On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification[*]

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

Luc Devroye
School of Computer Science, McGill University
Montreal, Canada H3A 2K6
lucdevroye@gmail.com

September 16, 2008

Abstract

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be identically distributed random vectors in $\mathbb{R}^d$, independently drawn according to some probability density. An observation $\mathbf{X}_i$ is said to be a layered nearest neighbour (LNN) of a point $\mathbf{x}$ if the hyperrectangle defined by $\mathbf{x}$ and $\mathbf{X}_i$ contains no other data points. We first establish consistency results on $L_n(\mathbf{x})$, the number

[*] The second author's research was sponsored by NSERC Grant A3456 and FQRNT Grant 90-ER-0291.

of LNN of $\mathbf{x}$. Then, given a sample $(\mathbf{X}, Y), (\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ of independent identically distributed random vectors from $\mathbb{R}^d \times \mathbb{R}$, one may estimate the regression function $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ by the LNN estimator $m_n(\mathbf{x})$, defined as an average over the Y_i's corresponding to those $\mathbf{X}_i$ with are LNN of $\mathbf{x}$. Under mild conditions on r, we establish consistency of $\mathbb{E}|r_n(\mathbf{x}) - r(\mathbf{x})|^p$ towards 0 as $n \to \infty$, for almost all $\mathbf{x}$ and all $p \geq 1$, and discuss the links between r_n and the random forest estimates of Breiman [7]. We finally show the universal consistency of the bagged (bootstrap-aggregated) nearest neighbour method for regression and classification.

1 Introduction

Let $\mathcal{D}_n = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ be a sample of independent and identically distributed (i.i.d.) random vectors in $\mathbb{R}^d$, $d \geq 2$. An observation $\mathbf{X}_i$ is said to be a *layered nearest neighbour* (LNN) of a target point $\mathbf{x}$ if the hyperrectangle defined by $\mathbf{x}$ and $\mathbf{X}_i$ contains no other data points. As illustrated in Figure 1 below, the number of LNN of $\mathbf{x}$ is typically larger than one and depends on the number and configuration of the sample points.

The LNN concept, which is briefly discussed in the monograph [12, Chapter 11, Problem 11.6], has strong connections with the notions of *dominance* and *maxima* in random vectors. Recall that a point $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})$ is said to be dominated by $\mathbf{X}_j$ if $X_{ik} \leq X_{jk}$ for all $k = 1, \ldots, d$, and a point $\mathbf{X}_i$ is called a maximum of $\mathcal{D}_n$ if none of the other points dominates. One can distinguish between *dominance* ($\mathbf{X}_{ik} \leq \mathbf{X}_{jk}$ for all k), *strong dominance* (at least one inequality is strict) and *strict dominance* (all inequalities are strict). The actual kind of dominance will not matter in this paper because we will assume throughout that the common distribution of the data has a density, so that equality of coordinates happens with zero probability. The study of the maxima of a set of points was initiated by Barndorff-Nielsen and Sobel [4] as an attempt to describe the boundary of a set of random points in $\mathbb{R}^d$. Dominance deals with the natural order relations for multivariate observations. Due to its close relationship with the convex hull, dominance is important in computational geometry, pattern classification, graphics, economics and data analysis. The reader is referred to Bai et al. [3] for more information

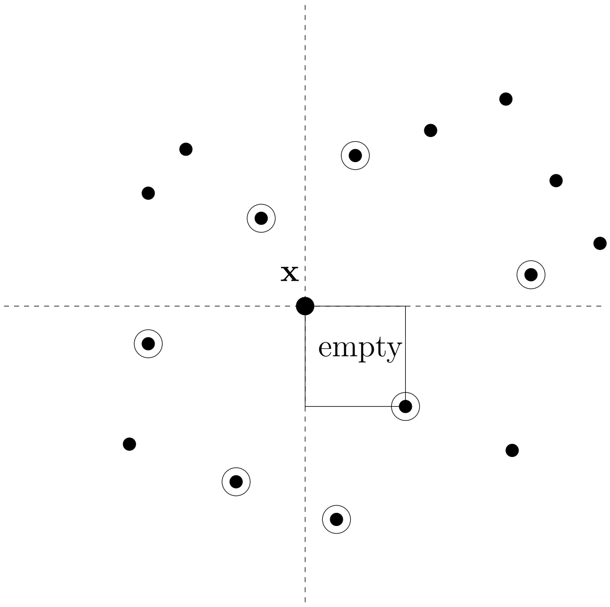

Figure 1: The layered nearest neighbours (LNN) of a point $\mathbf{x}$.

and references.

Denote by L_n the number of maxima in the set $\mathcal{D}_n$. Under the assumption that the components of each vector of $\mathcal{D}_n$ are independently and continuously distributed, a lot is known about the statistical properties of L_n (Barndorff-Nielsen and Sobel [4], Bai et al. [2, 3]). For example, it can be shown that

$$\mathbb{E}L_n = \frac{(\log n)^{d-1}}{(d-1)!} + \mathcal{O}\left((\log n)^{d-2}\right),$$

and

$$\frac{(d-1)!\, L_n}{(\log n)^{d-1}} \to 1 \quad \text{in probability},$$

as $n \to \infty$ and $d \geq 2$ is fixed. From this, one deduces that when the random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independently and uniformly distributed over $(0,1)^d$, then the number of LNN of any point $\mathbf{x}$ in $(0,1)^d$, denoted hereafter by $L_n(\mathbf{x})$, satisfies

$$\mathbb{E}L_n(\mathbf{x}) = \frac{2^d(\log n)^{d-1}}{(d-1)!} + \mathcal{O}\left((\log n)^{d-2}\right)$$

and

$$\frac{(d-1)!\, L_n(\mathbf{x})}{2^d (\log n)^{d-1}} \to 1 \quad \text{in probability as } n \to \infty.$$

Here, the extra factor represents the contribution of the 2^d quadrants surrounding the point $\mathbf{x}$.

On the other hand, to the best of our knowledge, little if nothing is known about the behavior of $L_n(\mathbf{x})$ under the much more general assumption that the sample points are distributed according to some arbitrary (i.e., non-necessarily uniform) probability density. A quick inspection reveals that the analysis of this important case is non-trivial and that it may not be readily deduced from the above-mentioned results. Thus, the first objective of this paper is to fill the gap and to offer consistency results about $L_n(\mathbf{x})$ when $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independently drawn according to some arbitrary probability density f. This will be the topic of section 2.

Next, we wish to emphasize that the LNN concept has also important statistical consequences. To formalize this idea, suppose that we are given a sequence $(\mathbf{X}, Y), (\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ of i.i.d. $\mathbb{R}^d \times \mathbb{R}$-valued random variables with $\mathbb{E}|Y| < \infty$. Then, denoting by $\mathcal{L}_n(\mathbf{x})$ the set of LNN of $\mathbf{x} \in \mathbb{R}^d$, the regression function $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ may be estimated by

$$r_n(\mathbf{x}) = \frac{1}{L_n(\mathbf{x})} \sum_{i=1}^{n} Y_i \mathbf{1}_{[\mathbf{X}_i \in \mathcal{L}_n(\mathbf{x})]}.$$

(Note that $L_n(\mathbf{x}) \geq 1$, so that the division makes sense). In other words, the estimate $r_n(\mathbf{x})$ just outputs the average of the Y_i's associated with the LNN of the target point $\mathbf{x}$.

The interest of studying the LNN regression estimate r_n, which was first mentioned in [12], is threefold. Firstly, we observe that this estimate has no smoothing parameter, a somewhat unusual situation in nonparametric estimation. Secondly, it is scale-invariant, which is clearly a desirable feature when the components of the vector represent physically different quantities. And thirdly, it turns out that r_n is intimately related to the *random forests* classification and regression estimates, which were defined by Breiman in [7].

Breiman [7] takes data $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ and partitions $\mathbb{R}^d$ randomly into "pure" rectangles, i.e., rectangles that each contain one data point. If $A(\mathbf{X})$ is the rectangle to which $\mathbf{X}$ belongs, then $\mathbf{X}$ votes "Y_i", where $\mathbf{X}_i$ is the unique data point in $A(\mathbf{X})$. Breiman repeats such voting and call the principle "random forests". Classification is done by a majority vote. Regression is done by averaging all Y_i's. Observe that each voting $\mathbf{X}_i$ is a LNN of $\mathbf{X}$, so that random forests lead to a weighted LNN estimate. In contrast, the estimate r_n above assigns uniform weights. Biau et al. [8] point out that the random forest method is not universally consistent, but the question of consistency remains open when $\mathbf{X}$ is assumed to have a density.

This paper is indeed concerned with minimal conditions of convergence. We say that a regression function estimate r_n is L_p-consistent ($p > 0$) if $\mathbb{E}|r_n(\mathbf{X}) - r(\mathbf{X})|^p \to 0$, as $n \to \infty$, where $\mathbf{X}$ is independent of and distributed as $\mathbf{X}_1$. It is universally L_p-consistent if this property is true for all distributions of $(\mathbf{X}_1, Y_1)$ with $\mathbb{E}|Y_1|^p < \infty$. Universal consistency was the driving theme of the monograph [12], and we try as much as possible to adhere to the style and notation of that text.

In classification, we have $Y \in \{0,1\}$, and construct a $\{0,1\}$-valued estimate $g_n(\mathbf{x})$ of Y. This is related to regression function estimation since one could use a regression function estimate $r_n(\mathbf{x})$ of $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$, and set

$$g_n(\mathbf{x}) = \mathbf{1}_{[r_n(\mathbf{x}) \geq 1/2]}. \tag{1}$$

That estimate has the property that if $\mathbb{E}|r_n(\mathbf{X}) - r(\mathbf{X})| \to 0$ as $n \to \infty$, then

$$\mathbb{P}(g_n(\mathbf{X}) \neq Y) \to \inf_{g:\mathbb{R}^d \to \{0,1\}} \mathbb{P}(g(\mathbf{X}) \neq Y),$$

a property which is called Bayes risk consistency (see [12]). It is universally Bayes risk consistent if this property is true for all distributions of $(\mathbf{X}_1, Y_1)$.

Random forests are some of the most successful ensemble methods that exhibit performance on the level of boosting and support vector machines. Fast and robust to noise, random forests do not overfit and offer possibilities for explanation and visualization of the input, such as variable selection. Moreover, random forests have been shown to give excellent performance on a number of practical problems and are among the most accurate general-purpose regression methods available.

An important attempt to investigate the driving force behind consistency of random forests is due to Lin and Jeon [17], who show that a forest can be seen as an adaptively weighted LNN regression estimate and argue that the LNN approach provides an interesting data-dependent way of measuring proximities between observations.

As a new step towards understanding random forests, we study the consistency of the (uniformly weighted) LNN regression estimate r_n and toroughly discuss the links between r_n and the random forest estimates of Breiman [7] (section 3). We finally show in section 4 the universal consistency of the bagged (bootstrap-aggregated) nearest neighbour method for regression and classification. Proofs of some technical results are gathered in section 5.

2 Some consistency properties of the LNN

Throughout this section, we let $\mathcal{D}_n = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ be $\mathbb{R}^d$-valued ($d \geq 2$) independent random variables, identically distributed according to some probability density f with respect to the Lebesgue measure λ. For any $\mathbf{x} \in \mathbb{R}^d$, we denote by $\mathcal{L}_n(\mathbf{x})$ the set of layered nearest neighbours (LNN) of $\mathbf{x}$ in $\mathcal{D}_n$, and we let $L_n(\mathbf{x})$ be the cardinality of $\mathcal{L}_n(\mathbf{x})$ (i.e., $L_n(\mathbf{x}) = |\mathcal{L}_n(\mathbf{x})|$). Finally, we denote the probability measure associated to f by μ.

We will prove the following two basic consistency theorems:

Theorem 2.1 *For μ-almost all $\mathbf{x} \in \mathbb{R}^d$, one has*

$$L_n(\mathbf{x}) \to \infty \quad \textit{in probability as } n \to \infty.$$

Theorem 2.2 *Suppose that f is λ-almost everywhere continuous. Then*

$$\frac{(d-1)!\,\mathbb{E}L_n(\mathbf{x})}{2^d(\log n)^{d-1}} \to 1 \quad \textit{as } n \to \infty,$$

at μ-almost all $\mathbf{x} \in \mathbb{R}^d$.

