribution of $(\mathbf{X},Y)$, especially on the smoothness of the regression function, we present in subsection 2.4 adaptive (i.e., data-dependent) choices of $k_n$ which preserve the minimax optimality of the estimate.

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 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 [18] and Steele [26], who also report encouraging numerical results in the context of regression and classification.

# 2 Rates of convergence

## 2.1 Weighted nearest neighbor estimates

As an appetizer, we start our analysis of the 1-NN bagged regression estimate from a larger point of view, by offering a general theorem on the rate of convergence of weighted nearest neighbor estimates, i.e., estimates of the form

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

with nonnegative weights satisfying the constraints $\sum_{i=1}^n V_i=1$ and $V_1\geq V_2\geq\ldots\geq V_n\geq 0$. Let us first 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 [22]). As put forward in Kulkarni and Posner [23], these quantities play a key role in 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 $\varepsilon$.

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

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

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

**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 nonincreasing discrete function of $r$.

Throughout the paper, we will denote by $\mu$ the distribution of $\mathbf{X}$, which will be assumed to be a bounded random variable. Recall that the support $\mathcal{S}(\mu)$ of $\mu$ is defined as the collection of all $\mathbf{x}$ with $\mu(\mathcal{B}(\mathbf{x}, \varepsilon)) > 0$ for all $\varepsilon > 0$. Letting $\rho = \mathcal{N}^{-1}(1, \mathcal{S}(\mu))$, we observe that $2\rho$ is an upper bound on the diameter of $\mathcal{S}(\mu)$. We are now in a position to state the main result of this subsection. We let the symbol $\lfloor . \rfloor$ denote the integer part function.

**Theorem 2.1** *Let $r_n(\mathbf{x}) = \sum_{i=1}^{n} V_i Y_{(i)}(\mathbf{x})$ be a weighted nearest neighbor estimate of $r(\mathbf{x})$. 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 positive constants $\sigma^2$ and $C$. Then* *(i) If* $d=1$,

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

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

$$\mathbb{E}\left[r_n(\mathbf{X})-r(\mathbf{X})\right]^2 \leq \sigma^2 \sum_{i=1}^n V_i^2 + 8\rho^2 C^2 \sum_{i=1}^n V_i \,\frac{i}{n}\left[1+\ln\left(\frac{n}{i}\right)\right].$$

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

$$\mathbb{E}\left[r_n(\mathbf{X})-r(\mathbf{X})\right]^2 \leq \sigma^2 \sum_{i=1}^n V_i^2 + \frac{8\rho^2 C^2}{1-2/d} \sum_{i=1}^n V_i \left\lfloor \frac{n}{i} \right\rfloor^{-2/d}.$$

**Proof of Theorem 2.1** Setting

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

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

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

The first term is easily bounded by noting that, for all $\mathbf{x} \in \mathbb{R}^d$,

$$\begin{aligned}
&\mathbb{E}\left[r_n(\mathbf{x})-\tilde{r}_n(\mathbf{x})\right]^2 \\
&= \mathbb{E}\left[\sum_{i=1}^n V_i \left(Y_{(i)}(\mathbf{x}) - r(\mathbf{X}_{(i)}(\mathbf{x}))\right)\right]^2 \\
&= \mathbb{E}\left[\sum_{i=1}^n V_i^2 \left(Y_{(i)}(\mathbf{x}) - r(\mathbf{X}_{(i)}(\mathbf{x}))\right)^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^n V_i^2 \sigma^2\left(\mathbf{X}_{(i)}(\mathbf{x})\right)\right] \\
&\leq \sigma^2 \sum_{i=1}^n V_i^2. \qquad (2.2)
\end{aligned}$$

To analyse the bias term in (2.1), we will need the following result, which bounds the convergence rate of the expected $i$-th nearest neighbor squared distance in terms of the metric covering radii of the support of the distribution $\mu$ of $\mathbf{X}$. Proposition 2.1 is a generalization of Theorem 1, page 1032 in Kulkarni and Posner [23], which only reports results for the rate of convergence of *the* nearest neighbor. Therefore, this result is interesting by itself.

