Springer Series in the Data Sciences

Gérard Biau
Luc Devroye

Lectures on the Nearest Neighbor Method

Springer

Springer Series in the Data Sciences

Springer Series in the Data Sciences focuses primarily on monographs and graduate level textbooks. The target audience includes students and researchers working in and across the fields of mathematics, theoretical computer science, and statistics.

Data Analysis and Interpretation is a broad field encompassing some of the fastest-growing subjects in interdisciplinary statistics, mathematics and computer science. It encompasses a process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision making. Data analysis has multiple facets and approaches, including diverse techniques under a variety of names, in different business, science, and social science domains. Springer Series in the Data Sciences addresses the needs of a broad spectrum of scientists and students who are utilizing quantitative methods in their daily research.

The series is broad but structured, including topics within all core areas of the data sciences. The breadth of the series reflects the variation of scholarly projects currently underway in the field of machine learning.

More information about this series at http://www.springer.com/series/13852

Gérard Biau • Luc Devroye

Lectures on the Nearest Neighbor Method

Springer

Gérard Biau
Université Pierre et Marie Curie
Paris, France

Luc Devroye
McGill University
Montreal, Quebec, Canada

ISSN 2365-5674 ISSN 2365-5682 (electronic)
Springer Series in the Data Sciences
ISBN 978-3-319-25386-2 ISBN 978-3-319-25388-6 (eBook)
DOI 10.1007/978-3-319-25388-6

Library of Congress Control Number: 2015956603

Mathematics Subject Classification (2010): 62G05, 62G07, 62G08, 62G20, 62G30, 62H30, 68T05, 68T10, 60C05, 60D05

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

Children learn effortlessly by example and exhibit a remarkable capacity of generalization. The field of machine learning, on the other hand, stumbles along clumsily in search of algorithms and methods, but nothing available today comes even close to an average two-year-old toddler. So, modestly, we present some results on one of the main paradigms in machine learning—nearest neighbor methods.

Rummaging through old data for the closest match seems like a sensible thing to do, and that primitive idea can be formalized and made rigorous. In the field of nonparametric statistics, where one is concerned with the estimation of densities, distribution functions, regression functions, and functionals, the nearest neighbor family of methods was in the limelight from the very beginning and has achieved some level of maturity.

We do not wish to survey the literature, but we think that it is important to bring the key statistical, probabilistic, combinatorial, and geometric ideas required in the analysis together under one umbrella. Our initial intent was to write chapters that roughly correspond each to a 90-minute lecture, but we were not disciplined enough to carry that through successfully.

The authors were influenced by many others who came before them and thank many colleagues and coauthors for their insight and help. We are particularly grateful to László Györfi.

The book was written in Montreal during Gérard's visits between 2013 and 2015. Bea was very supportive and deserves a special nod. And so do Marie France and Bernard. We also thank Montreal's best pâtissière, Birgit.

Gérard Biau
Luc Devroye
Montreal, February 2015

Contents

Part I
Density estimation

Chapter 1
Order statistics and nearest neighbors

1.1 Uniform order statistics

We start with some basic properties of uniform order statistics. For a general introduction to probability, see Grimmett and Stirzaker (2001). Some of the properties of order statistics presented in this chapter are covered by Rényi (1970); Galambos (1978), and Devroye (1986).

If $U_1, \ldots, U_n$ are i.i.d. uniform $[0, 1]$ random variables, then the order statistics for this sample are $U_{(1,n)}, \ldots, U_{(n,n)}$, where

$$U_{(1,n)} \leq \cdots \leq U_{(n,n)}$$

and $(U_{(1,n)}, \ldots, U_{(n,n)})$ is a permutation of $(U_1, \ldots, U_n)$. Ties occur with zero probability and may be broken arbitrarily, e.g., by declaring that U_i is smaller than U_j whenever $U_i = U_j$ and $i < j$. To simplify the notation, we omit the double indices when no confusion is possible and write $U_{(1)}, \ldots, U_{(n)}$ instead of $U_{(1,n)}, \ldots, U_{(n,n)}$.

By definition, the vector $(U_1, \ldots, U_n)$ is uniformly distributed in the unit cube $[0, 1]^n$. It follows that $(U_{(1)}, \ldots, U_{(n)})$ is also uniformly distributed in the simplex

$$A_n = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : 0 \leq x_1 \leq \cdots \leq x_n \leq 1\}.$$

Throughout the book, $\mathbb{1}_A$ stands for the indicator function of the set A.

Theorem 1.1. *The joint density of* $(U_{(1)}, \ldots, U_{(n)})$ *is*

$$f(x_1, \ldots, x_n) = n! \, \mathbb{1}_{A_n}(x_1, \ldots, x_n).$$

G. Biau, L. Devroye, *Lectures on the Nearest Neighbor Method*,
Springer Series in the Data Sciences, DOI 10.1007/978-3-319-25388-6_1

Proof. We denote by $(\sigma_1, \ldots, \sigma_n)$ the permutation of $(1, \ldots, n)$ such that $U_{(i)} = U_{\sigma_i}$ for all i. Let A be an arbitrary Borel set of $\mathbb{R}^d$. We have

$$\begin{aligned}
&\mathbb{P}\left\{(U_{(1)}, \ldots, U_{(n)}) \in A\right\} \\
&\quad = \sum_{\substack{\text{all permutations} \\ (\tau_1,\ldots,\tau_n) \text{ of } (1,\ldots,n)}} \mathbb{P}\left\{(U_{\sigma_1}, \ldots, U_{\sigma_n}) \in A, (\sigma_1, \ldots, \sigma_n) = (\tau_1, \ldots, \tau_n)\right\} \\
&\quad = n!\, \mathbb{P}\left\{(U_1, \ldots, U_n) \in A, (\sigma_1, \ldots, \sigma_n) = (1, \ldots, n)\right\} \\
&\qquad \text{(since all orderings have the same probability)} \\
&\quad = n! \int_A \mathbb{1}_{A_n}(x_1, \ldots, x_n) \mathrm{d}x_1 \ldots \mathrm{d}x_n.
\end{aligned}$$

The result follows by the arbitrariness of A. □

Next, let $U_1, \ldots, U_n$ be i.i.d. uniform $[0, 1]$ random variables with order statistics $U_{(1)} \leq \cdots \leq U_{(n)}$. The statistics S_i defined by

$$S_i = U_{(i)} - U_{(i-1)}, \quad 1 \leq i \leq n + 1,$$

where, by convention, $U_{(0)} = 0$ and $U_{(n+1)} = 1$ are called the uniform spacings.

Theorem 1.2. *The vector $(S_1, \ldots, S_n)$ is uniformly distributed in the simplex*

$$B_n = \left\{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_i \geq 0, \sum_{i=1}^{n} x_i \leq 1\right\},$$

and the vector $(S_1, \ldots, S_{n+1})$ is uniformly distributed in

$$\left\{(x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} : x_i \geq 0, \sum_{i=1}^{n+1} x_i = 1\right\}.$$

Proof. We know from Theorem 1.1 that $(U_{(1)}, \ldots, U_{(n)})$ is uniformly distributed in the simplex A_n. The transformation

$$\begin{cases}
s_1 = u_1 \\
s_2 = u_2 - u_1 \\
\quad \vdots \\
s_n = u_n - u_{n-1}
\end{cases}$$

has as inverse

$$\begin{cases} u_1 = s_1 \\ u_2 = s_1 + s_2 \\ \quad \vdots \\ u_n = s_1 + s_2 + \cdots + s_n, \end{cases}$$

and the Jacobian, i.e., the determinant of the matrix formed by $\frac{\partial s_j}{\partial u_i}$, is 1. This shows that the density of $(S_1, \ldots, S_n)$ is uniform on the set B_n. The second statement is clear. □

Proofs of this sort can often be obtained without the cumbersome transformations. For example, when $\mathbf{U}$ has the uniform density on a set $A \subseteq \mathbb{R}^d$, and T is a nonsingular linear transformation from $\mathbb{R}^d$ to $\mathbb{R}^d$, then $\mathbf{Z} = T\mathbf{U}$ is uniformly distributed in TA, as can be seen from the following argument: for any Borel set $B \subseteq \mathbb{R}^d$,

$$\begin{aligned} \mathbb{P}\{\mathbf{Z} \in B\} &= \mathbb{P}\{T\mathbf{U} \in B\} = \mathbb{P}\{\mathbf{U} \in T^{-1}B\} \\ &= \frac{\displaystyle\int_{(T^{-1}B)\cap A} d\mathbf{x}}{\displaystyle\int_A d\mathbf{x}} = \frac{\displaystyle\int_{B\cap(TA)} d\mathbf{x}}{\displaystyle\int_{TA} d\mathbf{x}}. \end{aligned}$$

Theorem 1.3. *The vector $(S_1, \ldots, S_{n+1})$ is distributed as*

$$\left(\frac{E_1}{\sum_{i=1}^{n+1} E_i}, \ldots, \frac{E_{n+1}}{\sum_{i=1}^{n+1} E_i} \right),$$

where $E_1, \ldots, E_{n+1}$ are independent standard exponential random variables.

The proof of Theorem 1.3 is based upon Lemma 1.1:

Lemma 1.1. *For any sequence of nonnegative real numbers $x_1, \ldots, x_{n+1}$, we have*

$$\mathbb{P}\{S_1 > x_1, \ldots, S_{n+1} > x_{n+1}\} = \left[\max\left(1 - \sum_{i=1}^{n+1} x_i, 0\right) \right]^n.$$

Proof. Assume, without loss of generality, that $\sum_{i=1}^{n+1} x_i < 1$ (for otherwise the lemma is obviously true). In the notation of Theorem 1.2, we start from the fact that $(S_1, \ldots, S_n)$ is uniformly distributed in B_n. Our probability is equal to

$$\mathbb{P}\left\{S_1 > x_1, \ldots, S_n > x_n, 1 - \sum_{i=1}^{n} S_i > x_{n+1}\right\},$$

that is,

$$\mathbb{P}\Big\{x_1 < S_1, \ldots, x_n < S_n, \sum_{i=1}^{n}(S_i - x_i) < 1 - \sum_{i=1}^{n+1} x_i\Big\}.$$

This is the probability of a set $B_n^\star$ that is a simplex just as B_n, except that its top is not at $(0, \ldots, 0)$ but rather at $(x_1, \ldots, x_n)$, and that its sides are not of length 1 but rather of length $1 - \sum_{i=1}^{n+1} x_i$. For uniform distributions, probabilities can be calculated as ratios of areas. In this case, we have

$$\frac{\int_{B_n^\star} \mathrm{d}\mathbf{x}}{\int_{B_n} \mathrm{d}\mathbf{x}} = \left(1 - \sum_{i=1}^{n+1} x_i\right)^n. \quad \square$$

Proof (Theorem 1.3). Let $G = G_{n+1}$ be the random variable $\sum_{i=1}^{n+1} E_i$. Note that we only need to show that the vector

$$\left(\frac{E_1}{G}, \ldots, \frac{E_n}{G}\right)$$

is uniformly distributed in B_n. The last component $\frac{E_{n+1}}{G}$ is taken care of by noting that it equals 1 minus the sum of the first n components. Let us use the symbols e_i, y, x_i for the running variables corresponding to E_i, G, $\frac{E_i}{G}$. We first compute the joint density of $(E_1, \ldots, E_n, G)$:

$$\begin{aligned} f(e_1, \ldots, e_n, y) &= \prod_{i=1}^{n} e^{-e_i} e^{-(y - e_1 - \cdots - e_n)} \mathbb{1}_{[\min_i e_i \geq 0]} \mathbb{1}_{[y \geq \sum_{i=1}^n e_i]} \\ &= e^{-y} \mathbb{1}_{[\min_i e_i \geq 0]} \mathbb{1}_{[y \geq \sum_{i=1}^n e_i]}. \end{aligned}$$

Here we used the fact that the joint density is the product of the first n variables and the density of G given $E_1 = e_1, \ldots, E_n = e_n$. Next, by the simple transformation of variables $x_1 = \frac{e_1}{y}, \ldots, x_n = \frac{e_n}{y}, y = y$, it is easily seen that the joint density of $(\frac{E_1}{G}, \ldots, \frac{E_n}{G}, G)$ is

$$y^n f(x_1 y, \ldots, x_n y, y) = y^n e^{-y} \mathbb{1}_{[y \geq 0]} \mathbb{1}_{B_n}(x_1, \ldots, x_n).$$

Finally, the density of $(\frac{E_1}{G}, \ldots, \frac{E_n}{G})$ is achieved by integrating the last density with respect to $\mathrm{d}y$, which gives us

$$\int_0^\infty y^n e^{-y} \mathrm{d}y \mathbb{1}_{B_n}(x_1, \ldots, x_n) = n!\, \mathbb{1}_{B_n}(x_1, \ldots, x_n). \quad \square$$

We end this section by two useful corollaries of Theorem 1.3.

Corollary 1.1. *The vector* $(U_{(1)}, \ldots, U_{(n)})$ *is distributed as*

$$\left(\frac{E_1}{\sum_{i=1}^{n+1} E_i}, \frac{E_1 + E_2}{\sum_{i=1}^{n+1} E_i}, \ldots, \frac{E_1 + \cdots + E_n}{\sum_{i=1}^{n+1} E_i} \right),$$

where $E_1, \ldots, E_{n+1}$ *are independent standard exponential random variables.*

Corollary 1.2. *The* i*-th order statistic* $U_{(i)}$ *has the beta density with parameters* i *and* $n + 1 - i$*. Its density is*

$$f(x) = \frac{n!}{(i-1)!(n-i)!} x^{i-1}(1-x)^{n-i}, \quad 0 \le x \le 1.$$

Proof. From Corollary 1.1, we deduce that

$$U_{(i)} \stackrel{\mathscr{D}}{=} \frac{G_i}{G_i + G_{n+1-i}},$$

where G_i and G_{n+1-i} are independent gamma random variables with parameters i and $n + 1 - i$, respectively. The conclusion follows from Lemma 20.9 in the Appendix. □

Of particular importance is the distribution of $U_{(n)} = \max(U_1, \ldots, U_n)$, with density nx^{n-1} on $[0, 1]$. Another important order statistic is the median. The median of $U_1, \ldots, U_{2n+1}$ is $U_{(n+1)}$. Its density is the symmetric beta,

$$f(x) = \frac{(2n+1)!}{n!^2} (x(1-x))^n, \quad 0 \le x \le 1.$$

1.2 The probability integral transform and the k-th order statistic

Throughout the book, the vector space $\mathbb{R}^d$ of all d-tuples $\mathbf{x} = (x_1, \ldots, x_d)$ is equipped with the Euclidean norm $\|\mathbf{x}\| = (x_1^2 + \cdots + x_d^2)^{1/2}$. For $\rho \ge 0$, we denote by $B(\mathbf{x}, \rho)$ the closed ball in $\mathbb{R}^d$ centered at $\mathbf{x}$ of radius ρ, i.e., $B(\mathbf{x}, \rho) = \{\mathbf{y} \in \mathbb{R}^d : \|\mathbf{y} - \mathbf{x}\| \le \rho\}$. Similarly, we define $B^\circ(\mathbf{x}, \rho)$ as the open ball centered at $\mathbf{x}$ of radius ρ, i.e., $B(\mathbf{x}, \rho) = \{\mathbf{y} \in \mathbb{R}^d : \|\mathbf{y} - \mathbf{x}\| < \rho\}$.

The probability integral transform states that if U is uniform $[0, 1]$ and the real-valued random variable X has continuous distribution function F, then

$$F(X) \stackrel{\mathscr{D}}{=} U.$$

Now, let $\mathbf{X}$ be a $\mathbb{R}^d$-valued random variable with distribution μ, and let $\mathbf{x}$ be a fixed point in $\mathbb{R}^d$. When μ has a density f with respect to the Lebesgue measure, then the random variable $\|\mathbf{X} - \mathbf{x}\|$ is continuous, and the probability integral transform implies that

$$\mu\big(B(\mathbf{x}, \|\mathbf{X} - \mathbf{x}\|)\big) \stackrel{\mathscr{D}}{=} U.$$

If $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d. with common distribution μ, and if $U_1, \ldots, U_n$ are i.i.d. uniform $[0, 1]$, then

$$\Big(\mu\big(B(\mathbf{x}, \|\mathbf{X}_1 - \mathbf{x}\|)\big), \ldots, \mu\big(B(\mathbf{x}, \|\mathbf{X}_n - \mathbf{x}\|)\big)\Big) \stackrel{\mathscr{D}}{=} (U_1, \ldots, U_n),$$

and using reordered samples with

$$\|\mathbf{X}_{(1)}(\mathbf{x}) - \mathbf{x}\| \le \cdots \le \|\mathbf{X}_{(n)}(\mathbf{x}) - \mathbf{x}\|$$

and

$$U_{(1)} \le \cdots \le U_{(n)},$$

we have

$$\Big(\mu\big(B(\mathbf{x}, \|\mathbf{X}_{(1)}(\mathbf{x}) - \mathbf{x}\|)\big), \ldots, \mu\big(B(\mathbf{x}, \|\mathbf{X}_{(n)}(\mathbf{x}) - \mathbf{x}\|)\big)\Big) \stackrel{\mathscr{D}}{=} \big(U_{(1)}, \ldots, U_{(n)}\big).$$

The study of $\mu(B(\mathbf{x}, \|\mathbf{X}_{(k)}(\mathbf{x}) - \mathbf{x}\|))$ is thus the study of $U_{(k)}$.

At numerous places in the book, the index k is allowed to vary with n. In this case, we continue to use the notation k (instead of k_n) and implicitly assume that $1 \le k \le n$. Observe that

$$\mathbb{E}U_{(k)} = \frac{k}{n+1},$$

since, by Corollary 1.2, $U_{(k)}$ is Beta$(k, n + 1 - k)$. Similarly (see Section 20.9 in the Appendix),

$$\mathbb{V}U_{(k)} = \frac{k(n+1-k)}{(n+1)^2(n+2)},$$

where the symbol $\mathbb{V}$ denotes variance. Therefore, by Chebyshev's inequality, for $\delta > 0$,

$$\begin{aligned}
\mathbb{P}\left\{\left|U_{(k)} - \frac{k}{n+1}\right| > \frac{\delta k}{n+1}\right\} &\le \frac{\mathbb{V}U_{(k)}}{(\delta k/(n+1))^2} \\
&= \frac{1}{\delta^2 k} \times \frac{n+1-k}{n+2} \\
&\le \frac{1}{\delta^2 k}.
\end{aligned}$$

Thus, we immediately have the law of large numbers:

Theorem 1.4. *If $k \to \infty$, then*

$$\frac{U_{(k)}}{k/n} \to 1 \quad \textit{in probability.}$$

The strong law of large numbers can be shown using tail inequalities for $U_{(k)}$. In particular, for $\delta > 0$,

$$\begin{aligned}
\mathbb{P}\left\{U_{(k)} > \frac{(1+\delta)k}{n}\right\} &= \mathbb{P}\left\{\mathrm{Bin}\left(n, \tfrac{(1+\delta)k}{n}\right) < k\right\} \\
&\le \exp\left(k - (1+\delta)k - k\log\left(\frac{k}{(1+\delta)k}\right)\right) \\
&\quad \text{(by Chernoff's bound—see Theorem 20.5 in the Appendix)} \\
&= \exp\left([\log(1+\delta) - \delta]k\right)
\end{aligned}$$

and, for $\delta \in (0, 1)$,

$$\begin{aligned}
\mathbb{P}\left\{U_{(k)} < \frac{(1-\delta)k}{n}\right\} &= \mathbb{P}\left\{\mathrm{Bin}\left(n, \tfrac{(1-\delta)k}{n}\right) \ge k\right\} \\
&\le \exp\left(k - (1-\delta)k - k\log\left(\frac{k}{(1-\delta)k}\right)\right) \\
&= \exp\left(\left[\delta + \log(1-\delta)\right]k\right).
\end{aligned}$$

Both upper bounds are of the form $e^{-\alpha k}$ for some $\alpha > 0$, and thus, by the Borel-Cantelli lemma:

Theorem 1.5. *If $k \to \infty$ such that, for all $\alpha > 0$,*

$$\sum_{n \ge 1} e^{-\alpha k} < \infty, \tag{1.1}$$

then

$$\frac{U_{(k)}}{k/n} \to 1 \quad \textit{almost surely.}$$

We note that $k/\log n \to \infty$ is sufficient for (1.1). It is a fun exercise to show that if k is monotonically increasing, then (1.1) implies $k/\log n \to \infty$ as well.

Theorem 1.5 implies the strong law of large numbers. However, something gets lost in the Borel-Cantelli argument. We offer the following "improved" strong law of large numbers, which is roughly equivalent to a result by Kiefer (1972).

Theorem 1.6. *If $k \to \infty$ such that, for all $n, m \geq 1$,*

$$1 \leq \frac{k_{n+m}}{k_n} \leq \psi\left(\frac{m}{n}\right), \tag{1.2}$$

where $\psi \geq 1$ is an increasing function with $\lim_{\delta \downarrow 0} \psi(\delta) = 1$, and if $k/(\log\log n) \to \infty$, then

$$\frac{U_{(k)}}{k/n} \to 1 \quad \textit{almost surely}.$$

We note that the smoothness condition (1.2) is satisfied for most monotone choices. In particular, it holds if $k_n = \lfloor n^\alpha \rfloor$ for any $\alpha > 0$, or if $k_n = \lfloor \log^\beta n \rfloor$ for any $\beta > 0$ ($\lfloor \cdot \rfloor$ is the floor function). Because the proof requires only minimal effort, and introduces a well-known sequencing trick, we offer it here.

Proof (Theorem 1.6). We partition the integers into sections defined by the thresholds

$$n_\ell = \left\lfloor (1+\delta)^\ell \right\rfloor,$$

where $\ell = 1, 2, \ldots$, and δ is a positive constant to be selected later. For all n large enough and

$$n_\ell < n \leq n_{\ell+1},$$

we have, using the fact that $U_{(k,n)}$ is increasing in k (for fixed n) and decreasing in n (for fixed k),

$$L_\ell \stackrel{\text{def}}{=} \frac{U_{(k_{n_\ell}+1, n_\ell+1)}}{k_{n_{\ell+1}}/(n_\ell+1)} \leq \frac{U_{(k_n,n)}}{k_n/n} \leq \frac{U_{(k_{n_{\ell+1}}, n_\ell+1)}}{k_{n_{\ell+1}}/n_{\ell+1}} \stackrel{\text{def}}{=} R_\ell.$$

By the Borel-Cantelli lemma, we are done if we can show that, for all $\varepsilon > 0$,

$$\sum_{\ell \geq 1} \mathbb{P}\{R_\ell > 1+\varepsilon\} < \infty$$

and, for all $\varepsilon \in (0,1)$,

$$\sum_{\ell \geq 1} \mathbb{P}\{L_\ell < 1-\varepsilon\} < \infty.$$

We show the part involving R_ℓ, and leave the other part as a small exercise. Observe that

$$\begin{aligned}
\mathbb{P}\{R_\ell > 1+\varepsilon\} &= \mathbb{P}\left\{U_{(k_{n_\ell+1}, n_\ell+1)} > \frac{(1+\varepsilon)k_{n_\ell+1}}{n_{\ell+1}}\right\} \\
&= \mathbb{P}\left\{\mathrm{Bin}\Big(n_\ell+1, \tfrac{(1+\varepsilon)k_{n_\ell+1}}{n_{\ell+1}}\Big) < k_{n_{\ell+1}}\right\}.
\end{aligned}$$

Since

$$\frac{n_\ell + 1}{n_{\ell+1}}(1+\varepsilon) \geq \frac{(1+\delta)^\ell(1+\varepsilon)}{(1+\delta)^{\ell+1}} = \frac{1+\varepsilon}{1+\delta},$$

then, for $\delta \leq \varepsilon$, by Chernoff's bound (Theorem 20.5),

$$\begin{aligned}&\mathbb{P}\{R_\ell > 1+\varepsilon\}\\ &\quad\leq \exp\left(k_{n_{\ell+1}} - \frac{n_\ell+1}{n_{\ell+1}}(1+\varepsilon)k_{n_\ell+1} - k_{n_\ell+1}\log\left(\frac{k_{n_{\ell+1}}}{k_{n_\ell+1}} \times \frac{n_{\ell+1}}{n_\ell+1} \times \frac{1}{1+\varepsilon}\right)\right).\end{aligned}$$

Using

$$\frac{n_\ell+1}{n_{\ell+1}} \geq \frac{(1+\delta)^\ell}{(1+\delta)^{\ell+1}} = \frac{1}{1+\delta}, \quad \frac{n_{\ell+1}}{n_\ell+1} \geq 1, \quad \frac{k_{n_{\ell+1}}}{k_{n_\ell+1}} \geq 1,$$

and

$$\frac{k_{n_{\ell+1}}}{k_{n_\ell+1}} \leq \psi\left(\frac{n_{\ell+1} - n_\ell - 1}{n_\ell+1}\right) \leq \psi(\delta),$$

we bound the last expression by

$$\exp\left(k_{n_\ell+1}\left[\psi(\delta) - \frac{1+\varepsilon}{1+\delta} + \psi(\delta)\log(1+\varepsilon)\right]\right).$$

The quantity in the square brackets tends to $\log(1+\varepsilon) - \varepsilon < 0$ as $\delta \downarrow 0$. So, choose $\delta > 0$ such that its value is $-\alpha$ for some $\alpha > 0$. Thus,

$$\mathbb{P}\{R_\ell > 1+\varepsilon\} \leq e^{-\alpha k_{n_\ell+1}}.$$

For all n large enough, we have $k_n > \frac{2}{\alpha}\log\log n$. Therefore, for all ℓ large enough,

$$\begin{aligned}\mathbb{P}\{R_\ell > 1+\varepsilon\} &\leq \exp\left(-2\log\log(n_\ell+1)\right)\\ &= \frac{1}{\log^2(n_\ell+1)}\\ &\leq \frac{1}{\ell^2\log^2(1+\delta)},\end{aligned}$$

so that

$$\sum_{\ell\geq 1}\mathbb{P}\{R_\ell > 1+\varepsilon\} < \infty,$$

as required. □

Chapter 2
The nearest neighbor distance

2.1 Consistency

Let $\mathbf{X}$ be a random variable taking values in $\mathbb{R}^d$, and let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be an i.i.d. sample drawn from $\mathbf{X}$. For fixed $\mathbf{x} \in \mathbb{R}^d$, we denote by $\mathbf{X}_{(1)}(\mathbf{x}), \ldots, \mathbf{X}_{(n)}(\mathbf{x})$ a reordering of $\mathbf{X}_1, \ldots, \mathbf{X}_n$ according to increasing values of $\|\mathbf{X}_i - \mathbf{x}\|$, that is,

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

If $\mathbf{X}_i$ and $\mathbf{X}_j$ are equidistant from $\mathbf{x}$, i.e., if $\|\mathbf{X}_i - \mathbf{x}\| = \|\mathbf{X}_j - \mathbf{x}\|$ for some $i \neq j$, then we have a distance tie. By convention, ties are broken by comparing indices, that is, by declaring that $\mathbf{X}_i$ is closer to $\mathbf{x}$ than $\mathbf{X}_j$ whenever $i < j$.

Let k be an integer comprised between 1 and n. A natural concern is to know whether the distance $\|\mathbf{X}_{(k)}(\mathbf{X}) - \mathbf{X}\|$ approaches zero in some probabilistic sense when the sample size tends to infinity and k may possibly vary with n. To answer this question, denote by μ the distribution of $\mathbf{X}$, and recall that the support of $\mathbf{X}$ (or support of μ) is defined by

$$\text{supp}(\mu) = \left\{\mathbf{x} \in \mathbb{R}^d : \mu\left(B(\mathbf{x}, \rho)\right) > 0 \text{ for all } \rho > 0\right\}.$$

Its properties are well known (see, e.g., Kallenberg, 2002):

(*i*) $\text{supp}(\mu)$ is a closed set.
(*ii*) $\text{supp}(\mu)$ is the smallest closed subset of $\mathbb{R}^d$ of μ-measure one.
(*iii*) One has $\mathbb{P}\{\mathbf{X} \in \text{supp}(\mu)\} = 1$.

A density is an equivalence class. For fixed f, its equivalence class consists of all g for which $\int_A g(\mathbf{x})\mathrm{d}\mathbf{x} = \int_A f(\mathbf{x})\mathrm{d}\mathbf{x}$ for all Borel sets A. In particular, if $f = g$ Lebesgue-almost everywhere, then g is in the equivalence class of f. Define

G. Biau, L. Devroye, *Lectures on the Nearest Neighbor Method*,
Springer Series in the Data Sciences, DOI 10.1007/978-3-319-25388-6_2

$$\underline{f}(\mathbf{x}) = \liminf_{\rho \downarrow 0} \frac{\int_{B(\mathbf{x},\rho)} f(\mathbf{y})\mathrm{d}\mathbf{y}}{\lambda\left(B(\mathbf{x},\rho)\right)}, \quad \bar{f}(\mathbf{x}) = \limsup_{\rho \downarrow 0} \frac{\int_{B(\mathbf{x},\rho)} f(\mathbf{y})\mathrm{d}\mathbf{y}}{\lambda\left(B(\mathbf{x},\rho)\right)},$$

where $B(\mathbf{x}, \rho)$ is the closed ball centered at $\mathbf{x}$ of radius ρ, and λ denotes the Lebesgue measure on $\mathbb{R}^d$. Both $\underline{f}$ and $\bar{f}$ are equivalent to f since $\underline{f} = \bar{f} = f$ at λ-almost all $\mathbf{x}$ by the Lebesgue differentiation theorem (see Theorem 20.18 in the Appendix). If the probability measure μ has density f, it is in general not true that

$$\mathrm{supp}(\mu) = \overline{\{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) > 0\}},$$

where $\bar{A}$ is the closure of the set A. However, we have the following:

Lemma 2.1. *Let the probability measure μ have a density f. If $A = \{\mathbf{x} \in \mathbb{R}^d : \underline{f}(\mathbf{x}) > 0\}$, then $\mathrm{supp}(\mu) = \bar{A}$.*

Proof. If $\mathbf{x} \in A$, then $\mu(B(\mathbf{x}, \rho)) > 0$ for all $\rho > 0$, and thus $A \subseteq \mathrm{supp}(\mu)$. Since $\mathrm{supp}(\mu)$ is closed, $\bar{A} \subseteq \mathrm{supp}(\mu)$.

Next, we take $\mathbf{x} \in \mathrm{supp}(\mu)$. We construct a sequence $\{\mathbf{x}_n\} \in A$ with $\mathbf{x}_n \to \mathbf{x}$, which shows that $\mathbf{x} \in \bar{A}$, and thus $\mathrm{supp}(\mu) \subseteq \bar{A}$. Since $\mathbf{x} \in \mathrm{supp}(\mu)$, we have $\mu(B(\mathbf{x}, 1/n)) > 0$ for all n. For fixed n, find $\mathbf{x}_n$ in $B(\mathbf{x}, 1/n)$ such that

$$\underline{f}(\mathbf{x}_n) \geq \frac{\int_{B(\mathbf{x},1/n)} \underline{f}(\mathbf{y})\mathrm{d}\mathbf{y}}{\lambda\left(B(\mathbf{x},1/n)\right)} = \frac{\int_{B(\mathbf{x},1/n)} f(\mathbf{y})\mathrm{d}\mathbf{y}}{\lambda\left(B(\mathbf{x},1/n)\right)} > 0,$$

so $\mathbf{x}_n \in A$, and $\|\mathbf{x}_n - \mathbf{x}\| \leq 1/n$. □

The support of the probability measure μ plays an important role in nearest neighbor analysis because of Lemma 2.2 below.

Lemma 2.2. *For $\mathbf{x} \in \mathbb{R}^d$, set*

$$\rho_{\mathbf{x}} = \inf\{\|\mathbf{y} - \mathbf{x}\| : \mathbf{y} \in \mathrm{supp}(\mu)\}.$$

If $k/n \to 0$, then $\|\mathbf{X}_{(k)}(\mathbf{x}) - \mathbf{x}\| \to \rho_{\mathbf{x}}$ almost surely. In particular, if $\mathbf{x} \in \mathrm{supp}(\mu)$ and $k/n \to 0$, then $\|\mathbf{X}_{(k)}(\mathbf{x}) - \mathbf{x}\| \to 0$ almost surely.

Proof. First note, since $\mathrm{supp}(\mu)$ is a closed set, that $\rho_{\mathbf{x}} = 0$ if and only if $\mathbf{x}$ belongs to $\mathrm{supp}(\mu)$. Moreover, by definition of the support, $\mathbf{X}_{(k)}(\mathbf{x})$ falls in $\mathrm{supp}(\mu)$ with probability one. Therefore, with probability one, $\|\mathbf{X}_{(k)}(\mathbf{x}) - \mathbf{x}\| \geq \rho_{\mathbf{x}}$. Now, let $\varepsilon > 0$ be arbitrary, and let

$$p_{\mathbf{x}} = \mathbb{P}\{\|\mathbf{X} - \mathbf{x}\| \leq \varepsilon + \rho_{\mathbf{x}}\} = \mu\left(B\left(\mathbf{x}, \varepsilon + \rho_{\mathbf{x}}\right)\right) > 0.$$

Then, for all n large enough,

$$\begin{aligned}\mathbb{P}\left\{\|\mathbf{X}_{(k)}(\mathbf{x})-\mathbf{x}\|-\rho_{\mathbf{x}}>\varepsilon\right\} &= \mathbb{P}\left\{\|\mathbf{X}_{(k)}(\mathbf{x})-\mathbf{x}\|>\varepsilon+\rho_{\mathbf{x}}\right\}\\ &= \mathbb{P}\left\{\mathrm{Bin}(n,p_{\mathbf{x}})<k\right\}\\ &\le \mathbb{P}\left\{\mathrm{Bin}(n,p_{\mathbf{x}})-np_{\mathbf{x}}<-\frac{np_{\mathbf{x}}}{2}\right\}.\end{aligned}$$

Thus, by Markov's inequality and Corollary 20.2 in the Appendix,

$$\begin{aligned}\mathbb{P}\left\{\|\mathbf{X}_{(k)}(\mathbf{x})-\mathbf{x}\|-\rho_{\mathbf{x}}>\varepsilon\right\} &\le \frac{\mathbb{E}|\mathrm{Bin}(n,p_{\mathbf{x}})-np_{\mathbf{x}}|^4}{(np_{\mathbf{x}}/2)^4}\\ &\le \frac{cn^2\mathbb{E}|\mathrm{Ber}(p_{\mathbf{x}})-p_{\mathbf{x}}|^4}{(np_{\mathbf{x}}/2)^4}\\ &\le \frac{16c}{n^2p_{\mathbf{x}}^4},\end{aligned} \tag{2.1}$$

where c is a positive constant. These probabilities are summable in n for all $\varepsilon>0$. Therefore, $\|\mathbf{X}_{(k)}(\mathbf{x})-\mathbf{x}\|\to\rho_{\mathbf{x}}$ almost surely. The second assertion of the lemma is clear. □

Remark 2.1. We leave it as an exercise to show, using Chernoff's bound (Theorem 20.5 in the Appendix), that for $k/n\to 0$ and all n large enough,

$$\mathbb{P}\left\{\|\mathbf{X}_{(k)}(\mathbf{x})-\mathbf{x}\|-\rho_{\mathbf{x}}>\varepsilon\right\}\le e^{-\alpha n}$$

for some $\alpha>0$ (depending upon $\mathbf{x}$ and ε). □

Since the support of $\mathbf{X}$ is of μ-measure one, we conclude from Lemma 2.2 that, at μ-almost all $\mathbf{x}$, $\|\mathbf{X}_{(k)}(\mathbf{x})-\mathbf{x}\|\to 0$ almost surely whenever $k/n\to 0$. In the same vein, we have the following lemma, which answers the question asked at the beginning of the chapter.

Lemma 2.3. *Assume that* $\mathbf{X}$ *is independent of the data* $\mathbf{X}_1,\ldots,\mathbf{X}_n$. *If* $k/n\to 0$, *then* $\|\mathbf{X}_{(k)}(\mathbf{X})-\mathbf{X}\|\to 0$ *almost surely.*

Proof. By independence of $\mathbf{X}$ and $\mathbf{X}_1,\ldots,\mathbf{X}_n$, we have, for all $\varepsilon>0$,

$$\mathbb{P}\left\{\|\mathbf{X}_{(k)}(\mathbf{X})-\mathbf{X}\|>\varepsilon\right\}=\int_{\mathbb{R}^d}\mathbb{P}\left\{\|\mathbf{X}_{(k)}(\mathbf{x})-\mathbf{x}\|>\varepsilon\right\}\mu(\mathrm{d}\mathbf{x}).$$

This last term converges to zero by Lemma 2.2 and the Lebesgue dominated convergence theorem. This shows the convergence in probability towards zero of $\|\mathbf{X}_{(k)}(\mathbf{X})-\mathbf{X}\|$. To establish the almost sure convergence, we use the more precise notation $\mathbf{X}_{(k,n)}(\mathbf{X})=\mathbf{X}_{(k)}(\mathbf{X})$ and prove that the sequence $\{\sup_{m\ge n}\|\mathbf{X}_{(k,m)}(\mathbf{X})-\mathbf{X}\|\}_{n\ge 1}$ tends to zero in probability. If k does not change with n, then $\|\mathbf{X}_{(k,m)}(\mathbf{X})-\mathbf{X}\|$ is monotonically decreasing, so that

$$\sup_{m \geq n} \|\mathbf{X}_{(k,m)}(\mathbf{X}) - \mathbf{X}\| = \|\mathbf{X}_{(k,n)}(\mathbf{X}) - \mathbf{X}\|$$

and the lemma is proved. If k is allowed to vary with n such that $k/n \to 0$, then, according to (2.1), at μ-almost all $\mathbf{x}$, for some positive C and all n large enough,

$$\mathbb{P}\left\{\sup_{m \geq n} \|\mathbf{X}_{(k,m)}(\mathbf{x}) - \mathbf{x}\| > \varepsilon\right\} \leq \sum_{m \geq n} \mathbb{P}\left\{\|\mathbf{X}_{(k,m)}(\mathbf{x}) - \mathbf{x}\| > \varepsilon\right\} \leq C \sum_{m \geq n} \frac{1}{m^2}.$$

This shows that $\sup_{m \geq n} \|\mathbf{X}_{(k,m)}(\mathbf{x}) - \mathbf{x}\|$ tends to zero in probability (as $n \to \infty$) at μ-almost all $\mathbf{x}$, and thus, by dominated convergence, that $\sup_{m \geq n} \|\mathbf{X}_{(k,m)}(\mathbf{X}) - \mathbf{X}\|$ tends to zero in probability. □

2.2 Rates of convergence

As in the preceding section, we let $\mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n$ be i.i.d. random vectors of $\mathbb{R}^d$, and let $\mathbf{X}_{(1)}(\mathbf{X}), \ldots, \mathbf{X}_{(n)}(\mathbf{X})$ be a reordering of $\mathbf{X}_1, \ldots, \mathbf{X}_n$ according to increasing values of $\|\mathbf{X}_i - \mathbf{X}\|$. For various applications and approximations, we will require information on the size of

$$\mathbb{E}\|\mathbf{X}_{(1)}(\mathbf{X}) - \mathbf{X}\|^2, \tag{2.2}$$

and, more generally, of $\mathbb{E}\|\mathbf{X}_{(k)}(\mathbf{X}) - \mathbf{X}\|^2$, for k an integer between 1 and n. This problem and related questions are explored in Evans et al. (2002); Liitiäinen et al. (2008a); Penrose and Yukich (2011)—see also Bickel and Breiman (1983). Nearest neighbor distances play a significant role in residual variance estimation (Devroye et al., 2003; Liitiäinen et al., 2007, 2008b, 2010), entropy estimation (Kozachenko and Leonenko, 1987; Leonenko et al., 2008; see also Chapter 7), and convergence analysis of estimates (Kulkarni and Posner, 1995; Kohler et al., 2006; Biau et al., 2010; see also Chapter 14).

By symmetry, (2.2) is the same as $\mathbb{E}\|\mathbf{X}_{(1,1)} - \mathbf{X}_1\|^2$, where $\mathbf{X}_1, \ldots, \mathbf{X}_{n+1}$ are i.i.d. random vectors, and $\mathbf{X}_{(1,1)}$ is the nearest neighbor of $\mathbf{X}_1$ among $\mathbf{X}_2, \ldots, \mathbf{X}_{n+1}$. Denoting by $\mathbf{X}_{(i,1)}$ the nearest neighbor of $\mathbf{X}_i$ among $\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_{n+1}$, this is also the same as

$$\frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{E}\|\mathbf{X}_{(i,1)} - \mathbf{X}_i\|^2. \tag{2.3}$$

The quantity (2.2) can be infinite—just on the real line, note that it is at least $\frac{1}{n+1}\mathbb{E}|X_{(n)} - X_{(n+1)}|^2$ if $X_{(1)} \leq \cdots \leq X_{(n+1)}$ are the order statistics for $X_1, \ldots, X_{n+1}$, because $X_{(n)}$ is the nearest neighbor of $X_{(n+1)}$. It is easy to construct long-tailed distributions with $\mathbb{E}|X_{(n)} - X_{(n+1)}|^2 = \infty$.

However, given that $\mathbf{X}_1 \in [0,1]^d$, there are universal (i.e., distribution-free) bounds for (2.2). This remarkable fact is captured in Theorem 2.1 below. Throughout, we let V_d be the volume of the unit Euclidean ball in $\mathbb{R}^d$, and recall that

$$V_d = \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)},$$

where $\Gamma(\cdot)$ stands for the gamma function, defined for $x > 0$ by $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\mathrm{d}t$.

Theorem 2.1. *Let* $\mathbf{X}$ *takes values in* $[0,1]^d$. *Then, for* $d \geq 2$,

$$\mathbb{E}\|\mathbf{X}_{(1)}(\mathbf{X}) - \mathbf{X}\|^2 \leq c_d \left(\frac{1}{n+1}\right)^{2/d},$$

where

$$c_d = \frac{4(1+\sqrt{d})^2}{V_d^{2/d}}.$$

For $d = 1$, *we have*

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

Remark 2.2. The set $[0,1]^d$ is chosen for simplicity of the analysis. We leave it to the reader to adapt the results to the case where $\mathbf{X}$ takes values in an arbitrary compact subset of $\mathbb{R}^d$. □

Proof (Theorem 2.1). For the proof, it is convenient to consider form (2.3). Let $R_i = \|\mathbf{X}_{(i,1)} - \mathbf{X}_i\|$. Let $B_i = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x} - \mathbf{X}_i\| < R_i/2\}$. Note that when $R_i = 0$, then $B_i = \emptyset$. Clearly, the B_i's are disjoint. Since $R_i \leq \sqrt{d}$, we see that $\cup_{i=1}^{n+1} B_i \subseteq [-\frac{\sqrt{d}}{2}, 1+\frac{\sqrt{d}}{2}]^d$. Therefore, if λ is the Lebesgue measure,

$$\lambda\left(\bigcup_{i=1}^{n+1} B_i\right) \leq (1+\sqrt{d})^d.$$

Hence,

$$\sum_{i=1}^{n+1} V_d \left(\frac{R_i}{2}\right)^d \leq (1+\sqrt{d})^d. \tag{2.4}$$

Now, for $d \geq 2$,

$$\left(\frac{1}{n+1}\sum_{i=1}^{n+1} R_i^2\right)^{d/2} \leq \frac{1}{n+1}\sum_{i=1}^{n+1} R_i^d$$

(by Jensen's inequality)

$$\leq \frac{1}{n+1} \times \frac{2^d(1+\sqrt{d})^d}{V_d}.$$

The theorem follows immediately for $d \geq 2$. For $d = 1$, we only have

$$\frac{1}{n+1}\sum_{i=1}^{n+1} R_i^2 \leq \left(\frac{1}{n+1}\sum_{i=1}^{n+1} R_i\right) \times \max_{1 \leq i \leq n+1} R_i$$

$$\leq \frac{2}{n+1}. \quad \square$$

Remark 2.3. It is a good exercise to show that, up to multiplicative constants possibly depending upon d, these bounds are best possible, in the sense that

$$\sup_{\text{all distributions of } \mathbf{X} \text{ on } [0,1]^d} \mathbb{E}\|\mathbf{X}_{(1)}(\mathbf{X}) - \mathbf{X}\|^2 \geq \begin{cases} \frac{\alpha_d}{(n+1)^{2/d}} & \text{for } d \geq 2 \\ \frac{\alpha_1}{n+1} & \text{for } d = 1, \end{cases}$$

for constants α_d. For $d = 1$, consider the distribution

$$\mathbb{P}\{X = 0\} = \frac{1}{n}, \quad \mathbb{P}\{X = 1\} = 1 - \frac{1}{n}.$$

Then

$$\mathbb{E}|X_{(1)}(X) - X|^2 \geq \mathbb{P}\{X = 0, X_1 = \cdots = X_n = 1\}$$

$$= \frac{1}{n}\left(1 - \frac{1}{n}\right)^n \sim \frac{1}{en} \quad \text{as } n \to \infty.$$

For $d \geq 2$, it suffices to consider the uniform distribution on $[0,1]^d$ and recall that $\mu(B(\mathbf{X}, \|\mathbf{X}_{(1)}(\mathbf{X}) - \mathbf{X}\|)) \stackrel{\mathscr{D}}{=} U_{(1)}$, where $U_{(1)} \leq \cdots \leq U_{(n)}$ are uniform $[0,1]$ order statistics (Chapter 1). Clearly, for this distribution, $\mu(B(\mathbf{X}, \|\mathbf{X}_{(1)}(\mathbf{X}) - \mathbf{X}\|)) \leq V_d\|\mathbf{X}_{(1)}(\mathbf{X}) - \mathbf{X}\|^d$. The conclusion follows by recalling that $U_{(1)}$ is Beta$(1, n)$ and by working out the moments of the beta law (see Section 20.9 in the Appendix). $\square$

For singular distributions (with respect to the Lebesgue measure), the behavior of (2.2) is better than predicted by the bounds.

Theorem 2.2. *If* $\mathbf{X}$ *is singular and takes values in* $[0,1]^d$*, then*

$$n^{1/d}\|\mathbf{X}_{(1)}(\mathbf{X}) - \mathbf{X}\| \to 0 \quad \textit{in probability.}$$

(This theorem is valid for all $d \geq 1$*.)*

Proof. By Theorem 20.20 in the Appendix, if μ is the singular probability measure of $\mathbf{X}$, and $B(\mathbf{x}, \rho)$ is the closed ball of radius ρ centered at $\mathbf{x}$, then, at μ-almost all $\mathbf{x}$,

$$\frac{\mu\left(B(\mathbf{x}, \rho)\right)}{\rho^d} \to \infty \quad \text{as } \rho \downarrow 0. \tag{2.5}$$

We show that for any $\varepsilon > 0$,

$$\mathbb{P}\left\{\|\mathbf{X}_{(1)}(\mathbf{X}) - \mathbf{X}\| > \frac{\varepsilon}{n^{1/d}}\right\} \to 0.$$

We have

$$\begin{aligned}
\limsup_{n\to\infty} \mathbb{P}\left\{\|\mathbf{X}_{(1)}(\mathbf{X}) - \mathbf{X}\| > \frac{\varepsilon}{n^{1/d}}\right\} &= \limsup_{n\to\infty} \mathbb{E}\left[\left(1 - \mu\left(B\left(\mathbf{X}, \frac{\varepsilon}{n^{1/d}}\right)\right)\right)^n\right] \\
&\leq \mathbb{E}\left[\limsup_{n\to\infty}\left(1 - \mu\left(B\left(\mathbf{X}, \frac{\varepsilon}{n^{1/d}}\right)\right)\right)^n\right] \\
&\quad \text{(by Fatou's lemma)} \\
&\leq \mathbb{E}\left[\limsup_{n\to\infty} \exp\left(-n\mu\left(B\left(\mathbf{X}, \frac{\varepsilon}{n^{1/d}}\right)\right)\right)\right] \\
&\quad \text{(since } 1 - u \leq e^{-u} \text{ for all } u\text{)} \\
&= \mathbb{E}\left[\exp\left(-\liminf_{n\to\infty} n\mu\left(B\left(\mathbf{X}, \frac{\varepsilon}{n^{1/d}}\right)\right)\right)\right] \\
&= 0
\end{aligned}$$

since μ is singular—see (2.5). □

Thus, finally, we consider absolutely continuous $\mathbf{X}$ with density f on $[0,1]^d$. (Note that in this case, by Hölder's inequality, $\int_{[0,1]^d} f^{1-2/d}(\mathbf{x})\mathrm{d}\mathbf{x} < \infty$ for $d \geq 2$.)

Theorem 2.3. *Let* $\mathbf{X}$ *have a density* f *on* $[0,1]^d$*. Then, for* $d > 2$*,*

$$n^{2/d}\mathbb{E}\|\mathbf{X}_{(1)}(\mathbf{X}) - \mathbf{X}\|^2 \to \frac{\Gamma\left(\frac{2}{d} + 1\right)}{V_d^{2/d}} \int_{[0,1]^d} f^{1-2/d}(\mathbf{x})\mathrm{d}\mathbf{x}.$$

For $d = 1$*,*

$$\liminf_{n\to\infty} n^2\mathbb{E}\|\mathbf{X}_{(1)}(\mathbf{X}) - \mathbf{X}\|^2 \geq \frac{1}{2}\int_{[0,1]^d; f>0} \frac{\mathrm{d}\mathbf{x}}{f(\mathbf{x})} > 0.$$

For $d = 2$, we have, for all densities f,

$$\frac{1}{\pi} + o(1) \le n\,\mathbb{E}\|\mathbf{X}_{(1)}(\mathbf{X}) - \mathbf{X}\|^2 < \frac{4}{\pi}(1+\sqrt{2})^2.$$

Remark 2.4. This theorem points out the special status of $\mathbb{R}^1$. On the real line, for absolutely continuous X, $|X_{(1)}(X) - X|^2$ converges at a rate above $1/n^2$. We recall that if X is singular, then $|X_{(1)}(X) - X|^2 = o_{\mathbb{P}}(1/n^2)$. Therefore, $|X_{(1)}(X) - X|^2$ can be used to distinguish between a purely singular distribution and an absolutely continuous distribution. □

Proof (Theorem 2.3). We will repeatedly make use of the gamma integral

$$\int_0^\infty e^{-\beta t^\alpha} \mathrm{d}t = \frac{\Gamma\left(\frac{1}{\alpha}+1\right)}{\beta^{1/\alpha}}, \quad \alpha, \beta > 0.$$

We have

$$\begin{aligned}
\mathbb{E}\|\mathbf{X}_{(1)}(\mathbf{X}) - \mathbf{X}\|^2 &= \int_0^\infty \mathbb{P}\left\{\|\mathbf{X}_{(1)}(\mathbf{X}) - \mathbf{X}\|^2 > t\right\} \mathrm{d}t \\
&= \int_0^\infty \mathbb{E}\left[\left(1 - \mu\left(B(\mathbf{X}, \sqrt{t})\right)\right)^n\right] \mathrm{d}t \\
&\qquad \text{(where } \mu \text{ is the distribution of } \mathbf{X}\text{)} \\
&= \frac{1}{n^{2/d}} \int_0^\infty \mathbb{E}\left[\left(1 - \mu\left(B\left(\mathbf{X}, \frac{\sqrt{t}}{n^{1/d}}\right)\right)\right)^n\right] \mathrm{d}t.
\end{aligned}$$

By the Lebesgue differentiation theorem (Theorem 20.18 in the Appendix), at Lebesgue-almost all $\mathbf{x}$, as $\rho \downarrow 0$,

$$\frac{\mu\left(B(\mathbf{x}, \rho)\right)}{V_d \rho^d} \to f(\mathbf{x}).$$

At such $\mathbf{x}$, therefore, we have for fixed t, as $n \to \infty$,

$$\left(1 - \mu\left(B\left(\mathbf{x}, \frac{\sqrt{t}}{n^{1/d}}\right)\right)\right)^n \to \exp(-f(\mathbf{x}) V_d t^{d/2}).$$

Fatou's lemma implies

$$\begin{aligned}
\liminf_{n\to\infty} n^{2/d}\mathbb{E}\|\mathbf{X}_{(1)}(\mathbf{X}) - \mathbf{X}\|^2 &\ge \int_{[0,1]^d} \int_0^\infty \exp(-f(\mathbf{x}) V_d t^{d/2}) f(\mathbf{x}) \mathrm{d}t \mathrm{d}\mathbf{x} \\
&= \Gamma\left(\frac{2}{d}+1\right) \int_{[0,1]^d; f>0} \frac{f(\mathbf{x})}{(f(\mathbf{x})V_d)^{2/d}} \mathrm{d}\mathbf{x} \\
&= \frac{\Gamma\left(\frac{2}{d}+1\right)}{V_d^{2/d}} \int_{[0,1]^d; f>0} f^{1-2/d}(\mathbf{x}) \mathrm{d}\mathbf{x}.
\end{aligned}$$

This shows the second assertion of the theorem, as well as the lower bound for the first and third statements. To establish the upper bound for $d > 2$, we take an arbitrary large constant R and note that

$$\begin{aligned} n^{2/d}\mathbb{E}\|\mathbf{X}_{(1)}(\mathbf{X}) - \mathbf{X}\|^2 &= n^{2/d}\mathbb{E}\left[\|\mathbf{X}_{(1)}(\mathbf{X}) - \mathbf{X}\|^2 \mathbb{1}_{[\|\mathbf{X}_{(1)}(\mathbf{X})-\mathbf{X}\| \le R/n^{1/d}]}\right] \\ &\quad + n^{2/d}\mathbb{E}\left[\|\mathbf{X}_{(1)}(\mathbf{X}) - \mathbf{X}\|^2 \mathbb{1}_{[\|\mathbf{X}_{(1)}(\mathbf{X})-\mathbf{X}\| > R/n^{1/d}]}\right] \\ &\stackrel{\text{def}}{=} \mathbf{I} + \mathbf{II}. \end{aligned}$$

By Fatou's lemma, we have

$$\begin{aligned} &\limsup_{n\to\infty} n^{2/d}\mathbb{E}\left[\|\mathbf{X}_{(1)}(\mathbf{X}) - \mathbf{X}\|^2 \mathbb{1}_{[\|\mathbf{X}_{(1)}(\mathbf{X})-\mathbf{X}\| \le R/n^{1/d}]}\right] \\ &\le \int_{[0,1]^d} \int_0^{R^2} \exp(-f(\mathbf{x})V_d t^{d/2}) \mathrm{d}t \mathrm{d}\mathbf{x} \\ &\le \frac{\Gamma\left(\frac{2}{d}+1\right)}{V_d^{2/d}} \int_{[0,1]^d} f^{1-2/d}(\mathbf{x}) \mathrm{d}\mathbf{x}. \end{aligned}$$

Next, $\mathbf{II}$ is small by choice of R. To see this, note again that, by (2.3),

$$\mathbf{II} = \frac{n^{2/d}}{n+1}\mathbb{E}\left[\sum_{i=1}^{n+1} \|\mathbf{X}_{(i,1)} - \mathbf{X}_i\|^2 \mathbb{1}_{[\|\mathbf{X}_{(i,1)}-\mathbf{X}_i\| > R/n^{1/d}]}\right].$$

Using R_i for $\|\mathbf{X}_{(i,1)} - \mathbf{X}_i\|$, we already observed in (2.4) that

$$\sum_{i=1}^{n+1} R_i^d \le a_d \stackrel{\text{def}}{=} \frac{2^d(1+\sqrt{d})^d}{V_d}.$$

Set $K = \sum_{i=1}^{n+1} \mathbb{1}_{[R_i > R/n^{1/d}]}$, and note that, necessarily, $K \le n a_d / R^d$. Then, by Jensen's inequality, whenever $K > 0$,

$$\left(\frac{1}{K}\sum_{i=1}^{n+1} R_i^2 \mathbb{1}_{[R_i > R/n^{1/d}]}\right)^{d/2} \le \frac{1}{K}\sum_{i=1}^{n+1} R_i^d \mathbb{1}_{[R_i > R/n^{1/d}]} \le \frac{a_d}{K}.$$

Thus,

$$\begin{aligned} \mathbf{II} &= \frac{n^{2/d}}{n+1}\mathbb{E}\left[\mathbb{1}_{[K>0]} \sum_{i=1}^{n+1} R_i^2 \mathbb{1}_{[R_i > R/n^{1/d}]}\right] \\ &\le \frac{n^{2/d}}{n}\mathbb{E}\left[K \times \left(\frac{a_d}{K}\right)^{2/d} \mathbb{1}_{[K>0]}\right] \end{aligned}$$

$$= a_d^{2/d} \mathbb{E}\left[\frac{K}{n}\right]^{1-2/d}$$
$$\leq \frac{a_d}{R^{d-2}},$$

which is as small as desired by choice of R when $d > 2$.

Finally, for $d = 2$, by (2.4),

$$\pi \sum_{i=1}^{n+1} R_i^2 \leq 4(1+\sqrt{2})^2.$$

Since

$$n\,\mathbb{E}\|\mathbf{X}_{(1)}(\mathbf{X}) - \mathbf{X}\|^2 = \frac{n}{n+1}\mathbb{E}\left[\sum_{i=1}^{n+1} R_i^2\right],$$

the upper bound follows. □

We conclude this section by providing an extension of Theorem 2.1 to the k-nearest neighbor ($1 \leq k \leq n$).

Theorem 2.4. *Let* $\mathbf{X}$ *takes values in* $[0,1]^d$. *Then, for* $d \geq 2$,

$$\mathbb{E}\|\mathbf{X}_{(k)}(\mathbf{X}) - \mathbf{X}\|^2 \leq c'_d \left(\frac{k}{n}\right)^{2/d},$$

where

$$c'_d = \frac{2^{3+\frac{2}{d}}(1+\sqrt{d})^2}{V_d^{2/d}}.$$

For $d = 1$, *we have*

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

Proof. The problem can be reduced to the nearest-neighbor inequality covered in Theorem 2.1. First, notice that $\mathbb{E}\|\mathbf{X}_{(k)}(\mathbf{X}) - \mathbf{X}\|^2 \leq d$. But, recalling that $V_d \leq 6$ for all $d \geq 1$, we have, for $2k > n$,

$$\inf_{d \geq 2} c'_d \left(\frac{k}{n}\right)^{2/d} \geq d$$

and $\frac{8k}{n} \geq 1$. Thus, the bounds are trivial for $2k > n$, so we assume that $2k \leq n$.

Partition the set $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ into $2k$ sets of sizes $n_1, \ldots, n_{2k}$, with

$$\sum_{j=1}^{2k} n_j = n \quad \text{and} \quad \left\lfloor \frac{n}{2k} \right\rfloor \le n_j \le \left\lfloor \frac{n}{2k} \right\rfloor + 1.$$

Let $\mathbf{X}^\star_{(1)}(j)$ be the nearest neighbor of $\mathbf{X}$ among all $\mathbf{X}_i$'s in the j-th group. Observe that, deterministically,

$$\|\mathbf{X}_{(k)}(\mathbf{X}) - \mathbf{X}\| \le \frac{1}{k} \sum_{j=1}^{2k} \|\mathbf{X}^\star_{(1)}(j) - \mathbf{X}\|$$

and, similarly,

$$\|\mathbf{X}_{(k)}(\mathbf{X}) - \mathbf{X}\|^2 \le \frac{1}{k} \sum_{j=1}^{2k} \|\mathbf{X}^\star_{(1)}(j) - \mathbf{X}\|^2,$$

because at least k of these nearest neighbors have values that are at least $\|\mathbf{X}_{(k)}(\mathbf{X}) - \mathbf{X}\|$. Apply Theorem 2.1 for $d \ge 2$ to obtain

$$\begin{aligned}
\mathbb{E}\|\mathbf{X}_{(k)}(\mathbf{X}) - \mathbf{X}\|^2 &\le \frac{1}{k} \sum_{j=1}^{2k} c_d \left(\frac{1}{n_j + 1} \right)^{2/d} \\
&\le \frac{1}{k} \sum_{j=1}^{2k} c_d \left(\frac{2k}{n} \right)^{2/d} \\
&= 2^{1+\frac{2}{d}} c_d \left(\frac{k}{n} \right)^{2/d}.
\end{aligned}$$

For $d = 1$, we argue similarly and get the upper bound

$$\frac{1}{k} \sum_{j=1}^{2k} \frac{2}{n_j + 1} \le \frac{1}{k} \sum_{j=1}^{2k} \frac{4k}{n} = \frac{8k}{n}. \quad \square$$

Chapter 3
The k-nearest neighbor density estimate

3.1 Nonparametric density estimation

A random vector $\mathbf{X}$ taking values in $\mathbb{R}^d$ has a (probability) density f with respect to the Lebesgue measure if, for all Borel sets $A \subseteq \mathbb{R}^d$, $\mathbb{P}\{\mathbf{X} \in A\} = \int_A f(\mathbf{x})\mathrm{d}\mathbf{x}$. In other words, if A is a small ball about $\mathbf{x}$, the probability that $\mathbf{X}$ falls in A is about $f(\mathbf{x})$ times the volume of A. It thus serves as a tool for computing probabilities of sets and, as a function that reveals the local concentration of probability mass, it may be used to visualize distributions of random variables.

The purpose of density estimation is to estimate an unknown density f from an i.i.d. sample drawn according to f. The view we take in this book is nonparametric, thereby assuming that f is largely unknown and that no assumptions can be made about its properties. Nonparametric estimation is particularly important when the common parametric forms—often unimodal—are suspect. For example, pattern recognition problems frequently involve densities that are multimodal and, even in the unimodal situation, there is rarely enough information available to warrant any parametric assumption. Thus, in the nonparametric context, a density estimate f_n is a Borel measurable function of $\mathbf{x}$ and the data $\mathbf{X}_1, \ldots, \mathbf{X}_n$:

$$f_n(\mathbf{x}) = f_n(\mathbf{x}; \mathbf{X}_1, \ldots, \mathbf{X}_n).$$

Often, but not always, f_n is itself a density in $\mathbf{x}$, i.e., it is Lebesgue-almost everywhere nonnegative and integrates to one. The choice of a density estimate is governed by a number of factors, like consistency (as the number of observations grows), smoothness, ease of computation, interpretability, and optimality for certain criteria.

The problem of nonparametric density estimation has a long and rich history, dating back to the pioneering works of Fix and Hodges (1951, 1952)—see also Fix and Hodges (1991a,b)—, Akaike (1954); Rosenblatt (1956); Whittle (1958); Parzen (1962); Watson and Leadbetter (1963), and Cacoullos (1966) in the late 50s and early 60s of the 20th century. The application scope is vast, as density estimates

G. Biau, L. Devroye, *Lectures on the Nearest Neighbor Method*,
Springer Series in the Data Sciences, DOI 10.1007/978-3-319-25388-6_3

are routinely employed across the entire and diverse range of applied statistics, including problems in exploratory data analysis, machine condition monitoring, pattern recognition, clustering, simulation, detection, medical diagnoses, financial investments, marketing, and econometrics. There are too many references to be included here, but the monographs by Rao (1983); Devroye and Györfi (1985); Silverman (1986); Devroye (1987); Scott (1992), and Tsybakov (2008) will provide the reader with introductions to the general subject area, both from a practical and theoretical perspective.

3.2 Distances between densities

The quality of a density estimate is measured by how well it performs the task at hand, estimating probabilities. In this respect, denoting by $\mathscr{B}$ the Borel sets of $\mathbb{R}^d$, the total variation criterion $d_{\mathrm{TV}}(f_n,f)$ is a natural distance:

$$d_{\mathrm{TV}}(f_n,f) = \sup_{A\in\mathscr{B}} \left| \int_A f_n(\mathbf{x})\mathrm{d}\mathbf{x} - \int_A f(\mathbf{x})\mathrm{d}\mathbf{x} \right|.$$

It should be noted that whenever f_n is a density, $0 \le d_{\mathrm{TV}}(f_n,f) \le 1$. When $d_{\mathrm{TV}}(f_n,f)$ is smaller than ε, then we know that for any Borel set A, the probability assigned to it by f differs at most by ε from the probability assigned to it by the estimate f_n. In other words, $d_{\mathrm{TV}}(f_n,f)$ is a practical easy-to-understand quantity.

Suppose now we look for an estimate f_n for which $d_{\mathrm{TV}}(f_n,f) \to 0$ in some probabilistic sense (e.g., almost surely, in probability, or in expectation) as $n \to \infty$. We see that this property will follow whenever

$$\int_{\mathbb{R}^d} |f_n(\mathbf{x}) - f(\mathbf{x})|\,\mathrm{d}\mathbf{x} \to 0$$

in the same probabilistic sense. Indeed,

$$d_{\mathrm{TV}}(f_n,f) \le \sup_{A\in\mathscr{B}} \int_A |f_n(\mathbf{x}) - f(\mathbf{x})|\,\mathrm{d}\mathbf{x} \le \int_{\mathbb{R}^d} |f_n(\mathbf{x}) - f(\mathbf{x})|\,\mathrm{d}\mathbf{x},$$

and, if f_n is itself a density,

$$d_{\mathrm{TV}}(f_n,f) = \frac{1}{2}\int_{\mathbb{R}^d} |f_n(\mathbf{x}) - f(\mathbf{x})|\,\mathrm{d}\mathbf{x}$$

(see, e.g., Devroye and Györfi, 1985). This shows the importance of $\int_{\mathbb{R}^d} |f_n(\mathbf{x}) - f(\mathbf{x})|\mathrm{d}\mathbf{x}$ for the study of the uniform convergence properties of the corresponding probability measures.

As might be expected, there are other possible global measures to assess the proximity between the estimate f_n and the target density f. First and foremost, there is the L^p distance

$$L^p(f_n,f) = \begin{cases} \left(\int_{\mathbb{R}^d} |f_n(\mathbf{x}) - f(\mathbf{x})|^p \, \mathrm{d}\mathbf{x} \right)^{1/p} & \text{for } 0 < p < \infty \\ \text{ess sup}_{\mathbf{x} \in \mathbb{R}^d} |f_n(\mathbf{x}) - f(\mathbf{x})| & \text{for } p = \infty, \end{cases}$$

where the essential supremum is with respect to the Lebesgue measure. The entropy-related Kullback-Leibler divergence is defined by

$$K(f,f_n) = \begin{cases} \displaystyle\int_{\mathbb{R}^d} f(\mathbf{x}) \log \left(\frac{f(\mathbf{x})}{f_n(\mathbf{x})} \right) \mathrm{d}\mathbf{x} & \text{if } f \ll f_n \\ \infty & \text{otherwise,} \end{cases}$$

where $f \ll f_n$ means that $\int_A f(\mathbf{x})\mathrm{d}\mathbf{x} = 0$ for every set A for which $\int_A f_n(\mathbf{x})\mathrm{d}\mathbf{x} = 0$.

For $p > 0$, the Hellinger distance takes the form

$$H^p(f_n,f) = \left(\int_{\mathbb{R}^d} \left| f_n^{1/p}(\mathbf{x}) - f^{1/p}(\mathbf{x}) \right|^p \mathrm{d}\mathbf{x} \right)^{1/p}.$$

Clearly, H^1 is the standard L^1 distance. For an account on the various properties and relationships between these and other global proximity measures, we refer to the introductory chapter of Devroye (1987).

The k-nearest neighbor density estimate that is discussed in this chapter has $\int_{\mathbb{R}^d} |f_n(\mathbf{x})| \, \mathrm{d}\mathbf{x} = \infty$ (Proposition 3.1 below), so that it is not suited for applications where one wants $\int_{\mathbb{R}^d} |f_n(\mathbf{x}) - f(\mathbf{x})| \mathrm{d}\mathbf{x}$ to converge to zero as n tends to infinity. Thus, for this or other reasons, we may also be interested in estimates f_n for which $f_n(\mathbf{x}) \to f(\mathbf{x})$ in some probabilistic sense at Lebesgue-almost all $\mathbf{x}$. We say that f_n is a weakly (strongly) pointwise consistent estimate of f on A if

$$f_n(\mathbf{x}) \to f(\mathbf{x}) \quad \text{in probability (almost surely) for all } \mathbf{x} \in A.$$

3.3 The k-nearest neighbor estimate

Our goal in this and the next few chapters is to investigate the properties of the k-nearest neighbor density estimate, which is defined below. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a sequence of i.i.d. random vectors taking values in $\mathbb{R}^d$, and assume that the common probability measure μ of the sequence is absolutely continuous with respect to the Lebesgue measure λ, with a density f.

By the Lebesgue differentiation theorem (Theorem 20.18 in the Appendix), we have, at λ-almost all $\mathbf{x}$,

$$f(\mathbf{x}) = \lim_{\rho \downarrow 0} \frac{\mu\left(B(\mathbf{x}, \rho)\right)}{\lambda\left(B(\mathbf{x}, \rho)\right)}.$$

In view of this relation, one can estimate $f(\mathbf{x})$ by the following method. Let k be an integer such that $1 \leq k \leq n$, let $R_{(k)}(\mathbf{x})$ be the distance from $\mathbf{x}$ to the k-th nearest neighbor in the data sequence, and let μ_n be the empirical distribution, i.e., for any Borel set $A \subseteq \mathbb{R}^d$,

$$\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[\mathbf{X}_i \in A]}.$$

The k-nearest neighbor density estimate is

$$f_n(\mathbf{x}) = \frac{\mu_n\left(B(\mathbf{x}, R_{(k)}(\mathbf{x}))\right)}{\lambda\left(B(\mathbf{x}, R_{(k)}(\mathbf{x}))\right)},$$

which, by construction, can also be written as follows:

$$f_n(\mathbf{x}) = \frac{k}{n\lambda\left(B(\mathbf{x}, R_{(k)}(\mathbf{x}))\right)}.$$

It was first introduced by Fix and Hodges (1951)—see also Fix and Hodges (1991a)—in the context of nonparametric discrimination, and further worked out by Loftsgaarden and Quesenberry (1965). The basic difference between this estimate and the kernel estimate of Akaike (1954); Rosenblatt (1956), and Parzen (1962) is that here a specific number of observations k is given and the distance to the k-th closest from $\mathbf{x}$ is measured. On the other hand, in the kernel approach, one counts the number of observations falling within a specified distance h from $\mathbf{x}$. Thus, the nearest neighbor and the kernel methods are somehow dual, the latter being equivalent to fix $R_{(k)}(\mathbf{x})$ and then determine k. A major practical advantage of the k-nearest neighbor estimate is that it is particularly easy to compute.

Let $\mathbf{X}_{(i)}(\mathbf{x})$ be the i-th nearest neighbor of $\mathbf{x}$ among $\mathbf{X}_1, \ldots, \mathbf{X}_n$. (Note that in this density context, distance ties happen with zero probability and may be broken arbitrarily.) Thus, $\|\mathbf{X}_{(1)}(\mathbf{x}) - \mathbf{x}\| \leq \cdots \leq \|\mathbf{X}_{(n)}(\mathbf{x}) - \mathbf{x}\|$. Denoting by V_d the volume of the unit ball in $(\mathbb{R}^d, \|\cdot\|)$, and observing that

$$\lambda\left(B(\mathbf{x}, R_{(k)}(\mathbf{x}))\right) = V_d \|\mathbf{X}_{(k)}(\mathbf{x}) - \mathbf{x}\|^d,$$

we have the following definition:

Definition 3.1. For $1 \leq k \leq n$, the k-nearest neighbor density estimate is defined by

$$f_n(\mathbf{x}) = \frac{k}{nV_d \|\mathbf{X}_{(k)}(\mathbf{x}) - \mathbf{x}\|^d}, \quad \mathbf{x} \in \mathbb{R}^d.$$

Adopting the convention $1/0 = \infty$, we note once and for all that, for fixed $\mathbf{x}$, $f_n(\mathbf{x})$ is positive and finite with probability one. Similarly, for fixed values of $\mathbf{X}_1, \ldots, \mathbf{X}_n$, $f_n(\mathbf{x})$ is positive and finite at λ-almost all $\mathbf{x}$. We also recall that in the d-dimensional Euclidean space,

$$V_d = \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)},$$

where $\Gamma(\cdot)$ is the gamma function.

3.4 First properties

The following proposition states some trivial properties of the k-nearest neighbor estimate.

Proposition 3.1. *Let f_n be the k-nearest neighbor density estimate. Then, for $0 \leq p \leq 1$,*

$$\int_{\mathbb{R}^d} f_n^p(\mathbf{x})\mathrm{d}\mathbf{x} = \infty.$$

On the other hand, for $p > 1$, with probability one,

$$\begin{cases} \displaystyle\int_{\mathbb{R}^d} f_n^p(\mathbf{x})\mathrm{d}\mathbf{x} = \infty & \text{if } k = 1 \\ \displaystyle\int_{\mathbb{R}^d} f_n^p(\mathbf{x})\mathrm{d}\mathbf{x} < \infty & \text{if } k > 1. \end{cases}$$

In addition, still with probability one,

$$\begin{cases} \operatorname{ess\,sup}_{\mathbf{x}\in\mathbb{R}^d} f_n(\mathbf{x}) = \infty & \text{if } k = 1 \\ \operatorname{ess\,sup}_{\mathbf{x}\in\mathbb{R}^d} f_n(\mathbf{x}) < \infty & \text{if } k > 1, \end{cases}$$

where the essential supremum is with respect to the Lebesgue measure.

Proof. Set

$$Z_n = \max_{1\leq i\leq n} \|\mathbf{X}_i\|.$$

Clearly, for fixed $\mathbf{x} \in \mathbb{R}^d$, by the triangle inequality,

$$\|\mathbf{X}_{(k)}(\mathbf{x}) - \mathbf{x}\| \leq \|\mathbf{x}\| + Z_n.$$

Thus,

$$\int_{\mathbb{R}^d} f_n^p(\mathbf{x})\mathrm{d}\mathbf{x} = \frac{k^p}{n^p V_d^p} \int_{\mathbb{R}^d} \frac{1}{\|\mathbf{X}_{(k)}(\mathbf{x}) - \mathbf{x}\|^{pd}}\,\mathrm{d}\mathbf{x}$$
$$\geq \frac{k^p}{n^p V_d^p} \int_{\mathbb{R}^d} \frac{1}{(\|\mathbf{x}\| + Z_n)^{pd}}\,\mathrm{d}\mathbf{x}.$$

Using a hyperspherical coordinate change of variables (see, e.g., Miller, 1964, Chapter 1), we obtain

$$\int_{\mathbb{R}^d} \frac{1}{(\|\mathbf{x}\| + Z_n)^{pd}}\,\mathrm{d}\mathbf{x} = \frac{2\pi^{d/2}}{\Gamma\left(\frac{d}{2}\right)} \int_0^\infty \frac{r^{d-1}}{(r + Z_n)^{pd}}\,\mathrm{d}r,$$

and the last integral is easily seen to be infinite whenever $0 \leq p \leq 1$.

Assume now that $p > 1$. Let

$$S_1 = \left\{\mathbf{x} \in \mathbb{R}^d : \mathbf{X}_{(1)}(\mathbf{x}) = \mathbf{X}_1\right\},$$

and observe that $\mathbf{X}_1$ belongs to the interior of S_1 with probability one. For $k = 1$, we have

$$f_n(\mathbf{x}) \geq \frac{1}{nV_d\|\mathbf{X}_1 - \mathbf{x}\|^d} \mathbb{1}_{[\mathbf{x}\in S_1]}.$$

Thus,

$$\int_{\mathbb{R}^d} f_n^p(\mathbf{x})\mathrm{d}\mathbf{x} \geq \frac{1}{n^p V_d^p} \int_{S_1} \frac{1}{\|\mathbf{X}_1 - \mathbf{x}\|^{pd}}\,\mathrm{d}\mathbf{x}$$
$$= \frac{1}{n^p V_d^p} \int_{S_1 - \mathbf{X}_1} \frac{1}{\|\mathbf{x}\|^{pd}}\,\mathrm{d}\mathbf{x},$$

where

$$S_1 - \mathbf{X}_1 = \{\mathbf{x} - \mathbf{X}_1 : \mathbf{x} \in S_1\}.$$

Using once again a hyperspherical coordinate change of variables, we see that the integral on the right-hand side is infinite for $p > 1$.

If $k > 1$, let, for $1 \leq i \leq n$,

$$S_i = \left\{\mathbf{x} \in \mathbb{R}^d : \mathbf{X}_{(k)}(\mathbf{x}) = \mathbf{X}_i\right\}.$$

It is easy to see that, on the event $E_n = [\mathbf{X}_j \neq \mathbf{X}_\ell, j \neq \ell]$, the sets $S_1, \ldots, S_n$ form a partition of $\mathbb{R}^d$. Thus, on the event E_n,

$$f_n(\mathbf{x}) = \frac{k}{nV_d} \sum_{i=1}^n \frac{1}{\|\mathbf{X}_i - \mathbf{x}\|^d} \mathbb{1}_{[\mathbf{x}\in S_i]},$$

and therefore,

$$\int_{\mathbb{R}^d} f_n^p(\mathbf{x})\mathrm{d}\mathbf{x} = \frac{k^p}{n^p V_d^p} \sum_{i=1}^{n} \int_{S_i} \frac{1}{\|\mathbf{X}_i - \mathbf{x}\|^{pd}}\,\mathrm{d}\mathbf{x}$$
$$= \frac{k^p}{n^p V_d^p} \sum_{i=1}^{n} \int_{S_i - \mathbf{X}_i} \frac{1}{\|\mathbf{x}\|^{pd}}\,\mathrm{d}\mathbf{x}.$$

On E_n, each $\mathbf{X}_i$ belongs to the interior of the complement of S_i and thus, by a hyperspherical coordinate change of variables, we obtain

$$\int_{S_i - \mathbf{X}_i} \frac{1}{\|\mathbf{x}\|^{pd}}\,\mathrm{d}\mathbf{x} < \infty.$$

The conclusion follows by observing that the event E_n has probability one. The assertion with the supremum uses similar arguments and is therefore omitted. □

Proposition 3.1 implies that it is impossible to study the k-nearest neighbor estimate properties in L^1 or, equivalently, that the k-nearest neighbor estimate is not attractive for applications where one wants $\int_{\mathbb{R}^d} |f_n(\mathbf{x}) - f(\mathbf{x})|\mathrm{d}\mathbf{x}$ to converge to zero in some probabilistic sense as n tends to infinity. Thus, in the remainder of the chapter, we focus on pointwise consistency properties of the estimate f_n.

3.5 Weak and strong pointwise consistency

Theorem 3.1 below summarizes results obtained by Loftsgaarden and Quesenberry (1965); Wagner (1973), and Moore and Yackel (1977a). However, our proof approach is different and relies on the results of Chapter 1 on uniform order statistics. It is stressed that it holds whenever $\mathbf{x}$ is a Lebesgue point of f, that is, an $\mathbf{x}$ for which

$$\lim_{\rho \downarrow 0} \frac{\mu\left(B(\mathbf{x}, \rho)\right)}{\lambda\left(B(\mathbf{x}, \rho)\right)} = \lim_{\rho \downarrow 0} \frac{\displaystyle\int_{B(\mathbf{x},\rho)} f(\mathbf{y})\mathrm{d}\mathbf{y}}{\displaystyle\int_{B(\mathbf{x},\rho)} \mathrm{d}\mathbf{y}} = f(\mathbf{x}).$$

As f is a density, we know that λ-almost all $\mathbf{x}$ satisfy this property (see Theorem 20.18 in the Appendix).

Theorem 3.1 (Pointwise consistency). *Let f_n be the k-nearest neighbor density estimate. If $k/n \to 0$, then, for λ-almost all $\mathbf{x} \in \mathbb{R}^d$,*

(i) $f_n(\mathbf{x}) \to f(\mathbf{x})$ in probability when $k \to \infty$;
(ii) $f_n(\mathbf{x}) \to f(\mathbf{x})$ almost surely when $k/\log n \to \infty$;

(*iii*) $f_n(\mathbf{x}) \to f(\mathbf{x})$ *almost surely when* $k/\log\log n \to \infty$ *and there exists an increasing function* $\psi \geq 1$ *with* $\lim_{\delta \downarrow 0} \psi(\delta) = 1$, *such that*

$$1 \leq \frac{k_{n+m}}{k_n} \leq \psi\left(\frac{m}{n}\right), \quad n, m \geq 1. \tag{3.1}$$

Proof. Set $R_{(k)}(\mathbf{x}) = \|\mathbf{X}_{(k)}(\mathbf{x}) - \mathbf{x}\|$. Recall (Chapter 1) that

$$\mu\left(B\left(\mathbf{x}, R_{(k)}(\mathbf{x})\right)\right) \stackrel{\mathscr{D}}{=} U_{(k)},$$

where $U_{(k)}$ is the k-th order statistic for a uniform sample. The k-nearest neighbor estimate can be written as

$$\begin{aligned} f_n(\mathbf{x}) &= \frac{k/n}{\lambda\left(B\left(\mathbf{x}, R_{(k)}(\mathbf{x})\right)\right)} \\ &= \frac{\mu\left(B\left(\mathbf{x}, R_{(k)}(\mathbf{x})\right)\right)}{\lambda\left(B\left(\mathbf{x}, R_{(k)}(\mathbf{x})\right)\right)} \times \frac{k/n}{\mu\left(B\left(\mathbf{x}, R_{(k)}(\mathbf{x})\right)\right)} \\ &\stackrel{\text{def}}{=} g_n(\mathbf{x}) \times h_n(\mathbf{x}). \end{aligned}$$

Assume that $\mathbf{x}$ is a Lebesgue point of f. When $\mathbf{x}$ is not in the support of μ, then $f(\mathbf{x}) = 0$ (since $\mathbf{x}$ is a Lebesgue point) and there exists $\rho_{\mathbf{x}} > 0$ such that $\|\mathbf{X}_{(k)}(\mathbf{x}) - \mathbf{x}\| \geq \rho_{\mathbf{x}}$ with probability one. Therefore, in this case,

$$f_n(\mathbf{x}) \leq \frac{k/n}{V_d \rho_{\mathbf{x}}^d} \to 0 \quad \text{as } k/n \to 0.$$

So, assume that $\mathbf{x}$ is in the support of μ. We note that, for fixed $\varepsilon > 0$,

$$|g_n(\mathbf{x}) - f(\mathbf{x})| \leq \varepsilon$$

if $R_{(k)}(\mathbf{x}) \leq \rho(\mathbf{x}, \varepsilon)$, for some positive $\rho(\mathbf{x}, \varepsilon)$. Thus,

$$\begin{aligned} |f_n(\mathbf{x}) - f(\mathbf{x})| &\leq |g_n(\mathbf{x}) - f(\mathbf{x})|\, h_n(\mathbf{x}) + f(\mathbf{x})|h_n(\mathbf{x}) - 1| \\ &\leq \varepsilon\, h_n(\mathbf{x}) + \infty.\mathbb{1}_{[R_{(k)}(\mathbf{x}) > \rho(\mathbf{x}, \varepsilon)]} + f(\mathbf{x})|h_n(\mathbf{x}) - 1| \\ &\leq (\varepsilon + f(\mathbf{x}))\, |h_n(\mathbf{x}) - 1| + \varepsilon + \infty.\mathbb{1}_{[R_{(k)}(\mathbf{x}) > \rho(\mathbf{x}, \varepsilon)]}. \end{aligned}$$

According to Theorem 1.4, the first term on the right-hand side tends to zero in probability if $f(\mathbf{x}) < \infty$ and $k \to \infty$. By Theorem 1.5, it tends to zero almost surely if $f(\mathbf{x}) < \infty$ and $k/\log n \to \infty$. Besides, it tends to zero almost surely if $f(\mathbf{x}) < \infty$, $k/\log\log n \to \infty$ and k satisfies (3.1), by Theorem 1.6. The second term is small by choice of ε. The third term tends to zero almost surely if $\mathbf{x}$ is in the support of μ (which we assume) by Lemma 2.2. The proof is finished by noting that λ-almost all $\mathbf{x}$ have $f(\mathbf{x}) < \infty$ and are Lebesgue points of f. □

Chapter 4
Uniform consistency

4.1 Bounded densities

This chapter is devoted to the study of the uniform consistency properties of the k-nearest neighbor density estimate f_n. Before embarking on the supremum norm convergence, it is useful to understand the behavior of f_n on bounded densities. We denote the essential supremum (with respect to the Lebesgue measure λ) of the density f by

$$\|f\|_\infty = \inf\left\{t \geq 0 : \lambda\left\{|f| > t\right\} = 0\right\}.$$

Theorem 4.1. *Assume that $\|f\|_\infty < \infty$. If $k/\log n \to \infty$, then the k-nearest neighbor density estimate f_n satisfies*

$$\sum_{n\geq 1} \mathbb{P}\left\{\sup_{\mathbf{x}\in\mathbb{R}^d} f_n(\mathbf{x}) > 2^{d+1}\|f\|_\infty\right\} < \infty.$$

In particular, with probability one, for all n large enough,

$$\sup_{\mathbf{x}\in\mathbb{R}^d} f_n(\mathbf{x}) \leq 2^{d+1}\|f\|_\infty.$$

Also, with probability one, for all n large enough,

$$\inf_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{X}_{(k)}(\mathbf{x}) - \mathbf{x}\| \geq \frac{1}{2}\left(\frac{1}{2nV_d\|f\|_\infty}\right)^{1/d}.$$

Before proving the theorem, we observe the following:

G. Biau, L. Devroye, *Lectures on the Nearest Neighbor Method*,
Springer Series in the Data Sciences, DOI 10.1007/978-3-319-25388-6_4

Lemma 4.1. *If φ is a nonnegative convex increasing function, $\varphi(0) = 0$, and $x_1, x_2, \ldots$ are nonnegative real numbers with $x_i \le x$ for some $x > 0$, then*

$$\sum_{i=1}^{\infty} \varphi(x_i) \le \frac{\varphi(x)}{x} \sum_{i=1}^{\infty} x_i.$$

Proof. By convexity, note that $\varphi(x_i) \le \varphi(x) x_i / x$. □

Proof (Theorem 4.1). The proof uses an infinite grid. Given a radius ρ, we place the centers of the grid at

$$\mathscr{G} = \frac{\rho}{\sqrt{d}} \mathbb{Z}^d,$$

where $\mathbb{Z}^d$ is the space of all integer-valued vectors in $\mathbb{R}^d$, and $\rho/\sqrt{d}$ is a scale factor. For $\mathbf{y} \in \mathscr{G}$, define $p_{\mathbf{y}} = \mu(B(\mathbf{y}, 2\rho))$, where μ is the distribution associated with f. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be n i.i.d. points drawn from μ on $\mathbb{R}^d$, and let

$$M_n = \max_{\mathbf{x} \in \mathbb{R}^d} |B(\mathbf{x}, \rho)|,$$

where

$$|B(\mathbf{x}, \rho)| \stackrel{\text{def}}{=} \sum_{i=1}^{n} \mathbb{1}_{[\mathbf{X}_i \in B(\mathbf{x}, \rho)]}.$$

For each $\mathbf{x} \in \mathbb{R}^d$, there exists $\mathbf{y} \in \mathscr{G}$ with $\|\mathbf{y} - \mathbf{x}\| \le \rho$. Take such a $\mathbf{y}$. Then

$$|B(\mathbf{x}, \rho)| \le |B(\mathbf{y}, 2\rho)|,$$

since $B(\mathbf{x}, \rho) \subseteq B(\mathbf{y}, 2\rho)$. Therefore,

$$M_n \le \max_{\mathbf{y} \in \mathscr{G}} |B(\mathbf{y}, 2\rho)|.$$

Note also that

$$\max_{\mathbf{y} \in \mathscr{G}} p_{\mathbf{y}} \le V_d (2\rho)^d \|f\|_\infty.$$

Finally, $\sum_{\mathbf{y} \in \mathscr{G}} p_{\mathbf{y}}$ is bounded from above by the maximal overlap among the balls $B(\mathbf{y}, 2\rho)$, $\mathbf{y} \in \mathscr{G}$. But a square cell of dimensions $4\rho \times \cdots \times 4\rho$ contains at most $(4\sqrt{d} + 1)^d$ points from $\mathscr{G}$, and thus, the overlap is bounded by that number too. Therefore,

$$\sum_{\mathbf{y} \in \mathscr{G}} p_{\mathbf{y}} \le (4\sqrt{d} + 1)^d.$$

Consider the function $\varphi(u) = e^{-u}u^k$. It is easy to verify that φ is convex on $[0, k - \sqrt{k}]$ and that $\varphi(0) = 0$. Thus, if

$$nV_d(2\rho)^d\|f\|_\infty \le \frac{k}{2} \le k - \sqrt{k}$$

(this is possible for all n large enough since, by our assumption, $k \to \infty$ as $n \to \infty$) we have, by Lemma 4.1,

$$\begin{aligned}\sum_{\mathbf{y}\in\mathscr{G}} e^{-np_\mathbf{y}}(np_\mathbf{y})^k &\le e^{-k/2}\left(\frac{k}{2}\right)^{k-1}\sum_{\mathbf{y}\in\mathscr{G}} np_\mathbf{y}\\ &\le e^{-k/2}\left(\frac{k}{2}\right)^{k-1}(4\sqrt{d}+1)^d n.\end{aligned}$$

With the preliminaries out of the way, we note that

$$\inf_{\mathbf{x}\in\mathbb{R}^d}\|\mathbf{X}_{(k)}(\mathbf{x}) - \mathbf{x}\| < \rho$$

implies that

$$\max_{\mathbf{x}\in\mathbb{R}^d}|B(\mathbf{x},\rho)| \ge k,$$

and thus that

$$\max_{\mathbf{y}\in\mathscr{G}}|B(\mathbf{y},2\rho)| \ge k.$$

Now, choose ρ such that

$$nV_d(2\rho)^d\|f\|_\infty = \frac{k}{2},$$

which implies that $\max_{\mathbf{y}\in\mathscr{G}} np_\mathbf{y} \le k/2$. Then

$$\begin{aligned}&\mathbb{P}\left\{\inf_{\mathbf{x}\in\mathbb{R}^d}\|\mathbf{X}_{(k)}(\mathbf{x}) - \mathbf{x}\| < \rho\right\}\\ &\quad\le \sum_{\mathbf{y}\in\mathscr{G}}\mathbb{P}\{|B(\mathbf{y},2\rho)| \ge k\}\\ &\quad\le \sum_{\mathbf{y}\in\mathscr{G}:p_\mathbf{y}>0}\mathbb{P}\left\{\mathrm{Bin}(n,p_\mathbf{y}) \ge k\right\}\\ &\quad\le \sum_{\mathbf{y}\in\mathscr{G}:p_\mathbf{y}>0}\exp\left(k - np_\mathbf{y} - k\log\left(\frac{k}{np_\mathbf{y}}\right)\right)\end{aligned}$$

(by Chernoff's bound—see Theorem 20.5 in the Appendix).