In the sequel, for $\mathbf{x} = (x_1, \ldots, x_d)$ and $\varepsilon > 0$, we let the hyperrectangle $\mathcal{R}_\varepsilon(\mathbf{x})$ be defined as $\mathcal{R}_\varepsilon(\mathbf{x}) = [x_1, x_1 + \varepsilon] \times \ldots \times [x_d, x_d + \varepsilon]$. The crucial result from real analysis that is needed in the proof of Theorem 2.1 and Theorem 2.2 is the following (see for instance Wheeden and Zygmund [22]):

Lemma 2.1 *Let g be locally integrable in $\mathbb{R}^d$. Then, for λ-almost all $\mathbf{x}$,*

$$\frac{1}{\varepsilon^d}\int_{\mathcal{R}_\varepsilon(\mathbf{x})}|g(\mathbf{y})-g(\mathbf{x})|\,d\mathbf{y}\to 0 \quad \textit{as } \varepsilon\to 0. \tag{2}$$

Thus also, at λ-almost all $\mathbf{x}$,

$$\frac{1}{\varepsilon^d}\int_{\mathcal{R}_\varepsilon(\mathbf{x})}g(\mathbf{y})d\mathbf{y}\to f(\mathbf{x}) \quad \textit{as } \varepsilon\to 0. \tag{3}$$

The following useful corollary may be easily deduced from Lemma 2.1 and the fact that $f(\mathbf{x})>0$ for μ-almost all $\mathbf{x}$:

Corollary 2.1 *Let (ε_n) be a sequence of positive real numbers such that $\varepsilon_n\to 0$ and $n\varepsilon_n^d\to\infty$ as $n\to\infty$. Then, for μ-almost all $\mathbf{x}\in\mathbb{R}^d$, one has*

$$n\mu\left(\mathcal{R}_{\varepsilon_n}(\mathbf{x})\right)\to\infty \quad \textit{as } n\to\infty.$$

Remark Lemma 2.1 only describes what happens if $\mathcal{R}_\varepsilon(\mathbf{x})$ is in the positive quadrant of $\mathbf{x}$. Trivially, it also holds for the 2^d-1 other quadrants centered at $\mathbf{x}$.

The proof of Theorem 2.1 uses a coupling argument. Roughly, the idea is to replace the sample $\mathbf{X}_1,\ldots,\mathbf{X}_n$ by a sample $\mathbf{Z}_1,\ldots,\mathbf{Z}_n$ which is (i) locally uniform around the point $\mathbf{x}$ and (ii) such that the probability of the event $[\mathbf{X}_i=\mathbf{Z}_i, i=1,\ldots,n]^c$ stays under control. This can be achieved via the following optimal coupling lemma (see for example Doeblin [14] or Rachev and Rüschendorf [20]):

Lemma 2.2 *Let f and g be probability densities on $\mathbb{R}^d$. Then there exist random variables $\mathbf{W}$ and $\mathbf{Z}$ with density f and g, respectively, such that*

$$\mathbb{P}(\mathbf{W}\neq\mathbf{Z})=\frac{1}{2}\int_{\mathbb{R}^d}|f(\mathbf{y})-g(\mathbf{y})|\,d\mathbf{y}.$$

We are now in a position to prove Theorem 2.1.

Proof of Theorem 2.1 Fix $\mathbf{x}$ for which (2) is true, and define the function g_ε as

$$g_\varepsilon(\mathbf{y}) = \begin{cases} \dfrac{\mu\left(\mathcal{R}_\varepsilon(\mathbf{x})\right)}{\varepsilon^d} & \text{if } \mathbf{y} \in \mathcal{R}_\varepsilon(\mathbf{x}) \\ f(\mathbf{y}) & \text{otherwise.} \end{cases}$$

Clearly, g_ε is a probability density on $\mathbb{R}^d$. Moreover,

$$\begin{aligned} &\int_{\mathbb{R}^d} |f(\mathbf{y}) - g_\varepsilon(\mathbf{y})|\, d\mathbf{y} \\ &\quad = \int_{\mathcal{R}_\varepsilon(\mathbf{x})} \left| f(\mathbf{y}) - \frac{1}{\varepsilon^d} \int_{\mathcal{R}_\varepsilon(\mathbf{x})} f(\mathbf{z}) d\mathbf{z} \right| d\mathbf{y} \\ &\quad \leq \int_{\mathcal{R}_\varepsilon(\mathbf{x})} |f(\mathbf{y}) - f(\mathbf{x})|\, d\mathbf{y} + \varepsilon^d \left| f(\mathbf{x}) - \frac{1}{\varepsilon^d} \int_{\mathcal{R}_\varepsilon(\mathbf{x})} f(\mathbf{z}) d\mathbf{z} \right| \\ &\quad \leq 2 \int_{\mathcal{R}_\varepsilon(\mathbf{x})} |f(\mathbf{y}) - f(\mathbf{x})|\, d\mathbf{y} \\ &\quad \leq 2\varepsilon^d \Phi(\varepsilon), \end{aligned} \tag{4}$$

where $\Phi(\varepsilon)$ is some nonnegative, nondecreasing function which has limit 0 as ε approaches 0.

According to Lemma 2.2 and inequality (4), there exist random variables $\mathbf{W}$ and $\mathbf{Z}$ with density f and g_ε, respectively, such that

$$\mathbb{P}(\mathbf{W} \neq \mathbf{Z}) \leq \varepsilon^d \Phi(\varepsilon).$$

Now, define $\mathbf{W}$ and $\mathbf{Z}$ samples by creating n independent $(\mathbf{W}_1, \mathbf{Z}_1), \ldots, (\mathbf{W}_n, \mathbf{Z}_n)$ pairs and assume, without loss of generality, that $(\mathbf{X}_1, \ldots, \mathbf{X}_n) = (\mathbf{W}_1, \ldots, \mathbf{W}_n)$. Thus, denoting by $\mathcal{E}_n$ the event

$$[\mathbf{X}_i = \mathbf{Z}_i, i = 1, \ldots, n],$$

we obtain, by construction of the optimal coupling,

$$\mathbb{P}(\mathcal{E}_n^c) \leq n\varepsilon^d \Phi(\varepsilon). \tag{5}$$

According to technical Lemma 5.1, there exists a sequence (ε_n) of positive real numbers such that $\varepsilon_n \to 0$, $n\varepsilon_n^d \to \infty$ and $n\varepsilon_n^d \Phi(\varepsilon_n) \to 0$ as $n \to \infty$. Thus, by choosing such a sequence, according to (5), the probability $\mathbb{P}(\mathcal{E}_n^c)$

can be made as small as desired for all n large enough.

To finish the proof of Theorem 2.1, denote by $L_{\varepsilon_n}(\mathbf{x})$ (respectively $L'_{\varepsilon_n}(\mathbf{x})$) the number of LNN of $\mathbf{x}$ in the sample $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ (respectively in the sample $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$) falling in $\mathcal{R}_{\varepsilon_n}(\mathbf{x})$. Clearly,

$$L_n(\mathbf{x}) \geq L_{\varepsilon_n}(\mathbf{x}), \tag{6}$$

and, on the event $\mathcal{E}_n$,

$$L_{\varepsilon_n}(\mathbf{x}) = L'_{\varepsilon_n}(\mathbf{x}). \tag{7}$$

By Lemma 5.2, since $n\varepsilon_n^d \to \infty$,

$$L'_{\varepsilon_n}(\mathbf{x}) \to \infty \quad \text{in probability as } n \to \infty,$$

at μ-almost all $\mathbf{x}$. This, together with (5)-(7) concludes the proof of the theorem. ■

Proof of Theorem 2.2 For $\mathbf{x} = (x_1, \ldots, x_d)$ and $\varepsilon > 0$, let $\mathcal{C}_\varepsilon(\mathbf{x})$ be the hypercube $[x_1-\varepsilon, x_1+\varepsilon] \times \ldots \times [x_d-\varepsilon, x_d+\varepsilon]$. Choose $\mathbf{x}$ in a set of μ-measure 1 such that $f(\mathbf{x}) > 0$, f is continuous at $\mathbf{x}$ and $\mu(\mathcal{C}_\varepsilon(\mathbf{x})) > 0$ for all $\varepsilon > 0$.

Fix $\delta \in (0, f(\mathbf{x}))$. Since f is continuous at $\mathbf{x}$, there exists $\varepsilon > 0$ such that $\mathbf{y} \in \mathcal{C}_\varepsilon(\mathbf{x})$ implies $|f(\mathbf{x}) - f(\mathbf{y})| < \delta$.

Let the hyperrectangle $\mathcal{R}_{(\mathbf{x},\mathbf{y})}$ be defined by $\mathbf{x}$ and $\mathbf{y}$. We note that

$$\begin{aligned}
\mathbb{E}L_n(\mathbf{x}) &= n \int_{\mathbb{R}^d} \left(1 - \mu(\mathcal{R}_{(\mathbf{x},\mathbf{y})})\right)^{n-1} f(\mathbf{y})d\mathbf{y} \\
&= n \int_{\mathbb{R}^d} \left(1 - \int_{\mathcal{R}_{(\mathbf{x},\mathbf{y})}} f(\mathbf{z})d\mathbf{z}\right)^{n-1} f(\mathbf{y})d\mathbf{y}.
\end{aligned}$$

Thus, using the continuity of f at $\mathbf{x}$, we obtain

$$\begin{aligned}
\mathbb{E}L_n(\mathbf{x}) &\geq n(f(\mathbf{x}) - \delta) \int_{\mathcal{C}_\varepsilon(\mathbf{x})} (1 - (f(\mathbf{x}) + \delta)\Pi|y_i - x_i|)^{n-1} d\mathbf{y} \\
&= n(f(\mathbf{x}) - \delta) \int_{\mathcal{C}_\varepsilon(\mathbf{0})} (1 - (f(\mathbf{x}) + \delta)\Pi|y_i|)^{n-1} d\mathbf{y}
\end{aligned}$$

$$
\begin{aligned}
&= n\,\frac{f(\mathbf{x})-\delta}{f(\mathbf{x})+\delta}\int_{\mathcal{C}_{\Delta\varepsilon}(\mathbf{0})} (1-\Pi|y_i|)^{n-1}\,d\mathbf{y} \\
&\quad \text{(with } \Delta=(f(\mathbf{x})+\delta)^{1/d}\text{)} \\
&= n\,2^d\,\frac{f(\mathbf{x})-\delta}{f(\mathbf{x})+\delta}\int_{\mathcal{R}_{\Delta\varepsilon}(\mathbf{0})} (1-\Pi y_i)^{n-1}\,d\mathbf{y},
\end{aligned}
$$

where the last equality follows from a symmetry argument. Thus, using technical Lemma 5.3, we conclude that

$$
\mathbb{E}L_n(\mathbf{x}) \ge 2^d\,\frac{f(\mathbf{x})-\delta}{f(\mathbf{x})+\delta}\left[\frac{(\log n)^{d-1}}{(d-1)!}+\mathcal{O}_{\Delta\varepsilon}(\log n)^{d-2}\right],
$$

where the notation $\mathcal{O}_{\Delta\varepsilon}$ means that the constant in the $\mathcal{O}$ term depends on $\Delta\varepsilon$. Letting $\delta\to 0$ shows that

$$
\liminf_{n\to\infty}\frac{(d-1)!\,\mathbb{E}L_n(\mathbf{x})}{2^d(\log n)^{d-1}} \ge 1.
$$

To show the opposite inequality, we write, using the continuity of f at $\mathbf{x}$,

$$
\begin{aligned}
\mathbb{E}L_n(\mathbf{x}) &= n\int_{\mathcal{C}_\varepsilon(\mathbf{x})}\left(1-\mu(\mathcal{R}_{(\mathbf{x},\mathbf{y})})\right)^{n-1} f(\mathbf{y})d\mathbf{y} \\
&\quad + n\int_{\mathbb{R}^d\setminus\mathcal{C}_\varepsilon(\mathbf{x})}\left(1-\mu(\mathcal{R}_{(\mathbf{x},\mathbf{y})})\right)^{n-1} f(\mathbf{y})d\mathbf{y} \\
&\le n\,2^d\,\frac{f(\mathbf{x})+\delta}{f(\mathbf{x})-\delta}\int_{\mathcal{R}_{\Delta\varepsilon}(\mathbf{0})}(1-\Pi y_i)^{n-1}d\mathbf{y} \\
&\quad \text{(with } \Delta=(f(\mathbf{x})-\delta)^{1/d}\text{)} \\
&\quad + n\int_{\mathbb{R}^d\setminus\mathcal{C}_\varepsilon(\mathbf{x})}\left(1-\mu(\mathcal{R}_{(\mathbf{x},\mathbf{y})})\right)^{n-1} f(\mathbf{y})d\mathbf{y}. \qquad (8)
\end{aligned}
$$