**Proposition 2.1** *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 $\mathcal{A}$ in the Euclidean $d$-space, the covering radius satisfies $\mathcal{N}^{-1}(r,\mathcal{A}) \leq \mathcal{N}^{-1}(1,\mathcal{A})r^{-1/d}$ (see [22]). Hence the following corollary:

**Corollary 2.1** *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* $d\geq 3$,

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

Thus, to prove Theorem 2.1, it suffices to note from (2.1) and (2.2) that

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

Next,

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

Therefore, integrating with respect to the distribution of $\mathbf{X}$, we obtain

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

and the conclusion follows by applying Corollary 2.1. ■

Theorem 2.1 offers a general result, which can be made more precise according to the weights definition. Taking for example

$$V_i = \begin{cases} 1/k_n & \text{if } i \leq k_n \\ 0 & \text{otherwise,} \end{cases}$$

we get the ordinary $k_n$-NN rule back. Here,

$$\sum_{i=1}^{n} V_i^2 = \frac{1}{k_n}$$

and

$$\begin{aligned} \sum_{i=1}^{n} V_i \left\lfloor \frac{n}{i} \right\rfloor^{-2/d} &= \frac{1}{k_n} \sum_{i=1}^{k_n} \left\lfloor \frac{n}{i} \right\rfloor^{-2/d} \\ &\leq \frac{1}{k_n} \sum_{i=1}^{k_n} \left\lfloor \frac{n}{k_n} \right\rfloor^{-2/d} \\ &= \left\lfloor \frac{n}{k_n} \right\rfloor^{-2/d} \\ &\leq \xi \left(\frac{n}{k_n}\right)^{-2/d} \end{aligned}$$

for some positive $\xi$. Therefore, in this context, according to Theorem 2.1, for $d \geq 3$, there exists a sequence $(k_n)$ with $k_n \propto n^{\frac{2}{d+2}}$ such that

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

for some positive constant $\Lambda$ independent of $\rho$, $C$ and $\sigma^2$. This is exactly Theorem 6.2, page 93 of Györfi, Kohler, Krzyżak and Walk [17], which states that the standard nearest neighbor estimate is of optimum rate for the class $\mathcal{F}$ of $(1, C, \rho, \sigma^2)$-smooth distributions $(\mathbf{X}, Y)$ such that $\mathbf{X}$ has compact support with covering radius $\rho$, the regression function $r$ is Lipschitz with constant

$C$ and, for all $\mathbf{x} \in \mathbb{R}^d$, $\sigma^2(\mathbf{x}) = \mathbb{V}[Y \,|\, \mathbf{X} = \mathbf{x}] \leq \sigma^2$ (note however that the ordinary $k_n$-NN predictor is *not* optimal for higher smoothness, see Problem 6.2 in [17]).

The adaptation of Theorem 2.1 to the 1-NN bagged regression estimate needs more careful attention. This will be the topic of the next two subsections.

## 2.2 Bagging with replacement

This bagging-type is sometimes called moon-bagging, standing for **m o**ut **o**f **n b**ootstrap **agg**regat**ing**. As seen in the introduction, in this case, the weighted nearest neighbor regression estimate takes the form

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

where

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

From now on, $\Gamma(t)$ will denote the Gamma function, i.e.,

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

In order to make full use of Theorem 2.1, we first need a careful control of the term $\sum_{i=1}^n V_i^2$. This is done in the next proposition.

**Proposition 2.2** *For $i = 1, \ldots, n$, let*

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

*Then*

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

The message of Proposition 2.2 is that, when resampling is done with replacement, the variance term of the bagged NN estimate is $\mathrm{O}(k_n/n)$. Let us now turn to the bias term analysis.