It follows that

$$\begin{aligned}\mathbb{P}\left\{\inf_{\mathbf{x}\in\mathbb{R}^d}\|\mathbf{X}_{(k)}(\mathbf{x})-\mathbf{x}\|<\rho\right\} &\le \left(\frac{e}{k}\right)^k \sum_{\mathbf{y}\in\mathscr{G}} e^{-np_{\mathbf{y}}}(np_{\mathbf{y}})^k \\ &\le (4\sqrt{d}+1)^d n\left(\frac{e}{k}\right)^k e^{-k/2}\left(\frac{k}{2}\right)^{k-1} \\ &= (4\sqrt{d}+1)^d \frac{2n}{k}\left(\frac{\sqrt{e}}{2}\right)^k.\end{aligned}$$

Since $k/\log n \to \infty$, we have

$$\sum_{n\ge 1} n\left(\frac{\sqrt{e}}{2}\right)^k < \infty.$$

Therefore, by the Borel-Cantelli lemma, with probability one, for all n large enough,

$$\inf_{\mathbf{x}\in\mathbb{R}^d}\|\mathbf{X}_{(k)}(\mathbf{x})-\mathbf{x}\| \ge \rho = \frac{1}{2}\left(\frac{k}{2nV_d\|f\|_\infty}\right)^{1/d}.$$

This shows the second statement of the theorem. The first one follows by observing that

$$\mathbb{P}\left\{\sup_{\mathbf{x}\in\mathbb{R}^d} f_n(\mathbf{x}) > 2^{d+1}\|f\|_\infty\right\} = \mathbb{P}\left\{\inf_{\mathbf{x}\in\mathbb{R}^d}\|\mathbf{X}_{(k)}(\mathbf{x})-\mathbf{x}\|<\rho\right\}. \quad \square$$

Remark 4.1. We leave it as a good exercise to show that the coefficient 2^{d+1} in the theorem can be replaced by $(1+\varepsilon)$ for any $\varepsilon > 0$. $\square$

4.2 Uniformly continuous densities

Since a continuity point of f is also a Lebesgue point, the proof of Theorem 3.1 reveals that $|f_n(\mathbf{x}) - f(\mathbf{x})| \to 0$ almost surely on the continuity set of f, provided $k/\log n \to \infty$ and $k/n \to 0$. Thus, if f is uniformly continuous, one is tempted to believe that $\sup_{\mathbf{x}\in\mathbb{R}^d}|f_n(\mathbf{x}) - f(\mathbf{x})| \to 0$ almost surely, under the same conditions on k. For $d = 1$, Moore and Henrichon (1969) showed that

$$\sup_{\mathbf{x}\in\mathbb{R}}|f_n(\mathbf{x}) - f(\mathbf{x})| \to 0 \quad \text{in probability}$$

if f is uniformly continuous and positive on $\mathbb{R}$, if $k/\log n \to \infty$ and if, additionally, $k/n \to 0$. Kim and Van Ryzin (1975), also for $d = 1$, proved the same result

for a slightly different type of estimate under essentially the same conditions. In the remainder of the section we prove the following theorem, due to Devroye and Wagner (1977), yet with a different approach.

Theorem 4.2 (Strong uniform consistency). *Assume that f is uniformly continuous. If $k/\log n \to \infty$ and $k/n \to 0$, then the k-nearest neighbor density estimate f_n is strongly uniformly consistent, that is,*

$$\sup_{\mathbf{x}\in\mathbb{R}^d} |f_n(\mathbf{x}) - f(\mathbf{x})| \to 0 \quad \textit{almost surely}.$$

Notice that $\sup_{\mathbf{x}\in\mathbb{R}^d} |f_n(\mathbf{x}) - f(\mathbf{x})|$ is indeed a random variable if f is continuous, since it is possible to replace the supremum over $\mathbb{R}^d$ by the supremum over a countable dense subset of $\mathbb{R}^d$ in view of the continuity of f and the shape of f_n, which is piecewise constant.

The proof of Theorem 4.2 begins with a lemma. As a prerequisite, we leave it as an exercise to show that the uniformly continuous density f is bounded and vanishes as $\|\mathbf{x}\|$ tends to infinity.

Lemma 4.2. *Assume that f is uniformly continuous and that $\varepsilon > 0$ is fixed. If $k \to \infty$ and $k/n \to 0$, then there exist a positive integer n_0 and $\alpha > 0$ such that*

$$\sup_{\mathbf{x}\in\mathbb{R}^d} \mathbb{P}\{|f_n(\mathbf{x}) - f(\mathbf{x})| \geq \varepsilon\} \leq e^{-\alpha k} \quad \textit{for all } n \geq n_0.$$

Proof. We introduce the modulus of continuity $\omega(t)$ of f:

$$\omega(t) = \sup_{\|\mathbf{y}-\mathbf{x}\|\leq t} |f(\mathbf{y}) - f(\mathbf{x})|,$$

and note that $w(t) \downarrow 0$ as $t \downarrow 0$. If $B(\mathbf{x}, \rho)$ is the closed ball centered at $\mathbf{x}$ of radius ρ, then

$$(f(\mathbf{x}) - \omega(\rho))\, V_d \rho^d \leq \int_{B(\mathbf{x},\rho)} f(\mathbf{y})\mathrm{d}\mathbf{y} \leq (f(\mathbf{x}) + \omega(\rho))\, V_d \rho^d.$$

Recall that

$$f_n(\mathbf{x}) = \frac{k}{n V_d R_{(k)}^d(\mathbf{x})},$$

where

$$R_{(k)}(\mathbf{x}) = \|\mathbf{X}_{(k)}(\mathbf{x}) - \mathbf{x}\|.$$

We have, for fixed $\varepsilon > 0$,

$$\mathbb{P}\{f_n(\mathbf{x}) \geq f(\mathbf{x}) + \varepsilon\} = \mathbb{P}\left\{R_{(k)}(\mathbf{x}) \leq t\right\},$$

with

$$t = \left(\frac{k}{nV_d(f(\mathbf{x}) + \varepsilon)} \right)^{1/d}.$$

Thus,

$$\begin{aligned}
\mathbb{P}\left\{f_n(\mathbf{x}) \geq f(\mathbf{x}) + \varepsilon\right\} &= \mathbb{P}\left\{\mathrm{Bin}\left(n, \textstyle\int_{B(\mathbf{x},t)} f(\mathbf{y})\mathrm{d}\mathbf{y}\right) \geq k\right\} \\
&\leq \mathbb{P}\left\{\mathrm{Bin}\left(n, \left(f(\mathbf{x}) + \omega(t)\right) V_d t^d\right) \geq k\right\} \\
&= \mathbb{P}\left\{\mathrm{Bin}\left(n, \tfrac{f(\mathbf{x})+\omega(t)}{f(\mathbf{x})+\varepsilon} \times \tfrac{k}{n}\right) \geq k\right\} \\
&\leq \mathbb{P}\left\{\mathrm{Bin}\left(n, \tfrac{f(\mathbf{x})+\omega\left(\left(\frac{k}{nV_d\varepsilon}\right)^{1/d}\right)}{f(\mathbf{x})+\varepsilon} \times \tfrac{k}{n}\right) \geq k\right\}.
\end{aligned}$$

Let n_0 be so large that $\omega\left(\left(\frac{k}{nV_d\varepsilon}\right)^{1/d}\right) \leq \frac{\varepsilon}{2}$ for all $n \geq n_0$. Then we have a further upper bound of

$$\begin{aligned}
\mathbb{P}\left\{\mathrm{Bin}\left(n, \tfrac{f(\mathbf{x})+\varepsilon/2}{f(\mathbf{x})+\varepsilon} \times \tfrac{k}{n}\right) \geq k\right\} &\leq \mathbb{P}\left\{\mathrm{Bin}\left(n, \tfrac{\|f\|_\infty+\varepsilon/2}{\|f\|_\infty+\varepsilon} \times \tfrac{k}{n}\right) \geq k\right\} \\
&\leq e^{-\beta_1 k}
\end{aligned}$$

by Chernoff's bound (Theorem 20.5), for β_1 depending upon $\|f\|_\infty$ and ε only.

Similarly, for $f(\mathbf{x}) > \varepsilon$,

$$\mathbb{P}\left\{f_n(\mathbf{x}) \leq f(\mathbf{x}) - \varepsilon\right\} = \mathbb{P}\left\{R_{(k)}(\mathbf{x}) \geq t\right\} \tag{4.1}$$

where now $t = \left(\frac{k}{nV_d(f(\mathbf{x})-\varepsilon)}\right)^{1/d}$. Expression (4.1) is bounded from above by

$$\mathbb{P}\left\{\mathrm{Bin}\left(n, \int_{B(\mathbf{x},t)} f(\mathbf{y})\mathrm{d}\mathbf{y}\right) \leq k\right\}.$$

Define $t^\star = \left(\frac{k}{nV_d\varepsilon}\right)^{1/d}$. Let n_0 be so large that $\omega(t^\star) \leq \frac{\varepsilon}{2}$ for all $n \geq n_0$. For $f(\mathbf{x}) \geq 2\varepsilon$, we have $t \leq t^\star$, and so

$$\begin{aligned}
\mathbb{P}\left\{\mathrm{Bin}\left(n, \textstyle\int_{B(\mathbf{x},t)} f(\mathbf{y})\mathrm{d}\mathbf{y}\right) \leq k\right\} &\leq \mathbb{P}\left\{\mathrm{Bin}\left(n, \tfrac{f(\mathbf{x})-\omega(t^\star)}{f(\mathbf{x})-\varepsilon} \times \tfrac{k}{n}\right) \leq k\right\} \\
&\leq \mathbb{P}\left\{\mathrm{Bin}\left(n, \tfrac{f(\mathbf{x})-\varepsilon/2}{f(\mathbf{x})-\varepsilon} \times \tfrac{k}{n}\right) \leq k\right\} \\
&\leq \mathbb{P}\left\{\mathrm{Bin}\left(n, \tfrac{\|f\|_\infty-\varepsilon/2}{\|f\|_\infty-\varepsilon} \times \tfrac{k}{n}\right) \leq k\right\} \\
&\leq e^{-\beta_2 k}
\end{aligned}$$

by Chernoff's bound, where β_2 depends upon $\|f\|_\infty$ and ε.

For $\varepsilon < f(\mathbf{x}) \leq 2\varepsilon$, we have

$$\mathbb{P}\{f(\mathbf{x}) - f_n(\mathbf{x}) \geq 2\varepsilon\} = 0$$

and thus, combining both results,

$$\sup_{\mathbf{x}\in\mathbb{R}^d} \mathbb{P}\{f(\mathbf{x}) - f_n(\mathbf{x}) \geq 2\varepsilon\} \leq e^{-\beta_2 k}$$

for all $n \geq n_0$. □

Proof (Theorem 4.2). According to Lemma 4.2, for every $\varepsilon > 0$, there exist n_0 and $\alpha > 0$ such that

$$\sup_{\mathbf{x}\in\mathbb{R}^d} \mathbb{P}\{|f_n(\mathbf{x}) - f(\mathbf{x})| > \varepsilon\} \leq e^{-\alpha k} \quad \text{for all } n \geq n_0.$$

This is quite powerful, but the supremum is, unfortunately, outside the probability. If $G = \{\mathbf{x}_1, \ldots, \mathbf{x}_{n^d}\}$ is a fixed set of n^d points, then the union bound implies

$$\mathbb{P}\left\{\max_{1\leq i\leq n} |f_n(\mathbf{x}_i) - f(\mathbf{x}_i)| > \varepsilon\right\} \leq n^d e^{-\alpha k} \quad \text{for all } n \geq n_0. \tag{4.2}$$

This is summable in n when $k/\log n \to \infty$. The proof is completed by relating our supremum to this discrete maximum.

Using the uniform continuity of f, we first find a constant $a > 0$ (and set $A = [-a, a]^d$, $A^\star = [-a-1, a+1]^d$) such that

$$\sup_{\mathbf{x}\notin A} f(\mathbf{x}) \leq \frac{\varepsilon}{2^{d+1}} \quad \text{and} \quad \frac{1}{(2a)^d} \leq \frac{\varepsilon}{2^{d+1}}.$$

Let

$$g(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & \text{if } \mathbf{x} \notin A \\ \frac{1}{(2a)^d} \displaystyle\int_A f(\mathbf{y})\mathrm{d}\mathbf{y} & \text{if } \mathbf{x} \in A. \end{cases}$$

Note that $\|g\|_\infty \leq \varepsilon/2^{d+1}$, by construction. Also, if $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d., drawn from f, and $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are i.i.d., drawn from the uniform law on A, then the following sample $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ is an i.i.d. sample drawn from g:

$$\mathbf{Y}_i = \begin{cases} \mathbf{X}_i & \text{if } \mathbf{X}_i \notin A \\ \mathbf{Z}_i & \text{if } \mathbf{X}_i \in A. \end{cases}$$

The verification is left to the reader. Note that

$$\sup_{\mathbf{x}\notin A^\star} (f(\mathbf{x}) - f_n(\mathbf{x})) \leq \sup_{\mathbf{x}\notin A^\star} f(\mathbf{x}) \leq \frac{\varepsilon}{2^{d+1}},$$

and

$$\sup_{\mathbf{x}\notin A^\star}(f_n(\mathbf{x})-f(\mathbf{x})) \le \sup_{\mathbf{x}\notin A^\star} f_n(\mathbf{x})$$
$$\le \max\left(\sup_{\mathbf{x}\notin A^\star: R_{(k)}(\mathbf{x})\ge 1} f_n(\mathbf{x}), \sup_{\mathbf{x}\notin A^\star: R_{(k)}(\mathbf{x})<1} f_n(\mathbf{x})\right),$$

where

$$R_{(k)}(\mathbf{x}) = \|\mathbf{X}_{(k)}(\mathbf{x})-\mathbf{x}\|.$$

Therefore,

$$\sup_{\mathbf{x}\notin A^\star}(f_n(\mathbf{x})-f(\mathbf{x})) \le \max\left(\frac{k}{nV_d}, \sup_{\mathbf{x}\notin A^\star} g_n(\mathbf{x})\right),$$

where $g_n(\mathbf{x})$ is the k-nearest neighbor estimate for $\mathbf{Y}_1,\dots,\mathbf{Y}_n$ (since, for $\mathbf{x}\notin A^\star$, $f_n(\mathbf{x}) = g_n(\mathbf{x})$ if $R_{(k)}(\mathbf{x}) < 1$). By Theorem 4.1 for bounded densities, it follows that, with probability one and for all n large enough,

$$\sup_{\mathbf{x}\notin A^\star}(f_n(\mathbf{x})-f(\mathbf{x})) \le \max\left(\frac{k}{nV_d}, 2^{d+1}\|g\|_\infty\right) \le \max\left(\frac{k}{nV_d}, \varepsilon\right).$$

Since $k/n\to 0$, we see that with probability one, for all n large enough,

$$\sup_{\mathbf{x}\notin A^\star}|f_n(\mathbf{x})-f(\mathbf{x})| \le \varepsilon.$$

Next, partition $A^\star = [-a-1, a+1]^d$ into n^d equal squares of volume $(\frac{2a+2}{n})^d$ each. Denote these squares by C_i, $1\le i\le n^d$, and let $G = \{\mathbf{x}_1,\dots,\mathbf{x}_{n^d}\}$ be the collection of their centers. We recall that, by inequality (4.2), with probability one,

$$\max_{1\le i\le n^d}|f_n(\mathbf{x}_i)-f(\mathbf{x}_i)| \le \varepsilon$$

for all n large enough. Define

$$\Delta_1 = \max_{1\le i\le n^d}\sup_{\mathbf{x}\in C_i}|f_n(\mathbf{x})-f_n(\mathbf{x}_i)| \quad\text{and}\quad \Delta_2 = \max_{1\le i\le n^d}\sup_{\mathbf{x}\in C_i}|f(\mathbf{x})-f(\mathbf{x}_i)|.$$

By the triangle inequality, we have with probability one and for all n large enough,

$$\sup_{\mathbf{x}\in A^\star}|f_n(\mathbf{x})-f(\mathbf{x})| \le \varepsilon+\Delta_1+\Delta_2.$$

Since $(2a+2)/n \to 0$, it is clear that $\Delta_2 \to 0$ by uniform continuity. Define

$$R_i = \inf_{\mathbf{x} \in C_i} R_{(k)}(\mathbf{x}),$$

and recall from Theorem 4.1 that with probability one, for all n large enough,

$$R_i \geq \rho \stackrel{\text{def}}{=} \frac{1}{2} \left(\frac{k}{2nV_d \|f\|_\infty} \right)^{1/d}.$$

Note that for any k, $\mathbf{x}$, $\mathbf{y}$, we have

$$\left| R_{(k)}(\mathbf{x}) - R_{(k)}(\mathbf{y}) \right| \leq \|\mathbf{x} - \mathbf{y}\|.$$

Thus,

$$\begin{aligned} \sup_{\mathbf{x},\mathbf{y} \in C_i} \left| R_{(k)}(\mathbf{x}) - R_{(k)}(\mathbf{y}) \right| &\leq \sup_{\mathbf{x},\mathbf{y} \in C_i} \|\mathbf{x} - \mathbf{y}\| \\ &= \text{diameter}(C_i) \\ &= \frac{2(a+1)\sqrt{d}}{n} \\ &\stackrel{\text{def}}{=} \gamma. \end{aligned}$$

In particular,

$$\sup_{\mathbf{x} \in C_i} R_{(k)}(\mathbf{x}) \leq \sup_{\mathbf{x} \in C_i} |R_{(k)}(\mathbf{x}) - R_i| + R_i \leq \gamma + R_i.$$

Also, for $\mathbf{x} \in C_i$,

$$|f_n(\mathbf{x}) - f_n(\mathbf{x}_i)| = \frac{k}{nV_d} \left(\frac{1}{\min^d \left(R_{(k)}(\mathbf{x}), R_{(k)}(\mathbf{x}_i) \right)} - \frac{1}{\max^d \left(R_{(k)}(\mathbf{x}), R_{(k)}(\mathbf{x}_i) \right)} \right),$$

so that, with probability one, for all n large enough,

$$\begin{aligned} \Delta_1 &\leq \frac{k}{nV_d} \max_{1 \leq i \leq n^d} \left(\frac{1}{R_i^d} - \frac{1}{(\gamma + R_i)^d} \right) \\ &\leq \frac{k}{nV_d \rho^d} \left(1 - \frac{1}{(1 + \gamma/\rho)^d} \right) \\ &\qquad \text{(since } R_i \geq \rho \text{ with probability one, for all } n \text{ large enough)} \\ &= \frac{k}{nV_d \rho^d} \times \text{o}(1) \end{aligned}$$

$$
\begin{aligned}
&\text{(since } \gamma/\rho \to 0\text{)} \\
&= 2^{d+1}\|f\|_\infty \times \mathrm{o}(1) \\
&\to 0 \quad \text{as } n \to \infty.
\end{aligned}
$$

This completes the proof. □

Remark 4.2. For $d = 1$, using empirical process techniques, Mack (1983) provides almost sure rates of convergence for the quantity $\sup_{\mathbf{x}\in J} |f_n(\mathbf{x}) - f(\mathbf{x})|$, where J is some suitably chosen interval. □

Chapter 5
Weighted k-nearest neighbor density estimates

5.1 Linear combinations

There are different ways to weigh or smooth the k-nearest neighbor density estimate. Some key ideas are surveyed in this chapter. For some of them, consistency theorems are stated.

Let f_{nk} denote the k-nearest neighbor estimate, where we temporarily make the dependence upon k explicit. One could consider linear combinations in a number of ways. If $(v_{n1}, \ldots, v_{nn})$ is a probability weight vector (i.e., each v_{nj} is nonnegative and $\sum_{j=1}^{n} v_{nj} = 1$), then a simple linear combination could be envisaged,

$$f_n = \sum_{j=1}^{n} v_{nj} f_{nj}.$$

Sufficient conditions for pointwise and uniform consistency can be derived almost effortlessly from the results of the preceding chapters.

There are, of course, many other ways of combining k-nearest neighbor estimates (see, e.g., Breiman et al., 1977; Moore and Yackel, 1977b; Rodríguez and Van Ryzin, 1985, 1986; Rodríguez, 1986, 2001; Biau et al., 2011). One of particular interest is the inverse average:

$$\frac{1}{f_n} = \sum_{j=1}^{n} \frac{v_{nj}}{f_{nj}}.$$

In general, for $p \in \mathbb{R}$, $p \neq 0$, we may consider

$$(f_n)^p = \sum_{j=1}^{n} v_{nj} (f_{nj})^p$$

G. Biau, L. Devroye, *Lectures on the Nearest Neighbor Method*,
Springer Series in the Data Sciences, DOI 10.1007/978-3-319-25388-6_5

as a way of averaging. All of these are consistent under modest conditions on $(v_{n1}, \ldots, v_{nn})$. It provides a bottomless source of student exercises. For simplicity, we only deal with the case $p = 1$.

5.2 Weak consistency

For now, we set

$$f_n = \sum_{j=1}^{n} v_{nj} f_{nj},$$

where $(v_{n1}, \ldots, v_{nn})$ is a probability weight vector (by convention, we let $v_{nj} = 0$ for $j > n$). As in the previous chapters, the distribution of the target density f is denoted by μ. Our first result concerns the weak consistency of f_n. ($\lceil \cdot \rceil$ is the ceiling function.)

Theorem 5.1. *Assume that, for all $\varepsilon > 0$,*

$$\sum_{j \le \lceil 1/\varepsilon \rceil} v_{nj} + \sum_{j \ge \lceil \varepsilon n \rceil} v_{nj} \to 0 \quad \text{as } n \to \infty.$$

(Or, equivalently, that there exist sequences $\{k\} = \{k_n\}$ and $\{\ell\} = \{\ell_n\}$ with $k \le \ell$, such that $k \to \infty$, $\ell/n \to 0$, and $\sum_{j=k}^{\ell} v_{nj} \to 1$.) Then, for λ-almost all $\mathbf{x} \in \mathbb{R}^d$,

$$f_n(\mathbf{x}) \to f(\mathbf{x}) \quad \text{in probability}$$

and indeed, if additionally $v_{n1} = v_{n2} = 0$, then, still for λ-almost all $\mathbf{x} \in \mathbb{R}^d$,

$$\mathbb{E}\,|f_n(\mathbf{x}) - f(\mathbf{x})| \to 0.$$

The proof of Theorem 5.1 starts with two lemmas.

Lemma 5.1. *If $\mathbf{x} \in \mathbb{R}^d$ is a Lebesgue point of f, then, for all $\varepsilon \in (0, 1)$,*

$$\liminf_{n \to \infty} \|\mathbf{X}_{(\lceil \varepsilon n \rceil)}(\mathbf{x}) - \mathbf{x}\| > 0 \quad \text{with probability one.}$$

Proof. If $\mathbf{x}$ is not in the support of μ, then with probability one $\|\mathbf{X}_{(\lceil \varepsilon n \rceil)}(\mathbf{x}) - \mathbf{x}\| \ge \rho_{\mathbf{x}}$ for some positive constant $\rho_{\mathbf{x}}$, and the result follows. So, let us assume that $\mathbf{x}$ belongs to the support of μ. In this case, for $\delta > 0$, $\mu(B(\mathbf{x}, \delta)) > 0$ and

$$\mathbb{P}\{\|\mathbf{X}_{(\lceil \varepsilon n \rceil)}(\mathbf{x}) - \mathbf{x}\| < \delta\} = \mathbb{P}\{\mathrm{Bin}(n, \mu\,(B(\mathbf{x}, \delta))) \ge \lceil \varepsilon n \rceil\},$$

which tends to zero exponentially quickly in n by Chernoff's bound (see Theorem 20.5 in the Appendix) whenever $\mu(B(\mathbf{x}, \delta)) < \varepsilon$. But, since $\mathbf{x}$ is a Lebesgue

point of f, $\mu(B(\mathbf{x},\delta))/(V_d\delta^d) \to f(\mathbf{x})$ as $\delta \downarrow 0$ by the Lebesgue differentiation theorem (Theorem 20.18). Thus, such a choice of δ exists and we conclude the proof by the Borel-Cantelli lemma. □

Lemma 5.2. *For any* $\mathbf{x} \in \mathbb{R}^d$ *and any* $\varepsilon \in (0,1)$,

$$\limsup_{n\to\infty} \|\mathbf{X}_{(\lceil \varepsilon n\rceil)}(\mathbf{x}) - \mathbf{x}\| < \infty \quad \textit{with probability one}.$$

Proof. For $0 < K < \infty$,

$$\mathbb{P}\left\{\|\mathbf{X}_{(\lceil \varepsilon n\rceil)}(\mathbf{x}) - \mathbf{x}\| > K\right\} = \mathbb{P}\left\{\mathrm{Bin}(n, \mu\left(B(\mathbf{x},K)\right)) < \lceil \varepsilon n\rceil\right\},$$

which tends to zero exponentially quickly in n by Chernoff's bound whenever $\mu(B(\mathbf{x},K)) > \varepsilon + \frac{1}{n}$. As $K \uparrow \infty$, $\mu(B(\mathbf{x},K)) \uparrow 1$, and thus, for all n large enough, it is possible to choose such a K. We conclude by the Borel-Cantelli lemma. □

Proof (Theorem 5.1). We leave the proof of the last part to the interested reader. Let us note only that it requires at all Lebesgue points $\mathbf{x}$ of f,

$$\sup_{n\geq 1} \max_{1\leq j\leq n} \mathbb{E} f_{nj}(\mathbf{x}) < \infty.$$

For the first part, let $\mathbf{x}$ be a Lebesgue point of f. Define

$$H(\rho) = \sup_{0<\delta\leq\rho} \left| \frac{\mu\left(B(\mathbf{x},\delta)\right)}{\lambda\left(B(\mathbf{x},\delta)\right)} - f(\mathbf{x}) \right|,$$

and recall that $H(\rho) \to 0$ as $\rho \downarrow 0$. We write $R_{(j)}(\mathbf{x}) = \|\mathbf{X}_{(j)}(\mathbf{x}) - \mathbf{x}\|$, and define

$$U_{(j)} = \mu\left(B\left(\mathbf{x}, R_{(j)}(\mathbf{x})\right)\right).$$

We have, for any $\varepsilon \in (0, 1/2)$,

$$\begin{aligned}
\left| \sum_{j=1}^{n} v_{nj} f_{nj}(\mathbf{x}) - f(\mathbf{x}) \right| &\leq \left| \sum_{j\leq\lceil \varepsilon n\rceil} v_{nj}\left(f_{nj}(\mathbf{x}) - f(\mathbf{x})\right) \right| \\
&\quad + \left(\sum_{j\geq\lceil \varepsilon n\rceil} v_{nj} \right) \left(f(\mathbf{x}) + \max_{j\geq\lceil \varepsilon n\rceil} f_{nj}(\mathbf{x}) \right) \\
&\stackrel{\text{def}}{=} \mathbf{I} + \mathbf{II}.
\end{aligned}$$

By assumption, the first factor of **II** is o(1). The second factor of **II** is not more than

$$f(\mathbf{x}) + \frac{1}{V_d R^d_{(\lceil \varepsilon n\rceil)}(\mathbf{x})}.$$

By Lemma 5.1,

$$\limsup_{n\to\infty} \frac{1}{R^d_{(\lceil \varepsilon n\rceil)}(\mathbf{x})} < \infty \quad \text{with probability one.}$$

Therefore, $\mathbf{II} \to 0$ almost surely. Next, using the representation

$$\begin{aligned} f_{nj}(\mathbf{x}) &= \frac{\mu\left(B\left(\mathbf{x}, R_{(j)}(\mathbf{x})\right)\right)}{\lambda\left(B\left(\mathbf{x}, R_{(j)}(\mathbf{x})\right)\right)} \times \frac{j/n}{U_{(j)}} \\ &= \left[f(\mathbf{x}) + \Theta H\left(R_{(\lceil \varepsilon n\rceil)}(\mathbf{x})\right)\right] \times \frac{j/n}{U_{(j)}}, \quad 1 \le j \le \lceil \varepsilon n\rceil, \end{aligned}$$

where Θ from here on represents an arbitrary random variable with $|\Theta| \le 1$, we see that

$$\mathbf{I} \le f(\mathbf{x}) \left| \sum_{j\le\lceil \varepsilon n\rceil} v_{nj}\left(\frac{j/n}{U_{(j)}} - 1\right)\right| + H\left(R_{(\lceil \varepsilon n\rceil)}(\mathbf{x})\right) \sum_{j\le\lceil \varepsilon n\rceil} \left(v_{nj} \times \frac{j/n}{U_{(j)}}\right).$$

Recall from Lemma 5.2 that $\limsup_{n\to\infty} H(R_{(\lceil \varepsilon n\rceil)}(\mathbf{x})) < \infty$ with probability one. Inspection of that proof shows that if $f(\mathbf{x}) > 0$, then

$$\limsup_{n\to\infty} H\left(R_{(\lceil \varepsilon n\rceil)}(\mathbf{x})\right) \le K_\varepsilon < \infty \quad \text{with probability one,}$$

where K_ε is any constant strictly larger than $H(\xi)$, with ξ the solution of

$$\mu\left(B(\mathbf{x}, \xi)\right) = 2\varepsilon.$$

If $\mathbf{x}$ belongs to the support of μ, then, clearly, K_ε is as small as desired by choice of ε. On the other hand, if $\mathbf{x}$ is not in the support of μ, then $f(\mathbf{x}) = 0$ (since $\mathbf{x}$ is a Lebesgue point) and $R_{(1)}(\mathbf{x}) \ge \rho_{\mathbf{x}}$ for some positive constant $\rho_{\mathbf{x}}$, with probability one, and we have

$$f_{nj}(\mathbf{x}) \le \frac{j}{nV_d\rho^d_{\mathbf{x}}}.$$

Thus, in this case,

$$\mathbf{I} = \sum_{j\le\lceil \varepsilon n\rceil} v_{nj} f_{nj}(\mathbf{x}) \le \frac{\varepsilon + 1/n}{V_d\rho^d_{\mathbf{x}}},$$

which again is as small as desired by choice of ε.

Therefore, the proof is complete if

$$\sum_{j=1}^{n} \left(v_{nj} \times \frac{j/n}{U_{(j)}}\right) = \mathrm{O}_{\mathbb{P}}(1) \tag{5.1}$$

and if

$$\sum_{j=1}^{n} v_{nj} \left| \frac{j/n}{U_{(j)}} - 1 \right| = \mathrm{o}_{\mathbb{P}}(1). \tag{5.2}$$

We treat (5.1) first, bounding it by

$$\max_{1 \le j \le n} \frac{j/n}{U_{(j)}}.$$

For fixed $t > 1$, we have

$$\begin{aligned}
\mathbb{P}\left\{ \max_{1 \le j \le n} \frac{j/n}{U_{(j)}} > t \right\} &\le \sum_{j=1}^{n} \mathbb{P}\left\{ U_{(j)} < \frac{j}{nt} \right\} \\
&= \sum_{j=1}^{n} \mathbb{P}\left\{ \mathrm{Bin}\left(n, \tfrac{j}{nt}\right) \ge j \right\} \\
&\le \sum_{j=1}^{n} \exp\left(j - \frac{j}{t} - j \log t \right) \\
&\quad \text{(by Chernoff's bound)} \\
&\le \sum_{j=1}^{\infty} e^{-\rho j},
\end{aligned}$$

where $\rho = 1/t + \log t - 1$, which is positive for $t > 1$. Hence,

$$\mathbb{P}\left\{ \max_{1 \le j \le n} \frac{j/n}{U_{(j)}} > t \right\} \le \frac{e^{-\rho}}{1 - e^{-\rho}},$$

and therefore

$$\sum_{j=1}^{n} \left(v_{nj} \times \frac{j/n}{U_{(j)}} \right) = \mathrm{O}_{\mathbb{P}}(1).$$

Finally, we turn to (5.2), which is bounded from above by

$$\left(1 + \max_{1 \le j \le n} \frac{j/n}{U_{(j)}} \right) \sum_{j \le \lceil 1/\delta \rceil} v_{nj} + \max_{j \ge \lceil 1/\delta \rceil} \left| \frac{j/n}{U_{(j)}} - 1 \right| \stackrel{\text{def}}{=} \mathbf{III} + \mathbf{IV},$$

where $\delta > 0$ will be chosen later. Now, by identity (5.1), one has **III**=$\mathrm{o}_{\mathbb{P}}(1)$ since $\sum_{j \le \lceil 1/\delta \rceil} v_{nj} = \mathrm{o}(1)$ for any fixed $\delta > 0$. Finally, for $t > 0$ arbitrary small,

$$\mathbb{P}\{\mathbf{IV} > t\} \le \sum_{j \ge \lceil 1/\delta \rceil} \left(\mathbb{P}\left\{ U_{(j)} < \frac{j}{n(1+t)} \right\} + \mathbb{P}\left\{ U_{(j)} > \frac{j}{n(1-t)} \right\} \right).$$

Applying Chernoff's bound to both probabilities on the right, we have

$$\mathbb{P}\{\mathbf{IV} > t\} \le \sum_{j \ge \lceil 1/\delta \rceil} \left(e^{-\beta j} + e^{-\gamma j} \right),$$

where

$$\beta = \frac{1}{1+t} + \log(1+t) - 1 \quad (\text{which is} > 0)$$

and

$$\gamma = \frac{1}{1-t} + \log(1-t) - 1 \quad (\text{which is} > 0).$$

So,

$$\mathbb{P}\{\mathbf{IV} > t\} \le \frac{e^{-\beta/\delta}}{1-e^{-\beta}} + \frac{e^{-\gamma/\delta}}{1-e^{-\gamma}},$$

which is as small as desired by choice of δ. This concludes the proof. □

5.3 Strong consistency

Strong pointwise and uniform consistency of weighted estimates can be problematic if too much weight is attached to the misbehaving members in the family. We offer a strong convergence theorem that takes care of this in a minimal manner—the most worrisome f_{nj}'s are of course those with small values of j. The technical condition imposed is that for all fixed $\delta > 0$,

$$\sum_{n \ge 1} \left(\log n \sum_{j \le \lceil (1/\delta) \log n \rceil} v_{nj} \right) < \infty. \tag{5.3}$$

An example of $(v_{n1}, \ldots, v_{nn})$ satisfying (5.3) is one in which $v_{nj} = 1/k$ for $j \le k$, where $k/\log^2 n \to \infty$. Another example has $v_{nj} = 0$ for $j < k$, and $k/\log n \to \infty$.

Theorem 5.2. *If in addition to the conditions of Theorem 5.1, we have (5.3), then, for λ-almost all $\mathbf{x} \in \mathbb{R}^d$,*

$$f_n(\mathbf{x}) \to f(\mathbf{x}) \quad \textit{almost surely}.$$

Proof. The proof is as that of the previous theorem, in which the only modification needed is in (5.1)–(5.2), which have to be replaced by

$$\sum_{j=1}^{n} \left(v_{nj} \times \frac{j/n}{U_{(j)}} \right) = \mathrm{O}(1) \quad \text{almost surely,} \tag{5.4}$$

and

$$\sum_{j=1}^{n} v_{nj} \left| \frac{j/n}{U_{(j)}} - 1 \right| \to 0 \quad \text{almost surely.} \tag{5.5}$$

This is achieved by replacing $1/\delta$ in the proof throughout by $(1/\delta)\log n$. We bound (5.4) by

$$\left(\max_{j \le \lceil (1/\delta)\log n \rceil} \frac{j/n}{U_{(j)}} \right) \sum_{j \le \lceil (1/\delta)\log n \rceil} v_{nj} + \max_{j \ge \lceil (1/\delta)\log n \rceil} \frac{j/n}{U_{(j)}} \stackrel{\text{def}}{=} \mathbf{V} + \mathbf{VI}.$$

Now, note that for $t > 1$,

$$\mathbb{P}\{\mathbf{VI} > t\} \le \sum_{j \ge \lceil (1/\delta)\log n \rceil} e^{-\rho j},$$

where $\rho = 1/t + \log t - 1$. Thus,

$$\mathbb{P}\{\mathbf{VI} > t\} \le \frac{n^{-\rho/\delta}}{1 - e^{-\rho}},$$

which is summable in n as soon as t is so large that $\rho > \delta$. Therefore, with probability one, for all n large enough, $\mathbf{VI} \le t$.

Next, for $t' > 1$,

$$\mathbb{P}\left\{ \max_{j \le \lceil (1/\delta)\log n \rceil} \frac{j/n}{U_{(j)}} > t' \right\} \le \left\lceil \frac{\log n}{\delta} \right\rceil e^{-\rho'},$$

where $\rho' = 1/t' + \log t' - 1$, and therefore, since $\rho' \ge \log(t'/e)$,

$$\mathbb{P}\left\{ \mathbf{V} > t' \sum_{j \le \lceil (1/\delta)\log n \rceil} v_{nj} \right\} \le \left\lceil \frac{\log n}{\delta} \right\rceil \frac{e}{t'}.$$

Set

$$t' = \frac{t}{\sum_{j \le \lceil (1/\delta)\log n \rceil} v_{nj}}$$

and note, since

$$\sum_{j \le \lceil (1/\delta) \log n \rceil} v_{nj} = \mathrm{o}\left(\frac{1}{\log n}\right) \tag{5.6}$$

by condition (5.3), that $t' > 1$ for all n large enough. Then

$$\mathbb{P}\{\mathbf{V} > t\} \le \frac{2e}{\delta t}\left(\sum_{j \le \lceil (1/\delta) \log n \rceil} v_{nj}\right) \log n.$$

By condition (5.3), this is summable in n. Thus, with probability one, for all n large enough,

$$\mathbf{V} + \mathbf{VI} \le 2t,$$

and t was arbitrary. Finally, we deal with (5.5) as we did with (5.2):

$$\begin{aligned}\sum_{j=1}^{n} v_{nj} \left| \frac{j/n}{U_{(j)}} - 1 \right| &\le \left(1 + \max_{1 \le j \le n} \frac{j/n}{U_{(j)}}\right) \sum_{j \le \lceil (1/\delta) \log n \rceil} v_{nj} \\ &\quad + \max_{j \ge \lceil (1/\delta) \log n \rceil} \left| \frac{j/n}{U_{(j)}} - 1 \right| \\ &\stackrel{\text{def}}{=} \mathbf{VII} + \mathbf{VIII}.\end{aligned}$$

Using (5.4) and (5.6), we see that $\mathbf{VII} = \mathrm{o}(1)$ with probability one. For arbitrary small $t > 0$,

$$\mathbb{P}\{\mathbf{VIII} > t\} \le \frac{e^{-(\beta/\delta) \log n}}{1 - e^{-\rho}} + \frac{e^{-(\gamma/\delta) \log n}}{1 - e^{-\gamma}}$$

using the same β and γ introduced in the previous proof. This is summable in n if $\delta < \min(\beta, \gamma)$. Thus, for such δ, with probability one and for all n large enough,

$$\mathbf{VII} + \mathbf{VIII} < 2t.$$

Again, t was arbitrary, and therefore,

$$\mathbf{VII} + \mathbf{VIII} \to 0 \quad \text{almost surely}$$

for all $\varepsilon > 0$ small enough. This concludes the proof. □

We conclude this section with the following simple theorem:

Theorem 5.3. *Assume that f is uniformly continuous. If there exist sequences $\{k\} = \{k_n\}$ and $\{\ell\} = \{\ell_n\}$ such that $k \le \ell$, $k/\log n \to \infty$, $\ell/n \to 0$, and $v_{nj} = 0$ for $j < k$ or $j > \ell$, then*

$$\sup_{\mathbf{x}\in\mathbb{R}^d} |f_n(\mathbf{x}) - f(\mathbf{x})| \to 0 \quad \textit{almost surely}.$$

Proof. Verify the proof for the standard k-nearest neighbor estimate (Theorem 4.2), and note that for any $\gamma > 0$, $\varepsilon > 0$,

$$\sum_{n\ge 1} n^\gamma \left(\max_{k\le j\le \ell} \mathbb{P}\left\{ \sup_{\mathbf{x}\in\mathbb{R}^d} |f_{nj}(\mathbf{x}) - f(\mathbf{x})| > \varepsilon \right\} \right) < \infty. \tag{5.7}$$

The theorem then follows by the union bound without further work. □

By (5.7), we also have the following result, which permits us to select the value of k in the k-nearest neighbor estimate depending upon the data.

Proposition 5.1. *Let $K_n = K_n(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ be a random variable such that almost surely, $K_n/\log n \to \infty$ and $K_n/n \to 0$. Then, if f is uniformly continuous,*

$$\sup_{\mathbf{x}\in\mathbb{R}^d} |f_n(\mathbf{x}) - f(\mathbf{x})| \to 0 \quad \textit{almost surely},$$

where f_n is the k-nearest neighbor density estimate with $k = K_n$.

Proof. Note that there exist sequences $\{k\}$ and $\{\ell\}$ such that $k \le \ell$, $k/\log n \to \infty$, $\ell/n \to 0$, and with probability one, $k \le K_n \le \ell$. Thus, for $\varepsilon > 0$, writing f_{nk} for the k-nearest neighbor estimate,

$$\mathbb{P}\left\{ \sup_{\mathbf{x}\in\mathbb{R}^d} |f_{nK_n}(\mathbf{x}) - f(\mathbf{x})| > \varepsilon \text{ i.o.} \right\} \le \mathbb{P}\{K_n < k \text{ i.o.}\} + \mathbb{P}\{K_n > \ell \text{ i.o.}\}$$
$$+ \mathbb{P}\left\{ \max_{k\le j\le \ell} \sup_{\mathbf{x}\in\mathbb{R}^d} |f_{nj}(\mathbf{x}) - f(\mathbf{x})| > \varepsilon \text{ i.o.} \right\}.$$

The first two probabilities are zero. The last one is zero by the Borel-Cantelli lemma if

$$\sum_{n\ge 1} \mathbb{P}\left\{ \max_{k\le j\le \ell} \sup_{\mathbf{x}\in\mathbb{R}^d} |f_{nj}(\mathbf{x}) - f(\mathbf{x})| > \varepsilon \right\} < \infty.$$

But the sum is bounded by

$$\sum_{n\ge 1} \left(n \max_{k\le j\le \ell} \mathbb{P}\left\{ \sup_{\mathbf{x}\in\mathbb{R}^d} |f_{nj}(\mathbf{x}) - f(\mathbf{x})| > \varepsilon \right\} \right) < \infty,$$

which is finite in view of (5.7). □

Chapter 6
Local behavior

6.1 The set-up

No study of a density estimate is complete without a discussion of the local behavior of it. That is, given a certain amount of smoothness at $\mathbf{x}$, how fast does $f_n(\mathbf{x})$ tend to $f(\mathbf{x})$? It is clear that for any sequence of density estimates, and any sequence $a_n \downarrow 0$, however slow, there exists a density f with $\mathbf{x}$ a Lebesgue point of f, such that

$$\limsup_{n\to\infty} \frac{\mathbb{E}\,|f_n(\mathbf{x}) - f(\mathbf{x})|}{a_n} \geq 1.$$

We will not show this, but just point to a similar theorem for the total variation error in density estimation (Devroye, 1987). However, under smoothness conditions, there is hope to get useful rates of convergence.

Let us begin by noting that to estimate $f(\mathbf{x})$ using only $\|\mathbf{X}_1 - \mathbf{x}\|, \ldots, \|\mathbf{X}_n - \mathbf{x}\|$, we might as well focus on the estimation of the density g of $Y \stackrel{\text{def}}{=} \|\mathbf{X} - \mathbf{x}\|^d$ at 0. So, the data are $Y_1, \ldots, Y_n$ with $Y_i = \|\mathbf{X}_i - \mathbf{x}\|^d$, where the exponent d is chosen for a reason that will become apparent below. For $d = 1$, we have

$$g(y) = f(x + y) + f(x - y), \quad y \geq 0,$$

and then, $g(0) = 2f(x)$. For general d, if $\mathbf{x}$ is a Lebesgue point of f,

$$g(0) = \lim_{\rho\downarrow 0} \frac{\mathbb{P}\left\{\|\mathbf{X} - \mathbf{x}\|^d \leq \rho\right\}}{\rho} = \lim_{\rho\downarrow 0} \frac{\mu\left(B(\mathbf{x}, \rho^{1/d})\right)}{\rho} = V_d f(\mathbf{x}),$$

where μ is the probability measure for f.

G. Biau, L. Devroye, *Lectures on the Nearest Neighbor Method*,
Springer Series in the Data Sciences, DOI 10.1007/978-3-319-25388-6_6

The most essential tool in our study is Taylor series expansion, which states that if g has $\ell \geq 0$ continuous derivatives $g^{(0)}(0), \ldots, g^{(\ell)}(0)$ at 0, then

$$g(y) = \sum_{j=0}^{\ell} \frac{g^{(j)}(0)}{j!} y^j + \mathrm{o}(y^\ell), \quad y \downarrow 0.$$

If f is a smooth density on $\mathbb{R}^d$, $d \geq 1$, then there is a multivariate Taylor series of f around $\mathbf{x}$ (see, e.g., Giaquinta and Modica, 2009), which translates into a different but related expansion for g. For $d = 1$, if f has ℓ continuous derivatives at x, then, in view of

$$g(y) = f(x + y) + f(x - y),$$

we have

$$g^{(\ell)}(0) = \begin{cases} 2f^{(\ell)}(x) & \text{for } \ell \text{ even} \\ 0 & \text{for } \ell \text{ odd,} \end{cases}$$

so all odd terms are absent. For $d > 1$ and $\ell = 2$, the correspondence is as follows. Let $\mathbf{x} = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$, set

$$f'(\mathbf{x}) = \left(\frac{\partial f}{\partial x_1}(\mathbf{x}), \ldots, \frac{\partial f}{\partial x_d}(\mathbf{x}) \right)^\top,$$

and let $f''(\mathbf{x})$ be the $d \times d$ Hessian matrix of partial derivatives

$$f''(\mathbf{x}) = \left(\frac{\partial^2 f}{\partial x_j \partial x_{j'}}(\mathbf{x}) \right)_{1 \leq j, j' \leq d},$$

where $\top$ denotes transposition and vectors are in column format. Then the multivariate Taylor series of order two at $\mathbf{x}$ is given by

$$f(\mathbf{x} + \mathbf{y}) = f(\mathbf{x}) + f'(\mathbf{x})^\top \mathbf{y} + \frac{1}{2} \mathbf{y}^\top f''(\mathbf{x}) \mathbf{y} + \mathrm{o}(\|\mathbf{y}\|^2), \quad \|\mathbf{y}\| \downarrow 0.$$

Now, $g(\rho)$ is the derivative with respect to ρ of

$$\begin{aligned} \mathbb{P}\left\{ \|\mathbf{X} - \mathbf{x}\|^d \leq \rho \right\} &= \int_{B(\mathbf{0}, \rho^{1/d})} f(\mathbf{x}) \mathrm{d}\mathbf{y} + \int_{B(\mathbf{0}, \rho^{1/d})} f'(\mathbf{x})^\top \mathbf{y} \, \mathrm{d}\mathbf{y} \\ &\quad + \frac{1}{2} \int_{B(\mathbf{0}, \rho^{1/d})} \mathbf{y}^\top f''(\mathbf{x}) \mathbf{y} \, \mathrm{d}\mathbf{y} + \int_{B(\mathbf{0}, \rho^{1/d})} \mathrm{o}(\|\mathbf{y}\|^2) \mathrm{d}\mathbf{y} \\ &= f(\mathbf{x}) V_d \rho + 0 + \frac{V_d \rho}{2} \int_{B(\mathbf{0}, \rho^{1/d})} \mathbf{y}^\top f''(\mathbf{x}) \mathbf{y} \frac{\mathrm{d}\mathbf{y}}{V_d \rho} + \mathrm{o}(\rho^{1+2/d}). \end{aligned}$$

The remaining integral can be written as

$$\frac{V_d \rho}{2} \mathbb{E}[\mathbf{Y}^\top f''(\mathbf{x})\mathbf{Y}],$$

where $\mathbf{Y}$ is uniformly distributed inside $B(\mathbf{0}, \rho^{1/d})$. By rescaling, this is

$$\frac{V_d \rho^{1+2/d}}{2} \mathbb{E}[\mathbf{Z}^\top f''(\mathbf{x})\mathbf{Z}],$$

where $\mathbf{Z}$ is uniform in $B(\mathbf{0}, 1)$. It is a straightforward exercise to show that

$$\mathbf{Z} \stackrel{\mathscr{D}}{=} U^{1/d} \times \frac{(N_1, \ldots, N_d)}{\sqrt{\sum_{i=1}^d N_i^2}},$$

where $N_1, \ldots, N_d$ are i.i.d. standard normal random variables, and U is uniform $[0, 1]$ and independent of the N_i's. Clearly,

$$\mathbb{E}\left[\frac{N_j N_{j'}}{\sum_{i=1}^d N_i^2}\right] = \begin{cases} \frac{1}{d} & \text{if } j = j' \\ 0 & \text{if } j \neq j'. \end{cases}$$

Therefore,

$$\begin{aligned} \frac{V_d \rho^{1+2/d}}{2} \mathbb{E}[\mathbf{Z}^\top f''(\mathbf{x})\mathbf{Z}] &= \frac{V_d \rho^{1+2/d}}{2d} \times \mathbb{E}U^{2/d} \times \sum_{j=1}^d \frac{\partial^2 f}{\partial x_j^2}(\mathbf{x}) \\ &= \frac{V_d \rho^{1+2/d}}{2d} \times \frac{1}{1+\frac{2}{d}} \times \mathrm{tr}(f''(\mathbf{x})) \\ &= \frac{V_d}{2d+4} \times \mathrm{tr}(f''(\mathbf{x}))\rho^{1+2/d}, \end{aligned}$$

where $\mathrm{tr}(\Delta)$ stands for the trace (i.e., the sum of the elements on the main diagonal) of the square matrix Δ. We conclude that for general d, if f is twice continuously differentiable in a neighborhood of $\mathbf{x}$, and $y \downarrow 0$,

$$\begin{aligned} g(y) &= \frac{d}{dy}\left[f(\mathbf{x})V_d y + \mathrm{tr}(f''(\mathbf{x}))\frac{V_d}{2d+4} y^{1+2/d}\right] + \mathrm{o}(y^{2/d}) \\ &= f(\mathbf{x})V_d + \mathrm{tr}(f''(\mathbf{x}))\frac{V_d}{2d} y^{2/d} + \mathrm{o}(y^{2/d}). \end{aligned}$$

These expressions are our points of departure. Although g is univariate, notice the dependence upon d in the expansion about 0.

6.2 The first example: univariate case

In this section, we discuss the estimation of $g(0)$ when $d = 1$, assuming that f (and thus g) has two continuous derivatives in a neighborhood of x. Since, for $y \downarrow 0$,

$$g(y) = g(0) + \frac{g''(0)}{2} y^2 + \mathrm{o}(y^2),$$

we have

$$\mu\left(B(x,\rho)\right) = \int_0^\rho g(y)\mathrm{d}y = g(0)\rho + \frac{g''(0)}{6}\rho^3 + \rho^3 w(\rho),$$

where $w(\rho) = \mathrm{o}(1)$ as $\rho \downarrow 0$. Four cases can be considered according to whether $g(0) = 0$, $g(0) \neq 0$, combined with $g''(0) = 0$, $g''(0) \neq 0$. The most important one is $g(0) \neq 0$, $g''(0) \neq 0$.

As in the previous chapters, we define $X_{(i)}(x)$, $1 \leq i \leq n$, by reordering the $\|X_i - x\|$'s, so that

$$\|X_{(1)}(x) - x\| \leq \cdots \leq \|X_{(n)}(x) - x\|.$$

We write simply $Y_{(1)} \leq \cdots \leq Y_{(n)}$, with

$$Y_{(i)} = \|X_{(i)}(x) - x\|.$$

Recall from Chapter 1 that

$$\left(\mu\left(B(x, Y_{(1)})\right), \ldots, \mu\left(B(x, Y_{(n)})\right)\right) \stackrel{\mathscr{D}}{=} \left(U_{(1)}, \ldots, U_{(n)}\right),$$

where $U_1, \ldots, U_n$ are i.i.d. uniform $[0, 1]$ random variables. Thus, in particular, we have

$$U_{(k)} \stackrel{\mathscr{D}}{=} g(0)Y_{(k)} + \frac{g''(0)}{6} Y_{(k)}^3 + Z_k,$$

where $Z_k = Y_{(k)}^3 w(Y_{(k)})$. When $g(0) \neq 0$, inversion of this formula yields

$$Y_{(k)} \stackrel{\mathscr{D}}{=} \frac{U_{(k)}}{g(0)} - \frac{g''(0)}{6g^4(0)} U_{(k)}^3 + U_{(k)}^3 \epsilon(U_{(k)}),$$

where $\epsilon(\rho) = \mathrm{o}(1)$ as $\rho \downarrow 0$ is a given function. We write $W_k = U_{(k)}^3 \epsilon(U_{(k)})$.

The ordinary k-nearest neighbor estimate of $g(0)$ is

$$g_n(0) = \frac{k/n}{Y_{(k)}}.$$

Thus,

$$\frac{1}{g_n(0)} = \frac{Y_{(k)}}{k/n} \overset{\mathscr{D}}{=} \frac{1}{g(0)} \frac{U_{(k)}}{k/n} - \frac{g''(0)}{6g^4(0)} \frac{U_{(k)}^3}{k/n} + \frac{W_k}{k/n}$$

and

$$\begin{aligned}\frac{1}{g_n(0)} - \frac{1}{g(0)} &\overset{\mathscr{D}}{=} \frac{1}{g(0)} \left(\frac{U_{(k)}}{k/n} - 1 \right) - \frac{g''(0)}{6g^4(0)} \frac{U_{(k)}^3}{k/n} + \frac{W_k}{k/n} \\ &\overset{\text{def}}{=} \mathbf{I} + \mathbf{II} + \mathbf{III}.\end{aligned}$$

We recall that, by Theorem 1.4,

$$\frac{U_{(k)}}{k/n} \to 1 \quad \text{in probability when } k \to \infty.$$

Thus,

$$\frac{\mathbf{II}}{(k/n)^2} \to -\frac{g''(0)}{6g^4(0)} \quad \text{in probability.}$$

Recall also (Lemma 2.2) that as $k/n \to 0$, $U_{(k)} \to 0$ almost surely, and thus,

$$\frac{W_k}{(k/n)^3} \to 0 \quad \text{almost surely,}$$

so that

$$\frac{\mathbf{III}}{(k/n)^2} \to 0 \quad \text{almost surely.}$$

Finally, we turn to $\mathbf{I}$ and recall (Corollary 1.1) that

$$U_{(k)} \overset{\mathscr{D}}{=} \frac{E_1 + \cdots + E_k}{E_1 + \cdots + E_{n+1}},$$

where $E_1, \ldots, E_{n+1}$ are independent standard exponential random variables. Decompose $\mathbf{I}$ as

$$\left(\frac{n}{E_1 + \cdots + E_{n+1}} - 1 \right) \frac{E_1 + \cdots + E_k}{g(0)k} + \frac{E_1 + \cdots + E_k - k}{g(0)k}.$$

By an application of the central limit theorem and the delta method (see, e.g., van der Vaart, 1998),

$$\frac{n}{E_1 + \cdots + E_{n+1}} - 1 = \mathrm{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{n}} \right).$$

In addition, by the law of large numbers,

$$\frac{E_1 + \cdots + E_k}{k} \to 1 \quad \text{in probability if } k \to \infty.$$

Next, the central limit theorem and Lemma 20.1 in the Appendix yield

$$\frac{E_1 + \cdots + E_k - k}{k} \stackrel{\mathscr{D}}{=} \frac{1}{\sqrt{k}}\big(N + \mathrm{o}_{\mathbb{P}}(1)\big),$$

where $\mathrm{o}_{\mathbb{P}}(1)$ is a random variable tending to 0 in probability as $k \to \infty$, and N is a standard normal random variable possibly dependent on the $\mathrm{o}_{\mathbb{P}}(1)$ term.

Combining all terms, we obtain, if $k \to \infty$ and $k/n \to 0$,

$$\mathbf{I} + \mathbf{II} + \mathbf{III} \stackrel{\mathscr{D}}{=} \frac{N}{g(0)\sqrt{k}} + \mathrm{o}_{\mathbb{P}}\left(\frac{1}{\sqrt{k}}\right) - \frac{g''(0)}{6g^4(0)}\left(\frac{k}{n}\right)^2 + \mathrm{o}_{\mathbb{P}}\left(\left(\frac{k}{n}\right)^2\right).$$

Using inversions and simple manipulations of $\mathrm{O}_{\mathbb{P}}$ and $\mathrm{o}_{\mathbb{P}}$ terms, we finally conclude:

Lemma 6.1. *Let $g_n(0)$ be the k-nearest neighbor density estimate of $g(0)$. If $g(0) \neq 0$, $k \to \infty$ and $k/n \to 0$, then*

$$\frac{g(0)}{g_n(0)} - 1 \stackrel{\mathscr{D}}{=} \frac{N}{\sqrt{k}} - \frac{g''(0)}{6g^3(0)}\left(\frac{k}{n}\right)^2 + \mathrm{o}_{\mathbb{P}}\left(\frac{1}{\sqrt{k}} + \left(\frac{k}{n}\right)^2\right),$$

where N is a standard normal random variable. Moreover,

$$\frac{g_n(0)}{g(0)} - 1 \stackrel{\mathscr{D}}{=} \frac{N}{\sqrt{k}} + \frac{g''(0)}{6g^3(0)}\left(\frac{k}{n}\right)^2 + \mathrm{o}_{\mathbb{P}}\left(\frac{1}{\sqrt{k}} + \left(\frac{k}{n}\right)^2\right).$$

This lemma suggests the following definition:

Definition 6.1. When $g(0) \neq 0$, we say that the rate of convergence of the k-nearest neighbor density estimate $g_n(0)$ of $g(0)$ is a_n if

$$\frac{\frac{g_n(0)}{g(0)} - 1}{a_n} \stackrel{\mathscr{D}}{\to} W$$

for some random variable W that is not identically 0 with probability one.

For example, if $k \sim \gamma\, n^{\alpha}$, $0 < \alpha < 1$, $\gamma > 0$, then Lemma 6.1 implies that

$$\begin{cases} \frac{\frac{g_n(0)}{g(0)} - 1}{n^{-\alpha/2}} \stackrel{\mathscr{D}}{\to} \frac{N}{\sqrt{\gamma}} & \text{if } 0 < \alpha < \frac{4}{5} \\ \frac{\frac{g_n(0)}{g(0)} - 1}{n^{2\alpha-2}} \stackrel{\mathscr{D}}{\to} \frac{g''(0)}{6g^3(0)}\,\gamma^2 & \text{if } \frac{4}{5} < \alpha < 1 \\ \frac{\frac{g_n(0)}{g(0)} - 1}{n^{-2/5}} \stackrel{\mathscr{D}}{\to} \frac{N}{\sqrt{\gamma}} + \frac{g''(0)}{6g^3(0)}\,\gamma^2 & \text{if } \alpha = \frac{4}{5}. \end{cases}$$

Thus, if $g''(0) \neq 0$, the best possible rate for the k-nearest neighbor estimate is $n^{-2/5}$, which is achieved for k such that $k \sim \gamma\, n^{4/5}$, $\gamma > 0$. If $g''(0) = 0$, then there exists a sequence $\{k\} = \{k_n\}$ such that the rate $o(n^{-2/5})$ is achievable. However, without further conditions on g, one cannot precisely determine the best possible rate.

This leaves us with the choice of γ. There are several possibilities, all dependent upon how one interprets the limit random variable

$$W = \frac{N}{\sqrt{\gamma}} + \frac{g''(0)}{6g^3(0)}\,\gamma^2.$$

One can opt to minimize $\mathbb{P}\{|W| \geq c\}$ for constant c, or minimize $\mathbb{E}|W|^\beta$ for suitable β. For example, minimizing

$$\mathbb{E}W^2 = \frac{1}{\gamma} + \frac{g''(0)^2}{36g^6(0)}\,\gamma^4$$

with respect to γ yields the choice

$$\gamma = \left(\frac{9g^6(0)}{g''(0)^2}\right)^{1/5},$$

whence $k \sim \gamma\, n^{4/5}$. Note that the quantity $\frac{|g''(0)|}{g^3(0)}$ is scale-invariant and measures the "difficulty" of the estimation problem at hand.

6.3 Bias elimination in weighted k-nearest neighbor estimates

Still for $d = 1$, consider the weighted estimate

$$\frac{1}{g_n(0)} = \frac{\alpha}{g_{nk}(0)} + \frac{1-\alpha}{g_{n,2k}(0)}, \tag{6.1}$$

where $\alpha \in \mathbb{R}$ is a weight and g_{nk} is the k-nearest neighbor estimate of $g(0)$ (it is implicitly assumed that $1 \leq k \leq n/2$). We can also consider

$$g_n(0) = \alpha g_{nk}(0) + (1-\alpha) g_{n,2k}(0), \tag{6.2}$$

but the conclusions will be the same and are thus left to the reader.

Theorem 6.1. *Assume that g has two continuous derivatives in a neighborhood of 0 and $g(0) \neq 0$. Then, if we take $\alpha = 4/3$, there exists a choice of $k = k_n$ such that the rate of convergence of the weighted nearest neighbor density estimates (6.1) and (6.2) is $o(n^{-2/5})$.*

Without further conditions on g, we cannot pin down the rate of convergence. On the other hand, with additional smoothness assumptions on g, we could determine that better rate, but are then faced with the fact that there exists another weighted estimate with an even better (but undetermined) rate of convergence.

Proof (Theorem 6.1). As in the proof of the previous section, we obtain

$$\begin{aligned}\frac{g(0)}{g_n(0)} - 1 &\stackrel{\mathscr{D}}{=} \alpha\left(\frac{U_{(k)}}{k/n} - 1\right) + (1-\alpha)\left(\frac{U_{(2k)}}{2k/n} - 1\right) \\ &\quad - \frac{g''(0)}{6g^3(0)}\left[\alpha\left(\frac{k}{n}\right)^2 + (1-\alpha)\left(\frac{2k}{n}\right)^2\right] + \mathrm{o}_{\mathbb{P}}\left(\left(\frac{k}{n}\right)^2\right) \\ &\stackrel{\text{def}}{=} \mathbf{I} + \mathbf{II} + \mathbf{III}.\end{aligned}$$

We can make $\mathbf{II} = 0$ by setting $\alpha + 4(1-\alpha) = 0$, i.e., $\alpha = 4/3$. This leaves $\mathbf{I}$. It is clear that $\mathbf{I} = \mathrm{O}_{\mathbb{P}}(1/\sqrt{k})$ by the central limit theorem, thereby establishing that

$$\mathbf{I} + \mathbf{II} + \mathbf{III} = \mathrm{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{k}}\right) + \mathrm{o}_{\mathbb{P}}\left(\left(\frac{k}{n}\right)^2\right),$$

if $k \to \infty$ and $k/n \to 0$. This concludes the proof. □

Remark 6.1. It is easy to verify that

$$\mathbf{I} \stackrel{\mathscr{D}}{=} \frac{1}{\sqrt{k}}\left[\alpha N + \frac{1-\alpha}{2}(N + N')\right] + \mathrm{o}_{\mathbb{P}}\left(\frac{1}{\sqrt{k}}\right)$$

where N, N' are independent standard normal random variables. This can be rewritten as

$$\mathbf{I} \stackrel{\mathscr{D}}{=} \frac{1}{\sqrt{k}}\sqrt{\frac{1+\alpha^2}{2}}\,N + \mathrm{o}_{\mathbb{P}}\left(\frac{1}{\sqrt{k}}\right).$$

For $\alpha \neq 4/3$, if we set $k \sim \gamma\, n^{4/5}$, $\gamma > 0$, and optimize γ as in the previous section (where $\mathbb{E}W^2$ in the limit was minimized), then $\mathbb{E}W^2$ is a function of α times the quantity

$$\left(\frac{g''(0)^2}{g^6(0)}\right)^{1/5}.$$

The function of α is a constant times

$$(1+\alpha^2)^{4/5}|4 - 3\alpha|^{2/5},$$

which is minimal on $[0, 1]$ for $\alpha = (8 - \sqrt{19})/15$. However, the overall minimum on $\mathbb{R}$ is at $\alpha = 4/3$. This implies that ordinary convex combinations ($\alpha \in [0, 1]$) are not as powerful as general combinations ($\alpha \in \mathbb{R}$). □

Remark 6.2. We leave it to the reader to find the best form of the weights v_{nj}, $1 \leq j \leq n$, in the weighted estimate

$$\frac{1}{g_n(0)} = \sum_{j=1}^{n} \frac{v_{nj}}{g_{nj}(0)}$$

that jointly makes the bias 0 and minimizes the variance term (the **I** in the proof). □

Remark 6.3. The weighted estimate with $\alpha = 4/3$ satisfies

$$\frac{1}{g_n(0)} = \frac{\frac{4}{3}Y_{(k)}}{k/n} - \frac{\frac{1}{3}Y_{(2k)}}{2k/n} = \frac{8Y_{(k)} - Y_{(2k)}}{6k/n}.$$

This is negative if $8Y_{(k)} < Y_{(2k)}$. However, for any x (in the support of μ or not), if $k/n \to 0$, then there exist $n_0(x) > 0$ and $\rho(x) > 0$ such that

$$\mathbb{P}\{8Y_{(k)} - Y_{(2k)} < 0\} \leq e^{-\rho(x)k}, \quad n \geq n_0(x).$$

This can be shown using the Chernoff's bound on binomials (Theorem 20.5) and is left as an exercise as well. □

6.4 Rates of convergence in $\mathbb{R}^d$

We have seen in Section 6.1 that if f is twice continuously differentiable in a neighborhood of $\mathbf{x}$, then $Y = \|\mathbf{X} - \mathbf{x}\|^d$ has density g on $[0, \infty)$ given by the expansion

$$g(y) = g(0) + c\, y^{2/d} + o(y^{2/d}), \quad y \downarrow 0,$$

where c is a function of d and the trace of the Hessian:

$$c = \mathrm{tr}(f''(\mathbf{x}))\frac{V_d}{2d}.$$

Averaging as for the case $d = 1$, we have

$$U_{(k)} \stackrel{\mathcal{D}}{=} g(0)Y_{(k)} + \frac{c}{1 + \frac{2}{d}} Y_{(k)}^{1+2/d} + Z_k,$$

with $Z_k = \mathrm{o}(Y_{(k)}^{1+2/d})$. Thus, if $g(0) \neq 0$, by inversion,

$$Y_{(k)} \stackrel{\mathscr{D}}{=} \frac{U_{(k)}}{g(0)} - \frac{c}{(1+\frac{2}{d})g^{2+2/d}(0)} U_{(k)}^{1+2/d} + W_k,$$

with $W_k = \mathrm{o}(U_{(k)}^{1+2/d})$. The k-nearest neighbor estimate of $g(0)$ is

$$g_n(0) = \frac{k/n}{Y_{(k)}},$$

and so,

$$\begin{aligned}\frac{g(0)}{g_n(0)} - 1 &\stackrel{\mathscr{D}}{=} \left(\frac{U_{(k)}}{k/n} - 1\right) - \frac{c}{\left(1+\frac{2}{d}\right) g^{1+2/d}(0)} \frac{U_{(k)}^{1+2/d}}{k/n} + \frac{W_k}{k/n} \\ &\stackrel{\text{def}}{=} \mathbf{I} + \mathbf{II} + \mathbf{III}.\end{aligned}$$

Arguing as for $d = 1$, one easily obtains

$$\frac{g(0)}{g_n(0)} - 1 \stackrel{\mathscr{D}}{=} \frac{N}{\sqrt{k}} - \frac{c}{(1+\frac{2}{d})g^{1+2/d}(0)} \left(\frac{k}{n}\right)^{2/d} + \mathrm{o}_{\mathbb{P}}\left(\frac{1}{\sqrt{k}} + \left(\frac{k}{n}\right)^{2/d}\right)$$

and

$$\frac{g_n(0)}{g(0)} - 1 \stackrel{\mathscr{D}}{=} \frac{N}{\sqrt{k}} + \frac{c}{(1+\frac{2}{d})g^{1+2/d}(0)} \left(\frac{k}{n}\right)^{2/d} + \mathrm{o}_{\mathbb{P}}\left(\frac{1}{\sqrt{k}} + \left(\frac{k}{n}\right)^{2/d}\right).$$

When we set $k \sim \gamma\, n^{\frac{4}{d+4}}$, $\gamma > 0$, the right-hand side is $W + \mathrm{o}_{\mathbb{P}}(n^{-\frac{2}{d+4}})$, where

$$W = n^{-\frac{2}{d+4}} \left[\frac{N}{\sqrt{\gamma}} + \frac{c}{\left(1+\frac{2}{d}\right) g^{1+2/d}(0)} \gamma^{2/d}\right].$$

Whenever $c \neq 0$, minimizing $\mathbb{E}W^2$ yields the following choice for γ:

$$\gamma = \left(\frac{g^{2d+4}(0)(d+2)^{2d}}{4^d c^{2d} d^d}\right)^{\frac{1}{d+4}}.$$

We conclude that the rate of convergence for the k-nearest neighbor estimate is $n^{-\frac{2}{d+4}}$ if $c \neq 0$, and is $\mathrm{o}(n^{-\frac{2}{d+4}})$ if $c = 0$.

As in the case $d = 1$, combining two estimates to eliminate the bias is possible. Define, for $\alpha \in \mathbb{R}$,

$$\frac{1}{g_n(0)} = \frac{\alpha}{g_{nk}(0)} + \frac{1-\alpha}{g_{n,2k}(0)}, \tag{6.3}$$

or

$$g_n(0) = \alpha g_{nk}(0) + (1-\alpha) g_{n,2k}(0). \tag{6.4}$$

Then note that the bias term for (6.3) becomes

$$\lambda \left[\alpha \frac{U_{(k)}^{1+2/d}}{k/n} + (1-\alpha) \frac{U_{(2k)}^{1+2/d}}{2k/n} \right],$$

where

$$\lambda = -\frac{c}{\left(1+\frac{2}{d}\right) g^{1+2/d}(0)}.$$

This term is

$$\lambda \left(\frac{k}{n}\right)^{2/d} \left[\alpha + (1-\alpha) 2^{2/d}\right] + \mathrm{o}_{\mathbb{P}}\left(\left(\frac{k}{n}\right)^{2/d}\right).$$

The main contribution to this bias term is 0 when

$$\alpha = \frac{4^{1/d}}{4^{1/d}-1}.$$

Theorem 6.2. *Assume that f is twice continuously differentiable in a neighborhood of $\mathbf{x}$ and $f(\mathbf{x}) \neq 0$. Then, if we take*

$$\alpha = \frac{4^{1/d}}{4^{1/d}-1},$$

there exists a choice of $k = k_n$ such that the rate of convergence of the weighted nearest neighbor density estimates (6.3) and (6.4) is $\mathrm{o}(n^{-\frac{2}{d+4}})$.

Remark 6.4. If we consider the standard k-nearest neighbor density estimate f_n, then our results show that whenever f is twice continuously differentiable in a neighborhood of $\mathbf{x}$ and $f(\mathbf{x}) \neq 0$, then

$$\sqrt{k} \frac{f_n(\mathbf{x}) - f(\mathbf{x})}{f(\mathbf{x})} \xrightarrow{\mathscr{D}} N,$$

provided $k \to \infty$ and $\frac{k}{n^{4/(d+4)}} \to 0$. This is precisely the asymptotic normality result of Moore and Yackel (1977a) (see also Mack, 1980, and Berlinet and Levallois, 2000). Note however that the condition $\frac{k}{n^{4/(d+4)}} \to 0$ is less severe than the condition $\frac{k}{n^{2/(d+2)}} \to 0$ that is imposed by these authors, yet with a less stringent smoothness assumption on f.

Following Biau et al. (2011), it is also possible to analyze the mean squared error development of f_n and show that if $k \to \infty$ and $k/n \to 0$, then

$$\mathbb{E}\,|f_n(\mathbf{x}) - f(\mathbf{x})|^2 = \frac{f^2(\mathbf{x})}{k} + \frac{c_{\mathbf{x}}^2}{f^{4/d}(\mathbf{x})}\left(\frac{k}{n}\right)^{4/d} + \mathrm{o}\left(\frac{1}{k} + \left(\frac{k}{n}\right)^{4/d}\right),$$

where

$$c_{\mathbf{x}} = \frac{\mathrm{tr}(f''(\mathbf{x}))}{2(d+2)V_d^{2/d}}.$$

For further references, see Mack and Rosenblatt (1979) and Hall (1983). Thus, for such $\mathbf{x}$, assuming that $c_{\mathbf{x}} \neq 0$, and for the choice

$$k = \min\left(\left\lceil\left(\frac{df^{2+4/d}(\mathbf{x})}{4c_{\mathbf{x}}^2}\right)n^{\frac{4}{d+4}}\right\rceil, n\right),$$

we have

$$\mathbb{E}\,|f_n(\mathbf{x}) - f(\mathbf{x})|^2 = \xi_{\mathbf{x}}\, n^{-\frac{4}{d+4}} + \mathrm{o}\big(n^{-\frac{4}{d+4}}\big),$$

where

$$\xi_{\mathbf{x}} = \left(1 + \frac{d}{4}\right)\left(\frac{4f^{4/d}(\mathbf{x})c_{\mathbf{x}}^2}{d}\right)^{\frac{d}{d+4}}.$$

□

6.5 Behavior of f near $\mathbf{x}$

The preceding discussions show the importance of the behavior of f near $\mathbf{x}$. We can in general imagine that

$$g(y) = g(0) + \lambda y^{\beta} + \mathrm{o}(y^{\beta}), \quad y \downarrow 0,$$

where $\beta > 0$, $\lambda \neq 0$. Of course, other behaviors are possible as well, such as

$$g(y) = g(0) + \frac{\lambda}{\log^{\beta}(1/y)} + \mathrm{o}\left(\frac{1}{\log^{\beta}(1/y)}\right)$$

or

$$g(y) = g(0) + e^{-\lambda/y^{\beta}} + \mathrm{o}\big(e^{-\lambda/y^{\beta}}\big),$$

and, indeed, many other situations can occur. A case-by-case study is clearly not productive—what one needs is a manner of adjusting k locally based on the data, so that one adapts automatically to the local smoothness.

Classically, one considers Taylor series expansions and assumes that the last term is continuous. Suppose, to simplify, that $d = 1$ and that f has 2ℓ continuous derivatives at x, $\ell \geq 1$. Then g has the Taylor series expansion

$$g(y) = \sum_{j=0}^{\ell} \frac{g^{(2j)}(0)}{(2j)!} y^{2j} + o(y^{2\ell}), \quad y \downarrow 0.$$

Arguing as in the previous sections, it is easy to see that the rate of convergence of the ordinary k-nearest neighbor estimate is $n^{-\frac{2m}{4m+1}}$, where

$$m = \inf \left\{ 1 \leq j \leq \ell : g^{(2j)}(0) \neq 0 \right\},$$

assuming at least one of the coefficients in the Taylor series expansion is nonzero. If $g^{(2j)}(0) = 0$, $1 \leq j \leq \ell$, then the rate is $o(n^{-\frac{2\ell}{4\ell+1}})$.

Most often, $m = 1$, and we rediscover the $n^{-2/5}$ rate. However, given that we know that 2ℓ continuous derivatives exist, even if all coefficients in the Taylor series are nonzero, we can define a weighted k-nearest neighbor estimate with rate $o(n^{-\frac{2\ell}{4\ell+1}})$, where, once again, the precise rate is impossible to pin down without further knowledge of the behavior of g near 0. It is a straightforward exercise to derive such weighted estimates. Indeed, we combine as follows, using $\ell + 1$ fixed weights $v_1, \ldots, v_{\ell+1}$:

$$\frac{1}{g_n(0)} = \sum_{j=1}^{\ell+1} \frac{v_j}{g_{n,jk}(0)}.$$

If $k \to \infty$, $k/n \to 0$, and $\sum_{j=1}^{\ell+1} v_j = 1$, then a slight adaptation of Theorem 5.1 shows that this estimate is weakly consistent at 0, even if some weights are negative. On the other hand, for $1 \leq j \leq \ell$, we want the bias terms to disappear. A sketch of how this is done is as follows. Define

$$g^{\star}(y) = \sum_{j=0}^{\ell} \frac{g^{(2j)}(0)}{(2j+1)!} y^{2j+1},$$

and note that

$$\left| g^{\star}(y) - \int_0^y g(z) \mathrm{d}z \right| = o(y^{2\ell+1}), \quad y \downarrow 0.$$

Then, by Lagrange inversion of polynomials, there exists a function h on $[0, \infty)$ with

$$h(z) = \sum_{j=0}^{\ell} \frac{h^{\star(2j)}(0)}{(2j+1)!} z^{2j+1} + o(z^{2\ell+1}), \quad z \downarrow 0,$$

such that locally, as $y \downarrow 0$, since $g^\star$ is invertible in a small enough neighborhood of 0, $g^\star(y) = z$ if and only if $z = h(y)$. The coefficients in the Taylor series expansion of h depend upon those of g. We thus have for $U_{(jk)}$, $1 \le j \le \ell + 1$, jointly,

$$U_{(jk)} \stackrel{\mathscr{D}}{=} g^\star(Y_{(jk)}) + \mathrm{o}(Y_{(jk)}^{2\ell+1}),$$

and therefore,

$$Y_{(jk)} \stackrel{\mathscr{D}}{=} h(U_{(jk)}) + \mathrm{o}(U_{(jk)}^{2\ell+1}).$$

The bias term in the expansion of $\frac{1}{g_n(0)} - \frac{1}{g(0)}$ is

$$\begin{aligned}
&\sum_{j=1}^{\ell+1} v_j \left[\frac{1}{jk/n} \sum_{i=1}^{\ell} \frac{h^{\star(2i)}(0)}{(2i+1)!} U_{(jk)}^{2i+1} \right] + \frac{\mathrm{o}_{\mathbb{P}}\left(U_{(\ell+1)k}^{2\ell+1}\right)}{k/n} \\
&\quad = \sum_{i=1}^{\ell} \left[\frac{h^{\star(2i)}(0)}{(2i+1)!} \sum_{j=1}^{\ell+1} v_j \frac{U_{(jk)}^{2i+1}}{jk/n} \right] + \mathrm{o}_{\mathbb{P}}\left(\left(\frac{k}{n} \right)^{2\ell} \right).
\end{aligned}$$

Using

$$\frac{U_{(jk)}}{jk/n} = 1 + \mathrm{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{k}} \right),$$

this is

$$\begin{aligned}
&\sum_{i=1}^{\ell} \frac{h^{\star(2i)}(0)}{(2i+1)!} \left[\sum_{j=1}^{\ell+1} v_j j^{2i} \left(\frac{k}{n} \right)^{2i} \left(1 + \mathrm{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{k}} \right) \right) \right] + \mathrm{o}_{\mathbb{P}}\left(\left(\frac{k}{n} \right)^{2\ell} \right) \\
&\quad = \sum_{i=1}^{\ell} \left[\frac{h^{\star(2i)}(0)}{(2i+1)!} \left(\frac{k}{n} \right)^{2i} \sum_{j=1}^{\ell+1} v_j j^{2i} \right] + \mathrm{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{k}} \right) + \mathrm{o}_{\mathbb{P}}\left(\left(\frac{k}{n} \right)^{2\ell} \right).
\end{aligned}$$

The first term disappears if for all $1 \le i \le \ell$, $\sum_{j=1}^{\ell+1} v_j j^{2i} = 0$. Therefore, the rate of convergence is $\mathrm{o}(n^{-\frac{2\ell}{4\ell+1}})$ if this constraint holds together with $\sum_{j=1}^{\ell+1} v_j = 1$. That leaves a linear system of $\ell + 1$ equations with $\ell + 1$ unknowns, and we can provide a general way of combining k-nearest neighbor estimates:

$$\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_{\ell+1} \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1^2 & 2^2 & \cdots & (\ell+1)^2 \\ \vdots & \vdots & \vdots & \vdots \\ 1^{2\ell} & 2^{2\ell} & \cdots & (\ell+1)^{2\ell} \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

The central matrix is a Vandermonde matrix of the form

$$\mathbb{V} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \cdots & \lambda_{\ell+1} \\ \lambda_1^2 & \lambda_2^2 & \cdots & \lambda_{\ell+1}^2 \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_1^\ell & \lambda_2^\ell & \cdots & \lambda_{\ell+1}^\ell \end{pmatrix},$$

where $\lambda_j = j^2$, $1 \le j \le \ell + 1$. A simple form of the inverse matrix $\mathbb{V}^{-1}$ is described in terms of $\mathbb{U}\mathbb{L}$, where $\mathbb{U}$ is an upper triangular matrix and $\mathbb{L}$ a lower triangular matrix (see, e.g., Turner, 1966). The explicit forms of $\mathbb{U}$ and $\mathbb{L}$ are

$$u_{ij} = \begin{cases} 0 & \text{if } i > j \\ 1 & \text{if } i = j = 1 \\ \displaystyle\prod_{\substack{k=1 \\ k \neq i}}^{j} \frac{1}{\lambda_i - \lambda_k} & \text{otherwise,} \end{cases}$$

and

$$\ell_{ij} = \begin{cases} 0 & \text{if } i < j \\ 1 & \text{if } i = j \\ \ell_{i-1,j-1} - \ell_{i-1,j}\lambda_{i-1} & \text{otherwise,} \end{cases}$$

with the convention $\ell_{i0} = 0$. In more pedestrian terms,

$$\mathbb{U} = \begin{pmatrix} 1 & \frac{1}{\lambda_1 - \lambda_2} & \frac{1}{(\lambda_1 - \lambda_2)(\lambda_1 - \lambda_3)} & \cdots \\ 0 & \frac{1}{\lambda_2 - \lambda_1} & \frac{1}{(\lambda_2 - \lambda_1)(\lambda_2 - \lambda_3)} & \cdots \\ 0 & 0 & \frac{1}{(\lambda_3 - \lambda_1)(\lambda_3 - \lambda_2)} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

and

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \cdots \\ -\lambda_1 & 1 & 0 & 0 \cdots \\ \lambda_1\lambda_2 & -(\lambda_1 + \lambda_2) & 1 & 0 \cdots \\ -\lambda_1\lambda_2\lambda_3 & \lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_2\lambda_3 & -(\lambda_1 + \lambda_2 + \lambda_3) & 1 \cdots \\ \vdots & \vdots & \vdots & \vdots \ddots \end{pmatrix}.$$

It is noted that the last line of $\mathbb{L}$ does not depend on $\lambda_{\ell+1}$ but only on $\lambda_1, \ldots, \lambda_\ell$. Elementary matrix operations show that

$$v_j = \begin{cases} 1 + \sum_{i=2}^{\ell+1} \left[(-1)^{i-1} \lambda_1 \ldots \lambda_{i-1} \prod_{k=2}^{i} \frac{1}{\lambda_1 - \lambda_k} \right] & \text{if } j = 1 \\ \sum_{i=j}^{\ell+1} \left[(-1)^{i-1} \lambda_1 \ldots \lambda_{i-1} \prod_{\substack{k=1 \\ k \neq j}}^{i} \frac{1}{\lambda_j - \lambda_k} \right] & \text{if } 1 < j \le \ell + 1, \end{cases}$$

that is,

$$v_j = \begin{cases} 1 + \sum_{i=2}^{\ell+1} \left[(-1)^{i-1} \left[(i-1)! \right]^2 \prod_{k=2}^{i} \frac{1}{1-k^2} \right] & \text{if } j = 1 \\ \sum_{i=j}^{\ell+1} \left[(-1)^{i-1} \left[(i-1)! \right]^2 \prod_{\substack{k=1 \\ k \neq j}}^{i} \frac{1}{j^2-k^2} \right] & \text{if } 1 < j \leq \ell+1. \end{cases}$$

Finally, using some manipulations on the gamma function that are left to the reader, we conclude that

$$v_j = \frac{2(-1)^j \left[(\ell+1)! \right]^2 (j-\ell-2)}{(j+\ell+1)!(\ell-j+2)!}, \quad 1 \leq j \leq \ell+1.$$

It is noteworthy that the proof does not require any knowledge of the actual coefficients of the inverse function h. There may be better combinations if one allows more than $\ell+1$ component estimates. This added freedom can be used to minimize the variance term, which is proportional to

$$\frac{1}{k} \sum_{j=1}^{\ell+1} \frac{v_j^2}{j}$$

in case of $\ell+1$ terms, and to

$$\frac{1}{k} \sum_{j=1}^{L} \frac{v_j^2}{j}$$

in case of $L > \ell+1$ terms. However, such optimization does not alter the rate of convergence.