By technical Lemma 5.3, we have

$$
\begin{aligned}
&n\,2^d\,\frac{f(\mathbf{x})+\delta}{f(\mathbf{x})-\delta}\int_{\mathcal{R}_{\Delta\varepsilon}(\mathbf{0})}(1-\Pi y_i)^{n-1}d\mathbf{y} \\
&\quad = 2^d\,\frac{f(\mathbf{x})+\delta}{f(\mathbf{x})-\delta}\left[\frac{(\log n)^{d-1}}{(d-1)!}+\mathcal{O}_{\Delta\varepsilon}(\log n)^{d-2}\right]. \qquad (9)
\end{aligned}
$$

Then, with respect to the second term in (8), we note that

$$
\mathbb{R}^d\setminus\mathcal{C}_\varepsilon(\mathbf{x}) = \bigcup_{j=0}^{d-1}\mathcal{C}_j,
$$

where, by definition, $\mathcal{C}_j$ denotes the collection of all $\mathbf{y}$ in $\mathbb{R}^d \backslash \mathcal{C}_\varepsilon(\mathbf{x})$ which have *exactly* j coordinates smaller than ε. Observe that, for each $j \in \{0, \ldots, d-1\}$,

$$\mathcal{C}_j = \bigcup_{\underline{j}} \mathcal{C}_{\underline{j}},$$

where the index $\underline{j}$ runs over the $\binom{d}{j}$ possible j-uples coordinate choices smaller than ε. Associated to each of these choices is a marginal density of f, that we denote by $f_{\underline{j}}$. For $j \geq 1$, with a slight abuse of notation, we let $\mathcal{C}_\varepsilon(\mathbf{x}_{\underline{j}})$ be the j-dimensional hypercube with center at the coordinates of $\mathbf{x}$ matching with $\underline{j}$ and side length 2ε. Finally, we choose ε small enough and $\mathbf{x}$ in a set of μ-measure 1 such that each marginal $f_{\underline{j}}$ is bounded over $\mathcal{C}_\varepsilon(\mathbf{x}_{\underline{j}})$ by, say, $\Lambda(\varepsilon)$.

Clearly, for $j = 0$,

$$\begin{aligned}
n & \int_{\mathcal{C}_0} \left(1 - \mu(\mathcal{R}_{(\mathbf{x},\mathbf{y})})\right)^{n-1} f(\mathbf{y}) d\mathbf{y} \\
& = n \int_{\mathcal{C}_0} \left(1 - \int_{\mathcal{R}_{(\mathbf{x},\mathbf{y})}} f(\mathbf{z}) d\mathbf{z}\right)^{n-1} f(\mathbf{y}) d\mathbf{y} \\
& \leq n(1 - (f(\mathbf{x}) - \delta)\varepsilon^d)^{n-1} \int_{\mathcal{C}_0} f(\mathbf{y}) d\mathbf{y} \\
& \leq n(1 - (f(\mathbf{x}) - \delta)\varepsilon^d)^{n-1} \\
& \quad \text{(since } f \text{ is a probability density)} \\
& \leq 1/\left[(f(\mathbf{x}) - \delta)\varepsilon^d\right],
\end{aligned}$$

where, in the last inequality, we used the fact that $\sup_{x \in [0,1]} x(1-x)^{n-1} \leq 1/n$. Similarly, for $j \in \{1, \ldots, d-1\}$, we may write

$$\begin{aligned}
n & \int_{\mathcal{C}_j} \left(1 - \mu(\mathcal{R}_{(\mathbf{x},\mathbf{y})})\right)^{n-1} f(\mathbf{y}) d\mathbf{y} \\
& = n \sum_{\underline{j}} \int_{\mathcal{C}_{\underline{j}}} \left(1 - \mu(\mathcal{R}_{(\mathbf{x},\mathbf{y})})\right)^{n-1} f(\mathbf{y}) d\mathbf{y} \\
& = n \sum_{\underline{j}} \int_{\mathcal{C}_{\underline{j}}} \left(1 - \int_{\mathcal{R}_{(\mathbf{x},\mathbf{y})}} f(\mathbf{z}) d\mathbf{z}\right)^{n-1} f(\mathbf{y}) d\mathbf{y}
\end{aligned}$$

$$\leq n \sum_{\underline{j}} \int_{\mathcal{C}_{\underline{j}}} \left(1 - (f(\mathbf{x}) - \delta)\varepsilon^{d-j}\Pi_\ell |y_\ell - x_\ell|\right)^{n-1} f(\mathbf{y})d\mathbf{y},$$

where the notation Π_ℓ means the product over the j coordinates which are smaller than ε. Thus, integrating the density f over those coordinates which are larger than ε, we obtain

$$\begin{aligned}
&\int_{\mathcal{C}_{\underline{j}}} \left(1 - (f(\mathbf{x}) - \delta)\varepsilon^{d-j}\Pi_\ell |y_\ell - x_\ell|\right)^{n-1} f(\mathbf{y})d\mathbf{y} \\
&\quad \leq \int_{\mathcal{C}_\varepsilon(\mathbf{x}_{\underline{j}})} \left(1 - (f(\mathbf{x}) - \delta)\varepsilon^{d-j}\Pi_\ell |y_\ell - x_\ell|\right)^{n-1} f_{\underline{j}}(\mathbf{y}_{\underline{j}})d\mathbf{y}_{\underline{j}}.
\end{aligned}$$

Using finally the fact that each marginal $f_{\underline{j}}$ is bounded by $\Lambda(\varepsilon)$ in the neighbourhood of $\mathbf{x}$, we obtain

$$\begin{aligned}
&\int_{\mathcal{C}_{\underline{j}}} \left(1 - (f(\mathbf{x}) - \delta)\varepsilon^{d-j}\Pi_\ell |y_\ell - x_\ell|\right)^{n-1} f(\mathbf{y})d\mathbf{y} \\
&\quad \leq \Lambda(\varepsilon) \int_{[0,\varepsilon]^j} \left(1 - (f(\mathbf{x}) - \delta)\varepsilon^{d-j} y_1 \dots y_j\right)^{n-1} dy_1 \dots dy_j \\
&\quad = \frac{\Lambda(\varepsilon)}{(f(\mathbf{x}) - \delta)\varepsilon^{d-j}} \int_{[0,\Delta\varepsilon^{d-j/j}]^j} (1 - y_1 \dots y_j)^{n-1} dy_1 \dots dy_j \\
&\qquad (\text{with } \Delta = (f(\mathbf{x}) - \delta)^{1/j}).
\end{aligned}$$

Therefore, by Lemma 5.3, for $j \in \{2, \dots, d-1\}$,

$$\begin{aligned}
&n \int_{\mathcal{C}_j} \left(1 - \mu(\mathcal{R}_{(\mathbf{x},\mathbf{y})})\right)^{n-1} f(\mathbf{y})d\mathbf{y} \\
&\quad \leq \frac{\binom{d}{j}\Lambda(\varepsilon)}{(f(\mathbf{x}) - \delta)\varepsilon^{d-j}} \left[\frac{(\log n)^{j-1}}{(j-1)!} + \mathcal{O}_{\Delta\varepsilon^{d-j/j}}(\log n)^{j-2} \right],
\end{aligned}$$

and clearly, for $j = 1$,

$$n \int_{\mathcal{C}_1} \left(1 - \mu(\mathcal{R}_{(\mathbf{x},\mathbf{y})})\right)^{n-1} f(\mathbf{y})d\mathbf{y} \leq \frac{\Lambda(\varepsilon)}{(f(\mathbf{x}) - \delta)\varepsilon^{d-1}}.$$

Putting all pieces together, we conclude that, for all $j \in \{0, \dots, d-1\}$,

$$\limsup_{n\to\infty} \frac{n}{(\log n)^{d-1}} \int_{\mathcal{C}_j} \left(1 - \mu(\mathcal{R}_{(\mathbf{x},\mathbf{y})})\right)^{n-1} f(\mathbf{y})d\mathbf{y} = 0,$$

and, consequently,

$$\limsup_{n\to\infty} \frac{n}{(\log n)^{d-1}} \int_{\mathbb{R}^d \setminus \mathcal{C}_\varepsilon(\mathbf{x})} \left(1-\mu(\mathcal{R}_{(\mathbf{x},\mathbf{y})})\right)^{n-1} f(\mathbf{y}) d\mathbf{y} = 0.$$

This, together with inequalities (8)-(9) and letting $\delta \to 0$ leads to

$$\limsup_{n\to\infty} \frac{(d-1)!\, \mathbb{E}L_n(\mathbf{x})}{2^d (\log n)^{d-1}} \leq 1.$$

■

3 LNN regression estimation

3.1 Consistency

Denote by $(\mathbf{X}, Y), (\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ i.i.d. random vectors of $\mathbb{R}^d \times \mathbb{R}$, and let $\mathcal{D}_n$ be the set of data defined by

$$\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}.$$

In this section we will assume that $|Y| \leq \gamma < \infty$ and that $\mathbf{X}$ has a density. We consider the general regression function estimation problem, where one wants to use the data $\mathcal{D}_n$ in order to construct an estimate $r_n : \mathbb{R}^d \to \mathbb{R}$ of the regression function $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$. Here $r_n(\mathbf{x}) = r_n(\mathbf{x}, \mathcal{D}_n)$ is a measurable function of $\mathbf{x}$ and the data. For simplicity, we will omit $\mathcal{D}_n$ in the notation and write $r_n(\mathbf{x})$ instead of $r_n(\mathbf{x}, \mathcal{D}_n)$.

As in section 2, for fixed $\mathbf{x} \in \mathbb{R}^d$, we denote by $\mathcal{L}_n(\mathbf{x})$ the LNN of $\mathbf{x}$ in the sample $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and let $L_n(\mathbf{x})$ be the cardinality of $\mathcal{L}_n(\mathbf{x})$ (i.e., $L_n(\mathbf{x}) = |\mathcal{L}_n(\mathbf{x})|$ — note that $L_n(\mathbf{x}) \geq 1$). As stated in the introduction, we will be concerned in this section with the consistency properties of the LNN regression estimate, which is defined by

$$r_n(\mathbf{x}) = \frac{1}{L_n(\mathbf{x})} \sum_{i=1}^n Y_i \mathbf{1}_{[\mathbf{X}_i \in \mathcal{L}_n(\mathbf{x})]}.$$

Our main result is the following theorem:

Theorem 3.1 (Pointwise L_p-consistency) *Assume that the regression function r is λ-almost everywhere continuous and that Y is bounded. Then, for μ-almost all $\mathbf{x} \in \mathbb{R}^d$ and all $p \geq 1$,*

$$\mathbb{E}\,|r_n(\mathbf{x}) - r(\mathbf{x})|^p \to 0 \quad \textit{as } n \to \infty.$$

The following corollary is a consequence of Theorem 3.1 and the dominated convergence theorem.

Theorem 3.2 (Gobal L_p-consistency) *Assume that the regression function r is λ-almost everywhere continuous and that Y is bounded. Then, for all $p \geq 1$,*

$$\mathbb{E}\,|r_n(\mathbf{X}) - r(\mathbf{X})|^p \to 0 \quad \textit{as } n \to \infty.$$

The theorems above are not universal — indeed, we assume that r is λ-almost everywhere continuous and that $\mathbf{X}$ has a density. It is noteworthy that no universal consistency result is possible. To see this, let $\mathbf{X}$ be $\mathbb{R}^2$-valued uniformly distributed on the diagonal $D = \{\mathbf{x} = (x_1, x_2) : 0 \leq x_1 \leq 1, x_2 = x_1\}$. Then the LNN regression estimate just takes an average over two observations. If Y is independent of $\mathbf{X}$ and uniform on $[-1, 1]$, it is easy to see that $r \equiv 0$, yet r_n has a constant nonzero variance and does not converge as $n \to \infty$ to 0 in probability. Equivalently, one could verify Stone's necessary and sufficient conditions (Stone [21]) for universal consistency of regression estimates.