**Proposition 2.3** *For $i = 1, \ldots, n$, let*

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

*Then*

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

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

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

$$\sum_{i=1}^{n} V_i \frac{i}{n}\left[1+\ln\left(\frac{n}{i}\right)\right] \leq \frac{2}{k_n}\left(1+\frac{1}{n}\right)^{k_n}\left[1+\ln(k_n+1)\right].$$

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

$$\sum_{i=1}^{n} V_i \left\lfloor \frac{n}{i} \right\rfloor^{-2/d} \leq \frac{1}{n^{k_n}} + \alpha_d \left(1+\frac{1}{n}\right)^{k_n} {k_n}^{-\frac{2}{d}},$$

*where*

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

The take-home message here is that, for $d \geq 3$, the squared bias is $\mathrm{O}(k_n^{-2/d})$. Finally, putting all the pieces together, we obtain

**Theorem 2.2** *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 positive constants* $\sigma^2$ *and* $C$. *Then*

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

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

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

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

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

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

*where*

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

By balancing the terms in Theorem 2.2, we are led to the following corollary:

**Corollary 2.2** *Under the assumptions of Theorem 2.2,*

(*i*) *If* $d = 1$*, there exists a sequence* $(k_n)$ *such that* $k_n \to \infty$*,* $k_n/n \to 0$*, and*

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

*for some positive constant* $\Lambda$ *independent of* $\rho$*,* $C$ *and* $\sigma^2$*.*

(*ii*) *If* $d = 2$*, there exists a sequence* $(k_n)$ *such that* $k_n \to \infty$*,* $k_n/n \to 0$*, and*

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

*for some positive constant* $\Lambda$ *independent of* $\rho$*,* $C$ *and* $\sigma^2$*.*

(*iii*) *If* $d \geq 3$*, there exists a sequence* $(k_n)$ *with* $k_n \propto n^{\frac{d}{d+2}}$ *such that*

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

*for some positive constant* $\Lambda$ *independent of* $\rho$*,* $C$ *and* $\sigma^2$*.*

Two important remarks are in order.

1. First, we note that, for $d \geq 3$ and a suitable choice of $k_n$, the bagged 1-NN estimate achieves both the minimax $n^{-2/(d+2)}$ rate and the optimal order of magnitude $((\rho C)^d\sigma^2)^{2/(d+2)}$ in the constant, for the class $\mathcal{F}$ of $(1, C, \rho, \sigma^2)$-smooth distributions $(\mathbf{X}, Y)$ such that $\mathbf{X}$ has compact support with covering radius $\rho$, the regression function $r$ is Lipschitz with constant $C$ and, for all $\mathbf{x} \in \mathbb{R}^d$, $\sigma^2(\mathbf{x}) = \mathbb{V}[Y \,|\, \mathbf{X} = \mathbf{x}] \leq \sigma^2$. Seconds, the bound is 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 $\rho$ and the smoothness of the regression function measured by the constant $C$. It is to correct this situation that we present adaptation results in subsection 2.4.

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 known to hold for the traditional $k_n$-NN regression estimate, which does not achieve the optimal rates in dimensions 1 and 2 (see Problems 6.1 and 6.7 in [17], Chapter 3).

## 2.3 Bagging without replacement

We briefly analyse in this subsection the rate of convergence of the bagged 1-NN regression 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 [4]. We know that, in this case, the weighted nearest neighbor regression estimate takes the form

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

where

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

Due to the fact that there is no repetition in the sampling process, the analysis turns out to be simpler. To prove Theorem 2.3 below, we start again by a control of the variance term $\sum_{i=1}^n V_i^2$.

**Proposition 2.4** *For $i = 1, \ldots, n$, let*

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

*Then*

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

Thus, as for bagging with replacement, the variance term of the without replacement bagged 1-NN estimate is $\mathrm{O}(k_n/n)$. The bias term may be treated by resorting to Theorem 2.1, via complex calculations due to the complicate form of the bagging weights. However, a much simpler route may be followed. Recall that

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

and observe that

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

where $\mathbf{X}^\star_{(1)}(\mathbf{x})$ is the nearest neighbor of $\mathbf{x}$ in a random subsample $\mathcal{S}_n$ drawn without replacement from $\{(\mathbf{X}_1,Y_1),\ldots,(\mathbf{X}_n,Y_n)\}$ 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 $\mathcal{D}_n$. This is the basic ingredient for the proof of the next proposition.