Remark 6.5. To conclude this section, we would like to point out the need to develop weighted k-nearest neighbor rules that adapt nicely to analytic densities, i.e., densities f (or g) that are completely determined by their (infinite) Taylor series expansions, and thus attain a rate of convergence equal to, or at least close to, $1/\sqrt{n}$. □

6.6 A nonlinear k-nearest neighbor estimate

The situation we are considering, by way of example, is that of the estimation of a density f on $\mathbb{R}$ that has 2ℓ continuous derivatives in a neighborhood of x for $\ell \geq 1$, and $f(x) > 0$. We know that the density g of $Y = \|X - x\|$ only has even-numbered terms in its Taylor series:

$$g(y) = \sum_{j=0}^{\ell} \frac{g^{(2j)}(0)}{(2j)!} y^{2j} + o(y^{2\ell}), \quad y \downarrow 0.$$

According to the previous section, by a linear combination of the $\ell+1$ k-nearest neighbor estimates, one can achieve a rate of convergence that is $o(n^{-\frac{2\ell}{4\ell+1}})$. The computations of the Lagrange inverse of a polynomial are cumbersome. Here we present a logical and simple nonlinear approach for bias elimination that has the same rate.

As before, we write $Y_{(1)} \le \cdots \le Y_{(n)}$ for the order statistics of the data. We also know that

$$\left(\mu\left(B(x, Y_{(1)})\right), \ldots, \mu\left(B(x, Y_{(n)})\right)\right) \stackrel{\mathscr{D}}{=} \left(U_{(1)}, \ldots, U_{(n)}\right),$$

a vector of uniform $[0,1]$ order statistics. Integrating g, we thus have, jointly for all $1 \le i \le n$,

$$U_{(i)} \stackrel{\mathscr{D}}{=} \sum_{j=0}^{\ell} \frac{g^{(2j)}(0)}{(2j+1)!} Y_{(i)}^{2j+1} + \psi(Y_{(i)}),$$

where $\psi(y) = o(y^{2\ell+1})$ when $y \downarrow 0$. Let us consider once again $\ell+1$ order statistics, $Y_{(k)}, Y_{(2k)}, \ldots, Y_{((\ell+1)k)}$, where $k = k_n \to \infty$ and $k/n \to 0$. Observe that $Y_{((\ell+1)k)} = O_{\mathbb{P}}(k/n)$ as $n \to \infty$. Thus, we have

$$\max_{1 \le i \le \ell+1} \left|\psi(Y_{(ik)})\right| = o_{\mathbb{P}}\left(\left(\frac{k}{n}\right)^{2\ell+1}\right).$$

In matrix notation,

$$\begin{pmatrix} Y_{(k)}^1 & Y_{(k)}^3 & \cdots & Y_{(k)}^{2\ell+1} \\ Y_{(2k)}^1 & Y_{(2k)}^3 & \cdots & Y_{(2k)}^{2\ell+1} \\ \vdots & \vdots & \vdots & \vdots \\ Y_{((\ell+1)k)}^1 & Y_{((\ell+1)k)}^3 & \cdots & Y_{((\ell+1)k)}^{2\ell+1} \end{pmatrix} \begin{pmatrix} \frac{g^{(0)}(0)}{1!} \\ \frac{g^{(2)}(0)}{3!} \\ \vdots \\ \frac{g^{(2\ell)}(0)}{(2\ell+1)!} \end{pmatrix} \stackrel{\mathscr{D}}{=} \begin{pmatrix} U_{(k)} \\ U_{(2k)} \\ \vdots \\ U_{((\ell+1)k)} \end{pmatrix} + o_{\mathbb{P}}\left(\left(\frac{k}{n}\right)^{2\ell+1}\right) \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Note that $U_{(i)}$ is close to i/n. More precisely, we recall (Chapter 1) that

$$\frac{U_{(ik)}}{ik/n} \to 1 \quad \text{in probability}, 1 \le i \le \ell+1,$$

and that

$$\frac{U_{(ik)} - ik/n}{\sqrt{ik/n}} \stackrel{\mathscr{D}}{=} N + o_{\mathbb{P}}(1), \quad 1 \le i \le \ell+1,$$

where N is a standard normal random variable. The exponential representation of order statistics (Corollary 1.1) implies the following, if we use $G_k(i)$ to denote independent Gamma(k) random variables, $1 \le i \le \ell + 1$:

$$\begin{aligned}
&\big(U_{(k)}, U_{(2k)} - U_{(k)}, \ldots, U_{((\ell+1)k)} - U_{(\ell k)}\big) \\
&\quad \stackrel{\mathscr{D}}{=} \frac{1}{E_1 + \cdots + E_{n+1}} \big(G_k(1), \ldots, G_k(\ell+1)\big) \\
&\quad = \frac{1 + \mathrm{O}_{\mathbb{P}}(1/\sqrt{n})}{n} \big(G_k(1), \ldots, G_k(\ell+1)\big) \\
&\quad \stackrel{\mathscr{D}}{=} \frac{1 + \mathrm{O}_{\mathbb{P}}(1/\sqrt{n})}{n} \left(k + \sqrt{k}\, N_1 + \mathrm{o}_{\mathbb{P}}(\sqrt{k}), \ldots, k + \sqrt{k}\, N_{\ell+1} + \mathrm{o}_{\mathbb{P}}(\sqrt{k})\right),
\end{aligned}$$

where $N_1, \ldots, N_{\ell+1}$ are i.i.d. standard normal random variables. Thus, setting up a matrix representation, we have

$$\begin{pmatrix} U_{(k)} \\ U_{(2k)} \\ \vdots \\ U_{((\ell+1)k)} \end{pmatrix} \stackrel{\mathscr{D}}{=} \frac{1+\mathrm{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right)}{n} \begin{pmatrix} k + \sqrt{k}\, N_1 \\ 2k + \sqrt{k}\,(N_1 + N_2) \\ \vdots \\ (\ell+1)k + \sqrt{k}\,(N_1 + \cdots + N_{\ell+1}) \end{pmatrix} + \mathrm{o}_{\mathbb{P}}\left(\frac{\sqrt{k}}{n}\right) \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

and thus,

$$\begin{pmatrix} \frac{U_{(k)} - k/n}{\sqrt{k}} \\ \frac{U_{(2k)} - 2k/n}{\sqrt{k}} \\ \vdots \\ \frac{U_{((\ell+1)k)} - (\ell+1)k/n}{\sqrt{k}} \end{pmatrix} \stackrel{\mathscr{D}}{=} \frac{1}{n} \begin{pmatrix} N_1 \\ N_1 + N_2 \\ \vdots \\ N_1 + \cdots + N_{\ell+1} \end{pmatrix} + \left(\mathrm{O}_{\mathbb{P}}\left(\frac{\sqrt{k}}{n^{3/2}}\right) + \mathrm{o}_{\mathbb{P}}\left(\frac{1}{n}\right)\right) \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

We note that $\mathrm{O}_{\mathbb{P}}(\sqrt{k}/n^{3/2}) = \mathrm{o}_{\mathbb{P}}(1/n)$. We jointly estimate $g^{(2j)}(0)$, $0 \le j \le \ell$, by estimates e_{2j}, $0 \le j \le \ell$, by mimicking the matrix representation of the $U_{(ik)}$'s, replacing $U_{(ik)}$ by its central value, ik/n. Thus,

$$\begin{pmatrix} Y^1_{(k)} & Y^3_{(k)} & \cdots & Y^{2\ell+1}_{(k)} \\ Y^1_{(2k)} & Y^3_{(2k)} & \cdots & Y^{2\ell+1}_{(2k)} \\ \vdots & \vdots & \vdots & \vdots \\ Y^1_{((\ell+1)k)} & Y^3_{((\ell+1)k)} & \cdots & Y^{2\ell+1}_{((\ell+1)k)} \end{pmatrix} \begin{pmatrix} \frac{e_0}{1!} \\ \frac{e_2}{3!} \\ \vdots \\ \frac{e_{2\ell}}{(2\ell+1)!} \end{pmatrix} \stackrel{\text{def}}{=} \begin{pmatrix} \frac{k}{n} \\ \frac{2k}{n} \\ \vdots \\ \frac{(\ell+1)k}{n} \end{pmatrix}.$$

Call the Y-matrix just $\mathbb{Y}$. Then

$$\begin{pmatrix} \frac{e_0}{1!} \\ \frac{e_2}{3!} \\ \vdots \\ \frac{e_{2\ell}}{(2\ell+1)!} \end{pmatrix} = \mathbb{Y}^{-1} \begin{pmatrix} \frac{k}{n} \\ \frac{2k}{n} \\ \vdots \\ \frac{(\ell+1)k}{n} \end{pmatrix}. \tag{6.5}$$

Theorem 6.3. *Assume that f has 2ℓ continuous derivatives in a neighborhood of x, for fixed $\ell \geq 1$, and $f(x) > 0$. Let g be the density of $\|X - x\|$. If $g(0)$ is estimated by e_0, which is defined by (6.5) in terms of the k-th, $2k$-th, ..., and $(\ell+1)$-th nearest neighbor, then the rate of convergence of e_0 towards $g(0)$ is $o(n^{-\frac{2\ell}{4\ell+1}})$.*

Proof. Note that the errors, jointly, are represented as follows:

$$\begin{pmatrix} \frac{g^{(0)}(0)}{1!} - \frac{e_0}{1!} \\ \frac{g^{(2)}(0)}{3!} - \frac{e_2}{3!} \\ \vdots \\ \frac{g^{(2\ell)}(0)}{(2\ell+1)!} - \frac{e_{2\ell}}{(2\ell+1)!} \end{pmatrix} \stackrel{\mathscr{D}}{=} \mathbb{Y}^{-1} \left\{ \begin{pmatrix} U_{(k)} - \frac{k}{n} \\ U_{(2k)} - \frac{2k}{n} \\ \vdots \\ U_{((\ell+1)k)} - \frac{(\ell+1)k}{n} \end{pmatrix} + o_{\mathbb{P}}\left(\left(\tfrac{k}{n}\right)^{2\ell+1}\right) \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \right\}.$$

Our representation for the $U_{(ik)}$-vector then yields

$$\begin{aligned} \text{Error} &\stackrel{\mathscr{D}}{=} \mathbb{Y}^{-1}\, \mathbb{C} \\ &\stackrel{\text{def}}{=} \mathbb{Y}^{-1} \left\{ \frac{\sqrt{k}}{n} \begin{pmatrix} N_1 \\ N_1 + N_2 \\ \vdots \\ N_1 + \cdots + N_{\ell+1} \end{pmatrix} + o_{\mathbb{P}}\left(\frac{\sqrt{k}}{n}\right) \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + o_{\mathbb{P}}\left(\left(\tfrac{k}{n}\right)^{2\ell+1}\right) \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \right\}. \end{aligned}$$

The error $g(0) - e_0$ is a linear combination of the entries of $\mathbb{C} = (C_1, \ldots, C_{\ell+1})^\top$. It is easy to see that

$$\max_{1 \leq i \leq \ell+1} |C_i| = O_{\mathbb{P}}\left(\frac{\sqrt{k}}{n}\right) + o_{\mathbb{P}}\left(\left(\frac{k}{n}\right)^{2\ell+1}\right).$$

Note that $\mathbb{Y}^{-1}$ consists of signed minors divided by $\det(\mathbb{Y})$. In particular, the elements of the first row of $\mathbb{Y}^{-1}$ are signed minors, each of which is in absolute value smaller than

$$\ell!\, Y_{((\ell+1)k)}^{(2\ell+1)+(2\ell-1)+\cdots+3} = \ell!\, Y_{((\ell+1)k)}^{(\ell+1)^2-1} = O_{\mathbb{P}}\left(\left(\frac{k}{n}\right)^{(\ell+1)^2-1}\right).$$

On the other hand, we have

$$\left|\frac{1}{\det(\mathbb{Y})}\right| = \mathrm{O}_{\mathbb{P}}\left(\left(\frac{n}{k}\right)^{(\ell+1)^2}\right) \tag{6.6}$$

(to be shown). Combined with the previous statement, this leads to

$$|g(0) - e_0| = \mathrm{O}_{\mathbb{P}}\left(\frac{n}{k}\right) \times \left[\mathrm{O}_{\mathbb{P}}\left(\frac{\sqrt{k}}{n}\right) + \mathrm{o}_{\mathbb{P}}\left(\left(\frac{k}{n}\right)^{2\ell+1}\right)\right],$$

that is,

$$|g(0) - e_0| = \mathrm{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{k}}\right) + \mathrm{o}_{\mathbb{P}}\left(\left(\frac{k}{n}\right)^{2\ell}\right).$$

This identity shows that the rate of convergence is $\mathrm{o}(n^{-\frac{2\ell}{4\ell+1}})$. Thus, we are done if we can prove (6.6).

Writing $a_i = Y_{(ik)}$, we note that

$$\mathbb{Y} = \begin{pmatrix} a_1^1 & a_1^3 & \cdots & a_1^{2\ell+1} \\ a_2^1 & a_2^3 & \cdots & a_2^{2\ell+1} \\ \vdots & \vdots & \vdots & \vdots \\ a_{\ell+1}^1 & a_{\ell+1}^3 & \cdots & a_{\ell+1}^{2\ell+1} \end{pmatrix}.$$

Its determinant is

$$a_1 \dots a_{\ell+1} \det \begin{pmatrix} 1 & a_1^2 & \cdots & a_1^{2\ell} \\ 1 & a_2^2 & \cdots & a_2^{2\ell} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & a_{\ell+1}^2 & \cdots & a_{\ell+1}^{2\ell} \end{pmatrix},$$

where the latter matrix is a Vandermonde matrix for elements a_i^2, $1 \le i \le \ell + 1$. Its determinant is

$$\begin{aligned} \prod_{1 \le i < j \le \ell+1} (a_j^2 - a_i^2) &\ge \prod_{1 \le i < j \le \ell+1} \left[(a_j - a_i)a_1\right] \\ &\ge \left[\min_{1 \le i < \ell+1} (a_{i+1} - a_i)a_1\right]^{\binom{\ell+1}{2}}. \end{aligned}$$

Thus,

$$\det(\mathbb{Y}) \ge a_1^{\ell+1+\binom{\ell+1}{2}} \left[\min_{1 \le i < \ell+1} (a_{i+1} - a_i)\right]^{\binom{\ell+1}{2}}.$$

Recall that

$$\frac{Y_{(ik)}}{ik/n} \to \frac{1}{g(0)} \quad \text{in probability.}$$

Therefore, with probability $1 - o_{\mathbb{P}}(1)$,

$$\det(\mathbb{Y}) \geq \left(\frac{k}{2g(0)n}\right)^{\ell+1+2\binom{\ell+1}{2}} = \left(\frac{k}{2g(0)n}\right)^{(\ell+1)^2}.$$

We conclude

$$\left|\frac{1}{\det(\mathbb{Y})}\right| = O_{\mathbb{P}}\left(\left(\frac{n}{k}\right)^{(\ell+1)^2}\right),$$

as required. □

Remark 6.6. It is an easy exercise to show that the errors of the estimates of the derivatives have worst rates of convergence. For $g^{(2)}(0) - e_2$ we obtain $o(n^{-\frac{2(\ell-1)}{4\ell+1}})$, and in general, for $g^{(2j)}(0) - e_{2j}$, we have the rate $o(n^{-\frac{2(\ell-j)}{4\ell+1}})$, $0 \leq j \leq \ell$, that is, for $g^{(2\ell)}(0) - e_{2\ell}$, the rate of convergence is $o(1)$ only. □

Chapter 7
Entropy estimation

7.1 Differential entropy

Differential entropy, or continuous entropy, is a concept in information theory related to the classical (Shannon) entropy (Shannon, 1948). For a random variable with density f on $\mathbb{R}^d$, it is defined by

$$\mathscr{E}(f) = -\int_{\mathbb{R}^d} f(\mathbf{x}) \log f(\mathbf{x}) \mathrm{d}\mathbf{x}, \tag{7.1}$$

when this integral exists (with the convention $0 \log 0 = 0$). If $d = 1$ and f is the uniform density on $[0, a]$, $a > 0$, then its differential entropy is

$$\mathscr{E}(f) = -\int_0^a \frac{1}{a} \log \frac{1}{a} \mathrm{d}x = \log a.$$

We see that for $a < 1$, $\log a < 0$, so that $\mathscr{E}(f)$ can be negative. The standard exponential has $\mathscr{E}(f) = 1$, and the standard Gaussian has $\mathscr{E}(f) = \log \sqrt{2\pi e}$, to give a few examples.

Since $u \log u \geq -1/e$ for all $u \geq 0$, the integral of the positive part of the function $-f \log f$ is finite as soon as f has compact support. Thus, for finite support densities, one can without any worries use (7.1), even when the integral diverges, for in that case, $\mathscr{E}(f) = -\infty$. In fact, we have

$$\mathscr{E}(f) \leq \frac{1}{e} \lambda \left(\mathrm{supp}(f) \right),$$

where λ denotes the Lebesgue measure on $\mathbb{R}^d$. The situation $\mathscr{E}(f) = -\infty$ can only occur for unbounded densities. Indeed, if $f \leq \|f\|_\infty < \infty$, then

$$\mathscr{E}(f) \geq -\log \|f\|_\infty.$$

G. Biau, L. Devroye, *Lectures on the Nearest Neighbor Method*,
Springer Series in the Data Sciences, DOI 10.1007/978-3-319-25388-6_7

An example of a density on $\mathbb{R}$ with $\mathscr{E}(f) = -\infty$ is

$$f(x) = \frac{\log 2}{x \log^2 x}, \quad 0 < x < 1/2.$$

There are other continuous entropies, most prominently the continuous version of Rényi's entropy (Rényi, 1961),

$$\mathscr{E}_q(f) = \frac{1}{1-q} \log \int_{\mathbb{R}^d} f^q(\mathbf{x}) d\mathbf{x}, \quad q \neq 1.$$

The quadratic entropy is $\mathscr{E}_2(f) = -\log \int_{\mathbb{R}^d} f^2(\mathbf{x}) d\mathbf{x}$. It is a good exercise to prove that, under appropriate conditions, $\lim_{q \to 1} \mathscr{E}_q(f) = \mathscr{E}(f)$. We will not be concerned with $\mathscr{E}_q(f)$ in this book, except perhaps for the observation that most of this chapter can be recast, with minor modifications, for $\mathscr{E}_q(f)$—see Section 7.5.

Entropy and related concepts play an important role in fields as diverse as physical sciences, source coding, texture classification, spectroscopy, image analysis, and signal processing, just to name a few (see, e.g., Leonenko et al., 2008, and the differential entropy handbook by Michalowicz et al., 2014—the latter monograph also provides a comprehensive collection of the entropy of most frequently used probability densities). The basic features of differential entropy are described in Cover and Thomas (2006). Of importance are its extremal properties: for example, if the density f is concentrated on the unit interval $[0, 1]$, then the differential entropy is maximal if and only if f is uniform on $[0, 1]$, and then $\mathscr{E}(f) = 0$; similarly, if the density has fixed variance, then $\mathscr{E}(f)$ is maximized by the Gaussian density.

Our objective in this chapter is to estimate $\mathscr{E}(f)$. This problem has various applications in goodness-of-fit testing, parameter estimation, quantization theory, and econometrics. There are numerous estimates in the literature, mostly listed in the survey by Beirlant et al. (1997). One of these, based on the nearest neighbor method, was proposed by Kozachenko and Leonenko (1987). This method, while natural and deceptively simple, has defied a thorough analysis. We hope to shed some light on this in the present chapter.

Given a generic density estimate f_n, the differential entropy can be estimated by one of two kinds of plug-in estimates,

$$-\int_{\mathbb{R}^d} f_n(\mathbf{x}) \log f_n(\mathbf{x}) d\mathbf{x},$$

or

$$-\frac{1}{n} \sum_{i=1}^{n} \log f_n(\mathbf{X}_i).$$

In most cases, the density estimate f_n itself is consistent. However, the rate of convergence of the plug-in estimate is limited by the rate of convergence of $\mathbb{E} f_n$ to f.

Kozachenko and Leonenko (1987) noted that if one uses an inconsistent estimate for plug-in, i.e., the 1-nearest neighbor estimate, then one can in fact improve the rate of convergence of the plug-in estimate.

7.2 The Kozachenko-Leonenko estimate

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ ($n \geq 3$) be i.i.d. random variables with density f on $\mathbb{R}^d$. Let R_i denote the distance between $\mathbf{X}_i$ and its nearest neighbor among $\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_n$:

$$R_i = \min_{j \neq i} \|\mathbf{X}_j - \mathbf{X}_i\|.$$

Let $B(\mathbf{x}, \rho)$, as always, denote the closed ball centered at $\mathbf{x}$ of radius ρ, and let μ be the probability measure corresponding to f.

Following Kozachenko and Leonenko (1987), we consider the estimate

$$\ell_n = \frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{1}{n\lambda\left(B(\mathbf{X}_i, R_i)\right)} \right),$$

where the lowercase ℓ is used as a mnemonic device. The rationale behind this definition is that

$$\ell_n = \log\left(\frac{n-1}{n}\right) + \frac{1}{n} \sum_{i=1}^{n} \log\left(f_{n-1}^{(-i)}(\mathbf{X}_i)\right),$$

where $f_{n-1}^{(-i)}$ is the 1-nearest neighbor density estimate of f based on the $n-1$ observations $\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_n$. Thus, under appropriate conditions, ℓ_n should not be asymptotically too far from

$$\mathbb{E} \log f(\mathbf{X}_1) = \int_{\mathbb{R}^d} f(\mathbf{x}) \log f(\mathbf{x}) \mathrm{d}\mathbf{x} = -\mathscr{E}(f).$$

This estimate is sensitive to large tails, as the following example explains.

Lemma 7.1. *There exists a function f on $\mathbb{R}$ with*

$$\int_{\mathbb{R}} f^p(x) \log^q f(x) \mathrm{d}x = 0, \quad \text{all } p, q > 0,$$

for which $\mathbb{E}\ell_n = -\infty$ for all $n \geq 2$.

Proof. Define

$$f(x) = \sum_{j=1}^{\infty} \mathbb{1}_{[2^{2^j}, 2^{2^j} + \frac{1}{j(j+1)}]}.$$

Observe that

$$\ell_n = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{1}{n\lambda\left(B(X_i, R_i)\right)}\right)$$
$$= \frac{1}{n}\sum_{i=1}^{n} \log^+\left(\frac{1}{n\lambda\left(B(X_i, R_i)\right)}\right) - \frac{1}{n}\sum_{i=1}^{n} \log^-\left(\frac{1}{n\lambda\left(B(X_i, R_i)\right)}\right). \tag{7.2}$$

It turns out that, for all n,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \log^+\left(\frac{1}{n\lambda\left(B(X_i, R_i)\right)}\right)\right] < \infty,$$

whereas

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \log^-\left(\frac{1}{n\lambda\left(B(X_i, R_i)\right)}\right)\right] = \infty.$$

To prove the first claim, denote by $X_{(1)} \le \cdots \le X_{(n)}$ the order statistics for the sample, and observe that

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \log^+\left(\frac{1}{n\lambda\left(B(X_i, R_i)\right)}\right)\right] \le \mathbb{E}\log^+\left(\frac{1}{\min_{1<i\le n}(X_{(i)} - X_{(i-1)})}\right)$$

(since the function $\log^+$ is increasing).

The minimum $\min_{1<i\le n}(X_{(i)} - X_{(i-1)})$ is stochastically minimized by the same quantity if f were the uniform distribution on $[0, 1]$—just squish the intervals in the definition of f together. That quantity behaves like the minimal uniform spacing, which is asymptotically distributed as E/n^2, where E is standard exponential. To be precise, denote by $U_{(1)} \le \cdots \le U_{(n)}$ uniform $[0, 1]$ order statistics, and let the associated spacings S_i be

$$S_i = U_{(i)} - U_{(i-1)}, \quad 1 \le i \le n+1,$$

where, by convention, $U_{(0)} = 0$ and $U_{(n+1)} = 1$. Recall that, by Theorem 1.3,

$$(S_1, \ldots, S_{n+1}) \overset{\mathscr{D}}{=} \left(\frac{E_1}{\sum_{i=1}^{n+1} E_i}, \ldots, \frac{E_{n+1}}{\sum_{i=1}^{n+1} E_i}\right),$$

where $E_1, \dots, E_{n+1}$ are independent standard exponential random variables. Therefore,

$$\begin{aligned}\mathbb{E}\log^+\left(\frac{1}{\min_{1<i\le n}(X_{(i)}-X_{(i-1)})}\right) &\le \mathbb{E}\log\left(\frac{1}{\min_{1<i\le n}(U_{(i)}-U_{(i-1)})}\right)\\ &\le \mathbb{E}\log\left(\frac{1}{\min_{1\le i\le n+1} S_i}\right)\\ &= \mathbb{E}\log\left(\frac{G_{n+1}}{\min_{1\le i\le n+1} E_i}\right),\end{aligned}$$

where $G_{n+1} = \sum_{i=1}^{n+1} E_i$ is Gamma$(n+1)$ distributed. Therefore,

$$\begin{aligned}\mathbb{E}\log^+\left(\frac{1}{\min_{1<i\le n}(X_{(i)}-X_{(i-1)})}\right) &\le \mathbb{E}\log G_{n+1} - \mathbb{E}\log\left(\min_{1\le i\le n+1} E_i\right)\\ &= \mathbb{E}\log G_{n+1} - \mathbb{E}\log\left(\frac{E}{n+1}\right)\\ &\quad \text{(where } E \text{ is standard exponential)}\\ &\le \log\left(\mathbb{E}G_{n+1}\right) - \mathbb{E}\log E + \log(n+1)\\ &\quad \text{(by Jensen's inequality)}\\ &= 2\log(n+1) + \gamma, v\end{aligned}$$

where $\gamma = -\mathbb{E}\log E = -\int_0^\infty e^{-t}\log t\,\mathrm{d}t = 0.577215664901532\dots$ is the Euler-Mascheroni constant. Thus, the first term in (7.2) satisfies

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\log^+\left(\frac{1}{n\lambda\left(B(X_i,R_i)\right)}\right)\right] < \infty.$$

Next, we show that the expectation of the second term of (7.2) is infinite. To this aim, we write, using the fact that the function $\log^-$ is decreasing,

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\log^-\left(\frac{1}{n\lambda\left(B(X_i,R_i)\right)}\right)\right] &\ge \frac{1}{n}\mathbb{E}\log^-\left(\frac{1}{X_{(n)}-X_{(n-1)}}\right)\\ &\ge \frac{1}{n}\mathbb{E}\log^-\left(\frac{1}{X_{(n)}-X_{(n-1)}}\right)\mathbb{1}_A\end{aligned}$$

for any event A. Let

$$A_j = \left[X_n = X_{(n)}, X_n \in j\text{-th interval}, X_1, \dots, X_{n-1} \in \text{intervals of index} < j\right]$$

and

$$A = \bigcup_{j=2n}^{\infty} A_j.$$

For $j \geq 2n$, writing $\beta_j = 2^{2^j} - 2^{2^{j-1}} - \frac{1}{(j-1)j}$ for the separating gap between the $(j-1)$-st and j-th intervals, we have

$$\begin{aligned}\mathbb{E}\log^-\left(\frac{1}{X_{(n)} - X_{(n-1)}}\right)\mathbb{1}_{A_j} &\geq (\log\beta_j)\mathbb{P}\{A_j\} \\ &= \frac{\log\beta_j}{nj(j+1)}\left(1-\frac{1}{j}\right)^{n-1} \\ &\geq \frac{\log\beta_j}{2nj(j+1)}.\end{aligned}$$

It is easy to see that $\sum_{j\geq 2n}\frac{\log\beta_j}{j(j+1)} = \infty$, so that, for every n,

$$\mathbb{E}\log^-\left(\frac{1}{X_{(n)} - X_{(n-1)}}\right)\mathbb{1}_A = \infty. \qquad \square$$

Thus, to avoid annoying conditions, we will simply assume that f has compact support, which in turn implies that

$$\mathscr{E}(f) = -\int_{\mathbb{R}^d} f(\mathbf{x})\log f(\mathbf{x})\mathrm{d}\mathbf{x}$$

is properly defined—it is either finite or $-\infty$. We leave it as an easy exercise to show that in this case, for all integers $p \geq 1$,

$$\int_{\mathbb{R}^d} f(\mathbf{x})\log^p f(\mathbf{x})\mathrm{d}\mathbf{x} < \infty \Leftrightarrow \int_{\mathbb{R}^d} f(\mathbf{x})\log^p(f(\mathbf{x})+1)\mathrm{d}\mathbf{x} < \infty.$$

The expression on the right-hand side is easier to handle since the integrand is always nonnegative. Our main result is as follows:

Theorem 7.1. *Assume that f has compact support and that $\int_{\mathbb{R}^d} f(\mathbf{x})\log(f(\mathbf{x})+1)\mathrm{d}\mathbf{x} < \infty$. Then*

$$\mathbb{E}\ell_n \to -\mathscr{E}(f) + \gamma,$$

where $\gamma = 0.577215664901532\ldots$ is the Euler-Mascheroni constant. Furthermore, if $\int_{\mathbb{R}^d} f(\mathbf{x})\log^2(f(\mathbf{x})+1)\mathrm{d}\mathbf{x} < \infty$, then

$$\mathbb{V}\ell_n \leq \frac{c}{n},$$

where c is a constant that depends upon f. Finally,

$$\ell_n \to -\mathscr{E}(f) + \gamma \quad \textit{in probability}.$$

Letting $\bar{\ell}_n = -\ell_n + \gamma$, we also conclude from Theorem 7.1 that $\mathbb{E}|\bar{\ell}_n - \mathscr{E}(f)|^2 \to 0$ as $n \to \infty$ (mean squared consistency). In the univariate case, Tsybakov and van der Meulen (1996) established the mean squared $O(1/\sqrt{n})$ rate of convergence of a truncated version of ℓ_n, for a class of densities with unbounded support and exponentially decreasing tails, such as the Gaussian density. Some analysis for bounded f is to be found in Leonenko et al. (2008). It should be noted that Theorem 7.1 does not settle the deeper question if, for all compact support densities f, under the unique condition $\int_{\mathbb{R}^d} f(\mathbf{x}) \log(f(\mathbf{x}) + 1) d\mathbf{x} < \infty$, $\ell_n \to -\mathscr{E}(f) + \gamma$ in probability.

Proof (Theorem 7.1). Define the following maximal functions:

$$f^\star(\mathbf{x}) = \sup_{\rho > 0} \left(\frac{\mu\left(B(\mathbf{x}, \rho)\right)}{\lambda\left(B(\mathbf{x}, \rho)\right)} \right) \quad \text{and} \quad g^\star(\mathbf{x}) = \sup_{\rho > 0} \left(\frac{\lambda\left(B(\mathbf{x}, \rho)\right)}{\mu\left(B(\mathbf{x}, \rho)\right)} \right).$$

Observe that

$$\begin{aligned}
\mathbb{E}\ell_n &= \mathbb{E} \log \left(\frac{1}{n\lambda\left(B(\mathbf{X}_1, R_1)\right)} \right) \\
&= \mathbb{E} \log \left(\frac{n\mu\left(B(\mathbf{X}_1, R_1)\right)}{n\lambda\left(B(\mathbf{X}_1, R_1)\right)} \right) + \mathbb{E} \log \left(\frac{1}{n\mu\left(B(\mathbf{X}_1, R_1)\right)} \right) \\
&\quad \text{(which is allowed since both expected values are finite as we will soon see)} \\
&\stackrel{\text{def}}{=} \mathbf{I} + \mathbf{II}.
\end{aligned}$$

Note that

$$-\log(g^\star(\mathbf{X}_1) + 1) \le \log \left(\frac{\mu\left(B(\mathbf{X}_1, R_1)\right)}{\lambda\left(B(\mathbf{X}_1, R_1)\right)} \right) \le \log(f^\star(\mathbf{X}_1) + 1),$$

and thus that

$$-\mathbb{E} \log(g^\star(\mathbf{X}) + 1) \le \mathbf{I} \le \mathbb{E} \log(f^\star(\mathbf{X}) + 1)$$

where $\mathbf{X} \stackrel{\mathscr{D}}{=} \mathbf{X}_1$. Assume that we can verify that

$$\mathbb{E} \log(f^\star(\mathbf{X}) + 1) = \int_{\mathbb{R}^d} f(\mathbf{x}) \log(f^\star(\mathbf{x}) + 1) d\mathbf{x} < \infty \tag{7.3}$$

and that

$$\mathbb{E} \log(g^\star(\mathbf{X}) + 1) = \int_{\mathbb{R}^d} f(\mathbf{x}) \log(g^\star(\mathbf{x}) + 1) d\mathbf{x} < \infty. \tag{7.4}$$

Then, by the Lebesgue dominated convergence theorem, we conclude that

$$\mathbf{I} \to \mathbb{E} \log f(\mathbf{X}) = -\mathscr{E}(f)$$

if

$$\frac{\mu\left(B(\mathbf{X}_1, R_1)\right)}{\lambda\left(B(\mathbf{X}_1, R_1)\right)} \to f(\mathbf{X}_1) \quad \text{in probability.}$$

This follows trivially since $R_1 \to 0$ almost surely (by Lemma 2.3) and μ-almost all $\mathbf{x}$ are Lebesgue points of f. Finally,

$$n\mu\left(B(\mathbf{X}_1, R_1)\right) \stackrel{\mathscr{D}}{=} nU_{(1)},$$

where $U_{(1)} \leq \cdots \leq U_{(n)}$ are uniform $[0, 1]$ order statistics (see Chapter 1). One can easily check (see Example 20.1 in the Appendix) that, for E a standard exponential random variable,

$$\mathbb{E} \log\left(\frac{1}{nU_{(1)}}\right) \to \mathbb{E} \log\left(\frac{1}{E}\right) = -\int_0^\infty e^{-t} \log t \, \mathrm{d}t = \gamma.$$

To verify (7.3) and (7.4), we use the results on maximal functions given in Section 20.8 of the Appendix. As f has compact support, we know by Lemma 20.8 that the condition $\int_{\mathbb{R}^d} f(\mathbf{x}) \log(f(\mathbf{x}) + 1)\mathrm{d}\mathbf{x} < \infty$ implies $\int_{\mathbb{R}^d} f(\mathbf{x}) \log(f^\star(\mathbf{x}) + 1)\mathrm{d}\mathbf{x} < \infty$. Next, denote by K a positive constant such that the support of f is included in $[-K, K]^d$. Then, similarly, since $R_1 \leq 2K\sqrt{d}$, we can replace λ in the definition of $g^\star$ by the Lebesgue measure truncated to the ball $B(\mathbf{0}, 3K\sqrt{d})$. By Lemma 20.7, for $t > 0$,

$$\mu\Big(\big\{\mathbf{x} \in \mathbb{R}^d : g^\star(\mathbf{x}) > t\big\}\Big) \leq \frac{c}{t}\lambda\Big(B\big(\mathbf{0}, 3K\sqrt{d}\big)\Big)$$

for a universal constant c. Note that

$$\begin{aligned}
\mathbb{E} \log(g^\star(\mathbf{X}) + 1) &= \int_0^\infty \mu\Big(\big\{\mathbf{x} \in \mathbb{R}^d : \log(g^\star(\mathbf{x}) + 1) > t\big\}\Big)\mathrm{d}t \\
&= \int_0^\infty \mu\Big(\big\{\mathbf{x} \in \mathbb{R}^d : g^\star(\mathbf{x}) + 1 > e^t\big\}\Big)\mathrm{d}t \\
&\leq \int_0^\infty \mu\Big(\Big\{\mathbf{x} \in \mathbb{R}^d : g^\star(\mathbf{x}) > \frac{e^t}{2}\Big\}\Big)\mathrm{d}t + \int_0^{\log 2} \mathrm{d}t \\
&\leq 2c\lambda\Big(B\big(\mathbf{0}, 3K\sqrt{d}\big)\Big) \int_0^\infty e^{-t}\mathrm{d}t + \log 2 \\
&= 2c\lambda\Big(B\big(\mathbf{0}, 3K\sqrt{d}\big)\Big) + \log 2,
\end{aligned}$$

so that (7.4) follows. This concludes the proof of the first part of the theorem. The third part follows from the first two by Chebyshev's inequality. The second part is shown in the next section. □

7.3 The variance of the Kozachenko-Leonenko estimate

The purpose of this section is to illustrate the power of some simple bounding methods—the application to the estimate at hand is only of secondary importance. We will apply the Efron-Stein inequality (Appendix, Theorem 20.10) to show that under the conditions of Theorem 7.1, $\mathbb{V}\ell_n = \mathrm{O}(1/n)$. To do so, a second independent sample is needed. We let $\mathbf{X}_1, \ldots, \mathbf{X}_n, \mathbf{X}'_1, \ldots, \mathbf{X}'_n$ be i.i.d., all distributed as $\mathbf{X}$, with density f on $\mathbb{R}^d$. Let ℓ_n be the Kozachenko-Leonenko estimate, based on $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Consider $\mathbf{X}_2, \ldots, \mathbf{X}_n$, and let $R_2(1), \ldots, R_n(1)$ be the nearest neighbor distances

$$R_i(1) = \min_{j>1, j\neq i} \|\mathbf{X}_j - \mathbf{X}_i\|.$$

Define

$$\ell_{n-1} = \frac{1}{n}\sum_{i=2}^{n} \log\left(\frac{1}{n\lambda\left(B\left(\mathbf{X}_i, R_i(1)\right)\right)}\right).$$

Finally, let $R'_1, \ldots, R'_n$ be the nearest neighbor distances based on $\mathbf{X}'_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$, and set

$$\ell'_n = \frac{1}{n}\left[\log\left(\frac{1}{n\lambda\left(B(\mathbf{X}'_1, R'_1)\right)}\right) + \sum_{i=2}^{n} \log\left(\frac{1}{n\lambda\left(B(\mathbf{X}_i, R'_i)\right)}\right)\right].$$

We have, by the Efron-Stein inequality,

$$\begin{aligned}\mathbb{V}\ell_n \le \frac{n}{2}\mathbb{E}|\ell_n - \ell'_n|^2 &\le n\mathbb{E}\left[|\ell_n - \ell_{n-1}|^2 + |\ell_{n-1} - \ell'_n|^2\right] \\ &= 2n\mathbb{E}|\ell_n - \ell_{n-1}|^2.\end{aligned}$$

Let $\tau_i \in \{1, \ldots, n\}$ denote the index of the nearest neighbor of $\mathbf{X}_i$, so that $R_i = \|\mathbf{X}_{\tau_i} - \mathbf{X}_i\|$. We have

$$\begin{aligned}n(\ell_n - \ell_{n-1}) = &\log\left(\frac{1}{n\lambda\left(B(\mathbf{X}_1, R_1)\right)}\right) \\ &+ \sum_{i=2}^{n} \mathbb{1}_{[\tau_i=1]}\left[\log\left(\frac{1}{n\lambda\left(B(\mathbf{X}_i, R_i)\right)}\right) - \log\left(\frac{1}{n\lambda\left(B\left(\mathbf{X}_i, R_i(1)\right)\right)}\right)\right]\end{aligned}$$

since only those data points $\mathbf{X}_i$, $i \geq 2$, with $\tau_i = 1$ can have $R_i(1) \neq R_i$ (and thus, $R_i(1) > R_i$). Squaring and using the c_r-inequality (Proposition 20.1 in the Appendix) shows that

$$\begin{aligned} & n^2(\ell_n - \ell_{n-1})^2 \\ & \leq \left(1 + \sum_{i=2}^{n} \mathbb{1}_{[\tau_i=1]}\right) \times \left[\log^2 \left(\frac{1}{n\lambda\left(B(\mathbf{X}_1, R_1)\right)} \right) \right. \\ & \left. +2 \sum_{i=2}^{n} \mathbb{1}_{[\tau_i=1]} \log^2 \left(\frac{1}{n\lambda\left(B(\mathbf{X}_i, R_i)\right)} \right) + 2 \sum_{i=2}^{n} \mathbb{1}_{[\tau_i=1]} \log^2 \left(\frac{1}{n\lambda\left(B\left(\mathbf{X}_i, R_1(1)\right)\right)} \right) \right]. \end{aligned} \tag{7.5}$$

According to Lemma 20.6, since $\mathbf{X}$ has a density, with probability one,

$$1 + \sum_{i=2}^{n} \mathbb{1}_{[\tau_i=1]} \leq \gamma_d,$$

where γ_d is the minimal number of cones of angle $\pi/6$ that cover $\mathbb{R}^d$. The first term on the right of (7.5) is bounded as follows in expectation:

$$\begin{aligned} & \mathbb{E} \log^2 \left(\frac{1}{n\lambda\left(B(\mathbf{X}_1, R_1)\right)} \right) \\ & = \mathbb{E} \left[\left(\log \left(\frac{n\mu\left(B(\mathbf{X}_1, R_1)\right)}{n\lambda\left(B(\mathbf{X}_1, R_1)\right)} \right) + \log \left(\frac{1}{n\mu\left(B(\mathbf{X}_1, R_1)\right)} \right) \right)^2 \right] \\ & \leq 2\mathbb{E} \log^2 \left(\frac{\mu\left(B(\mathbf{X}_1, R_1)\right)}{\lambda\left(B(\mathbf{X}_1, R_1)\right)} \right) + 2\mathbb{E} \log^2 \left(\frac{1}{n\mu\left(B(\mathbf{X}_1, R_1)\right)} \right). \end{aligned} \tag{7.6}$$

Arguing as in the proof of the first part of Theorem 7.1, and using the closeness of $n\mu(B(\mathbf{X}_1, R_1)) \stackrel{\mathscr{D}}{=} nU_{(1)}$ to a standard exponential random variable E, one checks (Example 20.1 in the Appendix) that the last term in (7.6) tends to $2\mathbb{E} \log^2 E = 2(\gamma^2 + \frac{\pi^2}{6})$. The first term in (7.6) is not larger than

$$2\mathbb{E} \log^2(f^\star(\mathbf{X}) + 1) + 2\mathbb{E} \log^2(g^\star(\mathbf{X}) + 1),$$

where $f^\star$ and $g^\star$ are as in the previous section. Using the arguments of that proof together with Lemma 20.8, we see that $\mathbb{E} \log^2(f^\star(\mathbf{X}) + 1) < \infty$ if $\int_{\mathbb{R}^d} f(\mathbf{x}) \log^2(f(\mathbf{x}) + 1) d\mathbf{x} < \infty$, and that $\mathbb{E} \log^2(g^\star(\mathbf{X}) + 1) < \infty$, for all densities f with compact support.

The middle term of (7.5) has expected value

$$\begin{aligned} & 2(n-1)\mathbb{E} \left[\mathbb{1}_{[\tau_2=1]} \log^2 \left(\frac{1}{n\lambda\left(B(\mathbf{X}_2, R_2)\right)} \right) \right] \\ & = 2(n-1) \times \frac{1}{n-1} \mathbb{E} \log^2 \left(\frac{1}{n\lambda\left(B(\mathbf{X}_2, R_2)\right)} \right) \end{aligned}$$

since $(\mathbf{X}_2, R_2)$ is independent of $\mathbb{1}_{[\tau_2=1]}$. Therefore, the middle term is in expectation twice the expectation of the first one.

Finally, we deal with the last term of (7.5). Its expectation is

$$2(n-1)\mathbb{E}\left[\mathbb{1}_{[\tau_2=1]}\log^2\left(\frac{1}{n\lambda\left(B(\mathbf{X}_2, R_2^\star)\right)}\right)\right]$$

(where $R_2^\star$ is the second nearest neighbor distance of $\mathbf{X}_2$ among $\mathbf{X}_1, \mathbf{X}_3, \ldots, \mathbf{X}_n$)

$$= 2(n-1) \times \frac{1}{n-1}\mathbb{E}\log^2\left(\frac{1}{n\lambda\left(B(\mathbf{X}_2, R_2^\star)\right)}\right)$$

(since $(\mathbf{X}_2, R_2^\star)$ is independent of $\mathbb{1}_{[\tau_2=1]}$).

We bound the last expected value by

$$2\mathbb{E}\log^2\left(\frac{\mu\left(B(\mathbf{X}_2, R_2^\star)\right)}{\lambda\left(B(\mathbf{X}_2, R_2^\star)\right)}\right) + 2\mathbb{E}\log^2\left(\frac{1}{n\mu\left(B(\mathbf{X}_2, R_2^\star)\right)}\right). \tag{7.7}$$

Since $n\mu(B(\mathbf{X}_2, R_2^\star)) \stackrel{\mathcal{D}}{=} nU_{(2)}$, the last term can be shown to converge to $2\mathbb{E}\log^2 G_2$ (where G_2 is a Gamma(2) random variable—see Example 20.1 in the Appendix), which is a finite constant. Finally, the first term of (7.7) is not larger than

$$2\mathbb{E}\log^2(f^\star(\mathbf{X}) + 1) + 2\mathbb{E}\log^2(g^\star(\mathbf{X}) + 1) < \infty.$$

This concludes the proof of Theorem 7.1. □

7.4 Study of the bias

The estimation of $\mathscr{E}(f)$ or $\int_{\mathbb{R}^d} f^\alpha(\mathbf{x})\mathrm{d}\mathbf{x}$ from i.i.d. observations has many applications, and so, the rate with which these functionals can be estimated is crucial. A thorough introduction and discussion can be found, e.g., in the work of Birgé and Massart (1995). Considering for example classes of densities bounded away from 0 and ∞ on $[0, 1]$ and satisfying the Lipschitz condition

$$\left|f(x) - f(x')\right| \leq c|x - x'|^\beta, \tag{7.8}$$

for fixed $c > 0$, $\beta \in (0, 1]$, they showed that most functionals cannot be estimated at a rate better than

$$\begin{cases} n^{-\frac{4\beta}{4\beta+1}} & \text{if } \beta < \frac{1}{4} \\ \frac{1}{\sqrt{n}} & \text{if } \frac{1}{4} \leq \beta \leq 1. \end{cases}$$

For $\int_{\mathbb{R}^d} f^2(\mathbf{x})d\mathbf{x}$, this was first observed by Bickel and Ritov (1988), who additionally provided estimates with matching rates of convergence (see also Laurent, 1996). For $\mathscr{E}(f)$, Donoho (1988) discusses the situation.

For the Kozachenko-Leonenko estimate ℓ_n, we know (by Theorem 7.1) that $\mathbb{V}\ell_n = \mathrm{O}(1/n)$, so the bias of the error, $\mathbb{E}\ell_n + \mathscr{E}(f) - \gamma$, is of interest. The phenomenon described above will be rediscovered: for sufficiently smooth f on $\mathbb{R}$, the rate $1/\sqrt{n}$ is achievable, while for unsmooth f, the bias dominates and makes the rate much slower. We provide a quick computation for the Lipschitz class given by (7.8), assuming for convenience that f is supported on $[0, 1]$.

The precise study of the bias will not be done here, as it is not essential for a better understanding. For a simple class of densities, we offer a quick-and-dirty upper bound.

Theorem 7.2. *Let $\alpha > 0$, $\beta \in (0, 1]$, $c > 0$ be given constants, and let the density f be supported on $[0, 1]$, with $\inf_{x\in[0,1]} f(x) = \alpha$ and, for all $x, x' \in [0, 1]$,*

$$\left|f(x) - f(x')\right| \le c|x - x'|^\beta.$$

Then

$$\mathbb{E}\ell_n = -\mathscr{E}(f) + \gamma + \mathrm{O}\left(\frac{1}{n^\beta}\right),$$

where $\gamma = 0.577215664901532\ldots$ is the Euler-Mascheroni constant.

Corollary 7.1. *Under the conditions of Theorem 7.2,*

$$\mathbb{E}\left|\ell_n + \mathscr{E}(f) - \gamma\right| = \begin{cases} \mathrm{O}\left(\frac{1}{n^\beta}\right) & \text{if } \beta < \frac{1}{2} \\ \mathrm{O}\left(\frac{1}{\sqrt{n}}\right) & \text{if } \frac{1}{2} \le \beta \le 1. \end{cases}$$

In other words, we rediscover the phenomenon described in Birgé and Massart (1995).

Proof (Theorem 7.2). We look at $\mathbb{E}\ell_n + \mathscr{E}(f) - \gamma$. Note that, if $R(x) = \min_{1\le i\le n} |X_i - x|$, $x \in \mathbb{R}$, then

$$\begin{aligned}
\mathbb{E}\ell_{n+1} &= \mathbb{E}\log\left(\frac{1}{2(n+1)R(X)}\right) \\
&\quad \text{(where } X \text{ is independent of } X_1, \ldots, X_n) \\
&= \mathbb{E}\log\left(\frac{\mu\left(B\left(X, R(X)\right)\right)}{2R(X)f(X)}\right) + \mathbb{E}\log f(X) + \mathbb{E}\log\left(\frac{1}{(n+1)\mu\left(B\left(X, R(X)\right)\right)}\right) \\
&\stackrel{\text{def}}{=} \mathbf{I} + \mathbf{II} + \mathbf{III}.
\end{aligned}$$

Clearly, $\mathbf{II} = -\mathscr{E}(f)$, which is finite since f is Lipschitz, hence bounded, and f is supported on $[0, 1]$. Next, recall (Chapter 1) that $(n+1)\mu(B(X, R(X)))$ is $(n+1)\text{Beta}(1, n)$ distributed, i.e., it has density

$$\frac{n}{n+1}\left(1-\frac{x}{n+1}\right)^{n-1}, \quad 0 \le x \le n+1.$$

For E a standard exponential random variable, we evaluate

$$\mathbf{IV} = \left|\mathbb{E}\log\left((n+1)\text{Beta}(1,n)\right) - \mathbb{E}\log E\right|$$

by using

$$e^{-x\frac{n-1}{n+1-x}} \le \left(1-\frac{x}{n+1}\right)^{n-1} \le e^{-\frac{n-1}{n+1}x}, \quad 0 \le x < n+1.$$

We have

$$\begin{aligned}
\mathbf{IV} &\le \left|\int_0^{n+1} \log x \left(\left(1-\frac{x}{n+1}\right)^{n-1} - e^{-\frac{n-1}{n+1}x}\right) \mathrm{d}x\right| + \mathrm{O}\left(\frac{1}{n}\right) \\
&\le \int_{2\log n}^{\infty} (\log x) e^{-\frac{n-1}{n+1}x} \mathrm{d}x + \int_0^{2\log n} |\log x| e^{-\frac{n-1}{n+1}x}\left(1-e^{-\frac{\frac{n-1}{n+1}x^2}{n+1-x}}\right)\mathrm{d}x + \mathrm{O}\left(\frac{1}{n}\right) \\
&\le \mathrm{O}\left(\frac{\log n}{n^2}\right) + \frac{1}{n+1-2\log n}\int_0^{\infty} |\log x| e^{-\frac{n-1}{n+1}x} x^2 \mathrm{d}x + \mathrm{O}\left(\frac{1}{n}\right) \\
&= \mathrm{O}\left(\frac{1}{n}\right).
\end{aligned}$$

Thus, $\mathbf{III} = \mathrm{O}(1/n) - \mathbb{E}\log E = \mathrm{O}(1/n) + \gamma$.

It remains to show that $\mathbf{I} = \mathrm{O}(1/n^{\beta})$. To this aim, note that, for any $\rho > 0$,

$$2\rho f(X) - \frac{2c\rho^{\beta+1}}{\beta+1} \le \mu\left(B(X,\rho)\right) \le 2\rho f(X) + \frac{2c\rho^{\beta+1}}{\beta+1},$$

and that

$$\mu\left(B(X,\rho)\right) \ge 2\alpha\rho.$$

Thus,

$$\begin{aligned}
\mathbf{I} &\le \mathbb{E}\log\left(1 + \frac{2cR^{\beta+1}(X)}{2(\beta+1)R(X)f(X)}\right) \\
&\le \frac{c}{(\beta+1)\alpha}\mathbb{E}R^{\beta}(X)
\end{aligned}$$

$$\begin{aligned}
&\text{(since } \log(1+u) \le u \text{ for all } u \ge 0) \\
&\le \frac{c}{(\beta+1)\alpha} \mathbb{E}^{\beta} R(X) \\
&\text{(by Jensen's inequality).}
\end{aligned}$$

Therefore, recalling the notation $R_i = \min_{j \neq i} |X_i - X_j|$, we are led to

$$\begin{aligned}
\mathbf{I} &\le \frac{c}{(\beta+1)\alpha} \mathbb{E}^{\beta} \left[\frac{1}{n+1} \sum_{i=1}^{n+1} R_i \right] \\
&\le \frac{c}{(\beta+1)\alpha} \times \left(\frac{2}{n+1} \right)^{\beta}.
\end{aligned}$$

Finally, define the event

$$A = \left[cR^{\beta}(X) < \frac{(\beta+1)\alpha}{2} \right].$$

On A, we have

$$(\beta+1) f(X) - cR^{\beta}(X) \ge \frac{(\beta+1)\alpha}{2}.$$

Thus,

$$\begin{aligned}
-\mathbf{I} &\le \mathbb{E} \log \left(\frac{2R(X) f(X)}{\max\left(\left(2R(X) f(X) - \frac{2c}{\beta+1} R^{\beta+1}(X)\right), 2\alpha R(X) \right)} \right) \\
&\le \mathbb{E} \log \left(1 + \frac{cR^{\beta}(X)}{(\beta+1) f(X) - cR^{\beta}(X)} \right) \mathbb{1}_A + \mathbb{E} \log \left(\frac{f(X)}{\alpha} \right) \mathbb{1}_{A^c} \\
&\le \mathbb{E} \log \left(1 + \frac{cR^{\beta}(X)}{(\beta+1)\alpha/2} \right) \mathbb{1}_A + \log \left(\frac{\psi(\beta, c)}{\alpha} \right) \mathbb{P}\{A^c\},
\end{aligned}$$

since, by Lemma 7.2 below,

$$f(x) \le \psi(\beta, c) \stackrel{\text{def}}{=} \left(\frac{\beta+1}{2\beta} \right)^{\frac{\beta}{\beta+1}} c^{\frac{1}{\beta+1}}.$$

So,

$$-\mathbf{I} \le \mathbb{E}\left[\frac{cR^{\beta}(X)}{(\beta+1)\alpha/2}\right] + \log\left(\frac{\psi(\beta,c)}{\alpha}\right)\mathbb{E}\left[\frac{cR^{\beta}(X)}{(\beta+1)\alpha/2}\right]$$

(by Markov's inequality)

$$= \mathrm{O}\left(\frac{1}{n^{\beta}}\right),$$

since $\mathbb{E}R^{\beta}(X) \le (\frac{2}{n+1})^{\beta}$, as noted above. □

Lemma 7.2. *If f is a Lipschitz density on $[0,1]$ satisfying $|f(x)-f(x')| \le c|x-x'|^{\beta}$ for $c > 0$ and $\beta \in (0,1]$, then*

$$\max_{x\in[0,1]} f(x) \le \psi(\beta,c) \stackrel{\text{def}}{=} \left(\frac{\beta+1}{2\beta}\right)^{\frac{\beta}{\beta+1}} c^{\frac{1}{\beta+1}}.$$

Proof. Let $M = f(0)$. Since $f(x) \ge \max(0, M - c|x|^{\beta})$, we have

$$\begin{aligned} 1 = \int_0^1 f(x)\mathrm{d}x &\ge \int_{-(\frac{M}{c})^{1/\beta}}^{(\frac{M}{c})^{1/\beta}} \left(M - c|x|^{\beta}\right)\mathrm{d}x \\ &= 2M\left(\frac{M}{c}\right)^{1/\beta} - \frac{2c}{\beta+1}\left(\frac{M}{c}\right)^{1+\frac{1}{\beta}} \\ &= \frac{2M^{1+\frac{1}{\beta}}}{c^{1/\beta}} \times \frac{\beta}{\beta+1}. \end{aligned}$$

Therefore,

$$M \le \left(\frac{\beta+1}{2\beta}\right)^{\frac{\beta}{\beta+1}} c^{\frac{1}{\beta+1}}.$$

□

7.5 Rényi's entropy

Rényi's entropy suggests that one should be able to mimic the Kozachenko-Leonenko methodology for estimating $\int_{\mathbb{R}^d} f^q(\mathbf{x})\mathrm{d}\mathbf{x}$ when q is close to 1. Assume that $q > 1$, and define

$$b_n = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{n\lambda\left(B(\mathbf{X}_i, R_i)\right)}\right)^{q-1},$$

where we adopt the notation of the previous sections. Then we have:

Theorem 7.3. *Assume that* $\int_{\mathbb{R}^d} f^q(\mathbf{x})\mathrm{d}\mathbf{x} < \infty$ *for fixed* $q \in (1,2)$. *Then*

$$\mathbb{E}b_n \to \Gamma(2-q)\int_{\mathbb{R}^d} f^q(\mathbf{x})\mathrm{d}\mathbf{x},$$

where Γ *is the gamma function. If, in addition,* $q \in (1,3/2)$ *and* $\int_{\mathbb{R}^d} f^{2q-1}(\mathbf{x})\mathrm{d}\mathbf{x} < \infty$, *then* $\mathbb{V}b_n = \mathrm{O}(1/n)$. *In that case,*

$$b_n \to \Gamma(2-q)\int_{\mathbb{R}^d} f^q(\mathbf{x})\mathrm{d}\mathbf{x} \quad \textit{in probability}.$$

Proof (Sketch). Note that if $f^\star$ is the maximal function of f, then

$$\begin{aligned}\mathbb{E}b_n &= \mathbb{E}\left[\left(\frac{1}{n\lambda\left(B(\mathbf{X}_1,R_1)\right)}\right)^{q-1}\right]\\ &= \mathbb{E}\left[\left(\frac{\mu\left(B(\mathbf{X}_1,R_1)\right)}{\lambda\left(B(\mathbf{X}_1,R_1)\right)}\right)^{q-1}\left(\frac{1}{n\mu\left(B(\mathbf{X}_1,R_1)\right)}\right)^{q-1}\right]\\ &\stackrel{\text{def}}{=} \mathbb{E}[A_nB_n].\end{aligned}$$

On the one hand,

$$A_n \le (f^\star(\mathbf{X}_1))^{q-1}, \quad A_n \to f^{q-1}(\mathbf{X}_1) \quad \text{in probability},$$

and, since $f(\mathbf{x}) \le f^\star(\mathbf{x})$ at λ-almost all $\mathbf{x}$,

$$\mathbb{E}\left[(f^\star(\mathbf{X}_1))^{q-1}\right] \le \int_{\mathbb{R}^d} f^{\star q}(\mathbf{x})\mathrm{d}\mathbf{x} < \infty$$

for $q > 1$ if $\int_{\mathbb{R}^d} f^q(\mathbf{x})\mathrm{d}\mathbf{x} < \infty$ (by the properties of maximal functions). On the other hand, B_n is independent of $\mathbf{X}_1$ (and thus of $f^\star(\mathbf{X}_1)$ or $f(\mathbf{X}_1)$),

$$B_n \stackrel{\mathscr{D}}{\to} \left(\frac{1}{E}\right)^{q-1} \quad \text{for } E \text{ standard exponential},$$

and, if $1 < q < 2$,

$$\sup_n \mathbb{E}[B_n \mathbb{1}_{[B_n>K]}] \to 0 \quad \text{as } K \to \infty$$

(a property of the sequence of beta distributions under consideration—see Example 20.1). Thus, by the generalized Lebesgue dominated convergence theorem (Lemma 20.2 in the Appendix),

$$\mathbb{E}b_n \to \mathbb{E}\left[f^{q-1}(\mathbf{X}_1)\right] \times \mathbb{E}\left[\left(\frac{1}{E}\right)^{q-1}\right]$$
$$= \Gamma(2-q)\int_{\mathbb{R}^d} f^q(\mathbf{x})\mathrm{d}\mathbf{x}.$$

One can use the Efron-Stein technique to show, in addition, that $\mathbb{V}b_n = \mathrm{O}(1/n)$ provided that

$$\mathbb{E}\left[(f^\star(\mathbf{X}_1))^{2q-2}\right] < \infty \quad \text{and} \quad \mathbb{E}\left[\left(\frac{1}{E}\right)^{2q-2}\right] < \infty.$$

Observe that $\mathbb{E}[(f^\star(\mathbf{X}_1))^{2q-2}] \leq \int_{\mathbb{R}^d}(f^\star(\mathbf{x}))^{2q-1}\mathrm{d}\mathbf{x}$, and that this integral is finite whenever $\int_{\mathbb{R}^d} f^{2q-1}(\mathbf{x})\mathrm{d}\mathbf{x} < \infty$ if $2q-1 > 1$, i.e., $q > 1$. Also,

$$\mathbb{E}\left[\left(\frac{1}{E}\right)^{2q-2}\right] = \Gamma(3-2q) < \infty$$

if $q < 3/2$. □

Part II
Regression estimation

Chapter 8
The nearest neighbor regression function estimate

8.1 Nonparametric regression function estimation

Let $(\mathbf{X}, Y)$ be a pair of random variables taking values in $\mathbb{R}^d \times \mathbb{R}$. The goal of regression analysis is to understand how the values of the response variable Y depend on the values of the observation vector $\mathbf{X}$. The objective is to find a Borel measurable function g such that $|Y - g(\mathbf{X})|$ is small, where "small" could be defined in terms of the L^p risk $\mathbb{E}|Y - g(\mathbf{X})|^p$ $(p > 0)$, for example. Of particular interest is the L^2 risk of g,

$$\mathbb{E}\left|Y - g(\mathbf{X})\right|^2. \tag{8.1}$$

One advantage of (8.1) is that the function g that minimizes the risk can be derived explicitly. To see this, assume that $\mathbb{E}Y^2 < \infty$ and note that we are interested in a measurable function $g^\star : \mathbb{R}^d \to \mathbb{R}$ such that

$$\mathbb{E}\left|Y - g^\star(\mathbf{X})\right|^2 = \inf_g \mathbb{E}\left|Y - g(\mathbf{X})\right|^2,$$

where the infimum is evaluated over all measurable functions $g : \mathbb{R}^d \to \mathbb{R}$ with $\mathbb{E}g^2(\mathbf{X}) < \infty$. Next, let

$$r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$$

be (a version of) the regression function of Y on $\mathbf{X}$. Observe that, for an arbitrary $g : \mathbb{R}^d \to \mathbb{R}$ such that $\mathbb{E}g^2(\mathbf{X}) < \infty$,

$$\begin{aligned}\mathbb{E}\left|Y - g(\mathbf{X})\right|^2 &= \mathbb{E}\left|Y - r(\mathbf{X}) + r(\mathbf{X}) - g(\mathbf{X})\right|^2 \\ &= \mathbb{E}|Y - r(\mathbf{X})|^2 + \mathbb{E}\left|r(\mathbf{X}) - g(\mathbf{X})\right|^2,\end{aligned}$$

G. Biau, L. Devroye, *Lectures on the Nearest Neighbor Method*,
Springer Series in the Data Sciences, DOI 10.1007/978-3-319-25388-6_8

where we have used

$$\begin{aligned}\mathbb{E}\left[(Y - r(\mathbf{X}))\,(r(\mathbf{X}) - g(\mathbf{X}))\right] &= \mathbb{E}\left[\mathbb{E}\left[(Y - r(\mathbf{X}))\,(r(\mathbf{X}) - g(\mathbf{X}))\mid \mathbf{X}\right]\right] \\ &= \mathbb{E}\left[(r(\mathbf{X}) - g(\mathbf{X}))\,\mathbb{E}\left[Y - r(\mathbf{X})\mid \mathbf{X}\right]\right] \\ &= \mathbb{E}\left[(r(\mathbf{X}) - g(\mathbf{X}))\,(r(\mathbf{X}) - r(\mathbf{X}))\right] \\ &= 0.\end{aligned}$$

Thus, denoting by μ the distribution of $\mathbf{X}$, we conclude that

$$\mathbb{E}\,|Y - g(\mathbf{X})|^2 = \mathbb{E}\,|Y - r(\mathbf{X})|^2 + \int_{\mathbb{R}^d} |g(\mathbf{x}) - r(\mathbf{x})|^2\,\mu(\mathrm{d}\mathbf{x}).$$

The second term on the right-hand side is called the L^2 error (integrated squared error) of g. It is always nonnegative and is zero if and only if $g(\mathbf{x}) = r(\mathbf{x})$ μ-almost surely. Therefore $g^\star(\mathbf{x}) = r(\mathbf{x})$ μ-almost surely, i.e., the optimal approximation (with respect to the L^2 risk) of Y by a square-integrable function of $\mathbf{X}$ is given by $r(\mathbf{X})$. It is known that the conditional expectation $\mathbb{E}[Y|\mathbf{X}]$ exists even if it is only assumed that Y has a finite first-order moment. Thus, in the sequel, we suppose that $\mathbb{E}|Y| < \infty$ and focus our attention on the function r as a good approximation of the link between $\mathbf{X}$ and Y.

In practice, however, the distribution of $(\mathbf{X}, Y)$ (and thus, the regression function) is usually unknown. Therefore, it is hopeless to predict Y using $r(\mathbf{X})$. But, fortunately, it is often possible to collect data according to the distribution of $(\mathbf{X}, Y)$ and to estimate the regression function from this data set. To be more precise, assume that we are given a sample $\mathscr{D}_n = ((\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n))$ of i.i.d. $\mathbb{R}^d \times \mathbb{R}$-valued random variables, independent of $(\mathbf{X}, Y)$ and distributed as this prototype pair. The objective is to use the data $\mathscr{D}_n$ in order to construct an estimate $r_n : \mathbb{R}^d \to \mathbb{R}$ of the function r. Here, $r_n(\mathbf{x}) = r_n(\mathbf{x}; \mathscr{D}_n)$ is a Borel measurable function of both $\mathbf{x}$ and the observations. However, for simplicity, we omit $\mathscr{D}_n$ in the notation and write $r_n(\mathbf{x})$ instead of $r_n(\mathbf{x}; \mathscr{D}_n)$.

In modern statistics, regression analysis is widely used for inference and forecasting, where its application has substantial overlap with the field of machine learning. Over the years, a large body of techniques for carrying out regression estimation has been developed. Customary methods such as linear and ordinary least squares regression are parametric, in the sense that the target function is defined in terms of a finite number of unknown parameters, which are estimated from the data. Parametric estimates usually depend only upon a few parameters and are therefore suitable even for small sample sizes n if the parametric model is appropriately chosen. On the other hand, regardless of the data, a parametric estimate cannot approximate the regression function better than the best function that has the assumed parametric structure. For example, a linear fit will produce a large error for every sample size if the true underlying regression function is not linear and cannot be well

approximated by linear functions. This inflexibility is avoided by the nonparametric estimates, which do not assume that the regression function can be described by finitely many parameters. Such procedures are therefore particularly appropriate when the joint distribution of $\mathbf{X}$ and Y cannot be safely assumed to belong to any specified parametric family of distributions.

The literature on nonparametric regression methods is too vast to permit anything like a fair summary within the confines of a short introduction. For a comprehensive presentation, we refer to the monograph by Györfi et al. (2002), which covers almost all known nonparametric regression techniques, such as classical local averaging procedures (including kernel, partitioning, and k-nearest neighbor estimates), least squares and penalized least squares estimates, local polynomial kernel estimates, and orthogonal series estimates. Our goal in this and the subsequent chapters is to offer an in-depth mathematical analysis of the nearest neighbor regression function estimate, a nonparametric method first discussed by Royall (1966) and Cover (1968) in the late 60s of the 20th century, and later by Stone (1977).

8.2 The nearest neighbor estimate

The data in our model can be rewritten as

$$Y_i = r(\mathbf{X}_i) + \varepsilon_i, \quad 1 \le i \le n,$$

where $\varepsilon_i = Y_i - r(\mathbf{X}_i)$ satisfies $\mathbb{E}[\varepsilon_i|\mathbf{X}_i] = 0$. Thus, each Y_i can be considered as the sum of the value of the regression function at $\mathbf{X}_i$ and some error ε_i, where the expected value of the error is zero. This motivates the construction of estimates by local averaging, i.e., estimation of $r(\mathbf{x})$ by the average of those Y_i's for which $\mathbf{X}_i$ is "close" to $\mathbf{x}$.

Of particular importance is the nearest neighbor estimate. Let us denote by $(\mathbf{X}_{(1)}(\mathbf{x}), Y_{(1)}(\mathbf{x})), \ldots, (\mathbf{X}_{(n)}(\mathbf{x}), Y_{(n)}(\mathbf{x}))$ a reordering of the data $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ according to increasing values of $\|\mathbf{X}_i - \mathbf{x}\|$, that is,

$$\|\mathbf{X}_{(1)}(\mathbf{x}) - \mathbf{x}\| \le \cdots \le \|\mathbf{X}_{(n)}(\mathbf{x}) - \mathbf{x}\|$$

and $Y_{(i)}(\mathbf{x})$ is the Y_i corresponding to $\mathbf{X}_{(i)}(\mathbf{x})$. When $\|\mathbf{X}_i - \mathbf{x}\| = \|\mathbf{X}_j - \mathbf{x}\|$ but $i \neq j$, then we have a distance tie, and a tie-breaking strategy must be defined to disambiguate the situation. There are different possible policies to reach this goal. The one we adopt is based on ranks: if $\|\mathbf{X}_i - \mathbf{x}\| = \|\mathbf{X}_j - \mathbf{x}\|$ but $i \neq j$, then $\mathbf{X}_i$ is considered to be closer to $\mathbf{x}$ if $i < j$. For more clarity, the next section is devoted to this messy distance tie issue. Nevertheless, it should be stressed that if μ has a density, then tie-breaking is needed with zero probability and becomes therefore irrelevant.

Definition 8.1. The (raw) nearest neighbor estimate is

$$r_n(\mathbf{x}) = \sum_{i=1}^{n} v_{ni} Y_{(i)}(\mathbf{x}), \tag{8.2}$$

where $(v_{n1}, \ldots, v_{nn})$ is a given (deterministic) weight vector summing to one.

Remark 8.1. Throughout and when needed, with a slight abuse of notation, we set $v_{ni} = 0$ for all $i > n$. □

Thus, in this procedure, the local averaging is done by weighing the Y_i's according to a sequence of specified weights and according to the rank of the distance from $\mathbf{X}_i$ to $\mathbf{x}$. The weights $(v_{n1}, \ldots, v_{nn})$ are always summing to one. They are usually nonnegative, but in some cases they can take negative values. Important subclasses include monotone weight vectors:

$$v_{n1} \geq v_{n2} \geq \cdots \geq v_{nn},$$

and the uniform weight vector:

$$v_{ni} = \begin{cases} \frac{1}{k} & \text{for } 1 \leq i \leq k \\ 0 & \text{for } k < i \leq n, \end{cases}$$

where k is a positive integer not exceeding n. In the latter case, we speak of the standard k-nearest neighbor estimate.

Nearest neighbor estimation is one of the oldest approaches to regression analysis and pattern recognition (Fix and Hodges, 1951, 1952; Cover, 1968). It is a widespread nonparametric method, with hundreds of research articles published on the topic since the 80s (Dasarathy, 1991, has provided a comprehensive collection of around 140 key papers). For implementation, it requires only a measure of distance in the sample space, hence its popularity as a starting point for refinement, improvement, and adaptation to new settings.

Remark 8.2. There are many ways of defining an ordering of data points. In the present text, we take the standard Euclidean (ℓ^2) distance. One can consider ℓ^p distances in general, and even distances skewed by an affine transformation. In particular, for $\mathbf{x} \in \mathbb{R}^d$, the norm of $\mathbf{x}$ can be defined as $\|A\mathbf{x}\|$, where A is a fixed positive definite $d \times d$ matrix.

A slightly different approach is taken by Devroye (1991b) (see also Olshen, 1977), where given $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and $\mathbf{x}$, one first ranks all data with respect to each coordinate. Let r_{ij} be the rank of the j-th coordinate of $\mathbf{X}_i$ among all j-th coordinates, and let r_j be the rank of the j-th coordinate of $\mathbf{x}$. Assume for now that all marginal distributions are nonatomic to avoid ties. Then form an ℓ^p metric

$$\text{distance}(\mathbf{X}_i, \mathbf{x}) = \begin{cases} \left(\sum_{j=1}^{d} |r_{ij} - r_j|^p \right)^{1/p} & \text{for } 1 \leq p < \infty \\ \max_{1 \leq j \leq d} |r_{ij} - r_j| & \text{for } p = \infty. \end{cases}$$

Most of the results of this book regarding regression function estimation and classification remain true for the nearest neighbor estimate based on this metric. In particular, the estimate is universally weakly consistent under the same condition on the v_{ni}'s given in Theorem 11.1. It has the advantage to be invariant under monotone transformations of the coordinate axes.

One can even construct regression function estimates that are invariant under all affine transformations of the data. One possibility is described by Biau et al. (2012). Here is another one. Let $\top$ denote transposition and assume that vectors are in column format. For a given set of D vectors $\mathbf{a}_1, \ldots, \mathbf{a}_D \in \mathbb{R}^d$, rank the data $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and $\mathbf{x}$ according to the values of $\mathbf{a}_j^\top \mathbf{X}_i$, $1 \le i \le n$, and $\mathbf{a}_j^\top \mathbf{x}$, and form the ℓ^p metric on the ranks described above, but now with D, not d, sets of ranks. The invariance under affine transformations follows when $\mathbf{a}_1, \ldots, \mathbf{a}_D$ are selected such that the ranks do not change. To achieve this, one can take $\mathbf{a}_j$ perpendicular to the $(d-1)$-dimensional hyperplane determined by $(\mathbf{X}_{(j-1)d+1}, \ldots, \mathbf{X}_{jd})$. This family of estimates has not yet been studied, but should be consistent for all fixed $D \ge d$, at least when $\mathbf{X}$ has a density. □

As for density estimation, there are different ways to measure the closeness of the regression estimate r_n to the true regression function r. These include global criteria, such as the distances in L^p

$$L^p(r_n, r) = \left(\int_{\mathbb{R}^d} |r_n(\mathbf{x}) - r(\mathbf{x})|^p \, \mu(\mathrm{d}\mathbf{x}) \right)^{1/p},$$

and the uniform deviation

$$L^\infty(r_n, r) = \operatorname*{ess\,sup}_{\mathbf{x} \in \mathbb{R}^d} |r_n(\mathbf{x}) - r(\mathbf{x})|$$

(with respect to the distribution μ). On the other hand, local criteria fix a query point $\mathbf{x}$ and look at the closeness between $r_n(\mathbf{x})$ and $r(\mathbf{x})$. The regression function estimate r_n is called weakly (strongly) pointwise consistent on A if

$$r_n(\mathbf{x}) \to r(\mathbf{x}) \quad \text{in probability (almost surely) for all } \mathbf{x} \in A.$$

From the pointwise convergence of r_n, one can often deduce results about the convergence of $\int_{\mathbb{R}^d} |r_n(\mathbf{x}) - r(\mathbf{x})|^p \mu(\mathrm{d}\mathbf{x})$, but the inverse deduction is not simple.

In the next few chapters, we will be interested in sufficient conditions on the weight vector $(v_{n1}, \ldots, v_{nn})$ for weak and strong consistency of the general nearest neighbor estimate, under minimal assumptions regarding the distribution of $(\mathbf{X}, Y)$.

8.3 Distance tie-breaking

When $\mathbf{X}$ has a density with respect to the Lebesgue measure on $\mathbb{R}^d$, distance ties occur with zero probability. However, in general, $\mathbf{X}$ does not have a density and ties can occur: the event $[\|\mathbf{X}_i - \mathbf{X}\| = \|\mathbf{X}_j - \mathbf{X}\|$ but $i \neq j]$ can happen with positive probability, so we have to find a strategy to break the ties.

Remark 8.3. To see that the density assumption cannot be relaxed to the condition that the distribution μ of $\mathbf{X}$ is merely nonatomic without facing possible distance ties, consider the following distribution on $\mathbb{R}^d \times \mathbb{R}^{d'}$ with $d, d' \geq 2$:

$$\mu = \frac{1}{2}(\tau_d \otimes \sigma_{d'}) + \frac{1}{2}(\sigma_d \otimes \tau_{d'}),$$

where τ_d denotes the uniform distribution on the surface of the unit ball of $\mathbb{R}^d$ and σ_d denotes the unit point mass at the origin of $\mathbb{R}^d$. Observe that if $\mathbf{X}$ has distribution $\tau_d \otimes \sigma_{d'}$ and $\mathbf{X}'$ has distribution $\sigma_d \otimes \tau_{d'}$, then $\|\mathbf{X} - \mathbf{X}'\| = \sqrt{2}$. Hence, if $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$, $\mathbf{X}_4$ are independent with distribution μ, then $\mathbb{P}\{\|\mathbf{X}_1 - \mathbf{X}_2\| = \|\mathbf{X}_3 - \mathbf{X}_4\|\} = 1/4$. □

There is no consensus on the best way to break distance ties. Some possible policies are discussed in Chapter 11 of the monograph by Devroye et al. (1996). The one we adopt throughout the book is by looking at the index, i.e., whenever $\|\mathbf{X}_i - \mathbf{x}\| = \|\mathbf{X}_j - \mathbf{x}\|$, $i \neq j$, we declare $\mathbf{X}_i$ closer to $\mathbf{x}$ if $i < j$. However, the possibility of distance ties leads to consider various possible versions of the nearest neighbor regression function estimate. Let us start by defining the rank Σ_i of $\mathbf{X}_i$ with respect to $\mathbf{x}$ as follows:

$$\Sigma_i = \sum_{j \neq i} \mathbb{1}_{[\|\mathbf{X}_j - \mathbf{x}\| < \|\mathbf{X}_i - \mathbf{x}\|]} + \sum_{j \leq i} \mathbb{1}_{[\|\mathbf{X}_j - \mathbf{x}\| = \|\mathbf{X}_i - \mathbf{x}\|]}.$$

Therefore, if $\|\mathbf{X}_i - \mathbf{x}\| = \|\mathbf{X}_j - \mathbf{x}\|$, then $\Sigma_i < \Sigma_j$ if and only if $i < j$. Since ties are broken by looking at the index, we see that the data are reordered by rank with respect to $\mathbf{x}$, i.e., that our ordering

$$\left((\mathbf{X}_{(1)}(\mathbf{x}), Y_{(1)}(\mathbf{x})\right), \ldots, \left((\mathbf{X}_{(1)}(\mathbf{x}), Y_{(1)}(\mathbf{x})\right)$$

is such that

$$(\mathbf{X}_i, Y_i) = \left(\mathbf{X}_{(\Sigma_i)}(\mathbf{x}), Y_{(\Sigma_i)}(\mathbf{x})\right).$$

There is an inverse rank vector $(\sigma_1, \ldots, \sigma_n)$, also a permutation of $(1, \ldots, n)$, which is defined by

$$\left(\mathbf{X}_{(i)}(\mathbf{x}), Y_{(i)}(\mathbf{x})\right) = (\mathbf{X}_{\sigma_i}, Y_{\sigma_i}).$$

Thus, $\sigma_{\Sigma_i} = i$ and $\Sigma_{\sigma_i} = i$ for all i. In the raw weighted nearest neighbor estimate, we have

$$r_n(\mathbf{x}) = \sum_{i=1}^{n} v_{ni} Y_{(i)}(\mathbf{x}) = \sum_{i=1}^{n} v_{ni} Y_{\sigma_i}$$
$$= \sum_{i=1}^{n} \sum_{j=1}^{n} v_{ni} Y_j \mathbb{1}_{[\Sigma_j = i]}.$$

The knee-jerk reaction of most statisticians is to break distance ties by averaging. If we write $r_n(\mathbf{x}; \mathbf{X}_1, \ldots, \mathbf{X}_n)$ for the raw weighted nearest neighbor estimate, then the averaged estimate, which removes any dependence on ties, is

$$\bar{r}_n(\mathbf{x}) = \frac{1}{n!} \sum_{\substack{\text{all permutations} \\ (\tau_1, \ldots, \tau_n) \text{ of } (1, \ldots, n)}} r_n(\mathbf{x}; \mathbf{X}_{\tau_1}, \ldots, \mathbf{X}_{\tau_n}).$$

If there are no distance ties, then $\bar{r}_n(\mathbf{x}) = r_n(\mathbf{x})$. However, if

$$\|\mathbf{X}_{(i-1)}(\mathbf{x}) - \mathbf{x}\| < \|\mathbf{X}_{(i)}(\mathbf{x}) - \mathbf{x}\| = \cdots = \|\mathbf{X}_{(j-1)}(\mathbf{x}) - \mathbf{x}\| < \|\mathbf{X}_{(j)}(\mathbf{x}) - \mathbf{x}\|,$$

then the weight of each of $Y_{(i)}(\mathbf{x}), \ldots, Y_{(j-1)}(\mathbf{x})$ in the definition of $\bar{r}_n(\mathbf{x})$ is the average weight

$$\frac{v_{ni} + \cdots + v_{nj-1}}{j - i}.$$

It is the same principle that is used in payouts in a golf tournament in case of ties—the prize money is averaged. The weight of $Y_{(i)}(\mathbf{x})$ in $\bar{r}_n(\mathbf{x})$ is written as V_{ni}, which now is a random variable:

$$\bar{r}_n(\mathbf{x}) = \sum_{i=1}^{n} V_{ni} Y_{(i)}(\mathbf{x}).$$

We still have $\sum_{i=1}^{n} V_{ni} = 1$, and if $(v_{ni}, 1 \le i \le n)$ is monotone, then so is $(V_{ni}, 1 \le i \le n)$.

Proving consistency and rates of convergence for r_n is harder and more informative than for $\bar{r}_n$ in view of the following trivial observations:

$$\mathbb{E}\,|\bar{r}_n(\mathbf{x}) - r(\mathbf{x})|^p \le \mathbb{E}\,|r_n(\mathbf{x}) - r(\mathbf{x})|^p\,,$$

$$\mathbb{E}\left[\sup_{\mathbf{x} \in \mathbb{R}^d} |\bar{r}_n(\mathbf{x}) - r(\mathbf{x})|\right] \le \mathbb{E}\left[\sup_{\mathbf{x} \in \mathbb{R}^d} |r_n(\mathbf{x}) - r(\mathbf{x})|\right],$$

or

$$\mathbb{E}\left|\bar{r}_n(\mathbf{X}) - r(\mathbf{X})\right|^p \leq \mathbb{E}\left|r_n(\mathbf{X}) - r(\mathbf{X})\right|^p.$$

Thus, all our results are for r_n.

We conclude this section with the following useful proposition, whose proof is a good illustration of the utility of the rank formalism. Further results of this sort are given in Kaufmann and Reiss (1992) and Cérou and Guyader (2006).