On the positive side, the results do not impose any condition on the density. They are also scale-free, i.e., the estimate does not change when all coordinates of $\mathbf{X}$ are transformed in a strictly monotone manner. In particular, one can without loss of generality assume that $\mathbf{X}$ is supported on $[0, 1]^d$.

The elementary result needed to prove Theorem 3.1 is:

Lemma 3.1 *If r is λ-almost everywhere continuous and Y is bounded, then, for fixed $p \geq 1$,*

$$\mathbb{E}\left[\frac{1}{L_n(\mathbf{x})}\sum_{i=1}^{n} \mathbf{1}_{[\mathbf{X}_i \in \mathcal{L}_n(\mathbf{x})]}\,|r(\mathbf{X}_i) - r(\mathbf{x})|^p\right] \quad \to 0 \quad \textit{as } n \to \infty,$$

at μ-almost all $\mathbf{x} \in \mathbb{R}^d$.

Proof of Lemma 3.1 Recall that $\mathcal{R}_\varepsilon(\mathbf{x}) = [x_1, x_1+\varepsilon] \times \ldots \times [x_d, x_d+\varepsilon]$. We can define $\mathcal{R}_\varepsilon(\mathbf{x}, \ell)$, $\ell = 1, \ldots, 2^d$, as $\mathcal{R}_\varepsilon(\mathbf{x})$ for the 2^d quadrants centered at $\mathbf{x}$. We then have $\mathcal{L}_n(\mathbf{x}, \ell)$ and $L_n(\mathbf{x}, \ell) = |\mathcal{L}_n(\mathbf{x}, \ell)|$. Also, on the ℓ-th quadrant, we have the sums

$$S_n(\mathbf{x}, \ell) = \sum_{i=1}^{n} \mathbf{1}_{[\mathbf{X}_i \in \mathcal{L}_n(\mathbf{x},\ell)]} \left| r(\mathbf{X}_i) - r(\mathbf{x}) \right|^p .$$

If

$$\mathbb{E}\left[\frac{S_n(\mathbf{x}, \ell)}{L_n(\mathbf{x}, \ell)}\right] \to 0 \quad \text{as } n \to \infty \text{ for all } \ell,$$

(with the convention $0 \times \infty = 0$ when $L_n(\mathbf{x}, \ell) = 0$), then

$$\mathbb{E}\left[\frac{\sum_{\ell=1}^{2^d} S_n(\mathbf{x}, \ell)}{\sum_{\ell=1}^{2^d} L_n(\mathbf{x}, \ell)}\right] \to 0 \quad \text{as } n \to \infty,$$

so that

$$\mathbb{E}\left[\frac{1}{L_n(\mathbf{x})} \sum_{i=1}^{n} \mathbf{1}_{[\mathbf{X}_i \in \mathcal{L}_n(\mathbf{x})]} \left| r(\mathbf{X}_i) - r(\mathbf{x}) \right|^p \right] \to 0 \quad \text{as } n \to \infty.$$

This follows from the fact that

$$\mathbb{E}\left[\frac{A_1 + \ldots + A_k}{B_1 + \ldots + B_k}\right] \to 0$$

if $\mathbb{E}[A_i/B_i] \to 0$ for all i, where the random variables A_i and B_i are nonnegative and satisfy $A_i \leq cB_i$ for some nonnegative c (again, we use the convention $0 \times \infty = 0$). So, we need only concentrate on the first quadrant.

For arbitrary $\varepsilon > 0$, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{L_n(\mathbf{x})} \sum_{i=1}^{n} \mathbf{1}_{[\mathbf{X}_i \in \mathcal{L}_n(\mathbf{x})]} \left| r(\mathbf{X}_i) - r(\mathbf{x}) \right|^p \right] \\
&\quad = \mathbb{E}\left[\frac{1}{L_n(\mathbf{x})} \sum_{i=1}^{n} \mathbf{1}_{[\mathbf{X}_i \in \mathcal{L}_n(\mathbf{x})]} \left| r(\mathbf{X}_i) - r(\mathbf{x}) \right|^p \mathbf{1}_{[\mathbf{X}_i \in \mathcal{R}_\varepsilon^c(\mathbf{x})]} \right] \\
&\qquad + \mathbb{E}\left[\frac{1}{L_n(\mathbf{x})} \sum_{i=1}^{n} \mathbf{1}_{[\mathbf{X}_i \in \mathcal{L}_n(\mathbf{x})]} \left| r(\mathbf{X}_i) - r(\mathbf{x}) \right|^p \mathbf{1}_{[\mathbf{X}_i \in \mathcal{R}_\varepsilon(\mathbf{x})]} \right]
\end{aligned}$$

$$\leq 2^p \gamma^p \mathbb{E}\left[\frac{1}{L_n(\mathbf{x})}\sum_{i=1}^n \mathbf{1}_{[\mathbf{X}_i\in\mathcal{L}_n(\mathbf{x})]}\mathbf{1}_{[\mathbf{X}_i\in\mathcal{R}_\varepsilon^c(\mathbf{x})]}\right]$$
$$+\left[\sup_{\mathbf{z}\in\mathbb{R}^d:\|\mathbf{z}-\mathbf{x}\|_\infty\leq\varepsilon}|r(\mathbf{z})-r(\mathbf{x})|\right]^p$$
$$(\text{since } |Y|\leq\gamma).$$

The rightmost term of the latter inequality tends to 0 as $\varepsilon \to 0$ at points $\mathbf{x}$ at which r is continuous. Thus, the lemma will be proven if we show that, for fixed $\varepsilon > 0$,

$$\mathbb{E}\left[\frac{1}{L_n(\mathbf{x})}\sum_{i=1}^n \mathbf{1}_{[\mathbf{X}_i\in\mathcal{L}_n(\mathbf{x})]}\mathbf{1}_{[\mathbf{X}_i\in\mathcal{R}_\varepsilon^c(\mathbf{x})]}\right] \to 0 \quad \text{as } n\to\infty.$$

To this aim, denote by N_n the (random) number of sample points falling in $\mathcal{R}_\varepsilon(\mathbf{x})$. For $N_n \geq 1$ and each $r = 1, \ldots, d$, let $\mathbf{X}_n^{\star(r)} = (X_{n,1}^{\star(r)}, \ldots, X_{n,d}^{\star(r)})$ be the observation in $\mathcal{R}_\varepsilon(\mathbf{x})$ whose r-coordinate is the closest to x_r (note that $\mathbf{X}_n^{\star(r)}$ is almost surely unique), and consider the set

$$\begin{aligned}\mathcal{P}_\varepsilon^{(r)} = [x_1+\varepsilon, +\infty[&\times \ldots \times [x_{r-1}+\varepsilon,+\infty[\\ &\times [x_r, X_{n,r}^{\star(r)}]\\ &\times [x_{r+1}+\varepsilon,+\infty[\times\ldots\times[x_d+\varepsilon,+\infty[\end{aligned}$$

(see Figure 2 for an illustration in dimension 2).

Take finally

$$\mathcal{P}_\varepsilon = \bigcup_{r=1}^d \mathcal{P}_\varepsilon^{(r)},$$

and define the random variable

$$Q_{n,\varepsilon} = \begin{cases} +\infty & \text{if } N_n = 0\\ \text{the number of sample points falling in } \mathcal{P}_\varepsilon & \text{if } N_n \geq 1.\end{cases}$$

It is shown in Lemma 5.5 that, for μ-almost all $\mathbf{x}$,

$$Q_{n,\varepsilon} = \mathcal{O}_{\mathbb{P}}(1),$$

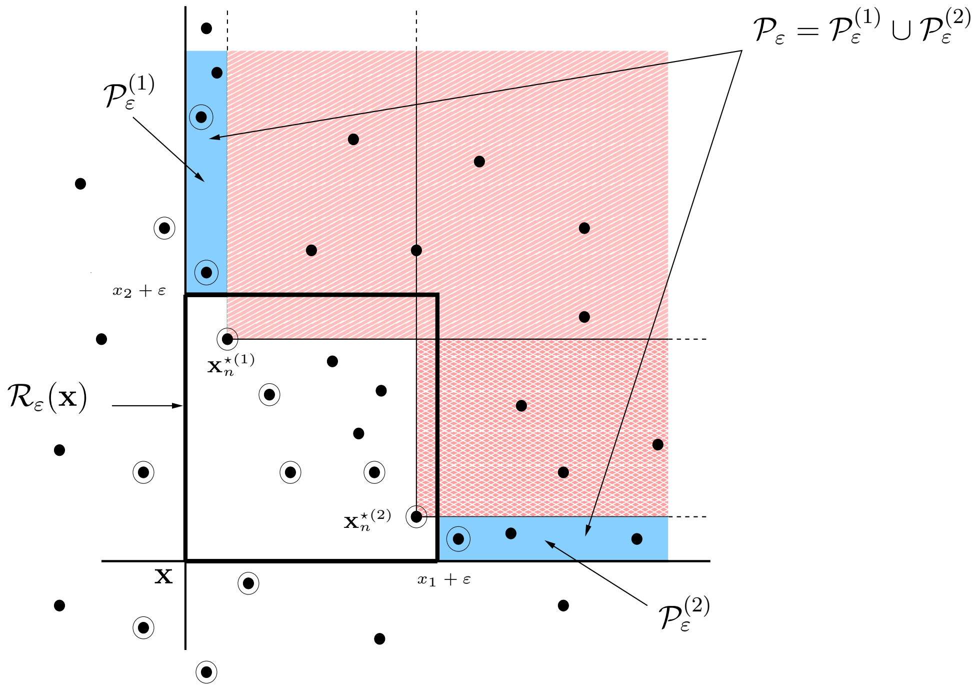
$\mathcal{P}_\varepsilon = \mathcal{P}_\varepsilon^{(1)} \cup \mathcal{P}_\varepsilon^{(2)}$
$\mathcal{P}_\varepsilon^{(1)}$
$x_2 + \varepsilon$
$\mathbf{x}_n^{\star(1)}$
$\mathcal{R}_\varepsilon(\mathbf{x})$
$\mathbf{x}_n^{\star(2)}$
$\mathbf{x}$
$x_1 + \varepsilon$
$\mathcal{P}_\varepsilon^{(2)}$

Figure 2: Notation in dimension $d = 2$. Here $N_n = 8$ and $Q_{n,\varepsilon} = 7$. Note that none of the points in the framed area can be a LNN of $\mathbf{x}$.

i.e., for any $\alpha > 0$, there exists $A > 0$ such that, for all n large enough,

$$\mathbb{P}(Q_{n,\varepsilon} \geq A) \leq \alpha. \tag{10}$$

Now, by definition of the LNN, on the event $[N_n \geq 1]$, we have

$$\mathbf{1}_{[\mathbf{X}_i \in \mathcal{L}_n(\mathbf{x})]} \mathbf{1}_{[\mathbf{X}_i \in \mathcal{R}^c_\varepsilon(\mathbf{x})]} \leq \mathbf{1}_{[\mathbf{X}_i \in \mathcal{P}_\varepsilon]},$$

and consequently,

$$\sum_{i=1}^n \mathbf{1}_{[\mathbf{X}_i \in \mathcal{L}_n(\mathbf{x})]} \mathbf{1}_{[\mathbf{X}_i \in \mathcal{R}^c_\varepsilon(\mathbf{x})]} \leq Q_{n,\varepsilon}. \tag{11}$$

Thus, for any $\alpha > 0$ and all n large enough, by (10) and (11),

$$\mathbb{E}\left[\frac{1}{L_n(\mathbf{x})} \sum_{i=1}^n \mathbf{1}_{[\mathbf{X}_i \in \mathcal{L}_n(\mathbf{x})]} \mathbf{1}_{[\mathbf{X}_i \in \mathcal{R}^c_\varepsilon(\mathbf{x})]}\right]$$

$$\begin{aligned}
&\leq \mathbb{E}\left[\frac{Q_{n,\varepsilon}}{L_n(\mathbf{x})}\,\mathbf{1}_{[Q_{n,\varepsilon}<A]}\right] + \mathbb{E}\mathbf{1}_{[Q_{n,\varepsilon}\geq A]}\\
&= \mathbb{E}\left[\frac{Q_{n,\varepsilon}}{L_n(\mathbf{x})}\,\mathbf{1}_{[Q_{n,\varepsilon}<A, L_n(\mathbf{x})\geq 1]}\right] + \mathbb{P}(Q_{n,\varepsilon}\geq A)\\
&\leq \mathbb{E}\left[\frac{A}{L_n(\mathbf{x})}\,\mathbf{1}_{[L_n(\mathbf{x})\geq 1]}\right] + \alpha.
\end{aligned}$$