**Proposition 2.5** *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* $\mathbf{x}' \in \mathbb{R}^d$,

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

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

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

$$\mathbb{E}\left[\tilde{r}_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 \le \frac{16\rho^2C^2}{k_n}.$$

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

$$\mathbb{E}\left[\tilde{r}_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 \le \frac{8\rho^2C^2}{k_n}(1+\ln k_n).$$

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

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

Thus, for $d\ge 3$, $\mathbb{E}\left[\tilde{r}_n^\star(\mathbf{X}) - r(\mathbf{X})\right]^2 = \mathrm{O}(k_n^{-2/d})$. Combining Proposition 2.4 and Proposition 2.5 leads to the desired theorem:

**Theorem 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* $\mathbf{x}' \in \mathbb{R}^d$,

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

*and*

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

*for some positive constants* $\sigma^2$ *and* $C$. *Then*

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

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

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

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

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

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

By balancing the variance and bias terms, we obtain the following useful corollary:

**Corollary 2.3** *Under the assumptions of Theorem 2.3,*

*(i) If* $d = 1$, *there exists a sequence* $(k_n)$ *such that* $k_n \to \infty$, $k_n/n \to 0$, *and*

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

*for some positive constant* $\Lambda$ *independent of* $\rho$, $C$ *and* $\sigma^2$.

*(ii) If* $d = 2$, *there exists a sequence* $(k_n)$ *such that* $k_n \to \infty$, $k_n/n \to 0$, *and*

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

*for some positive constant* $\Lambda$ *independent of* $\rho$, $C$ *and* $\sigma^2$.

*(iii) If $d \geq 3$, there exists a sequence $(k_n)$ with $k_n \propto n^{\frac{d}{d+2}}$ such that*

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

*for some positive constant $\Lambda$ independent of $\rho$, $C$ and $\sigma^2$.*

As in bagging with replacement, Corollary 2.3 expresses the fact that, for $d \geq 3$, the without replacement bagged 1-NN estimate asymptotically achieves both the minimax $n^{-2/(d+2)}$ rate of convergence and the optimal order of magnitude $((\rho C)^d \sigma^2)^{d/(d+2)}$ in the constant, for the class $\mathcal{F}$ of $(1, C, \rho, \sigma^2)$-smooth distributions $(\mathbf{X}, Y)$.

## 2.4 Adaptation

In the previous subsections, the parameter $k_n$ of the estimate with the optimal rate of convergence for the class $\mathcal{F}$ depends on the unknown distribution of $(\mathbf{X}, Y)$, especially on the smoothness of the regression function measured by the Lipschitz constant $C$. In this subsection, we present a data-dependent way of choosing the resampling size $k_n$ and show that, for bounded $Y$, the estimate with parameter chosen in such an adaptive way achieves the optimal rate of convergence (irrespectively of the resampling type). To this aim, we split the sample $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ in two parts of size $\lfloor n/2 \rfloor$ and $n - \lfloor n/2 \rfloor$, respectively (assuming $n \geq 2$). The first half is denoted by $\mathcal{D}_n^\ell$ (learning set) and is used to construct the bagged 1-NN estimate $r^\star_{\lfloor n/2 \rfloor}(\mathbf{x}, \mathcal{D}_n^\ell) = r^\star_{k, \lfloor n/2 \rfloor}(\mathbf{x}, \mathcal{D}_n^\ell)$ (for the sake of clarity, we make the dependence of the estimate upon $k$ explicit). The second half of the sample, denoted by $\mathcal{D}_n^t$ (testing set), is used to choose $k$ by picking $\hat{k}_n \in K = \{1, \ldots, \lfloor n/2 \rfloor\}$ to minimize the empirical risk

$$\frac{1}{n - \lfloor n/2 \rfloor} \sum_{i=\lfloor n/2 \rfloor + 1}^{n} \left(Y_i - r^\star_{k, \lfloor n/2 \rfloor}(\mathbf{X}_i)\right)^2.$$