Proposition 8.1. *Assume that* $\mathbb{E}|Y| < \infty$, *and let* $r(\mathbf{X}) = \mathbb{E}[Y|\mathbf{X}]$. *Then, conditional on* $\mathbf{X}_1, \ldots, \mathbf{X}_n$, *the random variables*

$$\left(\mathbf{X}_{(1)}(\mathbf{x}), Y_{(1)}(\mathbf{x})\right), \ldots, \left(\mathbf{X}_{(n)}(\mathbf{x}), Y_{(n)}(\mathbf{x})\right)$$

are independent. Moreover, for each $1 \leq i \leq n$,

$$\mathbb{E}\left[Y_{(i)}(\mathbf{x}) - r\left(\mathbf{X}_{(i)}(\mathbf{x})\right) \mid \mathbf{X}_1, \ldots, \mathbf{X}_n\right] = 0.$$

Proof. Set $\mathbf{Z}_i = (\mathbf{X}_i, Y_i)$ and $\mathbf{Z}_{(i)}(\mathbf{x}) = (\mathbf{X}_{(i)}(\mathbf{x}), Y_{(i)}(\mathbf{x}))$, $1 \leq i \leq n$. We have to prove that, for any Borel sets $A_1, \ldots, A_n$,

$$\mathbb{E}\left[\prod_{i=1}^n \mathbb{1}_{[\mathbf{Z}_{(i)}(\mathbf{x}) \in A_i]} \mid \mathbf{X}_1, \ldots, \mathbf{X}_n\right] = \prod_{i=1}^n \mathbb{E}\left[\mathbb{1}_{[\mathbf{Z}_{(i)}(\mathbf{x}) \in A_i]} \mid \mathbf{X}_1, \ldots, \mathbf{X}_n\right].$$

Let $(\sigma_1, \ldots, \sigma_n)$ be the random permutation of $(1, \ldots, n)$ such that $\mathbf{Z}_{(i)}(\mathbf{x}) = \mathbf{Z}_{\sigma_i}$. We may write

$$\begin{aligned}
&\mathbb{E}\left[\prod_{i=1}^n \mathbb{1}_{[\mathbf{Z}_{(i)}(\mathbf{x}) \in A_i]} \mid \mathbf{X}_1, \ldots, \mathbf{X}_n\right] \\
&\quad = \mathbb{E}\left[\prod_{i=1}^n \mathbb{1}_{[\mathbf{Z}_{\sigma_i} \in A_i]} \mid \mathbf{X}_1, \ldots, \mathbf{X}_n\right] \\
&\quad = \sum_{\substack{\text{all permutations} \\ (\tau_1, \ldots, \tau_n) \text{ of } (1, \ldots, n)}} \left[\mathbb{1}_{[(\sigma_1, \ldots, \sigma_n) = (\tau_1, \ldots, \tau_n)]} \times \mathbb{E}\left[\prod_{i=1}^n \mathbb{1}_{[\mathbf{Z}_{\tau_i} \in A_i]} \mid \mathbf{X}_1, \ldots, \mathbf{X}_n\right]\right],
\end{aligned}$$

where, in the last equality, we used the measurability of $(\sigma_1, \ldots, \sigma_n)$ with respect to $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Next, invoking the independence of $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$, it is a simple exercise to prove that $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are independent conditional on $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Thus,

$$\begin{aligned}
&\mathbb{E}\left[\prod_{i=1}^n \mathbb{1}_{[\mathbf{Z}_{(i)}(\mathbf{x}) \in A_i]} \mid \mathbf{X}_1, \ldots, \mathbf{X}_n\right] \\
&\quad = \sum_{\substack{\text{all permutations} \\ (\tau_1, \ldots, \tau_n) \text{ of } (1, \ldots, n)}} \left[\mathbb{1}_{[(\sigma_1, \ldots, \sigma_n) = (\tau_1, \ldots, \tau_n)]} \times \prod_{i=1}^n \mathbb{E}\left[\mathbb{1}_{[\mathbf{Z}_{\tau_i} \in A_i]} \mid \mathbf{X}_1, \ldots, \mathbf{X}_n\right]\right]
\end{aligned}$$

$$= \prod_{i=1}^{n} \mathbb{E}\left[\mathbb{1}_{[\mathbf{Z}_{\sigma_i} \in A_i]} \mid \mathbf{X}_1, \ldots, \mathbf{X}_n\right]$$
$$= \prod_{i=1}^{n} \mathbb{E}\left[\mathbb{1}_{[\mathbf{Z}_{(i)}(\mathbf{x}) \in A_i]} \mid \mathbf{X}_1, \ldots, \mathbf{X}_n\right].$$

This shows the first statement of the proposition. To prove the second one, notice that

$$\mathbb{E}\left[Y_{(i)}(\mathbf{x}) - r\left(\mathbf{X}_{(i)}(\mathbf{x})\right) \mid \mathbf{X}_1, \ldots, \mathbf{X}_n\right] = \mathbb{E}\left[Y_{(i)}(\mathbf{x}) \mid \mathbf{X}_1, \ldots, \mathbf{X}_n\right] - r\left(\mathbf{X}_{(i)}(\mathbf{x})\right).$$

But

$$\begin{aligned}
&\mathbb{E}\left[Y_{(i)}(\mathbf{x}) \mid \mathbf{X}_1, \ldots, \mathbf{X}_n\right] \\
&\quad = \sum_{j=1}^{n} \mathbb{1}_{[\sigma_i = j]} \mathbb{E}[Y_j \mid \mathbf{X}_1, \ldots, \mathbf{X}_n] \\
&\quad = \sum_{j=1}^{n} \mathbb{1}_{[\sigma_i = j]} \mathbb{E}[Y_j | \mathbf{X}_j] \\
&\qquad \text{(by independence of } (\mathbf{X}_j, Y_j) \text{ and } \mathbf{X}_1, \ldots, \mathbf{X}_{j-1}, \mathbf{X}_{j+1}, \ldots, \mathbf{X}_n) \\
&\quad = \sum_{j=1}^{n} \mathbb{1}_{[\sigma_i = j]} r\left(\mathbf{X}_j\right) \\
&\quad = r\left(\mathbf{X}_{(i)}(\mathbf{x})\right).
\end{aligned}$$

□

Chapter 9
The 1-nearest neighbor regression function estimate

9.1 Consistency and residual variance

Our objective in this short chapter is to analyze some elementary consistency properties of the 1-nearest neighbor regression function estimate. This will also offer the opportunity to familiarize the reader with concepts that will be encountered in the next few chapters. Recall that this very simple estimation procedure is defined by setting

$$r_n(\mathbf{x}) = Y_{(1)}(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d,$$

where $(\mathbf{X}_{(1)}(\mathbf{x}), Y_{(1)}(\mathbf{x})), \ldots, (\mathbf{X}_{(n)}(\mathbf{x}), Y_{(n)}(\mathbf{x}))$ is a reordering of the data according to increasing values of $\|\mathbf{X}_i - \mathbf{x}\|$, and distance ties are broken by looking at indices.

Assuming $\mathbb{E}Y^2 < \infty$, we have seen in the introduction of Chapter 8 that the regression function $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ achieves the minimal value of the L^2 risk over all Borel measurable functions $g : \mathbb{R}^d \to \mathbb{R}$ with $\mathbb{E}g^2(\mathbf{X}) < \infty$, that is,

$$\mathbb{E}\left|Y - r(\mathbf{X})\right|^2 = \inf_g \mathbb{E}\left|Y - g(\mathbf{X})\right|^2.$$

The quantity

$$L^\star = \mathbb{E}\left|Y - r(\mathbf{X})\right|^2 = \mathbb{E}Y^2 - \mathbb{E}r^2(\mathbf{X})$$

is called the residual variance (variance of the residual, noise variance). It is zero if and only if $Y = r(\mathbf{X})$ with probability one—such a situation is called noiseless.

In this chapter, we prove the following result:

Theorem 9.1. *Assume that $\mathbb{E}Y^2 < \infty$. Then the 1-nearest neighbor regression function estimate r_n satisfies*

$$\mathbb{E}\left|r_n(\mathbf{X}) - r(\mathbf{X})\right|^2 \to L^\star.$$

G. Biau, L. Devroye, *Lectures on the Nearest Neighbor Method*,
Springer Series in the Data Sciences, DOI 10.1007/978-3-319-25388-6_9

Thus, when $k = 1$, the mean integrated squared error $\mathbb{E}|r_n(\mathbf{X}) - r(\mathbf{X})|^2$ converges to the residual variance $L^\star$. This convergence is universal, in the sense that it happens for any distribution of $(\mathbf{X}, Y)$ with $\mathbb{E}Y^2 < \infty$. The 1-nearest neighbor estimate is L^2-consistent (that is, $\mathbb{E}|r_n(\mathbf{X}) - r(\mathbf{X})|^2 \to 0$) in the noiseless case only. To circumvent this problem, a possible strategy is to let the parameter k depend upon n. In the remaining chapters on regression, k will satisfy $k \to \infty$ and $k/n \to 0$ as $n \to \infty$ in order to obtain consistency.

9.2 Proof of Theorem 9.1

The proof of Theorem 9.1 begins with a useful technical lemma, which is a refinement of an inequality of Stone (1977) (see also Fritz, 1975).

Lemma 9.1. *Let $p \geq 1$, and let $g : \mathbb{R}^d \to \mathbb{R}$ be a Borel measurable function such that $\mathbb{E}|g(\mathbf{X})|^p < \infty$. Then*

$$\mathbb{E}\left|g\left(\mathbf{X}_{(1)}(\mathbf{X})\right)\right|^p \leq c_d \mathbb{E}\left|g(\mathbf{X})\right|^p,$$

where c_d is a positive constant depending upon the dimension d only.

Proof. Let $\sigma_1 = \sigma_1(\mathbf{X}; \mathbf{X}_1, \ldots, \mathbf{X}_n)$ be the index in $\{1, \ldots, n\}$ corresponding to the nearest neighbor pair $(\mathbf{X}_{(1)}(\mathbf{X}), Y_{(1)}(\mathbf{X}))$. Thus, $(\mathbf{X}_{(1)}(\mathbf{X}), Y_{(1)}(\mathbf{X})) = (\mathbf{X}_{\sigma_1}, Y_{\sigma_1})$. Write

$$\mathbb{E}\left|g\left(\mathbf{X}_{(1)}(\mathbf{X})\right)\right|^p = \mathbb{E}\left[\sum_{j=1}^n \mathbb{1}_{[\sigma_1 = j]}\left|g(\mathbf{X}_j)\right|^p\right].$$

Define the event

$$A = \bigcup_{i=1}^n [\mathbf{X} = \mathbf{X}_i].$$

Then

$$\begin{aligned}\mathbb{E}\left[\sum_{j=1}^n \mathbb{1}_{[\sigma_1 = j]}\mathbb{1}_A\left|g(\mathbf{X}_j)\right|^p\right] &= \mathbb{E}\left[|g(\mathbf{X})|^p \sum_{j=1}^n \mathbb{1}_{[\sigma_1 = j]}\mathbb{1}_A\right] \\ &\leq \mathbb{E}\left|g(\mathbf{X})\right|^p.\end{aligned}$$

So, only the case A^c is of interest. Cover $\mathbb{R}^d$ with a finite number γ_d of cones $\mathscr{C}_1, \ldots, \mathscr{C}_{\gamma_d}$ of angle $\theta = \pi/8$ (this is possible by virtue of Theorem 20.15). For each $1 \leq \ell \leq \gamma_d$, let $\mathscr{C}_\ell(\mathbf{X}) = \mathbf{X} + \mathscr{C}_\ell$ be the corresponding translated cone with

origin at $\mathbf{X}$. Within each translated cone, mark all $\mathbf{X}_j$'s of smallest radius $\|\mathbf{X}_j - \mathbf{X}\|$. In the ℓ-th cone, let N_ℓ be the number of marked $\mathbf{X}_j$'s. If the cone is empty, then $N_\ell = 0$. By symmetrization,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{j=1}^{n} \mathbb{1}_{[\sigma_1=j]} \left|g(\mathbf{X}_j)\right|^p \mathbb{1}_{[\cap_{i=1}^n [\mathbf{X}\neq\mathbf{X}_i]]}\right] \\
&= \sum_{j=1}^{n} \mathbb{E}\left[\mathbb{1}_{[\sigma_1(\mathbf{X}_j;\mathbf{X}_1,\ldots,\mathbf{X}_{j-1},\mathbf{X},\mathbf{X}_{j+1},\ldots,\mathbf{X}_n)=j]} \left|g(\mathbf{X})\right|^p \mathbb{1}_{[\cap_{\substack{i=1\\ i\neq j}}^n [\mathbf{X}_j\neq\mathbf{X}_i]]} \mathbb{1}_{[\mathbf{X}_j\neq\mathbf{X}]}\right] \\
&\le \mathbb{E}\left[\left|g(\mathbf{X})\right|^p \sum_{\ell=1}^{\gamma_d} \sum_{j:\mathbf{X}_j\in\mathscr{C}_\ell(\mathbf{X})} \mathbb{1}_{[\sigma_1(\mathbf{X}_j;\mathbf{X}_1,\ldots,\mathbf{X}_{j-1},\mathbf{X},\mathbf{X}_{j+1},\ldots,\mathbf{X}_n)=j]} \mathbb{1}_{[\cap_{\substack{i=1\\ i\neq j}}^n [\mathbf{X}_j\neq\mathbf{X}_i]]} \mathbb{1}_{[\mathbf{X}_j\neq\mathbf{X}]}\right].
\end{aligned}$$

Within cone $\mathscr{C}_\ell(\mathbf{X})$, on the event A^c, $\mathbf{X}$ can only be the nearest neighbor of at most one point, namely a marked point. In fact, if $N_\ell > 1$, then it cannot be the nearest neighbor of any point. To see this, assume that $\mathbf{X}_j$ is marked, that $\mathbf{X}_j \neq \mathbf{X}$, and let $\mathbf{X}_s$ belong to the cone (see Figure 9.1 for an illustration in dimension 2). Then, if $\|\mathbf{X}_s - \mathbf{X}\| \ge \|\mathbf{X}_j - \mathbf{X}\|$, one has $\|\mathbf{X}_j - \mathbf{X}_s\| < \|\mathbf{X} - \mathbf{X}_s\|$. Indeed, consider the triangle $(\mathbf{X}, \mathbf{X}_j, \mathbf{X}_s)$, and define the angles α, β, γ as in Figure 9.2. If $\alpha = 0$, the result is clear. If $\alpha > 0$, then

$$\frac{\|\mathbf{X}_j - \mathbf{X}_s\|}{\sin\alpha} = \frac{\|\mathbf{X} - \mathbf{X}_s\|}{\sin\gamma} = \frac{\|\mathbf{X}_j - \mathbf{X}\|}{\sin\beta}.$$

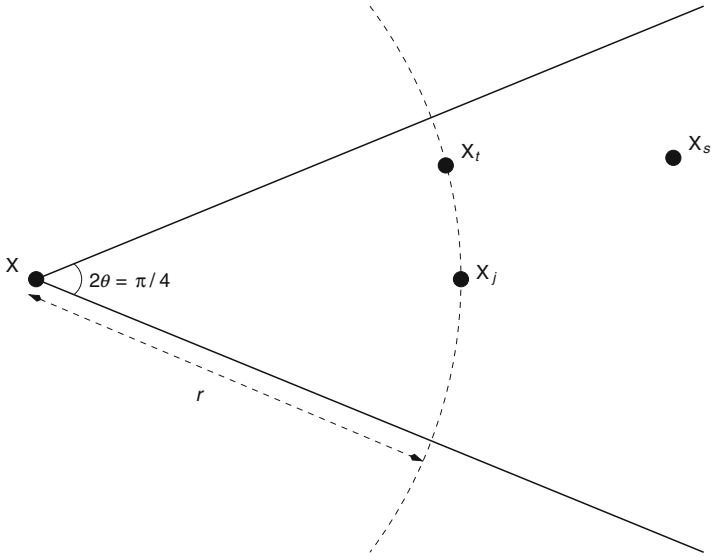

Fig. 9.1 A cone in dimension 2. Both $\mathbf{X}_j$ and $\mathbf{X}_t$ are marked points.

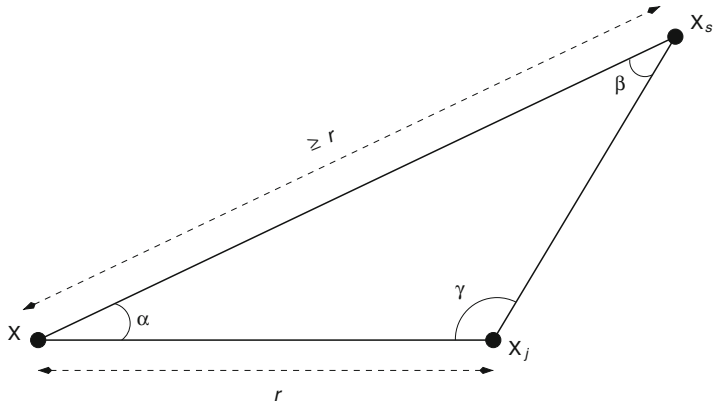

Fig. 9.2 Triangle $(\mathbf{X}, \mathbf{X}_j, \mathbf{X}_s)$ with $\mathbf{X}_j$ a marked point.

Now, $\beta + \gamma = \pi - \alpha$ and $\gamma \geq \beta$, so $\gamma \geq (\pi - \alpha)/2$. Hence,

$$\begin{aligned}
\|\mathbf{X}_j - \mathbf{X}_s\| &\leq \frac{\sin\alpha}{\sin\left(\frac{\pi-\alpha}{2}\right)}\|\mathbf{X} - \mathbf{X}_s\| \\
&= \frac{\sin\alpha}{\cos(\alpha/2)}\|\mathbf{X} - \mathbf{X}_s\| \\
&= 2\sin(\alpha/2)\|\mathbf{X} - \mathbf{X}_s\| \\
&\leq 2\sin(\pi/8)\|\mathbf{X} - \mathbf{X}_s\| \\
&< 2\sin(\pi/6)\|\mathbf{X} - \mathbf{X}_s\| \\
&= \|\mathbf{X} - \mathbf{X}_s\|.
\end{aligned}$$

Therefore, $\sigma_1(\mathbf{X}_s; \mathbf{X}_1, \ldots, \mathbf{X}_{s-1}, \mathbf{X}, \mathbf{X}_{s+1}, \ldots, \mathbf{X}_n) \neq s$ if $s \neq j$. We conclude

$$\sum_{j:\mathbf{X}_j \in \mathscr{C}_\ell(\mathbf{X})} \mathbb{1}_{[\sigma_1(\mathbf{X}_j;\mathbf{X}_1,\ldots,\mathbf{X}_{j-1},\mathbf{X},\mathbf{X}_{j+1},\ldots,\mathbf{X}_n)=j]} \mathbb{1}_{\left[\cap_{\substack{i=1 \\ i\neq j}}^n [\mathbf{X}_j \neq \mathbf{X}_i]\right]} \mathbb{1}_{[\mathbf{X}_j \neq \mathbf{X}]} \leq 1.$$

The inequality then follows easily with $c_d = \gamma_d + 1$. □

Lemma 9.2. *Let $p \geq 1$, and let $g : \mathbb{R}^d \to \mathbb{R}$ be a Borel measurable function such that $\mathbb{E}\,|g(\mathbf{X})|^p < \infty$. Then*

$$\mathbb{E}\left|g\left(\mathbf{X}_{(1)}(\mathbf{X})\right) - g(\mathbf{X})\right|^p \to 0 \quad \textit{and} \quad \mathbb{E}\left|g\left(\mathbf{X}_{(1)}(\mathbf{X})\right)\right|^p \to \mathbb{E}\,|g(\mathbf{X})|^p\,.$$

Proof. For $\varepsilon > 0$, find a uniformly continuous function g_ε with compact support such that $\mathbb{E}|g(\mathbf{X}) - g_\varepsilon(\mathbf{X})|^p \leq \varepsilon$ (see Theorem 20.17 in the Appendix). Then, using the elementary inequality $|a + b + c|^p \leq 3^{p-1}(|a|^p + |b|^p + |c|^p)$ valid for $p \geq 1$ (c_r-inequality—see Proposition 20.1 in the Appendix),

$$
\begin{aligned}
&\mathbb{E}\left|g\left(\mathbf{X}_{(1)}(\mathbf{X})\right)-g(\mathbf{X})\right|^p \\
&\quad\le 3^{p-1}\Big(\mathbb{E}\left|g\left(\mathbf{X}_{(1)}(\mathbf{X})\right)-g_\varepsilon\left(\mathbf{X}_{(1)}(\mathbf{X})\right)\right|^p+\mathbb{E}\left|g_\varepsilon\left(\mathbf{X}_{(1)}(\mathbf{X})\right)-g_\varepsilon(\mathbf{X})\right|^p \\
&\qquad\qquad +\mathbb{E}\left|g_\varepsilon(\mathbf{X})-g(\mathbf{X})\right|^p\Big) \\
&\quad\le 3^{p-1}(c_d+1)\mathbb{E}\left|g(\mathbf{X})-g_\varepsilon(\mathbf{X})\right|^p+3^{p-1}\mathbb{E}\left|g_\varepsilon\left(\mathbf{X}_{(1)}(\mathbf{X})\right)-g_\varepsilon(\mathbf{X})\right|^p \\
&\qquad \text{(by Lemma 9.1)} \\
&\quad\le 3^{p-1}(c_d+1)\varepsilon+3^{p-1}\mathbb{E}\left|g_\varepsilon\left(\mathbf{X}_{(1)}(\mathbf{X})\right)-g_\varepsilon(\mathbf{X})\right|^p.
\end{aligned}
$$

We find $\delta>0$ such that $|g_\varepsilon(\mathbf{y})-g_\varepsilon(\mathbf{x})|\le\varepsilon$ if $\|\mathbf{y}-\mathbf{x}\|\le\delta$. So,

$$
\mathbb{E}\left|g_\varepsilon\left(\mathbf{X}_{(1)}(\mathbf{X})\right)-g_\varepsilon(\mathbf{X})\right|^p\le\varepsilon^p+2^p\|g_\varepsilon\|_\infty^p\times\mathbb{P}\left\{\|\mathbf{X}_{(1)}(\mathbf{X})-\mathbf{X}\|>\delta\right\}.
$$

Let μ be the distribution of $\mathbf{X}$. By the Lebesgue dominated convergence theorem, we have

$$
\mathbb{P}\left\{\|\mathbf{X}_{(1)}(\mathbf{X})-\mathbf{X}\|>\delta\right\}\to 0
$$

if for μ-almost all $\mathbf{x}\in\mathbb{R}^d$,

$$
\mathbb{P}\left\{\|\mathbf{X}_{(1)}(\mathbf{x})-\mathbf{x}\|>\delta\right\}\to 0.
$$

If $\mathbf{x}$ belongs to the support of μ, then $\mathbb{P}\{\|\mathbf{X}-\mathbf{x}\|\le\delta\}>0$ for all $\delta>0$ by definition of the support. Thus,

$$
\mathbb{P}\left\{\|\mathbf{X}_{(1)}(\mathbf{x})-\mathbf{x}\|>\delta\right\}=(1-\mathbb{P}\{\|\mathbf{X}-\mathbf{x}\|\le\delta\})^n\to 0.
$$

But $\mu(\mathrm{supp}(\mu))=1$ (see Chapter 1) and, putting all the pieces together, we conclude that

$$
\mathbb{E}\left|g\left(\mathbf{X}_{(1)}(\mathbf{X})\right)-g(\mathbf{X})\right|^p\le 3^{p-1}\left((c_d+1)\varepsilon+\varepsilon^p\right)+\mathrm{o}(1),
$$

so that

$$
\mathbb{E}\left|g\left(\mathbf{X}_{(1)}(\mathbf{X})\right)-g(\mathbf{X})\right|^p\to 0.
$$

This shows the first assertion of the lemma. To prove the second one, just note that

$$
\left|\mathbb{E}\left|g\left(\mathbf{X}_{(1)}(\mathbf{X})\right)\right|^p-\mathbb{E}\left|g(\mathbf{X})\right|^p\right|\le\mathbb{E}\left|\left|g\left(\mathbf{X}_{(1)}(\mathbf{X})\right)\right|^p-\left|g(\mathbf{X})\right|^p\right|,
$$

and apply the first statement to the function $|g|^p$. □

Lemma 9.3. *Assume that* $\mathbb{E}Y^2<\infty$, *and let* $\sigma^2(\mathbf{x})=\mathbb{E}[|Y-r(\mathbf{X})|^2\,|\,\mathbf{X}=\mathbf{x}]$. *Then*

$$
\mathbb{E}\left|Y_{(1)}(\mathbf{X})-r\left(\mathbf{X}_{(1)}(\mathbf{X})\right)\right|^2=\mathbb{E}[\sigma^2(\mathbf{X}_{(1)}(\mathbf{X}))].
$$

Proof. As in the proof of Lemma 9.1, we let σ_1 be the index in $\{1,\dots,n\}$ such that $(\mathbf{X}_{(1)}(\mathbf{X}), Y_{(1)}(\mathbf{X})) = (\mathbf{X}_{\sigma_1}, Y_{\sigma_1})$. Then

$$\begin{aligned}\mathbb{E}\left|Y_{(1)}(\mathbf{X}) - r\left(\mathbf{X}_{(1)}(\mathbf{X})\right)\right|^2 &= \sum_{j=1}^{n} \mathbb{E}\left[\mathbb{1}_{[\sigma_1=j]}\left|Y_j - r(\mathbf{X}_j)\right|^2\right] \\ &= \sum_{j=1}^{n} \mathbb{E}\left[\mathbb{1}_{[\sigma_1=j]}\mathbb{E}\left[\left|Y_j - r(\mathbf{X}_j)\right|^2 \,\middle|\, \mathbf{X}, \mathbf{X}_1, \dots, \mathbf{X}_n\right]\right],\end{aligned}$$

where, in the second equality, we used the fact that σ_1 is measurable with respect to $\mathbf{X}, \mathbf{X}_1, \dots, \mathbf{X}_n$. Thus, using the independence between the pair $(\mathbf{X}_j, Y_j)$ and the observations $\mathbf{X}, \mathbf{X}_1, \dots, \mathbf{X}_{j-1}, \mathbf{X}_{j+1}, \dots, \mathbf{X}_n$, we are led to

$$\begin{aligned}\mathbb{E}\left|Y_{(1)}(\mathbf{X}) - r\left(\mathbf{X}_{(1)}(\mathbf{X})\right)\right|^2 &= \sum_{j=1}^{n} \mathbb{E}\left[\mathbb{1}_{[\sigma_1=j]}\mathbb{E}\left[\left|Y_j - r(\mathbf{X}_j)\right|^2 \,\middle|\, \mathbf{X}_j\right]\right] \\ &= \sum_{j=1}^{n} \mathbb{E}\left[\mathbb{1}_{[\sigma_1=j]}\sigma^2(\mathbf{X}_j)\right] \\ &= \mathbb{E}[\sigma^2(\mathbf{X}_{(1)}(\mathbf{X}))],\end{aligned}$$

as desired. □

We are now in a position to prove Theorem 9.1.

Proof (Theorem 9.1). We have

$$\begin{aligned}\mathbb{E}\left|r_n(\mathbf{X}) - r(\mathbf{X})\right|^2 &= \mathbb{E}\left|Y_{(1)}(\mathbf{X}) - r\left(\mathbf{X}_{(1)}(\mathbf{X})\right) + r\left(\mathbf{X}_{(1)}(\mathbf{X})\right) - r(\mathbf{X})\right|^2 \\ &= \mathbb{E}\left|Y_{(1)}(\mathbf{X}) - r\left(\mathbf{X}_{(1)}(\mathbf{X})\right)\right|^2 + \mathbb{E}\left|r\left(\mathbf{X}_{(1)}(\mathbf{X})\right) - r(\mathbf{X})\right|^2 \\ &\quad + 2\mathbb{E}\left[\left(Y_{(1)}(\mathbf{X}) - r\left(\mathbf{X}_{(1)}(\mathbf{X})\right)\right)\left(r\left(\mathbf{X}_{(1)}(\mathbf{X})\right) - r(\mathbf{X})\right)\right].\end{aligned}$$

The second term on the right-hand side tends to zero by Lemma 9.2 and the fact that $\mathbb{E}r^2(\mathbf{X}) \le \mathbb{E}Y^2 < \infty$. Define $\sigma^2(\mathbf{x}) = \mathbb{E}[|Y - r(\mathbf{X})|^2 \,|\, \mathbf{X} = \mathbf{x}]$, and note that the assumption $\mathbb{E}Y^2 < \infty$ implies $\mathbb{E}\sigma^2(\mathbf{X}) = \mathbb{E}|Y - r(\mathbf{X})|^2 < \infty$. According to Lemma 9.3, the first term is $\mathbb{E}[\sigma^2(\mathbf{X}_{(1)}(\mathbf{X}))]$, and it tends to $\mathbb{E}\sigma^2(\mathbf{X})$, again by Lemma 9.2. Finally, by the Cauchy-Schwarz inequality,

$$\begin{aligned}&\mathbb{E}^2\left[\left(Y_{(1)}(\mathbf{X}) - r\left(\mathbf{X}_{(1)}(\mathbf{X})\right)\right)\left(r\left(\mathbf{X}_{(1)}(\mathbf{X})\right) - r(\mathbf{X})\right)\right] \\ &\quad \le \mathbb{E}\left[\sigma^2\left(\mathbf{X}_{(1)}(\mathbf{X})\right)\right] \times \mathbb{E}\left|r\left(\mathbf{X}_{(1)}(\mathbf{X})\right) - r(\mathbf{X})\right|^2 \\ &\quad = \left(\mathbb{E}\sigma^2(\mathbf{X}) + \mathrm{o}(1)\right) \times \mathrm{o}(1) \\ &\quad = \mathrm{o}(1).\end{aligned}$$

□

Chapter 10
L^p-consistency and Stone's theorem

10.1 L^p-consistency

We know that, whenever $\mathbb{E}Y^2 < \infty$, the regression function $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ achieves the minimal value $L^\star = \mathbb{E}|Y - r(\mathbf{X})|^2$ of the L^2 risk over all square-integrable functions of $\mathbf{X}$. It is also easy to show, using the independence of $(\mathbf{X}, Y)$ and the sample $\mathscr{D}_n$, that the (conditional) L^2 risk $\mathbb{E}[|Y - r_n(\mathbf{X})|^2 \mid \mathscr{D}_n]$ of an estimate r_n of r satisfies

$$\mathbb{E}\big[\, |Y - r_n(\mathbf{X})|^2 \mid \mathscr{D}_n\big] = \mathbb{E}\,|Y - r(\mathbf{X})|^2 + \int_{\mathbb{R}^d} |r_n(\mathbf{x}) - r(\mathbf{x})|^2 \,\mu(\mathrm{d}\mathbf{x}),$$

where μ is the distribution of $\mathbf{X}$. This identity reveals that the L^2 risk of the estimate r_n is close to the optimal value if and only if the L^2 error $\int_{\mathbb{R}^d} |r_n(\mathbf{x}) - r(\mathbf{x})|^2 \mu(\mathrm{d}\mathbf{x})$ is close to zero. Therefore, the L^2 error (integrated squared error) is a nice criterion to measure the quality of an estimate.

Since r_n is a function of the data set $\mathscr{D}_n$, the L^2 error is itself a random variable. Most often, one is interested in the convergence to zero of the expectation of this random variable. The estimate r_n is said to be (globally) L^2-consistent (or mean integrated squared error consistent) if

$$\mathbb{E}\left[\int_{\mathbb{R}^d} |r_n(\mathbf{x}) - r(\mathbf{x})|^2 \,\mu(\mathrm{d}\mathbf{x})\right] \to 0 \quad \text{as } n \to \infty,$$

where the expectation $\mathbb{E}$ is evaluated with respect to the sample $\mathscr{D}_n$. Taking expectation with respect to both $\mathbf{X}$ and $\mathscr{D}_n$, this can be rewritten in a more compact form as

$$\mathbb{E}\,|r_n(\mathbf{X}) - r(\mathbf{X})|^2 \to 0.$$

G. Biau, L. Devroye, *Lectures on the Nearest Neighbor Method*,
Springer Series in the Data Sciences, DOI 10.1007/978-3-319-25388-6_10

More generally, denoting by p a positive real number and assuming that $\mathbb{E}|Y|^p < \infty$, the estimate r_n is (globally) L^p-consistent if

$$\mathbb{E}\,|r_n(\mathbf{X}) - r(\mathbf{X})|^p \to 0. \tag{10.1}$$

Note that (10.1) is a global proximity measure—from (10.1) we can merely conclude that $\liminf_{n\to\infty} \mathbb{E}|r_n(\mathbf{x}) - r(\mathbf{x})|^p = 0$ at μ-almost all $\mathbf{x}$ by Fatou's lemma. This does not imply that $r_n(\mathbf{x}) \to r(\mathbf{x})$ in probability.

Our main goal in this chapter is to show that the (raw) nearest neighbor regression function estimate defined in Chapter 8 is universally L^p-consistent, i.e., that (10.1) holds for all distributions of $(\mathbf{X}, Y)$ with $\mathbb{E}|Y|^p < \infty$. This result is a consequence of a fundamental theorem of Stone (1977), which provides necessary and sufficient conditions for consistency in L^p of general local averaging regression estimates. This is the topic of the next section.

10.2 Stone's theorem

A local averaging estimate is any estimate of the regression function that can be written as

$$r_n(\mathbf{x}) = \sum_{i=1}^{n} W_{ni}(\mathbf{x}) Y_i, \quad \mathbf{x} \in \mathbb{R}^d, \tag{10.2}$$

where $(W_{n1}(\mathbf{x}), \ldots, W_{nn}(\mathbf{x}))$ is a weight vector and each $W_{ni}(\mathbf{x})$ is a Borel measurable function of $\mathbf{x}$ and $\mathbf{X}_1, \ldots, \mathbf{X}_n$ (not $Y_1, \ldots, Y_n$). It is intuitively clear that the pairs $(\mathbf{X}_i, Y_i)$ such that $\mathbf{X}_i$ is "close" to $\mathbf{x}$ should provide more information about $r(\mathbf{x})$ than those "far" from $\mathbf{x}$. Therefore, the weights are typically larger in the neighborhood of $\mathbf{x}$, so that $r_n(\mathbf{x})$ is roughly a (weighted) mean of the Y_i's corresponding to $\mathbf{X}_i$'s in the neighborhood of $\mathbf{x}$. Thus, r_n can be viewed as a local averaging estimate. Often, but not always, the $W_{ni}(\mathbf{x})$'s are nonnegative and sum to one, so that $(W_{n1}(\mathbf{x}), \ldots, W_{nn}(\mathbf{x}))$ is a probability vector.

An example is the kernel estimate (Nadaraya, 1964, 1965; Watson, 1964), which is obtained by letting

$$W_{ni}(\mathbf{x}) = \frac{K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{X}_j}{h}\right)},$$

where K is a given nonnegative measurable function on $\mathbb{R}^d$ (called the kernel), and h is a positive number (called the bandwidth) depending upon n only. Put differently, for $\mathbf{x} \in \mathbb{R}^d$,

$$r_n(\mathbf{x}) = \frac{\sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right) Y_i}{\sum_{j=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{X}_j}{h}\right)}.$$

If both denominator and numerator are zero, then we set $r_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n Y_i$. Such a strategy ensures that r_n is a linear function of the Y_i's: if all Y_i's are replaced by $aY_i + b$, then $r_n(\mathbf{x})$ is replaced by $ar_n(\mathbf{x}) + b$. In particular, for the so-called naive kernel $K(\mathbf{z}) = \mathbb{1}_{[\|\mathbf{z}\|\le 1]}$, one obtains

$$r_n(\mathbf{x}) = \frac{\sum_{i=1}^n \mathbb{1}_{[\|\mathbf{x}-\mathbf{X}_i\|\le h]} Y_i}{\sum_{j=1}^n \mathbb{1}_{[\|\mathbf{x}-\mathbf{X}_j\|\le h]}},$$

i.e., $r(\mathbf{x})$ is estimated by averaging the Y_i's such that the distance between $\mathbf{x}$ and $\mathbf{X}_i$ is not larger than h. For a more general kernel K, the weight of Y_i (i.e., the influence of Y_i on the value of the estimate at $\mathbf{x}$) depends on the distance between $\mathbf{x}$ and $\mathbf{X}_i$ through the kernel shape. Popular kernels include the Epanechnikov kernel $K(\mathbf{z}) = (1 - \|\mathbf{z}\|^2)1_{[\|\mathbf{z}\|\le 1]}$ and the Gaussian kernel $K(\mathbf{z}) = e^{-\|\mathbf{z}\|^2}$.

A second important example is the nearest neighbor regression function estimate r_n, which we have introduced in Chapter 8. Recall that

$$r_n(\mathbf{x}) = \sum_{i=1}^n v_{ni} Y_{(i)}(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d, \tag{10.3}$$

where $(v_{n1}, \ldots, v_{nn})$ is a given weight vector summing to one, and the sequence $(\mathbf{X}_{(1)}(\mathbf{x}), Y_{(1)}(\mathbf{x})), \ldots, (\mathbf{X}_{(n)}(\mathbf{x}), Y_{(n)}(\mathbf{x}))$ is a permutation of $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ according to increasing values of $\|\mathbf{X}_i - \mathbf{x}\|$ (as usual, distance ties are broken by looking at indices). We see that the nearest neighbor estimate is indeed of the form (10.2), since it is obtained by putting

$$W_{ni}(\mathbf{x}) = v_{n\Sigma_i},$$

where $(\Sigma_1, \ldots, \Sigma_n)$ is a permutation of $(1, \ldots, n)$ such that $\mathbf{X}_i$ is the Σ_i-th nearest neighbor of $\mathbf{x}$ for all i.

Stone's theorem (Stone, 1977) offers general necessary and sufficient conditions on the weights in order to guarantee the universal L^p-consistency of local averaging estimates.

Theorem 10.1 (Stone, 1977). *Consider the following five conditions:*

(i) There is a constant C such that, for every Borel measurable function $g : \mathbb{R}^d \to \mathbb{R}$ with $\mathbb{E}|g(\mathbf{X})| < \infty$,

$$\mathbb{E}\left[\sum_{i=1}^n |W_{ni}(\mathbf{X})|\,|g(\mathbf{X}_i)|\right] \le C\,\mathbb{E}\,|g(\mathbf{X})| \quad \textit{for all } n \ge 1.$$

(ii) There is a constant $D \ge 1$ such that

$$\mathbb{P}\left\{\sum_{i=1}^n |W_{ni}(\mathbf{X})| \le D\right\} = 1 \quad \textit{for all } n \ge 1.$$

(iii) For all $a > 0$,

$$\sum_{i=1}^{n} |W_{ni}(\mathbf{X})| \, \mathbb{1}_{[\|\mathbf{X}_i - \mathbf{X}\| > a]} \to 0 \quad \textit{in probability.}$$

(iv) One has

$$\sum_{i=1}^{n} W_{ni}(\mathbf{X}) \to 1 \quad \textit{in probability.}$$

(v) One has

$$\max_{1 \le i \le n} |W_{ni}(\mathbf{X})| \to 0 \quad \textit{in probability.}$$

If (i)–(v) are satisfied for any distribution of $\mathbf{X}$, then the corresponding regression function estimate r_n is universally L^p-consistent ($p \ge 1$), that is,

$$\mathbb{E} \, |r_n(\mathbf{X}) - r(\mathbf{X})|^p \to 0$$

for all distributions of $(\mathbf{X}, Y)$ with $\mathbb{E}|Y|^p < \infty$, $p \ge 1$.

Suppose, conversely, that r_n is universally L^p-consistent. Then (iv) and (v) hold for any distribution of $\mathbf{X}$. Moreover, if the weights are nonnegative for all $n \ge 1$, then (iii) is satisfied. Finally, if the weights are nonnegative for all $n \ge 1$ and (ii) holds, then (i) holds as well.

Remark 10.1. It easily follows from the Lebesgue dominated convergence theorem that condition (v) of Theorem 10.1 may be replaced by

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}^2(\mathbf{X}) \right] \to 0. \qquad \square$$

Before we prove Theorem 10.1 in the next section, some comments are in order. Condition (i) is merely technical. It says in particular that for nonnegative weights and for a nonnegative, noiseless model (i.e., $Y = r(\mathbf{X}) \ge 0$ with probability one), the mean value of the estimate is bounded from above by some constant times the mean value of the regression function. The attentive reader may note that (i) is the condition that permits one to avoid placing a continuity assumption on r. Conditions (ii) and (iv) state that the sum of the weights is bounded and is asymptotically one. Condition (iii) requires that the overall weight of $\mathbf{X}_i$'s outside any ball of fixed radius centered at $\mathbf{X}$ must go to zero. In other words, it ensures that the estimate at a point $\mathbf{X}$ is asymptotically mostly influenced by the data close to $\mathbf{X}$. Finally, condition (v) states that asymptotically all weights become small. Thus, no single observation has a too large contribution to the estimate, so that the number of points encountered in the averaging must tend to infinity.

If $(W_{n1}(\mathbf{X}), \ldots, W_{nn}(\mathbf{X}))$ is a probability vector, then (ii) and (iv) hold automatically and the three remaining conditions are necessary and sufficient for consistency. This useful result is summarized in the following corollary.

Corollary 10.1. *Assume that the weights are nonnegative and sum to one:*

$$W_{ni}(\mathbf{X}) \geq 0 \quad \textit{and} \quad \sum_{i=1}^{n} W_{ni}(\mathbf{X}) = 1.$$

Then the corresponding regression function estimate is universally L^p-consistent $(p \geq 1)$ if and only if the following three conditions are satisfied for any distribution of $\mathbf{X}$*:*

(i) There is a constant C such that, for every Borel measurable function $g\colon \mathbb{R}^d \to \mathbb{R}$ with $\mathbb{E}|g(\mathbf{X})| < \infty$,

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(\mathbf{X}) \, |g(\mathbf{X}_i)| \right] \leq C \, \mathbb{E} \, |g(\mathbf{X})| \quad \textit{for all } n \geq 1.$$

(ii) For all $a > 0$,

$$\sum_{i=1}^{n} W_{ni}(\mathbf{X}) \mathbb{1}_{[\|\mathbf{X}_i - \mathbf{X}\| > a]} \to 0 \quad \textit{in probability.}$$

(iii) One has

$$\max_{1 \leq i \leq n} W_{ni}(\mathbf{X}) \to 0 \quad \textit{in probability.}$$

Technical condition (i) in Theorem 10.1 may be hard to verify for some families of weights. However, this requirement can be bypassed for a bounded Y, at the price of a stronger assumption on the regression function. Recall the definition

$$\|Y\|_\infty = \inf \big\{ t \geq 0 : \mathbb{P}\{|Y| > t\} = 0 \big\}.$$

Corollary 10.2. *Assume that $\|Y\|_\infty < \infty$ and that the regression function is uniformly continuous on $\mathbb{R}^d$. Assume, in addition, that for any distribution of* $\mathbf{X}$*, the weights satisfy the following four conditions:*

(i) There is a constant $D \geq 1$ such that

$$\mathbb{P}\left\{ \sum_{i=1}^{n} |W_{ni}(\mathbf{X})| \leq D \right\} = 1 \quad \textit{for all } n \geq 1.$$

(ii) For all $a > 0$,

$$\sum_{i=1}^{n} |W_{ni}(\mathbf{X})| \, \mathbb{1}_{[\|\mathbf{X}_i - \mathbf{X}\| > a]} \to 0 \quad \textit{in probability.}$$

(iii) One has

$$\sum_{i=1}^{n} W_{ni}(\mathbf{X}) \to 1 \quad \textit{in probability.}$$

(iv) One has

$$\max_{1 \le i \le n} |W_{ni}(\mathbf{X})| \to 0 \quad \textit{in probability.}$$

Then the corresponding regression function estimate is L^p ($p \ge 1$) consistent.

Proof. Verify that the conclusions of Lemma 10.1 and Lemma 10.2 in the next section hold without condition (i) of Theorem 10.1 as soon as $\|Y\|_\infty < \infty$ and r is uniformly continuous. □

10.3 Proof of Stone's theorem

For the sake of clarity, the proof of Theorem 10.1 is divided in a series of five lemmas. The first two lemmas concern the sufficiency, whereas the last three ones pertain to the necessity.

Lemma 10.1. *Let $p \ge 1$. Assume that $\mathbb{E}|Y|^p < \infty$. If conditions (i)-(ii) and (v) of Theorem 10.1 are satisfied, then*

$$\mathbb{E}\left| \sum_{i=1}^{n} W_{ni}(\mathbf{X}) \, (Y_i - r(\mathbf{X}_i)) \right|^p \to 0.$$

Proof. We first show that the conclusion of the lemma holds when $p = 2$. The general case is then obtained through a truncation argument. Define $\sigma^2(\mathbf{x}) = \mathbb{E}[|Y - r(\mathbf{X})|^2 \,|\, \mathbf{X} = \mathbf{x}]$, and note that the assumption $\mathbb{E}Y^2 < \infty$ implies $\mathbb{E}\sigma^2(\mathbf{X}) < \infty$. Next, write

$$\mathbb{E}\left| \sum_{i=1}^{n} W_{ni}(\mathbf{X}) \, (Y_i - r(\mathbf{X}_i)) \right|^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{E}\big[W_{ni}(\mathbf{X}) W_{nj}(\mathbf{X}) \, (Y_i - r(\mathbf{X}_i)) \, \big(Y_j - r(\mathbf{X}_j)\big) \big].$$

For $i \neq j$,

$$\begin{aligned}
&\mathbb{E}\big[W_{ni}(\mathbf{X})W_{nj}(\mathbf{X})\,(Y_i - r(\mathbf{X}_i))\,\big(Y_j - r(\mathbf{X}_j)\big)\,\big] \\
&\quad = \mathbb{E}\big[\mathbb{E}[W_{ni}(\mathbf{X})W_{nj}(\mathbf{X})\,(Y_i - r(\mathbf{X}_i))\,\big(Y_j - r(\mathbf{X}_j)\big)\;|\,\mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n, Y_i]\big] \\
&\quad = \mathbb{E}\big[W_{ni}(\mathbf{X})W_{nj}(\mathbf{X})\,(Y_i - r(\mathbf{X}_i))\,\mathbb{E}\left[Y_j - r(\mathbf{X}_j)\,|\,\mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n, Y_i\right]\big] \\
&\quad = \mathbb{E}\big[W_{ni}(\mathbf{X})W_{nj}(\mathbf{X})\,(Y_i - r(\mathbf{X}_i))\,\big(r(\mathbf{X}_j) - r(\mathbf{X}_j)\big)\,\big] \\
&\quad = 0.
\end{aligned}$$

In the third equality, we used the independence between the pair $(\mathbf{X}_j, Y_j)$ and $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_{j-1}, \mathbf{X}_{j+1}, \ldots, \mathbf{X}_n, Y_i$. Hence,

$$\begin{aligned}
\mathbb{E}\left|\sum_{i=1}^{n} W_{ni}(\mathbf{X})\,(Y_i - r(\mathbf{X}_i))\right|^2 &= \mathbb{E}\left[\sum_{i=1}^{n} W_{ni}^2(\mathbf{X})\,|Y_i - r(\mathbf{X}_i)|^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} W_{ni}^2(\mathbf{X})\sigma^2(\mathbf{X}_i)\right]. \qquad (10.4)
\end{aligned}$$

By Stone's conditions (ii) and (v), and the dominated convergence theorem,

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}^2(\mathbf{X})\right] \to 0 \qquad (10.5)$$

(see Remark 10.1). If Y is bounded, then so is σ^2, and (10.5) implies the desired result. For general $\sigma^2(\mathbf{x})$ and $\varepsilon > 0$, a denseness argument (cf. Theorem 20.17 in the Appendix) reveals that there exists a bounded Borel measurable function $\sigma_\varepsilon^2(\mathbf{x}) \leq L$ such that $\mathbb{E}|\sigma^2(\mathbf{X}) - \sigma_\varepsilon^2(\mathbf{X})| \leq \varepsilon$. Thus, by (ii),

$$\begin{aligned}
&\mathbb{E}\left|\sum_{i=1}^{n} W_{ni}(\mathbf{X})\,(Y_i - r(\mathbf{X}_i))\right|^2 \\
&\quad \leq \mathbb{E}\left[\sum_{i=1}^{n} W_{ni}^2(\mathbf{X})\,\left|\sigma^2(\mathbf{X}_i) - \sigma_\varepsilon^2(\mathbf{X}_i)\right|\right] + \mathbb{E}\left[\sum_{i=1}^{n} W_{ni}^2(\mathbf{X})\sigma_\varepsilon^2(\mathbf{X}_i)\right] \\
&\quad \leq D\,\mathbb{E}\left[\sum_{i=1}^{n} \left|W_{ni}(\mathbf{X})\right|\,\left|\sigma^2(\mathbf{X}_i) - \sigma_\varepsilon^2(\mathbf{X}_i)\right|\right] + L\,\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}^2(\mathbf{X})\right].
\end{aligned}$$

Thus, using (i) and (10.5), we obtain

$$\begin{aligned}
\limsup_{n\to\infty} \mathbb{E}\left|\sum_{i=1}^{n} W_{ni}(\mathbf{X})\,(Y_i - r(\mathbf{X}_i))\right|^2 &\leq CD\,\mathbb{E}\left|\sigma^2(\mathbf{X}) - \sigma_\varepsilon^2(\mathbf{X})\right| \\
&\leq CD\varepsilon.
\end{aligned}$$

Since ε is arbitrary, the lemma is proved for $p = 2$.

Consider now the general case $p \geq 1$. Given a positive number M, set $Y^{(M)} = Y\mathbb{1}_{[|Y|\leq M]}$ and $r^{(M)}(\mathbf{x}) = \mathbb{E}[Y^{(M)}|\mathbf{X} = \mathbf{x}]$. Using the elementary inequality $|a+b+c|^p \leq 3^{p-1}(|a|^p + |b|^p + |c|^p)$ valid for $p \geq 1$, we may write

$$\begin{aligned}\mathbb{E}\left|\sum_{i=1}^n W_{ni}(\mathbf{X})\,(Y_i - r(\mathbf{X}_i))\right|^p &\leq 3^{p-1}\mathbb{E}\left|\sum_{i=1}^n W_{ni}(\mathbf{X})\big(Y_i - Y_i^{(M)}\big)\right|^p \\ &\quad + 3^{p-1}\mathbb{E}\left|\sum_{i=1}^n W_{ni}(\mathbf{X})\big(Y_i^{(M)} - r^{(M)}(\mathbf{X}_i)\big)\right|^p \\ &\quad + 3^{p-1}\mathbb{E}\left|\sum_{i=1}^n W_{ni}(\mathbf{X})\big(r^{(M)}(\mathbf{X}_i) - r(\mathbf{X}_i)\big)\right|^p. \end{aligned} \tag{10.6}$$

For $p > 1$, by the triangle and Hölder's inequalities, setting $q = p/(p-1)$, we have

$$\begin{aligned}\mathbb{E}\left|\sum_{i=1}^n W_{ni}(\mathbf{X})\big(Y_i - Y_i^{(M)}\big)\right|^p &\leq \mathbb{E}\left[\left(\sum_{i=1}^n \big|W_{ni}(\mathbf{X})\big|\big|Y_i - Y_i^{(M)}\big|\right)^p\right] \\ &= \mathbb{E}\left[\left(\sum_{i=1}^n \big|W_{ni}(\mathbf{X})\big|^{1/q}\big|W_{ni}(\mathbf{X})\big|^{1/p}\big|Y_i - Y_i^{(M)}\big|\right)^p\right] \\ &\leq \mathbb{E}\left[\left(\sum_{i=1}^n \big|W_{ni}(\mathbf{X})\big|\right)^{p/q}\sum_{i=1}^n \big|W_{ni}(\mathbf{X})\big|\big|Y_i - Y_i^{(M)}\big|^p\right].\end{aligned}$$

Thus, by conditions (i)–(ii),

$$\mathbb{E}\left|\sum_{i=1}^n W_{ni}(\mathbf{X})\big(Y_i - Y_i^{(M)}\big)\right|^p \leq CD^{p/q}\,\mathbb{E}\big|Y - Y^{(M)}\big|^p.$$

On the other hand, for $p = 1$,

$$\mathbb{E}\left|\sum_{i=1}^n W_{ni}(\mathbf{X})\big(Y_i - Y_i^{(M)}\big)\right| \leq C\,\mathbb{E}\big|Y - Y^{(M)}\big|.$$

Using the fact that $\mathbb{E}|Y - Y^{(M)}|^p \to 0$ as $M \to \infty$, we conclude that, for all $p \geq 1$,

$$\lim_{M\to\infty}\sup_{n\geq 1}\mathbb{E}\left|\sum_{i=1}^n W_{ni}(\mathbf{X})\big(Y_i - Y_i^{(M)}\big)\right|^p = 0.$$

Similarly, for $p > 1$,

$$\begin{aligned}\mathbb{E}\left|\sum_{i=1}^{n} W_{ni}(\mathbf{X})\big(r^{(M)}(\mathbf{X}_i) - r(\mathbf{X}_i)\big)\right|^p &\le CD^{p/q}\,\mathbb{E}\big|r(\mathbf{X}) - r^{(M)}(\mathbf{X})\big|^p \\ &= CD^{p/q}\,\mathbb{E}\big|\mathbb{E}[Y - Y^{(M)} \mid \mathbf{X}]\big|^p \\ &\le CD^{p/q}\,\mathbb{E}\big|Y - Y^{(M)}\big|^p \\ &\quad \text{(by Jensen's inequality)},\end{aligned}$$

whereas for $p = 1$,

$$\mathbb{E}\left|\sum_{i=1}^{n} W_{ni}(\mathbf{X})\big(r^{(M)}(\mathbf{X}_i) - r(\mathbf{X}_i)\big)\right| \le C\,\mathbb{E}\big|Y - Y^{(M)}\big|.$$

For $p \ge 1$, the term $\mathbb{E}|Y - Y^{(M)}|^p$ approaches zero as $M \to \infty$, uniformly in n. It follows that the first and third terms in (10.6) can be made arbitrary small for all M large enough, independently of n. Thus, to prove that the conclusion of the lemma holds for Y, it is enough to show that it holds for $Y^{(M)}$. In other words, without loss of generality, it can be assumed that Y is bounded. But if Y is bounded, then to prove the result for all $p \ge 1$, it is enough to show that it is true for $p = 2$. Since this has already been done, the proof of Lemma 10.1 is complete. □

Lemma 10.2. *Let $p \ge 1$, and let $g : \mathbb{R}^d \to \mathbb{R}$ be a Borel measurable function such that $\mathbb{E}|g(\mathbf{X})|^p < \infty$. If conditions (i)–(iii) of Theorem 10.1 are satisfied, then*

$$\mathbb{E}\left|\sum_{i=1}^{n} W_{ni}(\mathbf{X})\,(g(\mathbf{X}_i) - g(\mathbf{X}))\right|^p \to 0.$$

Proof. For $p > 1$, by the triangle and Hölder's inequalities, setting $q = p/(p-1)$, we have

$$\begin{aligned}&\mathbb{E}\left|\sum_{i=1}^{n} W_{ni}(\mathbf{X})\,(g(\mathbf{X}_i) - g(\mathbf{X}))\right|^p \\ &\le \mathbb{E}\left[\left(\sum_{i=1}^{n} |W_{ni}(\mathbf{X})|\,|g(\mathbf{X}_i) - g(\mathbf{X})|\right)^p\right] \\ &= \mathbb{E}\left[\left(\sum_{i=1}^{n} |W_{ni}(\mathbf{X})|^{1/q}\,|W_{ni}(\mathbf{X})|^{1/p}\,|g(\mathbf{X}_i) - g(\mathbf{X})|\right)^p\right] \\ &\le \mathbb{E}\left[\left(\sum_{i=1}^{n} |W_{ni}(\mathbf{X})|\right)^{p/q} \sum_{i=1}^{n} |W_{ni}(\mathbf{X})|\,|g(\mathbf{X}_i) - g(\mathbf{X})|^p\right].\end{aligned}$$

Thus, by condition (ii),

$$\mathbb{E}\left| \sum_{i=1}^{n} W_{ni}(\mathbf{X}) \left(g(\mathbf{X}_i) - g(\mathbf{X})\right) \right|^p \leq D^{p/q}\, \mathbb{E}\left[\sum_{i=1}^{n} |W_{ni}(\mathbf{X})|\, |g(\mathbf{X}_i) - g(\mathbf{X})|^p \right].$$

On the other hand, for $p = 1$,

$$\mathbb{E}\left| \sum_{i=1}^{n} W_{ni}(\mathbf{X}) \left(g(\mathbf{X}_i) - g(\mathbf{X})\right) \right| \leq \mathbb{E}\left[\sum_{i=1}^{n} |W_{ni}(\mathbf{X})|\, |g(\mathbf{X}_i) - g(\mathbf{X})| \right].$$

Therefore, the result is proved if we show that, for all $p \geq 1$,

$$\mathbb{E}\left[\sum_{i=1}^{n} |W_{ni}(\mathbf{X})|\, |g(\mathbf{X}_i) - g(\mathbf{X})|^p \right] \to 0.$$

Choose $\varepsilon > 0$. Let g_ε be a continuous function on $\mathbb{R}^d$ having compact support and such that $\mathbb{E}|g(\mathbf{X}) - g_\varepsilon(\mathbf{X})|^p \leq \varepsilon$ (such a choice is possible for $p \geq 1$ by a denseness argument—see Theorem 20.17 in the Appendix). Because $|a+b+c|^p \leq 3^{p-1}(|a|^p + |b|^p + |c|^p)$, we may write

$$\begin{aligned} \mathbb{E}\left[\sum_{i=1}^{n} |W_{ni}(\mathbf{X})|\, |g(\mathbf{X}_i) - g(\mathbf{X})|^p \right] &\leq 3^{p-1}\mathbb{E}\left[\sum_{i=1}^{n} |W_{ni}(\mathbf{X})|\, |g(\mathbf{X}_i) - g_\varepsilon(\mathbf{X}_i)|^p \right] \\ &\quad + 3^{p-1}\mathbb{E}\left[\sum_{i=1}^{n} |W_{ni}(\mathbf{X})|\, |g_\varepsilon(\mathbf{X}_i) - g_\varepsilon(\mathbf{X})|^p \right] \\ &\quad + 3^{p-1}\mathbb{E}\left[\sum_{i=1}^{n} |W_{ni}(\mathbf{X})|\, |g_\varepsilon(\mathbf{X}) - g(\mathbf{X})|^p \right]. \end{aligned} \tag{10.7}$$

By condition (i),

$$\mathbb{E}\left[\sum_{i=1}^{n} |W_{ni}(\mathbf{X})|\, |g(\mathbf{X}_i) - g_\varepsilon(\mathbf{X}_i)|^p \right] \leq C\, \mathbb{E}\, |g(\mathbf{X}) - g_\varepsilon(\mathbf{X})|^p \leq C\varepsilon.$$

Moreover, it follows from (ii) that

$$\mathbb{E}\left[\sum_{i=1}^{n} |W_{ni}(\mathbf{X})|\, |g_\varepsilon(\mathbf{X}) - g(\mathbf{X})|^p \right] \leq D\, \mathbb{E}\, |g_\varepsilon(\mathbf{X}) - g(\mathbf{X})|^p \leq D\varepsilon.$$

Thus, to prove that the conclusion of Lemma 10.2 holds for g, it suffices to prove that the second term in (10.7) can be made arbitrarily small as n tends to infinity.

Since g_ε is continuous with compact support, it is bounded and uniformly continuous as well. Choose $\delta > 0$. There is an $a > 0$ such that $|g_\varepsilon(\mathbf{y}) - g_\varepsilon(\mathbf{x})|^p \le \delta$ if $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{y} \in \mathbb{R}^d$, and $\|\mathbf{y} - \mathbf{x}\| \le a$. Then, by (ii),

$$\mathbb{E}\left[\sum_{i=1}^n |W_{ni}(\mathbf{X})| \, |g_\varepsilon(\mathbf{X}_i) - g_\varepsilon(\mathbf{X})|^p \right] \le 2^p \|g_\varepsilon\|_\infty^p \, \mathbb{E}\left[\sum_{i=1}^n |W_{ni}(\mathbf{X})| \, \mathbb{1}_{[\|\mathbf{X}_i - \mathbf{X}\| > a]} \right] + D\delta.$$

It follows from (ii) and (iii) that

$$\mathbb{E}\left[\sum_{i=1}^n |W_{ni}(\mathbf{X})| \, \mathbb{1}_{[\|\mathbf{X}_i - \mathbf{X}\| > a]} \right] \to 0.$$

Thus

$$\limsup_{n \to \infty} \mathbb{E}\left[\sum_{i=1}^n |W_{ni}(\mathbf{X})| \, |g_\varepsilon(\mathbf{X}_i) - g_\varepsilon(\mathbf{X})|^p \right] \le D\delta.$$

Putting all the pieces together, we obtain

$$\limsup_{n \to \infty} \mathbb{E}\left[\sum_{i=1}^n |W_{ni}(\mathbf{X})| \, |g(\mathbf{X}_i) - g(\mathbf{X})|^p \right] \le 3^{p-1} \left((C + D)\varepsilon + D\delta \right).$$

Since ε and δ are arbitrarily small, the conclusion of Lemma 10.2 holds. □

Lemma 10.3. *Assume that the weights satisfy the following property: there is a sequence $Z_1, \ldots, Z_n$ of independent standard normal random variables such that $(Z_1, \ldots, Z_n)$ is independent of $(\mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n)$, and*

$$\sum_{i=1}^n W_{ni}(\mathbf{X}) Z_i \to 0 \quad \textit{in probability.}$$

Then

$$\max_{1 \le i \le n} |W_{ni}(\mathbf{X})| \to 0 \quad \textit{in probability.}$$

Proof. The conditional distribution of $\sum_{i=1}^n W_{ni}(\mathbf{X}) Z_i$ given $\mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n$ is normal with zero mean and variance $\sum_{i=1}^n W_{ni}^2(\mathbf{X})$. Thus, for $\varepsilon > 0$,

$$\begin{aligned} \mathbb{P}\left\{ \left| \sum_{i=1}^n W_{ni}(\mathbf{X}) Z_i \right| > \varepsilon \right\} &= \mathbb{P}\left\{ |Z_1| \sqrt{\sum_{i=1}^n W_{ni}^2(\mathbf{X})} > \varepsilon \right\} \\ &\ge \left(\frac{2}{\sqrt{2\pi}} \int_1^\infty e^{-x^2/2} \mathrm{d}x \right) \mathbb{P}\left\{ \sum_{i=1}^n W_{ni}^2(\mathbf{X}) > \varepsilon^2 \right\}, \end{aligned}$$

and hence

$$\mathbb{P}\left\{ \sum_{i=1}^{n} W_{ni}^2(\mathbf{X}) > \varepsilon^2 \right\} \to 0.$$

The conclusion of the lemma follows from $\max_{1\le i\le n} W_{ni}^2(\mathbf{X}) \le \sum_{i=1}^{n} W_{ni}^2(\mathbf{X})$. □

Lemma 10.4. *Assume that the weights are nonnegative and that, for every bounded and continuous nonnegative function $g : \mathbb{R}^d \to \mathbb{R}$,*

$$\sum_{i=1}^{n} W_{ni}(\mathbf{X}) g(\mathbf{X}_i) \to g(\mathbf{X}) \quad \textit{in probability.}$$

Then condition (iii) of Theorem 10.1 is satisfied.

Proof. Let $a > 0$ be given. Choose $\mathbf{x}_0 \in \mathbb{R}^d$, and let g be a bounded and continuous nonnegative function on $\mathbb{R}^d$ such that $g(\mathbf{x}) = 0$ for $\|\mathbf{x} - \mathbf{x}_0\| \le a/3$ and $g(\mathbf{x}) = 1$ for $\|\mathbf{x} - \mathbf{x}_0\| \ge 2a/3$. Then on the event $[\|\mathbf{X} - \mathbf{x}_0\| \le a/3]$, $g(\mathbf{X}) = 0$ and

$$\sum_{i=1}^{n} W_{ni}(\mathbf{X}) g(\mathbf{X}_i) \ge \sum_{i=1}^{n} W_{ni}(\mathbf{X}) \mathbb{1}_{[\|\mathbf{X}_i - \mathbf{X}\| > a]}.$$

Therefore,

$$\mathbb{1}_{[\|\mathbf{X}-\mathbf{x}_0\|\le a/3]} \sum_{i=1}^{n} W_{ni}(\mathbf{X}) \mathbb{1}_{[\|\mathbf{X}_i - \mathbf{X}\| > a]} \to 0 \quad \text{in probability.}$$

Thus, since any compact subset of $\mathbb{R}^d$ can be covered by a finite number of closed balls of radius $a/3$, we conclude that for every compact subset B of $\mathbb{R}^d$,

$$\mathbb{1}_B(\mathbf{X}) \sum_{i=1}^{n} W_{ni}(\mathbf{X}) \mathbb{1}_{[\|\mathbf{X}_i - \mathbf{X}\| > a]} \to 0 \quad \text{in probability.}$$

Therefore, since $\mathbb{P}\{\|\mathbf{X}\| > M\}$ tends to zero as $M \to \infty$, (*iii*) holds as desired. □

Lemma 10.5. *Assume that the weights are nonnegative and satisfy the following property: for every Borel measurable function $g : \mathbb{R}^d \to \mathbb{R}$ such that $\mathbb{E}|g(\mathbf{X})| < \infty$,*

$$\limsup_{n\to\infty} \mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(\mathbf{X}) \, |g(\mathbf{X}_i)| \right] < \infty.$$

Then there is a positive integer n_0 and a positive constant C such that, for every Borel measurable function $g : \mathbb{R}^d \to \mathbb{R}$ with $\mathbb{E}|g(\mathbf{X})| < \infty$,

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(\mathbf{X}) \, |g(\mathbf{X}_i)| \right] \le C \, \mathbb{E}\,|g(\mathbf{X})| \quad \textit{for all } n \ge n_0.$$

Proof. Suppose that the conclusion of the lemma is false. Then there is a strictly increasing sequence $\{n_\ell\}$ of positive integers and a sequence $\{g_\ell\}$ of Borel measurable functions on $\mathbb{R}^d$ such that $\mathbb{E}|g_\ell(\mathbf{X})| = 2^{-\ell}$ and

$$\mathbb{E}\left[\sum_{i=1}^{n_\ell} W_{n_\ell i}(\mathbf{X})\,|g_\ell(\mathbf{X}_i)|\right] \geq \ell.$$

Set $g = \sum_{\ell=1}^{\infty} |g_\ell|$. Then g is a measurable function on $\mathbb{R}^d$, $\mathbb{E}|g(\mathbf{X})| = 1 < \infty$, and

$$\mathbb{E}\left[\sum_{i=1}^{n_\ell} W_{n_\ell i}(\mathbf{X})\,|g(\mathbf{X}_i)|\right] \geq \mathbb{E}\left[\sum_{i=1}^{n_\ell} W_{n_\ell i}(\mathbf{X})\,|g_\ell(\mathbf{X}_i)|\right] \geq \ell.$$

Thus

$$\limsup_{n\to\infty} \mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(\mathbf{X})\,|g(\mathbf{X}_i)|\right] = \infty,$$

which contradicts the hypothesis. Therefore the lemma is valid. □

We are now ready to prove Theorem 10.1.

Proof (Theorem 10.1). Using the inequality $|a+b+c|^p \leq 3^{p-1}(|a|^p + |b|^p + |c|^p)$, we may write

$$\begin{aligned}\mathbb{E}\,|r_n(\mathbf{X}) - r(\mathbf{X})|^p \leq\ & 3^{p-1}\mathbb{E}\left|\sum_{i=1}^{n} W_{ni}(\mathbf{X})\,(Y_i - r(\mathbf{X}_i))\right|^p \\ & + 3^{p-1}\mathbb{E}\left|\sum_{i=1}^{n} W_{ni}(\mathbf{X})\,(r(\mathbf{X}_i) - r(\mathbf{X}))\right|^p \qquad (10.8)\\ & + 3^{p-1}\mathbb{E}\left|\Big(\sum_{i=1}^{n} W_{ni}(\mathbf{X}) - 1\Big) r(\mathbf{X})\right|^p.\end{aligned}$$

The first term of inequality (10.8) tends to zero by Lemma 10.1, whereas the second one tends to zero by Lemma 10.2. Concerning the third term, we have

$$\mathbb{E}\left|\Big(\sum_{i=1}^{n} W_{ni}(\mathbf{X}) - 1\Big) r(\mathbf{X})\right|^p \to 0$$

by conditions (ii), (iv) and the Lebesgue dominated convergence theorem.

To prove the second assertion of Theorem 10.1, first note that if r_n is L^p-consistent for some $p \geq 1$ and all distributions of $(\mathbf{X}, Y)$ with $\mathbb{E}|Y|^p < \infty$, then

$$\sum_{i=1}^{n} W_{ni}(\mathbf{X}) Y_i \to r(\mathbf{X}) \quad \text{in probability.} \qquad (10.9)$$

The first two necessity statements follow from (10.9). The first one is shown by taking $Y = 1$, independently of $\mathbf{X}$, so that $r(\mathbf{X}) = 1$. The second one is an implication of (10.9) and Lemma 10.3, by letting Y be a standard normal random variable independent of $\mathbf{X}$, for which in particular $r(\mathbf{X}) = 0$.

To prove the third necessity assertion, assume that the weights are nonnegative and take $Y = r(\mathbf{X}) = g(\mathbf{X})$, where g is a bounded and continuous nonnegative function on $\mathbb{R}^d$. The conclusion immediately follows from (10.9) and Lemma 10.4.

The last necessity result is implied by Lemma 10.5. To see this, assume that the weights are nonnegative, and let g be a Borel measurable function on $\mathbb{R}^d$ such that $\mathbb{E}|g(\mathbf{X})| < \infty$. Take $Y = r(\mathbf{X}) = |g(\mathbf{X})|$. Then, by the triangle inequality,

$$\left| \mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}(\mathbf{X}) \, |g(\mathbf{X}_i)| \Big] - \mathbb{E} \, |g(\mathbf{X})| \right| \le \mathbb{E} \left| \sum_{i=1}^{n} W_{ni}(\mathbf{X}) \, |g(\mathbf{X}_i)| - |g(\mathbf{X})| \right|.$$

This inequality and the universal L^1-consistency of the estimate imply

$$\limsup_{n \to \infty} \mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}(\mathbf{X}) \, |g(\mathbf{X}_i)| \Big] < \infty.$$

Therefore, by Lemma 10.5,

$$\mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}(\mathbf{X}) \, |g(\mathbf{X}_i)| \Big] \le C \, \mathbb{E} \, |g(\mathbf{X})| \quad \text{for all } n \ge n_0.$$

This inequality is true for all $n \ge 1$ (with a different constant) if condition (*ii*) is satisfied. □

10.4 The nearest neighbor estimate

In this section, we apply Stone's theorem 10.1 to the nearest neighbor regression function estimate (10.3).

Theorem 10.2 (Universal L^p-consistency). *Let $(v_{n1}, \dots, v_{nn})$ be a probability weight vector such that $v_{n1} \ge \cdots \ge v_{nn}$ for all n. Then the corresponding nearest neighbor regression function estimate is universally L^p-consistent ($p \ge 1$) if and only if there exists a sequence of integers $\{k\} = \{k_n\}$ such that*

$$\begin{array}{ll} (i) & k \to \infty \quad \text{and} \quad k/n \to 0; \\ (ii) & \sum_{i>k} v_{ni} \to 0; \\ (iii) & v_{n1} \to 0. \end{array} \tag{10.10}$$

Remark 10.2. Conditions (*i*) and (*ii*) of Theorem 10.2 may be replaced by the following equivalent one: for all $\varepsilon > 0$, $\sum_{i>\varepsilon n} v_{ni} \to 0$ (see Lemma 20.3 in the Appendix for a proof of this equivalence result). □

For the standard k-nearest neighbor estimate, $v_{ni} = 1/k$ for $1 \le i \le k$ and $v_{ni} = 0$ otherwise, where $\{k\} = \{k_n\}$ is a sequence of positive integers not exceeding n.

Corollary 10.3. *The k-nearest neighbor regression function estimate is universally L^p-consistent ($p \ge 1$) if and only if $k \to \infty$ and $k/n \to 0$.*

This is a nice result, since no condition on $(\mathbf{X}, Y)$ other than $\mathbb{E}|Y|^p < \infty$ is required. This type of distribution-free result is called universal. The concept of universal consistency is important because the use of a nonparametric estimate is usually a consequence of the partial or total lack of information regarding the distribution of $(\mathbf{X}, Y)$. Since in many situations we do not have any prior knowledge about this distribution, it is therefore essential to design estimates that perform well for all distributions.

The crucial results needed to prove Theorem 10.2 are gathered in the next two lemmas. If $\mathbf{x}, \mathbf{x}_1, \ldots, \mathbf{x}_n$ are elements of $\mathbb{R}^d$ and $(v_{n1}, \ldots, v_{nn})$ is a weight vector, we let

$$W_{ni}(\mathbf{x}; \mathbf{x}_1, \ldots, \mathbf{x}_n) = v_{nk}, \quad 1 \le i \le n,$$

whenever $\mathbf{x}_i$ is the k-th nearest neighbor of $\mathbf{x}$ from among $\mathbf{x}_1, \ldots, \mathbf{x}_n$ (distance ties are broken by comparing indices).

Lemma 10.6. *Let $\mathbf{x}, \mathbf{x}_1, \ldots, \mathbf{x}_n$ be vectors of $\mathbb{R}^d$, and let $(v_{n1}, \ldots, v_{nn})$ be a probability weight vector such that $v_{n1} \ge \cdots \ge v_{nn}$ for all n. Then*

$$\sum_{i=1}^{n} W_{ni}(\mathbf{x}_i; \mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_n) \le 2\gamma_d,$$

where γ_d is the minimal number of cones of angle $\pi/12$ that cover $\mathbb{R}^d$.

Proof. By Theorem 20.15, there exists a finite collection of minimal cardinality $\mathscr{C}_1, \ldots, \mathscr{C}_{\gamma_d}$ of cones of angle $\pi/12$, with different central directions, such that their union covers $\mathbb{R}^d$:

$$\bigcup_{\ell=1}^{\gamma_d} \mathscr{C}_\ell = \mathbb{R}^d.$$

Similarly, for $\mathbf{x} \in \mathbb{R}^d$,

$$\bigcup_{\ell=1}^{\gamma_d} (\mathbf{x} + \mathscr{C}_\ell) = \mathbb{R}^d.$$

To facilitate the notation, we let $\mathscr{C}_\ell(\mathbf{x}) = \mathbf{x} + \mathscr{C}_\ell$, and denote by $\mathscr{C}_\ell^\star(\mathbf{x})$ the translated cone $\mathscr{C}_\ell(\mathbf{x})$ minus the point $\mathbf{x}$. Introduce the set $A = \{i : 1 \le i \le n, \mathbf{x}_i = \mathbf{x}\}$, and let $|A|$ be its cardinality.

Let $\mathbf{x}_i$ be a point falling in $\mathscr{C}_\ell^\star(\mathbf{x})$. The key observation is that if $\mathbf{x}_i$ is the nearest neighbor of $\mathbf{x}$ from among those $\mathbf{x}_1, \ldots, \mathbf{x}_n$ that belong to $\mathscr{C}_\ell^\star(\mathbf{x})$, then $\mathbf{x}$ can be the nearest neighbor of only $\mathbf{x}_i$—or, equivalently, it cannot be the nearest neighbor of any other point $\mathbf{x}_j$ in $\mathscr{C}_\ell^\star(\mathbf{x})$. To see this, just note that if $\|\mathbf{x}_i - \mathbf{x}\| > 0$ and $\|\mathbf{x}_i - \mathbf{x}\| \le \|\mathbf{x}_j - \mathbf{x}\|$, then, by Lemma 20.5, $\|\mathbf{x}_i - \mathbf{x}_j\| < \|\mathbf{x} - \mathbf{x}_j\|$.

By induction, we conclude that if $\mathbf{x}_i$ is the $(|A| + k)$-th nearest neighbor of $\mathbf{x}$ among those $\mathbf{x}_1, \ldots, \mathbf{x}_n$ that belong to $\mathscr{C}_\ell(\mathbf{x})$, then $\mathbf{x}$ is at least the k-th nearest neighbor of $\mathbf{x}_i$ among $\{\mathbf{x}, \mathbf{x}_j : j \neq i \text{ and } \mathbf{x}_j \in \mathscr{C}_\ell^\star(\mathbf{x})\}$. So, using the monotonicity condition on the v_{ni}'s, we have

$$\sum_{i:\mathbf{x}_i \in \mathscr{C}_\ell^\star(\mathbf{x})} W_{ni}(\mathbf{x}_i; \mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_n) \le \sum_{i=1}^{n-|A|} v_{ni} \le 1.$$

Similarly,

$$\sum_{i \in A} W_{ni}(\mathbf{x}_i; \mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_n) = \sum_{i=1}^{|A|} v_{ni} \le 1.$$

Since there are γ_d cones, the lemma is proved. □

Lemma 10.7 (Stone's lemma). *Let $(v_{n1}, \ldots, v_{nn})$ be a probability weight vector such that $v_{n1} \ge \cdots \ge v_{nn}$ for all n. Then, for every Borel measurable function $g : \mathbb{R}^d \to \mathbb{R}$ such that $\mathbb{E}|g(\mathbf{X})| < \infty$,*

$$\mathbb{E}\left[\sum_{i=1}^{n} v_{ni} \left| g\left(\mathbf{X}_{(i)}(\mathbf{X})\right) \right| \right] \le 2\gamma_d \mathbb{E}\left|g(\mathbf{X})\right|,$$

where γ_d is the minimal number of cones of angle $\pi/12$ that cover $\mathbb{R}^d$.

Proof. Let g be a Borel measurable function on $\mathbb{R}^d$ such that $\mathbb{E}|g(\mathbf{X})| < \infty$. Notice that

$$\mathbb{E}\left[\sum_{i=1}^{n} v_{ni} \left| g\left(\mathbf{X}_{(i)}(\mathbf{X})\right) \right| \right] = \mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(\mathbf{X}) \left|g(\mathbf{X}_i)\right| \right],$$

where, with the notation of Lemma 10.6, $W_{ni}(\mathbf{X}) = W_{ni}(\mathbf{X}; \mathbf{X}_1, \ldots, \mathbf{X}_n)$. Therefore,

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n} v_{ni} \left| g\left(\mathbf{X}_{(i)}(\mathbf{X})\right) \right| \right] &= \mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(\mathbf{X}; \mathbf{X}_1, \ldots, \mathbf{X}_n) \left|g(\mathbf{X}_i)\right| \right] \\
&= \mathbb{E}\left[\left|g(\mathbf{X})\right| \sum_{i=1}^{n} W_{ni}(\mathbf{X}_i; \mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_n) \right] \\
&\qquad \text{(by symmetrization)}
\end{aligned}$$

and thus, by Lemma 10.6,

$$\mathbb{E}\left[\sum_{i=1}^{n} v_{ni}\left|g\left(\mathbf{X}_{(i)}(\mathbf{X})\right)\right|\right] \leq 2\gamma_d \mathbb{E}\left|g(\mathbf{X})\right|.$$

□

Proof (Theorem 10.2).

The sufficiency. We proceed by checking the conditions of Corollary 10.1. The weights $W_{ni}(\mathbf{X})$ in Corollary 10.1 are obtained by putting

$$W_{ni}(\mathbf{X}) = v_{n\Sigma_i},$$

where $(\Sigma_1, \ldots, \Sigma_n)$ is a permutation of $(1, \ldots, n)$ such that $\mathbf{X}_i$ is the Σ_i-th nearest neighbor of $\mathbf{X}$ for all i. Condition (iii) is obvious according to (10.10)(iii). For condition (ii), take n so large that $1 \leq k < n$ (this is possible in view of the first requirement of (10.10)), and observe that, for each $a > 0$,

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(\mathbf{X})\mathbb{1}_{[\|\mathbf{X}_i-\mathbf{X}\|>a]}\right] &= \mathbb{E}\left[\sum_{i=1}^{n} v_{ni}\mathbb{1}_{[\|\mathbf{X}_{(i)}(\mathbf{X})-\mathbf{X}\|>a]}\right] \\
&= \int_{\mathbb{R}^d} \mathbb{E}\left[\sum_{i=1}^{n} v_{ni}\mathbb{1}_{[\|\mathbf{X}_{(i)}(\mathbf{x})-\mathbf{x}\|>a]}\right]\mu(\mathrm{d}\mathbf{x}) \\
&\leq \int_{\mathbb{R}^d} \mathbb{P}\left\{\|\mathbf{X}_{(k)}(\mathbf{x}) - \mathbf{x}\| > a\right\}\mu(\mathrm{d}\mathbf{x}) + \sum_{i>k} v_{ni}.
\end{aligned}$$

Thus, using (10.10)(ii), the second condition of Corollary 10.1 is satisfied when

$$\int_{\mathbb{R}^d} \mathbb{P}\left\{\|\mathbf{X}_{(k)}(\mathbf{x}) - \mathbf{x}\| > a\right\}\mu(\mathrm{d}\mathbf{x}) \to 0.$$

But, by Lemma 2.2 and the Lebesgue dominated convergence theorem, this is true for all $a > 0$ whenever $k/n \to 0$. Let us finally consider condition (i) of Corollary 10.1. We have to show that for any Borel measurable function g on $\mathbb{R}^d$ such that $\mathbb{E}|g(\mathbf{X})| < \infty$,

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(\mathbf{X})\left|g(\mathbf{X}_i)\right|\right] = \mathbb{E}\left[\sum_{i=1}^{n} v_{ni}\left|g\left(\mathbf{X}_{(i)}(\mathbf{X})\right)\right|\right] \leq C\,\mathbb{E}\left|g(\mathbf{X})\right|,$$

for all $n \geq 1$ and some constant C. We have precisely proved in Lemma 10.7 that this inequality always holds with $C = 2\gamma_d$. Thus, condition (i) is verified.

The necessity. By the necessity part of Stone's theorem 10.1, we already know that the requirement $v_{n1} = \max_i v_{ni} \to 0$ is necessary. Thus, it remains to show that conditions (10.10)(*i*) and (10.10)(*ii*) are necessary as well. According to Remark 10.2, this is equivalent to proving that, for all $\varepsilon > 0$, $\sum_{i>\varepsilon n} v_{ni} \to 0$.

Let X be uniform on $[0, 1]$, and let $Y = X^2$. Take $0 < x < 1/2$, and define $Z_n = \sum_{i=1}^n v_{ni}(X_{(i)}^2(x) - x^2)$. Note that $y^2 - x^2 = (y - x)^2 + 2x(y - x)$. Since $x \le 1/2$, $\mathbb{E}[X_{(i)}(x) - x] \ge 0$ (which we leave as an exercise). Thus,

$$\mathbb{E}\left[X_{(i)}^2(x) - x^2\right] \ge \mathbb{E}\left|X_{(i)}(x) - x\right|^2.$$

Now, for $t > 0$, $t < x$,

$$\mathbb{P}\left\{|X_{(i)}(x) - x| > t\right\} = \mathbb{P}\left\{\text{Bin}(n, 2t) < i\right\}$$

and for $t \ge x$,

$$\mathbb{P}\left\{|X_{(i)}(x) - x| > t\right\} = \mathbb{P}\{X_{(i)}(0) > x + t\} = \mathbb{P}\left\{\text{Bin}(n, x + t) < i\right\}.$$

By the duality between the binomial and beta distributions (see Section 20.9 in the Appendix),

$$\mathbb{P}\left\{\text{Bin}(n, 2t) < i\right\} = \mathbb{P}\left\{\text{Beta}(i, n + 1 - i) > 2t\right\}, \quad 0 < t < x,$$

and

$$\mathbb{P}\left\{\text{Bin}(n, x + t) < i\right\} = \mathbb{P}\left\{\text{Beta}(i, n + 1 - i) > t + x\right\}, \quad t \ge x.$$

In any case, for all $t > 0$,

$$\mathbb{P}\left\{|X_{(i)}(x) - x| > t\right\} \ge \mathbb{P}\left\{\text{Beta}(i, n + 1 - i) > 2t\right\}.$$

Recalling that $\mathbb{E}Z = \int_0^\infty \mathbb{P}\{Z > t\}\mathrm{d}t$ for any nonnegative random variable Z, we obtain

$$\begin{aligned}\mathbb{E}|X_{(i)}(x) - x|^2 &\ge \frac{1}{4}\mathbb{E}\left[\text{Beta}^2(i, n + 1 - i)\right] \\ &= \frac{1}{4} \times \frac{i(i + 1)}{(n + 1)(n + 2)}.\end{aligned}$$

Therefore,

$$\mathbb{E}Z_n = \sum_{i=1}^n v_{ni}\mathbb{E}\left[X_{(i)}^2(x) - x^2\right] \ge \sum_{i=1}^n \left[v_{ni} \times \frac{i(i + 1)}{4(n + 1)(n + 2)}\right] \ge \sum_{i=1}^n \frac{i^2 v_{ni}}{4(n + 2)^2}.$$

It suffices now to show that $Z_{n'}$ cannot converge to zero in probability along a subsequence n' of n when for some $\varepsilon > 0$, $\delta > 0$,

$$\sum_{i>\varepsilon n'} v_{n'i} \geq \delta > 0$$

along this subsequence. Indeed,

$$\sum_{i=1}^{n'} i^2 v_{n'i} \geq \sum_{i>\varepsilon n'} i^2 v_{n'i} \geq (\varepsilon n' + 1)^2 \delta,$$

whence

$$\mathbb{E}Z_{n'} \geq \frac{(\varepsilon n' + 1)^2 \delta}{4(n' + 2)^2}.$$

Also,

$$\begin{aligned} \mathbb{P}\left\{ Z_{n'} > \frac{(\varepsilon n' + 1)^2 \delta}{8(n' + 2)^2} \right\} &\geq \mathbb{P}\{Z_{n'} > \mathbb{E}Z_{n'}/2\} \\ &\geq \mathbb{E}Z_{n'}/2 \\ &\geq \frac{(\varepsilon n' + 1)^2 \delta}{8(n' + 2)^2} \sim \frac{\delta \varepsilon^2}{8}, \end{aligned}$$

where the second inequality uses technical Lemma 10.8 below. Therefore, for all $\varepsilon > 0$, $\sum_{i>\varepsilon n} v_{ni} \to 0$ as $n \to \infty$. □

Lemma 10.8. *If $0 \leq a \leq 1$ and $0 < c$ are constants, then any $[0, c]$-valued random variable Z satisfies*

$$\mathbb{P}\{Z > a\mathbb{E}Z\} \geq \frac{1 - a}{c} \mathbb{E}Z.$$

Proof. Just note that

$$\mathbb{E}Z = \mathbb{E}[Z\mathbb{1}_{[Z \leq a\mathbb{E}Z]}] + \mathbb{E}[Z\mathbb{1}_{[Z > a\mathbb{E}Z]}] \leq a\mathbb{E}Z + c\mathbb{P}\{Z > a\mathbb{E}Z\}.$$

□

Remark 10.3 (Nonmonotone weights). Devroye (1981a) proved the following theorem (notation u^+ means $\max(u, 0)$):

Theorem 10.3. *Let $p \geq 1$, and let $(v_{n1}, \ldots, v_{nn})$ be a probability weight vector. Assume that $\mathbb{E}[|Y|^p \log^+ |Y|] < \infty$ and that there exists a sequence of integers $\{k\} = \{k_n\}$ such that*

$$\begin{aligned} &(i) \quad k \to \infty \quad \textit{and} \quad k/n \to 0; \\ &(ii) \quad v_{ni} = 0 \quad \textit{when } i > k; \\ &(iii) \quad \sup_n (k \max_i v_{ni}) < \infty. \end{aligned} \tag{10.11}$$

Then the corresponding nearest neighbor regression function estimate is L^p-consistent.

The condition put on Y in Theorem 10.3 is stricter than the condition $\mathbb{E}|Y|^p < \infty$ needed in Theorem 10.2. However, the conditions on the sequence of weights are not strictly nested: the monotonicity constraint is absent in Theorem 10.3, but (10.11)(*ii*) is stricter than (10.10)(*ii*). □

Chapter 11
Pointwise consistency

11.1 Weak pointwise consistency

Theorem 11.1 below is a slight extension of a theorem due to Devroye (1981a). It offers sufficient conditions on the probability weight vector guaranteeing that the (raw) nearest neighbor estimate (8.2) satisfies, for all $p \geq 1$,

$$\mathbb{E}\,|r_n(\mathbf{x}) - r(\mathbf{x})|^p \to 0 \quad \text{at } \mu\text{-almost all } \mathbf{x},$$

under the sole requirement that $\mathbb{E}|Y|^p < \infty$. Since convergence in L^1 implies convergence in probability, this theorem also shows that the nearest neighbor estimate is universally weakly consistent at μ-almost all $\mathbf{x}$, provided $\mathbb{E}|Y| < \infty$.