By Theorem 2.1,

$$L_n(\mathbf{x}) \to \infty \quad \text{in probability as } n \to \infty,$$

at μ-almost all $\mathbf{x}$. This implies

$$\mathbb{E}\left[\frac{1}{L_n(\mathbf{x})}\,\mathbf{1}_{[L_n(\mathbf{x})\geq 1]}\right] \to 0 \text{ as } n \to \infty,$$

which concludes the proof of Lemma 3.1. ■

Proof of Theorem 3.1 Because $|a+b|^p \leq 2^{p-1}\left(|a|^p + |b|^p\right)$ for $p \geq 1$, we see that

$$\begin{aligned}
&\mathbb{E}\left|r_n(\mathbf{x}) - r(\mathbf{x})\right|^p\\
&\quad\leq 2^{p-1}\mathbb{E}\left|\frac{1}{L_n(\mathbf{x})}\sum_{i=1}^{n}\mathbf{1}_{[\mathbf{X}_i\in\mathcal{L}_n(\mathbf{x})]}\left(Y_i - r(\mathbf{X}_i)\right)\right|^p\\
&\qquad + 2^{p-1}\mathbb{E}\left|\frac{1}{L_n(\mathbf{x})}\sum_{i=1}^{n}\mathbf{1}_{[\mathbf{X}_i\in\mathcal{L}_n(\mathbf{x})]}\left(r(\mathbf{X}_i) - r(\mathbf{x})\right)\right|^p.
\end{aligned}$$

Thus, by Jensen's inequality,

$$\begin{aligned}
&\mathbb{E}\left|r_n(\mathbf{x}) - r(\mathbf{x})\right|^p\\
&\quad\leq 2^{p-1}\mathbb{E}\left|\frac{1}{L_n(\mathbf{x})}\sum_{i=1}^{n}\mathbf{1}_{[\mathbf{X}_i\in\mathcal{L}_n(\mathbf{x})]}\left(Y_i - r(\mathbf{X}_i)\right)\right|^p\\
&\qquad + 2^{p-1}\mathbb{E}\left[\frac{1}{L_n(\mathbf{x})}\sum_{i=1}^{n}\mathbf{1}_{[\mathbf{X}_i\in\mathcal{L}_n(\mathbf{x})]}\left|r(\mathbf{X}_i) - r(\mathbf{x})\right|^p\right]. \qquad (12)
\end{aligned}$$

The rightmost term in (12) tends to 0 for μ-almost all $\mathbf{x}$ by Lemma 3.1. Thus, it remains to show that the first term tends to 0 at μ-almost all $\mathbf{x}$. By successive applications of inequalities of Marcinkiewicz and Zygmund [18] (see also Petrov [19, pages 59-60]), we have for some positive constant C_p depending only on p,

$$
\begin{aligned}
&\mathbb{E}\left|\frac{1}{L_n(\mathbf{x})}\sum_{i=1}^{n}\mathbf{1}_{[\mathbf{X}_i\in\mathcal{L}_n(\mathbf{x})]}\left(Y_i - r(\mathbf{X}_i)\right)\right|^p \\
&\le C_p\,\mathbb{E}\left[\frac{1}{L_n^2(\mathbf{x})}\sum_{i=1}^{n}\mathbf{1}_{[\mathbf{X}_i\in\mathcal{L}_n(\mathbf{x})]}\left(Y_i - r(\mathbf{X}_i)\right)^2\right]^{p/2} \\
&\le (2\gamma)^p C_p\,\mathbb{E}\left[\frac{1}{L_n^2(\mathbf{x})}\sum_{i=1}^{n}\mathbf{1}_{[\mathbf{X}_i\in\mathcal{L}_n(\mathbf{x})]}\right]^{p/2} \\
&\quad (\text{since } |Y|\le\gamma) \\
&= (2\gamma)^p C_p\,\mathbb{E}\left[\frac{1}{L_n(\mathbf{x})}\sum_{i=1}^{n}\frac{1}{L_n(\mathbf{x})}\mathbf{1}_{[\mathbf{X}_i\in\mathcal{L}_n(\mathbf{x})]}\right]^{p/2} \\
&= (2\gamma)^p C_p\,\mathbb{E}\left[\frac{1}{L_n^{p/2}(\mathbf{x})}\right].
\end{aligned}
$$

By Theorem 2.1,

$$L_n(\mathbf{x}) \to \infty \quad \text{in probability as } n\to\infty,$$

at μ-almost all $\mathbf{x}$. Since $L_n(\mathbf{x}) \ge 1$, this implies

$$\mathbb{E}\left[\frac{1}{L_n^{p/2}(\mathbf{x})}\right] \to 0 \text{ as } n\to\infty,$$

and the proof is complete. ■

Thus, in particular, since (1) is equivalent to taking a majority vote over LNN, we have Bayes risk consistency whenever r is λ-almost everywhere continuous and $\mathbf{X}$ has a density. This partially solves an exercise in [12].

In view of Theorem 2.2, averaging in the LNN is never over more than $\mathcal{O}((\log n)^{d-1})$ elements. One cannot expect a great rate of convergence for these estimates. The same is true, mutatis mutandis, for Breiman's random forests because averaging is over a subset of size $\mathcal{O}((\log n)^{d-1})$. However, one can hope to improve the averaging rate by the judicious use of subsampling in bagging (bootstrap-aggregation). Bagging, which was suggested by Breiman in [5], is a simple way of randomizing and averaging predictors in order to improve their performance. In bagging, randomization is achieved by generating many bootstrap samples from the original data set. This is illustrated in the next section on 1-nearest neighbour bagging.

3.2 Random forests and LNN

As stated in the introduction, a random forest is a tree-ensemble learning algorithm, where each tree depends on the values of a random vector sampled independently and with the same distribution for all trees. Thus, a random forest consists of many decision trees and outputs the average of the decisions provided by individual trees. Random forests have been shown to give excellent performance on a number of practical problems. They work fast, generally exhibit a substantial performance improvement over single tree algorithms such as CART, and yield generalization error rates that compare favorably to traditional statistical methods. In fact, random forests are among the most accurate general-purpose learning algorithms available (Breiman [7]).

Algorithms for inducing a random forest were first developed by Breiman and Cutler, and "Random Forests" is their trademark. The web page

http://www.stat.berkeley.edu/users/breiman/RandomForests

provides a collection of downloadable technical reports, and gives an overview of random forests as well as comments on the features of the method.

Following Biau et al. [8], who study consistency of various versions of random forests and other randomized ensemble classifiers, a regression forest may be modelled as follows. Assume that $\Theta_1, \ldots, \Theta_m$ are i.i.d. draws of some randomizing variable Θ, independent of the sample. Then, a random forest is a collection of m randomized regression trees $t_1(\mathbf{x}, \Theta_1, \mathcal{D}_n), \ldots, t_m(\mathbf{x}, \Theta_m, \mathcal{D}_n)$, which are finally combined to form the aggregated regression estimate

$$r_n(\mathbf{x}) = \frac{1}{m} \sum_{j=1}^{m} t_j(\mathbf{x}, \Theta_j, \mathcal{D}_n).$$

The randomizing variable Θ is used to determine how the successive cuts are performed when building the tree, such as selection the coordinate to split and position of the split. In the model we have in mind, each individual randomized tree $t_j(\mathbf{x}, \Theta_j, \mathcal{D}_n)$ is typically constructed without pruning, that is, the tree building process continues until each terminal node contains no more than k data points, where k is some prespecified positive integer. Different random forests differ in how randomness is introduced in the tree building process, ranging from extreme random splitting strategies (Breiman [6], Cutler and Zhao [10]) to more involved data-dependent strategies (Amit and Geman [1], Breiman [7], Dietterich [13]). However, as pointed out by Lin and Jeon [17], no matter what splitting strategy is used, if the nodes of the individual trees define rectangular cells, then a random forest with $k = 1$ can be viewed as a *weighted* LNN regression estimate. Besides, if the randomized splitting scheme is independent of the responses $Y_1, \ldots, Y_n$ — such a scheme is called *non-adaptive* in [17] — then so are the weights. One example of such a scheme is the purely random splitting where, for each internal node, we randomly choose a variable to split on, and the split point is chosen uniformly at random over all possible split points on that variable. Thus, for such non-adaptive strategies,

$$r_n(\mathbf{x}) = \sum_{i=1}^{n} Y_i W_{ni}(\mathbf{x}),$$

where the weights $(W_{n1}(\mathbf{x}), \ldots, W_{nn}(\mathbf{x}))$ are nonnegative Borel measurable functions of $\mathbf{x}$, $\mathbf{X}_1, \ldots, \mathbf{X}_n, \Theta_1, \ldots, \Theta_m$, and such that $W_{ni}(\mathbf{x}) = 0$ if $\mathbf{X}_i \notin \mathcal{L}_n(\mathbf{x})$ and

$$\sum_{i=1}^{n} W_{ni}(\mathbf{x}) = \sum_{i=1}^{n} W_{ni}(\mathbf{x}) \mathbf{1}_{[\mathbf{X}_i \in \mathcal{L}_n(\mathbf{x})]} = 1.$$

The next proposition states a lower bound on the rate of convergence of the mean squared error of a random forest with non-adaptive splitting scheme. In this proposition, the symbol $\mathbb{V}$ denotes variance and $\mathbb{E}$ denotes expectation with respect to $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and $\Theta_1, \ldots, \Theta_m$.

Proposition 3.1 *For any $\mathbf{x} \in \mathbb{R}^d$, assume that $\sigma^2 = \mathbb{V}[Y|\mathbf{X} = \mathbf{x}]$ is independent of $\mathbf{x}$. Then*

$$\mathbb{E}\left[r_n(\mathbf{x}) - r(\mathbf{x})\right]^2 \geq \frac{\sigma^2}{\mathbb{E}L_n(\mathbf{x})}.$$

Proof of Proposition 3.1 We may write, using the independence of $\mathcal{D}_n$ and $\Theta_1, \ldots, \Theta_m$,

$$\begin{aligned}
\mathbb{E}\left[r_n(\mathbf{x}) - r(\mathbf{x})\right]^2 &\geq \mathbb{E}\left[\mathbb{V}[r_n(\mathbf{x})|\mathbf{X}_1, \ldots, \mathbf{X}_n, \Theta_1, \ldots, \Theta_m]\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} W_{ni}^2(\mathbf{x}) \mathbb{V}[Y_i|\mathbf{X}_1, \ldots, \mathbf{X}_n, \Theta_1, \ldots, \Theta_m]\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} W_{ni}^2(\mathbf{x}) \mathbb{V}[Y_i|\mathbf{X}_i]\right] \\
&= \sigma^2\, \mathbb{E}\left[\sum_{i=1}^{n} W_{ni}^2(\mathbf{x})\right] \\
&\geq \sigma^2\, \mathbb{E}\left[\frac{1}{L_n(\mathbf{x})}\left(\sum_{i=1}^{n} W_{ni}(\mathbf{x})\right)^2\right] \\
&\quad \text{(by the Cauchy-Schwarz inequality)} \\
&= \sigma^2\, \mathbb{E}\left[\frac{1}{L_n(\mathbf{x})}\right],
\end{aligned}$$

where, in the last equality, we used the fact that $\sum_{i=1}^{n} W_{ni}(\mathbf{x}) = 1$. The conclusion follows from Jensen's inequality. ■

Proposition 3.1 is thrown in here because we know that $\mathbb{E}L_n(\mathbf{x}) \sim 2^d(\log n)^{d-1}/(d-1)!$ at μ-almost all $\mathbf{x}$, when f is λ-almost everywhere continuous (Theorem 2.2). Thus, under this additional condition on f, at an $\mathbf{x}$ for which Theorem 2.2 is valid, we have

$$\begin{aligned}
\mathbb{E}\left[r_n(\mathbf{x}) - r(\mathbf{x})\right]^2 &\geq \frac{\sigma^2}{\mathbb{E}L_n(\mathbf{x})} \\
&\sim \frac{\sigma^2(d-1)!}{2^d(\log n)^{d-1}},
\end{aligned}$$

which is rather slow as a function of n.