Define the estimate

$$r_n^\star(\mathbf{x}) = r^\star_{\hat{k}_n, \lfloor n/2 \rfloor}(\mathbf{x}, \mathcal{D}_n^\ell),$$

and note that $r_n^\star$ depends on the entire data $\mathcal{D}_n$. If $|Y| \leq L < \infty$ almost surely, a straightforward adaptation of Theorem 7.1 in [17] shows that, for any $\delta > 0$,

$$\mathbb{E}[r_n^\star(\mathbf{X}) - r(\mathbf{X})]^2 \leq (1 + \delta) \inf_{k \in K} \mathbb{E}[r^\star_{k, \lfloor n/2 \rfloor}(\mathbf{X}) - r(\mathbf{X})]^2 + \Xi \frac{\ln n}{n},$$

for some positive constant $\Xi$ depending only on $L$, $d$ and $\delta$. Immediately from Corollary 2.2 and Corollary 2.3 we can conclude:

**Theorem 2.4** *Suppose that $|Y| \leq L$ almost surely, and let $r_n^\star$ be the bagged 1-NN estimate with $k \in K = \{1, \ldots, \lfloor n/2 \rfloor\}$ chosen by data-splitting, irrespectively of the resampling type. Then $(\ln n)^{(d+2)/(2d)} n^{-1/2} \leq \rho C$ together with $d \geq 3$ implies, for $n \geq 2$,*

$$\mathbb{E}[r_n^\star(\mathbf{X}) - r(\mathbf{X})]^2 \leq (\Lambda + \mathrm{o}(1)) \left( \frac{(\rho C)^d}{n} \right)^{\frac{2}{d+2}},$$

*for some positive constant $\Lambda$ which depends only on $L$ and $d$.*

Thus, the expected error of the estimate obtained via data-splitting is bounded from above up to a constant by the corresponding minimax lower bound for the class $\mathcal{F}$ of regression functions, with the optimal dependence in $C$ and $\rho$.

# 3 Proofs

## 3.1 Proof of Proposition 2.1

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

$$D_{(i)}(\mathbf{X}) = \|\mathbf{X}_{(i)}(\mathbf{X}) - \mathbf{X}\|.$$

Clearly,

$$\begin{aligned}\mathbb{P}\left(D_{(i)}^2(\mathbf{X}) > \varepsilon\right) &= \mathbb{E}\left[\mathbb{P}\left(D_{(i)}(\mathbf{X}) > \sqrt{\varepsilon} \,|\, \mathbf{X}\right)\right] \\ &= \mathbb{E}\left[\sum_{j=0}^{i-1} \binom{n}{j} \left[F_{\mathbf{X}}\left(\sqrt{\varepsilon}\right)\right]^j \left[1 - F_{\mathbf{X}}\left(\sqrt{\varepsilon}\right)\right]^{n-j}\right]. \qquad (3.1)\end{aligned}$$

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,

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

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

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

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

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

leading to

$$\mathbb{E}\left[\frac{1}{F_{\mathbf{X}}\left(\sqrt{\varepsilon}\right)}\right] \leq \mathcal{N}\left(\frac{\sqrt{\varepsilon}}{2}\right). \tag{3.2}$$

Consequently, combining inequalities (3.1), (3.2) and technical Lemma 3.1, we obtain

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

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

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

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

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

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

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

as desired.

## 3.2 Proof of Corollary 2.1

For any bounded set $\mathcal{A}$ in the Euclidean $d$-space, the covering radius satisfies $\mathcal{N}^{-1}(r, \mathcal{A}) \leq \mathcal{N}^{-1}(1, \mathcal{A})r^{-1/d}$ (see [22]). Consequently, using Proposition 2.1, 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 $d \geq 3$,

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

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

## 3.3 Proof of Proposition 2.2

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,

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

Clearly,

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

Putting all the 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 concludes the proof of the proposition.

### 3.4 Proof of Proposition 2.3

We distinguish between the cases $d = 1$, $d = 2$ and $d \geq 3$.

$(i)$ If $d = 1$, for $i = 1, \ldots, n$, by definition of the $V_i$'s,

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

Thus

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

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

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

which is clearly true for $j = k_n$, without the factor 2 in front of the integral. For $j < k_n$, it is illustrated in Figure 1, where we have plotted the function $f(x) = x(1-x)^{k_n - j}$. The factor 2 is necessary because $f$ is not monotonic on $[0, 1]$.