It is assumed throughout that $v_{ni} \geq 0$, $1 \leq i \leq n$, and $\sum_{i=1}^n v_{ni} = 1$. As in the previous chapters, we let μ be the distribution of $\mathbf{X}$.

Theorem 11.1 (Universal weak pointwise consistency). *Let $p \geq 1$. Assume that $\mathbb{E}|Y|^p < \infty$ and that there exists a sequence of integers $\{k\} = \{k_n\}$ such that*

$$\begin{array}{ll} (i) & k \to \infty \quad \text{and} \quad k/n \to 0; \\ (ii) & v_{ni} = 0 \quad \text{when } i > k; \\ (iii) & \sup_n (k \max_i v_{ni}) < \infty. \end{array} \tag{11.1}$$

Then the corresponding nearest neighbor regression function estimate r_n satisfies

$$\mathbb{E}\,|r_n(\mathbf{x}) - r(\mathbf{x})|^p \to 0 \quad \text{at } \mu\text{-almost all } \mathbf{x} \in \mathbb{R}^d.$$

In particular, the nearest neighbor estimate is universally weakly consistent at μ-almost all $\mathbf{x}$, that is,

$$r_n(\mathbf{x}) \to r(\mathbf{x}) \quad \text{in probability at } \mu\text{-almost all } \mathbf{x} \in \mathbb{R}^d$$

for all distributions of $(\mathbf{X}, Y)$ with $\mathbb{E}|Y| < \infty$.

G. Biau, L. Devroye, *Lectures on the Nearest Neighbor Method*,
Springer Series in the Data Sciences, DOI 10.1007/978-3-319-25388-6_11

As an important by-product, Theorem 11.1 implies that the standard k-nearest neighbor estimate ($v_{ni} = 1/k$ for $1 \le i \le k$ and $v_{ni} = 0$ otherwise) is universally weakly pointwise consistent at μ-almost all $\mathbf{x}$ when $k \to \infty$ and $k/n \to 0$. Other examples are the triangular weight k-nearest neighbor estimate ($v_{ni} = (k-i+1)/b_n$ for $1 \le i \le k$ and $v_{ni} = 0$ otherwise, where $b_n = k(k+1)/2$) and the quadratic weight estimate ($v_{ni} = (k^2 - (i-1)^2)/b_n$ for $1 \le i \le k$ and $v_{ni} = 0$ otherwise, where $b_n = k(k+1)(4k-1)/6$).

Corollary 11.1. *If $k \to \infty$ and $k/n \to 0$, then the k-nearest neighbor regression function estimate is universally weakly consistent at μ-almost all $\mathbf{x} \in \mathbb{R}^d$.*

The elementary result needed to prove Theorem 11.1 is Lemma 11.1 below. When the distribution of $\mathbf{X}$ is continuous, the proof is easy. However, in the general case, we have to take care of the messy problem of distance ties, which introduces additional technical difficulties.

Lemma 11.1. *Let $p \ge 1$, and let $g : \mathbb{R}^d \to \mathbb{R}$ be a Borel measurable function such that $\mathbb{E}|g(\mathbf{X})|^p < \infty$. Assume that there exists a sequence of integers $\{k\} = \{k_n\}$ such that*

$$
\begin{array}{ll}
(i) & k/n \to 0; \\
(ii) & v_{ni} = 0 \quad \textit{when } i > k; \\
(iii) & \sup_n (k \max_i v_{ni}) < \infty.
\end{array}
$$

Then

$$
\mathbb{E}\left| \sum_{i=1}^{n} v_{ni} g\left(\mathbf{X}_{(i)}(\mathbf{x})\right) - g(\mathbf{x}) \right|^p \to 0 \quad \textit{at } \mu\textit{-almost all } \mathbf{x} \in \mathbb{R}^d.
$$

Proof. Take $\mathbf{x}$ such that $\mathbf{x} \in \mathrm{supp}(\mu)$, and that $\mathbf{x}$ satisfies

$$
\max\left(\frac{1}{\mu\left(B(\mathbf{x},\rho)\right)} \int_{B(\mathbf{x},\rho)} |g(\mathbf{y}) - g(\mathbf{x})|^p \, \mu(\mathrm{d}\mathbf{y}), \right.
$$
$$
\left. \frac{1}{\mu\left(B^\circ(\mathbf{x},\rho)\right)} \int_{B^\circ(\mathbf{x},\rho)} |g(\mathbf{y}) - g(\mathbf{x})|^p \, \mu(\mathrm{d}\mathbf{y}) \right) \to 0 \quad \text{as } \rho \downarrow 0
$$

(where B° is the open ball). Since $\mathbb{E}|g(\mathbf{X})|^p < \infty$, we have by the generalized version of the Lebesgue differentiation theorem (see Theorem 20.19 in the Appendix) that μ-almost all $\mathbf{x}$ satisfy this property. Fix $\varepsilon > 0$ and choose $\delta > 0$ such that the maximum above is smaller than ε for all $0 < \rho \le \delta$.

Observe, by Jensen's inequality (which is valid here since $(v_{n1}, \ldots, v_{nn})$ is a probability vector and $p \ge 1$) and conditions (ii)–(iii), that for some constant $\alpha > 0$,

$$
\mathbb{E}\left| \sum_{i=1}^{n} v_{ni} g\left(\mathbf{X}_{(i)}(\mathbf{x})\right) - g(\mathbf{x}) \right|^p \le \frac{\alpha}{k} \mathbb{E}\left[\sum_{j=1}^{n} \left|g(\mathbf{X}_j) - g(\mathbf{x})\right|^p \mathbb{1}_{[\Sigma_j \le k]} \right], \tag{11.2}
$$

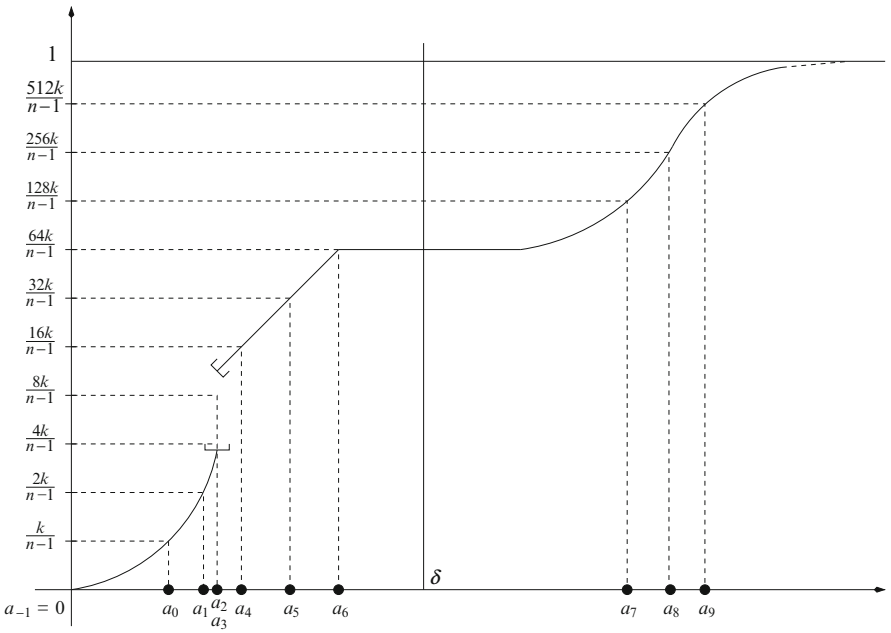

Fig. 11.1 An example of function F and the associated a_i's.

where Σ_j is the rank of $\mathbf{X}_j$ with respect to the distances to $\mathbf{x}$ if ties are broken by index comparisons (see Chapter 8). The objective is to bound (11.2) and show that it can be made as small as desired as $n \to \infty$.

Define $Z = \|\mathbf{X} - \mathbf{x}\|$, and let $F(z) = \mathbb{P}\{Z \le z\}$, $F_0(z) = \mathbb{P}\{Z < z\}$, $p(z) = F(z) - F_0(z) = \mathbb{P}\{Z = z\}$. For $u \in [0, 1]$, define

$$F^{-1}(u) = \inf\{t \ge 0 : F(t) \ge u\}.$$

Assume that $n \ge 2$ and let $I = \max\{i \ge 0 : \frac{2^i k}{n-1} < 1\}$. The sequence of points of interest to us are $a_{-1} = 0 \le a_0 \le a_1 \le a_2 \le \cdots \le a_I$, where

$$a_i = F^{-1}\left(\frac{2^i k}{n-1}\right), \quad i \ge 0$$

(see Figure 11.1 for an example).

For $i \ge 0$, we have

$$F_0(a_i) \le \frac{2^i k}{n-1} \le F(a_i) = F_0(a_i) + p(a_i).$$

It is possible to have $a_i = a_{i+1}$ (because of possible atoms in F), and this causes the principal technical problem. We address this by introducing the set of large atoms, A:

$$A = \left\{z \ge 0 : p(z) \ge \frac{1}{4}F(z)\right\}.$$

So, we have $z \in A$ if and only if $F_0(z) \leq \frac{3}{4}F(z)$. The range of z, $[0, \infty)$, is partitioned as follows, for fixed $\delta > 0$:

$$\{0\} \cup \bigcup_{i \geq 0 : a_i \in A, a_i \leq \delta} (a_{i-1}, a_i) \cup \bigcup_{i \geq 0 : a_i \notin A, a_i \leq \delta} (a_{i-1}, a_i] \cup (A \cap (0, \delta]) \cup (\delta, \infty). \tag{11.3}$$

We implicitly assume that $\delta > 0$ is one of the a_i's—this is always possible for all n large enough since $k/n \to 0$, as soon as $\mathbf{x}$ is not an isolated atom of μ. (If $\mathbf{x}$ is an isolated atom, we just write

$$[0, \infty) = \{0\} \cup (0, \delta] \cup (\delta, \infty),$$

and observe that, for δ small enough, $\mathbb{P}\{\|\mathbf{X} - \mathbf{x}\| \in (0, \delta]\} = 0$. Thus, the analysis is simpler in this case and left to the reader.)

We define

$$\psi(z) = \frac{1}{\mu\left(\partial B(\mathbf{x}, z)\right)} \int_{\partial B(\mathbf{x}, z)} |g(\mathbf{y}) - g(\mathbf{x})|^p \, \mu(\mathrm{d}\mathbf{y}),$$

where $\partial B(\mathbf{x}, z)$ is the boundary of the closed ball $B(\mathbf{x}, z)$ centered at $\mathbf{x}$ of radius z: $\partial B(\mathbf{x}, z) = \{\mathbf{y} \in \mathbb{R}^d : \|\mathbf{y} - \mathbf{x}\| = z\}$. Note that $F(z) = \mu(B(\mathbf{x}, z))$ and $p(z) = \mu(\partial B(\mathbf{x}, z))$. Thus, if $z \in A$, $0 < z \leq \delta$, we have

$$\psi(z) \leq \frac{1}{\frac{1}{4}\mu\left(B(\mathbf{x}, z)\right)} \int_{B(\mathbf{x}, z)} |g(\mathbf{y}) - g(\mathbf{x})|^p \, \mu(\mathrm{d}\mathbf{y}) \leq 4\varepsilon,$$

by choice of δ. Now, for each set B in the partition (11.3), we bound

$$h(B) \stackrel{\text{def}}{=} \frac{\alpha}{k} \mathbb{E}\left[\sum_{j=1}^{n} \left| g(\mathbf{X}_j) - g(\mathbf{x}) \right|^p \mathbb{1}_{[\Sigma_j \leq k]} \mathbb{1}_{[\|\mathbf{X}_j - \mathbf{x}\| \in B]} \right].$$

Clearly, $h(\{0\}) = 0$. Assume next $B = A \cap (0, \delta]$. Then

$$\begin{aligned} h(B) &= \sum_{z \in B} \frac{\alpha}{k} \mathbb{E}\left[\sum_{j=1}^{n} \left| g(\mathbf{X}_j) - g(\mathbf{x}) \right|^p \mathbb{1}_{[\Sigma_j \leq k]} \mathbb{1}_{[\|\mathbf{X}_j - \mathbf{x}\| = z]} \right] \\ &= \sum_{z \in B} \frac{\alpha}{k} \psi(z) \mathbb{E}\left[\sum_{j=1}^{n} \mathbb{1}_{[\Sigma_j \leq k]} \mathbb{1}_{[\|\mathbf{X}_j - \mathbf{x}\| = z]} \right] \\ &\qquad \text{(using our index-based distance tie-breaking convention)} \\ &\leq \frac{4\alpha\varepsilon}{k} \mathbb{E}\left[\sum_{j=1}^{n} \mathbb{1}_{[\Sigma_j \leq k]} \mathbb{1}_{[\|\mathbf{X}_j - \mathbf{x}\| \in B]} \right] \end{aligned}$$

$$\leq \frac{4\alpha\varepsilon}{k}\mathbb{E}\left[\sum_{j=1}^{n}\mathbb{1}_{[\Sigma_j\leq k]}\right]$$

$$= 4\alpha\varepsilon.$$

Consider now $B_i = (a_{i-1}, a_i)$, $a_i \in A$, $a_i \leq \delta$, $i \geq 0$. Clearly, by our way of breaking distance ties,

$$h(B_i) \leq \frac{\alpha n}{k}\mathbb{E}\left[|g(\mathbf{X}_1) - g(\mathbf{x})|^p\, \mathbb{1}_{[\Sigma_1\leq k]}\mathbb{1}_{[\|\mathbf{X}_1-\mathbf{x}\|\in B_i]}\right].$$

Thus, for $i \geq 1$, after conditioning on $\mathbf{X}_1$,

$$h(B_i) \leq \frac{\alpha n}{k}\mathbb{E}\left[|g(\mathbf{X}) - g(\mathbf{x})|^p\, \mathbb{1}_{[\|\mathbf{X}-\mathbf{x}\|\in B_i]}\mathbb{1}_{[\mathrm{Bin}(n-1,\,F(a_{i-1}))<k]}\right]$$

(where the binomial random variable is independent of $\mathbf{X}_1$)

$$\leq \frac{\alpha n}{k}\int_{B_i}|g(\mathbf{y}) - g(\mathbf{x})|^p\,\mu(\mathrm{d}\mathbf{y}) \times \mathbb{P}\left\{\mathrm{Bin}\big(n-1,\,\tfrac{2^{i-1}k}{n-1}\big) < k\right\}$$

(since $F(a_{i-1}) \geq \frac{2^{i-1}k}{n-1}$).

(Note that we use the notation $\int_{B_i}$ as a shortcut for $\int_{\{\mathbf{y}:\|\mathbf{y}-\mathbf{x}\|\in B_i\}}$.) Similarly, for $i = 0$,

$$h(B_0) \leq \frac{\alpha n}{k}\int_{B_0}|g(\mathbf{y}) - g(\mathbf{x})|^p\,\mu(\mathrm{d}\mathbf{y}).$$

Now, by choice of δ, for $i \geq 0$,

$$\begin{aligned}\int_{B_i}|g(\mathbf{y}) - g(\mathbf{x})|^p\,\mu(\mathrm{d}\mathbf{y}) &\leq \int_{B^\circ(\mathbf{0},a_i)}|g(\mathbf{y}) - g(\mathbf{x})|^p\,\mu(\mathrm{d}\mathbf{y}) \\ &\leq \varepsilon\mu\left(B^\circ(\mathbf{0}, a_i)\right) \\ &\leq \varepsilon\frac{2^i k}{n-1} \\ &\quad \text{(since } F_0(a_i) \leq \tfrac{2^i k}{n-1}\text{)}.\end{aligned} \tag{11.4}$$

Therefore, for $i \geq 1$,

$$h(B_i) \leq 2^i\alpha\varepsilon\frac{n}{n-1} \times \mathbb{P}\left\{\mathrm{Bin}\big(n-1,\,\tfrac{2^{i-1}k}{n-1}\big) < k\right\},$$

and for $i = 0$,

$$h(B_0) \leq \alpha\varepsilon\frac{n}{n-1}.$$

Applying Chernoff's bound (Theorem 20.5 in the Appendix), we have

$$\begin{aligned}\mathbb{P}\left\{\mathrm{Bin}\big(n-1, \tfrac{2^{i-1}k}{n-1}\big) < k\right\} &\le \exp(k - 2^{i-1}k + k\log 2^{i-1}) \\ &= \exp\left(k\left(1 + (i-1)\log 2 - 2^{i-1}\right)\right).\end{aligned}$$

Clearly, the exponent is nonpositive for all $i \ge 1$. Also,

$$\begin{aligned}\sum_{i=1}^{\infty} 2^i \exp\left(k\left(1 + (i-1)\log 2 - 2^{i-1}\right)\right) &\le \sum_{i=1}^{\infty} \exp(1 - \log 2 + 2i\log 2 - 2^{i-1}) \\ &\stackrel{\text{def}}{=} c < \infty.\end{aligned}$$

Hence,

$$\sum_{i\ge 0: a_i \in A, a_i \le \delta} h(B_i) \le \alpha\varepsilon \frac{n}{n-1}(1+c).$$

Consider $B_i = (a_{i-1}, a_i]$, $a_i \notin A$, $a_i \le \delta$, $i \ge 0$. Then the bounding is as above, i.e., the binomial argument remains unchanged, and only the passage via (11.4) requires a modification. Indeed,

$$\begin{aligned}\int_{B_i} |g(\mathbf{y}) - g(\mathbf{x})|^p \, \mu(\mathrm{d}\mathbf{y}) &\le \int_{B(\mathbf{0}, a_i)} |g(\mathbf{y}) - g(\mathbf{x})|^p \\ &\qquad \text{(where } B \text{ is the closed ball)} \\ &\le \varepsilon \mu\left(B(\mathbf{0}, a_i)\right) \\ &\le \varepsilon \frac{2^i k}{n-1} \times \frac{4}{3} \\ &\qquad \text{(as noted earlier).}\end{aligned}$$

The factor $4/3$ carries through. So,

$$\sum_{i\ge 0: a_i \notin A, a_i \le \delta} h(B_i) \le \alpha\varepsilon \frac{n}{n-1}(1+c) \times \frac{4}{3}.$$

Collecting bounds, uniformly over all $k \ge 1$,

$$h\left([0,\delta]\right) \le \alpha\varepsilon\Big(4 + \frac{7}{3}\frac{n}{n-1}(1+c)\Big),$$

which is as small as desired by choice of ε.

It remains to bound $h((\delta,\infty))$. Observe the following:

$$
\begin{aligned}
h\left((\delta,\infty)\right) &\le \frac{\alpha n}{k}\mathbb{E}\left[|g(\mathbf{X})-g(\mathbf{x})|^p\,\mathbb{1}_{[\|\mathbf{X}-\mathbf{x}\|>\delta]}\mathbb{1}_{[\mathrm{Bin}(n-1,F(\delta))<k]}\right] \\
&\qquad \text{(by conditioning on } \mathbf{X}_1\text{)} \\
&\le \frac{\alpha n}{k}\mathbb{E}\,|g(\mathbf{X})-g(\mathbf{x})|^p \times \mathbb{P}\left\{\mathrm{Bin}(n-1,F(\delta))<k\right\} \\
&\le \frac{\alpha n}{k}\times 2^{p-1}\left(\mathbb{E}\,|g(\mathbf{X})|^p+|g(\mathbf{x})|^p\right) \\
&\qquad \times \exp\left(k-(n-1)F(\delta)-k\log\left(\frac{k}{(n-1)F(\delta)}\right)\right) \\
&\qquad \text{(if } k<(n-1)F(\delta)\text{, which is valid for } n \text{ large enough since } k/n\to 0 \\
&\qquad \text{and } F(\delta)>0\text{)} \\
&= \mathrm{o}(1) \quad \text{as } n\to\infty,
\end{aligned}
$$

because $\mathbb{E}|g(\mathbf{X})|^p<\infty$, and for all n large enough, the exponent in $\exp(\cdot)$ is smaller than a negative constant times n in view of $k/n\to 0$ and $F(\delta)>0$.

Therefore

$$
h\left([0,\infty)\right) \le \alpha\varepsilon\Big(4+\frac{14}{3}(1+c)\Big)+\mathrm{o}(1),
$$

and the proof is complete. □

Proof (Theorem 11.1). Because $|a+b|^p\le 2^{p-1}(|a|^p+|b|^p)$ for $p\ge 1$, we see that

$$
\begin{aligned}
\mathbb{E}\left|\sum_{i=1}^n v_{ni}Y_{(i)}(\mathbf{x})-r(\mathbf{x})\right|^p &\le 2^{p-1}\mathbb{E}\left|\sum_{i=1}^n v_{ni}\left[Y_{(i)}(\mathbf{x})-r\left(\mathbf{X}_{(i)}(\mathbf{x})\right)\right]\right|^p \\
&\quad + 2^{p-1}\mathbb{E}\left|\sum_{i=1}^n v_{ni}r\left(\mathbf{X}_{(i)}(\mathbf{x})\right)-r(\mathbf{x})\right|^p. \qquad (11.5)
\end{aligned}
$$

We note that $\mathbb{E}|r(\mathbf{X})|^p<\infty$, by Jensen's inequality. Thus, the second term on the right-hand side of (11.5) tends to zero at μ-almost all $\mathbf{x}$ by Lemma 11.1. We show that the first term tends to zero at μ-almost all $\mathbf{x}$ when $p\ge 2$. The case $1\le p<2$ is then obtained through a truncation argument.

Conditional on $\mathbf{X}_1,\dots,\mathbf{X}_n$, the differences $Y_{(i)}(\mathbf{x})-r(\mathbf{X}_{(i)}(\mathbf{x}))$, $1\le i\le n$, are independent and centered (see Proposition 8.1). Theorem 20.13 in the Appendix thus implies that for some positive constant C_p depending only upon p,

$$
\mathbb{E}\left|\sum_{i=1}^n v_{ni}\left[Y_{(i)}(\mathbf{x})-r\left(\mathbf{X}_{(i)}(\mathbf{x})\right)\right]\right|^p \le C_p\,\mathbb{E}\left|\sum_{i=1}^n v_{ni}^2\left|Y_{(i)}(\mathbf{x})-r\left(\mathbf{X}_{(i)}(\mathbf{x})\right)\right|^2\right|^{p/2}.
$$

Let Σ_i be the rank of $\mathbf{X}_i$ in the reordering of the data according to increasing values of $\|\mathbf{X}_i - \mathbf{x}\|$. So, $\mathbf{X}_i$ receives weight $v_{n\Sigma_i}$. Since

$$\mathbb{E}\left| \sum_{i=1}^{n} v_{ni}^2 \left| Y_{(i)}(\mathbf{x}) - r\left(\mathbf{X}_{(i)}(\mathbf{x})\right)\right|^2 \right|^{p/2} = \mathbb{E}\left| \sum_{i=1}^{n} v_{n\Sigma_i}^2 \left| Y_i - r(\mathbf{X}_i)\right|^2 \right|^{p/2},$$

we have

$$\begin{aligned}\mathbb{E}\left| \sum_{i=1}^{n} v_{ni} \left[Y_{(i)}(\mathbf{x}) - r\left(\mathbf{X}_{(i)}(\mathbf{x})\right)\right] \right|^p &\le C_p (\max_i v_{ni})^{p/2} \mathbb{E}\left| \sum_{i=1}^{n} v_{n\Sigma_i} \left| Y_i - r(\mathbf{X}_i)\right|^2 \right|^{p/2} \\ &\le C_p (\max_i v_{ni})^{p/2} \mathbb{E}\left[\sum_{i=1}^{n} v_{n\Sigma_i} \left| Y_i - r(\mathbf{X}_i)\right|^p \right], \end{aligned} \tag{11.6}$$

where we used Jensen's inequality (which is valid here since $p \ge 2$). Next, letting $g(\mathbf{x}) = \mathbb{E}[|Y - r(\mathbf{X})|^p \,|\, \mathbf{X} = \mathbf{x}]$, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{n} v_{n\Sigma_i} \left| Y_i - r(\mathbf{X}_i)\right|^p \right] &= \mathbb{E}\left[\sum_{i=1}^{n} v_{n\Sigma_i} \mathbb{E}\left[\left| Y_i - r(\mathbf{X}_i)\right|^p \,\middle|\, \mathbf{X}_1, \ldots, \mathbf{X}_n \right] \right] \\ &= \mathbb{E}\left[\sum_{i=1}^{n} v_{n\Sigma_i} g(\mathbf{X}_i) \right]. \end{aligned}$$

It follows that

$$\mathbb{E}\left| \sum_{i=1}^{n} v_{ni} \left[Y_{(i)}(\mathbf{x}) - r\left(\mathbf{X}_{(i)}(\mathbf{x})\right)\right] \right|^p \le C_p (\max_i v_{ni})^{p/2} \mathbb{E}\left[\sum_{i=1}^{n} v_{ni} g\left(\mathbf{X}_{(i)}(\mathbf{x})\right) \right].$$

Since $\mathbb{E}|g(\mathbf{X})| < \infty$, the quantity $\mathbb{E}[\sum_{i=1}^{n} v_{ni} g(\mathbf{X}_{(i)}(\mathbf{x}))]$ remains bounded for μ-almost all $\mathbf{x}$ by Lemma 11.1. Hence, using $\max_i v_{ni} \to 0$ as $n \to \infty$ (condition (*iii*)), we conclude that the first term in (11.5) tends to zero at μ-almost all $\mathbf{x}$. Thus, Theorem 11.1 is proved for $p \ge 2$.

Consider now the case $1 \le p < 2$. Define for integer $M > 0$, $Y^{(M)} = Y \mathbb{1}_{[|Y| \le M]}$, $Z^{(M)} = Y - Y^{(M)}$, $r^{(M)}(\mathbf{x}) = \mathbb{E}[Y^{(M)} | \mathbf{X} = \mathbf{x}]$ and $s^{(M)}(\mathbf{x}) = \mathbb{E}[Z^{(M)} | \mathbf{X} = \mathbf{x}]$. Then, with obvious notation,

$$\begin{aligned}\mathbb{E}\left| \sum_{i=1}^{n} v_{ni} \left[Y_{(i)}(\mathbf{x}) - r\left(\mathbf{X}_{(i)}(\mathbf{x})\right)\right] \right|^p &\le 2^{p-1} \mathbb{E}\left| \sum_{i=1}^{n} v_{ni} \left[Y_{(i)}^{(M)}(\mathbf{x}) - r^{(M)}\left(\mathbf{X}_{(i)}(\mathbf{x})\right)\right] \right|^p \\ &\quad + 2^{p-1} \mathbb{E}\left[\sum_{i=1}^{n} v_{ni} \left| Z_{(i)}^{(M)}(\mathbf{x}) - s^{(M)}\left(\mathbf{X}_{(i)}(\mathbf{x})\right)\right|^p \right], \end{aligned} \tag{11.7}$$

where we used Jensen's inequality again.

Since $Y^{(M)}$ is bounded, the first term of (11.7) tends to zero at μ-almost all $\mathbf{x}$. To see this, it suffices to note that if Y is bounded, then to prove the result for all $p \geq 1$, it is enough to show that it holds for $p = 2$. This has already been done in the first part of the proof.

With respect to the last term of (11.7), we may write

$$\mathbb{E}\left[\sum_{i=1}^{n} v_{ni}\left|Z_{(i)}^{(M)}(\mathbf{x}) - s^{(M)}\left(\mathbf{X}_{(i)}(\mathbf{x})\right)\right|^{p}\right]$$
$$\leq 2^{p-1}\mathbb{E}\left[\sum_{i=1}^{n} v_{ni}\left|Z_{(i)}^{(M)}(\mathbf{x})\right|^{p}\right] + 2^{p-1}\mathbb{E}\left[\sum_{i=1}^{n} v_{ni}\left|s^{(M)}\left(\mathbf{X}_{(i)}(\mathbf{x})\right)\right|^{p}\right].$$

Thus, letting $g^{(M)}(\mathbf{x}) = \mathbb{E}[|Z^{(M)}|^{p} \mid \mathbf{X} = \mathbf{x}]$,

$$\mathbb{E}\left[\sum_{i=1}^{n} v_{ni}\left|Z_{(i)}^{(M)}(\mathbf{x}) - s^{(M)}\left(\mathbf{X}_{(i)}(\mathbf{x})\right)\right|^{p}\right]$$
$$\leq 2^{p-1}\mathbb{E}\left[\sum_{i=1}^{n} v_{ni}\left|Z_{(i)}^{(M)}(\mathbf{x})\right|^{p}\right] + 2^{p-1}\mathbb{E}\left[\sum_{i=1}^{n} v_{ni}\, g^{(M)}\left(\mathbf{X}_{(i)}(\mathbf{x})\right)\right].$$

Next, as in the first part of the proof, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n} v_{ni}\left|Z_{(i)}^{(M)}(\mathbf{x})\right|^{p}\right] &= \mathbb{E}\left[\sum_{i=1}^{n} v_{n\Sigma_i}\left|Z_{i}^{(M)}\right|^{p}\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} v_{n\Sigma_i}\mathbb{E}\left[\left|Z_{i}^{(M)}\right|^{p}\,\middle|\,\mathbf{X}_1,\ldots,\mathbf{X}_n\right]\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} v_{n\Sigma_i}\, g^{(M)}(\mathbf{X}_i)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} v_{ni} g^{(M)}\left(\mathbf{X}_{(i)}(\mathbf{x})\right)\right].
\end{aligned}$$

Therefore, we conclude that the last term of (11.7) is not greater than

$$2^{2p-1}\mathbb{E}\left[\sum_{i=1}^{n} v_{ni}\, g^{(M)}\left(\mathbf{X}_{(i)}(\mathbf{x})\right)\right].$$

Let A_M be the set of all $\mathbf{x}$ for which the first term of (11.7) tends to zero and the quantity $\mathbb{E}[\sum_{i=1}^{n} v_{ni} g^{(M)}(\mathbf{X}_{(i)}(\mathbf{x}))]$ tends to $g^{(M)}(\mathbf{x})$ as $n \to \infty$. We have already shown (see Lemma 11.1) that, for each fixed M, $\mu(A_M) = 1$. Let B be the set of all

$\mathbf{x}$ with $g^{(M)}(\mathbf{x}) \to 0$ as $M \to \infty$. Clearly, $\mu(B) = 1$ because $\mathbb{E}[g^{(M)}(\mathbf{X})] \to 0$ as $M \to \infty$ and $g^{(M)}$ is monotone in M. For all $\mathbf{x}$ in $B \cap (\cap_M A_M)$, we claim that (11.7) tends to zero: first pick M large enough so that $g^{(M)}(\mathbf{x})$ is small, and then let n grow large. Since this set has μ-measure 1, the theorem is proved. □

We conclude this section with the following theorem, which states that the nearest neighbor estimate is weakly pointwise consistent at μ-almost all $\mathbf{x}$ for a broader family of weights, with no condition on $(\mathbf{X}, Y)$ other than the boundedness of Y. Its proof is an easy adaptation of the proof of Theorem 11.1.

Theorem 11.2 (Weak pointwise consistency). *Let $p \geq 1$. Assume that $\|Y\|_\infty < \infty$ and that there exists a sequence of integers $\{k\} = \{k_n\}$ such that*

$$\begin{array}{ll} (i) & k \to \infty \quad \text{and} \quad k/n \to 0; \\ (ii) & \sum_{i>k} v_{ni} \to 0; \\ (iii) & \sup_n (k \max_i v_{ni}) < \infty. \end{array} \tag{11.8}$$

Then the corresponding nearest neighbor regression function estimate r_n satisfies

$$\mathbb{E}\,|r_n(\mathbf{x}) - r(\mathbf{x})|^p \to 0 \quad \text{at } \mu\text{-almost all } \mathbf{x} \in \mathbb{R}^d.$$

In particular, if $\|Y\|_\infty < \infty$, then

$$r_n(\mathbf{x}) \to r(\mathbf{x}) \quad \text{in probability at } \mu\text{-almost all } \mathbf{x} \in \mathbb{R}^d.$$

Remark 11.1. Conditions (11.8) in Theorem 11.2 may be replaced by the following equivalent ones: there exists a positive constant α such that

$$\begin{array}{ll} (i) & \sum_{i>\alpha/\max_i v_{ni}} v_{ni} \to 0; \\ (ii) & \sum_{i>\varepsilon n} v_{ni} \to 0, \quad \text{all } \varepsilon > 0; \\ (iii) & \max_i v_{ni} \to 0. \end{array}$$

(See Lemma 20.4 in the Appendix for a proof of this equivalency.) An example of weights that satisfy (11.8) includes the geometric choice

$$v_{ni} = \frac{p_n(1-p_n)^{i-1}}{1-(1-p_n)^n}, \quad 1 \leq i \leq n,$$

with $p_n \in (0,1)$, $p_n \to 0$ and $np_n \to \infty$. □

Proof (Theorem 11.2). When g is μ-almost surely bounded, the replacement of (11.1)(*ii*) by (11.8)(*ii*) does not upset the conclusion of Lemma 11.1. In the proof of Theorem 11.1, take $p = 2$, and estimate (11.6) from above by $c \max_i v_{ni}$ for some constant c. □

11.2 Concentration of measure and its consequences

Today, there are powerful tools for controlling the variation and stability of random variables, and especially of functions of n independent random variables. The book by Boucheron et al. (2013) summarizes this subject. The objective of this section, by way of introduction, is to highlight the use of these tools for the nearest neighbor regression function estimate

$$r_n(\mathbf{x}) = \sum_{i=1}^{n} v_{ni} Y_{(i)}(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d,$$

with $\min_i v_{ni} \geq 0$ and $\sum_{i=1}^{n} v_{ni} = 1$.

The first notion is that of a self-bounding function. One says that a function $g : \mathbb{R}^n \to [0, \infty)$ has the self-bounding property if there exist functions $g_i : \mathbb{R}^{n-1} \to [0, \infty)$ such that, for all $x_1, \ldots, x_n$ and all $1 \leq i \leq n$,

$$0 \leq g(x_1, \ldots, x_n) - g_i(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \leq 1$$

and

$$\sum_{i=1}^{n} (g(x_1, \ldots, x_n) - g_i(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)) \leq g(x_1, \ldots, x_n).$$

For a self-bounding function g,

$$\mathbb{V}g(X_1, \ldots, X_n) \leq \mathbb{E}g(X_1, \ldots, X_n),$$

where $X_1, \ldots, X_n$ are independent real-valued random variables. This is an immediate corollary of the Efron-Stein inequality (see Theorem 20.10 in the Appendix).

Assume that $\|Y\|_\infty \leq 1$. Write $Y = Y^+ - Y^-$, where $Y^+ = \max(Y, 0)$ and $Y^- = -\min(Y, 0)$. Then

$$r_n(\mathbf{x}) = r_n^+(\mathbf{x}) - r_n^-(\mathbf{x}),$$

with

$$r_n^+(\mathbf{x}) = \sum_{i=1}^{n} v_{ni} Y_{(i)}^+(\mathbf{x}) \quad \text{and} \quad r_n^-(\mathbf{x}) = \sum_{i=1}^{n} v_{ni} Y_{(i)}^-(\mathbf{x}).$$

Take

$$g = \frac{r_n^+(\mathbf{x})}{\max_i v_{ni}} \quad \text{and} \quad g_i = \frac{1}{\max_i v_{ni}} \sum_{j \neq i} v_{nj} Y_{(j)}^+(\mathbf{x}).$$

Clearly,

$$g - g_i = \frac{v_{ni}}{\max_i v_{ni}} Y^+_{(i)}(\mathbf{x}),$$

so that we have a self-bounding function. Therefore,

$$\mathbb{V}\left[\frac{r_n^+(\mathbf{x})}{\max_i v_{ni}}\right] \le \mathbb{E}\left[\frac{r_n^+(\mathbf{x})}{\max_i v_{ni}}\right],$$

and so

$$\mathbb{V}r_n^+(\mathbf{x}) \le (\max_i v_{ni})\mathbb{E}r_n^+(\mathbf{x}) \le \max_i v_{ni}.$$

Similarly,

$$\mathbb{V}r_n^-(\mathbf{x}) \le (\max_i v_{ni})\mathbb{E}r_n^-(\mathbf{x}) \le \max_i v_{ni}.$$

We have, without further work, using $(a+b)^2 \le 2(a^2+b^2)$,

$$\mathbb{V}r_n(\mathbf{x}) \le 2\left(\mathbb{V}r_n^+(\mathbf{x}) + \mathbb{V}r_n^-(\mathbf{x})\right) \le 4\max_i v_{ni}.$$

Thus, if $\|Y\|_\infty \le 1$,

$$\mathbb{E}\left|r_n(\mathbf{x}) - r(\mathbf{x})\right|^2 = \mathbb{V}r_n(\mathbf{x}) + (\mathbb{E}r_n(\mathbf{x}) - r(\mathbf{x}))^2 \le 4\max_i v_{ni} + (\mathbb{E}r_n(\mathbf{x}) - r(\mathbf{x}))^2.$$

The second term is the bias term, where we note that $\mathbb{E}r_n(\mathbf{x}) = \sum_{i=1}^n v_{ni} r(\mathbf{X}_{(i)}(\mathbf{x}))$. By Lemma 11.1 and Lemma 20.4, the mere conditions $\sum_{i>\alpha/\max_i v_{ni}}$ for some $\alpha > 0$ and $\sum_{i>\varepsilon n} v_{ni} \to 0$ for all $\varepsilon > 0$ (besides $v_{ni} \ge 0$, $\sum_{i=1}^n v_{ni} = 1$) imply that the bias term tends to zero μ-almost surely. The variance term tends to zero when $\max_i v_{ni} \to 0$. For the standard k-nearest neighbor estimate, $k/n \to 0$ is needed for the bias, and $k \to \infty$ is needed for the variance.

Of course, we already know from Theorem 11.2 that $\mathbb{E}|r_n(\mathbf{x}) - r(\mathbf{x})|^2 \to 0$ at μ-almost all $\mathbf{x}$. The power of the concentration inequalities will be clear when we can say something about

$$\mathbb{P}\left\{|r_n(\mathbf{x}) - \mathbb{E}r_n(\mathbf{x})| \ge \varepsilon\right\}.$$

Indeed, we can do better than Chebyshev's inequality,

$$\mathbb{P}\left\{|r_n(\mathbf{x}) - \mathbb{E}r_n(\mathbf{x})| \ge \varepsilon\right\} \le \frac{\mathbb{V}r_n(\mathbf{x})}{\varepsilon^2} \le \frac{4\max_i v_{ni}}{\varepsilon^2}.$$

The following exponential inequality was proved by Boucheron et al. (2013, page 182) by the entropy method: for a self-bounding function g,

$$\begin{aligned}&\mathbb{P}\left\{g(X_1,\ldots,X_n)-\mathbb{E}g(X_1,\ldots,X_n)\geq t\right\}\\&\quad\leq\exp\left(-h\left(\frac{t}{\mathbb{E}g(X_1,\ldots,X_n)}\right)\mathbb{E}g(X_1,\ldots,X_n)\right),\quad t>0,\end{aligned}$$

and

$$\begin{aligned}&\mathbb{P}\left\{g(X_1,\ldots,X_n)-\mathbb{E}g(X_1,\ldots,X_n)\leq -t\right\}\\&\quad\leq\exp\left(-h\left(-\frac{t}{\mathbb{E}g(X_1,\ldots,X_n)}\right)\mathbb{E}g(X_1,\ldots,X_n)\right),\quad 0<t\leq\mathbb{E}g(X_1,\ldots,X_n),\end{aligned}$$

where $h(u)=(1+u)\log(1+u)-u$, $u\geq -1$. Recalling $h(u)\geq u^2/(2+2u/3)$ for $u\geq 0$ and $h(u)\geq u^2/2$ for $u\leq 0$, we have

$$\begin{aligned}&\mathbb{P}\left\{g(X_1,\ldots,X_n)-\mathbb{E}g(X_1,\ldots,X_n)\geq t\right\}\\&\quad\leq\exp\left(-\frac{t^2}{2\mathbb{E}g(X_1,\ldots,X_n)+2t/3}\right),\quad t>0,\end{aligned}$$

and

$$\begin{aligned}&\mathbb{P}\left\{g(X_1,\ldots,X_n)-\mathbb{E}g(X_1,\ldots,X_n)\leq -t\right\}\\&\quad\leq\exp\left(-\frac{t^2}{2\mathbb{E}g(X_1,\ldots,X_n)}\right),\quad 0<t\leq\mathbb{E}g(X_1,\ldots,X_n).\end{aligned}$$

Apply this to $r_n^+(\mathbf{x})/\max_i v_{ni}$, which is a self-bounding function when $\|Y\|_\infty\leq 1$. Then we obtain

$$\mathbb{P}\left\{r_n^+(\mathbf{x})-\mathbb{E}r_n^+(\mathbf{x})\geq\varepsilon\right\}\leq\exp\left(-\frac{\varepsilon^2}{2+2\varepsilon/3}\times\frac{1}{\max_i v_{ni}}\right),\quad\varepsilon>0,$$

and

$$\mathbb{P}\left\{r_n^+(\mathbf{x})-\mathbb{E}r_n^+(\mathbf{x})\leq -\varepsilon\right\}\leq\begin{cases}\exp\left(-\frac{\varepsilon^2}{2}\times\frac{1}{\max_i v_{ni}}\right) & \text{for } 0<\varepsilon\leq\mathbb{E}r_n^+(\mathbf{x})\\ 0 & \text{for } \varepsilon>\mathbb{E}r_n^+(\mathbf{x}).\end{cases}$$

Therefore,

$$\mathbb{P}\left\{\left|r_n^+(\mathbf{x})-\mathbb{E}r_n^+(\mathbf{x})\right|\geq\varepsilon\right\}\leq 2\exp\left(-\frac{\varepsilon^2}{2+2\varepsilon/3}\times\frac{1}{\max_i v_{ni}}\right),\quad\varepsilon>0.$$

Obviously, the same inequality is valid for $r_n^-(\mathbf{x})$. Thus, summarizing, we have proved the following interesting proposition:

Proposition 11.1. *Assume that* $\|Y\|_\infty \le 1$. *Then, at all* $\mathbf{x} \in \mathbb{R}^d$, *for all* $\varepsilon > 0$,

$$\mathbb{P}\{|r_n(\mathbf{x}) - \mathbb{E}r_n(\mathbf{x})| \ge \varepsilon\} \le 4\exp\left(-\frac{\varepsilon^2}{8+4\varepsilon/3} \times \frac{1}{\max_i v_{ni}}\right).$$

Note that the right-hand side is summable in n for all $\varepsilon > 0$ if

$$(\log n) \max_i v_{ni} \to 0. \tag{11.9}$$

Hence, by the Borel-Cantelli lemma,

$$r_n(\mathbf{x}) - \mathbb{E}r_n(\mathbf{x}) \to 0 \quad \text{almost surely at all } \mathbf{x},$$

if (11.9) holds. Observing that

$$|r_n(\mathbf{x}) - r(\mathbf{x})| \le |r_n(\mathbf{x}) - \mathbb{E}r_n(\mathbf{x})| + |\mathbb{E}r_n(\mathbf{x}) - r(\mathbf{x})|$$

and that the second term tends to zero at μ-almost all $\mathbf{x}$ whenever $\sum_{i>\alpha/\max_i v_{ni}}$ for some $\alpha > 0$ and $\sum_{i>\varepsilon n} v_{ni} \to 0$ for all $\varepsilon > 0$, we conclude that

$$r_n(\mathbf{x}) \to r(\mathbf{x}) \quad \text{almost surely at } \mu\text{-almost all } \mathbf{x},$$

if, additionally, $\|Y\|_\infty \le 1$ and (11.9) holds. Almost sure (or strong) consistency of $r_n(\mathbf{x})$ towards $r(\mathbf{x})$ is the topic of the next section.

11.3 Strong pointwise consistency

In some applications, data arrive sequentially—these include internet data, measurements from monitoring stations, and stock ticker data, for example. One can construct a regression function estimate r_n of r, and ask what happens to r_n when more data come in, as a function of n. In other words, the sequence $\{r_1, r_2, r_3, \ldots\}$ itself is of interest. If we know, for example, that $r_n(\mathbf{x}) \to r(\mathbf{x})$ almost surely, then $|r_n(\mathbf{x}) - r(\mathbf{x})| \le \varepsilon$ for all n greater than some n_0, with probability one. The weak consistency of $r_n(\mathbf{x})$ to $r(\mathbf{x})$ does not offer such guarantees. One could in fact have $\limsup_{n\to\infty} |r_n(\mathbf{x}) - r(\mathbf{x})| = \infty$ almost surely—a disaster in terms of sequential applications—while $|r_n(\mathbf{x}) - r(\mathbf{x})| \to 0$ in probability.

The next theorem (Devroye, 1982) establishes the strong pointwise consistency of the nearest neighbor estimate. The requirements that are imposed on the sequence of weights are similar to the ones of Theorem 11.2, except that the condition $k \to \infty$ is now replaced by the slightly stronger one $k/\log n \to \infty$. For the proof, we refer to the previous section.

Theorem 11.3 (Strong pointwise consistency). *Assume that* $\|Y\|_\infty < \infty$ *and that there exists a sequence of integers* $\{k\} = \{k_n\}$ *such that*

$$\begin{array}{ll} (i) & k/\log n \to \infty \quad \textit{and} \quad k/n \to 0; \\ (ii) & \sum_{i>k} v_{ni} \to 0; \\ (iii) & \sup_n (k \max_i v_{ni}) < \infty. \end{array} \tag{11.10}$$

Then the corresponding nearest neighbor regression function estimate r_n *satisfies*

$$r_n(\mathbf{x}) \to r(\mathbf{x}) \quad \textit{almost surely at } \mu\textit{-almost all } \mathbf{x} \in \mathbb{R}^d.$$

Remark 11.2. Conditions (11.10) in Theorem 11.3 may be replaced by the following equivalent ones (see Lemma 20.4 in the Appendix): there exists a positive constant α such that

$$\begin{array}{ll} (i) & \sum_{i>\alpha/\max_i v_{ni}} \to 0; \\ (ii) & \sum_{i>\varepsilon n} v_{ni} \to 0, \quad \text{all } \varepsilon > 0; \\ (iii) & \sup_n ((\log n) \max_i v_{ni}) \to 0. \end{array}$$

□

Theorem 11.3 is rephrased for the standard k-nearest neighbor estimate as follows:

Corollary 11.2. *Assume that* $\|Y\|_\infty < \infty$. *If* $k/\log n \to \infty$ *and* $k/n \to 0$, *then the k-nearest neighbor regression function estimate is strongly consistent at* μ*-almost all* $\mathbf{x} \in \mathbb{R}^d$.

The condition $k/\log n \to \infty$ in the previous corollary is suboptimal. One can show that under regularity assumptions on k, the conditions $k/\log\log n \to \infty$ and $k/n \to 0$ are necessary and sufficient for $r_n(\mathbf{x}) \to r(\mathbf{x})$ almost surely at μ-almost all $\mathbf{x}$. A similar result with $(\log\log n) \max_i v_{ni}$ replacing $k/\log\log n \to \infty$ exists for the general nearest neighbor estimate—see Devroye (1982).

Theorem 11.4 (Strong pointwise consistency of the k-nearest neighbor estimate). *Assume that* $\|Y\|_\infty < \infty$. *Assume, in addition, that the sequence* $\{k\} = \{k_n\}$ *is increasing and regularly varying, i.e., for all* $\theta \in (0, 1]$, $k_{\lceil \theta n \rceil}/k_n \to c > 0$ *for some finite c. Then, if* $k/\log\log n \to \infty$ *and* $k/n \to 0$, *the k-nearest neighbor regression function estimate is strongly consistent at* μ*-almost all* $\mathbf{x} \in \mathbb{R}^d$.

Remark 11.3. The regular variation condition implies that $c = \theta^\rho$ for some $\rho \geq 0$ (see Bingham et al., 1987). When $\rho = 0$, and thus $c = 1$, the sequence is called slowly varying. □

Before embarking on the proof of Theorem 11.3, we make a short detour through the theory of records. Consider i.i.d. uniform $[0,1]$ random variables $U_1, \ldots, U_n$ and define $Z_i = F^{-1}(U_i)$, where F is a distribution function on $[0,\infty)$ and

$$F^{-1}(u) = \inf\{t \geq 0 : F(t) \geq u\}, \quad u \in [0,1].$$

We define the rank R_i of U_i at the moment of its birth by

$$R_i = \sum_{j=1}^{i} \mathbb{1}_{[U_j \leq U_i]}.$$

Since ties happen with zero probability, the R_i's are well defined. We cannot do the same with the Z_i's because F may be atomic. Let us break ties by indices. So, Z_i is placed before Z_j if $Z_i < Z_j$, or if $Z_i = Z_j$ and $i < j$. The rank of Z_i at the moment of its birth is

$$R_i' = \sum_{j=1}^{i} \mathbb{1}_{[Z_j \leq Z_i]}.$$

Since $U_j \leq U_i$ implies $Z_j \leq Z_i$ (but not vice versa), we have, for the coupled sequences, $R_i \leq R_i'$ for all i.

It is well known that $R_1, \ldots, R_n$ are independent and R_i is uniformly distributed on $\{1, \ldots, i\}$. In what follows, we need a tail bound for the quantity

$$\sum_{i=m+1}^{n} \mathbb{1}_{[R_i' \leq k]},$$

where $k \leq m < n$ are given parameters. We have, for $u > 0$,

$$\begin{aligned}
\mathbb{P}\left\{ \sum_{i=m+1}^{n} \mathbb{1}_{[R_i' \leq k]} \geq u \right\} &\leq \mathbb{P}\left\{ \sum_{i=m+1}^{n} \mathbb{1}_{[R_i \leq k]} \geq u \right\} \\
&\leq \mathbb{P}\left\{ \sum_{i=m+1}^{n} \mathrm{Ber}\left(\tfrac{k}{i}\right) \geq u \right\} \\
&\leq \mathbb{P}\left\{ \mathrm{Bin}\left(n-m, \tfrac{k}{m}\right) \geq u \right\} \\
&\leq \exp\left(u - \frac{n-m}{m} k - u \log\left(\frac{u}{\frac{n-m}{m} k} \right) \right)
\end{aligned}$$

for $u \geq \frac{n-m}{m} k$, by Chernoff's bound on binomials—see Theorem 20.5 in the Appendix. Setting $u = e \times \frac{n-m}{m} k$, we have

$$\mathbb{P}\left\{ \sum_{i=m+1}^{n} \mathbb{1}_{[R_i' \leq k]} \geq e \times \frac{n-m}{m} k \right\} \leq e^{-\frac{n-m}{m} k}.$$

This inequality is the only probabilistic fact needed to prove Theorem 11.4.

Proof (Theorem 11.4). Without loss of generality, it is assumed that $\|Y\|_\infty \leq 1$. The groundwork was laid in the previous theorem. We recall (see Lemma 11.1) that $\mathbb{E}r_n(\mathbf{x}) \to r(\mathbf{x})$ at μ-almost all $\mathbf{x}$, so we take such $\mathbf{x}$. We recall from Proposition 11.1 that

$$\mathbb{P}\left\{|r_n(\mathbf{x}) - \mathbb{E}r_n(\mathbf{x})| \geq \varepsilon\right\} \leq 4e^{-\alpha k_n}$$

for some positive constant α depending upon ε only. Define

$$n_\ell = \lfloor N(1+\delta)^\ell \rfloor, \quad \ell \geq 0,$$

where $\delta > 0$ and $N \in \mathbb{N}^\star$ are fixed and to be determined. Let $\varepsilon > 0$ be fixed and arbitrary. Note that we can find N large enough (depending upon ε) such that

$$\begin{aligned}\sum_{\ell=0}^{\infty} \mathbb{P}\left\{|r_{n_\ell}(\mathbf{x}) - \mathbb{E}r_{n_\ell}(\mathbf{x})| \geq \varepsilon\right\} &\leq \sum_{\ell=0}^{\infty} \frac{1}{\log^2 n_\ell} \\ &\leq \sum_{\ell=0}^{\infty} \frac{1}{(-1 + \log N + \ell \log(1+\delta))^2} \\ &< \infty.\end{aligned}$$

Here we used the condition $k/\log\log n \to \infty$. By the Borel-Cantelli lemma, we conclude that

$$r_{n_\ell}(\mathbf{x}) \to r(\mathbf{x}) \quad \text{almost surely at } \mu\text{-almost all } \mathbf{x},$$

when $\ell \to \infty$. The proof is complete if we can show that, as $\ell \to \infty$,

$$\max_{n_{\ell-1} < j \leq n_\ell} \left|r_j(\mathbf{x}) - r_{n_\ell}(\mathbf{x})\right| \to 0 \quad \text{almost surely at } \mu\text{-almost all } \mathbf{x}.$$

This "trick" of partitioning the integers into suitable intervals—in this case of exponentially growing size—is standard in the literature on strong convergence.

Observe the following, for $\ell \geq 1$:

$$\begin{aligned}\frac{n_{\ell-1}}{n_\ell} \geq \frac{N(1+\delta)^{\ell-1} - 1}{N(1+\delta)^\ell} &\geq \frac{1}{1+\delta} - \frac{1}{N} \\ &\geq \frac{1}{1+2\delta},\end{aligned}$$

where the last inequality is valid for all $N \geq 3 + 2\delta + \frac{1}{\delta}$. Similarly, for all $N \geq 2/\delta$,

$$\frac{n_\ell}{n_{\ell-1}} \geq 1 + \frac{\delta}{2}.$$

Hence, for all $N \geq 3 + 2\delta + \frac{2}{\delta}$,

$$\frac{\delta}{2} \leq \frac{n_\ell - n_{\ell-1}}{n_{\ell-1}} \leq 2\delta. \tag{11.11}$$

Moreover, by regular variation of k_n, we have

$$\frac{k_{n_{\ell-1}}}{k_{n_\ell}} \sim \left(\frac{n_{\ell-1}}{n_\ell}\right)^c \sim \frac{1}{(1+\delta)^c}$$

for some $c > 0$ as $\ell \to \infty$. Thus,

$$\frac{k_{n_\ell} - k_{n_{\ell-1}}}{k_{n_\ell}} \to 1 - \frac{1}{(1+\delta)^c} \quad \text{as } \ell \to \infty.$$

Because we are observing the regression function estimate over time, it is necessary to introduce time as an index. Thus, the reordered data at time n are

$$\left(\mathbf{X}_{(1,n)}(\mathbf{x}), Y_{(1,n)}(\mathbf{x})\right), \ldots, \left(\mathbf{X}_{(n,n)}(\mathbf{x}), Y_{(n,n)}(\mathbf{x})\right).$$

The rank of $\mathbf{X}_i$ among $\mathbf{X}_1, \ldots, \mathbf{X}_n$ after reordering is $\Sigma_{i,n}$, the last index always referring to time. Note that $\Sigma_{i,n}$ is increasing in n. We have

$$r_n(\mathbf{x}) = \frac{1}{k_n} \sum_{i=1}^{n} Y_i \mathbb{1}_{[\Sigma_{i,n} \leq k_n]} \quad \text{and} \quad r_j(\mathbf{x}) = \frac{1}{k_j} \sum_{i=1}^{j} Y_i \mathbb{1}_{[\Sigma_{i,j} \leq k_j]}.$$

Let $n_{\ell-1} < j \leq n_\ell$. Define

$$r'_j(\mathbf{x}) = \frac{1}{k_{n_\ell}} \sum_{i=1}^{j} Y_i \mathbb{1}_{[\Sigma_{i,j} \leq k_j]}.$$

For such j, we have, with probability one,

$$\begin{aligned}
\left|r_j(\mathbf{x}) - r'_j(\mathbf{x})\right| &\leq \left(\frac{1}{k_j} - \frac{1}{k_{n_\ell}}\right) \sum_{i=1}^{j} |Y_i| \mathbb{1}_{[\Sigma_{i,j} \leq k_j]} \\
&\leq 1 - \frac{k_j}{k_{n_\ell}} \\
&\quad \text{(since } \|Y\|_\infty \leq 1\text{)} \\
&\leq \frac{k_{n_\ell} - k_{n_{\ell-1}}}{k_{n_\ell}} \\
&\to 1 - \frac{1}{(1+\delta)^c} \quad \text{as } \ell \to \infty.
\end{aligned}$$

We conclude that, with probability one,

$$\left|r_j(\mathbf{x}) - r_{n_\ell}(\mathbf{x})\right| \le \mathrm{o}(1) + \left(1 - \frac{1}{(1+\delta)^c}\right) + \left|r'_j(\mathbf{x}) - r_{n_\ell}(\mathbf{x})\right|.$$

Note that for $n_{\ell-1} < i \le j \le n_\ell$, $\mathbb{1}_{[\Sigma_{i,j} \le k_j]} \le \mathbb{1}_{[\Sigma_{i,i} \le k_{n_\ell}]}$. Thus, with probability one,

$$\begin{aligned} k_{n_\ell}\left|r'_j(\mathbf{x}) - r_{n_\ell}(\mathbf{x})\right| &\le \sum_{i=1}^{n_\ell} \left|\mathbb{1}_{[\Sigma_{i,n_\ell} \le k_{n_\ell}]} - \mathbb{1}_{[\Sigma_{i,j} \le k_j]}\right| \\ &\le \sum_{i=1}^{n_{\ell-1}} \left|\mathbb{1}_{[\Sigma_{i,n_\ell} \le k_{n_\ell}]} - \mathbb{1}_{[\Sigma_{i,j} \le k_j]}\right| + \sum_{i=n_{\ell-1}+1}^{n_\ell} \mathbb{1}_{[\Sigma_{i,i} \le k_{n_\ell}]}. \end{aligned}$$

Now, $\Sigma_{1,1}, \ldots, \Sigma_{n_\ell,n_\ell}$ are distributed like the local ranks $R'_1, \ldots, R'_{n_\ell}$. We recall

$$\mathbb{P}\left\{\frac{1}{k_{n_\ell}} \sum_{i=n_{\ell-1}+1}^{n_\ell} \mathbb{1}_{[\Sigma_{i,i} \le k_{n_\ell}]} \ge e \times \frac{n_\ell - n_{\ell-1}}{n_{\ell-1}}\right\} \le e^{-\frac{n_\ell - n_{\ell-1}}{n_{\ell-1}} k_{n_\ell}},$$

and thus, according to inequalities (11.11), for all N large enough,

$$\mathbb{P}\left\{\frac{1}{k_{n_\ell}} \sum_{i=n_{\ell-1}+1}^{n_\ell} \mathbb{1}_{[\Sigma_{i,i} \le k_{n_\ell}]} \ge e(2\delta)\right\} \le e^{-k_{n_\ell}\delta/2}. \tag{11.12}$$

Using $k/\log\log n \to \infty$, and our definition of n_ℓ, this is summable in ℓ for all $\delta > 0$.

The proof is finished if we can show that

$$\sup_{n_{\ell-1} < j \le n_\ell} \frac{1}{k_{n_\ell}} \sum_{i=1}^{n_{\ell-1}} \left|\mathbb{1}_{[\Sigma_{i,n_\ell} \le k_{n_\ell}]} - \mathbb{1}_{[\Sigma_{i,j} \le k_j]}\right| \to 0 \quad \text{almost surely in } \ell. \tag{11.13}$$

To get rid of the dependence upon j, observe that

$$\mathbb{1}_{[\Sigma_{i,n_\ell} \le k_{n_{\ell-1}}]} \le \mathbb{1}_{[\Sigma_{i,j} \le k_j]} \le \mathbb{1}_{[\Sigma_{i,n_{\ell-1}} \le k_{n_\ell}]}.$$

Thus, the supremum in (11.13) is not larger than

$$\begin{aligned} &\frac{1}{k_{n_\ell}} \sum_{i=1}^{n_{\ell-1}} \left(\mathbb{1}_{[\Sigma_{i,n_\ell} \le k_{n_\ell}]} - \mathbb{1}_{[\Sigma_{i,n_\ell} \le k_{n_{\ell-1}}]}\right) + \frac{1}{k_{n_\ell}} \sum_{i=1}^{n_{\ell-1}} \left(\mathbb{1}_{[\Sigma_{i,n_{\ell-1}} \le k_{n_\ell}]} - \mathbb{1}_{[\Sigma_{i,n_\ell} \le k_{n_\ell}]}\right) \\ &\le \frac{1}{k_{n_\ell}} \sum_{i=1}^{n_{\ell-1}} \mathbb{1}_{[k_{n_{\ell-1}} < \Sigma_{i,n_\ell} \le k_{n_\ell}]} + \frac{1}{k_{n_\ell}} \sum_{i=n_{\ell-1}+1}^{n_\ell} \mathbb{1}_{[\Sigma_{i,n_\ell} \le k_{n_\ell}]} \end{aligned}$$

$$\leq \frac{k_{n_\ell} - k_{n_{\ell-1}}}{k_{n_\ell}} + \frac{1}{k_{n_\ell}} \sum_{i=n_{\ell-1}+1}^{n_\ell} \mathbb{1}_{[\Sigma_{i,n_\ell} \leq k_{n_\ell}]}$$

$$\leq \mathrm{o}(1) + 1 - \frac{1}{(1+\delta)^c} + \frac{1}{k_{n_\ell}} \sum_{i=n_{\ell-1}+1}^{n_\ell} \mathbb{1}_{[\Sigma_{i,i} \leq k_{n_\ell}]}.$$

Calling the last term L_ℓ, recall from (11.12) that, for all N large enough,

$$\mathbb{P}\{L_\ell \geq e(2\delta)\} \leq e^{-k_{n_\ell}\delta/2}.$$

Since δ was arbitrary, the proof is complete. □

Remark 11.4.

(*i*) Walk (2008) proved the strong pointwise consistency of the k-nearest neighbor estimate under the sole condition $\mathbb{E}|Y| < \infty$ (universal consistency). The sequence $\{k_n\}$ is assumed to be regularly varying with exponent $\beta \in (0, 1]$, that is k_n is of the form $k_n = n^\beta L(n)$, where the function $L : (0, \infty) \to (0, \infty)$ varies slowly at infinity, i.e.,

$$\frac{L(tx)}{L(t)} \to 1 \quad \text{as } t \to \infty,$$

for every $x > 0$. Examples include $k_n = \lfloor n^\beta \rfloor$ $(0 < \beta < 1)$ and $k_n = \lfloor n/\log(n+1) \rfloor$.

Theorem 11.5 (Walk, 2008). *Assume that $\mathbb{E}|Y| < \infty$ and that the sequence $\{k\} = \{k_n\}$ is increasing and regularly varying with exponent $\beta \in (0, 1]$. Then, if $k \to \infty$ and $k/n \to 0$, the k-nearest neighbor regression function estimate is strongly consistent at μ-almost all $\mathbf{x} \in \mathbb{R}^d$.*

(*ii*) The k-nearest neighbor regression function estimate

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

can be regarded as the uniform kernel case of the more general estimate defined by

$$s_n(\mathbf{x}) = \frac{\sum_{i=1}^n K\left(\frac{\mathbf{x}-\mathbf{X}_i}{\|\mathbf{X}_{(k)}(\mathbf{x})-\mathbf{x}\|}\right) Y_i}{\sum_{j=1}^n K\left(\frac{\mathbf{x}-\mathbf{X}_j}{\|\mathbf{X}_{(k)}(\mathbf{x})-\mathbf{x}\|}\right)}, \tag{11.14}$$

which allows unequal weights to be given to the observations. In the spirit of Moore and Yackel (1977a,b) results for density estimation, Collomb (1980,

1981) provide some weak and strong pointwise consistency results for the estimate (11.14). The proofs involve a general lemma, showing that the properties of this estimate are consequences of the same properties for the standard Nadaraya-Watson estimate (kernel method). □

Chapter 12
Uniform consistency

12.1 Uniform consistency

In the present chapter, we consider the uniform convergence of

$$r_n(\mathbf{x}) = \sum_{i=1}^{n} v_{ni} Y_{(i)}(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d,$$

where $(v_{n1}, \ldots, v_{nn})$ is a vector of weights summing to one, and the sequence $(\mathbf{X}_{(1)}(\mathbf{x}), Y_{(1)}(\mathbf{x})), \ldots, (\mathbf{X}_{(n)}(\mathbf{x}), Y_{(n)}(\mathbf{x}))$ is a reordering of the data according to increasing Euclidean distances $\|\mathbf{X}_i - \mathbf{x}\|$. We are focusing on the convergence to zero of $\sup_{\mathbf{x} \in S} |r_n(\mathbf{x}) - r(\mathbf{x})|$ for all distributions of $(\mathbf{X}, Y)$ for which $\mathbf{X} \in \mathbb{R}^d$ has compact support S, and with some conditions on the tails of Y.

The supremum creates two problems—first of all, by moving $\mathbf{x}$ about $\mathbb{R}^d$, the data ordering changes. We will count the number of possible data permutations in the second section. Second, we need a uniform condition on the "noise" $Y - r(\mathbf{X})$ so that the averaging done by the weights v_{ni} is strong enough. This is addressed in the third section. In the fourth section, we prove our main theorem, which generalizes a result from Devroye (1978):

Theorem 12.1 (Strong uniform consistency). *Let* $\mathbf{X}$ *have distribution* μ *of compact support* S *on* $\mathbb{R}^d$*, and let the regression function* r *be continuous on* S*. Assume that the random variable* $Y - r(\mathbf{X})$ *given* $\mathbf{X} = \mathbf{x}$ *satisfies the uniform noise condition: there exists* $\lambda > 0$ *such that*

$$\sup_{\mathbf{x} \in \mathbb{R}^d} \mathbb{E}\left[e^{\lambda |Y - r(\mathbf{X})|} \,|\, \mathbf{X} = \mathbf{x} \right] < \infty.$$

Assume furthermore that, for all $\varepsilon > 0$*,*

G. Biau, L. Devroye, *Lectures on the Nearest Neighbor Method*,
Springer Series in the Data Sciences, DOI 10.1007/978-3-319-25388-6_12

$$\sum_{i>\varepsilon n} |v_{ni}| \to 0,$$

and that

$$\sum_{i=1}^{n} v_{ni} = 1, \quad \sup_n \sum_{i=1}^{n} |v_{ni}| < \infty,$$

and

$$(\log n)\big(\max_i |v_{ni}|\big) = o(1).$$

Then the corresponding nearest neighbor regression function estimate r_n satisfies

$$\sup_{\mathbf{x}\in S} |r_n(\mathbf{x}) - r(\mathbf{x})| \to 0 \quad \textit{almost surely}.$$

Remark 12.1.

(*i*) We allow weights that take negative values. As we will see later, there could be a benefit from such a choice.

(*ii*) For the standard k-nearest neighbor estimate, the conditions of Theorem 12.1 are equivalent to $k/\log n \to \infty$ and $k/n \to 0$.

(*iii*) For the supremum norm convergence, it is possible to widen the class of noise distributions at the expense of more restricted assumptions on k. For example, if we have $\mathbb{E}|Y|^p < \infty$ for $p > 0$, then the condition $k/\log n \to \infty$ for the k-nearest neighbor estimate r_n should be replaced by $k \log n/n^{1/p} \to \infty$. For $p = 2$, this has been done, e.g., by Cheng (1984). The idea is as follows: since

$$\mathbb{E}|Y|^p = \int_0^\infty \mathbb{P}\big\{|Y| > t^{1/p}\big\} \mathrm{d}t < \infty,$$

we see that in the sequence $Y_1, \ldots, Y_n$, $|Y_i| > i^{1/p}$ happens finitely often with probability one. The Y_i's with $|Y_i| > i^{1/p}$ are thus harmless. Then argue as in the proof of Theorem 12.1, using explicit exponential bounds for the random variable $|Y_i| \mathbb{1}_{[|Y_i| \le i^{1/p}]}$. □

Finally, in the last section, we show by means of a simple example that the conditions on k given above for the k-nearest neighbor estimate are necessary, even if Y remains bounded and $\mathbf{X}$ is uniform on $[0, 1]$. In other words, the conditions of Theorem 12.1 are optimal.

12.2 The number of reorderings of the data

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ be vectors with possibly some duplicates. For fixed $\mathbf{x} \in \mathbb{R}^d$, reorder these vectors according to increasing values of $\|\mathbf{x}_i - \mathbf{x}\|$, breaking, as always, ties by considering indices. Let $(\sigma_1, \ldots, \sigma_n)$ be the permutation thus obtained:

$$\|\mathbf{x}_{\sigma_1} - \mathbf{x}\| \le \cdots \le \|\mathbf{x}_{\sigma_n} - \mathbf{x}\|.$$

The inverse is the rank Σ_i, i.e.,

$$\mathbf{x}_i = \mathbf{x}_{\sigma_{\Sigma_i}}, \quad 1 \le i \le n.$$

Let

$$\mathscr{W} = \left\{(\Sigma_1, \ldots, \Sigma_n) : \mathbf{x} \in \mathbb{R}^d\right\}$$

be the set of all rank vectors one can observe by moving $\mathbf{x}$ around in space. Similarly,

$$\mathscr{S} = \left\{(\sigma_1, \ldots, \sigma_n) : \mathbf{x} \in \mathbb{R}^d\right\}.$$

Notation $|\mathscr{W}|$ (respectively $|\mathscr{S}|$) stands for the cardinality of $\mathscr{W}$ (respectively $\mathscr{S}$).

Theorem 12.2. *One has*

$$|\mathscr{W}| = |\mathscr{S}| \le \left(\frac{25}{d}\right)^d n^{2d} \quad \textit{for all } n \ge 2d.$$

Proof. The hyperplane $\|\mathbf{x} - \mathbf{x}_i\|^2 = \|\mathbf{x} - \mathbf{x}_j\|^2$ generates a sign

$$p_{ij}(\mathbf{x}) = \begin{cases} 1 & \text{if } \|\mathbf{x} - \mathbf{x}_i\|^2 > \|\mathbf{x} - \mathbf{x}_j\|^2 \\ 0 & \text{if } \|\mathbf{x} - \mathbf{x}_i\|^2 = \|\mathbf{x} - \mathbf{x}_j\|^2 \\ -1 & \text{if } \|\mathbf{x} - \mathbf{x}_i\|^2 < \|\mathbf{x} - \mathbf{x}_j\|^2. \end{cases}$$

The collection of signs $(p_{ij}(\mathbf{x}), 1 \le i < j \le n)$, called the sign pattern, determines the ordering of $\|\mathbf{x} - \mathbf{x}_i\|^2$ and identifies all ties. There are $3^{\binom{n}{2}}$ possible sign patterns, but not all of them are possible in $\mathbb{R}^d$.

For $d = 1$, it is easy to see that the number of sign patterns of N polynomials of degree not exceeding D is at most $2ND + 1$. For larger dimensions, the Milnor-Thom theorem (Petrovskiĭ and Oleĭnik, 1952; Milnor, 1964; Thom, 1965) states that the maximal number of sign patterns of N polynomials of degree at most D in $\mathbb{R}^d$ is

$$\left(\frac{50DN}{d}\right)^d \quad \text{for all } N \ge d \ge 2$$

(see also Warren, 1968, and Pollack and Roy, 1993). Better bounds are known for hyperplane arrangements (i.e., when $D = 1$), but they are still $O(N^d)$ for d fixed—see the discussion in Matoušek (2002, Chapter 6), or in Grünbaum (1972). For our example, $D = 1, N = \binom{n}{2}$, so that for any $d \geq 1$ and all $n \geq 2d$, the number of sign patterns is not more than

$$\left(\frac{25}{d}\right)^d n^{2d}. \qquad \square$$

12.3 A uniform exponential tail condition

In regression function estimation, the residual $Y - r(\mathbf{X})$ is sometimes called the "noise." It measures the departure from the regression function r. In this text, we have several conditions on the noise, starting with the standard one, $\mathbb{E}|Y|^p < \infty$ for some $p > 0$. In some places, we assume $\|Y\|_\infty < \infty$, where $\|Y\|_\infty$ is the essential supremum of Y.

A practical assumption that captures many important applications is the uniform exponential condition: there exists $\lambda > 0$ such that

$$\sup_{\mathbf{x} \in \mathbb{R}^d} \mathbb{E}\left[e^{\lambda |Y - r(\mathbf{X})|} \mid \mathbf{X} = \mathbf{x}\right] < \infty.$$

This class contains all homoscedastic cases, i.e., cases in which $Y - r(\mathbf{X})$ given $\mathbf{X} = \mathbf{x}$ is distributed as Z with $\mathbb{E}Z = 0$, and for which $\mathbb{E}e^{\lambda |Z|} < \infty$ for some $\lambda > 0$. Examples for Z include the Gaussian and exponential distributions, as well as all bounded distributions.

Lemma 12.1. *Let $Z_1, Z_2, \ldots$ be a sequence of independent zero-mean real-valued random variables with*

$$\sup_{n \geq 1} \mathbb{E}e^{\lambda |Z_n|} \leq c < \infty, \tag{12.1}$$

for a given $\lambda > 0$ and some constant $c > 0$. Let $(v_{n1}, \ldots, v_{nn})$ be a weight vector, with $v_n = \max_i |v_{ni}| > 0$. Let $\Delta_n = \sum_{i=1}^n |v_{ni}|$. Then, for all $\varepsilon > 0$,

$$\mathbb{P}\left\{\sum_{i=1}^n v_{ni} Z_i \geq \varepsilon\right\} \leq \exp\left(-\frac{\varepsilon^2 \lambda^2}{8c\Delta_n v_n}\right), \quad \varepsilon \leq \frac{\Delta_n}{\lambda} \min(1, 2c).$$

Similarly,

$$\mathbb{P}\left\{\sum_{i=1}^{n} v_{ni}Z_i \le -\varepsilon\right\} \le \exp\left(-\frac{\varepsilon^2\lambda^2}{8c\Delta_n v_n}\right), \quad \varepsilon \le \frac{\Delta_n}{\lambda}\min(1, 2c).$$

Proof. Fix $\lambda > 0$ so that (12.1) holds. Then note that by Taylor series expansion with remainder,

$$\begin{aligned}\mathbb{E}e^{\alpha Z_n} &\le 1 + \alpha\mathbb{E}Z_n + \frac{\alpha^2}{2}\mathbb{E}Z_n^2 + \frac{|\alpha|^3}{6}\mathbb{E}\left[|Z_n|^3 e^{|\alpha Z_n|}\right] \\ &\le 1 + \frac{\alpha^2}{2}\mathbb{E}Z_n^2 + \frac{c|\alpha|^3}{6}\mathbb{E}|Z_n|^3, \quad \text{if } |\alpha| \le \lambda.\end{aligned}$$

Observe that

$$\frac{\lambda^2}{2}\mathbb{E}Z_n^2 \le \mathbb{E}e^{\lambda|Z_n|} \le c.$$

So,

$$\sup_{n\ge 1}\mathbb{E}Z_n^2 \le \frac{2c}{\lambda^2}.$$

Similarly,

$$\frac{\lambda^3}{6}\mathbb{E}|Z_n|^3 \le \mathbb{E}e^{\lambda|Z_n|} \le c,$$

and thus

$$\sup_{n\ge 1}\mathbb{E}|Z_n|^3 \le \frac{6c}{\lambda^3}.$$

We conclude that for $|\alpha| \le \lambda$,

$$\mathbb{E}e^{\alpha Z_n} \le 1 + \frac{c\alpha^2}{\lambda^2} + \frac{c^2|\alpha|^3}{\lambda^3}.$$

By Chernoff's bounding method (Theorem 20.5 in the Appendix), for $\varepsilon > 0$, $\gamma > 0$,

$$\begin{aligned}\mathbb{P}\left\{\sum_{i=1}^{n} v_{ni}Z_i \ge \varepsilon\right\} &\le e^{-\gamma\varepsilon}\mathbb{E}e^{\gamma\sum_{i=1}^{n} v_{ni}Z_i} \\ &= e^{-\gamma\varepsilon}\prod_{i=1}^{n}\mathbb{E}e^{\gamma\, v_{ni}Z_i}\end{aligned}$$

$$\leq e^{-\gamma\varepsilon} \prod_{i=1}^{n} \left(1 + \frac{c\gamma^2 v_{ni}^2}{\lambda^2} + \frac{c^2\gamma^3 |v_{ni}|^3}{\lambda^3}\right)$$

(assuming $\gamma\, v_n \leq \lambda$)

$$\leq \exp\left(-\gamma\varepsilon + \sum_{i=1}^{n} \left(\frac{c\gamma^2 v_{ni}^2}{\lambda^2} + \frac{c^2\gamma^3 |v_{ni}|^3}{\lambda^3}\right)\right)$$

(since $1 + u \leq e^u$ for all u)

$$\leq \exp\left(-\gamma\varepsilon + \frac{c\gamma^2 \Delta_n v_n}{\lambda^2} + \frac{c^2\gamma^3 \Delta_n v_n^2}{\lambda^3}\right).$$

Put $\gamma = \frac{\varepsilon\lambda^2}{2c\Delta_n v_n}$ (which minimizes the sum of the first two terms in the exponent). Then the bound is

$$\exp\left(-\frac{\varepsilon^2\lambda^2}{4c\Delta_n v_n} + \frac{\varepsilon^3\lambda^3}{8c\Delta_n^2 v_n}\right) \leq \exp\left(-\frac{\varepsilon^2\lambda^2}{8c\Delta_n v_n}\right)$$

if $\varepsilon\lambda/\Delta_n \leq 1$. Note also that $\gamma\, v_n \leq \lambda$ implies that we should have $\varepsilon\lambda/\Delta_n \leq 2c$. □

12.4 Proof of Theorem 12.1

We shall begin the proof of Theorem 12.1 with a few preliminary lemmas.

Lemma 12.2. *Assume that the support S of μ is compact. Then, for all $\varepsilon > 0$,*

$$\inf_{\mathbf{x}\in S} \mu\left(B(\mathbf{x}, \varepsilon)\right) > 0.$$

Proof. Suppose that the result is false and that $\inf_{\mathbf{x}\in S} \mu(B(\mathbf{x}, \varepsilon)) = 0$ for some $\varepsilon > 0$. Then there exists a sequence $\{\mathbf{x}_i\}$ from S with $\mu(B(\mathbf{x}_i, \varepsilon)) \to 0$. Since S is compact, the sequence $\{\mathbf{x}_i\}$ must have a cluster point $\mathbf{y}$ in S. Therefore, there exists a further subsequence $\{\mathbf{x}_i^\star\}$ such that

$$\mu(B(\mathbf{x}_i^\star, \varepsilon)) \to 0 \quad \text{and} \quad \|\mathbf{x}_i^\star - \mathbf{y}\| \leq \varepsilon/2 \quad \text{for all } i.$$

Thus, $B(\mathbf{y}, \varepsilon/2)$ is contained in the intersection of all the $B(\mathbf{x}_i^\star, \varepsilon)$. Hence

$$\mu\left(B(\mathbf{y}, \varepsilon/2)\right) \leq \liminf_{i\to\infty} \mu\left(B(\mathbf{x}_i^\star, \varepsilon)\right) = 0,$$

which contradicts the fact that $\mathbf{y}$ belongs to S. □

Lemma 12.3. *Assume that the support S of μ is compact and that the regression function r is continuous on S. If, for all $\varepsilon > 0$, $\sum_{i>\varepsilon n} |v_{ni}| \to 0$, $\sum_{i=1}^n v_{ni} = 1$ and $\sup_n \sum_{i=1}^n |v_{ni}| < \infty$, then*

$$\sup_{\mathbf{x}\in S} |\tilde{r}_n(\mathbf{x}) - r(\mathbf{x})| \to 0 \quad \textit{almost surely},$$

where

$$\tilde{r}_n(\mathbf{x}) = \sum_{i=1}^n v_{ni} r\left(\mathbf{X}_{(i)}(\mathbf{x})\right).$$

Proof. Let $\varepsilon > 0$ be arbitrary, and find a $\delta > 0$ such that $|r(\mathbf{y}) - r(\mathbf{x})| \le \varepsilon$ for all $\mathbf{x}, \mathbf{y} \in S$ such that $\|\mathbf{y} - \mathbf{x}\| \le \delta$ (use the uniform continuity of r on S). Let $r^\star = \|r\|_\infty \sup_n \sum_{i=1}^n |v_{ni}|$, let $\theta \in (0, 1/2]$ be arbitrary, and let A_n be the event

$$A_n = \left[\|\mathbf{X}_{(\lceil\theta n\rceil)}(\mathbf{x}) - \mathbf{x}\| \le \delta \text{ for all } \mathbf{x} \in S\right].$$

Clearly,

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{\mathbf{x}\in S} |\tilde{r}_n(\mathbf{x}) - r(\mathbf{x})| > 2\varepsilon\right\} \\
&\quad \le \mathbb{P}\{A_n^c\} + \mathbb{P}\left\{A_n, \sup_{\mathbf{x}\in S} |\tilde{r}_n(\mathbf{x}) - r(\mathbf{x})| > 2\varepsilon\right\} \\
&\quad = \mathbb{P}\{A_n^c\} + \mathbb{P}\left\{A_n, \sup_{\mathbf{x}\in S} \left| \sum_{i=1}^{\lceil\theta n\rceil} v_{ni} \left(r\left(\mathbf{X}_{(i)}(\mathbf{x})\right) - r(\mathbf{x})\right) \right| + r^\star \sum_{i>\lceil\theta n\rceil} |v_{ni}| > 2\varepsilon\right\}.
\end{aligned}$$

Thus, for all n large enough,

$$\mathbb{P}\left\{\sup_{\mathbf{x}\in S} |\tilde{r}_n(\mathbf{x}) - r(\mathbf{x})| > 2\varepsilon\right\} \le \mathbb{P}\{A_n^c\}$$

since, on the event A_n, $|r(\mathbf{X}_{(i)}(\mathbf{x})) - r(\mathbf{x})| \le \varepsilon$ for all $\mathbf{x} \in S$ and $1 \le i \le \lceil\theta n\rceil$, and $r^\star \sum_{i>\lceil\theta n\rceil} |v_{ni}| \le \varepsilon$ for all n large enough by our assumption.

Next,

$$\begin{aligned}
A_n^c &\subseteq \left[\mu_n\left(B(\mathbf{x}, \delta)\right) < \frac{\lceil\theta n\rceil}{n} \text{ for some } \mathbf{x} \in S\right] \\
&\subseteq \left[\mu_n\left(B(\mathbf{x}, \delta)\right) < 2\theta \text{ for some } \mathbf{x} \in S\right] \\
&\qquad \text{(for all } n \text{ large enough, since } \theta \in (0, 1/2]),
\end{aligned}$$

where $\mu_n(B) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{[\mathbf{X}_i \in B]}$ is the empirical measure of a Borel set B with $\mathbf{X}_1, \ldots, \mathbf{X}_n$. From Lemma 12.2,

$$\inf_{\mathbf{x}\in S} \mu\left(B(\mathbf{x}, \delta/2)\right) = c > 0.$$

Since S is compact, we can find a finite number N of points $\mathbf{x}_1, \ldots, \mathbf{x}_N$ from S with the property that, for every $\mathbf{x}$ in S, there exists an $\mathbf{x}_i$, $1 \le i \le N$, with $\|\mathbf{x} - \mathbf{x}_i\| \le \delta/2$. Thus,

$$\begin{aligned}
\mathbb{P}\Big\{\sup_{\mathbf{x}\in S} |\tilde{r}_n(\mathbf{x}) - r(\mathbf{x})| > 2\varepsilon\Big\} &\le \mathbb{P}\Big\{\inf_{\mathbf{x}\in S} \mu_n\left(B(\mathbf{x}, \delta)\right) < 2\theta\Big\} \\
&\le \mathbb{P}\Big\{\bigcup_{i=1}^N \big\{\mu_n\left(B(\mathbf{x}_i, \delta/2)\right) < 2\theta\big\}\Big\} \\
&\le N \sup_{\mathbf{x}\in S} \mathbb{P}\big\{\mu_n\left(B(\mathbf{x}, \delta/2)\right) < 2\theta\big\} \\
&= N \sup_{\mathbf{x}\in S} \mathbb{P}\big\{\mathrm{Bin}(n, p_{\mathbf{x}}) < 2n\theta\big\},
\end{aligned}$$

where $p_{\mathbf{x}} = \mu(B(\mathbf{x}, \delta/2)) > 0$. Therefore, if $\theta < c/4$, then, by Chernoff's bound (Theorem 20.5 in the Appendix),

$$\begin{aligned}
\mathbb{P}\Big\{\sup_{\mathbf{x}\in S} |\tilde{r}_n(\mathbf{x}) - r(\mathbf{x})| > 2\varepsilon\Big\} &\le N \sup_{\mathbf{x}\in S} \mathbb{P}\big\{\mathrm{Bin}(n, p_{\mathbf{x}}) < nc/2\big\} \\
&\le N \sup_{\mathbf{x}\in S} \exp\left(\frac{nc}{2} - np_{\mathbf{x}} - \frac{nc}{2}\log\left(\frac{c}{2p_{\mathbf{x}}}\right)\right) \\
&\le N \exp\left(-\frac{nc}{2}(1 - \log 2)\right),
\end{aligned}$$

where, in the last inequality, we used the fact that $\inf_{p\in[c,1]}(p + \frac{c}{2}\log(\frac{c}{2p})) = c + \frac{c}{2}\log\frac{1}{2}$. In conclusion, for all n large enough,

$$\mathbb{P}\Big\{\sup_{\mathbf{x}\in S} |\tilde{r}_n(\mathbf{x}) - r(\mathbf{x})| > 2\varepsilon\Big\} \le Ne^{-\alpha n}$$

for some $\alpha > 0$, and Lemma 12.3 now follows by the Borel-Cantelli lemma since $e^{-\alpha n}$ is summable with respect to n. □

We are now in a position to prove Theorem 12.1.