As mentioned above, there are two related methods to possibly get a better rate of convergence:

(i) One can modify the splitting method and stop as soon as a future rectangle split would cause a sub-rectangle to have fewer than k points. In this manner, if $k \to \infty$, $k/n \to 0$, one can obtain consistent regression function estimates and classifiers with variances of errors that are of the order $1/[k(\log n)^{d-1}]$. In a sense, this generalizes the classical k-nearest neighbour (k-NN) approach (Györfi et al. [16, Chapter 6]).

(ii) One could resort to bagging and randomize using small random subsamples. In the next section, we illustrate how this can be done for the 1-NN rule of Fix and Hodges [15] (see also Cover and Hart [9]), thereby extending previous results of [8]. A random subsample of size k is drawn, and the method is repeated m times. The regression estimate takes the average over the m Y-values corresponding to the nearest neighbours. In classification, a majority vote is taken. It is shown that for appropriate k and m, this 1-NN bagging is universally consistent, and indeed, that it corresponds to a weighted 1-NN rule, roughly speaking, with geometrically decreasing weights (fore more on weighted NN rules, see Stone [21], Devroye [11] or Györfi et al. [16]). Because of this equivalence, one can optimize using standard bias/variance trade-off methods, such as used, e.g., in [16].

4 The bagged 1-NN rule

Breiman's bagging principle has a simple application in the context of nearest neighbour methods. We proceed as follows, via a randomized basic regression estimate $r_{n,k}$ in which $1 \le k \le n$ is a parameter. The predictor $r_{n,k}$ is the 1-NN rule for a random sample $\mathcal{S}_n$ drawn with (without) replacement from $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$, with $|\mathcal{S}_n| = k$. Clearly, $r_{n,k}$ is not generally universally consistent.

We apply bagging, that is, we repeat the random sampling m times, and take the average of the individual outcomes. Formally, if $Z_j = r_{n,k}(\mathbf{x})$ is the prediction in the j-th round of bagging, we let the bagged regression estimate $r^\star$ be defined as

$$r^\star(\mathbf{x}) = \frac{1}{m}\sum_{j=1}^{m} Z_j,$$

where $Z_1, \ldots, Z_m$ are the outcomes in the individual rounds. In the context of classification, $Z_j \in \{0, 1\}$, and we classify $\mathbf{x}$ as being in class 1 if $r^\star(\mathbf{x}) \ge 1/2$, that is

$$\sum_{j=1}^{m} \mathbf{1}_{[Z_j=1]} \geq \sum_{j=1}^{m} \mathbf{1}_{[Z_j=0]}.$$

The corresponding bagged classifier is denoted by $g_n^\star$.

Theorem 4.1 *If $m \to \infty$ (or $m = \infty$), $k \to \infty$ and $k/n \to 0$, then $r_n^\star$ is universally L_p-consistent for all $p \geq 1$.*

Corollary 4.1 *If $m \to \infty$ (or $m = \infty$), $k \to \infty$ and $k/n \to 0$, then $g_n^\star$ is universally Bayes risk consistent.*

Remark In the theorem, the fact that sampling was done with/without replacement is irrelevant.

Before proving Theorem 4.1, recall that if we let $V_{n1} \geq V_{n2} \geq \ldots \geq V_{nn} \geq 0$ denote weights that sum to one, and $V_{n1} \to 0$, $\sum_{i>\varepsilon n} V_{ni} \to 0$ for all $\varepsilon > 0$ as $n \to \infty$, then the regression estimate

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

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 neighbour regression estimate.* It is universally L_p-consistent for all $p \geq 1$ (Stone [21], and Problems 11.7, 11.8 of Devroye et al. [12]). In the sequel, to shorten notation, we omit the index n in the weights and write, for instance, V_1 instead of V_{n1}.

Proof of Theorem 4.1 We first observe that if $m = \infty$, $r_n^\star$ is in fact a weighted nearest neighbour estimate with

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

To avoid trouble, we have a unique way of breaking distance ties, that is, any tie is broken by using indices to declare a winner. Then, a moment's thought shows that for the “without replacement” sampling, V_i is hypergeometric:

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

We have

$$\begin{aligned} V_i &= \frac{k}{n-k+1} \cdot \frac{n-i}{n} \cdot \frac{n-i-1}{n-1} \cdots \frac{n-i-k+2}{n-k+2} \\ &= \frac{k}{n-k+1} \prod_{j=0}^{k-2} \left(1 - \frac{i}{n-j}\right) \\ &\in \left[\frac{k}{n-k+1} \exp\left(\frac{-i(k-1)}{n-k-i+2}\right), \frac{k}{n-k+1} \exp\left(\frac{-i(k-1)}{n}\right)\right], \end{aligned}$$

where we used $\exp(-u/(1-u)) \leq 1-u \leq \exp(-u)$, $0 \leq u < 1$. Clearly, V_i is nonincreasing, with

$$V_1 \leq \frac{k}{n-k} \to 0.$$

Also,

$$\begin{aligned} \sum_{i>\varepsilon n} V_i &\leq \frac{k}{n-k+1} \sum_{i>\varepsilon n} e^{-i(k-1)/n} \\ &\leq \frac{k}{n-k+1} \cdot \frac{e^{-\varepsilon(k-1)}}{(1-e^{-(k-1)/n})} \\ &\sim e^{-\varepsilon(k-1)} \to 0 \quad \text{as } k \to \infty. \end{aligned}$$

For sampling with replacement,

$$\begin{aligned} V_i &= \left(1 - \frac{i-1}{n}\right)^k - \left(1 - \frac{i}{n}\right)^k \\ &= \left(1 - \frac{i-1}{n}\right)^k \left[1 - \left(1 - \frac{1}{n-i+1}\right)^k\right] \\ &\in \left[e^{-(i-1)k/(n-i+1)} \left[\frac{k}{n-i+1} - \frac{k(k-1)}{2} \left(\frac{1}{n-i+1}\right)^2\right],\right. \end{aligned}$$

$$e^{-(i-1)k/n} \cdot \frac{k}{n-i+1}\Bigg],$$

where we used $1-\alpha u \le (1-u)^\alpha \le 1-\alpha u+\alpha(\alpha-1)u^2/2$ for integer $\alpha \ge 1$, $0 \le u \le 1$. Again, V_i is nonincreasing, and

$$V_1 = 1 - \left(1-\frac{1}{n}\right)^k \le \frac{k}{n} \to 0.$$

Also

$$\sum_{i>\varepsilon n} V_i = \left(1-\frac{\lfloor \varepsilon n\rfloor}{n}\right)^k \to 0$$

since $\varepsilon > 0$ is fixed and $k \to \infty$.

Remark For $\varepsilon > 1$, note that uniformly over $1 \le i \le \varepsilon n$,

$$\sup_{1\le i\le \varepsilon n} \left| \frac{V_i}{e^{-ik/n} \,.\, k/n} - 1 \right| \to 0,$$

so the weights behave as $\rho \exp(-\rho i)$, $\rho = k/n$.

For $m < \infty$, $m \to \infty$, the weights of the neighbours are random variables $(W_1,\ldots,W_n)$, with $\sum_{i=1}^n W_i = 1$, and, in fact,

$$(W_1,\ldots,W_n) \overset{\mathcal{L}}{=} \frac{\text{Multinomial } (m\,;\,V_1,\ldots,V_n)}{m}.$$

We note that this random vector is *independent* of the data!

In the proof of the consistency result below, we use Stone's [21] general consistency theorem for locally weighted average estimates, see also [12, Theorem 6.3]. According to Stone's theorem, consistency holds if the following three conditions are satisfied:

(i)

$$\mathbb{E}\left[\max_{i=1,\ldots,n} W_i\right] \to 0 \quad \text{as } n \to \infty.$$

(*ii*) For all $\varepsilon > 0$,

$$\mathbb{E}\left[\sum_{i>\varepsilon n} W_i\right] \to 0 \quad \text{as } n \to \infty.$$

(*iii*) There is a constant C such that, for every nonnegative measurable function f satisfying $\mathbb{E}f(\mathbf{X}) < \infty$,

$$\mathbb{E}\left[\sum_{i=1}^{n} W_i f(\mathbf{X}_i)\right] \leq C\,\mathbb{E}f(\mathbf{X}).$$

Checking Stone's conditions of convergence requires only minor work. To show (*i*), note that

$$\begin{aligned}
&\mathbb{P}\left(\max_{i=1,\ldots,n} W_i \geq \varepsilon\right)\\
&\quad \leq \sum_{i=1}^{n} \mathbb{P}(W_i \geq \varepsilon)\\
&\quad = \sum_{i=1}^{n} \mathbb{P}\left(\text{Bin }(m, V_i) \geq m\varepsilon\right)\\
&\quad = \sum_{i=1}^{n} \mathbb{P}\left(\text{Bin }(m, V_i) \geq mV_i + m(\varepsilon - V_i)\right)\\
&\quad \leq \sum_{i=1}^{n} \frac{\mathbb{V}\left[\text{Bin }(m, V_i)\right]}{\left(m(\varepsilon - V_i)\right)^2}\\
&\qquad \text{(by Chebyshev's inequality, for all } n \text{ large enough),}\\
&\quad \leq \frac{\sum_{i=1}^{n} mV_i}{m^2(\varepsilon - V_1)^2} = \frac{1}{m(\varepsilon - V_1)^2} \to 0.
\end{aligned}$$

Secondly, for (*ii*), we set $p = \sum_{i>\varepsilon n} V_i$, and need only show that $\mathbb{E}[\text{Bin }(m,p)/m] \to 0$. But this follows from $p \to 0$. Condition (*iii*) reduces to

$$\mathbb{E}\left[\sum_{i=1}^{n} V_i f(\mathbf{X}_i)\right],$$

which we know is bounded by a constant times $\mathbb{E}f(\mathbf{X})$ for any sequence of nonincreasing nonnegative weights V_i that sum to one (Stone [21], and [12, Chapter 11, Problems 11.7 and 11.8].

This concludes the proof. ■

5 Some technical lemmas

Throughout this section, for $\mathbf{x} = (x_1, \ldots, x_d)$ and $\varepsilon > 0$, $\mathcal{R}_\varepsilon(\mathbf{x})$ refers to the hyperrectangle

$$\mathcal{R}_\varepsilon(\mathbf{x}) = [x_1, x_1 + \varepsilon] \times \ldots \times [x_d, x_d + \varepsilon].$$

Lemma 5.1 *Let $\Phi : (0, \infty) \to [0, \infty)$ be a nondecreasing function with limit 0 at 0. Then there exists a sequence (ε_n) of positive real numbers such that $n\varepsilon_n^d \to \infty$ and $n\varepsilon_n^d \Phi(\varepsilon_n) \to 0$ as $n \to \infty$.*

Proof of Lemma 5.1 Note first that if such a sequence (ε_n) exists, then $\varepsilon_n \to 0$ as $n \to \infty$. Indeed, if this is not the case, then $\varepsilon_n \geq C$ for some positive C and infinitely many n. Consequently, using the fact that Φ is nondecreasing, one obtains $n\varepsilon_n^d \Phi(\varepsilon_n) \geq n\varepsilon_n^d \Phi(C)$ for infinitely many n, and this is impossible.

For any integer $\ell \geq 1$, set $e_\ell = \Phi(1/\ell)$ and observe that the sequence (e_ℓ) is nonincreasing and tends to 0 as $\ell \to \infty$. Let $\varphi_\ell = \ell^d / \sqrt{e_\ell}$. Clearly, the sequence (φ_ℓ) is nondecreasing and satisfies $\varphi_\ell / \ell^d \to \infty$ and $[\varphi_\ell / \ell^d] \times \Phi(1/\ell) = \sqrt{e_\ell} \to 0$ as $\ell \to \infty$.