Consequently,

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

Recalling the general formula

$$\int_0^1 x^{p-1}(1-x)^{q-1} \mathrm{d}x = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}, \quad p, q > 0, \tag{3.3}$$

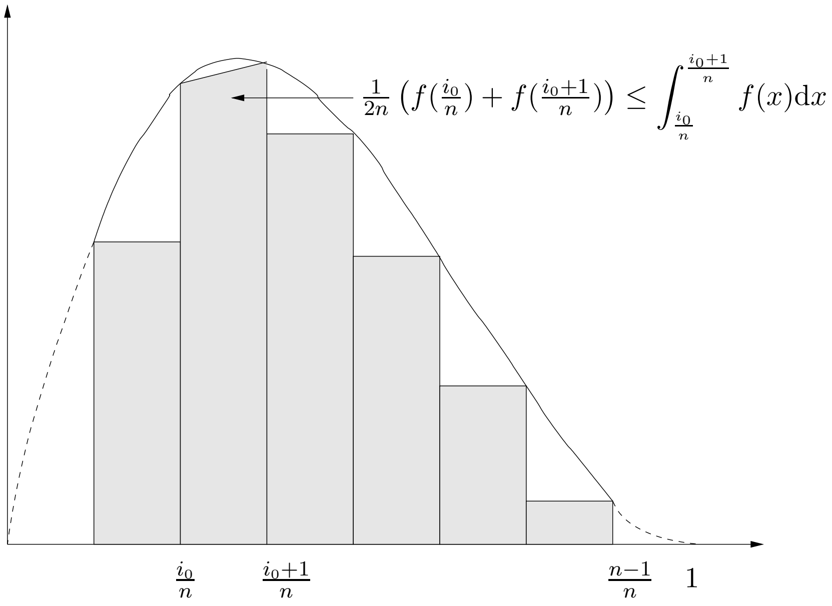
$\frac{1}{2n}\left(f(\frac{i_0}{n})+f(\frac{i_0+1}{n})\right) \leq \int_{\frac{i_0}{n}}^{\frac{i_0+1}{n}} f(x)\mathrm{d}x$
$\frac{i_0}{n}$
$\frac{i_0+1}{n}$
$\frac{n-1}{n}$
1

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

we obtain

$$\begin{aligned}
\sum_{i=1}^{n} V_i\,\frac{i}{n} &\leq 2\sum_{j=1}^{k_n}\binom{k_n}{j}\frac{1}{n^{j-1}}\,\frac{\Gamma(2)\Gamma(k_n-j+1)}{\Gamma(k_n-j+3)}\\
&= 2\sum_{j=1}^{k_n}\binom{k_n}{j}\frac{1}{n^{j-1}}\,\frac{1}{(k_n-j+1)(k_n-j+2)}\\
&= 2\sum_{j=1}^{k_n}\binom{k_n}{j-1}\frac{1}{n^{j-1}}\,\frac{1}{j(k_n-j+2)}\\
&= 2\sum_{j=0}^{k_n-1}\binom{k_n}{j}\frac{1}{n^{j}}\,\frac{1}{(j+1)(k_n-j+1)}.
\end{aligned}$$

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

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

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

$$\begin{aligned}\sum_{i=1}^{n} V_i \frac{i}{n}\left[1+\ln\left(\frac{n}{i}\right)\right] & \\ \leq 2\Bigg[\frac{1}{k_n}\left(1+\frac{1}{n}\right)^{k_n} & \\ -\sum_{j=1}^{k_n}\binom{k_n}{j}\frac{1}{n^{j-1}}\int_0^1 x(1-x)^{k_n-j}\ln x\,\mathrm{d}x\Bigg]. & \qquad (3.4)\end{aligned}$$