Proof (Theorem 12.1). In view of Lemma 12.3, we only need to show that

$$\sum_{n\ge 1} \mathbb{P}\Big\{\sup_{\mathbf{x}\in S} |r_n(\mathbf{x}) - \tilde{r}_n(\mathbf{x})| > \varepsilon\Big\} < \infty$$

for all $\varepsilon > 0$, where we recall that

$$\tilde{r}_n(\mathbf{x}) = \sum_{i=1}^{n} v_{ni} r\left(\mathbf{X}_{(i)}(\mathbf{x})\right).$$

Note that

$$r_n(\mathbf{x}) - \tilde{r}_n(\mathbf{x}) = \sum_{i=1}^{n} v_{ni}\left[Y_{(i)}(\mathbf{x}) - r\left(\mathbf{X}_{(i)}(\mathbf{x})\right)\right].$$

According to Proposition 8.1, conditional on $\mathbf{X}_1, \ldots, \mathbf{X}_n$, the random variables $Y_{(1)}(\mathbf{x}) - r(\mathbf{X}_{(1)}(\mathbf{x})), \ldots, Y_{(n)}(\mathbf{x}) - r(\mathbf{X}_{(n)}(\mathbf{x}))$ are independent with zero mean. Moreover, they satisfy the uniform noise condition

$$\sup_{(\mathbf{x}_1,\ldots,\mathbf{x}_n)\in\mathbb{R}^{dn}} \mathbb{E}\left[e^{\lambda\left|Y_{(i)}(\mathbf{x})-r(\mathbf{X}_{(i)}(\mathbf{x}))\right|} \,\middle|\, (\mathbf{X}_1,\ldots,\mathbf{X}_n) = (\mathbf{x}_1,\ldots,\mathbf{x}_n)\right] \le c < \infty$$

for some $\lambda, c > 0$. By Lemma 12.1,

$$\begin{aligned}
&\sup_{(\mathbf{x}_1,\ldots,\mathbf{x}_n)\in\mathbb{R}^{dn}} \mathbb{P}\left\{\left|\sum_{i=1}^{n} v_{ni}\left|Y_{(i)}(\mathbf{x}) - r\left(\mathbf{X}_{(i)}(\mathbf{x})\right)\right|\right| > \varepsilon \,\middle|\, (\mathbf{X}_1,\ldots,\mathbf{X}_n) = (\mathbf{x}_1,\ldots,\mathbf{x}_n)\right\} \\
&\quad \le 2\exp\left(-\frac{\varepsilon^2\lambda^2}{8c\Delta_n v_n}\right), \quad \varepsilon \le \frac{\Delta_n}{\lambda}\min(1, 2c), \qquad (12.2)
\end{aligned}$$

where $v_n = \max_i |v_{ni}|$ and $\Delta_n = \sum_{i=1}^{n} |v_{ni}|$.

While this inequality is true for all $\mathbf{x}$, the supremum over $\mathbf{x}$ is under the probability sign. Thus, letting

$$\mathscr{W} = \left\{(\sigma_1, \ldots, \sigma_n) : \text{all permutations of } (1, \ldots, n) \text{ obtainable by moving } \mathbf{x} \text{ in } \mathbb{R}^d\right\},$$

we may write

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{\mathbf{x}\in S} |r_n(\mathbf{x}) - \tilde{r}_n(\mathbf{x})| > \varepsilon \,\middle|\, (\mathbf{X}_1,\ldots,\mathbf{X}_n) = (\mathbf{x}_1,\ldots,\mathbf{x}_n)\right\} \\
&\quad \le \mathbb{P}\left\{\bigcup_{(\sigma_1,\ldots,\sigma_n)\in\mathscr{W}} \left|\sum_{i=1}^{n} v_{ni}\left(Y_{\sigma_i} - r(\mathbf{X}_{\sigma_i})\right)\right| > \varepsilon \,\middle|\, (\mathbf{X}_1,\ldots,\mathbf{X}_n) = (\mathbf{x}_1,\ldots,\mathbf{x}_n)\right\} \\
&\quad \le \sum_{(\sigma_1,\ldots,\sigma_n)\in\mathscr{W}} \mathbb{P}\left\{\left|\sum_{i=1}^{n} v_{ni}\left(Y_{\sigma_i} - r(\mathbf{X}_{\sigma_i})\right)\right| > \varepsilon \,\middle|\, (\mathbf{X}_1,\ldots,\mathbf{X}_n) = (\mathbf{x}_1,\ldots,\mathbf{x}_n)\right\} \\
&\quad \le \left(\frac{25}{d}\right)^d n^{2d} \times 2\exp\left(-\frac{\varepsilon^2\lambda^2}{8c\Delta_n v_n}\right), \quad \varepsilon \le \frac{\Delta_n}{\lambda}\min(1, 2c),
\end{aligned}$$

by combining Theorem 12.2 and (12.2). Observe that $\sup_n \Delta_n < \infty$ and that $\Delta_n \geq |\sum_{i=1}^n v_{ni}| = 1$. Thus, for all $\varepsilon > 0$, the upper bound is summable in n when $(\log n)v_n \to 0$. □

Remark 12.2. The proof of Theorem 12.1 uses the Milnor-Thom theorem (see Section 12.2). For the standard k-nearest neighbor estimate, what matters in the proof is the number of ways of "grabbing" k points by balls $B(\mathbf{x}, \rho)$, $\mathbf{x} \in \mathbb{R}^d$, $\rho \geq 0$. Since the k-nearest neighbor estimate can only use one of these sets, one needs an upper bound for

$$\left| \left\{ \{\mathbf{X}_1, \ldots, \mathbf{X}_n\} \cap B(\mathbf{x}, \rho) : \mathbf{x} \in \mathbb{R}^d, \rho \geq 0 \right\} \right|.$$

If $\mathscr{A}$ is a class of sets on $\mathbb{R}^d$ (such as all balls $B(\mathbf{x}, \rho)$) then the shatter coefficient of $\mathscr{A}$ is

$$\mathbf{S}(\mathscr{A}, n) = \max_{(\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{dn}} \left| \left\{ \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \cap A : A \in \mathscr{A} \right\} \right|.$$

It plays a crucial role in the work of Vapnik and Chervonenkis (Vapnik and Chervonenkis, 1971—see also Devroye and Lugosi, 2001) for finding uniform bounds for empirical processes.

For example, for all hyperplanes in $\mathbb{R}^d$, we have

$$\mathbf{S}(\mathscr{A}, n) \leq 2n^d + 2$$

(Devroye et al., 1996, page 223), and for the class of closed balls in $\mathbb{R}^d$,

$$\mathbf{S}(\mathscr{A}, n) \leq 2n^{d+1} + 2.$$

The latter result suffices to prove Theorem 12.1 for the k-nearest neighbor estimate. □

12.5 The necessity of the conditions on k

Consider the standard k-nearest neighbor estimate when $\mathbf{X}$ has compact support, $|Y| \leq 1$, and r is uniformly continuous on the support of $\mathbf{X}$. It is a simple exercise to show that $k/n \to 0$ is necessary for weak uniform consistency, because it is even necessary for weak pointwise consistency at one point of the support. The necessity of $k/\log n \to \infty$ for supremum norm convergence can be shown by a simple example on $[0, 1]$.

Let Y be independent of X and Rademacher, i.e., $Y = \pm 1$ with equal probability, and let X be uniform on $[0, 1]$. We have $r(x) = 0$, and

$$r_n(x) = \frac{1}{k}\sum_{i=1}^{k} Y_{(i)}(x) \stackrel{\mathscr{D}}{=} \frac{1}{k}\left(2\mathrm{Bin}\left(k, \tfrac{1}{2}\right) - k\right),$$

for all x. Define

$$x_j = \frac{2jk}{n}, \quad 1 \le j \le \frac{n}{2k} - 1,$$

where we assume without loss of generality that n is a multiple of $2k$ and $\frac{n}{2k} \ge 2$. Then for $\varepsilon > 0$, assume that $\mathbb{P}\{\sup_{x\in\mathbb{R}} |r_n(x)| > \varepsilon\} \to 0$. Define

$$A = \left\{ j \in \left\{1, \ldots, \frac{n}{2k} - 1\right\} : |X_{(k)}(x_j) - x_j| \le \frac{k}{n} \right\}.$$

Clearly, given A, the $r_n(x_j)$'s, $j \in A$, are all independent because the k-neighborhoods do not overlap, and thus different Y_i's are used in $r_n(x_j)$, $j \in A$. Let $N = |A|$. Then N is the number of components (among $\frac{n}{2k} - 1$) of a multinomial random vector $(N_1, \ldots, N_{\frac{n}{2k}-1})$ whose value is larger than or equal to k, where $(N_1, \ldots, N_{\frac{n}{2k}})$ has parameters $(n, \frac{2k}{n}, \ldots, \frac{2k}{n})$. We have, conditioning on A,

$$\mathbb{P}\left\{ \sup_{x\in\mathbb{R}} |r_n(x)| \le \varepsilon \right\} \le \mathbb{E}\left[\mathbb{P}^N \left\{ |r_n(x_1)| \le \varepsilon \right\} \right] \stackrel{\text{def}}{=} \mathbb{E}q^N.$$

Thus,

$$\begin{aligned}
\mathbb{E}q^N = \mathbb{E}\left[\prod_{j=1}^{\frac{n}{2k}-1} q^{\mathbb{1}_{[N_j \ge k]}} \right] &= \mathbb{E}\left[\prod_{j=1}^{\frac{n}{2k}-1} \left(q\mathbb{1}_{[N_j \ge k]} + \mathbb{1}_{[N_j < k]} \right) \right] \\
&= \mathbb{E}\left[\prod_{j=1}^{\frac{n}{2k}-1} \left(1 - (1-q)\mathbb{1}_{[N_j \ge k]} \right) \right] \\
&\le \mathbb{E}\left[\prod_{j=1}^{\frac{n}{2k}-1} \left(1 - (1-q)\xi_j \right) \right],
\end{aligned}$$

where $\xi_1, \ldots, \xi_{\frac{n}{2k}-1}$ are independent Bernoulli random variables with success probability $\mathbb{P}\{N_j \ge k\}$. Here we used the negative association property of the multinomial law—see Marshall and Olkin (1979), or Tong (1980). Next, by the Chebyshev-Cantelli inequality (Theorem 20.11 in the Appendix),

$$\mathbb{P}\{N_j \ge k\} = \mathbb{P}\left\{ \mathrm{Bin}\left(n, \tfrac{2k}{n}\right) \ge k \right\} \ge \frac{k}{k+2} \stackrel{\text{def}}{=} p.$$

Therefore,

$$\begin{aligned}\mathbb{E}q^N &\le (1-(1-q)p)^{\frac{n}{2k}-1}\\ &\le \exp\left(-(1-q)p\left(\frac{n}{2k}-1\right)\right)\\ &\quad \text{(since } 1-u\le e^{-u} \text{ for all } u\text{)}\\ &= \exp\left(-(1-q)\frac{k}{k+2}\left(\frac{n}{2k}-1\right)\right).\end{aligned}$$

By assumption, this term tends to one for every $\varepsilon > 0$, and hence, the exponent must tend to zero. Therefore $(1-q)\frac{n}{2k} \to 0$, or,

$$\frac{n}{2k} \times \mathbb{P}\left\{|r_n(x_1)| > \varepsilon\right\} \to 0.$$

But, for $\varepsilon \in (0,1)$, this is

$$\begin{aligned}\frac{n}{2k} \times \mathbb{P}\left\{\left|\mathrm{Bin}\left(k, \tfrac{1}{2}\right) - \frac{k}{2}\right| > \frac{\varepsilon k}{2}\right\} &\ge \frac{n}{2k} \times \mathbb{P}\left\{\mathrm{Bin}\left(k, \tfrac{1}{2}\right) > \frac{k}{2}(1+\varepsilon)\right\}\\ &= \frac{n}{2k} \times \Theta\left(\frac{1}{\sqrt{k}}\right) \times \left(\frac{1}{(1+\varepsilon)^{1+\varepsilon}(1-\varepsilon)^{1-\varepsilon}}\right)^k,\end{aligned}$$

by properties of the tail of the binomial, where $\Theta(1/\sqrt{k})$ denotes a factor sandwiched between $c_1/\sqrt{k}$ and $c_2/\sqrt{k}$ for $0 < c_1 < c_2 < \infty$.

Assume that $k \le \delta \log n$ for some $0 < \delta < \infty$ along a subsequence. Then along that subsequence, the lower bound is larger than

$$\Theta\left(\frac{n}{\log^{3/2} n}\right) \times \left(\frac{1}{(1+\varepsilon)^{1+\varepsilon}(1-\varepsilon)^{1-\varepsilon}}\right)^{\delta \log n}.$$

We can find ε small enough such that this does not tend to zero along that subsequence, and thus obtain a contradiction. Therefore, $k/\log n \to \infty$.

Chapter 13
Advanced properties of uniform order statistics

13.1 Moments

Various properties of $U_{(1)}, \ldots, U_{(n)}$, uniform [0, 1] order statistics, will be needed in the analysis that follows. These are collected in the present chapter. The first group of properties is directly related to $U_{(i)}$ $(1 \leq i \leq n)$, while the second group deals with random linear combinations of them.

Recall (Corollary 1.1) that

$$\left(U_{(1)}, \ldots, U_{(n)}\right) \stackrel{\mathscr{D}}{=} \left(\frac{G_1}{G_{n+1}}, \ldots, \frac{G_n}{G_{n+1}}\right),$$

where $G_i = \sum_{j=1}^{i} E_j$, $1 \leq i \leq n+1$, and $E_1, \ldots, E_{n+1}$ are independent standard exponential random variables. In particular, G_i is Gamma(i) and $U_{(i)}$ is Beta$(i, n+1-i)$ distributed, i.e., $U_{(i)}$ has density

$$\frac{x^{i-1}(1-x)^{n-i}}{B(i, n+1-i)}, \quad 0 \leq x \leq 1,$$

where

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

We start with the following simple lemmas.

Lemma 13.1. *For $\alpha > 0$,*

$$\mathbb{E}U_{(i)}^{\alpha} = \frac{\Gamma(i+\alpha)\Gamma(n+1)}{\Gamma(i)\Gamma(n+1+\alpha)}. \tag{13.1}$$

G. Biau, L. Devroye, *Lectures on the Nearest Neighbor Method*,
Springer Series in the Data Sciences, DOI 10.1007/978-3-319-25388-6_13

Furthermore,

$$\max_{1\le i\le n} \mathbb{E}\left[\left(\frac{U_{(i)}}{i/n}\right)^{\alpha}\right] \le (1+\alpha)^{1+\alpha}.$$

Proof. The first statement follows by working out the moments of the beta distribution (see Section 20.9 in the Appendix). Using Theorem 20.14, (13.1) is not larger than

$$\frac{\left(1+\frac{\alpha}{i}\right)^{i}}{\left(1+\frac{\alpha}{n+1}\right)^{n+1}} \times \left(\frac{i+\alpha}{n+1+\alpha}\right)^{\alpha} \times \sqrt{\frac{i}{i+\alpha}} \times \sqrt{\frac{n+1+\alpha}{n+1}}.$$

Since $(1+\frac{\alpha}{u})^u$ is increasing in $u>0$, and $1\le i\le n$, this is not more than

$$\left(1+\frac{\alpha}{2}\right)\left(\frac{i+\alpha}{n+1+\alpha}\right)^{\alpha}.$$

We conclude that

$$\begin{aligned}
\max_{1\le i\le n} \mathbb{E}\left[\left(\frac{U_{(i)}}{i/n}\right)^{\alpha}\right] &\le \left(1+\frac{\alpha}{2}\right) \max_{1\le i\le n}\left(\frac{i+\alpha}{i}\right)^{\alpha}\left(\frac{n}{n+1+\alpha}\right)^{\alpha} \\
&\le \left(1+\frac{\alpha}{2}\right)(1+\alpha)^{\alpha} \\
&\le (1+\alpha)^{1+\alpha}. \quad \square
\end{aligned}$$

Lemma 13.2. *For $\alpha>0$,*

$$\max_{k\le i\le n} \mathbb{E}\left[\left(\frac{U_{(i)}}{i/n}\right)^{\alpha}\right] \le \left(1+\frac{\alpha}{k}\right)^{\max(\alpha,1)}$$

and, if $k>\alpha$,

$$\min_{k\le i\le n} \mathbb{E}\left[\left(\frac{U_{(i)}}{i/n}\right)^{\alpha}\right] \ge \left(\frac{n}{n+1}\right)^{\alpha} \exp\left(-\frac{\alpha}{k}-\frac{\alpha^2}{2(k-\alpha)}\right).$$

Thus,

$$\max_{k\le i\le n}\left|\mathbb{E}\left[\left(\frac{U_{(i)}}{i/n}\right)^{\alpha}\right]-1\right| = \mathrm{O}\left(\frac{1}{k}\right) \quad \text{as } k\to\infty.$$

Proof. Fix $i\in\{k,\ldots,n\}$. By Theorem 20.14,

$$\mathbb{E}U_{(i)}^{\alpha} \le \frac{\left(1+\frac{\alpha}{i}\right)^{i}}{\left(1+\frac{\alpha}{n+1}\right)^{n+1}} \times \left(\frac{i+\alpha}{n+1+\alpha}\right)^{\alpha} \times \frac{n+1+\alpha}{n+1}.$$

Therefore,

$$\mathbb{E}\left[\left(\frac{U_{(i)}}{i/n}\right)^{\alpha}\right] \le \left(\frac{1+\frac{\alpha}{i}}{1+\frac{\alpha}{n+1}}\right)^{\alpha}\left(1+\frac{\alpha}{n+1}\right)$$
$$= \left(1+\frac{\alpha}{i}\right)^{\alpha}\left(1+\frac{\alpha}{n+1}\right)^{1-\alpha}$$
$$\le \left(1+\frac{\alpha}{i}\right)^{\max(\alpha,1)}.$$

For the lower bound, note that

$$\mathbb{E}U_{(i)}^{\alpha} \ge \frac{\left(1+\frac{\alpha}{i}\right)^{i}}{\left(1+\frac{\alpha}{n+1}\right)^{n+1}} \times \left(\frac{i+\alpha}{n+1+\alpha}\right)^{\alpha} \times \frac{i}{i+\alpha},$$

so that

$$\mathbb{E}\left[\left(\frac{U_{(i)}}{i/n}\right)^{\alpha}\right] \ge \left(\frac{n}{n+1}\right)^{\alpha} e^{-\alpha}\left(1+\frac{\alpha}{i}\right)^{i}\left(\frac{1+\frac{\alpha}{i}}{1+\frac{\alpha}{n+1}}\right)^{\alpha} \times \frac{1}{1+\frac{\alpha}{i}}$$
(since $1+u \le e^{u}$ for all u)
$$\ge \left(\frac{n}{n+1}\right)^{\alpha} e^{-\alpha}\left(1+\frac{\alpha}{k}\right)^{k-1}$$
$$\ge \left(\frac{n}{n+1}\right)^{\alpha} \exp\left(-\frac{\alpha}{k}-\frac{\alpha^2}{2(k-\alpha)}\right)$$
(where we used $\log(1+u) \ge u - \frac{u^2}{2(1-u)}$ for $0 < u < 1$). □

13.2 Large deviations

We will need large deviation bounds for $U_{(i)}$. These are easy to derive by Chernoff's bounding method applied to the gamma distribution (see Section 20.3.2 in the Appendix). We first summarize the result:

Theorem 13.1. *For $\delta \in (0, 1/2]$, define*

$$\varphi(\delta) = -\frac{\delta}{3} - \log\left(1-\frac{\delta}{3}\right).$$

Then

$$\max_{1\le i\le n}\left(\frac{\mathbb{P}\left\{U_{(i)}\ge \frac{i}{n}(1+\delta)\right\}}{\exp\left(-i\varphi(\delta)\right)}\right)\le 2.$$

For $\delta\in(0,1)$*, define*

$$\psi(\delta)=\frac{\delta}{2}-\log\left(1+\frac{\delta}{2}\right).$$

Then

$$\max_{1\le i\le n}\left(\frac{\mathbb{P}\left\{U_{(i)}\le \frac{i}{n}(1-\delta)\right\}}{\exp\left(-i\psi(\delta)\right)}\right)\le 2.$$

Recalling that $U_{(i)}\overset{\mathscr{D}}{=}G_i/G_n$, it is easy to be convinced that since G_n is close to n, the tail behavior of $U_{(i)}$ can be uniformly bounded in n. This is what Theorem 13.1 captures.

Proof (Theorem 13.1). The proof is based on Theorem 20.6 in the Appendix (Chernoff's bounding method for the gamma distribution). For $\delta\in(0,1/2]$, we have

$$\begin{aligned}
\mathbb{P}\left\{U_{(i)}\ge\frac{i}{n}(1+\delta)\right\}&=\mathbb{P}\left\{\frac{G_i}{G_n}\ge\frac{i}{n}(1+\delta)\right\}\\
&\le\mathbb{P}\left\{G_i\ge i\left(1+\frac{\delta}{2}\right)\right\}+\mathbb{P}\left\{G_n\le n\frac{1+\frac{\delta}{2}}{1+\delta}\right\}\\
&\le\mathbb{P}\left\{G_i\ge i\left(1+\frac{\delta}{2}\right)\right\}+\mathbb{P}\left\{G_n\le n\left(1-\frac{\delta}{3}\right)\right\}\\
&\quad\text{(valid for } 0<\delta\le 1/2)\\
&\le\exp\left(-i\left(\frac{\delta}{2}-\log\left(1+\frac{\delta}{2}\right)\right)\right)\\
&\qquad+\exp\left(-n\left(-\frac{\delta}{3}-\log\left(1-\frac{\delta}{3}\right)\right)\right)\\
&\quad\text{(by Theorem 20.6)}\\
&\le 2\exp\left(-i\left(-\frac{\delta}{3}-\log\left(1-\frac{\delta}{3}\right)\right)\right)\\
&\quad\text{(since } \tfrac{\delta}{2}-\log(1+\tfrac{\delta}{2})\ge-\tfrac{\delta}{3}-\log(1-\tfrac{\delta}{3})\ge 0 \text{ on } (0,1/2]).
\end{aligned}$$

Finally, for $\delta \in (0,1)$,

$$\begin{aligned}
\mathbb{P}\left\{U_{(i)} \le \frac{i}{n}(1-\delta)\right\} &\le \mathbb{P}\left\{G_i \le i\left(1-\frac{\delta}{2}\right)\right\} + \mathbb{P}\left\{G_n \ge n\frac{1-\frac{\delta}{2}}{1-\delta}\right\} \\
&\le \mathbb{P}\left\{G_i \le i\left(1-\frac{\delta}{2}\right)\right\} + \mathbb{P}\left\{G_n \ge n\left(1+\frac{\delta}{2}\right)\right\} \\
&\le \exp\left(-i\left(-\frac{\delta}{2} - \log\left(1-\frac{\delta}{2}\right)\right)\right) \\
&\quad + \exp\left(-n\left(\frac{\delta}{2} - \log\left(1+\frac{\delta}{2}\right)\right)\right) \\
&\le 2\exp\left(-i\left(\frac{\delta}{2} - \log\left(1+\frac{\delta}{2}\right)\right)\right) \\
&\quad \text{(since } -\tfrac{\delta}{2} - \log(1-\tfrac{\delta}{2}) \ge \tfrac{\delta}{2} - \log(1+\tfrac{\delta}{2}) \ge 0 \text{ on } (0,1)). \quad \square
\end{aligned}$$

13.3 Sums of functions of uniform order statistics

For fixed $\alpha > 0$, the quantities that will matter most to us come in the general form

$$\frac{1}{k}\sum_{i=1}^{k} U_{(i)}^{\alpha} L_i,$$

where k depends upon n ($1 \le k \le n$), and $L_1, \ldots, L_k$ are i.i.d. random variables independent of $U_{(1)}, \ldots, U_{(n)}$. Three special cases are of interest:

(*i*) $L_1 \equiv 1$;
(*ii*) $\mathbb{E}L_1 = \mu > 0$;
(*iii*) $\mathbb{E}L_1 = 0, \mathbb{V}L_1 = \sigma^2 > 0$.

Clearly, (*i*) and (*ii*) can be handled together.

Theorem 13.2. *Let $k \to \infty$ as $n \to \infty$. If $\mathbb{E}L_1 = \mu < \infty$, then*

$$\mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k}\left(\frac{U_{(i)}}{k/n}\right)^{\alpha} L_i\right] \to \frac{\mu}{1+\alpha}.$$

Furthermore, if $\mathbb{V}L_1 < \infty$,

$$\frac{1}{k}\sum_{i=1}^{k}\left(\frac{U_{(i)}}{k/n}\right)^{\alpha} L_i \to \frac{\mu}{1+\alpha} \quad \textit{in probability}.$$

Proof. The first part follows if we can show that

$$\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\left[\left(\frac{U_{(i)}}{k/n}\right)^{\alpha}\right]\to\frac{1}{1+\alpha}.$$

Let $\ell_k\to\infty$ be a sequence with $1\le\ell_k<k$ and $\ell_k=\mathrm{o}(k)$. Then, by Lemma 13.1,

$$\begin{aligned}\frac{1}{k}\sum_{i=1}^{\ell_k}\mathbb{E}\left[\left(\frac{U_{(i)}}{k/n}\right)^{\alpha}\right]&\le(1+\alpha)^{1+\alpha}\frac{1}{k}\sum_{i=1}^{\ell_k}\frac{(i/n)^{\alpha}}{(k/n)^{\alpha}}\\&=\mathrm{O}\left(\left(\frac{\ell_k}{k}\right)^{1+\alpha}\right)=\mathrm{o}(1).\end{aligned}$$

Next,

$$\begin{aligned}\frac{1}{k}\sum_{i=\ell_k+1}^{k}\mathbb{E}\left[\left(\frac{U_{(i)}}{k/n}\right)^{\alpha}\right]&=\frac{1}{k}\sum_{i=\ell_k+1}^{k}\mathbb{E}\left[\left(\frac{U_{(i)}}{i/n}\right)^{\alpha}\right]\left(\frac{i}{k}\right)^{\alpha}\\&=\frac{1}{k}\sum_{i=\ell_k+1}^{k}\left(1+\mathrm{O}\left(\frac{1}{\ell_k}\right)\right)\left(\frac{i}{k}\right)^{\alpha}\\&\qquad\text{(where the ``O'' is uniform over } i \text{ and } n \text{ by Lemma 13.2)}\\&=\frac{1}{k}\sum_{i=\ell_k+1}^{k}\left(\frac{i}{k}\right)^{\alpha}(1+\mathrm{o}(1))\\&\sim\frac{1}{k^{1+\alpha}}\times\frac{k^{1+\alpha}-\ell_k^{1+\alpha}}{1+\alpha}\\&=\frac{1+\mathrm{o}(1)}{1+\alpha}.\end{aligned}$$

This proves the first assertion of the theorem. The second one follows by analyzing $\mathbf{I}+\mathbf{II}$, where

$$\mathbf{I}=\frac{1}{k}\sum_{i=1}^{k}\left(\frac{U_{(i)}}{k/n}\right)^{\alpha}(L_i-\mu)\quad\text{and}\quad\mathbf{II}=\frac{\mu}{k}\sum_{i=1}^{k}\left(\frac{U_{(i)}}{k/n}\right)^{\alpha}.$$

Let $\mathbf{U}=(U_{(1)},\dots,U_{(n)})$. Then, if $\sigma^2=\mathbb{V}L_1$,

$$\mathbb{E}[\mathbf{I}^2\,|\,\mathbf{U}]=\frac{\sigma^2}{k^2}\sum_{i=1}^{k}\left(\frac{U_{(i)}}{k/n}\right)^{2\alpha}.$$

By Lemma 13.1,

$$\mathbb{E}\mathbf{I}^2 \le (1+2\alpha)^{1+2\alpha}\frac{\sigma^2}{k^2}\sum_{i=1}^{k}\left(\frac{i/n}{k/n}\right)^{2\alpha} = \mathrm{O}\left(\frac{1}{k}\right),$$

so that by Chebyshev's inequality, $\mathbf{I} \to 0$ in probability. We conclude by showing that

$$\frac{1}{k}\sum_{i=1}^{k}\left(\frac{U_{(i)}}{k/n}\right)^{\alpha} \to \frac{1}{1+\alpha} \quad \text{in probability.}$$

Let $\ell_k = \lfloor \log^2 k \rfloor$. For $\delta > 0$ smaller than $1/2$, define

$$A = \bigcup_{i=\ell_k+1}^{k}\left[\left|\frac{U_{(i)}}{i/n} - 1\right| > \delta\right].$$

By Theorem 13.1,

$$\mathbb{P}\{A\} \le 4\sum_{i=\ell_k+1}^{k}\exp\big(-i\min\left(\varphi(\delta),\psi(\delta)\right)\big),$$

where φ and ψ are defined in Theorem 13.1. Clearly, for fixed $\delta > 0$, $\mathbb{P}\{A\} = \mathrm{o}(1/k^2)$. If A^c occurs, then

$$\frac{1}{k}\sum_{i=\ell_k+1}^{k}\left(\frac{i}{k}\right)^{\alpha}(1-\delta) \le \frac{1}{k}\sum_{i=\ell_k+1}^{k}\left(\frac{U_{(i)}}{k/n}\right)^{\alpha} \le \frac{1}{k}\sum_{i=\ell_k+1}^{k}\left(\frac{i}{k}\right)^{\alpha}(1+\delta).$$

Since

$$\frac{1}{k}\sum_{i=\ell_k+1}^{k}\left(\frac{i}{k}\right)^{\alpha} \to \frac{1}{1+\alpha}$$

and $\delta > 0$ is arbitrary, it suffices to show that

$$\frac{1}{k}\sum_{i=1}^{\ell_k}\left(\frac{U_{(i)}}{k/n}\right)^{\alpha} \to 0 \quad \text{in probability.}$$

The expected value of the left-hand side is not more than

$$(1+\alpha)^{1+\alpha}\frac{1}{k}\sum_{i=1}^{\ell_k}\left(\frac{i}{k}\right)^{\alpha} = \mathrm{O}\left(\left(\frac{\ell_k}{k}\right)^{1+\alpha}\right) = \mathrm{o}(1),$$

which is sufficient. □

Theorem 13.3. *Let $k \to \infty$ and $k/n \to 0$. If $\mathbb{E}L_1 = 0$ and $0 < \mathbb{V}L_1 = \sigma^2 < \infty$, then*

$$\frac{\sqrt{1+2\alpha}}{\sigma} \times \frac{1}{\sqrt{k}} \sum_{i=1}^{k} \left(\frac{U_{(i)}}{k/n} \right)^{\alpha} L_i \xrightarrow{\mathscr{D}} N,$$

where N is a standard normal random variable.

Proof. We first observe that

$$\frac{\sqrt{1+2\alpha}}{\sigma} \times \frac{1}{\sqrt{k}} \sum_{i=1}^{k} \left(\frac{i}{k} \right)^{\alpha} L_i \xrightarrow{\mathscr{D}} N.$$

This follows from a particular version of Lindeberg's central limit theorem (Lindeberg, 1920; see, e.g., Petrov, 1975), which states that if $X_1, \ldots, X_k$ are independent zero-mean random variables with finite variances, and $\sigma_k^2 \stackrel{\text{def}}{=} \sum_{i=1}^{k} \mathbb{V}X_i > 0$, then

$$\frac{1}{\sigma_k} \sum_{i=1}^{k} X_i \xrightarrow{\mathscr{D}} N$$

if, for every $\varepsilon > 0$,

$$\frac{1}{\sigma_k^2} \sum_{i=1}^{k} \mathbb{E}[X_i^2 \mathbb{1}_{[|X_i| > \varepsilon \sigma_k]}] \to 0 \quad \text{as } k \to \infty. \tag{13.2}$$

Replacing X_i by $i^{\alpha} L_i$, we have

$$\sigma_k^2 = \sigma^2 \sum_{i=1}^{k} i^{2\alpha} \sim \frac{\sigma^2 k^{1+2\alpha}}{1+2\alpha}.$$

Also,

$$\mathbb{E}\left[i^{2\alpha} L_i^2 \mathbb{1}_{[|i^{\alpha} L_i| > \varepsilon \sigma_k]} \right] \le k^{2\alpha} \mathbb{E}\left[L_1^2 \mathbb{1}_{[L_1^2 > \varepsilon \sigma_k / k^{\alpha}]} \right] = \mathrm{o}(k^{2\alpha})$$

since $\mathbb{E}L_1^2 = \sigma^2 < \infty$ and

$$\frac{\sigma_k}{k^{\alpha}} \sim \frac{\sigma\sqrt{k}}{\sqrt{1+2\alpha}} \to \infty.$$

Thus, Lindeberg's condition (13.2) is satisfied, and therefore,

$$\frac{1}{\sigma_k} \sum_{i=1}^{k} i^{\alpha} L_i \xrightarrow{\mathscr{D}} N,$$

which is equivalent to

$$\frac{\sqrt{1+2\alpha}}{\sigma} \times \frac{1}{\sqrt{k}} \sum_{i=1}^{k} \left(\frac{i}{k}\right)^{\alpha} L_i \xrightarrow{\mathscr{D}} N.$$

Next, we show that

$$W_n \stackrel{\text{def}}{=} \frac{1}{\sqrt{k}} \sum_{i=1}^{k} \left(\left(\frac{U_{(i)}}{k/n}\right)^{\alpha} - \left(\frac{i/n}{k/n}\right)^{\alpha} \right) L_i \to 0 \quad \text{in probability.}$$

This follows by proving that $\mathbb{E}W_n^2 = \mathrm{o}(1)$. Given $\mathbf{U} = (U_{(1)}, \ldots, U_{(n)})$, we have

$$\mathbb{E}[W_n^2 \,|\, \mathbf{U}] = \frac{\sigma^2}{k} \sum_{i=1}^{k} \left(\left(\frac{U_{(i)}}{k/n}\right)^{\alpha} - \left(\frac{i/n}{k/n}\right)^{\alpha} \right)^2,$$

so that

$$\begin{aligned}
\mathbb{E}W_n^2 &= \frac{\sigma^2}{k} \sum_{i=1}^{k} \frac{1}{(k/n)^{2\alpha}} \mathbb{E}\left[\left(U_{(i)}^{\alpha} - \left(\frac{i}{n}\right)^{\alpha} \right)^2 \right] \\
&= \frac{\sigma^2}{k} \sum_{i=1}^{k} \left(\frac{i}{k}\right)^{2\alpha} \mathbb{E}\left[\left(\left(\frac{U_{(i)}}{i/n}\right)^{\alpha} - 1 \right)^2 \right] \\
&= \frac{\sigma^2}{k} \sum_{i=1}^{k} \left(\frac{i}{k}\right)^{2\alpha} \left(\mathbb{E}\left[\left(\frac{U_{(i)}}{i/n}\right)^{2\alpha} \right] + 1 - 2\mathbb{E}\left[\left(\frac{U_{(i)}}{i/n}\right)^{\alpha} \right] \right) \\
&= \frac{\sigma^2}{k} \sum_{i=1}^{k} \left(\frac{i}{k}\right)^{2\alpha} \left(1 + \frac{\theta}{i} + 1 - 2\left(1 + \frac{\theta}{i}\right) \right),
\end{aligned}$$

by Lemma 13.2, where $\theta \in [-c, c]$ is a constant taking different values whenever used, and c is a universal constant. We conclude

$$\mathbb{E}W_n^2 \le \frac{\sigma^2}{k} \sum_{i=1}^{k} \left(\frac{i}{k}\right)^{2\alpha} \frac{3c}{i} = \mathrm{O}\left(\frac{1}{k}\right). \quad \square$$

Chapter 14
Rates of convergence

14.1 The finer behavior of the nearest neighbor regression function estimate

In this chapter, we study the local rate of convergence of $r_n(\mathbf{x})$ to $r(\mathbf{x})$. We obtain full information on the first asymptotic term of $r_n(\mathbf{x}) - r(\mathbf{x})$, and are rewarded with (*i*) a central limit theorem for $r_n(\mathbf{x}) - r(\mathbf{x})$, and (*ii*) a way of helping the user decide how to choose the weights v_{ni} of the estimate.

While it is true that the best sequence $(v_{n1}, \ldots, v_{nn})$ may depend upon $\mathbf{x}$, it is interesting that for sufficiently smooth problems (in a sense to be made precise in this chapter), there is a universally good way of picking the v_{ni}'s.

To simplify the notation, it is assumed throughout that $\mathbf{x} = \mathbf{0}$, without loss of generality. Moreover, to keep all unnecessary distractions from the reader, we study only the weak convergence properties of $r_n(\mathbf{0}) - r(\mathbf{0})$. We let the conditional variance of Y be

$$\sigma^2(\mathbf{x}) = \mathbb{E}\left[|Y - r(\mathbf{X})|^2 \, \middle| \, \mathbf{X} = \mathbf{x} \right],$$

and assume the following:

(*i*) There exists a sequence of positive integers $\{k\} = \{k_n\}$ with $k \to \infty$, $k/n \to 0$, and a positive constant c such that

$$|v_{ni}| \le \begin{cases} \frac{c}{k} & \text{for } 1 \le i \le k \\ 0 & \text{otherwise.} \end{cases}$$

It is stressed that the v_{ni}'s may have an arbitrary sign. However, as always, we assume that $\sum_{i=1}^n v_{ni} = 1$.

(*ii*) The random variable $\mathbf{X}$ has a density f on $\mathbb{R}^d$ that is twice continuously differentiable in a neighborhood of $\mathbf{0}$. Also, $f(\mathbf{0}) > 0$.

G. Biau, L. Devroye, *Lectures on the Nearest Neighbor Method*,
Springer Series in the Data Sciences, DOI 10.1007/978-3-319-25388-6_14

(*iii*) The regression function r is twice continuously differentiable in a neighborhood of $\mathbf{0}$.

(*iv*) One has $\|Y\|_\infty \leq 1$. This condition can be weakened to either $\|Y - r(\mathbf{X})\|_\infty \leq 1$ or even

$$\sup_{\mathbf{x} \in \mathbb{R}^d} \mathbb{E}\left[|Y - r(\mathbf{X})|^3 \mid \mathbf{X} = \mathbf{x} \right] < \infty.$$

(*v*) The function σ is continuous in a neighborhood of $\mathbf{0}$ and $\sigma^2(\mathbf{0}) > 0$.

The case of a continuous σ and $\sigma(\mathbf{0}) = 0$ requires additional scrutiny that has no place in these lecture notes.

14.2 The projection to the halfline

For pointwise analysis of the nearest neighbor estimate, it is essential to understand that from the vantage point of $\mathbf{0} \in \mathbb{R}^d$, the problem is one dimensional. To set this up, we define the one-dimensional quantity $Z = \|\mathbf{X}\|$, which has density g on $[0, \infty)$. Observe that

$$g(0) = \lim_{\rho \downarrow 0} \frac{\mathbb{P}\{Z \leq \rho\}}{\rho} = \lim_{\rho \downarrow 0} \frac{\mathbb{P}\{\|\mathbf{X}\| \leq \rho\}}{\rho}.$$

Therefore, since $\frac{\mathbb{P}\{\|\mathbf{X}\| \leq \rho\}}{V_d \rho^d} \to f(\mathbf{0})$ by continuity,

$$g(0) = \begin{cases} 2f(\mathbf{0}) & \text{for } d = 1 \\ 0 & \text{for } d > 1. \end{cases}$$

Let G denote the distribution function of Z. Then, as $z \downarrow 0$,

$$G(z) = \mathbb{P}\{\|\mathbf{X}\| \leq z\} = \int_{B(\mathbf{0},z)} f(\mathbf{y}) \mathrm{d}\mathbf{y} \sim f(\mathbf{0}) V_d z^d.$$

We can reorder $\mathbf{X}_1, \ldots, \mathbf{X}_n$ by increasing values of $\|\mathbf{X}_i\|$ to obtain

$$Z_{(1)} = \|\mathbf{X}_{(1)}\| \leq \cdots \leq Z_{(n)} = \|\mathbf{X}_{(n)}\|.$$

(For simplicity of notation, we drop the dependence upon the query point $\mathbf{0}$, and write $\mathbf{X}_{(i)}$, $Y_{(i)}$ and $Z_{(i)}$ instead of $\mathbf{X}_{(i)}(\mathbf{0})$, $Y_{(i)}(\mathbf{0})$ and $Z_{(i)}(\mathbf{0})$.) If $U_{(1)} \leq \cdots \leq U_{(n)}$ are uniform $[0, 1]$ order statistics, then we also have (see Chapter 1)

$$\left(Z_{(1)}, \ldots, Z_{(n)}\right) \stackrel{\mathscr{D}}{=} \left(G^{-1}(U_{(1)}), \ldots, G^{-1}(U_{(n)})\right),$$

where

$$G^{-1}(u) = \inf\{t \geq 0 : G(t) \geq u\}, \quad u \in [0, 1].$$

Since

$$G^{-1}(u) = \left(\frac{u}{V_d f(\mathbf{0})}\right)^{1/d} + \psi(u),$$

where $\psi(u) = \mathrm{o}(u^{1/d})$ as $u \downarrow 0$, it will be convenient to replace $Z_{(i)}$ by

$$Z_{(i)} \stackrel{\mathscr{D}}{=} \left(\frac{U_{(i)}}{V_d f(\mathbf{0})}\right)^{1/d} + \psi(U_{(i)}). \tag{14.1}$$

The second quantity we require is the projected regression function

$$m(z) \stackrel{\text{def}}{=} \mathbb{E}[Y \,\big|\, \|\mathbf{X}\| = z] = \mathbb{E}[r(\mathbf{X}) \,\big|\, \|\mathbf{X}\| = z].$$

Note that $m(0) = r(\mathbf{0})$. We will be using the variation/bias decomposition

$$Y_i - r(\mathbf{0}) = (Y_i - m(Z_i)) + (m(Z_i) - r(\mathbf{0}))$$

instead of the d-dimensional and more standard decomposition

$$Y_i - r(\mathbf{0}) = (Y_i - r(\mathbf{X}_i)) + (r(\mathbf{X}_i) - r(\mathbf{0}))\,.$$

These are different! In fact, we have

$$Y_i - r(\mathbf{0}) = (Y_i - r(\mathbf{X}_i)) + (r(\mathbf{X}_i) - m(Z_i)) + (m(Z_i) - r(\mathbf{0}))\,.$$

The new middle term, $r(\mathbf{X}_i) - m(Z_i)$, acts as some sort of noise in the projected decomposition.

For our study, what matters is the local behavior of m, which can be obtained via analytical methods. In the following analysis, we place ourselves in a small ball $B(\mathbf{0}, \varepsilon)$ around $\mathbf{0}$ in which both f and r are twice continuously differentiable, and assume furthermore that $f(\mathbf{0}) > 0$. We observe that

$$m(z) = \lim_{\delta \downarrow 0} \frac{\mathbb{E}[r(\mathbf{X}) \mathbb{1}_{[z \leq \|\mathbf{X}\| \leq z+\delta]}]}{\mathbb{P}\{z \leq \|\mathbf{X}\| \leq z + \delta\}}. \tag{14.2}$$

To compute this, we make use of the Taylor series expansions (Giaquinta and Modica, 2009) of f and r about $\mathbf{0}$. For $\mathbf{x} = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$, we let

$$f'(\mathbf{0}) = \left(\frac{\partial f}{\partial x_1}(\mathbf{0}), \ldots, \frac{\partial f}{\partial x_d}(\mathbf{0})\right)^\top \quad \text{and} \quad f''(\mathbf{0}) = \left(\frac{\partial^2 f}{\partial x_j \partial x_{j'}}(\mathbf{0})\right)_{1 \leq j, j' \leq d},$$

where $\top$ denotes transposition and vectors are in column format. Similarly,

$$r'(\mathbf{0}) = \left(\frac{\partial r}{\partial x_1}(\mathbf{0}), \ldots, \frac{\partial r}{\partial x_d}(\mathbf{0})\right)^\top \quad \text{and} \quad r''(\mathbf{0}) = \left(\frac{\partial^2 r}{\partial x_j \partial x_{j'}}(\mathbf{0})\right)_{1 \le j,j' \le d}.$$

The symbol $\mathrm{tr}(\Delta)$ stands for the trace of the square matrix Δ, and λ denotes the Lebesgue measure on $\mathbb{R}^d$. We show the following proposition.

Proposition 14.1. *Assume that f and r are twice continuously differentiable in a neighborhood of $\mathbf{0}$, and $f(\mathbf{0}) > 0$. Then, as $z \downarrow 0$,*

$$m(z) = r(\mathbf{0}) + \alpha z^2 + \mathrm{o}(z^2),$$

where

$$\alpha \stackrel{\text{def}}{=} \frac{f(\mathbf{0})\mathrm{tr}(r''(\mathbf{0})) + 2r'(\mathbf{0})^\top f'(\mathbf{0})}{2df(\mathbf{0})}.$$

The following two lemmas are useful for proving the result.

Lemma 14.1. *For the ball $B(\mathbf{0}, \rho) \subseteq \mathbb{R}^d$ and a $d \times d$ matrix A, we have*

$$\int_{B(\mathbf{0},\rho)} \mathbf{x}^\top A\mathbf{x}\, d\mathbf{x} = \frac{\rho^2 \lambda\left(B(\mathbf{0}, \rho)\right)}{d+2} \times \mathrm{tr}(A).$$

Proof. Note that if $\mathbf{Y}$ is uniform in $B(\mathbf{0}, \rho)$, then

$$\mathbf{Y} \stackrel{\mathscr{D}}{=} \rho U^{1/d} \mathbf{Z},$$

where U is uniform on $[0, 1]$, and $\mathbf{Z}$ is uniform on the unit surface and independent of U. Thus,

$$\begin{aligned}\frac{1}{\lambda\left(B(\mathbf{0}, \rho)\right)} \int_{B(\mathbf{0},\rho)} \mathbf{x}^\top A\mathbf{x}\, d\mathbf{x} &= \rho^2 \mathbb{E}U^{2/d} \times \mathbb{E}[\mathbf{Z}^\top A\mathbf{Z}] \\ &= \frac{\rho^2}{1 + \frac{2}{d}} \times \mathbb{E}\left[\sum_{j,j'} Z_j Z_{j'} A_{jj'}\right],\end{aligned}$$

where $A_{jj'}$ is the (j, j')-th element of A and Z_j is the j-th coordinate of $\mathbf{Z}$. Therefore,

$$\frac{1}{\lambda\left(B(\mathbf{0}, \rho)\right)} \int_{B(\mathbf{0},\rho)} \mathbf{x}^\top A\mathbf{x}\, d\mathbf{x} = \frac{\rho^2 d}{d+2} \sum_{j=1}^{d} A_{jj} \mathbb{E}Z_j^2 = \frac{\rho^2}{d+2} \times \mathrm{tr}(A). \quad \square$$

Lemma 14.2. *For $z, \delta > 0$, let $B_{z,\delta}(\mathbf{0}) \stackrel{\text{def}}{=} B(\mathbf{0}, z+\delta) - B(\mathbf{0}, z)$. Then, for fixed $z > 0$,*

$$\frac{1}{\lambda\left(B_{z,\delta}(\mathbf{0})\right)} \int_{B_{z,\delta}(\mathbf{0})} \mathbf{x}^\top A \mathbf{x} \, \mathrm{d}\mathbf{x} \to \frac{z^2}{d} \times \mathrm{tr}(A) \quad \text{as } \delta \downarrow 0.$$

Proof. We have

$$\lambda\left(B_{z,\delta}(\mathbf{0})\right) = \left((z+\delta)^d - z^d\right) V_d = d\delta z^{d-1} V_d + \mathrm{O}(\delta^2).$$

Also, by Lemma 14.1,

$$\begin{aligned} \int_{B_{z,\delta}(\mathbf{0})} \mathbf{x}^\top A \mathbf{x} \, \mathrm{d}\mathbf{x} &= \frac{(z+\delta)^{d+2} - z^{d+2}}{d+2} \times V_d \times \mathrm{tr}(A) \\ &= \delta z^{d+1} V_d \times \mathrm{tr}(A) + \mathrm{O}(\delta^2), \end{aligned}$$

and the limit follows. □

Proof (Proposition 14.1). We use $\varphi(\mathbf{x})$ and $\psi(\mathbf{x})$ to denote functions that are $\mathrm{o}(\|\mathbf{x}\|^2)$ as $\|\mathbf{x}\| \downarrow 0$. By Lemma 14.2, if $B_{z,\delta}(\mathbf{0}) \stackrel{\text{def}}{=} B(\mathbf{0}, z+\delta) - B(\mathbf{0}, z)$, then for fixed z, as $\delta \downarrow 0$,

$$\begin{aligned} &\mathbb{P}\{z \le \|\mathbf{X}\| \le z + \delta\} \\ &\quad = \int_{B_{z,\delta}(\mathbf{0})} \left(f(\mathbf{0}) + f'(\mathbf{0})^\top \mathbf{x} + \frac{1}{2} \mathbf{x}^\top f''(\mathbf{0}) \mathbf{x} + \varphi(\mathbf{x}) \right) \mathrm{d}\mathbf{x} \\ &\quad = \lambda\left(B_{z,\delta}(\mathbf{0})\right) \left[f(\mathbf{0}) + \frac{z^2}{2d} \mathrm{tr}(f''(\mathbf{0}))\,(1 + \mathrm{o}(1)) \right] + \int_{B_{z,\delta}(\mathbf{0})} \varphi(\mathbf{x}) \mathrm{d}\mathbf{x}, \end{aligned}$$

and

$$\begin{aligned} &\mathbb{E}[r(\mathbf{X}) \mathbb{1}_{[\mathbf{X} \in B_{z,\delta}(\mathbf{0})]}] \\ &\quad = \int_{B_{z,\delta}(\mathbf{0})} \left(r(\mathbf{0}) + r'(\mathbf{0})^\top \mathbf{x} + \frac{1}{2} \mathbf{x}^\top r''(\mathbf{0}) \mathbf{x} + \psi(\mathbf{x}) \right) \\ &\qquad \times \left(f(\mathbf{0}) + f'(\mathbf{0})^\top \mathbf{x} + \frac{1}{2} \mathbf{x}^\top f''(\mathbf{0}) \mathbf{x} + \varphi(\mathbf{x}) \right) \mathrm{d}\mathbf{x} \\ &\quad = \lambda\left(B_{z,\delta}(\mathbf{0})\right) \left[r(\mathbf{0}) f(\mathbf{0}) + \frac{z^2}{2d} \mathrm{tr}\left(f(\mathbf{0}) r''(\mathbf{0}) + r(\mathbf{0}) f''(\mathbf{0}) + 2 r'(\mathbf{0}) f'(\mathbf{0})^\top \right) \right. \\ &\qquad \left. \times (1 + \mathrm{o}(1)) \right] + \int_{B_{z,\delta}(\mathbf{0})} \left(\psi(\mathbf{x}) f(\mathbf{x}) + \varphi(\mathbf{x}) r(\mathbf{x}) \right) \mathrm{d}\mathbf{x}. \end{aligned}$$

Consider the ratio in (14.2) and let $\delta \downarrow 0$ to obtain

$$m(z) = \frac{r(\mathbf{0})f(\mathbf{0}) + \frac{z^2}{2d}\mathrm{tr}\big(f(\mathbf{0})r''(\mathbf{0}) + r(\mathbf{0})f''(\mathbf{0}) + 2r'(\mathbf{0})f'(\mathbf{0})^\top\big) + \mathrm{o}(z^2)}{f(\mathbf{0}) + \frac{z^2}{2d}\mathrm{tr}(f''(\mathbf{0})) + \mathrm{o}(z^2)}$$

$$= r(\mathbf{0}) + \frac{z^2}{2df(\mathbf{0})}\Big[f(\mathbf{0})\mathrm{tr}(r''(\mathbf{0})) + 2\mathrm{tr}\big(r'(\mathbf{0})f'(\mathbf{0})^\top\big)\Big] + \mathrm{o}(z^2),$$

valid as $z \downarrow 0$. The conclusion follows by observing that

$$\mathrm{tr}\big(r'(\mathbf{0})f'(\mathbf{0})^\top\big) = r'(\mathbf{0})^\top f'(\mathbf{0}). \quad \square$$

Remark 14.1. We note that $f''(\mathbf{0})$ is absent in the Taylor series expansion of m. In fact, the expansion remains valid if f is just assumed to be continuously differentiable. That refinement is implicit in the analysis by coupling methods given in the next chapter. $\square$

Our study is split into two parts, that of the variation

$$V_n \stackrel{\text{def}}{=} \sum_{i=1}^{n} v_{ni}\left(Y_{(i)} - m(Z_{(i)})\right),$$

and that of the bias

$$B_n \stackrel{\text{def}}{=} \sum_{i=1}^{n} v_{ni}\left(m(Z_{(i)}) - m(0)\right),$$

where we note that $r_n(\mathbf{0}) - r(\mathbf{0}) = V_n + B_n$. Here, as before,

$$(\mathbf{X}_{(1)}, Y_{(1)}, Z_{(1)}), \ldots, (\mathbf{X}_{(n)}, Y_{(n)}, Z_{(n)})$$

is a reordering of $(\mathbf{X}_1, Y_1, Z_1), \ldots, (\mathbf{X}_n, Y_n, Z_n)$ according to increasing values of $Z_i = \|\mathbf{X}_i\|$.

14.3 Study of the bias

Let $U_{(1)} \le \cdots \le U_{(n)}$ be the order statistics corresponding to a uniform $[0,1]$ sample. The following quantity plays a central role in the study of the bias of the nearest neighbor regression function estimate, and thus deserves a special treatment:

$$W_n = \sum_{i=1}^{n} v_{ni} U_{(i)}^{2/d}.$$

Recall that conditions (i)-(v) are defined in Section 14.1.

Proposition 14.2. *Assume that condition* (i) *is satisfied. Then*

$$W_n = \left(\frac{k}{n}\right)^{2/d} \left(\sum_{i=1}^{k} v_{ni} \left(\frac{i}{k}\right)^{2/d}\right) (1 + o_{\mathbb{P}}(1)) + o_{\mathbb{P}}\left(\left(\frac{k}{n}\right)^{2/d}\right).$$

Remark 14.2. There is in general no guarantee that the sequence $\sum_{i=1}^{k} v_{ni}(i/k)^{2/d}$ has a limit as $n \to \infty$. Note however that, in all cases, $\sum_{i=1}^{k} v_{ni}(i/k)^{2/d} \leq 1$. Also, for nonnegative weights,

$$0 < \frac{1}{2(2c)^{2/d}} \leq \liminf_{n\to\infty} \left(\sum_{i=1}^{k} v_{ni} \left(\frac{i}{k}\right)^{2/d}\right) \leq \limsup_{n\to\infty} \left(\sum_{i=1}^{k} v_{ni} \left(\frac{i}{k}\right)^{2/d}\right) \leq 1.$$

Thus, for nonnegative weights, we can replace the result of Proposition 14.2 by

$$\frac{W_n}{(k/n)^{2/d} \sum_{i=1}^{k} v_{ni}(i/k)^{2/d}} \to 1 \quad \text{in probability.} \quad \square$$

An example of sequence that satisfies condition (i) and has $\sum_{i=1}^{k} v_{ni} i^{2/d} = 0$ is as follows. Let $\{k\} = \{k_n\}$ be a sequence of multiples of 4, with $k \in \{1, \ldots, n\}$, $k \to \infty$, and $k/n \to 0$. Set

$$v_{ni} = \begin{cases} 0 & \text{for } 1 \leq i \leq \frac{k}{2} \\ \frac{c_k}{i^{2/d}} \times \frac{k^{2/d}}{k} & \text{for } \frac{k}{2} < i \leq \frac{3k}{4} \\ -\frac{c_k}{i^{2/d}} \times \frac{k^{2/d}}{k} & \text{for } \frac{3k}{4} < i \leq k \\ 0 & \text{for } k < i \leq n, \end{cases}$$

where c_k is a constant depending upon k. Obviously, $\sum_{i=1}^{k} v_{ni} i^{2/d} = 0$. Also, if we adjust c_k to ensure $\sum_{i=1}^{n} v_{ni} = 1$, we obtain for $d > 2$,

$$\lim_{k\to\infty} c_k = \frac{\frac{d-2}{d}}{2\left(\frac{3}{4}\right)^{1-\frac{2}{d}} - \left(\frac{1}{2}\right)^{1-\frac{2}{d}} - 1},$$

so that (i) holds. For $d = 2$, we find $\lim_{k\to\infty} c_k = \frac{1}{\log(9/8)}$, and for $d = 1$, $\lim_{k\to\infty} c_k = 3$.

Proof (Proposition 14.2). From Corollary 1.1, we recall that

$$\left(U_{(1)}, \ldots, U_{(n)}\right) \stackrel{\mathscr{D}}{=} \left(\frac{G_1}{G_{n+1}}, \ldots, \frac{G_n}{G_{n+1}}\right),$$

where $G_i = \sum_{j=1}^{i} E_j$, $1 \leq i \leq n+1$, and $E_1, \ldots, E_{n+1}$ are independent standard exponential random variables. Therefore, we rewrite W_n in the form

$$W_n \stackrel{\mathscr{D}}{=} \sum_{i=1}^{k} v_{ni} \left(\frac{G_i}{G_{n+1}}\right)^{2/d} = \sum_{i=1}^{k} v_{ni} \left(\frac{G_i}{k}\right)^{2/d} \left(\frac{k}{n}\right)^{2/d} \left(\frac{n}{G_{n+1}}\right)^{2/d}.$$

By the law of large numbers, $G_{n+1}/n \to 1$ in probability. Thus,

$$\left(\frac{n}{k}\right)^{2/d} W_n = \sum_{i=1}^{k} v_{ni} \left(\frac{G_i}{k}\right)^{2/d} (1 + o_{\mathbb{P}}(1)) .$$

Note that, for $\varepsilon > 0$,

$$\begin{aligned}
&\left| \sum_{i=1}^{k} v_{ni} \left(\frac{G_i}{k}\right)^{2/d} - \sum_{i=1}^{k} v_{ni} \left(\frac{i}{k}\right)^{2/d} \right| \\
&\quad \le \frac{c}{k} \sum_{i=1}^{k} \left| \left(\frac{G_i}{k}\right)^{2/d} - \left(\frac{i}{k}\right)^{2/d} \right| \\
&\quad \le c\varepsilon + \frac{c}{k} \sum_{i=1}^{k} \left| \left(\frac{G_i}{k}\right)^{2/d} - \left(\frac{i}{k}\right)^{2/d} \right| \mathbb{1}_{[|(G_i/k)^{2/d} - (i/k)^{2/d}| > \varepsilon]} \\
&\quad \le c\varepsilon + \infty . \mathbb{1}_{[G_k > 2k]} + \frac{c 2^{2/d}}{k} \sum_{i=1}^{k} \mathbb{1}_{[|(G_i/k)^{2/d} - (i/k)^{2/d}| > \varepsilon]} .
\end{aligned}$$

Choose ε so small that the first term is small. Observe that the second term vanishes if $G_k \le 2k$, an event that happens with probability going to one. Finally, the last term is $o_{\mathbb{P}}(1)$ as $k \to \infty$. To see this, just note that

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{k} \sum_{i=1}^{k} \mathbb{1}_{[|(G_i/k)^{2/d} - (i/k)^{2/d}| > \varepsilon]} \right] &= \frac{1}{k} \sum_{i=1}^{k} \mathbb{P}\left\{ \left| \left(\frac{G_i}{k}\right)^{2/d} - \left(\frac{i}{k}\right)^{2/d} \right| > \varepsilon \right\} \\
&\le \frac{1}{k} \sum_{i=1}^{k} \mathbb{P}\left\{ \left| \left(\frac{G_i}{i}\right)^{2/d} - 1 \right| > \varepsilon \right\} .
\end{aligned}$$

This is a Cesàro mean, which tends to zero since $G_n/n \to 1$ in probability as $n \to \infty$.

The proposition follows from what we just showed, i.e., that

$$\sum_{i=1}^{k} v_{ni} \left(\frac{G_i}{k}\right)^{2/d} - \sum_{i=1}^{k} v_{ni} \left(\frac{i}{k}\right)^{2/d} \to 0 \quad \text{in probability.} \quad \square$$

We are now ready for the bias term B_n.

Theorem 14.1. *Assume that conditions* (*i*), (*ii*), *and* (*iii*) *are satisfied. Then*

$$B_n = \beta \left(\frac{k}{n}\right)^{2/d} \left(\sum_{i=1}^{k} v_{ni} \left(\frac{i}{k}\right)^{2/d} \right) (1 + o_{\mathbb{P}}(1)) + o_{\mathbb{P}}\left(\left(\frac{k}{n}\right)^{2/d} \right),$$

where

$$\beta \stackrel{\text{def}}{=} \frac{f(\mathbf{0})\mathrm{tr}(r''(\mathbf{0})) + 2r'(\mathbf{0})^\top f'(\mathbf{0})}{2dV_d^{2/d} f^{1+2/d}(\mathbf{0})}.$$

Proof. By Proposition 14.1, where α is defined, we have

$$\begin{aligned} B_n = \sum_{i=1}^n v_{ni}\left(m(Z_{(i)}) - m(\mathbf{0})\right) &= \alpha \sum_{i=1}^n v_{ni} Z_{(i)}^2 + \sum_{i=1}^n v_{ni}\varphi(Z_{(i)}) \\ &\quad (\text{where } \varphi(z) = \mathrm{o}(z^2) \text{ as } z \downarrow 0) \\ &\stackrel{\text{def}}{=} \mathbf{I} + \mathbf{II}. \end{aligned}$$

Clearly,

$$\begin{aligned} |\mathbf{II}| \le \sum_{i=1}^n |v_{ni}| \sup_{0<z\le Z_{(k)}} |\varphi(z)| &\le \sum_{i=1}^n |v_{ni}| Z_{(k)}^2 \sup_{0<z\le Z_{(k)}} \left|\frac{\varphi(z)}{z^2}\right| \\ &\le c\, Z_{(k)}^2 \sup_{0<z\le Z_{(k)}} \left|\frac{\varphi(z)}{z^2}\right| \\ &= \mathrm{o}_{\mathbb{P}}(Z_{(k)}^2) \end{aligned}$$

since $Z_{(k)} \to 0$ in probability (by Lemma 2.2 and the fact that $\mathbf{0}$ belongs to the support of $\mathbf{X}$—see condition (*ii*)).

Next, recall the decomposition (14.1)

$$Z_{(i)} \stackrel{\mathscr{D}}{=} \left(\frac{U_{(i)}}{V_d f(\mathbf{0})}\right)^{1/d} + \psi(U_{(i)}),$$

where $\psi(u) = \mathrm{o}(u^{1/d})$ as $u \downarrow 0$. Thus,

$$\mathbf{I} \stackrel{\mathscr{D}}{=} \beta \sum_{i=1}^n v_{ni} U_{(i)}^{2/d} + 2\alpha \sum_{i=1}^n v_{ni} \left(\frac{U_{(i)}}{V_d f(\mathbf{0})}\right)^{1/d} \psi(U_{(i)}) + \sum_{i=1}^n v_{ni}\psi^2(U_{(i)}),$$

where $\beta = \frac{\alpha}{V_d^{2/d} f^{2/d}(\mathbf{0})}$. Using the fact that $U_{(k)} \to 0$ in probability and $|v_{ni}| \le c/k$ for $1 \le i \le k$, it is easy to see that

$$\mathbf{I} \stackrel{\mathscr{D}}{=} \beta \sum_{i=1}^n v_{ni} U_{(i)}^{2/d} + \mathrm{o}_{\mathbb{P}}(U_{(k)}^{2/d}) = \beta \sum_{i=1}^n v_{ni} U_{(i)}^{2/d} + \mathrm{o}_{\mathbb{P}}\left(\left(\frac{k}{n}\right)^{2/d}\right),$$

by the well-known fact (Theorem 1.4) that $U_{(k)} = \mathrm{O}_{\mathbb{P}}(k/n)$. Combining this result with Proposition 14.2 proves the theorem. □

14.4 Study of the variation

The conditional variance of Y is defined by

$$\sigma^2(\mathbf{x}) = \mathbb{E}\left[|Y - r(\mathbf{X})|^2 \mid \mathbf{X} = \mathbf{x}\right],$$

but in our analysis, after the projection to the halfline, we must work with

$$\tau^2(z) = \mathbb{E}\left[|Y - m(z)|^2 \mid \|\mathbf{X}\| = z\right], \quad z \geq 0.$$

The relationship between these quantities is

$$\tau^2(z) = \mathbb{E}\left[|r(\mathbf{X}) - m(z)|^2 + \sigma^2(\mathbf{X}) \mid \|\mathbf{X}\| = z\right].$$

In any case, if $\|Y\|_\infty \leq 1$ (condition (iv)), then r and m are similarly bounded, and thus, $\sigma^2 \leq 4$ and $\tau^2 \leq 4$ μ-almost surely.

Conditional on $Z_1, \ldots, Z_n$, the variation term

$$V_n = \sum_{i=1}^{n} v_{ni} \left(Y_{(i)} - m(Z_{(i)})\right)$$

is a weighted sum of independent zero-mean random variables (Proposition 8.1) bounded by 2 in absolute value with probability one. Let Σ_i be the rank of Z_i in the reordering of the data according to increasing values of $Z_i = \|\mathbf{X}_i\|$. So, Z_i receives weight $v_{n\Sigma_i}$ and

$$V_n = \sum_{i=1}^{n} v_{n\Sigma_i} \left(Y_i - m(Z_i)\right).$$

The conditional variance of V_n is given by

$$\begin{aligned}
\mathbb{V}[V_n \mid Z_1, \ldots, Z_n] &= \sum_{i=1}^{n} v_{n\Sigma_i}^2 \mathbb{E}\left[|Y_i - m(Z_i)|^2 \mid Z_1, \ldots, Z_n\right] \\
&= \sum_{i=1}^{n} v_{n\Sigma_i}^2 \mathbb{E}\left[|Y_i - m(Z_i)|^2 \mid Z_i\right] \\
&\qquad \text{(by independence between } (Z_i, Y_i) \text{ and } Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n) \\
&= \sum_{i=1}^{n} v_{n\Sigma_i}^2 \tau^2(Z_i) \\
&= \sum_{i=1}^{n} v_{ni}^2 \tau^2(Z_{(i)}).
\end{aligned}$$

Since $\tau^2 \le 4$, we see that, with probability one,

$$\mathbb{E}[V_n^2 \,|\, Z_1, \ldots, Z_n] \le 4 \sum_{i=1}^{n} v_{ni}^2 \le \frac{4c^2}{k}.$$

This can be used as a means of getting rough upper bounds on the rate of convergence to zero of $r_n(\mathbf{0}) - r(\mathbf{0})$. However, $\mathbb{E}V_n^2$ may be much smaller than $O(1/k)$—this all depends upon the behavior of τ^2, and thus σ^2, near $\mathbf{0}$. Note that $\sigma(\mathbf{0}) = \tau(0)$ and that τ too is continuous in a neighborhood of 0 under condition (v). Lemma 11.1 shows that

$$\frac{\mathbb{V}[V_n \,|\, Z_1, \ldots, Z_n]}{\sigma^2(\mathbf{0}) \sum_{i=1}^{n} v_{ni}^2} = \frac{\sum_{i=1}^{n} v_{ni}^2 \tau^2(Z_{(i)})}{\tau^2(0) \sum_{i=1}^{n} v_{ni}^2} \to 1 \quad \text{in probability.}$$

However, V_n satisfies a central limit result, given by the following theorem:

Theorem 14.2. *Assume that conditions (i), (iv), and (v) are satisfied. Then*

$$\frac{V_n}{\sigma(\mathbf{0}) \sqrt{\sum_{i=1}^{n} v_{ni}^2}} \xrightarrow{\mathscr{D}} N,$$

where N is a standard normal random variable.

Thus, evoking Lemma 20.1 in the Appendix, we may write

$$V_n \stackrel{\mathscr{D}}{=} \sigma(\mathbf{0}) \sqrt{\sum_{i=1}^{n} v_{ni}^2}\, (N + o_{\mathbb{P}}(1)).$$

The asymptotic normality of the nearest neighbor regression function estimate has first been established by Royall (1966). Later, Mack (1981) (see also Lai, 1977) derived the rate of convergence for the bias and variance, as well as the asymptotic normality, for estimates of the form

$$s_n(\mathbf{x}) = \frac{\sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{X}_i}{\|\mathbf{X}_{(k)}(\mathbf{x})-\mathbf{x}\|}\right) Y_i}{\sum_{j=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{X}_j}{\|\mathbf{X}_{(k)}(\mathbf{x})-\mathbf{x}\|}\right)},$$

where K is a bounded density (kernel) satisfying $K(\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathbb{R}^d$ with $\|\mathbf{x}\| > 1$. This class of estimates incorporates the features of both the kernel and the k-nearest neighbor methods (see also Stute, 1984, who studies the asymptotic normality of a smoothed nearest neighbor-type estimate).

Proof (Theorem 14.2). It is useful to recall the Berry-Esseen inequality (see Berry, 1941, Esseen, 1942, or the book by Petrov, 1975) for sums of independent random variables $W_1, \ldots, W_n$ such that $\mathbb{E}W_i = 0$, $\sum_{i=1}^n \mathbb{E}W_i^2 > 0$, and $\mathbb{E}|W_i|^3 < \infty$:

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}\left\{ \frac{\sum_{i=1}^n W_i}{\sqrt{\sum_{i=1}^n \mathbb{E}W_i^2}} \le t \right\} - \mathbb{P}\{N \le t\} \right| \le \frac{\gamma \sum_{i=1}^n \mathbb{E}|W_i|^3}{\left(\sum_{i=1}^n \mathbb{E}W_i^2\right)^{3/2}}, \tag{14.3}$$

for some universal constant $\gamma > 0$.

We apply this inequality with the formal replacement

$$W_i \stackrel{\text{def}}{=} v_{ni}\left(Y_{(i)} - m(Z_{(i)})\right),$$

conditional on $Z_1, \ldots, Z_n$. Since, conditional on $Z_1, \ldots, Z_n$,

$$\mathbb{E}W_i^2 = v_{ni}^2 \tau^2(Z_{(i)}) \quad \text{and} \quad \mathbb{E}|W_i|^3 \le \frac{8c}{k} v_{ni}^2,$$

the bound in (14.3) becomes

$$\begin{aligned} \frac{8c\gamma \sum_{i=1}^n v_{ni}^2}{k\left(\sum_{i=1}^n v_{ni}^2 \tau^2(Z_{(i)})\right)^{3/2}} &\le \frac{8c\gamma}{k\sqrt{\sum_{i=1}^n v_{ni}^2} \times \min^{3/2}\left(\tau^2(Z_{(1)}), \ldots, \tau^2(Z_{(k)})\right)} \\ &\le \frac{8c\gamma}{k^{1/2}\min^{3/2}\left(\tau^2(Z_{(1)}), \ldots, \tau^2(Z_{(k)})\right)} \\ &\qquad \text{(since } \textstyle\sum_{i=1}^n v_{ni}^2 \ge \frac{1}{k}\text{, by the Cauchy-Schwarz inequality).} \end{aligned}$$

Observe that

$$\begin{aligned} \frac{\sum_{i=1}^n v_{ni}\left(Y_{(i)} - m(Z_{(i)})\right)}{\tau(0)\sqrt{\sum_{i=1}^n v_{ni}^2}} &= \frac{\sum_{i=1}^n v_{ni}\left(Y_{(i)} - m(Z_{(i)})\right)}{\sqrt{\sum_{i=1}^n v_{ni}^2 \tau^2(Z_{(i)})}} \times \frac{\sqrt{\sum_{i=1}^n v_{ni}^2 \tau^2(Z_{(i)})}}{\tau(0)\sqrt{\sum_{i=1}^n v_{ni}^2}} \\ &\stackrel{\text{def}}{=} \mathbf{I} \times \mathbf{II}. \end{aligned}$$

Now $\mathbf{II} \to 1$ in probability as noted earlier. For $\mathbf{I}$, we have

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}\{\mathbf{I} \le t \,|\, Z_1, \ldots, Z_n\} - \mathbb{P}\{N \le t\}\right| = \frac{\mathrm{O}(1/\sqrt{k})}{\min^{3/2}\left(\tau^2(Z_{(1)}), \ldots, \tau^2(Z_{(k)})\right)}.$$

Hence,

$$\begin{aligned}
&\sup_{t\in\mathbb{R}} \left|\mathbb{P}\{\mathbf{I} \le t\} - \mathbb{P}\{N \le t\}\right| \\
&\quad = \sup_{t\in\mathbb{R}} \left|\mathbb{E}\mathbb{P}\{\mathbf{I} \le t \,|\, Z_1, \ldots, Z_n\} - \mathbb{P}\{N \le t\}\right| \\
&\quad \le \frac{\mathrm{O}(1/\sqrt{k})}{\tau^3(0)} + \mathbb{P}\left\{\min\left(\tau^2(Z_{(1)}), \ldots, \tau^2(Z_{(k)})\right) < \frac{\tau^2(0)}{2}\right\}.
\end{aligned}$$

The latter probability tends to zero since $\tau(0) > 0$, τ is continuous at 0, and $Z_{(k)} \to 0$ in probability. Thus, $\mathbf{I} \xrightarrow{\mathscr{D}} N$, so that $\mathbf{I} \times \mathbf{II} \xrightarrow{\mathscr{D}} N$. □

14.5 Combining all results

Merging Theorem 14.1 and Theorem 14.2 leads to the main result of the chapter.

Theorem 14.3 (Pointwise rate of convergence). *Assume that conditions* (i)-(v) *are satisfied. Then the corresponding nearest neighbor regression function estimate* r_n *satisfies*

$$\begin{aligned}
r_n(\mathbf{0}) - r(\mathbf{0}) &\stackrel{\mathscr{D}}{=} \sigma(\mathbf{0}) \sqrt{\sum_{i=1}^{n} v_{ni}^2}\,(N + \mathrm{o}_{\mathbb{P}}(1)) \\
&+ \beta \left(\frac{k}{n}\right)^{2/d} \left(\sum_{i=1}^{k} v_{ni} \left(\frac{i}{k}\right)^{2/d}\right)(1 + \mathrm{o}_{\mathbb{P}}(1)) + \mathrm{o}_{\mathbb{P}}\left(\left(\frac{k}{n}\right)^{2/d}\right), \qquad (14.4)
\end{aligned}$$

where N *is a standard normal random variable and*

$$\beta \stackrel{\text{def}}{=} \frac{f(\mathbf{0})\mathrm{tr}(r''(\mathbf{0})) + 2r'(\mathbf{0})^\top f'(\mathbf{0})}{2dV_d^{2/d} f^{1+2/d}(\mathbf{0})}.$$

For the standard k-nearest neighbor estimate, one has

$$v_{ni} = \begin{cases} \frac{1}{k} & \text{for } 1 \le i \le k \\ 0 & \text{for } k < i \le n, \end{cases}$$

where $\{k\} = \{k_n\}$ is a sequence of integers such that $1 \le k \le n$. In this case,

$$\sum_{i=1}^{n} v_{ni}^2 = \frac{1}{k} \quad \text{and} \quad \sum_{i=1}^{k} v_{ni} \left(\frac{i}{k}\right)^{2/d} = \frac{d}{d+2}\,(1 + \mathrm{o}(1)).$$

Corollary 14.1. *Assume that conditions (ii)-(v) are satisfied. If $k \to \infty$ and $k/n \to 0$, then the k-nearest neighbor regression function estimate r_n satisfies*

$$r_n(\mathbf{0}) - r(\mathbf{0}) \stackrel{\mathscr{D}}{=} \frac{\sigma(\mathbf{0})}{\sqrt{k}} N + \xi \left(\frac{k}{n}\right)^{2/d} + o_{\mathbb{P}}\left(\frac{1}{\sqrt{k}} + \left(\frac{k}{n}\right)^{2/d}\right),$$

where N is a standard normal random variable and

$$\xi \stackrel{\text{def}}{=} \frac{f(\mathbf{0})\text{tr}(r''(\mathbf{0})) + 2r'(\mathbf{0})^\top f'(\mathbf{0})}{2(d+2)V_d^{2/d} f^{1+2/d}(\mathbf{0})}.$$

The result of Theorem 14.3 can form the basis of a further discussion regarding the choice of the weights. We essentially have two categories of weights, one in which we are not restricting v_{ni} to be nonnegative, and one in which $v_{ni} \geq 0$ for all i. If we choose

$$\sum_{i=1}^{k} v_{ni} i^{2/d} = 0, \tag{14.5}$$

a choice discussed earlier, then, in view of the fact that $\frac{1}{k} \leq \sum_{i=1}^{n} v_{ni}^2 \leq \frac{c^2}{k}$, we have the following:

Theorem 14.4. *Assume that conditions (i)-(v) are satisfied and, in addition, that (14.5) holds. Then there exists a sequence $\{k\} = \{k_n\}$ with the property that*

$$\frac{k}{n^{\frac{4}{d+4}}} \to \infty, \quad \frac{k}{n} \to 0,$$

such that

$$r_n(\mathbf{0}) - r(\mathbf{0}) = o_{\mathbb{P}}\left(n^{-\frac{2}{d+4}}\right).$$

The trouble with this result is that we cannot specify the choice of k. To do so would require an analysis of the next term in the bias. But knowledge of that next term comes only at the expense of higher order derivative conditions on f and r, and under the new conditions, one can push things further and get even better rates than those of Theorem 14.4.

If, on the other hand, we introduce the additional requirement $\inf_i v_{ni} \geq 0$ for all n, then the rate of convergence of $r_n(\mathbf{0}) - r(\mathbf{0})$ is limited to $n^{-\frac{2}{d+4}}$. The only good thing in that case is that one can optimize both k and the shape of the weight vector based on Theorem 14.3. Besides, the best shape of $(v_{n1}, \ldots, v_{nn})$ and the choice of k can be determined separately and universally for all regression estimation problems satisfying (i)-(v).

If we set $k = Kn^{\frac{4}{d+4}}$ for some constant $K > 0$, then (14.4) can be recast as follows:

$$\frac{r_n(\mathbf{0}) - r(\mathbf{0})}{n^{\frac{2}{d+4}}} \mathrel{\mathscr{D}}{=} \frac{\sigma(\mathbf{0})}{\sqrt{K}} \sqrt{k \sum_{i=1}^{n} v_{ni}^2 N + \beta \left(\sum_{i=1}^{k} v_{ni} \left(\frac{i}{k} \right)^{2/d} \right) K^{2/d}} + \mathrm{o}_{\mathbb{P}}(1). \tag{14.6}$$

Several points of view may be taken now. As noted earlier, both expressions involving the weights are bounded away from 0 and ∞ for nonnegative weights. Since $\sum_{i=1}^{k} v_{ni} \left(\frac{i}{k} \right)^{2/d}$ is minimal for monotonically decreasing v_{ni}, and since rearrangements of indices do not alter $\sum_{i=1}^{n} v_{ni}^2$, it suffices to consider only monotonically decreasing weights in (14.6).

If we take the expected value of the square of the last term in (14.6)—thus ignoring the $\mathrm{o}_{\mathbb{P}}(1)$ term—then we have

$$\liminf_{n \to \infty} \frac{\mathbb{E} \left| r_n(\mathbf{0}) - r(\mathbf{0}) \right|^2}{n^{\frac{4}{d+4}}} > 0.$$

Doing this, and putting

$$V = k \sum_{i=1}^{k} v_{ni}^2, \quad B = \left(\sum_{i=1}^{k} v_{ni} \left(\frac{i}{k} \right)^{2/d} \right)^2,$$

leads us to minimize the quantity

$$\frac{\sigma^2(\mathbf{0})}{K} V + \beta^2 B K^{4/d}. \tag{14.7}$$

The optimal choice for K is

$$K^\star = \left(\frac{d \sigma^2(\mathbf{0}) V}{4 \beta^2 B} \right)^{\frac{d}{d+4}}.$$

With that choice, (14.7) becomes

$$\left[\left(\frac{4}{d} \right)^{\frac{d}{d+4}} + \left(\frac{d}{4} \right)^{\frac{4}{d+4}} \right] \left(\sigma^2(\mathbf{0}) \right)^{\frac{4}{d+4}} \beta^{\frac{2d}{d+4}} V^{\frac{4}{d+4}} B^{\frac{d}{d+4}}. \tag{14.8}$$

Interestingly, the optimization of (14.8) leads to the problem of finding the best vector $v_{ni} \geq 0$, i.e., the one that minimizes $V^4 B^d$:

$$V^4 B^d = \left(k \sum_{i=1}^{k} v_{ni}^2 \right)^4 \left(\sum_{i=1}^{k} v_{ni} \left(\frac{i}{k} \right)^{2/d} \right)^d. \tag{14.9}$$

For any $v_{ni} \geq 0$ satisfying (iv), the expression in (14.9) is sandwiched between two strictly positive constants, uniformly over all k. The actual best form of $v_{n1} \geq \cdots \geq v_{nk}$ with $\sum_{i=1}^{k} v_{ni} = 1$ that minimizes (14.9) is unknown to us, except for $d = 1$. In this case, the literature on kernel density estimation (see, e.g., Devroye, 1987, or Tsybakov, 2008) permits one to show that (14.9) is asymptotically minimized by the Epanechnikov kernel, i.e.,

$$v_{ni} = \frac{c_k}{k}\left(1 - \left(\frac{i}{k}\right)^2\right), \quad 1 \leq i \leq k,$$

where $\lim_{k\to\infty} c_k = 4/3$.

14.6 Supplement: L^2 rates of convergence

This chapter has focused on pointwise rates of convergence of the nearest neighbor estimate. Of course, it is also possible to study rates of convergence for integrated criteria, such as the mean integrated squared error $\mathbb{E}|r_n(\mathbf{X}) - r(\mathbf{X})|^2$. This topic is sketched hereafter. To simplify the analysis, we focus on rates of convergence results for the class of smooth distributions of $(\mathbf{X}, Y)$ such that $\mathbf{X}$ takes values in $[0,1]^d$, $\mathbb{E}Y^2 < \infty$, the regression function r is Lipschitz, and the conditional variance function $\sigma^2(\mathbf{x}) = \mathbb{E}[|Y - r(\mathbf{X})|^2 \,|\, \mathbf{X} = \mathbf{x}]$ is uniformly bounded on $\mathbb{R}^d$ (for L^2 rates of convergence under more general conditions, see Kohler et al., 2006).

Theorem 14.5 (L^2 rates of convergence). *Let $r_n(\mathbf{x}) = \sum_{i=1}^{n} v_{ni} Y_{(i)}(\mathbf{x})$ be the nearest neighbor regression function estimate, where $(v_{n1}, \ldots, v_{nn})$ is a probability weight vector. Assume that $\mathbf{X}$ takes values in $[0,1]^d$. Assume, in addition, that for all $\mathbf{x}$ and $\mathbf{x}' \in \mathbb{R}^d$,*

$$\left|r(\mathbf{x}) - r(\mathbf{x}')\right| \leq L\|\mathbf{x} - \mathbf{x}'\|$$

and

$$\sup_{\mathbf{x}\in\mathbb{R}^d} \sigma^2(\mathbf{x}) \leq \sigma^2,$$

for some positive constants L and σ^2. Then

(i) For $d = 1$,

$$\mathbb{E}\,|r_n(\mathbf{X}) - r(\mathbf{X})|^2 \leq \sigma^2 \sum_{i=1}^{n} v_{ni}^2 + 8L^2 \sum_{i=1}^{n} v_{ni}\,\frac{i}{n}.$$

(ii) For $d \geq 2$,

$$\mathbb{E}\,|r_n(\mathbf{X}) - r(\mathbf{X})|^2 \leq \sigma^2 \sum_{i=1}^{n} v_{ni}^2 + c'_d L^2 \sum_{i=1}^{n} v_{ni}\left(\frac{i}{n}\right)^{2/d},$$

where

$$c'_d = \frac{2^{3+\frac{2}{d}}(1+\sqrt{d})^2}{V_d^{2/d}}.$$

For the standard k-nearest neighbor estimate, we have the following corollary:

Corollary 14.2. *Let r_n be the k-nearest neighbor regression function estimate. Then, under the conditions of Theorem 14.5,*

(i) For $d = 1$, there exists a sequence $\{k\} = \{k_n\}$ with $k \sim \sqrt{\frac{\sigma^2}{L^2}\, n}$ such that

$$\mathbb{E}\,|r_n(\mathbf{X}) - r(\mathbf{X})|^2 \le \Lambda_1 \sqrt{\frac{\sigma^2 L^2}{n}},$$

for some positive universal constant Λ_1.

(ii) For $d \ge 2$, there exists a sequence $\{k\} = \{k_n\}$ with $k \sim (\frac{\sigma^2}{L^2})^{\frac{d}{d+2}} n^{\frac{2}{d+2}}$ such that

$$\mathbb{E}\,|r_n(\mathbf{X}) - r(\mathbf{X})|^2 \le \Lambda_d \left(\frac{\sigma^2 L^d}{n}\right)^{\frac{2}{d+2}},$$

for some positive universal constant Λ_d.

The explicit bounds of Corollary 14.2 are valid for all finite sample sizes. On the other hand, the estimates with the optimal rate of convergence depend upon the unknown distribution of $(\mathbf{X}, Y)$—it is to correct this situation that we present adaptation results in Chapter 16. Also, we encounter here the phenomenon called the curse of dimensionality: in order to achieve the error $\mathbb{E}|r_n(\mathbf{X}) - r(\mathbf{X})|^2 \approx \varepsilon$, we need a sample of size $n \approx (1/\varepsilon)^{1+\frac{d}{2}}$, which is exponentially large in d. Thus, to get good error rates, the number of data points should grow exponentially with the number of components of $\mathbf{X}$. A possible route to address this shortcoming is feature selection—see Chapter 16.

Proof (Theorem 14.5). The proof of Theorem 14.5 relies on Theorem 2.4, which bounds the expected square distance between $\mathbf{X}$ and its i-th nearest neighbor. Letting

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

we start with the variance/bias decomposition

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

To bound the first term, note that

$$\mathbb{E}\,|r_n(\mathbf{X}) - \tilde{r}_n(\mathbf{X})|^2 = \mathbb{E}\left|\sum_{i=1}^n W_{ni}(\mathbf{X})\,(Y_i - r(\mathbf{X}_i))\right|^2,$$

where $W_{ni}(\mathbf{X}) = v_{n\Sigma_i}$ and $(\Sigma_1, \ldots, \Sigma_n)$ is a permutation of $(1, \ldots, n)$ such that $\mathbf{X}_i$ is the Σ_i-th nearest neighbor of $\mathbf{X}$ for all i. But we have already shown in (10.4) that

$$\mathbb{E}\left|\sum_{i=1}^n W_{ni}(\mathbf{X})\,(Y_i - r(\mathbf{X}_i))\right|^2 = \mathbb{E}\left[\sum_{i=1}^n W_{ni}^2(\mathbf{X})\sigma^2(\mathbf{X}_i)\right],$$

so that

$$\mathbb{E}\,|r_n(\mathbf{X}) - \tilde{r}_n(\mathbf{X})|^2 \le \sigma^2 \sum_{i=1}^n v_{ni}^2.$$

Finally,

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

The conclusion follows by applying Theorem 2.4. □

Chapter 15
Regression: the noiseless case

15.1 Noiseless estimation

Classical function estimation deals with the estimation of a function r on $\mathbb{R}^d$ from a finite number of points $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Some applications are concerned with L^p errors with respect to the Lebesgue measure on compacts. Others use it for Monte Carlo purposes, wanting to estimate $\int_A r(\mathbf{x})\mathrm{d}\mathbf{x}$ over a compact set A. The model we study here takes a sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of i.i.d. random vectors with a density f on A that is not known. We observe

$$Y_i = r(\mathbf{X}_i), \quad 1 \le i \le n,$$

and study nearest neighbor-style estimates of r. If $\mathbf{X}_{(i)}(\mathbf{x})$ is the i-th nearest neighbor of $\mathbf{x}$ among $\mathbf{X}_1, \ldots, \mathbf{X}_n$, and $Y_{(i)}(\mathbf{x}) = r(\mathbf{X}_{(i)}(\mathbf{x}))$, then the general estimate is

$$r_n(\mathbf{x}) = \sum_{i=1}^{n} v_{ni} Y_{(i)}(\mathbf{x}) = \sum_{i=1}^{n} v_{ni} r\left(\mathbf{X}_{(i)}(\mathbf{x})\right),$$

where $(v_{n1}, \ldots, v_{nn})$ is a weight vector summing to one. To simplify the analysis, we set $v_{ni} = \frac{1}{k}\mathbb{1}_{[1 \le i \le k]}$, where $\{k\} = \{k_n\}$ is a sequence of integers between 1 and n. Thus, in this chapter,

$$r_n(\mathbf{x}) = \frac{1}{k} \sum_{i=1}^{k} r\left(\mathbf{X}_{(i)}(\mathbf{x})\right). \tag{15.1}$$

The knee-jerk reaction in this noiseless situation is to take $k = 1$. Indeed, how can one do better than taking the nearest neighbor? However, as we will see below, one can in fact outperform the 1-nearest neighbor in dimensions 2 and above. That point will be made by a careful analysis of the pointwise error $r_n(\mathbf{x}) - r(\mathbf{x})$.