For each $n \geq 1$, let ℓ_n be the largest positive integer ℓ such that $\varphi_\ell \leq n$, and let $\varepsilon_n = 1/\ell_n$. Then the sequence (ε_n) satisfies

$$n\varepsilon_n^d \geq \varphi_{\ell_n} / \ell_n^d \to \infty$$

and

$$n\varepsilon_n^d \Phi(\varepsilon_n) \geq [\varphi_{\ell_n} / \ell_n^d] \times \Phi(1/\ell_n) \to 0$$

as $n \to \infty$. ■

Lemma 5.2 *Suppose that μ has a probability density f. For $\mathbf{x} \in \mathbb{R}^d$, let g_ε be the probability density defined by*

$$g_\varepsilon(\mathbf{y}) = \begin{cases} \dfrac{\mu\left(\mathcal{R}_\varepsilon(\mathbf{x})\right)}{\varepsilon^d} & \textit{if } \mathbf{y} \in \mathcal{R}_\varepsilon(\mathbf{x}) \\ f(\mathbf{y}) & \textit{otherwise,} \end{cases}$$

and let $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ *be independent random vectors distributed according to* g_ε*. Let* (ε_n) *be a sequence of positive real numbers such that* $\varepsilon_n \to 0$ *and* $n\varepsilon_n^d \to \infty$ *as* $n \to \infty$*. Then, denoting by* $L'_{\varepsilon_n}(\mathbf{x})$ *the number of LNN of* $\mathbf{x}$ *in the sample* $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$ *falling in* $\mathcal{R}_{\varepsilon_n}(\mathbf{x})$*, one has*

$$L'_{\varepsilon_n}(\mathbf{x}) \to \infty \quad \textit{in probability as } n \to \infty,$$

at μ*-almost all* $\mathbf{x}$*.*

Proof of Lemma 5.2 To ligthen notation a bit, we set $p_\varepsilon(\mathbf{x}) = \mu(\mathcal{R}_\varepsilon(\mathbf{x}))$. Choose $\mathbf{x}$ in a set of μ-measure 1 such that $\mu(\mathcal{R}_\varepsilon(\mathbf{x})) > 0$ for all $\varepsilon > 0$ and $np_{\varepsilon_n}(\mathbf{x}) \to \infty$ as $n \to \infty$ (by Corollary 2.1 this is possible).

The number of sample points falling in $\mathcal{R}_{\varepsilon_n}(\mathbf{x})$ is distributed according to some binomial random variable N_n with parameters n and $p_{\varepsilon_n}(\mathbf{x})$. Thus, we may write, for all $A > 0$,

$$\begin{aligned}
\mathbb{P}(N_n < A) &\leq \mathbb{P}(N_n < np_{\varepsilon_n}(\mathbf{x})/2) \\
&\quad \text{(for all } n \text{ large enough)} \\
&= \mathbb{P}(N_n - np_{\varepsilon_n}(\mathbf{x}) < -np_{\varepsilon_n}(\mathbf{x})/2) \\
&\leq 4/\left(np_{\varepsilon_n}(\mathbf{x})\right) \\
&\quad \text{(by Chebyshev's inequality)},
\end{aligned}$$

from which we deduce that $N_n \to \infty$ in probability as $n \to \infty$. This implies that

$$\mathbb{E}\left[\frac{1}{(\log N_n)^{d-1}}\mathbf{1}_{[N_n \geq 2]}\right] \to 0 \text{ as } n \to \infty. \tag{13}$$

Now, denote by K_m the number of maxima in a sequence of m i.i.d. points chosen uniformly at random from $(0,1)^d$. Using the fact that the $\mathbf{Z}_i$'s which fall in $\mathcal{R}_{\varepsilon_n}(\mathbf{x})$ are uniformly distributed on $\mathcal{R}_{\varepsilon_n}(\mathbf{x})$, we note that $L'_{\varepsilon_n}(\mathbf{x})$ and K_{N_n} have the same distribution. Therefore, the theorem will be proven if we show that $K_{N_n} \to \infty$ in probability as $n \to \infty$.

A straightforward adaptation of the arguments in Barndorff-Nielsen and Sobel [4] and Bai et al. [2, 3] shows that there exist two positive constants Δ_1 and Δ_2 such that, on the event $[N_n \geq 2]$,

$$\mathbb{E}[K_{N_n} | N_n] \geq \Delta_1 (\log N_n)^{d-1} \tag{14}$$

and

$$\mathbb{V}[K_{N_n}|N_n] \leq \Delta_2 (\log N_n)^{d-1}. \tag{15}$$

Fix $A > 0$ and $\alpha > 0$, and let the event $\mathcal{E}_n$ be defined as

$$\mathcal{E}_n = \left[N_n < e^{(2A/\Delta_1)^{1/(d-1)}} \vee 2\right].$$

Since $N_n \to \infty$ in probability, one has $\mathbb{P}(\mathcal{E}_n) \leq \alpha$ for all n large enough. Using (14), we may write, conditionally on N_n,

$$\mathbb{P}(K_{N_n} < A|N_n) \leq \mathbb{P}\left(K_{N_n} < \mathbb{E}[K_{N_n}|N_n]/2 \,|\, N_n\right) \mathbf{1}_{\mathcal{E}_n^c} + \mathbf{1}_{\mathcal{E}_n}.$$

Thus, by Chebyshev's inequality and inequalities (14)-(15),

$$\mathbb{P}(K_{N_n} < A|N_n) \leq \frac{\Delta}{(\log N_n)^{d-1}} \mathbf{1}_{\mathcal{E}_n^c} + \mathbf{1}_{\mathcal{E}_n}$$

for some positive constant Δ. Taking expectations on both sides, we finally obtain, for all n large enough,

$$\mathbb{P}(K_{N_n} < A) \leq \mathbb{E}\left[\frac{\Delta}{(\log N_n)^{d-1}} \mathbf{1}_{[N_n \geq 2]}\right] + \alpha,$$

which, together with (13), completes the proof of the lemma. ■

Lemma 5.3 *Let $\Delta \in (0, 1)$. Then, for all $n \geq 1$,*

$$n \int_{[0,\Delta]^d} (1 - \Pi y_i)^{n-1} d\mathbf{y} = \frac{(\log n)^{d-1}}{(d-1)!} + \mathcal{O}_\Delta\left((\log n)^{d-2}\right),$$

where the notation $\mathcal{O}_\Delta$ means that the constant in the $\mathcal{O}$ term depends on Δ.

Proof of Lemma 5.3 The proof starts with the observation (see for example Bai et al. [2]) that

$$n \int_{[0,1]^d} (1 - \Pi y_i)^{n-1} d\mathbf{y} = \frac{(\log n)^{d-1}}{(d-1)!} + \mathcal{O}\left((\log n)^{d-2}\right). \tag{16}$$

To show the result, we proceed by induction on $d \geq 2$. For $d = 2$, we may write

$$n \int_{[0,\Delta]^2} (1 - y_1 y_2)^{n-1} dy_1 dy_2$$

$$
\begin{aligned}
&= n\int_{[0,1]^2}(1-y_1y_2)^{n-1}dy_1dy_2 - n\int_{[0,1]^2\setminus[0,\Delta]^2}(1-y_1y_2)^{n-1}dy_1dy_2\\
&= \log n + \mathcal{O}(1) - n\int_{[0,1]^2\setminus[0,\Delta]^2}(1-y_1y_2)^{n-1}dy_1dy_2\\
&\quad \text{(by identity (16)).}
\end{aligned}
$$

Observing that

$$
\begin{aligned}
n&\int_{[0,1]^2\setminus[0,\Delta]^2}(1-y_1y_2)^{n-1}dy_1dy_2\\
&\le 2n\int_{[0,1]}(1-\Delta y)^{n-1}dy\\
&\le 2/\Delta
\end{aligned}
$$

yields

$$
n\int_{[0,\Delta]^2}(1-y_1y_2)^{n-1}dy_1dy_2 = \log n + \mathcal{O}_\Delta(1),
$$

as desired. Having disposed of this preliminary step, suppose that, for all positive $\Delta \in (0,1)$,

$$
n\int_{[0,\Delta]^d}(1-\Pi y_i)^{n-1}d\mathbf{y} = \frac{(\log n)^{d-1}}{(d-1)!} + \mathcal{O}_\Delta\left((\log n)^{d-2}\right). \tag{17}
$$

Then, for $d+1$,

$$
\begin{aligned}
n&\int_{[0,\Delta]^{d+1}}(1-\Pi y_i)^{n-1}d\mathbf{y}\\
&= n\int_{[0,1]^{d+1}}(1-\Pi y_i)^{n-1}d\mathbf{y} - n\int_{[0,1]^{d+1}\setminus[0,\Delta]^{d+1}}(1-\Pi y_i)^{n-1}d\mathbf{y}\\
&= \frac{(\log n)^d}{d!} + \mathcal{O}\left((\log n)^{d-1}\right) - n\int_{[0,1]^{d+1}\setminus[0,\Delta]^{d+1}}(1-\Pi y_i)^{n-1}d\mathbf{y}\\
&\quad \text{(by identity (16)).}
\end{aligned}
$$

With respect to the rightmost term, we note that

$$
n\int_{[0,1]^{d+1}\setminus[0,\Delta]^{d+1}}(1-\Pi y_i)^{n-1}d\mathbf{y}
$$

$$
\begin{aligned}
&\leq nd \int_{[0,1]^d} (1-\Delta\Pi y_i)^{n-1} d\mathbf{y}\\
&= n(d/\Delta) \int_{[0,\Delta^{1/d}]^d} (1-\Pi y_i)^{n-1} d\mathbf{y}\\
&= \frac{d(\log n)^{d-1}}{\Delta(d-1)!} + \mathcal{O}_\Delta\left((\log n)^{d-2}\right)\\
&\quad \text{(by induction hypothesis (17))}\\
&= \mathcal{O}_\Delta\left((\log n)^{d-1}\right).
\end{aligned}
$$

Putting all pieces together, we obtain

$$
n \int_{[0,\Delta]^{d+1}} (1-\Pi y_i)^{n-1} d\mathbf{y} = \frac{(\log n)^d}{d!} + \mathcal{O}_\Delta\left((\log n)^{d-1}\right),
$$

as desired. ■

For a better understanding of the next two lemmas, the reader should refer to Figure 2.

Lemma 5.4 *Suppose that μ has a probability density f. Fix $\mathbf{x} = (x_1, \ldots, x_d)$, $\varepsilon > 0$, and denote by N_n the (random) number of sample points falling in $\mathcal{R}_\varepsilon(\mathbf{x})$. For $N_n \geq 1$ and each $r = 1, \ldots, d$, let $\mathbf{X}_n^{\star(r)} = (X_{n,1}^{\star(r)}, \ldots, X_{n,d}^{\star(r)})$ be the observation in $\mathcal{R}_\varepsilon(\mathbf{x})$ whose r-coordinate is the closest to x_r. Define the random variables*

$$
M_{n,r} = \begin{cases} +\infty & \text{if } N_n = 0\\ X_{n,r}^{\star(r)} - x_r & \text{if } N_n \geq 1. \end{cases}
$$

Then, for μ-almost all $\mathbf{x}$,

$$
M_{n,r} = \mathcal{O}_{\mathbb{P}}\left(\frac{1}{n}\right),
$$

i.e., for any $\alpha > 0$, there exists $A > 0$ such that, for all n large enough,

$$
\mathbb{P}\left(M_{n,r} \geq \frac{A}{n}\right) \leq \alpha.
$$

Proof of Lemma 5.4 Note first that $\mathbf{X}_n^{\star(r)}$ is almost surely uniquely defined. Choose $\mathbf{x}$ in a set of μ-measure 1 such that $\mu\left(\mathcal{R}_\varepsilon(\mathbf{x})\right) > 0$ and set $p_\varepsilon(\mathbf{x}) = \mu(\mathcal{R}_\varepsilon(\mathbf{x}))$. For any $r = 1, \ldots, d$, let $\mathcal{T}_\varepsilon^r(\mathbf{x})$ be the $d-1$-dimensional rectangle defined by

$$\mathcal{T}_\varepsilon^{(r)}(\mathbf{x}) = \{\mathbf{y} = (y_1, \ldots, y_{r-1}, y_{r+1}, \ldots, y_d) \in \mathbb{R}^{d-1} : x_j \leq y_j \leq x_j + \varepsilon, j \neq r\},$$

and let

$$\begin{aligned} & f_{\varepsilon,\mathbf{x}}^{(r)}(z) \\ & \quad = \frac{\mathbf{1}_{[0\leq z \leq \varepsilon]}}{\mu\left(\mathcal{R}_\varepsilon(\mathbf{x})\right)} \int_{T_\varepsilon^{(r)}(\mathbf{x})} f(y_1, \ldots, y_{r-1}, z, y_{r+1}, \ldots, y_d) dy_1 \ldots dy_{r-1} dy_{r+1} \ldots dy_d \end{aligned}$$

be the marginal density of the distribution μ conditioned by the event $[\mathbf{X} \in \mathcal{R}_\varepsilon(\mathbf{x})]$. Note that we can still choose $\mathbf{x}$ in a set of μ-measure 1 such that, for any $r = 1, \ldots, d$, $f_{\varepsilon,\mathbf{x}}^{(r)}(x_r) > 0$ and $f_{\varepsilon,\mathbf{x}}^{(r)}(z)$ satisfies (3) at x_r, i.e.,

$$\int_{x_r}^{x_r+t} f_{\varepsilon,\mathbf{x}}^{(r)}(z) dz = t f_{\varepsilon,\mathbf{x}}^{(r)}(x_r) + t\zeta_r(t), \quad \text{with } \lim_{t\to 0^+} \zeta_r(t) = 0.$$