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

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

we have, for all $m \geq 0$ (see for example [15], formula (4.253.1)),

$$-\int_0^1 x(1-x)^m \ln x\,\mathrm{d}x = \frac{H_{m+2}-1}{(m+1)(m+2)}.$$

Thus we may write

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

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,

$$\begin{aligned}
&-\sum_{j=1}^{k_n}\binom{k_n}{j}\frac{1}{n^{j-1}}\int_0^1 x(1-x)^{k_n-j}\ln x\,\mathrm{d}x\\
&\leq \frac{\ln(k_n+1)}{k_n}\sum_{j=0}^{k_n-1}\binom{k_n}{j}\frac{1}{n^j}\\
&\leq \frac{\ln(k_n+1)}{k_n}\left(1+\frac{1}{n}\right)^{k_n}. \qquad (3.5)
\end{aligned}$$

Combining inequalities (3.4) and (3.5) leads to the desired result.

(*iii*) For $d \geq 3$, we note that for all $i = 1, \ldots, n-1$,

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

and set consequently

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

We obtain

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

Applying formula (3.3) 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

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

Thus, by technical Lemma 3.2,

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

This concludes the proof of Proposition 2.3.

## 3.5 Proof of Proposition 2.4

We have, 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

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

Observing finally that

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

we conclude that

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

## 3.6 Proof of Proposition 2.5

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 $\{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ 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,

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

Since $\mathrm{Card}(\mathcal{S}_n) = k_n$, we conclude by applying Corollary 2.1, with $i = 1$ and replacing $n$ by $k_n$.

## 3.7 Two technical lemmas

**Lemma 3.1** *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 $i = 1, \ldots, n$,*

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

**Proof of Lemma 3.1** 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 3.2** *For each $d \geq 3$, each $k_n \geq 1$, and $j = 0, \ldots, k_n - 1$, we have*

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

**Proof of Lemma 3.2** 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

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

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

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

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

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

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

**Acknowledgments.** The authors are greatly indebted to Luc Devroye for valuable comments and insightful suggestions on the first draft of the paper. They also thank two referees and the Acting Editor for their careful reading of the paper and thoughtful critical comments.

# 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. `http://www.lsta.upmc.fr/BIAU/bd4.pdf`

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

[3] Breiman, L. (2001). Random forests, *Machine Learning*, **45**, 5-32.

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

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

[6] Cover, T.M. (1968a). Estimation by the nearest neighbor rule, *IEEE Transactions on Information Theory*, **14**, 50-55.

[7] Cover, T.M. (1968b). Rates of convergence for nearest neighbor procedures, in *Proceedings of the Hawaii International Conference on Systems Sciences*, 413-415, Honolulu.

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

[9] Devroye, L. (1981). On the almost everywhere convergence of nonparametric regression function estimates, *The Annals of Statistics*, **9**, 1310-1319.

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

[11] 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.

[12] 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.

[13] 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.

[14] Friedman, J.H. and Hall, P. (2000). On bagging and nonlinear estimation, *Journal of Statistical Planning and Inference*, **137**, 669-683.

[15] Gradshteyn, I.S. and Ryzhik, I.M. (2007). *Table of Integrals, Series, and Products*, Academic Press, New York.

[16] Györfi, L. (1978). On the rate of convergence of nearest neighbor rules, *IEEE Transactions on Information Theory*, **29**, 509-512.

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

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

[19] Ibragimov, I.A. and Khasminskii, R.Z. (1980). On nonparametric estimation of regression, *Doklady Akademii Nauk SSSR*, **252**, 780-784.

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

[21] Ibragimov, I.A. and Khasminskii, R.Z. (1982). On the bounds for quality of nonparametric regression function estimation, *Theory of Probability and its Applications*, **27**, 81-94.

[22] 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.

[23] 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.

[24] Lin, Y. and Jeon, Y. (2006). Random forests and adaptive nearest neighbours, *Journal of the American Statistical Association*, **101**, 578-590.

[25] Psaltis, D., Snapp, R.R. and Venkatesh, S.S. (1994). On the finite sample performance of the nearest neighbor classifier, *IEEE Transactions on Information Theory*, **40**, 820-837.

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

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

[28] Venkatesh, S.S., Snapp, R.R. and Psaltis, D. (1992). Bellman strikes again! The growth rate of sample complexity with dimension for the nearest neighbor classifier, in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 93-102, Pittsburgh.