G. Biau, L. Devroye, *Lectures on the Nearest Neighbor Method*,
Springer Series in the Data Sciences, DOI 10.1007/978-3-319-25388-6_15

15.2 A local limit law

Since we consider the local behavior at $\mathbf{x}$, we assume that $\mathbf{x} = \mathbf{0}$. Throughout, for simplicity of notation, we drop the dependence upon the query point $\mathbf{0}$, and write $\mathbf{X}_{(i)}$ and $Y_{(i)}$ instead of $\mathbf{X}_{(i)}(\mathbf{0})$ and $Y_{(i)}(\mathbf{0})$. The objective of this introductory section is to study the limit behavior in $\mathbb{R}^{dk}$ of the vector $(\mathbf{X}_{(1)}, \ldots, \mathbf{X}_{(k)})$ when k is constant. These limit laws will be used later in the chapter to study the asymptotic behavior of $r_n(\mathbf{0}) - r(\mathbf{0})$, first when k is constant, and next when $k \to \infty$ while $k/n \to 0$.

We assume that $\mathbf{0}$ is a Lebesgue point of f and $f(\mathbf{0}) > 0$, but no other conditions are necessary for the main result of this section. As always, we denote by μ the common distribution of the $\mathbf{X}_i$'s.

Let $E_1, \ldots, E_{n+1}$ be independent standard exponential random variables. We know (Corollary 1.1) that the order statistics for n uniform points in $[0, 1]$, denoted by $U_{(1)} \leq \cdots \leq U_{(n)}$, are distributed as follows:

$$\left(U_{(1)}, \ldots, U_{(n)}\right) \stackrel{\mathcal{D}}{=} \left(\frac{G_1}{G_{n+1}}, \ldots, \frac{G_n}{G_{n+1}}\right), \tag{15.2}$$

where $G_i = \sum_{j=1}^{i} E_j$, $1 \leq i \leq n+1$. Since

$$\left(\mu\left(B(\mathbf{0}, \|\mathbf{X}_{(1)}\|)\right), \ldots, \mu\left(B(\mathbf{0}, \|\mathbf{X}_{(n)}\|)\right)\right) \stackrel{\mathcal{D}}{=} \left(U_{(1)}, \ldots, U_{(n)}\right), \tag{15.3}$$

and $\mu(B(\mathbf{0}, \rho)) \sim f(\mathbf{0}) V_d \rho^d$ as $\rho \downarrow 0$ (because $\mathbf{0}$ is a Lebesgue point), it is immediate (and a corollary of a stronger statement proved below) that for fixed k,

$$f(\mathbf{0}) V_d n\left(\|\mathbf{X}_{(1)}\|^d, \ldots, \|\mathbf{X}_{(k)}\|^d\right) \stackrel{\mathcal{D}}{\to} (G_1, \ldots, G_k).$$

The following theorem, proved by coupling, is thus not surprising.

Theorem 15.1. *Assume that $\mathbf{0}$ is a Lebesgue point of f, $f(\mathbf{0}) > 0$, and k is fixed. Then*

$$(f(\mathbf{0}) V_d n)^{1/d} (\mathbf{X}_{(1)}, \ldots, \mathbf{X}_{(k)}) \stackrel{\mathcal{D}}{\to} \left(\mathbf{Z}_1 G_1^{1/d}, \ldots, \mathbf{Z}_k G_k^{1/d}\right),$$

where $G_1, \ldots, G_k$ are as above, and $\mathbf{Z}_1, \ldots, \mathbf{Z}_k$ are independent random vectors uniformly distributed on the surface of $B(\mathbf{0}, 1)$.

Proof. Let K be a positive constant to be chosen later. Consider a density $g_{K,n}$ related to f as follows: let

$$p = \int_{B\left(\mathbf{0}, \frac{K}{n^{1/d}}\right)} f(\mathbf{z}) \mathrm{d}\mathbf{z},$$

and set

$$g_{K,n}(\mathbf{x}) = \begin{cases} f(\mathbf{0}) & \text{for } \mathbf{x} \in B\left(\mathbf{0}, \frac{K}{n^{1/d}}\right) \\ f(\mathbf{x}) \left(\frac{1 - f(\mathbf{0}) V_d \frac{K^d}{n}}{1-p} \right) & \text{otherwise,} \end{cases}$$

which is a proper density (i.e., nonnegative and integrating to one) for all n large enough. We assume that n is indeed large enough for this to happen. Note that

$$\begin{aligned} \int_{\mathbb{R}^d} |g_{K,n}(\mathbf{x}) - f(\mathbf{x})| \, d\mathbf{x} &= \int_{B\left(\mathbf{0}, \frac{K}{n^{1/d}}\right)} |f(\mathbf{0}) - f(\mathbf{x})| \, d\mathbf{x} \\ &\quad + \int_{B^c\left(\mathbf{0}, \frac{K}{n^{1/d}}\right)} f(\mathbf{x}) \left| \frac{1 - f(\mathbf{0}) V_d \rho^d}{1-p} - 1 \right| d\mathbf{x} \\ &= \int_{B\left(\mathbf{0}, \frac{K}{n^{1/d}}\right)} |f(\mathbf{0}) - f(\mathbf{x})| \, d\mathbf{x} + \left| p - f(\mathbf{0}) V_d \frac{K^d}{n} \right| \\ &\le 2 \int_{B\left(\mathbf{0}, \frac{K}{n^{1/d}}\right)} |f(\mathbf{0}) - f(\mathbf{x})| \, d\mathbf{x} \\ &= o\left(\frac{1}{n}\right) \end{aligned}$$

(since $\mathbf{0}$ is a Lebesgue point of f).

Therefore, by Doeblin's coupling method (Doeblin, 1937; see, e.g., Rachev and Rüschendorf, 1998), there exist random variables $\mathbf{X}$ and $\mathbf{Y}$ with density f and $g_{K,n}$, respectively, such that

$$\mathbb{P}\{\mathbf{Y} \neq \mathbf{X}\} = \frac{1}{2} \int_{\mathbb{R}^d} |g_{K,n}(\mathbf{x}) - f(\mathbf{x})| \, d\mathbf{x} = o\left(\frac{1}{n}\right).$$

Repeating this n times, we create two coupled samples of random variables that are i.i.d. within the sample. The sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ is drawn from the distribution of $\mathbf{X}$, and the sample $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ is drawn from the distribution of $\mathbf{Y}$.

Recall that the total variation distance between two random vectors $\mathbf{W}, \mathbf{W}' \in \mathbb{R}^d$ is defined by

$$d_{\mathrm{TV}}(\mathbf{W}, \mathbf{W}') = \sup_{A \in \mathscr{B}} \left| \mathbb{P}\{\mathbf{W} \in A\} - \mathbb{P}\{\mathbf{W}' \in A\} \right|,$$

where $\mathscr{B}$ denotes the Borel sets of $\mathbb{R}^d$. Let $\|\mathbf{Y}_{(1)}\| \le \cdots \le \|\mathbf{Y}_{(n)}\|$ and $\|\mathbf{X}_{(1)}\| \le \cdots \le \|\mathbf{X}_{(n)}\|$ be the reordered samples. Then

$$\begin{aligned} d_{\mathrm{TV}}\big((\mathbf{Y}_{(1)},\ldots,\mathbf{Y}_{(k)}),(\mathbf{X}_{(1)},\ldots,\mathbf{X}_{(k)})\big) &\le d_{\mathrm{TV}}\big((\mathbf{Y}_1,\ldots,\mathbf{Y}_n),(\mathbf{X}_1,\ldots,\mathbf{X}_n)\big) \\ &\le \sum_{i=1}^{n} \mathbb{P}\{\mathbf{Y}_i \neq \mathbf{X}_i\} \\ &\le n \times \mathrm{o}\Big(\frac{1}{n}\Big) \\ &= \mathrm{o}(1) \quad \text{as } n \to \infty. \end{aligned}$$

Define

$$U_{(i)} = \nu\left(B(\mathbf{0}, \|\mathbf{Y}_{(i)}\|)\right),$$

where ν is the probability measure of $\mathbf{Y}$. We recall that $U_{(1)},\ldots,U_{(n)}$ are uniform order statistics, and thus

$$n\big(U_{(1)},\ldots,U_{(k)}\big) \xrightarrow{\mathscr{D}} \big(G_1,\ldots,G_k\big)$$

(see Corollary 1.1). In fact, this convergence is also in the d_{TV} sense. Also, if $\|\mathbf{Y}_{(k)}\| \le K/n^{1/d}$, then

$$U_{(i)} = f(\mathbf{0})V_d\|\mathbf{Y}_{(i)}\|^d.$$

Thus,

$$\begin{aligned} & d_{\mathrm{TV}}\left(\big(\|\mathbf{Y}_{(1)}\|,\ldots,\|\mathbf{Y}_{(k)}\|\big),\left(\Big(\frac{U_{(1)}}{f(\mathbf{0})V_d}\Big)^{1/d},\ldots,\Big(\frac{U_{(k)}}{f(\mathbf{0})V_d}\Big)^{1/d}\right)\right) \\ &\quad \le \mathbb{P}\Big\{\|\mathbf{Y}_{(k)}\| > \frac{K}{n^{1/d}}\Big\} \\ &\quad = \mathbb{P}\Big\{\mathrm{Bin}\big(n, f(\mathbf{0})V_d\tfrac{K^d}{n}\big) < k\Big\} \\ &\quad = \mathbb{P}\Big\{\mathrm{Poisson}(f(\mathbf{0})V_dK^d) < k\Big\} + \mathrm{o}(1) \quad (\text{as } n \to \infty) \\ &\quad \le \varepsilon + \mathrm{o}(1) \end{aligned}$$

for all K large enough, depending upon ε.

Let $\mathbf{Z}_1,\ldots,\mathbf{Z}_k$ be i.i.d. random vectors uniformly distributed on the surface of $B(\mathbf{0},1)$. Then

$$\begin{aligned} & d_{\mathrm{TV}}\left((f(\mathbf{0})V_dn)^{1/d}(\mathbf{X}_{(1)},\ldots,\mathbf{X}_{(k)}),(\mathbf{Z}_1G_1^{1/d},\ldots,\mathbf{Z}_kG_k^{1/d})\right) \\ &\quad \le d_{\mathrm{TV}}\big((\mathbf{X}_{(1)},\ldots,\mathbf{X}_{(k)}),(\mathbf{Y}_{(1)},\ldots,\mathbf{Y}_{(k)})\big) \\ &\qquad + d_{\mathrm{TV}}\left((f(\mathbf{0})V_dn)^{1/d}(\mathbf{Y}_{(1)},\ldots,\mathbf{Y}_{(k)}),(\mathbf{Z}_1G_1^{1/d},\ldots,\mathbf{Z}_kG_k^{1/d})\right) \end{aligned}$$

$$\begin{aligned}
&\le \mathrm{o}(1) + d_{\mathrm{TV}}\big((\mathbf{Y}_{(1)},\dots,\mathbf{Y}_{(k)}),(\mathbf{Z}_1\|\mathbf{Y}_{(1)}\|,\dots,\mathbf{Z}_k\|\mathbf{Y}_{(k)}\|)\big)\\
&\quad + d_{\mathrm{TV}}\left((f(\mathbf{0})V_d n)^{1/d}(\mathbf{Z}_1\|\mathbf{Y}_{(1)}\|,\dots,\mathbf{Z}_k\|\mathbf{Y}_{(k)}\|),(\mathbf{Z}_1 G_1^{1/d},\dots,\mathbf{Z}_k G_k^{1/d})\right)\\
&\le \mathrm{o}(1) + \mathbb{P}\left\{\|\mathbf{Y}_{(k)}\| > \frac{K}{n^{1/d}}\right\}\\
&\quad + d_{\mathrm{TV}}\left((\|\mathbf{Y}_{(1)}\|,\dots,\|\mathbf{Y}_{(k)}\|),\left(\left(\frac{U_{(1)}}{f(\mathbf{0})V_d}\right)^{1/d},\dots,\left(\frac{U_{(k)}}{f(\mathbf{0})V_d}\right)^{1/d}\right)\right)\\
&\quad + d_{\mathrm{TV}}\big(n^{1/d}(\mathbf{Z}_1 U_{(1)}^{1/d},\dots,\mathbf{Z}_k U_{(k)}^{1/d}),(\mathbf{Z}_1 G_1^{1/d},\dots,\mathbf{Z}_k G_k^{1/d})\big)
\end{aligned}$$

since on $[\|\mathbf{Y}\| \le K/n^{1/d}]$, $\mathbf{Y}$ has a radially symmetric distribution. Therefore,

$$\begin{aligned}
&d_{\mathrm{TV}}\left((f(\mathbf{0})V_d n)^{1/d}(\mathbf{X}_{(1)},\dots,\mathbf{X}_{(k)}),(\mathbf{Z}_1 G_1^{1/d},\dots,\mathbf{Z}_k G_k^{1/d})\right)\\
&\quad \le \mathrm{o}(1) + 2\varepsilon + d_{\mathrm{TV}}\big(n(U_{(1)},\dots,U_{(k)}),(G_1,\dots,G_k)\big)\\
&\quad = \mathrm{o}(1) + 2\varepsilon.
\end{aligned}$$

This concludes the proof. □

Writing $\mathbf{Z}_i = (Z_{i,1},\dots,Z_{i,d})$, we have the following corollary:

Corollary 15.1. *Assume that* $\mathbf{0}$ *is a Lebesgue point of* f, $f(\mathbf{0}) > 0$, *and* k *is fixed. Then*

$$f(\mathbf{0})V_d n\big(\|\mathbf{X}_{(1)}\|^d,\dots,\|\mathbf{X}_{(k)}\|^d\big) \xrightarrow{\mathscr{D}} (G_1,\dots,G_k).$$

Also, if $a_1,\dots,a_d$ *are real numbers, then, writing* $\mathbf{X}_{(i)} = (X_{(i,1)},\dots,X_{(i,d)})$,

$$\begin{aligned}
&(f(\mathbf{0})V_d n)^{1/d}\left(\sum_{j=1}^d a_j X_{(1,j)},\dots,\sum_{j=1}^d a_j X_{(k,j)}\right)\\
&\xrightarrow{\mathscr{D}} \left(G_1^{1/d}\sum_{j=1}^d a_j Z_{1,j},\dots,G_k^{1/d}\sum_{j=1}^d a_j Z_{k,j}\right).
\end{aligned}$$

15.3 Analysis for fixed k

In this section, we still assume that k is held fixed and study the asymptotic behavior of $r_n(\mathbf{0}) - r(\mathbf{0})$, where r_n is the k-nearest neighbor estimate (15.1). The standing conditions for this section are the following ones:

(1) $\mathbf{0}$ is a Lebesgue point of the density f and $f(\mathbf{0}) > 0$.
(2) The regression function r is continuously differentiable in a neighborhood of $\mathbf{0}$.

Recall the notation $\mathbf{X}_{(i)} = (X_{(i,1)}, \ldots, X_{(i,d)})$, $\mathbf{Z}_i = (Z_{i,1}, \ldots, Z_{i,d})$ ($\mathbf{Z}_1, \ldots, \mathbf{Z}_k$ are independent random vectors uniformly distributed on the surface of $B(\mathbf{0}, 1)$), and let $r'(\mathbf{0}) = (r'_1(\mathbf{0}), \ldots, r'_d(\mathbf{0}))$, with $r'_j(\mathbf{0}) = \frac{\partial r}{\partial x_j}(\mathbf{0})$, $1 \le j \le d$. Then, by a Taylor series approximation,

$$\begin{aligned} r_n(\mathbf{0}) - r(\mathbf{0}) &= \frac{1}{k} \sum_{i=1}^{k} \big(r(\mathbf{X}_{(i)}) - r(\mathbf{0}) \big) \\ &= \frac{1}{k} \sum_{i=1}^{k} \left(\sum_{j=1}^{d} r'_j(\mathbf{0}) X_{(i,j)} + \psi(\mathbf{X}_{(i)}) \right) \end{aligned}$$

(where $\psi(\mathbf{x}) = \mathrm{o}(\|\mathbf{x}\|)$ as $\|\mathbf{x}\| \downarrow 0$). Observe that

$$\frac{\frac{1}{k} \sum_{i=1}^{k} \psi(\mathbf{X}_{(i)})}{\|\mathbf{X}_{(k)}\|} \to 0 \quad \text{in probability}$$

if $\|\mathbf{X}_{(k)}\| \to 0$ in probability. But, by Corollary 15.1, $\|\mathbf{X}_{(k)}\| = \mathrm{O}_{\mathbb{P}}((\frac{k}{n})^{1/d})$, and therefore,

$$\frac{1}{k} \sum_{i=1}^{k} \psi(\mathbf{X}_{(i)}) = \mathrm{o}_{\mathbb{P}} \left(\left(\frac{k}{n} \right)^{1/d} \right).$$

Still by Corollary 15.1,

$$\begin{aligned} &(f(\mathbf{0}) V_d n)^{1/d} \left(\sum_{j=1}^{d} r'_j(\mathbf{0}) X_{(1,j)}, \ldots, \sum_{j=1}^{d} r'_j(\mathbf{0}) X_{(k,j)} \right) \\ &\xrightarrow{\mathscr{D}} \left(G_1^{1/d} \sum_{j=1}^{d} r'_j(\mathbf{0}) Z_{1,j}, \ldots, G_k^{1/d} \sum_{j=1}^{d} r'_j(\mathbf{0}) Z_{k,j} \right), \end{aligned}$$

so that

$$(f(\mathbf{0}) V_d n)^{1/d} \left(\frac{1}{k} \sum_{i=1}^{k} \sum_{j=1}^{d} r'_j(\mathbf{0}) X_{(i,j)} \right) \xrightarrow{\mathscr{D}} \frac{1}{k} \sum_{i=1}^{k} G_i^{1/d} \left(\sum_{j=1}^{d} r'_j(\mathbf{0}) Z_{i,j} \right).$$

In particular, for $k = 1$, the limit law is $E^{1/d} \sum_{j=1}^{d} r'_j(\mathbf{0}) Z_{1,j}$, where E denotes a standard exponential random variable. By radial symmetry,

$$\sum_{j=1}^{d} r'_j(\mathbf{0}) Z_{1,j} \stackrel{\mathscr{D}}{=} \|r'(\mathbf{0})\| Z_{1,1},$$

where $\|r'(\mathbf{0})\|^2 = \sum_{j=1}^d r_j'^2(\mathbf{0})$. Thus, for $k = 1$, we have

$$r_n(\mathbf{0}) - r(\mathbf{0}) \stackrel{\mathscr{D}}{=} \frac{\|r'(\mathbf{0})\|}{(f(\mathbf{0})V_d n)^{1/d}} E^{1/d} Z_{1,1} + \mathrm{o}_{\mathbb{P}}\left(\frac{1}{n^{1/d}}\right)$$
$$= \begin{cases} \mathrm{o}_{\mathbb{P}}\left(\frac{1}{n^{1/d}}\right) & \text{if } \|r'(\mathbf{0})\| = 0 \\ \mathrm{O}_{\mathbb{P}}\left(\frac{1}{n^{1/d}}\right) & \text{if } \|r'(\mathbf{0})\| > 0. \end{cases}$$

Note that for $d = 1$, $EZ_{1,1}$ has the Laplace distribution with density $\frac{1}{2}e^{-|x|}$, $x \in \mathbb{R}$.

Remark 15.1. It is easy to see that if $N_1, \ldots, N_d$ are i.i.d. standard Gaussian random variables, then, with $\mathbf{N} = (N_1, \ldots, N_d)$, the normalized random vector $\mathbf{N}/\|\mathbf{N}\|$ is uniformly distributed on the surface of $B(\mathbf{0}, 1)$. Hence,

$$Z_{i,1} \stackrel{\mathscr{D}}{=} \frac{N_1}{\|\mathbf{N}\|} = \frac{N_1}{\sqrt{\sum_{j=1}^d N_j^2}}.$$

Therefore, for $d > 1$,

$$Z_{i,1}^2 \stackrel{\mathscr{D}}{=} \frac{N_1^2}{\sum_{j=1}^d N_j^2} \stackrel{\mathscr{D}}{=} \frac{G_{1/2}}{G_{1/2} + G_{(d-1)/2}},$$

where $G_{1/2}$ is Gamma$(1/2)$ and $G_{(d-1)/2}$ is Gamma$(\frac{d-1}{2})$ and independent of $G_{1/2}$ (see Section 20.9 in the Appendix). Thus, by Lemma 20.9,

$$Z_{i,1}^2 \stackrel{\mathscr{D}}{=} \mathrm{Beta}\left(\tfrac{1}{2}, \tfrac{d-1}{2}\right),$$

so that $Z_{i,1}$ has density

$$\frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{d-1}{2}\right)}(1 - z^2)^{\frac{d-3}{2}}, \quad |z| \le 1.$$

We recover the well-known fact that $Z_{i,1}$ is uniform on $[-1, 1]$ for $d = 3$. □

Summarizing all of the above, we conclude:

Theorem 15.2. *Assume that conditions* (1) *and* (2) *are satisfied. Then, for the fixed k-nearest neighbor regression function estimate r_n,*

$$(f(\mathbf{0})V_d n)^{1/d}\,(r_n(\mathbf{0}) - r(\mathbf{0})) \stackrel{\mathscr{D}}{\to} L_k \|r'(\mathbf{0})\|,$$

where

$$L_k \stackrel{\text{def}}{=} \frac{1}{k} \sum_{i=1}^k G_i^{1/d} Z_{i,1}.$$

Analysis of L_k suggests that

$$r_n(\mathbf{0}) - r(\mathbf{0}) = \mathrm{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{k}}\left(\frac{k}{n}\right)^{1/d}\right). \tag{15.4}$$

To see this we note below that, as $k \to \infty$,

$$\mathbb{E}L_k^2 \sim \frac{k^{\frac{2}{d}-1}}{d+2}.$$

Combined with Theorem 15.2, one may be tempted to conclude (15.4). This is an erroneous reasoning because Theorem 15.2 does not permit $k \to \infty$. A more careful analysis shows that (15.4) is indeed true if $k = \mathrm{O}(n^{\frac{2}{d+2}})$, under conditions that are stricter than (1) and (2)—see the next section.

Let us return to the computation of $\mathbb{E}L_k^2$. We have

$$\mathbb{E}[L_k^2 \,|\, G_1, \ldots, G_k] = \frac{1}{k^2}\sum_{i=1}^{k} G_i^{2/d}\mathbb{E}Z_{i,1}^2 = \mathbb{E}Z_{1,1}^2 \times \frac{1}{k^2}\sum_{i=1}^{k} G_i^{2/d}.$$

Thus, since $\mathbb{E}Z_{1,1}^2 = \frac{\mathbb{E}\|\mathbf{Z}_1\|^2}{d} = 1/d$,

$$\mathbb{E}L_k^2 = \frac{1}{dk^2}\sum_{i=1}^{k}\mathbb{E}G_i^{2/d} = \frac{1}{dk^2}\sum_{i=1}^{k}\frac{\Gamma\left(i+\frac{2}{d}\right)}{\Gamma(i)} \sim \frac{k^{\frac{2}{d}-1}}{d+2}.$$

For the latter equivalence, we used the fact that

$$\frac{\Gamma\left(i+\frac{2}{d}\right)}{\Gamma(i)} = i^{2/d}\left(1+\mathrm{O}\left(\frac{1}{i}\right)\right),$$

which is implied by Theorem 20.14 in the Appendix. The quantity $\mathbb{E}L_k^2$ is minimal for $k = 1$ when $d = 1$. It is basically invariant under the choice of k for $d = 2$, but, surprisingly, for $d > 2$, $\mathbb{E}L_k^2$ is minimized for $k = n$. Of course, this argument fails because the delicate limit law we derived fails to hold. However, this motivates the study of $r_n(\mathbf{0}) - r(\mathbf{0})$ in $\mathbb{R}^d$, $d > 2$, with k depending upon n such that $k \to \infty$.

15.4 Analysis for diverging k

Most of the arguments used in the previous sections for finite k do not carry over to the case $k \to \infty$. It is nevertheless possible to extend the coupling argument to this situation. The standing conditions are as follows:

(*i*) $r_n(\mathbf{0}) = \frac{1}{k}\sum_{i=1}^{k} r(\mathbf{X}_{(i)})$, with $k/\log n \to \infty$ and $k/n \to 0$.
(*ii*) The density f is continuously differentiable in a neighborhood of $\mathbf{0}$ and $f(\mathbf{0}) > 0$.
(*iii*) The regression function r is twice continuously differentiable in a neighborhood of $\mathbf{0}$.
(*iv*) The function r is bounded.

For $\mathbf{x} = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$, we let

$$f'(\mathbf{0}) = \left(\frac{\partial f}{\partial x_1}(\mathbf{0}), \ldots, \frac{\partial f}{\partial x_d}(\mathbf{0})\right)^\top, \quad r'(\mathbf{0}) = \left(\frac{\partial r}{\partial x_1}(\mathbf{0}), \ldots, \frac{\partial r}{\partial x_d}(\mathbf{0})\right)^\top,$$

and

$$r''(\mathbf{0}) = \left(\frac{\partial^2 r}{\partial x_j \partial x_{j'}}(\mathbf{0})\right)_{1 \le j,j' \le d}.$$

Theorem 15.3 (Noiseless rate of convergence). *Assume that conditions* (*i*)-(*iv*) *are satisfied. Then the k-nearest neighbor regression function estimate* r_n *satisfies*

$$r_n(\mathbf{0}) - r(\mathbf{0}) \stackrel{\mathscr{D}}{=} \frac{\zeta}{\sqrt{k}}\left(\frac{k}{n}\right)^{1/d} N + \xi\left(\frac{k}{n}\right)^{2/d} + o_{\mathbb{P}}\left(\frac{1}{\sqrt{k}}\left(\frac{k}{n}\right)^{1/d} + \left(\frac{k}{n}\right)^{2/d}\right),$$

where N is a standard normal random variable,

$$\zeta \stackrel{\text{def}}{=} \frac{\|r'(\mathbf{0})\|}{\sqrt{d+2}\,(V_d f(\mathbf{0}))^{1/d}},$$

and

$$\xi \stackrel{\text{def}}{=} \frac{f(\mathbf{0})\mathrm{tr}(r''(\mathbf{0})) + 2r'(\mathbf{0})^\top f(\mathbf{0})}{2(d+2)V_d^{2/d} f^{1+2/d}(\mathbf{0})}.$$

Theorem 15.3 can be used to determine the best choice of k. If we take the sum of the expected values of the squares of the main terms on the right-hand side of the limit theorem as a yardstick, then the best k would minimize

$$\frac{\zeta^2}{k}\left(\frac{k}{n}\right)^{2/d} + \xi^2\left(\frac{k}{n}\right)^{4/d}. \tag{15.5}$$

For $d = 1$, the value $k = 1$ is best. The case of constant k was dealt with in the previous sections and is well understood. Besides, Theorem 15.3 requires $k \to \infty$ to be valid. For $d = 2$, (15.5) is

$$\frac{\zeta^2}{n} + \xi^2\left(\frac{k}{n}\right)^2,$$

Table 15.1 Optimal rates of convergence of $r_n(\mathbf{0}) - r(\mathbf{0})$ for the k-nearest neighbor regression function estimate in the noisy and noiseless cases.

	Noisy estimation	**Noiseless case**
$d = 1$	$\Theta(n^{-2/5})$	$\Theta(n^{-1})$
$d = 2$	$\Theta(n^{-1/3})$	$\Theta(n^{-1/2})$
$d > 2$	$\Theta(n^{-\frac{2}{d+4}})$	$\Theta(n^{-\frac{2}{d+2}})$

which is also minimal for $k = 1$, although any choice $k = o(\sqrt{n})$ will do almost as well. At $d = 2$, there is a phase transition, since for $d > 2$, (15.5) is minimized for a sequence k that grows with n roughly as $n^{\frac{2}{d+2}}$:

$$k \sim \left(\frac{\zeta^2}{\xi^2} \times \frac{d-2}{4} \right)^{\frac{d}{d+2}} n^{\frac{2}{d+2}}.$$

With that choice, (15.5) becomes

$$n^{-\frac{4}{d+2}} \left[\left(\frac{d-2}{4} \right)^{\frac{4}{d+2}} + \left(\frac{d-2}{4} \right)^{\frac{2-d}{2+d}} \right] (\zeta^2)^{\frac{4}{d+2}} (\xi^2)^{\frac{d-2}{d+2}}.$$

The influence of the local behavior of f and r is plainly visible in the multiplicative constant. It is perhaps worthwhile to juxtapose the results regarding optimal rates for the k-nearest neighbor estimate in the noisy and noiseless cases, as in Table 15.1. Here Θ stands for a rate "in probability" as described in Theorem 15.3. For every dimension d, the noiseless rate of convergence is, unsurprisingly, better, although the difference tends to decrease as the dimension increases.

Remark 15.2. We note that under our conditions, it is possible that one or both of ζ and ξ are zero, in which case there is a choice of k that makes $r_n(\mathbf{0}) - r(\mathbf{0}) = o_{\mathbb{P}}(n^{-\frac{2}{d+2}})$. So as to keep the notation and argument transparent, we did not consider the weighted nearest neighbor estimate here, but with appropriate weights, under the conditions of Theorem 15.3, one should be able to obtain $r_n(\mathbf{0}) - r(\mathbf{0}) = o_{\mathbb{P}}(n^{-\frac{2}{d+2}})$. □

Proof (Theorem 15.3). Theorem 15.3 is proved by using a coupling argument. We create two coupled samples of random variables that are i.i.d. within the sample. The sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ is drawn from the distribution of $\mathbf{X}$, which has density f. The sample $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ is drawn from the distribution of $\mathbf{Y}$, which has density g_ρ described below—g_ρ is nearly uniform on the ball $B(\mathbf{0}, \rho)$ and matches f up to a multiplicative constant outside $B(\mathbf{0}, \rho)$:

$$g_\rho(\mathbf{x}) = \begin{cases} f(\mathbf{0}) + f'(\mathbf{0})^\top \mathbf{x} & \text{for } \mathbf{x} \in B(\mathbf{0}, \rho) \\ f(\mathbf{x}) \left(\frac{1 - f(\mathbf{0}) V_d \rho^d}{1 - \int_{B(\mathbf{0}, \rho)} f(\mathbf{z}) d\mathbf{z}} \right) & \text{otherwise.} \end{cases}$$

Here, as elsewhere, vectors are column vectors and $\top$ denotes transposition. The radius is $\rho = K(k/n)^{1/d}$, where $K = K(\varepsilon)$ is picked so large that $\mathbb{P}\{\|\mathbf{X}_{(k)}\| > \rho\} \le \varepsilon$. This can be done since

$$\|\mathbf{X}_{(k)}\| = \mathrm{O}_{\mathbb{P}}\left(\left(\frac{k}{n}\right)^{1/d}\right)$$

(by (15.2), (15.3), and the fact that $\mathbf{0}$ is a Lebesgue point of f with $f(\mathbf{0}) > 0$—see condition (ii)). Since $\rho \downarrow 0$, g_ρ is a proper density for all n large enough (in particular, for ρ so small that $\|f'(\mathbf{0})\|\rho \le f(\mathbf{0})$).

We compare

$$r_n(\mathbf{0}) = \frac{1}{k}\sum_{i=1}^{k} r(\mathbf{X}_{(i)}) \quad \text{with} \quad s_n(\mathbf{0}) = \frac{1}{k}\sum_{i=1}^{k} r(\mathbf{Y}_{(i)}),$$

and prove that

$$r_n(\mathbf{0}) - s_n(\mathbf{0}) = \mathrm{o}_{\mathbb{P}}\left(\left(\frac{k}{n}\right)^{2/d}\right). \tag{15.6}$$

This is so small that it cannot possibly asymptotically dominate the last term shown in Theorem 15.3. So, we will establish Theorem 15.3 for $s_n(\mathbf{0})$.

The Taylor series expansion of r about $\mathbf{0}$ is

$$r(\mathbf{x}) = r(\mathbf{0}) + r'(\mathbf{0})^\top \mathbf{x} + \frac{1}{2}\mathbf{x}^\top r''(\mathbf{0})\mathbf{x} + \|\mathbf{x}\|^2 w(\mathbf{x}),$$

where $w(\mathbf{x}) = \mathrm{o}(1)$ as $\|\mathbf{x}\| \downarrow 0$. Thus,

$$\begin{aligned} s_n(\mathbf{0}) - s(\mathbf{0}) &= r'(\mathbf{0})^\top \frac{1}{k}\sum_{i=1}^{k} \mathbf{Y}_{(i)} + \frac{1}{2k}\sum_{i=1}^{k} \mathbf{Y}_{(i)}^\top r''(\mathbf{0})\mathbf{Y}_{(i)} \\ &\quad + \frac{1}{k}\sum_{i=1}^{k} \|\mathbf{Y}_{(i)}\|^2 w(\mathbf{Y}_{(i)}) \\ &\stackrel{\text{def}}{=} \mathbf{I} + \mathbf{II} + \mathbf{III}. \end{aligned}$$

We show that

$$\mathbf{III} = \mathrm{o}_{\mathbb{P}}\left(\left(\frac{k}{n}\right)^{2/d}\right), \tag{15.7}$$

so that it too has no influence in the asymptotic statement of Theorem 15.3. Terms **I** and **II** will be related, by appropriate probabilistic representations, to sums of uniform order statistics dealt with in Chapter 13, and this will establish Theorem 15.3. The four items, (15.6), (15.7), **I** and **II**, are treated in the next few paragraphs.

Proof of (15.6). Letting

$$p = \int_{B(\mathbf{0},\rho)} f(\mathbf{z})\mathrm{d}\mathbf{z},$$

we have

$$\begin{aligned}
\int_{\mathbb{R}^d} \left|g_\rho(\mathbf{x}) - f(\mathbf{x})\right| \mathrm{d}\mathbf{x} &= \int_{B(\mathbf{0},\rho)} \left|g_\rho(\mathbf{x}) - f(\mathbf{x})\right| \mathrm{d}\mathbf{x} \\
&\quad + \int_{B^c(\mathbf{0},\rho)} f(\mathbf{x}) \left| \frac{1 - f(\mathbf{0}) V_d \rho^d}{1-p} - 1 \right| \mathrm{d}\mathbf{x} \\
&= \int_{B(\mathbf{0},\rho)} \left|g_\rho(\mathbf{x}) - f(\mathbf{x})\right| \mathrm{d}\mathbf{x} + \left| p - f(\mathbf{0}) V_d \rho^d \right| \\
&\le 2 \int_{B(\mathbf{0},\rho)} \left|g_\rho(\mathbf{x}) - f(\mathbf{x})\right| \mathrm{d}\mathbf{x}.
\end{aligned}$$

By Doeblin's coupling lemma (Doeblin, 1937; Rachev and Rüschendorf, 1998), there exist random variables $\mathbf{X}$ and $\mathbf{Y}$ with density f and g_ρ, respectively, such that

$$\begin{aligned}
\mathbb{P}\{\mathbf{Y} \neq \mathbf{X}\} &= \frac{1}{2} \int_{\mathbb{R}^d} \left|g_\rho(\mathbf{x}) - f(\mathbf{x})\right| \mathrm{d}\mathbf{x} \\
&\le \int_{B(\mathbf{0},\rho)} \left|g_\rho(\mathbf{x}) - f(\mathbf{x})\right| \mathrm{d}\mathbf{x} \\
&= 2 \int_{B(\mathbf{0},\rho)} |f(\mathbf{x}) - f(\mathbf{0}) - f'(\mathbf{0})^\top \mathbf{x}| \mathrm{d}\mathbf{x} \\
&= \rho^{d+1} \epsilon(\rho),
\end{aligned}$$

where $\epsilon(\rho) = \mathrm{o}(1)$ as $\rho \downarrow 0$, by the Taylor series expansion of f about $\mathbf{0}$.

Let $(\mathbf{X}_i, \mathbf{Y}_i)$, $1 \le i \le n$, be independently drawn from the distribution of $(\mathbf{X}, \mathbf{Y})$. Then,

$$\begin{aligned}
&|r_n(\mathbf{0}) - s_n(\mathbf{0})| \, \mathbb{1}_{[\|\mathbf{X}_{(k)}\| \le \rho]} \mathbb{1}_{[\|\mathbf{Y}_{(k)}\| \le \rho]} \\
&\quad \le \frac{1}{k} \left(\sum_{i=1}^n \mathbb{1}_{[\mathbf{X}_i \neq \mathbf{Y}_i]} \sup_{\mathbf{x},\mathbf{y} \in B(\mathbf{0},\rho)} |r(\mathbf{x}) - r(\mathbf{y})| \right) \mathbb{1}_{[\|\mathbf{X}_{(k)}\| \le \rho]} \mathbb{1}_{[\|\mathbf{Y}_{(k)}\| \le \rho]}
\end{aligned}$$

$$\leq \frac{2}{k}\sum_{i=1}^{n} \mathbb{1}_{[\mathbf{X}_i \neq \mathbf{Y}_i]} \sup_{\mathbf{x}\in B(\mathbf{0},\rho)} |r(\mathbf{x}) - r(\mathbf{0})|$$
$$= \frac{\mathrm{O}(\rho)}{k} \times \mathrm{O}_{\mathbb{P}}\left(n\int_{\mathbb{R}^d} \left|g_\rho(\mathbf{x}) - f(\mathbf{x})\right| \mathrm{d}\mathbf{x}\right)$$

since $\sum_{i=1}^{n} \mathbb{1}_{[\mathbf{X}_i \neq \mathbf{Y}_i]}$ is $\mathrm{Bin}(n, \frac{1}{2}\int_{\mathbb{R}^d} |g_\rho(\mathbf{x}) - f(\mathbf{x})|\mathrm{d}\mathbf{x})$. Therefore,

$$|r_n(\mathbf{0}) - s_n(\mathbf{0})|\, \mathbb{1}_{[\|\mathbf{X}_{(k)}\|\leq\rho]}\mathbb{1}_{[\|\mathbf{Y}_{(k)}\|\leq\rho]} = \mathrm{O}_{\mathbb{P}}\left(\frac{\rho^{d+2}\epsilon(\rho)n}{k}\right) = \mathrm{o}_{\mathbb{P}}\left(\left(\frac{k}{n}\right)^{2/d}\right).$$

Finally,

$$T_n \stackrel{\text{def}}{=} |r_n(\mathbf{0}) - s_n(\mathbf{0})|\,(\mathbb{1}_{[\|\mathbf{X}_{(k)}\|>\rho]} + \mathbb{1}_{[\|\mathbf{Y}_{(k)}\|>\rho]})$$
$$\leq 2\|r\|_\infty(\mathbb{1}_{[\|\mathbf{X}_{(k)}\|>\rho]} + \mathbb{1}_{[\|\mathbf{Y}_{(k)}\|>\rho]}).$$

By Chernoff's bound (Theorem 20.5), for some constants $c, c' > 0$,

$$\mathbb{P}\{\|\mathbf{X}_{(k)}\| > \rho\} = \mathbb{P}\{\mathrm{Bin}(n,p) < k\}$$
$$\leq e^{-cnp} \leq e^{-c'n\rho^d} = e^{-c'K^d k}$$

if $k \leq np/2$ (which is the case for K large enough). A similar argument applies to $\|\mathbf{Y}_{(k)}\|$, and therefore, $\mathbb{E}T_n \leq e^{-c''k}$ for some positive constant c''. By Markov's inequality,

$$T_n = \mathrm{O}_{\mathbb{P}}\big(e^{-c'''k}\big)$$

for another constant $c''' > 0$. Thus,

$$|r_n(\mathbf{0}) - s_n(\mathbf{0})| = \mathrm{o}_{\mathbb{P}}\left(\left(\frac{k}{n}\right)^{2/d}\right) + \mathrm{O}_{\mathbb{P}}\big(e^{-C'''k}\big) = \mathrm{o}_{\mathbb{P}}\left(\left(\frac{k}{n}\right)^{2/d}\right),$$

since $k/\log n \to \infty$.

Proof of (15.7). We have

$$\mathbf{III} \leq \|\mathbf{Y}_{(k)}\|^2 \sup_{\mathbf{y}\in B(\mathbf{0},\|\mathbf{Y}_{(k)}\|)} w(\mathbf{y}).$$

The result follows from

$$\|\mathbf{Y}_{(k)}\| = \mathrm{O}_{\mathbb{P}}\left(\left(\frac{k}{n}\right)^{1/d}\right).$$

Study of I and II. We describe a random variable distributed as $\mathbf{Y}$. Let U, W be i.i.d. uniform $[0,1]$ random variables, and let $\mathbf{Z}$ be uniformly distributed on $\{\mathbf{z} \in \mathbb{R}^d : \|\mathbf{z}\| = 1\}$, the surface of $B(\mathbf{0},1)$, where $\mathbf{Z}$ is independent of U and W. Define

$$\mathbf{Y} = \begin{cases} \text{distributed as } \mathbf{X} \text{ conditional on } \|\mathbf{X}\| > \rho, & \text{if } U > f(\mathbf{0})V_d\rho^d \\ \left(\frac{U}{f(\mathbf{0})V_d}\right)^{1/d}\mathbf{Z}, & \text{if } U \le f(\mathbf{0})V_d\rho^d \text{ and } W > \left(\frac{-f'(\mathbf{0})^\top\left(\frac{U}{f(\mathbf{0})V_d}\right)^{1/d}\mathbf{Z}}{f(\mathbf{0})}\right)^+ \\ -\left(\frac{U}{f(\mathbf{0})V_d}\right)^{1/d}\mathbf{Z}, & \text{if } U \le f(\mathbf{0})V_d\rho^d \text{ and } W \le \left(\frac{-f'(\mathbf{0})^\top\left(\frac{U}{f(\mathbf{0})V_d}\right)^{1/d}\mathbf{Z}}{f(\mathbf{0})}\right)^+, \end{cases}$$

and the coupled random variable

$$\mathbf{Y}^\star = \begin{cases} \text{distributed as } \mathbf{X} \text{ conditional on } \|\mathbf{X}\| > \rho, & \text{if } U > f(\mathbf{0})V_d\rho^d \\ \left(\frac{U}{f(\mathbf{0})V_d}\right)^{1/d}\mathbf{Z}, & \text{otherwise.} \end{cases}$$

But, for $\mathbf{x} \in B(\mathbf{0},\rho)$,

$$\mathbb{P}\{\mathbf{Y} = \mathbf{x} \,|\, \mathbf{Y}^\star = \mathbf{x}\} = 1 - \left(\frac{-f'(\mathbf{0})^\top\mathbf{x}}{f(\mathbf{0})}\right)^+,$$

and

$$\mathbb{P}\{\mathbf{Y} = \mathbf{x} \,|\, \mathbf{Y}^\star = -\mathbf{x}\} = \left(\frac{f'(\mathbf{0})^\top\mathbf{x}}{f(\mathbf{0})}\right)^+.$$

Therefore, the density of $\mathbf{Y}$ on $B(\mathbf{0},\rho)$ is

$$f(\mathbf{0})\left(1 - \left(\frac{-f'(\mathbf{0})^\top\mathbf{x}}{f(\mathbf{0})}\right)^+ + \left(\frac{f'(\mathbf{0})^\top\mathbf{x}}{f(\mathbf{0})}\right)^+\right) = f(\mathbf{0}) + f'(\mathbf{0})^\top\mathbf{x}.$$

Consider first a quadratic form of $\mathbf{Y}$ for a $d \times d$ matrix A. We have, for $\|\mathbf{Y}\| \le \rho$,

$$\begin{aligned}\mathbf{Y}^\top A\mathbf{Y} &= \left(\frac{U}{f(\mathbf{0})V_d}\right)^{2/d}\mathbf{Z}^\top A\mathbf{Z}\left(1 - 2\mathbb{1}_{\left[W \le \left(\frac{-f'(\mathbf{0})^\top\left(\frac{U}{f(\mathbf{0})V_d}\right)^{1/d}\mathbf{Z}}{f(\mathbf{0})}\right)^+\right]}\right)^2 \\ &= \left(\frac{U}{f(\mathbf{0})V_d}\right)^{2/d}\mathbf{Z}^\top A\mathbf{Z}.\end{aligned}$$

In other words, the bias introduced by the "W trick" cancels out. For a linear form with vector $\mathbf{a} \in \mathbb{R}^d$, for $\|\mathbf{Y}\| \le \rho$,

$$\mathbf{a}^\top\mathbf{Y} = \left(\frac{U}{f(\mathbf{0})V_d}\right)^{1/d}\mathbf{a}^\top\mathbf{Z} - 2\left(\frac{U}{f(\mathbf{0})V_d}\right)^{1/d}\mathbf{a}^\top\mathbf{Z}\mathbb{1}_{\left[W \le \left(\frac{-f'(\mathbf{0})^\top\left(\frac{U}{f(\mathbf{0})V_d}\right)^{1/d}\mathbf{Z}}{f(\mathbf{0})}\right)^+\right]}.$$

In preparation for the finale, we note that if $\mathbf{Z} = (Z_1, \ldots, Z_d)^\top$,

$$\mathbb{E}[\mathbf{Z}^\top A \mathbf{Z}] = \sum_{j=1}^{d} A_{jj} \mathbb{E} Z_j^2 = \frac{\mathbb{E}[\mathbf{Z}^\top \mathbf{Z}]}{d} \times \mathrm{tr}(A) = \frac{\mathrm{tr}(A)}{d},$$

where the elements of A are denoted by $A_{jj'}$. Here we used the fact that

$$\mathbb{E} Z_1^2 = \cdots = \mathbb{E} Z_d^2 = \frac{\mathbb{E}[Z_1^2 + \cdots + Z_d^2]}{d} = \frac{1}{d}.$$

Also note that $|\mathbf{Z}^\top A \mathbf{Z}| \le d^2 \max_{j,j'} |A_{jj'}|$. Next, for $\mathbf{a} = (a_1, \ldots, a_d)^\top$, $\mathbb{E}[\mathbf{a}^\top \mathbf{Z}] = 0$, and

$$\mathbb{E}\big[(\mathbf{a}^\top \mathbf{Z})^2\big] = \mathbb{E}[\mathbf{a}^\top \mathbf{Z}\mathbf{Z}^\top \mathbf{a}] = \mathbf{a}^\top \mathbb{E}[\mathbf{Z}\mathbf{Z}^\top]\mathbf{a} = \frac{\mathbf{a}^\top \mathbf{a}}{d} = \frac{\|\mathbf{a}\|^2}{d}.$$

Finally, if $\|f'(\mathbf{0})\| \big(\frac{U}{f(\mathbf{0})V_d}\big)^{1/d} \le f(\mathbf{0})$,

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{a}^\top \mathbf{Z} \mathbb{1}_{\left[W \le \left(\frac{-f'(\mathbf{0})^\top \left(\frac{U}{f(\mathbf{0})V_d} \right)^{1/d} \mathbf{Z}}{f(\mathbf{0})} \right)^+ \right]} \,\middle|\, U \right] \\
&\quad = \mathbb{E}\left[\mathbf{a}^\top \mathbf{Z} \left(\frac{-f'(\mathbf{0})^\top \left(\frac{U}{f(\mathbf{0})V_d} \right)^{1/d} \mathbf{Z}}{f(\mathbf{0})} \right)^+ \,\middle|\, U \right] \\
&\quad = \frac{-\left(\frac{U}{f(\mathbf{0})V_d} \right)^{1/d}}{f(\mathbf{0})} \times \mathbb{E}\left[\mathbf{a}^\top \mathbf{Z} f'(\mathbf{0})^\top \mathbf{Z} \mathbb{1}_{[f'(\mathbf{0})^\top \mathbf{Z} < 0]} \right] \\
&\quad = \frac{-\left(\frac{U}{f(\mathbf{0})V_d} \right)^{1/d}}{2f(\mathbf{0})} \times \mathbb{E}\left[\mathbf{a}^\top \mathbf{Z} f'(\mathbf{0})^\top \mathbf{Z} \right].
\end{aligned}$$

Therefore, if $\|f'(\mathbf{0})\| \big(\frac{U}{f(\mathbf{0})V_d}\big)^{1/d} \le f(\mathbf{0})$,

$$\begin{aligned}
\mathbb{E}\left[\mathbf{a}^\top \mathbf{Z} \mathbb{1}_{\left[W \le \left(\frac{-f'(\mathbf{0})^\top \left(\frac{U}{f(\mathbf{0})V_d} \right)^{1/d} \mathbf{Z}}{f(\mathbf{0})} \right)^+ \right]} \,\middle|\, U \right] &= \frac{-\left(\frac{U}{f(\mathbf{0})V_d} \right)^{1/d}}{2f(\mathbf{0})} \times \mathbb{E}\left[\mathbf{a}^\top \mathbf{Z}\mathbf{Z}^\top f'(\mathbf{0}) \right] \\
&= \frac{-\left(\frac{U}{f(\mathbf{0})V_d} \right)^{1/d}}{2df(\mathbf{0})} \times \mathbf{a}^\top f'(\mathbf{0}). \qquad (15.8)
\end{aligned}$$

A last coupling will tame the analysis into submission. We consider a sample of i.i.d. random variables $(U_1, W_1, \mathbf{Z}_1), \ldots, (U_n, W_n, \mathbf{Z}_n)$ that are distributed as $(U, W, \mathbf{Z})$ in the definition of $\mathbf{Y}$. Define

$$\mathbf{Y}^{\star\star} = \begin{cases} \left(\frac{U}{f(\mathbf{0})V_d}\right)^{1/d} \mathbf{Z} & \text{if } W > \left(\frac{-f'(\mathbf{0})^\top \left(\frac{U}{f(\mathbf{0})V_d}\right)^{1/d} \mathbf{Z}}{f(\mathbf{0})}\right)^+ \\ -\left(\frac{U}{f(\mathbf{0})V_d}\right)^{1/d} \mathbf{Z} & \text{otherwise.} \end{cases}$$

Couple $\mathbf{Y}$ and $\mathbf{Y}^{\star\star}$ by using the same values of $(U, W, \mathbf{Z})$. Draw i.i.d. samples $(\mathbf{Y}_1, \mathbf{Y}_1^{\star\star}), \ldots, (\mathbf{Y}_n, \mathbf{Y}_n^{\star\star})$ from this coupled distribution, and look at

$$s_n(\mathbf{0}) = \frac{1}{k} \sum_{i=1}^{k} r(\mathbf{Y}_{(i)}) \quad \text{and} \quad s_n^{\star\star}(\mathbf{0}) = \frac{1}{k} \sum_{i=1}^{k} r(\mathbf{Y}_{(i)}^{\star\star}),$$

where, as usual, $\mathbf{Y}_{(1)}, \ldots, \mathbf{Y}_{(n)}$ and $\mathbf{Y}_{(1)}^{\star\star}, \ldots, \mathbf{Y}_{(n)}^{\star\star}$ are reordered by increasing values of $\|\mathbf{Y}_i\|$, respectively $\|\mathbf{Y}_i^{\star\star}\|$. In this case, this means that the $U_{(i)}$ values are increasing. In particular, if $(\frac{U_{(k)}}{f(\mathbf{0})V_d})^{1/d} \le \rho$, we have

$$(\mathbf{Y}_{(1)}, \ldots, \mathbf{Y}_{(k)}) = (\mathbf{Y}_{(1)}^{\star\star}, \ldots, \mathbf{Y}_{(k)}^{\star\star}),$$

and therefore,

$$\mathbb{P}\left\{s_n(\mathbf{0}) \neq s_n^{\star\star}(\mathbf{0})\right\} \le \mathbb{P}\{U_{(k)} > f(\mathbf{0})V_d\rho^d\}. \tag{15.9}$$

In the definition of $\rho = K(k/n)^{1/d}$, using Chernoff's bound we can choose K so large that the probability (15.9) is smaller than e^{-ck} for some constant $c > 0$. Thus, we only need to look at $s_n^{\star\star}(\mathbf{0})$.

We have, on $[s_n(\mathbf{0}) = s_n^{\star\star}(\mathbf{0})]$,

$$\begin{aligned} \mathbf{I} + \mathbf{II} &= r'(\mathbf{0})^\top \frac{1}{k} \sum_{i=1}^{k} \mathbf{Y}_{(i)}^{\star\star} + \frac{1}{2k} \sum_{i=1}^{k} \mathbf{Y}_{(i)}^{\star\star T} r''(\mathbf{0}) \mathbf{Y}_{(i)}^{\star\star} \\ &= \frac{1}{k} \sum_{i=1}^{k} \left(\frac{U_{(i)}}{V_d f(\mathbf{0})}\right)^{1/d} r'(\mathbf{0})^\top \mathbf{Z}_{(i)} + \frac{1}{2k} \sum_{i=1}^{k} \left(\frac{U_{(i)}}{V_d f(\mathbf{0})}\right)^{2/d} \mathbf{Z}_{(i)}^\top r''(\mathbf{0}) \mathbf{Z}_{(i)} \\ &\quad - \frac{2}{k} \sum_{i=1}^{k} \left(\frac{U_{(i)}}{V_d f(\mathbf{0})}\right)^{1/d} r'(\mathbf{0})^\top \mathbf{Z}_{(i)} \mathbb{1}_{\left[W_{(i)} \le \left(\frac{-f'(\mathbf{0})^\top \left(\frac{U_{(i)}}{f(\mathbf{0})V_d}\right)^{1/d} \mathbf{Z}_{(i)}}{f(\mathbf{0})}\right)^+\right]} \\ &\stackrel{\text{def}}{=} \mathbf{T}_1 + \mathbf{T}_2 + \mathbf{T}_3. \end{aligned}$$

By Theorem 13.3, if

$$\sigma^2 = \mathbb{E}\left[\left(r'(\mathbf{0})^\top \mathbf{Z}\right)^2\right] = \frac{\|r'(\mathbf{0})\|^2}{d},$$

then

$$\begin{aligned}\mathrm{T}_1 &\stackrel{\mathscr{D}}{=} \frac{1}{\sqrt{k}}\left(\frac{k}{n}\right)^{1/d} \frac{\sigma}{\sqrt{1+\frac{2}{d}}} \times \frac{1}{(V_d f(\mathbf{0}))^{1/d}} \left(N + \mathrm{o}_{\mathbb{P}}(1)\right) \\ &= \frac{1}{\sqrt{k}}\left(\frac{k}{n}\right)^{1/d} \frac{\|r'(\mathbf{0})\|}{\sqrt{d+2}} \times \frac{1}{(V_d f(\mathbf{0}))^{1/d}} \left(N + \mathrm{o}_{\mathbb{P}}(1)\right).\end{aligned}$$

By Theorem 13.2 , with $\mu = \mathbb{E}[\mathbf{Z}^\top r''(\mathbf{0})\mathbf{Z}] = \frac{\mathrm{tr}(r''(\mathbf{0}))}{d}$,

$$\begin{aligned}\mathrm{T}_2 &= \left(\frac{k}{n}\right)^{2/d} \frac{\mu}{2(1+\frac{2}{d})} \times \frac{1}{(V_d f(\mathbf{0}))^{2/d}} \left(1 + \mathrm{o}_{\mathbb{P}}(1)\right) \\ &= \left(\frac{k}{n}\right)^{2/d} \frac{\mathrm{tr}(r''(\mathbf{0}))}{2(d+2)} \times \frac{1}{(V_d f(\mathbf{0}))^{2/d}} \left(1 + \mathrm{o}_{\mathbb{P}}(1)\right).\end{aligned}$$

Again by Theorem 13.2, using (15.8), using a small additional argument, and putting all the pieces together,

$$\begin{aligned}\mathrm{T}_3 &= \left(\frac{k}{n}\right)^{2/d} \frac{r'(\mathbf{0})^\top f'(\mathbf{0})}{d f(\mathbf{0})} \times \frac{1}{1+\frac{2}{d}} \times \frac{1}{(V_d f(\mathbf{0}))^{2/d}} \left(1 + \mathrm{o}_{\mathbb{P}}(1)\right) \\ &= \left(\frac{k}{n}\right)^{2/d} \frac{r'(\mathbf{0})^\top f'(\mathbf{0})}{(d+2) f(\mathbf{0})} \times \frac{1}{(V_d f(\mathbf{0}))^{2/d}} \left(1 + \mathrm{o}_{\mathbb{P}}(1)\right).\end{aligned}$$

This finishes the proof of the theorem. □

Remark 15.3. In the noisy case, we have

$$\begin{aligned}r_n(\mathbf{0}) - r(\mathbf{0}) &= \frac{1}{k}\sum_{i=1}^{k} Y_{(i)} - r(\mathbf{0}) \\ &= \frac{1}{k}\sum_{i=1}^{k} \left(Y_{(i)} - r(\mathbf{X}_{(i)})\right) + \frac{1}{k}\sum_{i=1}^{k} r(\mathbf{X}_{(i)}) - r(\mathbf{0}), \qquad (15.10)\end{aligned}$$

where now $Y_{(1)}, \ldots, Y_{(n)}$ are as in Chapter 14. We already covered the last term of (15.10) in Theorem 15.3. By virtue of Theorem 14.2, the first term on the right in (15.10) is asymptotically distributed as

$$\frac{\sigma(\mathbf{0})}{\sqrt{k}} \left(N + \mathrm{o}_{\mathbb{P}}(1)\right), \qquad (15.11)$$

where $\sigma(\mathbf{x}) = \mathbb{V}[Y|\mathbf{X} = \mathbf{x}]$. The only conditions needed for this are that $\sigma(\mathbf{0}) > 0$, that Y is almost surely bounded (a condition that can be relaxed), $k \to \infty$, $k/n \to 0$, and that σ is continuous at $\mathbf{0}$. Adding (15.11) to the expression in Theorem 15.3 makes the $\frac{1}{\sqrt{k}}(\frac{k}{n})^{1/d}$ term asymptotically irrelevant. In other words, we basically rediscover Corollary 14.1. The readers have two very different proofs of the same result—one more analytic and classical, and a coupling proof that explains each term in the asymptotic behavior. □

Chapter 16
The choice of a nearest neighbor estimate

16.1 Parameter selection

Selecting the estimate within a class of estimates that is optimal in a certain sense is perhaps the ultimate goal of nonparametric estimation. It assumes that the class of estimates is sufficiently rich within the universe of all possible estimates. That the nearest neighbor regression function estimate is rich as a class follows not only from the universality, but also from the fact that it achieves rates of convergence for various criteria that are close to the best possible over certain classes of distributions on $(\mathbf{X}, Y)$, a property that is studied in minimax theory (Stone, 1980, 1982).

In this chapter, we take a class of nearest neighbor estimates. Examples include:

(*i*) **The k-nearest neighbor estimate**:

$$v_{ni} = \begin{cases} \frac{1}{k} & \text{for } 1 \le i \le k \\ 0 & \text{otherwise.} \end{cases}$$

This class is parametrized by $k \in \{1, \ldots, n\}$.

(*ii*) **The monotone weight estimate**:

$$v_{n1} \ge v_{n2} \ge \cdots \ge v_{nn} \ge 0, \quad \sum_{i=1}^{n} v_{ni} = 1.$$

(*iii*) **The discretized weight estimate**:

$$v_{ni} = \frac{\alpha(i,n)}{\beta(n)}, \quad 1 \le i \le n,$$

G. Biau, L. Devroye, *Lectures on the Nearest Neighbor Method*,
Springer Series in the Data Sciences, DOI 10.1007/978-3-319-25388-6_16

where $\beta(n)$ is a fixed positive integer sequence, and $\alpha(i,n)$ is an integer between $-B\beta(n)$ and $B\beta(n)$, with $B > 0$ another fixed integer. Also, the requirement $\sum_{i=1}^n v_{ni} = 1$ implies that $\sum_{i=1}^n \alpha(i,n) = \beta(n)$. This class of estimates is discrete and has not more than $(2B\beta(n)+1)^n$ members.

(*iv*) **The kernel-weighted estimate**: let K be a fixed real-valued function defined on $[0,1]$ with $\int_0^1 K(x)\mathrm{d}x = 1$, which we call the kernel; the family of estimates is parametrized by $k \in \{1,\ldots,n\}$, and uses

$$v_{ni} = \begin{cases} \displaystyle\int_{(i-1)/k}^{i/k} K(x)\mathrm{d}x & \text{for } 1 \le i \le k \\ 0 & \text{otherwise,} \end{cases}$$

so that, automatically, $\sum_{i=1}^n v_{ni} = \int_0^1 K(x)\mathrm{d}x = 1$.

Taking these classes as examples, we will set out to study data-dependent choices of the best estimates within each class.

16.2 Oracle inequality

The first simplification comes from data splitting, a technique that affords us valuable independence for making our choices. In particular, we assume that the (training) data are $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$, as before, and that we have test data $(\mathbf{X}'_1, Y'_1), \ldots, (\mathbf{X}'_m, Y'_m)$, which are independent of the data and consist of independent pairs all distributed as $(\mathbf{X}_1, Y_1)$. In many cases—illustrated below—one can get away with the choice $m = \mathrm{o}(n)$, requiring only $m = \Theta(n^\alpha)$ for some $0 < \alpha < 1$.

Now, we need to settle on the criterion. The most natural one in the regression function setting is the L^2 criterion

$$\mathbb{E}\,|r_n(\mathbf{X}) - r(\mathbf{X})|^2\,, \tag{16.1}$$

where, as usual, r_n denotes the estimate of r and $(\mathbf{X}, Y)$ denotes an independent pair, distributed as $(\mathbf{X}_1, Y_1)$. We recall (Chapter 10) that

$$\mathbb{E}\,|Y - r_n(\mathbf{X})|^2 = \mathbb{E}\,|r_n(\mathbf{X}) - r(\mathbf{X})|^2 + \mathbb{E}\,|Y - r(\mathbf{X})|^2\,, \tag{16.2}$$

where the last term does not depend upon the estimate or the data. Thus, minimizing (16.1) is equivalent to minimizing (16.2) and even (16.3):

$$\mathbb{E}[r_n^2(\mathbf{X}) - 2r_n(\mathbf{X})Y] = \mathbb{E}\,|Y - r_n(\mathbf{X})|^2 - \mathbb{E}Y^2. \tag{16.3}$$

Assume that the estimates are parametrized by a parameter $\theta \in \mathscr{A}$, with $|\mathscr{A}| < \infty$. For example, for the k-nearest neighbor estimate, $\theta = k$ and $\mathscr{A} = \{1, \ldots, n\}$. We make the dependence of $r_n(\mathbf{x})$ on θ explicit when we write $r_{n,\theta}(\mathbf{x})$. Define

$$L_{n,\theta} = \mathbb{E}\left[|r_{n,\theta}(\mathbf{X}) - r(\mathbf{X})|^2 \,\middle|\, \mathbf{X}_1, Y_1, \ldots, \mathbf{X}_n, Y_n \right], \tag{16.4}$$

and

$$L_n^\star = \min_{\theta \in \mathscr{A}} L_{n,\theta}, \quad \theta^\star \in \arg\min_{\theta \in \mathscr{A}} L_{n,\theta}.$$

Thus, $\theta^\star$ is the best parameter for the given data. Let the estimate of (16.4) based on the test data be

$$\hat{L}_{n,\theta} = \frac{1}{m} \sum_{j=1}^{m} \left(r_{n,\theta}^2(\mathbf{X}_j') - 2 r_{n,\theta}(\mathbf{X}_j') Y_j' \right). \tag{16.5}$$

We choose $\hat{\theta}^\star$ so as to minimize $\hat{L}_{n,\theta}$:

$$\hat{L}_{n,\hat{\theta}^\star} = \min_{\theta \in \mathscr{A}} \hat{L}_{n,\theta}, \quad \hat{\theta}^\star \in \arg\min_{\theta \in \mathscr{A}} \hat{L}_{n,\theta}.$$

Ties are broken in a canonical manner. Since $\mathscr{A}$ is finite, we can map $\mathscr{A}$ to the integers $\{1, \ldots, |\mathscr{A}|\}$, and break ties by choosing the parameter of smallest integer value. Therefore, "arg min" returns a unique $\theta \in \mathscr{A}$. It should also be noted that $\hat{\theta}^\star$ depends on both the training data and the test data.

What matters to us is the difference $L_{n,\hat{\theta}^\star} - L_n^\star$. In fact, the relative difference is of primary interest. This leads one naturally to oracle inequalities, such as the one presented in the next theorem.

Theorem 16.1. *Assume that* $1 < |\mathscr{A}| < \infty$ *and* $\|Y\|_\infty < \infty$. *Then*

$$\mathbb{E}[L_{n,\hat{\theta}^\star} \,|\, \mathbf{X}_1, Y_1, \ldots, \mathbf{X}_n, Y_n] \le L_n^\star + \frac{M}{\sqrt{m}} \left(\sqrt{2 \log |\mathscr{A}|} + \frac{1}{\sqrt{2 \log |\mathscr{A}|}} \right),$$

where

$$M = \|Y\|_\infty^2 \left(1 + \sum_{i=1}^{n} |v_{ni}| \right)^2.$$

For a related discussion, see Györfi et al. (2002, Theorem 7.1). For classification—a special case of regression—data splitting was similarly analyzed by Devroye (1988).

Proof (Theorem 16.1). Let $\mathscr{D}_n = ((\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n))$. Using the independence between $(\mathbf{X}, Y)$ and $\mathscr{D}_n$, it is easy to see that

$$L_{n,\theta} = \mathbb{E}\left[r_{n,\theta}^2(\mathbf{X}) - 2r_{n,\theta}(\mathbf{X})Y + r^2(\mathbf{X}) \mid \mathscr{D}_n\right].$$

We apply Hoeffding's inequality (Theorem 20.7 in the Appendix) to the sum

$$\begin{aligned}
&\frac{1}{m}\sum_{j=1}^{m}\Big(\left[r_{n,\theta}^2(\mathbf{X}_j') - 2r_{n,\theta}(\mathbf{X}_j')Y_j' + r^2(\mathbf{X}_j')\right] \\
&\qquad\qquad - \mathbb{E}\left[r_{n,\theta}^2(\mathbf{X}) - 2r_{n,\theta}(\mathbf{X})Y + r^2(\mathbf{X}) \mid \mathscr{D}_n\right]\Big) \\
&= \tilde{L}_{n,\theta} - L_{n,\theta},
\end{aligned}$$

where $\tilde{L}_{n,\theta} = \hat{L}_{n,\theta} + \frac{1}{m}\sum_{j=1}^{m} r^2(\mathbf{X}_j')$. To do so, note that $|r_{n,\theta}(\mathbf{x})| \le \|Y\|_\infty \sum_{i=1}^{n} |v_{ni}|$, and thus that

$$\begin{aligned}
\left|r_{n,\theta}^2(\mathbf{X}) - 2r_{n,\theta}(\mathbf{X})Y + r^2(\mathbf{X})\right| &\le \|Y\|_\infty^2\left(\left(\sum_{i=1}^{n}|v_{ni}|\right)^2 + 2\sum_{i=1}^{n}|v_{ni}| + 1\right) \\
&= \|Y\|_\infty^2\left(1 + \sum_{i=1}^{n}|v_{ni}|\right)^2 \\
&\stackrel{\text{def}}{=} M.
\end{aligned}$$

Thus, by Hoeffding's inequality, for $t > 0$,

$$\mathbb{P}\{L_{n,\theta} - \tilde{L}_{n,\theta} \ge t \mid \mathscr{D}_n\} \le \exp\left(-\frac{mt^2}{2M^2}\right). \tag{16.6}$$

Since $\mathbb{E}[\tilde{L}_{n,\theta^\star} \mid \mathscr{D}_n] = L_n^\star$, we only need to show that

$$\mathbb{E}[L_{n,\hat{\theta}^\star} - \tilde{L}_{n,\hat{\theta}^\star} \mid \mathscr{D}_n] \le \frac{M}{\sqrt{m}}\left(\sqrt{2\log|\mathscr{A}|} + \frac{1}{\sqrt{2\log|\mathscr{A}|}}\right). \tag{16.7}$$

Fix $t > 0$. Then, clearly, by (16.6),

$$\begin{aligned}
\mathbb{P}\{L_{n,\hat{\theta}^\star} - \tilde{L}_{n,\hat{\theta}^\star} \ge t \mid \mathscr{D}_n\} &\le \mathbb{P}\left\{\max_{\theta\in\mathscr{A}}(L_{n,\theta} - \tilde{L}_{n,\theta}) \ge t \mid \mathscr{D}_n\right\} \\
&\le \sum_{\theta\in\mathscr{A}} \exp\left(-\frac{mt^2}{2M^2}\right).
\end{aligned}$$

Setting $\delta^\star = \sqrt{\frac{2M^2}{m}\log|\mathscr{A}|}$, it follows that

$$\begin{aligned}\mathbb{E}[L_{n,\hat{\theta}^\star} - \tilde{L}_{n,\hat{\theta}^\star} \mid \mathscr{D}_n] &\le \int_0^\infty \mathbb{P}\{L_{n,\hat{\theta}^\star} - \tilde{L}_{n,\hat{\theta}^\star} \ge t \mid \mathscr{D}_n\}\mathrm{d}t \\ &\le \delta^\star + \int_{\delta^\star}^\infty |\mathscr{A}| \exp\left(-\frac{mt^2}{2M^2}\right)\mathrm{d}t \\ &\le \delta^\star + \frac{M^2}{m\delta^\star} \\ &= \frac{M}{\sqrt{m}}\left(\sqrt{2\log|\mathscr{A}|} + \frac{1}{\sqrt{2\log|\mathscr{A}|}}\right).\end{aligned}$$

Thus, (16.7) is verified, and the theorem is proved. □

16.3 Examples

The oracle inequality of Theorem 16.1 permits the user to obtain asymptotically optimal estimates. For example, if r is Lipschitz and $\mathbf{X}$ and Y are both bounded random variables in $\mathbb{R}^d$ ($d \ge 2$) and $\mathbb{R}$, respectively, then any k-nearest neighbor regression estimate $\mathbb{E}|r_n(\mathbf{X}) - r(\mathbf{X})|^2$ can be made to converge to zero at the rate $\mathrm{O}(n^{-\frac{2}{d+2}})$ (Corollary 14.2) if we know that $(\mathbf{X}, Y)$ satisfies these general conditions. This rate of convergence is best possible within that class of estimates and distributions (Stone, 1982).

If we select $k \in \{1, \ldots, n\}$ (so, $|\mathscr{A}| = n$) by the data splitting method, then Theorem 16.1 above guarantees that if we take

$$\frac{\log n}{m} = \mathrm{o}\big(n^{-\frac{4}{d+2}}\big)$$

(i.e., $m \gg n^{\frac{4}{d+2}}\log n$), then

$$\mathbb{E}[L_{n,\hat{\theta}^\star} \mid \mathbf{X}_1, Y_1, \ldots, \mathbf{X}_n, Y_n] \le L_n^\star\,(1 + \mathrm{o}(1))\,,$$

and therefore do we not only have an error that is $\mathrm{O}(n^{-\frac{2}{d+2}})$, but we even have the correct asymptotic coefficient in the rate of convergence, i.e.,

$$\frac{L_{n,\hat{\theta}^\star}}{L_n^\star} \to 1 \quad \text{in probability.}$$

In addition, this optimality property carries over to many other classes of distributions.

Györfi et al. (2002, Chapter 7) use Bernstein's inequality (Theorem 20.8 in the Appendix) instead of Hoeffding's to obtain an improved version of the oracle inequality of Theorem 16.1, for $\hat{\theta}^\star$ defined as in (16.5). For fixed $\delta > 0$, and

$$\xi(\delta) \stackrel{\text{def}}{=} \|Y\|_\infty^2 \Big(\frac{16}{\delta} + 35 + 19\delta \Big),$$

they show that

$$\mathbb{E}[L_{n,\hat{\theta}^\star} \,|\, \mathbf{X}_1, Y_1, \ldots, \mathbf{X}_n, Y_n] \le (1+\delta) L_n^\star + \frac{\xi(\delta)\,(1+\log|\mathscr{A}|)}{m}.$$

This inequality improves on Theorem 16.1 when $L_n^\star$ is small, but only if one can "guess" the approximate value of $L_n^\star$ so as to pick δ optimally.

There are other ways of selecting the best parameter within a class of parameters. The most celebrated among these is the leave-one-out or cross-validation method. Here $\mathbb{E}[r_n^2(\mathbf{X}) - 2r_n(\mathbf{X})Y]$ is estimated by using the data set itself. If $r_n(i, \mathbf{x})$ denotes $r_n(\mathbf{x})$ but with $(\mathbf{X}_i, Y_i)$ removed from the data, then one hopes that

$$\frac{1}{n} \sum_{i=1}^{n} \big(r_n^2(i, \mathbf{X}_i) - 2r_n(i, \mathbf{X}_i) Y_i \big)$$

is a good approximation of $\mathbb{E}[r_n^2(\mathbf{X}) - 2r_n(\mathbf{X})r(\mathbf{X})]$ $(= \mathbb{E}\,|r_n(\mathbf{X}) - r(\mathbf{X})|^2 - \mathbb{E}r^2(\mathbf{X}))$.

16.4 Feature selection

Feature selection, also known as variable selection, is the process of choosing relevant components of the vector $\mathbf{X}$ for use in model construction. There are many potential benefits of such an operation: facilitating data visualization and data understanding, reducing the measurement and storage requirements, decreasing training and utilization times, and defying the curse of dimensionality to improve prediction performance. In addition, it is often the case that finding a correct subset of variables is an important problem in its own right. For example, physicians may make a decision based on the selected features whether a dangerous surgery is necessary for treatment or not.

Feature selection has been an active research area in the statistics, machine learning, and data mining communities. Many attempts have been made to develop efficient algorithms for selecting the "best" (depending on the context) subset of components—for an overview of the problem, see, e.g., Guyon and Elisseeff, 2003, and the monograph by Hastie et al., 2009. General recipes are hard to give as the solutions depend on the specific problem, and some methods put more emphasis on one aspect than another. However, there are some rules of thumb that should be

followed. One such rule is that noisy measurements, that is, components that are independent of Y, should be avoided. Also, adding a component that is a function of other components is useless.

In selecting subsets of variables, a sensible objective is to make the mean integrated squared error $\mathbb{E}|r_n(\mathbf{X})-r(\mathbf{X})|^2$ as small as possible. This depends on many things, such as the joint distribution of the selected components of $\mathbf{X}$ and the answer Y, the sample size, and the regression estimate r_n itself. To make things a bit simpler, we first investigate the residual variance corresponding to the selected components. This approach makes sense since, in view of (16.2), the residual variance is the theoretical limit of the L^2 performance $\mathbb{E}|Y - r_n(\mathbf{X})|^2$ of any regression method.

In this data-independent context, we may formulate the feature selection problem as follows. Let $\mathbf{X} = (X_1, \ldots, X_d)$ represent d measurements. For a set $A \subseteq \{1, \ldots, d\}$ of indices, let $\mathbf{X}_A$ denote the $|A|$-dimensional random vector whose components are the X_j's with $j \in A$ (in the order of increasing indices). Define

$$L^\star(A) = \mathbb{E}\,|Y - \mathbb{E}[Y|\mathbf{X}_A]|^2\,,$$

that is, the minimum risk that can be achieved using the features in A as explanatory variables. Obviously, $L^\star(A) \le L^\star(B)$ whenever $B \subseteq A$, $L^\star(\{1, \ldots, d\}) = L^\star = \mathbb{E}|Y - r(\mathbf{X})|^2$, and $L^\star(\emptyset) = \mathbb{V}Y$. Thus, the problem is to find an efficient way of selecting an index set A with $|A| = p$, whose corresponding error is the smallest. Here $p \le d$ is a fixed integer. Exhaustive evaluation over all variable subsets of size p is often computationally prohibitive, as the number of subsets to be considered grows very rapidly with the number of features—for example, $\binom{12}{6}$ is 924, while $\binom{24}{12}$ is 2,704,156. A wide range of search strategies can be used, including best-first, branch and bound, simulated annealing, and genetic algorithms (see Kohavi and John, 1997, or Kumar and Minz, 2014, for reviews).

It is easy to see that the best p individual features—that is, components corresponding to the p smallest values of $L^\star(\{j\})$—do not necessarily constitute the best p-dimensional vector. Indeed, the following simple example shows that a combination of "good" single features may lead to a larger risk than a combination of "worse" features. Let $\mathbf{X} = (X_1, X_2, X_3)^\top$ be jointly Gaussian with nonsingular variance-covariance matrix Σ, and let $Y = \mathbf{a}^\top \mathbf{X}$ for some $\mathbf{a} \in \mathbb{R}^3$ to be chosen later. For $A \subseteq \{1, 2, 3\}$, we have by the elementary properties of the multivariate normal distribution

$$L^\star(A) = \mathbf{a}^\top \Sigma \mathbf{a} - \mathbf{a}^\top \Sigma P^\top (P \Sigma P^\top)^{-1} P \Sigma \mathbf{a},$$

where $P = P(A)$ consists of the rows of the 3×3 identity matrix with row labels in A. Take

$$\Sigma = \begin{pmatrix} 1 & -0.7 & 0 \\ -0.7 & 1 & -0.7 \\ 0 & -0.7 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{a} = \begin{pmatrix} 2 \\ 2.5 \\ 1 \end{pmatrix},$$

to obtain the following ordering of minimum risks:

$$L^\star(\{1\}) = \frac{11}{16} > L^\star(\{2\}) = \frac{59}{100} > L^\star(\{3\}) = \frac{3}{16}$$

and

$$L^\star(\{1,2\}) = \frac{2}{51} < L^\star(\{1,3\}) = \frac{1}{8} < L^\star(\{2,3\}) = \frac{8}{51}.$$

Thus, the (individually) best two components (X_2 and X_3) become the worst two-dimensional ones, whereas the worst two single components (X_1 and X_2) jointly represent the best feature pair.

Remark 16.1. Antos et al. (1999) (see also Devroye et al., 2003) proved that for any sequence $\{\phi_n\}$ of estimates and any sequence $\{a_n\}$ of positive real numbers converging to zero, there exists a distribution of $(\mathbf{X}, Y)$ such that $\mathbb{E}|\phi_n - L^\star| \geq a_n$ infinitely often. Thus, any estimate of $L^\star$ is doomed to converge arbitrarily slowly for some distribution of $(\mathbf{X}, Y)$, and no method can guarantee universally good performance. Error bounds or confidence bands for $L^\star$ can only be constructed under additional assumptions on the distribution of the data. □

The following theorem says that every feature selection algorithm that finds the best p-element subset has to search exhaustively through all $\binom{d}{p}$ subsets for some distributions—any other method is condemned to failure, no matter how many simulations are performed and no matter how large the sample sizes are.

Theorem 16.2. *For every* $1 \leq p \leq d$, *let* $\mathrm{rank}(A)$ *be the desired rank of* $A \subseteq \{1,\ldots,d\}$, $|A| = p$, *in the ordering of* $\{L^\star(A) : |A| = p\}$. *[Thus,* $1 \leq \mathrm{rank}(A) \leq \binom{d}{|A|}$*.] Then there exists a distribution of the random variable* $(\mathbf{X}, Y) \in \mathbb{R}^d \times \mathbb{R}$ *for which these ranks can be achieved.*

Proof. The distribution of $\mathbf{X}$ is uniform on the hypercube $\{-1,1\}^d$, while $Y = g(\mathbf{X})$, where

$$g(\mathbf{x}) = \sum_B \alpha_B \prod_{j\in B} x_j,$$

$B \subseteq \{1,\ldots,d\}$, and $\alpha_B \geq 0$ are given numbers to be determined later. We note first that if $\mathbf{x}_A = (x_{j_1},\ldots,x_{j_p})$ for $\mathbf{x} = (x_1,\ldots,x_d)$ and $A = \{j_1,\ldots,j_p\}$, then

$$\begin{aligned} g(\mathbf{x}_A) &\stackrel{\text{def}}{=} \int g(\mathbf{x}) \prod_{j\notin A} \mathrm{d}x_j \\ &= \mathbb{E}[g(\mathbf{X}) \,|\, \mathbf{X}_A = \mathbf{x}_A] \\ &= \sum_{B:B\subseteq A} \alpha_B \prod_{j\in B} x_j. \end{aligned}$$

Observe that if $A = \emptyset$, then $g(\mathbf{x}_\emptyset) = \mathbb{E}g(\mathbf{X}) = 0$, and that if $A = \{1, \ldots, d\}$, then $g(\mathbf{x}_A) = g(\mathbf{x})$. One can also see that

$$g(\mathbf{x}) - g(\mathbf{x}_A) = \sum_B \alpha_B \prod_{j \in B} x_j - \sum_{B: B \subseteq A} \alpha_B \prod_{j \in B} x_j,$$

and, since $\mathbb{E}[\prod_{j \in B} X_j \prod_{j \in B'} X_j] = \mathbb{1}_{[B=B']}$, that

$$\mathbb{E}g^2(\mathbf{X}) = \sum_B \alpha_B^2$$

and

$$\mathbb{E}g^2(\mathbf{X}_A) = \sum_{B: B \subseteq A} \alpha_B^2 = \mathbb{E}\left[g(\mathbf{X})g(\mathbf{X}_A)\right].$$

Therefore,

$$\begin{aligned} L^\star(A) = \mathbb{E}\left|g(\mathbf{X}) - g(\mathbf{X}_A)\right|^2 &= \sum_B \alpha_B^2 - \sum_{B: B \subseteq A} \alpha_B^2 \\ &\stackrel{\text{def}}{=} \sum_B \alpha_B^2 - \varphi(A). \end{aligned}$$

One can find values of $\{\alpha_A : A \subseteq \{1, \ldots, d\}\}$ that give an ordering of $L^\star(A)$ (and thus $\varphi(A)$) that is consistent with any given ordering within the subsets of equal cardinality. This can be done incrementally for all sets $\{A : |A| = p\}$ as p increases from 1 to d.

We explain the inductive step that fills in the values of α_A, $|A| = p$. Define

$$\lambda_{p-1} = 1 + \sum_{A: |A| < p} \alpha_A^2.$$

Then, for $|A| = p$, define

$$\alpha_A = \sqrt{\text{rank}(A) \times \lambda_{p-1}}.$$

To show that this suffices, take $|A| = |A'| = p$, with $\text{rank}(A) < \text{rank}(A')$. Then

$$\begin{aligned} \varphi(A) = \sum_{B: B \subseteq A} \alpha_B^2 &= \sum_{B: B \subsetneq A} \alpha_B^2 + \alpha_A^2 \\ &\leq \lambda_{p-1} - 1 + (\text{rank}(A) \times \lambda_{p-1}), \end{aligned}$$

while

$$\varphi(A') \geq \alpha_{A'}^2 = \text{rank}(A') \times \lambda_{p-1},$$

so that $\varphi(A) < \varphi(A')$ if and only if $\text{rank}(A) < \text{rank}(A')$. Finally, we verify the nesting property: if $|A| = p$, $|A'| < p$, then

$$\begin{aligned}\varphi(A) &\geq \text{rank}(A) \times \lambda_{p-1} \geq \lambda_{p-1} \\ &> \sum_{B:|B|<p} \alpha_B^2 \geq \varphi(A'). \quad \square\end{aligned}$$

The previous negative result parallels a similar negative result for pattern recognition by Cover and Van Campenhout (1977) (see also Devroye et al., 1996, Chapter 32), where $Y \in \{0, 1\}$,

$$L^\star = \inf_{g:\mathbb{R}^d \to \{0,1\}} \mathbb{E}\,|Y - g(\mathbf{X})|^2$$

and thus

$$L^\star(A) = \inf_{g_A:\mathbb{R}^{|A|} \to \{0,1\}} \mathbb{E}\left|Y - g_A(X_j : j \in A)\right|^2.$$

Of course, in practice, the real measure of the goodness of the selected feature set is the mean integrated squared error $\mathbb{E}|r_n(\mathbf{X}) - r(\mathbf{X})|^2$ of the regression estimate designed by using training data. If we know what estimate will be used after feature selection, then the best strategy is to select a set of coordinates based on comparing estimates of the error. This can typically be achieved by using the data splitting technique discussed in the previous sections. Assuming for example that r_n is the k-nearest neighbor estimate and that the set of candidate features is described by a collection $A_1, \ldots, A_q$ of subsets of $\{1, \ldots, d\}$, then we may simultaneously select both the best parameter k in $\mathscr{A}_1 = \{1, \ldots, n\}$ and the best component subset A in $\mathscr{A}_2 = \{A_1, \ldots, A_q\}$. To do this, we let $\theta = (k, A) \in \mathscr{A} = \mathscr{A}_1 \times \mathscr{A}_2$, and minimize in θ over $\mathscr{A}$

$$\hat{L}_{n,\theta} = \frac{1}{m} \sum_{j=1}^{m} \left(r_{n,\theta}^2(\mathbf{X}_j') - 2r_{n,\theta}(\mathbf{X}_j')Y_j'\right)$$

via the test data $(\mathbf{X}_1', Y_1'), \ldots, (\mathbf{X}_m', Y_m')$, where $r_{n,\theta}$ is the k-nearest neighbor estimate with parameter k in the space of variables described by the set A. If $\hat{\theta}_n^\star$ denotes the minimum over $\mathscr{A}$ of the above quantity, then, according to Theorem 16.1, we conclude that

$$\mathbb{E}[L_{n,\hat{\theta}^\star} \mid \mathbf{X}_1, Y_1, \ldots, \mathbf{X}_n, Y_n] \leq \min_{\theta=(k,A)} L_{n,\theta} + \mathrm{O}\left(\sqrt{\frac{\log(nq)}{m}}\right).$$

Part III
Supervised classification

Chapter 17
Basics of classification

17.1 Introduction

Supervised classification (also called pattern recognition, discrimination, or class prediction) is a specific regression problem, where the observation $\mathbf{X}$ takes values in $\mathbb{R}^d$ and the random response Y takes values in $\{0, 1\}$. Given $\mathbf{X}$, one has to guess the value of Y (also termed the label or class), and this guess is called a decision. Pattern recognition is important in different scientific disciplines, such as medicine, biology, finance, and meteorology. In medicine, for example, one needs to evaluate patients according to their disease risk, and the typical questions for classification are: "Is this person infected?," "Will this patient respond to the treatment?," or "Will this patient have serious side effects from using the drug?"—in all these cases, a yes/no or $0/1$ decision has to be made.