Since N_n is binomial with parameters n and $p_{\varepsilon_n}(\mathbf{x})$, we have for any $r = 1, \ldots, d$ and $t > 0$,

$$\begin{aligned} & \mathbb{P}(M_{n,r} \geq t) \\ & \quad = \mathbb{E}\left[\mathbb{P}(M_{n,r} > t | N_n)\right] \\ & \quad \leq \mathbb{E}\left[\mathbf{1}_{[N_n>0]} \mathbb{P}(M_{n,r} > t | N_n)\right] + \mathbb{P}(N_n = 0) \\ & \quad \leq \mathbb{E}\left[\left(1 - \int_{x_r}^{x_r+t} f_{\varepsilon,\mathbf{x}}^{(r)}(z) dz\right)^{N_n}\right] + \left(1 - p_\varepsilon(\mathbf{x})\right)^n \\ & \quad = \left[\left(1 - \int_{x_r}^{x_r+t} f_{\varepsilon,\mathbf{x}}^{(r)}(z) dz\right) p_\varepsilon(\mathbf{x}) + 1 - p_\varepsilon(\mathbf{x})\right]^n + \left(1 - p_\varepsilon(\mathbf{x})\right)^n \\ & \quad = \left(1 - p_\varepsilon(\mathbf{x}) \int_{x_r}^{x_r+t} f_{\varepsilon,\mathbf{x}}^{(r)}(z) dz\right)^n + \left(1 - p_\varepsilon(\mathbf{x})\right)^n \\ & \quad \leq \exp\left(-n p_\varepsilon(\mathbf{x}) \int_{x_r}^{x_r+t} f_{\varepsilon,\mathbf{x}}^{(r)}(z) dz\right) + \exp\left(-n p_\varepsilon(\mathbf{x})\right) \\ & \quad = \exp\left(-nt p_\varepsilon(\mathbf{x}) \left(f_{\varepsilon,\mathbf{x}}^{(r)}(x_r) + \zeta_r(t)\right)\right) + \exp\left(-n p_\varepsilon(\mathbf{x})\right). \end{aligned}$$

This shows that

$$M_{n,r} = \mathcal{O}_{\mathbb{P}}\left(\frac{1}{n}\right),$$

as desired. ■

With the notation of Lemma 5.4, we define the random variable

$$Q_{n,\varepsilon} = \begin{cases} +\infty & \text{if } N_n = 0 \\ \text{the number of sample points falling in } \mathcal{P}_\varepsilon & \text{if } N_n \geq 1, \end{cases}$$

where, in the second statement,

$$\mathcal{P}_\varepsilon = \bigcup_{r=1}^{d} \mathcal{P}_\varepsilon^{(r)}$$

and

$$\begin{aligned} \mathcal{P}_\varepsilon^{(r)} &= [x_1 + \varepsilon, +\infty[\times \ldots \times [x_{r-1} + \varepsilon, +\infty[\\ &\quad \times [x_r, X_{n,r}^{\star(r)}] \\ &\quad \times [x_{r+1} + \varepsilon, +\infty[\times \ldots \times [x_d + \varepsilon, +\infty[. \end{aligned}$$

Lemma 5.5 *Suppose that μ has a probability density f. For μ-almost all $\mathbf{x}$,*

$$Q_{n,\varepsilon} = \mathcal{O}_{\mathbb{P}}(1).$$

Proof of Lemma 5.5 For $N_n \geq 1$, denote by $Q_{n,\varepsilon}^{(r)}$ the number of sample points falling in $\mathcal{P}_\varepsilon^{(r)}$, and set $Q_{n,\varepsilon}^{(r)} = +\infty$ otherwise. Then, clearly,

$$Q_{n,\varepsilon} = \sum_{r=1}^{d} Q_{n,\varepsilon}^{(r)}.$$

Therefore, the result will be proven if we show that, for μ-almost all $\mathbf{x}$ and all $r = 1, \ldots, d$,

$$Q_{n,\varepsilon}^{(r)} = \mathcal{O}_{\mathbb{P}}(1).$$

We fix $\mathbf{x}$ for which Lemma 5.4 is satisfied and fix $r \in \{1, \ldots, d\}$.

Let $\alpha > 0$. According to Lemma 5.4, there exists $A > 0$ such that, for all n large enough,

$$\mathbb{P}\left(M_{n,r} \geq \frac{A}{n}\right) \leq \alpha.$$

Denoting by $\mathcal{E}_n$ the event

$$\left[M_{n,r} < \frac{A}{n}\right],$$

we obtain, for all $t > 0$,

$$\begin{aligned}
&\mathbb{P}\left(Q_{n,\varepsilon}^{(r)} \geq t\right)\\
&\quad = \mathbb{E}\left[\mathbb{P}\left(Q_{n,\varepsilon}^{(r)} \geq t | M_{n,r}\right)\right]\\
&\quad \leq \mathbb{E}\left[\mathbf{1}_{\mathcal{E}_n}\mathbb{P}\left(Q_{n,\varepsilon}^{(r)} \geq t | M_{n,r}\right)\right] + \mathbb{P}(\mathcal{E}_n^c)\\
&\quad \leq \mathbb{E}\left[\mathbf{1}_{\mathcal{E}_n}\mathbb{P}\left(Q_{n,\varepsilon}^{(r)} \geq t | M_{n,r}\right)\right] + \alpha\\
&\qquad \text{(for all n large enough)}\\
&\quad \leq \frac{\mathbb{E}\left[\mathbf{1}_{\mathcal{E}_n}\mathbb{E}[Q_{n,\varepsilon}^{(r)} | M_{n,r}]\right]}{t} + \alpha\\
&\qquad \text{(by Markov's inequality).}
\end{aligned}$$

With respect to the first term in the last inequality we may write, using the definition of $\mathcal{E}_n$,

$$\begin{aligned}
&\mathbf{1}_{\mathcal{E}_n}\mathbb{E}[Q_{n,\varepsilon}^{(r)} | M_{n,r}]\\
&\quad = n\mathbf{1}_{\mathcal{E}_n}\int_{x_1+\varepsilon}^{\infty}\cdots\int_{x_{r-1}+\varepsilon}^{\infty}\int_{x_r}^{x_r+M_{n,r}}\int_{x_{r+1}+\varepsilon}^{\infty}\cdots\int_{x_d+\varepsilon}^{\infty} f(\mathbf{y})d\mathbf{y}\\
&\quad \leq n\int_{x_1+\varepsilon}^{\infty}\cdots\int_{x_{r-1}+\varepsilon}^{\infty}\int_{x_r}^{x_r+A/n}\int_{x_{r+1}+\varepsilon}^{\infty}\cdots\int_{x_d+\varepsilon}^{\infty} f(\mathbf{y})d\mathbf{y}.
\end{aligned}$$

Let

$$\begin{aligned}
&g_{\varepsilon,\mathbf{x}}^{(r)}(z)\\
&\quad = \int_{x_1+\varepsilon}^{\infty}\cdots\int_{x_{r-1}+\varepsilon}^{\infty}\\
&\quad \times \int_{x_{r+1}+\varepsilon}^{\infty}\cdots\int_{x_d+\varepsilon}^{\infty} f(y_1,\ldots,y_{r-1},z,y_{r+1},\ldots,y_d)dy_1\ldots dy_{r-1}dy_{r+1}\ldots dy_d,
\end{aligned}$$

and observe that we can still choose $\mathbf{x}$ in a set of μ-measure 1 such that $g^{(r)}_{\varepsilon,\mathbf{x}}(z)$ satisfies (3), i.e.,

$$\int_{x_r}^{x_r+t} g^{(r)}_{\varepsilon,\mathbf{x}}(z)dz = tg^{(r)}_{\varepsilon,\mathbf{x}}(x_r) + t\zeta_r(t), \quad \text{with } \lim_{t\to 0^+} \zeta_r(t) = 0.$$

Thus, for $\delta > 0$, we can take n large enough to ensure

$$n\int_{x_r}^{x_r+A/n} g^{(r)}_{\varepsilon,\mathbf{x}}(z)dz \le A(1+\delta)g^{(r)}_{\varepsilon,\mathbf{x}}(x_r).$$

Putting all pieces together, we obtain, for any $t > 0$, $\delta > 0$, $\alpha > 0$, and all n large enough,

$$\mathbb{P}\left(Q^{(r)}_{n,\varepsilon} \ge t\right) \le \frac{1}{t}A(1+\delta)g^{(r)}_{\varepsilon,\mathbf{x}}(x_r) + \alpha.$$

This shows that

$$Q^{(r)}_{n,\varepsilon} = \mathcal{O}_{\mathbb{P}}(1).$$

■

References

[1] Amit, Y. and Geman, D. (1997). Shape quantization and recognition with randomized trees, *Neural Computation*, **9**, 1545-1588.

[2] Bai, Z.-D., Devroye, L., Hwang, H.-K. and Tsai, T.-S. (2005). Maxima in hypercubes, *Random Structures and Algorithms*, **27**, 290-309.

[3] Bai, Z.-D., Chao, C.-C., Hwang, H.-K. and Liang, W.-Q. (1998). On the variance of the number of maxima in random vectors and its applications, *The Annals of Applied Probability*, **8**, 886-895.

[4] Barndorff-Nielsen, O. and Sobel, M. (1966). On the distribution of admissible points in a vector random sample, *Theory of Probability and its Applications*, **11**, 249-269.

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

[6] Breiman, L. (2000). Some infinite theory for predictor ensembles, *Technical Report 577*, Statistics Department, UC Berkeley. http://www.stat.berkeley.edu/~breiman .

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

[8] Biau, G., Devroye, L. and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers, *Technical Report*, Université Pierre et Marie Curie, Paris VI. http://www.lsta.upmc.fr/BIAU/bdl2.pdf .

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

[10] Cutler, A. and Zhao, G. (2001). Pert – Perfect random tree ensembles, *Computing Science and Statistics*, **33**, 490-497.

[11] Devroye, L. (1978). The uniform convergence of nearest neighbour regression function estimators and their application in optimization, *IEEE Transactions on Information Theory*, **24**, 142-151.

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

[13] Dietterich, T.G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization, *Machine Learning*, **40**, 139-157.

[14] Doeblin, W. (1937). Exposé de la théorie des chaînes simples constantes de Markov à un nombre fini d'états, *Revue Mathématique de l'Union Interbalkanique*, **2**, 77-105.

[15] Fix, E. and Hodges, J. (1951). Discriminatory analysis. Nonparametric discrimination: Consistency properties, *Technical Report 4*, USAF School of Aviation Medicine, Randolph Field, Texas.

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

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

[18] Marcinkiewicz, J. and Zygmund, A. (1937). Sur les fonctions indépendantes, *Fundamenta Mathematicae*, **29**, 60-90.

[19] Petrov, V.V. (1975). *Sums of Independent Random Variables*, Springer-Verlag, Berlin.

[20] Rachev, S.T. and Rüschendorf, L. (1998). *Mass Transportation Problems, Volume I: Theory*, Springer, New York.

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

[22] Wheeden, R.L. and Zygmund, A. (1977). *Measure and Integral. An Introduction to Real Analysis*, Marcel Dekker, New York.