Mathematically, the decision is a Borel measurable function $g : \mathbb{R}^d \to \{0, 1\}$, called a classifier. An error occurs if $g(\mathbf{X}) \neq Y$, and the error probability for a classifier g is

$$L(g) = \mathbb{P}\{g(\mathbf{X}) \neq Y\}.$$

Of particular interest is the Bayes decision function

$$g^\star(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbb{P}\{Y = 1|\mathbf{X} = \mathbf{x}\} > \mathbb{P}\{Y = 0|\mathbf{X} = \mathbf{x}\} \\ 0 & \text{otherwise,} \end{cases}$$

which minimizes the error probability (ties are broken, by convention, in favor of class 0).

Lemma 17.1. *For any decision function $g : \mathbb{R}^d \to \mathbb{R}$, one has*

$$L(g^\star) \leq L(g),$$

that is, $g^\star$ is the optimal decision.

G. Biau, L. Devroye, *Lectures on the Nearest Neighbor Method*,
Springer Series in the Data Sciences, DOI 10.1007/978-3-319-25388-6_17

Proof. Let $g : \mathbb{R}^d \to \{0, 1\}$ be an arbitrary Borel measurable function. Then

$$\mathbb{P}\{g(\mathbf{X}) \neq Y\} = 1 - \mathbb{P}\{g(\mathbf{X}) = Y\}.$$

Thus,

$$\begin{aligned}\mathbb{P}\{g(\mathbf{X}) \neq Y\} - \mathbb{P}\{g^\star(\mathbf{X}) \neq Y\} &= \mathbb{P}\{g^\star(\mathbf{X}) = Y\} - \mathbb{P}\{g(\mathbf{X}) = Y\} \\ &= \mathbb{E}\left[\mathbb{P}\{g^\star(\mathbf{X}) = Y|\mathbf{X}\} - \mathbb{P}\{g(\mathbf{X}) = Y|\mathbf{X}\}\right] \\ &\geq 0,\end{aligned}$$

since, by definition of $g^\star$,

$$\begin{aligned}\mathbb{P}\{g^\star(\mathbf{X}) = Y|\mathbf{X}\} &= \mathbb{P}\{g^\star(\mathbf{X}) = 1, Y = 1|\mathbf{X}\} + \mathbb{P}\{g^\star(\mathbf{X}) = 0, Y = 0|\mathbf{X}\} \\ &= \mathbb{1}_{[g^\star(\mathbf{X})=1]}\mathbb{P}\{Y = 1|\mathbf{X}\} + \mathbb{1}_{[g^\star(\mathbf{X})=0]}\mathbb{P}\{Y = 0|\mathbf{X}\} \\ &= \max\left(\mathbb{P}\{Y = 0|\mathbf{X}\}, \mathbb{P}\{Y = 1|\mathbf{X}\}\right). \quad \square\end{aligned}$$

The error $L^\star \stackrel{\text{def}}{=} L(g^\star)$ is referred to as the Bayes probability of error (or Bayes error):

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

We stress that $L^\star = 0$ if and only if $Y = g^\star(\mathbf{X})$ with probability one, i.e., Y is a Borel measurable function of $\mathbf{X}$. In the design of classifiers, the probabilities $\mathbb{P}\{Y = 0|\mathbf{X} = \mathbf{x}\}$ and $\mathbb{P}\{Y = 1|\mathbf{X} = \mathbf{x}\}$ are called the a posteriori probabilities. Observe that

$$\mathbb{P}\{Y = 1|\mathbf{X} = \mathbf{x}\} = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}] = r(\mathbf{x}),$$

so that the Bayes decision function also takes the form

$$g^\star(\mathbf{x}) = \begin{cases} 1 & \text{if } r(\mathbf{x}) > 1/2 \\ 0 & \text{otherwise.} \end{cases} \tag{17.1}$$

Remark 17.1. Clearly,

$$\begin{aligned}L(g) = 1 - \mathbb{P}\{g(\mathbf{X}) = Y\} &= 1 - \mathbb{E}\left[\mathbb{P}\{g(\mathbf{X}) = Y|\mathbf{X}\}\right] \\ &= 1 - \mathbb{E}\left[\mathbb{1}_{[g(\mathbf{X})=1]}r(\mathbf{X}) + \mathbb{1}_{[g(\mathbf{X})=0]}\left(1 - r(\mathbf{X})\right)\right].\end{aligned}$$

Therefore,

$$L^\star = 1 - \mathbb{E}\left[\mathbb{1}_{[r(\mathbf{X})>1/2]}r(\mathbf{X}) + \mathbb{1}_{[r(\mathbf{X})\leq 1/2]}\left(1 - r(\mathbf{X})\right)\right].$$

This may be rewritten as

$$L^\star = \mathbb{E}\left[\min\left(r(\mathbf{X}), 1 - r(\mathbf{X})\right)\right] = \frac{1}{2} - \frac{1}{2}\mathbb{E}|2r(\mathbf{X}) - 1|.$$

In some special cases, we may obtain other helpful forms. For example, if $\mathbf{X}$ has a density f with respect to the Lebesgue measure on $\mathbb{R}^d$, then

$$\begin{aligned} L^\star &= \int_{\mathbb{R}^d} \min\left(r(\mathbf{x}), 1 - r(\mathbf{x})\right) f(\mathbf{x}) \mathrm{d}\mathbf{x} \\ &= \int_{\mathbb{R}^d} \min\left((1-p) f_0(\mathbf{x}), p f_1(\mathbf{x})\right) \mathrm{d}\mathbf{x}, \end{aligned}$$

where $p = \mathbb{P}\{Y = 1\}$ and f_i is the density of $\mathbf{X}$ given that $Y = i$. The probabilities p and $1 - p$ are called the class probabilities, and f_0 and f_1 are the class-conditional densities. If f_0 and f_1 are nonoverlapping, that is, $\int_{\mathbb{R}^d} f_0(\mathbf{x}) f_1(\mathbf{x}) \mathrm{d}\mathbf{x} = 0$, then obviously $L^\star = 0$. Besides, assuming that $p = 1/2$, we have

$$\begin{aligned} L^\star &= \frac{1}{2} \int_{\mathbb{R}^d} \min\left(f_0(\mathbf{x}), f_1(\mathbf{x})\right) \mathrm{d}\mathbf{x} \\ &= \frac{1}{2} - \frac{1}{4} \int_{\mathbb{R}^d} |f_0(\mathbf{x}) - f_1(\mathbf{x})| \, \mathrm{d}\mathbf{x} \end{aligned}$$

(since $\min(a, b) = \frac{a+b}{2} - \frac{|a-b|}{2}$). Thus, the Bayes error is directly related to the L^1 distances between the densities. □

Most of the time, the distribution of $(\mathbf{X}, Y)$ is unknown, so that the optimal decision $g^\star$ is unknown too. We do not consult an expert to try to reconstruct $g^\star$, but have access to a database $\mathscr{D}_n = ((\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n))$ of i.i.d. copies of $(\mathbf{X}, Y)$, observed in the past. We assume that $\mathscr{D}_n$ and $(\mathbf{X}, Y)$ are independent. A classifier, or classification rule, $g_n(\mathbf{x}; \mathscr{D}_n)$ is a Borel measurable function of $\mathbf{x}$ and $\mathscr{D}_n$ that attempts to estimate Y from $\mathbf{x}$ and $\mathscr{D}_n$. For simplicity, we omit $\mathscr{D}_n$ in the notation and write $g_n(\mathbf{x})$ instead of $g_n(\mathbf{x}; \mathscr{D}_n)$. The process of constructing g_n is sometimes called learning, supervised learning, or learning with a teacher.

The error probability of a given classifier g_n is the random variable

$$L(g_n) = \mathbb{P}\{g_n(\mathbf{X}) \neq Y | \mathscr{D}_n\}.$$

So, $L(g_n)$ averages over the distribution of $(\mathbf{X}, Y)$, but the data set is held fixed. It measures the future performance of the rule with the given data.

17.2 Weak, strong, and universal consistency

Generally, we cannot hope to design a function that achieves the Bayes error probability $L^\star$, but it is possible that the limit behavior of $L(g_n)$ compares favorably to $L^\star$. This idea is encapsulated in the notion of consistency:

Definition 17.1 (Weak and strong consistency). A classification rule g_n is (weakly) consistent (or asymptotically Bayes-risk efficient) for a certain distribution of $(\mathbf{X}, Y)$ if

$$\mathbb{E}L(g_n) = \mathbb{P}\{g_n(\mathbf{X}) \neq Y\} \to L^\star \quad \text{as } n \to \infty,$$

and strongly consistent if

$$L(g_n) \to L^\star \quad \text{almost surely.}$$

Remark 17.2. Noting that $L(g_n) \geq L^\star$, consistency may alternatively be defined as the convergence in L^1 of $L(g_n)$, that is, $\mathbb{E}|L(g_n) - L^\star| \to 0$. Since the random variable $L(g_n)$ is bounded, this convergence is equivalent to the convergence of $L(g_n)$ to $L^\star$ in probability, which means that, for all $\varepsilon > 0$,

$$\mathbb{P}\{|L(g_n) - L^\star| > \varepsilon\} \to 0 \quad \text{as } n \to \infty.$$

Moreover, since almost sure convergence always implies convergence in probability (see the Appendix), strong consistency implies consistency. □

A consistent rule guarantees that by increasing the amount of data, the probability that the error probability is within a small distance of the optimal achievable gets arbitrarily close to one. Strong consistency means that by using more data, the error probability gets arbitrarily close to the optimum for every training sequence, except for a set of sequences that has zero probability altogether.

If the statistician has a fair amount of a priori knowledge about the distribution of $(\mathbf{X}, Y)$, then he may be able to construct a parametric model for this distribution, determine the parameters in the model that best fit the data, and use this particular version of the model with $\mathbf{X}$ to obtain an estimate of Y. However, if the model is not exact, then it is usually impossible to design an asymptotically optimal discrimination rule in this manner. Thus, in the absence of sufficient knowledge about the distribution of $(\mathbf{X}, Y)$, is it still possible to set up a (nonparametric) asymptotically optimal classification rule? The answer is affirmative. Besides, since in many situations we definitely do not have any prior information, it is clearly essential to have a rule that gives good performance for all distributions of $(\mathbf{X}, Y)$. This strong requirement of universal goodness is formulated as follows:

Definition 17.2 (Universal consistency). A classification rule is called universally (strongly) consistent if it is (strongly) consistent for any distribution of $(\mathbf{X}, Y)$.

Universal consistency was the driving theme of the monograph by Devroye et al. (1996), and we try in the present introductory chapter as much as possible to adhere to the style and notation of that textbook.

17.3 Classification and regression estimation

We show in this section how consistent classification rules can be deduced from consistent regression function estimates. Indeed, a natural approach to classification is to first assess the regression function $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ from the training data $\mathscr{D}_n$ by some regression estimate r_n, and then use the plug-in rule

$$g_n(\mathbf{x}) = \begin{cases} 1 & \text{if } r_n(\mathbf{x}) > 1/2 \\ 0 & \text{otherwise.} \end{cases} \tag{17.2}$$

The next theorem (see, e.g., Van Ryzin, 1966; Wolverton and Wagner, 1969; Glick, 1973; Csibi, 1975; Györfi, 1976, 1978; Devroye, 1982; Devroye and Györfi, 1985) states that if r_n is close to the true regression function r in an L^p sense, then the error probability of the associated decision g_n is close to the Bayes probability of error. As in the preceding chapters, we denote by μ the distribution of $\mathbf{X}$.

Theorem 17.1 (Classification and regression). *Let r_n be a regression function estimate of r, and let g_n be the corresponding plug-in classification rule. Then*

$$0 \le L(g_n) - L^\star \le 2 \int_{\mathbb{R}^d} |r_n(\mathbf{x}) - r(\mathbf{x})| \, \mu(\mathrm{d}\mathbf{x}).$$

In particular, for all $p \ge 1$,

$$0 \le L(g_n) - L^\star \le 2 \left(\int_{\mathbb{R}^d} |r_n(\mathbf{x}) - r(\mathbf{x})|^p \, \mu(\mathrm{d}\mathbf{x}) \right)^{1/p},$$

and

$$0 \le \mathbb{E}L(g_n) - L^\star \le 2 \, \mathbb{E}^{1/p} |r_n(\mathbf{X}) - r(\mathbf{X})|^p .$$

Proof. Proceeding as in the proof of Lemma 17.1, we may write

$$\begin{aligned}
&\mathbb{P}\{g_n(\mathbf{X}) \neq Y|\mathbf{X}, \mathscr{D}_n\} \\
&\quad = 1 - \mathbb{P}\{g_n(\mathbf{X}) = Y|\mathbf{X}, \mathscr{D}_n\} \\
&\quad = 1 - (\mathbb{P}\{g_n(\mathbf{X}) = 1, Y = 1|\mathbf{X}, \mathscr{D}_n\} + \mathbb{P}\{g_n(\mathbf{X}) = 0, Y = 0|\mathbf{X}, \mathscr{D}_n\}) \\
&\quad = 1 - \left(\mathbb{1}_{[g_n(\mathbf{X})=1]} \mathbb{P}\{Y = 1|\mathbf{X}, \mathscr{D}_n\} + \mathbb{1}_{[g_n(\mathbf{X})=0]} \mathbb{P}\{Y = 0|\mathbf{X}, \mathscr{D}_n\} \right) \\
&\quad = 1 - \left(\mathbb{1}_{[g_n(\mathbf{X})=1]} r(\mathbf{X}) + \mathbb{1}_{[g_n(\mathbf{X})=0]} \left(1 - r(\mathbf{X})\right) \right),
\end{aligned}$$

where, in the last equality, we used the independence of $(\mathbf{X}, Y)$ and $\mathscr{D}_n$. Similarly,

$$\mathbb{P}\{g^\star(\mathbf{X}) \neq Y|\mathbf{X}\} = 1 - \left(\mathbb{1}_{[g^\star(\mathbf{X})=1]} r(\mathbf{X}) + \mathbb{1}_{[g^\star(\mathbf{X})=0]} \left(1 - r(\mathbf{X})\right)\right).$$

Therefore

$$\begin{aligned}
&\mathbb{P}\{g_n(\mathbf{X}) \neq Y|\mathbf{X}, \mathscr{D}_n\} - \mathbb{P}\{g^\star(\mathbf{X}) \neq Y|\mathbf{X}\} \\
&\quad = r(\mathbf{X}) \left(\mathbb{1}_{[g^\star(\mathbf{X})=1]} - \mathbb{1}_{[g_n(\mathbf{X})=1]}\right) + \left(1 - r(\mathbf{X})\right) \left(\mathbb{1}_{[g^\star(\mathbf{X})=0]} - \mathbb{1}_{[g_n(\mathbf{X})=0]}\right) \\
&\quad = \left(2r(\mathbf{X}) - 1\right) \left(\mathbb{1}_{[g^\star(\mathbf{X})=1]} - \mathbb{1}_{[g_n(\mathbf{X})=1]}\right) \\
&\quad = |2r(\mathbf{X}) - 1| \mathbb{1}_{[g_n(\mathbf{X}) \neq g^\star(\mathbf{X})]}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{P}\{g_n(\mathbf{X}) \neq Y|\mathscr{D}_n\} - L^\star &= 2 \int_{\mathbb{R}^d} |r(\mathbf{x}) - 1/2| \mathbb{1}_{[g_n(\mathbf{x}) \neq g^\star(\mathbf{x})]} \mu(\mathrm{d}\mathbf{x}) \\
&\leq 2 \int_{\mathbb{R}^d} |r_n(\mathbf{x}) - r(\mathbf{x})| \, \mu(\mathrm{d}\mathbf{x}),
\end{aligned}$$

since $g_n(\mathbf{x}) \neq g^\star(\mathbf{x})$ implies $|r_n(\mathbf{x}) - r(\mathbf{x})| \geq |r(\mathbf{x}) - 1/2|$. The other assertions follow from Hölder's and Jensen's inequality, respectively. □

Theorem 17.1 implies that a regression function estimate r_n with small L^p error automatically leads to a decision g_n with small misclassification probability. In particular, the mere existence of a regression estimate r_n for which

$$\int_{\mathbb{R}^d} |r_n(\mathbf{x}) - r(\mathbf{x})|^2 \, \mu(\mathrm{d}\mathbf{x}) \to 0 \tag{17.3}$$

in probability or almost surely implies that the corresponding plug-in decision g_n is consistent or strongly consistent, respectively. The standard consistency proof for a classification rule based on r_n usually involves showing (17.3), or its L^1 version.

Remark 17.3.

(*i*) If the bounds of Theorem 17.1 are useful for proving consistency, they are not tight and almost useless when it comes to studying rates of convergence. For (17.2) to be a good approximation of (17.1), it is not important that $r_n(\mathbf{x})$ be close to $r(\mathbf{x})$ everywhere. What is critical is that $r_n(\mathbf{x})$ should be on the same side of the decision boundary as $r(\mathbf{x})$, i.e., that $r_n(\mathbf{x}) > 1/2$ whenever $r(\mathbf{x}) > 1/2$ and $r_n(\mathbf{x}) \leq 1/2$ whenever $r(\mathbf{x}) \leq 1/2$. It is proved in Devroye et al. (1996, Theorem 6.5) that for consistent rules, rates of convergence of $\mathbb{E}L(g_n)$ to $L^\star$ are always orders of magnitude better than rates of convergence of $\mathbb{E}^{1/2}|r_n(\mathbf{X}) - r(\mathbf{X})|^2$ to zero. Pattern recognition is thus easier than regression function estimation, in the sense that, to achieve acceptable results in classification, we can do more with smaller sample sizes than in regression estimation. This is a consequence of the fact that less is required in pattern recognition.

(*ii*) The behavior of $r(\mathbf{x})$ at those $\mathbf{x}$'s where $r(\mathbf{x}) \approx 1/2$ is sometimes expressed by a so-called margin condition, which takes the form

$$\mathbb{P}\{|r(\mathbf{X}) - 1/2| \leq t\} \leq C\, t^{\alpha},$$

for some positive constants C and α, and all $0 < t \leq t^{\star}$, where $t^{\star} \leq 1/2$ (see, e.g., Tsybakov, 2004; Massart and Nédélec, 2006; Audibert and Tsybakov, 2007; Kohler and Krzyżak, 2007; Samworth, 2012; Gada et al., 2014). This assumption offers a useful characterization of the behavior of the regression function r in the vicinity of the boundary set $\{\mathbf{x} \in \mathbb{R}^d : r(\mathbf{x}) = 1/2\}$. □

Stone's theorem 10.1 provides us with conditions ensuring universal L^p-consistency of local averaging regression function estimates. Thus, by virtue of Theorem 17.1, the same theorem allows us to deduce universal consistency of the corresponding plug-in rules.

Recall that a local averaging estimate of the regression function takes the form

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

where $(W_{n1}(\mathbf{x}), \ldots, W_{nn}(\mathbf{x}))$ is a weight vector, and each $W_{ni}(\mathbf{x})$ is a Borel measurable function of $\mathbf{x}$ and $\mathbf{X}_1, \ldots, \mathbf{X}_n$ (not $Y_1, \ldots, Y_n$). Equivalently, in our binary classification setting,

$$r_n(\mathbf{x}) = \sum_{i=1}^{n} W_{ni}(\mathbf{x}) \mathbb{1}_{[Y_i=1]}.$$

The companion plug-in classification rule is defined as

$$g_n(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_{i=1}^{n} W_{ni}(\mathbf{x}) Y_i > 1/2 \\ 0 & \text{otherwise,} \end{cases}$$

or, equivalently, whenever $\sum_{i=1}^{n} W_{ni}(\mathbf{x}) = 1$,

$$g_n(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_{i=1}^{n} W_{ni}(\mathbf{x}) \mathbb{1}_{[Y_i=1]} > \sum_{i=1}^{n} W_{ni}(\mathbf{x}) \mathbb{1}_{[Y_i=0]} \\ 0 & \text{otherwise.} \end{cases}$$

As in the regression setting, it is clear that the pairs $(\mathbf{X}_i, Y_i)$ such that $\mathbf{X}_i$ is "close" to $\mathbf{x}$ should provide more information about $r(\mathbf{x})$ than those "far" from $\mathbf{x}$. Thus, the weights are typically larger in the neighborhood of $\mathbf{x}$. Examples of such rules include the histogram, kernel, and nearest neighbor rules. Theorem 17.2 below follows directly from Theorem 17.1 and Stone's theorem 10.1.

Theorem 17.2 (Stone's theorem for classification). *Assume that for any distribution of* $\mathbf{X}$*, the weights satisfy the following four conditions:*

(*i*) *There is a constant* C *such that, for every Borel measurable function* $h : \mathbb{R}^d \to \mathbb{R}$*,*

$$\mathbb{E}\left[\sum_{i=1}^{n} |W_{ni}(\mathbf{X})|\,|h(\mathbf{X}_i)|\right] \le C\,\mathbb{E}\,|h(\mathbf{X})| \quad \textit{for all } n \ge 1.$$

(*ii*) *There is a constant* $D \ge 1$ *such that*

$$\mathbb{P}\left\{\sum_{i=1}^{n} |W_{ni}(\mathbf{X})| \le D\right\} = 1 \quad \textit{for all } n \ge 1.$$

(*iii*) *For all* $a > 0$*,*

$$\sum_{i=1}^{n} |W_{ni}(\mathbf{X})|\,\mathbb{1}_{[\|\mathbf{X}_i - \mathbf{X}\| > a]} \to 0 \quad \textit{in probability}.$$

(*iv*) *One has*

$$\sum_{i=1}^{n} W_{ni}(\mathbf{X}) \to 1 \quad \textit{in probability}.$$

(*v*) *One has*

$$\max_{1 \le i \le n} |W_{ni}(\mathbf{X})| \to 0 \quad \textit{in probability}.$$

Then the corresponding plug-in classification rule g_n *is universally consistent, that is,*

$$\mathbb{E}L(g_n) \to L^\star$$

for all distributions of $(\mathbf{X}, Y)$*.*

17.4 Supplement: multi-label classification

The supervised classification theory can be generalized without difficulty from the binary case to the multi-label case, where Y takes $M \ge 2$ distinct values, say $\{1, \ldots, M\}$. The Bayes decision function can be computed via the a posteriori probabilities $r_j(\mathbf{x}) = \mathbb{P}\{Y = j | \mathbf{X} = \mathbf{x}\}$, $1 \le j \le M$:

$$g^\star(\mathbf{x}) \in \arg\max_{1 \le j \le M} r_j(\mathbf{x}),$$

where, by convention, ties are broken in favor of smaller indices. As in the binary case, the performance with a certain discrimination rule g_n is measured by its probability of error $L(g_n) = \mathbb{P}\{g_n(\mathbf{X}) \neq Y | \mathscr{D}_n\}$ and, in any case, $L(g_n)$ cannot be smaller than the Bayes error

$$L^\star = L(g^\star) = \inf_{g:\mathbb{R}^d \to \{1,\ldots,M\}} \mathbb{P}\{g(\mathbf{X}) \neq Y\}.$$

Observing that

$$r_j(\mathbf{x}) = \mathbb{E}[\mathbb{1}_{[Y=j]} | \mathbf{X} = \mathbf{x}], \quad 1 \leq j \leq M,$$

the unknown $r_j(\mathbf{x})$'s can be approximated by estimates $r_{nj}(\mathbf{x})$ constructed from the data sets

$$\mathscr{D}_{nj} = \left((\mathbf{X}_1, \mathbb{1}_{[Y_1=j]}), \ldots, (\mathbf{X}_n, \mathbb{1}_{[Y_n=j]})\right),$$

and the plug-in estimate is

$$g_n(\mathbf{x}) \in \arg\max_{1 \leq j \leq M} r_{nj}(\mathbf{x}). \tag{17.4}$$

A generalized version of Theorem 17.1 asserts that

$$\begin{aligned} 0 \leq L(g_n) - L^\star &\leq 2 \sum_{j=1}^{M} \int_{\mathbb{R}^d} \left| r_{nj}(\mathbf{x}) - r(\mathbf{x}) \right| \mu(\mathrm{d}\mathbf{x}) \\ &\leq 2 \sum_{j=1}^{M} \left(\int_{\mathbb{R}^d} \left| r_{nj}(\mathbf{x}) - r(\mathbf{x}) \right|^p \mu(\mathrm{d}\mathbf{x}) \right)^{1/p}, \end{aligned}$$

for all $p \geq 1$. Thus, if the estimates r_{nj} are close to the a posteriori probabilities r_j, then again the error of the plug-in estimate (17.4) is close to the optimal error.

Chapter 18
The nearest neighbor rule: fixed k

18.1 Introduction

In this chapter, $(\mathbf{X}, Y) \in \mathbb{R}^d \times \{0, 1\}$, and $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ are reordered according to increasing values of $\|\mathbf{X}_i - \mathbf{x}\|$. Ties are broken as for regression. The reordered sequence is denoted by $(\mathbf{X}_{(1)}(\mathbf{x}), Y_{(1)}(\mathbf{x})), \ldots, (\mathbf{X}_{(n)}(\mathbf{x}), Y_{(n)}(\mathbf{x}))$. As usual, we let $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ and recall, since $Y \in \{0, 1\}$, that $r(\mathbf{x}) = \mathbb{P}\{Y = 1|\mathbf{X} = \mathbf{x}\}$.

Definition 18.1. Let $(v_{n1}, \ldots, v_{nn})$ be a given weight vector summing to one. The nearest neighbor classification rule (or nearest neighbor classifier) is defined for $\mathbf{x} \in \mathbb{R}^d$ by

$$g_n(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_{i=1}^n v_{ni} Y_{(i)}(\mathbf{x}) > 1/2 \\ 0 & \text{otherwise,} \end{cases}$$

or, equivalently,

$$g_n(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_{i=1}^n v_{ni} \mathbb{1}_{[Y_{(i)}(\mathbf{x})=1]} > \sum_{i=1}^n v_{ni} \mathbb{1}_{[Y_{(i)}(\mathbf{x})=0]} \\ 0 & \text{otherwise.} \end{cases}$$

In other words, $g_n(\mathbf{x})$ takes a weighted vote among the labels of the nearest neighbors of $\mathbf{x}$. For the particular choice $(v_{n1}, \ldots, v_{nn}) = (1/k, \ldots, 1/k, 0, \ldots, 0)$, we obtain the standard k-nearest neighbor rule (Fix and Hodges, 1951, 1991a, 1952, 1991b; Cover and Hart, 1967; Stone, 1977), which corresponds to a majority vote:

$$g_n(\mathbf{x}) = \begin{cases} 1 & \text{if } \frac{1}{k} \sum_{i=1}^k \mathbb{1}_{[Y_{(i)}(\mathbf{x})=1]} > \frac{1}{k} \sum_{i=1}^k \mathbb{1}_{[Y_{(i)}(\mathbf{x})=0]} \\ 0 & \text{otherwise,} \end{cases}$$

G. Biau, L. Devroye, *Lectures on the Nearest Neighbor Method*,
Springer Series in the Data Sciences, DOI 10.1007/978-3-319-25388-6_18

or more simply, observing that the $1/k$ terms do not play a role in the decision,

$$g_n(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_{i=1}^k \mathbb{1}_{[Y_{(i)}(\mathbf{x})=1]} > \sum_{i=1}^k \mathbb{1}_{[Y_{(i)}(\mathbf{x})=0]} \\ 0 & \text{otherwise.} \end{cases}$$

As an appetizer, and in the spirit of Chapter 9, we will be concerned in this chapter with convergence issues for the k-nearest neighbor rule when k does not change with n. In particular, we will see that for all distributions, the expected error probability $\mathbb{E}L(g_n)$ tends to a limit $L_{k\text{NN}}$ that is in general close to but larger than $L^\star$. The methodology for obtaining this result is based on Stone's lemma 10.7 and is interesting in its own right. The expression for $L_{k\text{NN}}$ is then worked out, and several inequalities such as $L_{k\text{NN}} \leq L^\star + 1/\sqrt{ke}$ are discussed. For surveys of various aspects of the nearest neighbor or related methods, see Devijver (1980), Devroye and Wagner (1982), Dasarathy (1991), and Devroye et al. (1996).

18.2 Behavior for fixed k

The main result of the chapter is Theorem 18.1 below. Under various regularity conditions ($\mathbf{X}$ has a density f, and both f and r are almost everywhere continuous), it is due to Cover and Hart (1967). In the present generality, the theorem essentially appears in Stone (1977)—see also Fritz (1975) and Devroye (1981b).

Theorem 18.1. *Let $k \in \{1, \ldots, n\}$ be odd and fixed. Let g_n be the k-nearest neighbor classification rule. Then, for all distributions of $(\mathbf{X}, Y)$,*

$$\mathbb{E}L(g_n) \to L_{k\text{NN}} \quad \text{as } n \to \infty,$$

where

$$L_{k\text{NN}} \stackrel{\text{def}}{=} \mathbb{E}\left[\sum_{j=0}^{k} \binom{k}{j} r(\mathbf{X})^j \left(1 - r(\mathbf{X})\right)^{k-j} \left(r(\mathbf{X}) \mathbb{1}_{[j<k/2]} + \left(1 - r(\mathbf{X})\right) \mathbb{1}_{[j>k/2]} \right) \right].$$

In particular, for $k = 1$,

$$\mathbb{E}L(g_n) \to L_{1\text{NN}} \stackrel{\text{def}}{=} 2\mathbb{E}\left[r(\mathbf{X}) \left(1 - r(\mathbf{X})\right) \right],$$

and $L^\star \leq L_{1\text{NN}} \leq 2L^\star(1 - L^\star) \leq 2L^\star$.

Thus, whenever $k = 1$, the theorem says that the 1-nearest neighbor classification rule is asymptotically at most twice as bad as the Bayes rule—especially for small $L^\star$, this property should be useful. It is noteworthy that this convergence is universal, in the sense that it happens for any distribution of $(\mathbf{X}, Y)$. On the other hand, recalling that $L^\star = \mathbb{E}[\min(r(\mathbf{X}), 1 - r(\mathbf{X}))]$, we see that the 1-nearest neighbor

classifier is consistent (that is, $\mathbb{E}L(g_n) \to L^\star$) if $r(\mathbf{X}) \in \{0, 1/2, 1\}$ with probability one. The noiseless case occurs when $r(\mathbf{X}) \in \{0, 1\}$. The independent case occurs when $r(\mathbf{X}) = 1/2$ (since Y is a random coin flip independent of $\mathbf{X}$ then). Thus, logically, in the next chapter we will allow k to grow with n in order to obtain universally good consistency properties.

To prove Theorem 18.1, we first need the following lemma, which generalizes Lemma 9.2.

Lemma 18.1. *Let $h : \mathbb{R}^d \to \mathbb{R}$ be a Borel measurable function such that $\mathbb{E}|h(\mathbf{X})| < \infty$. If $k/n \to 0$, then*

$$\frac{1}{k}\sum_{i=1}^{k} \mathbb{E}\left|h\left(\mathbf{X}_{(i)}(\mathbf{X})\right) - h(\mathbf{X})\right| \to 0 \quad \textit{and} \quad \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}\left|h\left(\mathbf{X}_{(i)}(\mathbf{X})\right)\right| \to \mathbb{E}\left|h(\mathbf{X})\right|.$$

Proof. We only need to prove the first assertion, since the second one follows by the triangle inequality. Given $\varepsilon > 0$, find a uniformly continuous function h_ε with compact support such that $\mathbb{E}|h(\mathbf{X}) - h_\varepsilon(\mathbf{X})| \le \varepsilon$ (this is possible by Theorem 20.17 in the Appendix). Then there is a $\delta > 0$, depending upon ε only, such that $\|\mathbf{y} - \mathbf{x}\| \le \delta$ implies $|h_\varepsilon(\mathbf{y}) - h_\varepsilon(\mathbf{x})| \le \varepsilon$. Thus, if γ_d denotes a constant depending upon d only,

$$\begin{aligned}
&\frac{1}{k}\sum_{i=1}^{k} \mathbb{E}\left|h\left(\mathbf{X}_{(i)}(\mathbf{X})\right) - h(\mathbf{X})\right| \\
&\quad \le \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}\left|h\left(\mathbf{X}_{(i)}(\mathbf{X})\right) - h_\varepsilon\left(\mathbf{X}_{(i)}(\mathbf{X})\right)\right| + \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}\left|h_\varepsilon\left(\mathbf{X}_{(i)}(\mathbf{X})\right) - h_\varepsilon(\mathbf{X})\right| \\
&\qquad + \mathbb{E}\left|h_\varepsilon(\mathbf{X}) - h(\mathbf{X})\right| \\
&\quad \le (2\gamma_d + 1)\mathbb{E}\left|h(\mathbf{X}) - h_\varepsilon(\mathbf{X})\right| + \varepsilon + 2\|h_\varepsilon\|_\infty \times \mathbb{P}\left\{\|\mathbf{X}_{(k)}(\mathbf{X}) - \mathbf{X}\| > \delta\right\} \\
&\qquad \text{(by Stone's lemma 10.7, with } (v_{n1}, \ldots, v_{nn}) = (1/k, \ldots, 1/k, 0, \ldots, 0)) \\
&\quad \le 2(\gamma_d + 1)\varepsilon + \|h_\varepsilon\|_\infty \times \mathbb{P}\left\{\|\mathbf{X}_{(k)}(\mathbf{X}) - \mathbf{X}\| > \delta\right\}.
\end{aligned}$$

By the Lebesgue dominated convergence theorem and Lemma 2.2, the probability on the right-hand side vanishes as n tends to infinity, so that

$$\frac{1}{k}\sum_{i=1}^{k} \mathbb{E}\left|h\left(\mathbf{X}_{(i)}(\mathbf{X})\right) - h(\mathbf{X})\right| \to 0. \qquad \square$$

Proof (Theorem 18.1). Assume that we are given i.i.d. pairs $(\mathbf{X}_1, U_1), \ldots, (\mathbf{X}_n, U_n)$, all distributed as $(\mathbf{X}, U)$, where $\mathbf{X}$ is as before, and U is uniformly distributed on $[0, 1]$ and independent of $\mathbf{X}$. For fixed $\mathbf{x} \in \mathbb{R}^d$, we define, for all $1 \le i \le n$,

$$Y_i = \mathbb{1}_{[U_i \le r(\mathbf{X}_i)]} \quad \text{and} \quad Y_i'(\mathbf{x}) = \mathbb{1}_{[U_i \le r(\mathbf{x})]}.$$

Observe that $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ is an i.i.d. sequence with each of the n pairs distributed as our prototype $(\mathbf{X}, Y)$. We now have an i.i.d. sequence with i-th vector given by $\mathbf{X}_i$, Y_i, $Y_i'(\mathbf{x})$, U_i. Reordering the data sequence according to increasing values of $\|\mathbf{X}_i - \mathbf{x}\|$ yields a new sequence with the i-th vector denoted by $\mathbf{X}_{(i)}(\mathbf{x})$, $Y_{(i)}(\mathbf{x})$, $Y'_{(i)}(\mathbf{x})$, and $U_{(i)}(\mathbf{x})$. Studying the k-nearest neighbor classification rule g_n turns out to be almost equivalent to studying the approximate rule g_n':

$$g_n'(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_{i=1}^n Y'_{(i)}(\mathbf{x}) > k/2 \\ 0 & \text{otherwise.} \end{cases}$$

The latter classifier is of no practical value because it requires the knowledge of $r(\mathbf{x})$. Interestingly however, it is easier to study, as $Y'_{(1)}(\mathbf{x}), \ldots, Y'_{(n)}(\mathbf{x})$ are i.i.d., whereas $Y_{(1)}(\mathbf{x}), \ldots, Y_{(n)}(\mathbf{x})$ are not. In particular, we have that

$$\mathbb{P}\left\{g_n(\mathbf{X}) \neq g_n'(\mathbf{X})\right\} \leq \sum_{i=1}^k \mathbb{E}\left|r\left(\mathbf{X}_{(i)}(\mathbf{X})\right) - r(\mathbf{X})\right|. \tag{18.1}$$

To prove (18.1), observe that for fixed $\mathbf{x} \in \mathbb{R}^d$, we have

$$\begin{aligned}
&\mathbb{P}\left\{g_n(\mathbf{x}) \neq g_n'(\mathbf{x})\right\} \\
&\quad \leq \mathbb{P}\left\{\sum_{i=1}^n Y_{(i)}(\mathbf{x}) \neq \sum_{i=1}^n Y'_{(i)}(\mathbf{x})\right\} \\
&\quad \leq \mathbb{P}\left\{\left(Y_{(1)}(\mathbf{x}), \ldots, Y_{(n)}(\mathbf{x})\right) \neq \left(Y'_{(1)}(\mathbf{x}), \ldots, Y'_{(n)}(\mathbf{x})\right)\right\} \\
&\quad \leq \mathbb{P}\left\{\bigcup_{i=1}^k \left[r\left(\mathbf{X}_{(i)}(\mathbf{x})\right) < U_{(i)}(\mathbf{x}) \leq r(\mathbf{x})\right] \cup \bigcup_{i=1}^k \left[r(\mathbf{x}) < U_{(i)}(\mathbf{x}) \leq r\left(\mathbf{X}_{(i)}(\mathbf{x})\right)\right]\right\}.
\end{aligned}$$

Inequality (18.1) follows by the union bound and the fact that the $U_{(i)}(\mathbf{x})$'s are uniform on $[0, 1]$, independent of the $\mathbf{X}_{(i)}(\mathbf{x})$'s.

Next, let $\mathscr{D}_n' = ((\mathbf{X}_1, Y_1, U_1), \ldots, (\mathbf{X}_n, Y_n, U_n))$ be the i.i.d. data augmented by the uniform random variables $U_1, \ldots, U_n$, as described above. For the decision g_n based on $\mathscr{D}_n$, we have the probability of error

$$L(g_n) = \mathbb{P}\{g_n(\mathbf{X}) \neq Y | \mathscr{D}_n\} = \mathbb{P}\{g_n(\mathbf{X}) \neq Y | \mathscr{D}_n'\},$$

whereas for g_n' we have

$$L(g_n') = \mathbb{P}\{g_n'(\mathbf{X}) \neq Y | \mathscr{D}_n'\}.$$

Clearly,

$$\mathbb{E}\left|L(g_n) - L(g_n')\right| \leq \mathbb{P}\left\{g_n(\mathbf{X}) \neq g_n'(\mathbf{X})\right\} = \mathrm{o}(1),$$

by inequality (18.1) and Lemma 18.1. We have just shown that $\mathbb{E}L(g_n) - \mathbb{E}L(g'_n) \to 0$. Thus, to prove the result, it is enough to establish that $\mathbb{E}L(g'_n) \to L_{k\text{NN}}$, where the rule g'_n is defined by

$$g'_n(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_{i=1}^k Z_i > k/2 \\ 0 & \text{otherwise,} \end{cases}$$

for $Z_1, \ldots, Z_k$ independent Bernoulli random variables with success probability $r(\mathbf{x})$. But, for every n,

$$\begin{aligned} \mathbb{E}L(g'_n) &= \mathbb{P}\left\{Z_1 + \cdots + Z_k > \frac{k}{2}, Y = 0\right\} + \mathbb{P}\left\{Z_1 + \cdots + Z_k < \frac{k}{2}, Y = 1\right\} \\ &= \mathbb{P}\left\{Z_1 + \cdots + Z_k > \frac{k}{2}, Z_0 = 0\right\} + \mathbb{P}\left\{Z_1 + \cdots + Z_k < \frac{k}{2}, Z_0 = 1\right\} \\ &\quad (Z_0, \ldots, Z_k \text{ are i.i.d. Ber}(\mu(\mathbf{x}))), \end{aligned}$$

which leads directly to the first part of the theorem. To prove the second point, observe that for all $p \in [0,1]$, $2p(1-p) \geq \min(p, 1-p)$ as $2\max(p, 1-p) \geq 1$. Thus, recalling that $L^\star = \mathbb{E}A(\mathbf{X})$, where $A(\mathbf{X}) = \min(r(\mathbf{X}), 1 - r(\mathbf{X}))$, we conclude that

$$\begin{aligned} L^\star \leq L_{1\text{NN}} &= 2\mathbb{E}\left[r(\mathbf{X})\left(1 - r(\mathbf{X})\right)\right] \\ &= 2\mathbb{E}\left[A(\mathbf{X})\left(1 - A(\mathbf{X})\right)\right] \\ &= 2\mathbb{E}\left[A(\mathbf{X})\right]\mathbb{E}\left[1 - A(\mathbf{X})\right] - 2\mathbb{V}A(\mathbf{X}) \\ &\leq 2L^\star(1 - L^\star) \leq 2L^\star. \end{aligned}$$

□

The limit result in Theorem 18.1 is distribution-free, and the limit $L_{k\text{NN}}$ depends upon $r(\mathbf{X})$ only. The continuity or lack of smoothness of r is irrelevant—it only matters for the speed with which $\mathbb{E}L(g_n)$ approaches the limit $L_{k\text{NN}}$. Until now we assumed throughout that k was odd, so that voting ties were avoided. Extensions to even k and to more general weighted nearest neighbor rules are available in the literature—see, e.g., Bailey and Jain (1978), Devijver (1978), and Devroye et al. (1996, Chapter 5).

Returning to the case where k is odd, several useful representations of $L_{k\text{NN}}$ may be obtained. For example, we have

$$\begin{aligned} L_{k\text{NN}} &= \mathbb{E}\left[r(\mathbf{X})\mathbb{P}\left\{\text{Bin}(k, r(\mathbf{X})) < \frac{k}{2} \,\middle|\, \mathbf{X}\right\}\right] \\ &\quad + \mathbb{E}\left[\left(1 - r(\mathbf{X})\right)\mathbb{P}\left\{\text{Bin}(k, r(\mathbf{X})) > \frac{k}{2} \,\middle|\, \mathbf{X}\right\}\right] \\ &= \mathbb{E}\left[\min\left(r(\mathbf{X}), 1 - r(\mathbf{X})\right)\right] \\ &\quad + \mathbb{E}\left[\left(1 - 2\min\left(r(\mathbf{X}), 1 - r(\mathbf{X})\right)\right)\mathbb{P}\left\{\text{Bin}\left(k, \min\left(r(\mathbf{X}), 1 - r(\mathbf{X})\right)\right) > \frac{k}{2} \,\middle|\, \mathbf{X}\right\}\right]. \end{aligned}$$

Put differently,

$$L_{k\text{NN}} = \mathbb{E}\left[\alpha_k\left(r(\mathbf{X})\right)\right],$$

where, for all $p \in [0,1]$,

$$\alpha_k(p) = \min(p, 1-p) + |2p-1|\mathbb{P}\left\{\text{Bin}(k, \min(p,1-p)) > \frac{k}{2}\right\}.$$

Since $L^\star = \mathbb{E}\left[\min\left(r(\mathbf{X}), 1-r(\mathbf{X})\right)\right]$, this expression may be exploited to obtain a host of inequalities on $L_{k\text{NN}} - L^\star$. For example, using Hoeffding's bound on binomials (Corollary 20.1 in the Appendix), we may write

$$\begin{aligned}
L_{k\text{NN}} - L^\star &\le \sup_{0\le p\le 1/2} (1-2p)\mathbb{P}\left\{\text{Bin}(k,p) > \frac{k}{2}\right\} \\
&\le \sup_{0\le p\le 1/2} (1-2p)\mathbb{P}\left\{\text{Bin}(k,p) - kp > k\left(\frac{1}{2}-p\right)\right\} \\
&\le \sup_{0\le p\le 1/2} (1-2p)e^{-2k(1/2-p)^2}.
\end{aligned}$$

Therefore,

$$L_{k\text{NN}} - L^\star \le \sup_{0\le u\le 1} ue^{-ku^2/2} \le \frac{1}{\sqrt{ke}}.$$

Bounds of this type have been obtained by Györfi and Györfi (1978), Devijver (1979), and Devroye (1981c). A sharp version, due to Devroye (1981c), asserts that for all distributions and all odd $k \ge 3$,

$$L_{k\text{NN}} \le L^\star\left(1 + \frac{\gamma}{\sqrt{k}}\left(1 + \text{O}(k^{-1/6})\right)\right),$$

where $\gamma = \sup_{t>0} 2t\mathbb{P}\{N > t\} = 0.33994241\ldots$, N is standard normal, and $\text{O}(\cdot)$ refers to $k \to \infty$.

Let us finally mention that it is also instructive to look at the behavior of $L_{k\text{NN}}$ when $L^\star$ is zero or, at least, small. Devroye et al. (1996, Theorem 5.4) show that, for all distributions,

$$L^\star \le \cdots \le L_{(2k+1)\text{NN}} \le L_{(2k-1)\text{NN}} \le \cdots \le L_{3\text{NN}} \le L_{1\text{NN}} \le 2L^\star.$$

We retain from this inequality that if $L^\star = 0$, then $L_{k\text{NN}} = 0$ for all odd k. Remarkably, then, for every fixed k, the k-nearest neighbor rule is consistent. To analyze the behavior of $L_{k\text{NN}}$ when $L^\star$ is small, recall that

$$
\begin{aligned}
L_{k\text{NN}} &= \mathbb{E}\Big[\sum_{j=0}^{k} \binom{k}{j} r(\mathbf{X})^j \left(1 - r(\mathbf{X})\right)^{k-j} \left(r(\mathbf{X}) \mathbb{1}_{[j<k/2]} + \left(1 - r(\mathbf{X})\right) \mathbb{1}_{[j>k/2]} \right) \Big] \\
&= \mathbb{E}\Big[\sum_{j<k/2} \binom{k}{j} r(\mathbf{X})^{j+1} \left(1 - r(\mathbf{X})\right)^{k-j} + \sum_{j>k/2} \binom{k}{j} r(\mathbf{X})^j \left(1 - r(\mathbf{X})\right)^{k-j+1} \Big] \\
&= \sum_{j<k/2} \binom{k}{j} \mathbb{E}\Big[\big(r(\mathbf{X}) \left(1 - r(\mathbf{X})\right) \big)^{j+1} \Big(\left(1 - r(\mathbf{X})\right)^{k-2j-1} + r(\mathbf{X})^{k-2j-1} \Big) \Big].
\end{aligned}
$$

As $p^a + (1-p)^a$ is a function of $p(1-p)$ for integer a, this may be further reduced to simplified forms such as

$$
\begin{aligned}
L_{1\text{NN}} &= 2\mathbb{E}\left[r(\mathbf{X}) \left(1 - r(\mathbf{X})\right) \right] \\
L_{3\text{NN}} &= \mathbb{E}\left[r(\mathbf{X}) \left(1 - r(\mathbf{X})\right) \right] + 4\mathbb{E}\Big[\left(r(\mathbf{X}) \left(1 - r(\mathbf{X})\right) \right)^2 \Big] \\
L_{5\text{NN}} &= \mathbb{E}\left[r(\mathbf{X}) \left(1 - r(\mathbf{X})\right) \right] + \mathbb{E}\Big[\left(r(\mathbf{X}) \left(1 - r(\mathbf{X})\right) \right)^2 \Big] + 12\mathbb{E}\Big[\left(r(\mathbf{X}) \left(1 - r(\mathbf{X})\right) \right)^3 \Big],
\end{aligned}
$$

and so on. The behavior of $\alpha_k(p)$ as p approaches zero is very informative. Indeed, as $p \downarrow 0$, we have

$$
\begin{aligned}
\alpha_1(p) &= 2p(1-p) \sim 2p \\
\alpha_3(p) &= p(1-p)\left(1 + 4p(1-p)\right) \sim p + 3p^2 \\
\alpha_5(p) &= p(1-p)\left(1 + p(1-p) + 12p^2(1-p)^2\right) \sim p + 10p^3,
\end{aligned}
$$

while for the Bayes error, $L^\star = \mathbb{E}[\min(r(\mathbf{X}), 1 - r(\mathbf{X}))] = \mathbb{E}[\alpha_\infty(r(\mathbf{X}))]$, where $\alpha_\infty = \min(p, 1-p) \sim p$ as $p \downarrow 0$. Assuming, for example, that $r(\mathbf{x}) = p$ at all $\mathbf{x}$, we conclude that, as $p \downarrow 0$,

$$
L_{1\text{NN}} \sim 2L^\star \quad \text{and} \quad L_{3\text{NN}} \sim L^\star.
$$

Moreover, $L_{1\text{NN}} - L^\star \sim L^\star$, $L_{3\text{NN}} - L^\star \sim 3L^{\star 2}$. Assume that $L^\star = p = 0.01$. Then $L_{1\text{NN}} - L^\star \approx 0.01$, whereas $L_{3\text{NN}} - L^\star \approx 0.0003$. Thus, for all practical purposes, the 3-nearest neighbor rule is virtually perfect. For this reason, the 3-nearest neighbor rule is highly recommended. Little is gained by considering the 5-nearest neighbor rule when p is small, as $L_{5\text{NN}} - L^\star \approx 0.00001$.

Chapter 19
The nearest neighbor rule: variable k

19.1 Universal consistency

Given weights $(v_{n1}, \ldots, v_{nn})$ satisfying $\sum_{i=1}^{n} v_{ni} = 1$, the nearest neighbor classifier is defined for $\mathbf{x} \in \mathbb{R}^d$ by

$$g_n(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_{i=1}^{n} v_{ni} Y_{(i)}(\mathbf{x}) > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

The next theorem provides necessary and sufficient conditions on the weight sequence for this rule to be universally consistent, i.e., $\mathbb{E}L(g_n) \to L^\star$ for all distributions of $(\mathbf{X}, Y)$. It starts with the observation that the nearest neighbor rule is a local averaging classifier, where the weights $W_{ni}(\mathbf{x})$ are obtained by putting

$$W_{ni}(\mathbf{x}) = v_{n\Sigma_i},$$

where $(\Sigma_1, \ldots, \Sigma_n)$ is a permutation of $(1, \ldots, n)$ such that $\mathbf{X}_i$ is the Σ_i-th nearest neighbor of $\mathbf{x}$ for all i. Thus, this decision rule falls within the scope of Stone's theorem 17.2 for classification, so that we just need to check the conditions of that theorem. This has already been done in Chapter 10, in the context of regression estimation.

Theorem 19.1 (Universal consistency). *Let $(v_{n1}, \ldots, v_{nn})$ be a probability weight vector such that $v_{n1} \geq \cdots \geq v_{nn}$ for all n. Then the corresponding nearest neighbor classification rule is universally consistent if and only if there exists a sequence of integers $\{k\} = \{k_n\}$ such that*

$$\begin{array}{ll} (i) & k \to \infty \quad \textit{and} \quad k/n \to 0; \\ (ii) & \sum_{i>k} v_{ni} \to 0; \\ (iii) & v_{n1} \to 0. \end{array} \tag{19.1}$$

G. Biau, L. Devroye, *Lectures on the Nearest Neighbor Method*,
Springer Series in the Data Sciences, DOI 10.1007/978-3-319-25388-6_19

For the k-nearest neighbor classifier, we set $v_{ni} = 1/k$ for $1 \le i \le k$ and $v_{ni} = 0$ otherwise, where k is a positive integer not exceeding n. The following corollary, which appeared in Stone's 1997 paper (with a different distance tie-breaking strategy), was the first universal consistency result for any rule.

Corollary 19.1. *The k-nearest neighbor classification rule is universally consistent if and only if $k \to \infty$ and $k/n \to 0$.*

For the proof of Theorem 19.1, we first need a lemma.

Lemma 19.1. *Let c be a positive constant. Then any $[-c, c]$-valued random variable Z satisfies, for all $t > \mathbb{E}Z$,*

$$\mathbb{P}\{Z \le t\} \ge \frac{t - \mathbb{E}Z}{c + t}.$$

Proof. Just note that

$$\mathbb{E}Z = \mathbb{E}[Z\mathbb{1}_{[-c \le Z \le t]}] + \mathbb{E}[Z\mathbb{1}_{[t < Z \le c]}] \ge -c\mathbb{P}\{Z \le t\} + t\mathbb{P}\{Z > t\},$$

so that

$$(c + t)\mathbb{P}\{Z \le t\} \ge t - \mathbb{E}Z. \quad \square$$

Proof (Theorem 19.1). We only need to prove the necessity part. We start by proving that if g_n is universally consistent, then the requirement $v_{n1} = \max_i v_{ni} \to 0$ is necessary. Letting $Y^\star = 2Y - 1$ and $Y_i^\star = 2Y_i - 1$, it is more convenient to consider the distribution of $(\mathbf{X}, Y^\star) \in \mathbb{R}^d \times \{-1, 1\}$, and rewrite the nearest neighbor rule as

$$g_n(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_{i=1}^n v_{ni} Y_{(i)}^\star(\mathbf{x}) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

We take $\mathbf{X} \equiv 0$, $Y^\star = 1$ with probability $p \in (1/2, 1)$, and $Y^\star = -1$ with probability $1 - p$. If (iii) does not hold, there exists $\delta > 0$ and a subsequence $\{n_\ell\}$ such that

$$v_{n_\ell 1} \ge \delta > 0, \quad \ell \ge 1.$$

Note that for all $\mathbf{x}$, $r_n(\mathbf{x}) = \sum_{i=1}^n v_{ni} Y_i^\star$ by our way of breaking ties.

Observe that $L^\star = 1 - p$, and that

$$\mathbb{E}L(g_n) = p\mathbb{P}\left\{\sum_{i=1}^n v_{ni} Y_i^\star \le 0\right\} + (1 - p)\mathbb{P}\left\{\sum_{i=1}^n v_{ni} Y_i^\star > 0\right\}.$$

Thus $\mathbb{E}L(g_n) - L^\star \to 0$ implies

$$(2p - 1)\mathbb{P}\left\{\sum_{i=1}^n v_{ni} Y_i^\star \le 0\right\} \to 0,$$

and therefore,

$$\mathbb{P}\left\{\sum_{i=1}^{n} v_{ni} Y_i^\star \le 0\right\} \to 0.$$

Writing $n = n_\ell$ to avoid an extra level of indexing, we see that

$$\mathbb{P}\left\{\sum_{i=1}^{n} v_{ni} Y_i^\star \le 0\right\} \ge p\mathbb{P}\left\{\sum_{i=2}^{n} v_{ni} Y_i^\star \le v_{n1}\right\}.$$

Note that if we set $Z = \sum_{i=2}^{n} v_{ni} Y_i^\star$, then

$$\mathbb{E}Z = (2p-1)\sum_{i=2}^{n} v_{ni} = (2p-1)(1-v_{n1}).$$

Also, $|Z| \le 1$. Therefore, by Lemma 19.1,

$$\mathbb{P}\left\{\sum_{i=1}^{n} v_{ni} Y_i^\star \le 0\right\} \ge p \times \frac{v_{n1} - (2p-1)(1-v_{n1})}{1+v_{n1}},$$

provided that $(2p-1)(1-v_{n1}) \le v_{n1}$, i.e., that $2p-1 \le 2pv_{n1}$ (since $v_{n1} \ge \delta$, it suffices to take p so close to $1/2$ that $\frac{2p-1}{2p} < \delta$). In conclusion,

$$\begin{aligned}\mathbb{P}\left\{\sum_{i=1}^{n} v_{ni} Y_i^\star \le 0\right\} &\ge p \times \frac{2pv_{n1} - (2p-1)}{1+v_{n1}} \\ &\ge p\,(2p\delta - (2p-1)) \\ &> 0.\end{aligned}$$

Thus, along the subsequence $\{n_\ell\}$, we cannot have

$$\mathbb{P}\left\{\sum_{i=1}^{n} v_{n_\ell i} Y_i^\star \le 0\right\} \to 0,$$

which provides us with a contradiction.

To complete the proof, it remains to show that conditions (i) and (ii) are necessary as well. Following Remark 10.2, this is equivalent to proving that, for all $\varepsilon > 0$, $\sum_{i>\varepsilon n} v_{ni} \to 0$. By making all components of $\mathbf{X}$ zero except one, it is easy to see that we can restrict our counterexample to $d = 1$. So, to argue by contradiction, we assume that there exists a pair $\varepsilon > 0$, $\delta > 0$ such that along a subsequence $\{n_\ell\}$,

$$\sum_{i \ge \varepsilon n} v_{ni} \ge \delta, \quad n \in \{n_1, n_2, \ldots\}.$$

Take $n = n_\ell$ for any ℓ in this subsequence. Consider the triatomic distribution given by

$$(X,Y) = \begin{cases} (0,1) & \text{w.p. } \frac{\varepsilon(1+\delta)}{2} \\ (0,0) & \text{w.p. } \frac{\varepsilon(1-\delta)}{2} \\ (1,0) & \text{w.p. } 1-\varepsilon. \end{cases}$$

Note that $L^\star = \frac{\varepsilon(1-\delta)}{2}$. Also, letting $N = \sum_{i=1}^n \mathbb{1}_{[X_i=0]}$ and $Z_i \stackrel{\mathscr{D}}{=} \text{Ber}(\frac{1+\delta}{2})$,

$$\begin{aligned} \mathbb{E}L(g_n) &\geq \mathbb{P}\Big\{X=0, Y=1, \sum_{i=1}^N v_{ni} Z_i \leq 1/2\Big\} + \mathbb{P}\Big\{X=0, Y=0, \sum_{i=1}^N v_{ni} Z_i > 1/2\Big\} \\ &= \frac{\varepsilon(1+\delta)}{2} \times \mathbb{P}\Big\{ \sum_{i=1}^N v_{ni} Z_i \leq 1/2\Big\} + \frac{\varepsilon(1-\delta)}{2} \times \mathbb{P}\Big\{ \sum_{i=1}^N v_{ni} Z_i > 1/2\Big\}. \end{aligned}$$

Thus,

$$\mathbb{E}L(g_n) - L^\star \geq \varepsilon\delta \times \mathbb{P}\Big\{ \sum_{i=1}^N v_{ni} Z_i \leq 1/2\Big\}.$$

Since $\mathbb{E}L(g_n) - L^\star \to 0$, we must have

$$\mathbb{P}\Big\{ \sum_{i=1}^N v_{ni} Z_i \leq 1/2\Big\} \to 0.$$

Observe that

$$\begin{aligned} \mathbb{P}\Big\{ \sum_{i=1}^N v_{ni} Z_i \leq 1/2\Big\} &\geq \mathbb{P}\big\{N < \varepsilon n\big\} \times \mathbb{P}\Big\{ \sum_{i<\varepsilon n} v_{ni} Z_i \leq 1/2\Big\} \\ &\qquad \text{(by the positivity of the } v_{ni}\text{'s)} \\ &= \mathbb{P}\big\{\text{Bin}(n,\varepsilon) < \varepsilon n\big\} \times \mathbb{P}\Big\{ \sum_{i<\varepsilon n} v_{ni} Z_i \leq 1/2\Big\} \\ &= (1/2 + \text{o}(1))\, \mathbb{P}\Big\{ \sum_{i<\varepsilon n} v_{ni} Z_i \leq 1/2\Big\}. \end{aligned}$$

Thus,

$$\mathbb{P}\Big\{ \sum_{i<\varepsilon n} v_{ni} Z_i \leq 1/2\Big\} \to 0.$$

Now, set $Z = \sum_{i<\varepsilon n} v_{ni} Z_i$, and note that

$$\mathbb{E}Z = \frac{1+\delta}{2} \sum_{i<\varepsilon n} v_{ni} \leq \frac{(1+\delta)(1-\delta)}{2} = \frac{1-\delta^2}{2},$$

and

$$\mathbb{V}Z = \frac{(1+\delta)(1-\delta)}{4} \sum_{i<\varepsilon n} v_{ni}^2 \leq \frac{(1-\delta^2)(1-\delta)}{4},$$

and so, by the Chebyshev-Cantelli inequality (see Theorem 20.11 in the Appendix),

$$\begin{aligned}
\mathbb{P}\{Z > 1/2\} &= \mathbb{P}\{Z - \mathbb{E}Z > 1/2 - \mathbb{E}Z\} \\
&\leq \mathbb{P}\{Z - \mathbb{E}Z \geq \delta^2/2\} \\
&\leq \frac{\mathbb{V}Z}{\mathbb{V}Z + \delta^4/4}.
\end{aligned}$$

We conclude

$$\mathbb{P}\{Z \leq 1/2\} \geq \frac{\delta^4/4}{\mathbb{V}Z + \delta^4/4} \geq \frac{\delta^4}{\delta^4 + (1-\delta^2)(1-\delta)} > 0,$$

which is a contradiction, since this cannot tend to zero. □

Remark 19.1. The monotonicity condition on the v_{ni}'s is in fact not needed in the necessity part. Moreover, we leave it as an exercise that without requiring $v_{ni} \geq 0$, but with $\sup_n \sum_{i=1}^n |v_{ni}| \leq c < \infty$, then conditions (19.1) can be shown to be necessary. □

19.2 An exponential inequality

This section is devoted to the proof of Theorem 19.2, which offers a beautiful exponential inequality on the difference $L(g_n) - L^\star$ for the nearest neighbor classification rule g_n. We assume the existence of a density for μ (the distribution of $\mathbf{X}$), so that we can avoid messy technicalities necessary to handle distance ties. It is stressed that, by the Borel-Cantelli lemma, Theorem 19.2 implies strong consistency of the k-nearest neighbor rule whenever $\mathbf{X}$ has an absolutely continuous distribution, provided $k \to \infty$ and $k/n \to 0$. Earlier versions of this result appeared in Beck (1979), Devroye and Györfi (1985), and Zhao (1987). In its present form, our theorem is an extension of Devroye et al. (1994, Theorem 1), who proved a density-free version under an appropriate distance tie-breaking strategy (different from ours).

Theorem 19.2 (Strong consistency). *Let $(v_{n1}, \ldots, v_{nn})$ be a probability weight vector such that $v_{n1} \geq \cdots \geq v_{nn}$ for all n, and let g_n be the corresponding nearest neighbor classification rule. Assume that μ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$. Assume, in addition, that there exists a sequence of integers $\{k\} = \{k_n\}$ such that*

$$\begin{array}{ll} (i) & k \to \infty \quad \textit{and} \quad k/n \to 0; \\ (ii) & v_{ni} = 0 \quad \textit{when } i > k; \\ (iii) & \sup_n (k v_{n1}) \leq \alpha, \quad \alpha > 0. \end{array}$$

Then, for every $\varepsilon > 0$, there is a positive integer n_0 such that, for all $n \geq n_0$,

$$\mathbb{P}\{L(g_n) - L^\star \geq \varepsilon\} \leq 4e^{-n\varepsilon^2/(50\alpha^2\gamma_d^2)},$$

where $\gamma_d = C_{\pi/6}$ (see Theorem 20.15) depends only upon d. Thus, the nearest neighbor classification rule is strongly consistent.

The basic result that will be needed in the proof of Theorem 19.2 is the following one:

Lemma 19.2. *For $a \geq 0$, let*

$$S_a(\mathbf{x}) = \left\{\mathbf{x}' \in \mathbb{R}^d : \mu\big(B(\mathbf{x}', \|\mathbf{x} - \mathbf{x}'\|)\big) \leq a\right\}.$$

Then, for all $\mathbf{x} \in \mathbb{R}^d$,

$$\mu\left(S_a(\mathbf{x})\right) \leq \gamma_d a.$$

Proof. Let $\mathscr{C}_1(\mathbf{x}), \ldots, \mathscr{C}_{\gamma_d}(\mathbf{x})$ be a collection of cones of angle $\pi/6$ covering $\mathbb{R}^d$, all centered at $\mathbf{x}$ but with different central directions (such a covering is possible by Theorem 20.15). Then

$$\mu\left(S_a(\mathbf{x})\right) \leq \sum_{\ell=1}^{\gamma_d} \mu\big(\mathscr{C}_\ell(\mathbf{x}) \cap S_a(\mathbf{x})\big).$$

Let $\mathbf{x}' \in \mathscr{C}_\ell(\mathbf{x}) \cap S_a(\mathbf{x})$. Then, by the geometrical property of cones shown in Lemma 20.5, we have

$$\mu\left(\mathscr{C}_\ell(\mathbf{x}) \cap B(\mathbf{x}, \|\mathbf{x}' - \mathbf{x}\|) \cap S_a(\mathbf{x})\right) \leq \mu\left(B(\mathbf{x}', \|\mathbf{x} - \mathbf{x}'\|)\right) \leq a,$$

where we used the fact that $\mathbf{x}' \in S_a(\mathbf{x})$. Since $\mathbf{x}'$ was arbitrary,

$$\mu\big(\mathscr{C}_\ell(\mathbf{x}) \cap S_a(\mathbf{x})\big) \leq a. \quad \square$$

Proof (Theorem 19.2). We are now ready to prove the theorem. The decision rule g_n may be rewritten as

$$g_n(\mathbf{x}) = \begin{cases} 1 & \text{if } r_n(\mathbf{x}) > 1/2 \\ 0 & \text{otherwise,} \end{cases}$$

where r_n is the companion regression function estimate, that is,

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

or, equivalently,

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

where $W_{ni}(\mathbf{x}) = v_{n\Sigma_i}$ and $(\Sigma_1, \ldots, \Sigma_n)$ is a permutation of $(1, \ldots, n)$ such that $\mathbf{X}_i$ is the Σ_i-th nearest neighbor of $\mathbf{x}$ for all i (note, since $\mathbf{x}$ has a density, that distance ties do not matter). Thus, by Theorem 17.1, the statement follows if we show that for sufficiently large n,

$$\mathbb{P}\left\{ \int_{\mathbb{R}^d} |r_n(\mathbf{x}) - r(\mathbf{x})| \, \mu(\mathrm{d}\mathbf{x}) \geq \frac{\varepsilon}{2} \right\} \leq 4e^{-n\varepsilon^2/(50\alpha^2\gamma_d^2)}.$$

Set $\rho_n(\mathbf{x})$ to satisfy

$$\mu\big(B\left(\mathbf{x}, \rho_n(\mathbf{x})\right)\big) = \frac{k}{n}.$$

Notice that the solution always exists, by the absolute continuity of μ. (This is the only point in the proof where we use this assumption.) Also define

$$r_n^\star(\mathbf{x}) = \sum_{i=1}^{n} W_{ni}(\mathbf{x}) Y_i \mathbb{1}_{[\mathbf{X}_i \in B(\mathbf{x}, \rho_n(\mathbf{x}))]}.$$

The basis of the proof is the decomposition

$$|r_n(\mathbf{x}) - r(\mathbf{x})| \leq \left|r_n^\star(\mathbf{x}) - r_n(\mathbf{x})\right| + \left|r_n^\star(\mathbf{x}) - r(\mathbf{x})\right|. \tag{19.2}$$

For the second term on the right-hand side, set $R_{(k)}(\mathbf{x}) = \|\mathbf{X}_{(k)}(\mathbf{x}) - \mathbf{x}\|$ and observe that, by conditions (*ii*) and (*iii*),

$$\begin{aligned}
\left|r_n^\star(\mathbf{x}) - r_n(\mathbf{x})\right| &= \left| \sum_{i=1}^n W_{ni}(\mathbf{x}) Y_i \mathbb{1}_{[\mathbf{X}_i \in B(\mathbf{x}, \rho_n(\mathbf{x}))]} - \sum_{i=1}^n W_{ni}(\mathbf{x}) Y_i \mathbb{1}_{[\mathbf{X}_i \in B(\mathbf{x}, R_{(k)}(\mathbf{x}))]} \right| \\
&\le \frac{\alpha}{k} \sum_{i=1}^n \left| \mathbb{1}_{[\mathbf{X}_i \in B(\mathbf{x}, \rho_n(\mathbf{x}))]} - \mathbb{1}_{[\mathbf{X}_i \in B(\mathbf{x}, R_{(k)}(\mathbf{x}))]} \right| \\
&\le \frac{\alpha}{k} \left| n\mu_n\big(B\left(\mathbf{x}, \rho_n(\mathbf{x})\right)\big) - k \right| \\
&= \frac{\alpha n}{k} \left| \mu_n\big(B\left(\mathbf{x}, \rho_n(\mathbf{x})\right)\big) - \mu\big(B\left(\mathbf{x}, \rho_n(\mathbf{x})\right)\big) \right| \\
&\stackrel{\text{def}}{=} Z_n(\mathbf{x}),
\end{aligned}$$

where μ_n is the empirical measure. Note that

$$\begin{aligned}
\mathbb{E} Z_n(\mathbf{x}) &\le \frac{\alpha n}{k} \sqrt{\mathbb{V} \mu_n\big(B\left(\mathbf{x}, \rho_n(\mathbf{x})\right)\big)} \\
&= \frac{\alpha n}{k} \sqrt{\frac{k}{n}\left(1 - \frac{k}{n}\right)\frac{1}{n}} \\
&\le \frac{\alpha}{\sqrt{k}},
\end{aligned}$$

uniformly over $\mathbf{x}$. Also, since $\mathbb{E}|r_n(\mathbf{X}) - r(\mathbf{X})| = \mathrm{o}(1)$ (Theorem 10.2), we have

$$\mathbb{E}\left|r_n^\star(\mathbf{X}) - r(\mathbf{X})\right| \le \frac{\alpha}{\sqrt{k}} + \mathrm{o}(1) < \frac{\epsilon}{20}$$

for all n large enough. Thus, by (19.2),

$$\begin{aligned}
&\mathbb{P}\left\{ \int_{\mathbb{R}^d} |r_n(\mathbf{x}) - r(\mathbf{x})| \, \mu(\mathrm{d}\mathbf{x}) \ge \frac{\varepsilon}{2} \right\} \\
&\quad \le \mathbb{P}\left\{ \left| \int_{\mathbb{R}^d} Z_n(\mathbf{x}) \mu(\mathrm{d}\mathbf{x}) - \mathbb{E} \int_{\mathbb{R}^d} Z_n(\mathbf{x}) \mu(\mathrm{d}\mathbf{x}) \right| \ge \frac{\varepsilon}{5} \right\} \\
&\qquad + \mathbb{P}\left\{ \left| \int_{\mathbb{R}^d} \left|r_n^\star(\mathbf{x}) - r(\mathbf{x})\right| \mu(\mathrm{d}\mathbf{x}) - \mathbb{E} \int_{\mathbb{R}^d} \left|r_n^\star(\mathbf{x}) - r(\mathbf{x})\right| \mu(\mathrm{d}\mathbf{x}) \right| \ge \frac{\varepsilon}{5} \right\}.
\end{aligned} \tag{19.3}$$

To begin with, we get an exponential bound for the second probability on the right-hand side of (19.3) by the bounded difference inequality (Theorem 20.9 in the Appendix). Fix the data and replace $(\mathbf{x}_i, y_i)$ by $(\hat{\mathbf{x}}_i, \hat{y}_i)$, changing the value of $r_n^\star(\mathbf{x})$ to $r_{ni}^\star(\mathbf{x})$. Then

$$\left| \int_{\mathbb{R}^d} \left|r_n^\star(\mathbf{x}) - r(\mathbf{x})\right| \mu(\mathrm{d}\mathbf{x}) - \int_{\mathbb{R}^d} \left|r_{ni}^\star(\mathbf{x}) - r(\mathbf{x})\right| \mu(\mathrm{d}\mathbf{x}) \right| \le \int_{\mathbb{R}^d} \left|r_n^\star(\mathbf{x}) - r_{ni}^\star(\mathbf{x})\right| \mu(\mathrm{d}\mathbf{x}).$$

But $|r_n^\star(\mathbf{x}) - r_{ni}^\star(\mathbf{x})|$ is bounded by α/k and can differ from zero only if $\mathbf{x}_i \in B(\mathbf{x}, \rho_n(\mathbf{x}))$ or $\hat{\mathbf{x}}_i \in B(\mathbf{x}, \rho_n(\mathbf{x}))$. Observe that $\mathbf{x}_i \in B(\mathbf{x}, \rho_n(\mathbf{x}))$ if and only if

$$\mu\big(B(\mathbf{x}, \|\mathbf{x}_i - \mathbf{x}\|)\big) \leq \frac{k}{n}.$$

However, the measure of such $\mathbf{x}$'s is bounded by $\gamma_d k/n$ by Lemma 19.2. Therefore,

$$\sup_{\mathbf{x}_1, y_1, \ldots, \mathbf{x}_n, y_n, \hat{\mathbf{x}}_i, \hat{y}_i} \int_{\mathbb{R}^d} \left| r_n^\star(\mathbf{x}) - r_{ni}^\star(\mathbf{x}) \right| \mu(\mathrm{d}\mathbf{x}) \leq \frac{\alpha}{k} \times \frac{2\gamma_d k}{n} = \frac{2\alpha\gamma_d}{n},$$

and, by the bounded difference inequality,

$$\mathbb{P}\left\{ \left| \int_{\mathbb{R}^d} \left| r_n^\star(\mathbf{x}) - r(\mathbf{x}) \right| \mu(\mathrm{d}\mathbf{x}) - \mathbb{E} \int_{\mathbb{R}^d} \left| r_n^\star(\mathbf{x}) - r(\mathbf{x}) \right| \mu(\mathrm{d}\mathbf{x}) \right| \geq \frac{\varepsilon}{5} \right\} \leq 2e^{-n\varepsilon^2/(50\alpha^2\gamma_d^2)}.$$

Finally, we need a bound for the first term on the right-hand side of (19.3). This probability may be bounded by the bounded difference inequality in exactly the same way as for the first one, obtaining

$$\mathbb{P}\left\{ \left| \int_{\mathbb{R}^d} Z_n(\mathbf{x}) \mu(\mathrm{d}\mathbf{x}) - \mathbb{E} \int_{\mathbb{R}^d} Z_n(\mathbf{x}) \mu(\mathrm{d}\mathbf{x}) \right| \geq \frac{\varepsilon}{5} \right\} \leq 2e^{-n\varepsilon^2/(50\alpha^2\gamma_d^2)},$$

and the proof is complete. □

Chapter 20
Appendix

20.1 Some basic concepts

For any real-valued function g, define $g^+ = \max(g, 0)$ and $g^- = -\min(g, 0)$. These are called the positive and negative parts of g, respectively, and satisfy the relations

$$g^+, g^- \geq 0, \quad |g| = g^+ + g^-, \quad \text{and} \quad g = g^+ - g^-.$$

We recall that a real-valued random variable X is said to be integrable if $\mathbb{E}|X| < \infty$, or, equivalently, if $\mathbb{E}X^+ < \infty$ and $\mathbb{E}X^- < \infty$. In that case, $\mathbb{E}X = \mathbb{E}X^+ - \mathbb{E}X^-$. We use $\|X\|_\infty$ to denote the essential supremum of X:

$$\|X\|_\infty = \inf\left\{t \geq 0 : \mathbb{P}\{|X| > t\} = 0\right\}.$$

There are several notions of convergence for random variables, summarized below.

Definition 20.1. Let $\{X_n\}$ be a sequence of real-valued random variables.

(*i*) We say that

$$X_n \to X \quad \text{in probability}$$

if, for all $\varepsilon > 0$,

$$\mathbb{P}\{|X_n - X| > \varepsilon\} \to 0 \quad \text{as } n \to \infty.$$

(*ii*) We say that

$$X_n \to X \quad \text{almost surely}$$

if

$$\mathbb{P}\Big\{\lim_{n\to\infty} X_n = X\Big\} = 1.$$

G. Biau, L. Devroye, *Lectures on the Nearest Neighbor Method*,
Springer Series in the Data Sciences, DOI 10.1007/978-3-319-25388-6_20

(*iii*) Let $p > 0$. We say that

$$X_n \to X \quad \text{in } L^p$$

if

$$\mathbb{E}|X_n - X|^p \to 0 \quad \text{as } n \to \infty.$$

Almost sure convergence implies convergence in probability. Moreover, convergence in L^p implies convergence in probability, and none of these implications is an equivalence.

For any sequence of events $\{A_n\}$, the event $[A_n \text{ i.o.}]$ ("i.o." stands for "infinitely often") is defined by

$$[A_n \text{ i.o.}] = \limsup_{n\to\infty} A_n \stackrel{\text{def}}{=} \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m.$$

The Borel-Cantelli lemma states that if

$$\sum_{n\geq 1} \mathbb{P}\{A_n\} < \infty,$$

then

$$\mathbb{P}\{A_n \text{ i.o.}\} = 0.$$

In particular, if

$$\sum_{n\geq 1} \mathbb{P}\left\{|X_n - X| > \varepsilon\right\} < \infty \tag{20.1}$$

holds for all $\varepsilon > 0$ small enough, then

$$\mathbb{P}\Big\{|X_n - X| > \varepsilon \text{ i.o.}\Big\} = \mathbb{P}\Big\{\limsup_{n\to\infty} |X_n - X| > \varepsilon\Big\} = 0.$$

Thus, with probability one, $\lim_{n\to\infty} |X_n - X| = 0$, i.e., $X_n \to X$ almost surely. In other words, condition (20.1) is sufficient for almost sure convergence.

Definition 20.2. Let $\{u_n\}$ be a sequence of real numbers.

(*i*) We say that $\{X_n\}$ is $\mathrm{o}_{\mathbb{P}}(u_n)$ as $n \to \infty$ if $X_n/u_n \to 0$ in probability.
(*ii*) We say that $\{X_n\}$ is $\mathrm{O}_{\mathbb{P}}(u_n)$ if, for all $\varepsilon > 0$, there exists a finite $M > 0$ such that, for all n large enough,

$$\mathbb{P}\left\{|X_n/u_n| > M\right\} \leq \varepsilon.$$

In particular, $\{X_n\}$ is $\mathrm{o}_{\mathbb{P}}(1)$ if $X_n \to 0$ in probability, and it is $\mathrm{O}_{\mathbb{P}}(1)$ if the sequence is bounded in probability.

The concept of convergence in distribution is central in modern statistics.

Definition 20.3. The real-valued sequence $\{X_n\}$ converges in distribution to X (written $X_n \xrightarrow{\mathscr{D}} X$) if

$$\mathbb{P}\{X_n \le x\} \to \mathbb{P}\{X \le x\}$$

for all $x \in \mathbb{R}$ at which $F(x) = \mathbb{P}\{X \le x\}$ is continuous.

A sequence $\{X_n\}$ is asymptotically (standard) normal if for all x,

$$\mathbb{P}\{X_n \le x\} \to \phi(x) \stackrel{\text{def}}{=} \mathbb{P}\{N \le x\},$$

where N denotes a standard normal random variable, and

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} \mathrm{d}t$$

is the normal distribution function. If $X_1, X_2, \ldots$ are i.i.d. with finite mean μ, then the law of large numbers (respectively the strong law of large numbers) asserts that

$$\frac{1}{n} \sum_{i=1}^{n} X_i \to \mu \quad \text{in probability (respectively almost surely).}$$

Moreover, if the X_i's have finite variance $\sigma^2 > 0$, then the central limit theorem states that

$$\sqrt{n}\, \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu}{\sigma} \xrightarrow{\mathscr{D}} N.$$

Notice that this implies that $(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu)/\sigma = \mathrm{O}_{\mathbb{P}}(1/\sqrt{n})$.

The reader should keep in mind the following equivalence:

Lemma 20.1. *We have $X_n \xrightarrow{\mathscr{D}} X$ if and only if there exists a sequence of random variables $\{Y_n\}$ such that*

$$Y_n \to 0 \quad \textit{in probability} \quad \textit{and} \quad X_n \stackrel{\mathscr{D}}{=} X + Y_n. \tag{20.2}$$

This is an easy implication of Skorohod's representation theorem (see Billingsley, 1995). We write $X_n \stackrel{\mathscr{D}}{=} X + \mathrm{o}_{\mathbb{P}}(1)$ and stress that Y_n in (20.2) is generally dependent on X. Let us finally mention the following result, which is frequently encountered in consistency proofs.

Theorem 20.1 (Slutsky's theorem). *Let $\{X_n\}$, $\{Y_n\}$, and X be real-valued random variables. Assume that $X_n \xrightarrow{\mathscr{D}} X$ and that $Y_n \to y$ in probability for some $y \in \mathbb{R}$. Then $X_n Y_n \xrightarrow{\mathscr{D}} Xy$ and $X_n/Y_n \xrightarrow{\mathscr{D}} X/y$ if $y \neq 0$.*

20.2 Convergence theorems

The following two convergence theorems are central in measure and integration theory. They are used at various places in the book.

Theorem 20.2 (Fatou's lemma). *Let $(\Omega, \mathscr{A}, \mu)$ be a measure space, and let $g_n : \Omega \to [0, \infty)$ be a sequence of nonnegative measurable functions. Then*

$$\int_\Omega \liminf_{n\to\infty} g_n \, \mathrm{d}\mu \le \liminf_{n\to\infty} \int_\Omega g_n \, \mathrm{d}\mu.$$

(Reverse Fatou's lemma). *In addition, if $\sup_n g_n \le g$ μ-almost everywhere and $\int_\Omega |g| \mathrm{d}\mu < \infty$, then*

$$\limsup_{n\to\infty} \int_\Omega g_n \, \mathrm{d}\mu \le \int_\Omega \limsup_{n\to\infty} g_n \, \mathrm{d}\mu.$$

Theorem 20.3 (The Lebesgue dominated convergence theorem). *Let $(\Omega, \mathscr{A}, \mu)$ be a measure space, and let $g_n : \Omega \to [0, \infty)$ be a sequence of measurable functions. Assume that $g_n \to g$ μ-almost everywhere and that $\sup_n |g_n| \le h$ μ-almost everywhere, with $\int_\Omega |h| \mathrm{d}\mu < \infty$. Then $\int_\Omega |g| \mathrm{d}\mu < \infty$ and*

$$\int_\Omega |g_n - g| \mathrm{d}\mu \to 0.$$

In particular,

$$\int_\Omega g_n \, \mathrm{d}\mu \to \int_\Omega g \, \mathrm{d}\mu.$$

The next theorem is an extension of the Lebesgue dominated convergence theorem tailored to random variables.

Theorem 20.4. *Let $\{X_n\}$ be a sequence of real-valued random variables. Assume that*

$$\sup_n \mathbb{E}\left[|X_n| \mathbb{1}_{[|X_n|>K]}\right] \to 0 \quad \text{as } K \to \infty. \tag{20.3}$$

If $X_n \xrightarrow{\mathscr{D}} X$, then $\mathbb{E}|X| < \infty$ and $\mathbb{E}X_n \to \mathbb{E}X$. If $X_n \to X$ in probability, then $\mathbb{E}|X| < \infty$ and $\mathbb{E}|X_n - X| \to 0$. Conversely, if $\mathbb{E}|X| < \infty$ and $\mathbb{E}|X_n - X| \to 0$, then $X_n \to X$ in probability and (20.3) is satisfied.

A sequence $\{X_n\}$ of random variables verifying condition (20.3) is called uniformly integrable. This property is satisfied, for example, if $\sup_n |X_n| \le Z$ almost surely with $\mathbb{E}|Z| < \infty$, or if $\sup_n \mathbb{E}|X_n|^{1+\varepsilon} < \infty$ for some $\varepsilon > 0$. In the same spirit, we have the following lemma:

Lemma 20.2 (A generalized Lebesgue dominated convergence theorem). *Let $\{X_n\}$, $\{Y_n\}$, X, Y and Z be nonnegative real-valued random variables. Assume that $X_n \to X$ in probability, that $\sup_n X_n \le Z$ with probability one, that $Y_n \xrightarrow{\mathscr{D}} Y$, that (X, Z) and Y_n are independent for all n large enough, and that $\mathbb{E}Z < \infty$. Assume furthermore that*

$$\sup_n \mathbb{E}[Y_n \mathbb{1}_{[Y_n > K]}] \to 0 \quad \text{as } K \to \infty. \tag{20.4}$$

Then $\mathbb{E}X < \infty$, $\mathbb{E}Y < \infty$, and $\mathbb{E}[X_n Y_n] \to \mathbb{E}X\mathbb{E}Y$.

Proof. For $\varepsilon > 0$, $\delta > 0$,

$$\begin{aligned}
\mathbb{E}[X_n Y_n] &\le \mathbb{E}\left[(X+\varepsilon)Y_n\right] + \mathbb{E}[ZY_n \mathbb{1}_{[X_n > X+\varepsilon]}] \\
&\le \mathbb{E}[X+\varepsilon]\mathbb{E}Y_n + K\,\mathbb{E}[Z\mathbb{1}_{[X_n > X+\varepsilon]}] + \mathbb{E}[ZY_n \mathbb{1}_{[Y_n > K]}] \\
&\quad \text{(for arbitrary positive } K\text{)} \\
&= (\mathbb{E}X + \varepsilon)\left(\mathbb{E}Y + \mathrm{o}(1)\right) + \mathrm{o}(1) + \mathbb{E}Z\,\mathbb{E}[Y_n \mathbb{1}_{[Y_n > K]}]
\end{aligned}$$

since $\mathbb{E}Y_n \to \mathbb{E}Y$ by (20.4), and by the Lebesgue dominated convergence theorem for the middle term. Therefore, for all n large enough,

$$\mathbb{E}[X_n Y_n] < \mathbb{E}X\mathbb{E}Y + \delta$$

by choice of ε and K. A matching lower bound is obtained in a similar fashion. □

20.3 Chernoff's bounds

20.3.1 Binomial random variables

The following theorem offers bounds for the upper and lower tail probabilities of a binomial random variable. It is due to Chernoff (1952) (see also Karp, 1988, and Hagerup and Rüb, 1990).

Theorem 20.5 (Chernoff, 1952). *Let Z be a binomial random variable with parameters $n \ge 1$ and $p \in (0, 1]$.*

(i) Let

$$\psi(t) = \left(\frac{np}{t}\right)^t \left(\frac{n(1-p)}{n-t}\right)^{n-t}, \quad 0 < t < n.$$

Then

$$\mathbb{P}\{Z \ge t\} \le \psi(t) \quad \text{for } np \le t < n$$

and

$$\mathbb{P}\{Z \le t\} \le \psi(t) \quad \textit{for } 0 < t < np.$$

(*ii*) *Let* $\varphi(t) = t - np - t\log(t/np)$. *Then*

$$\mathbb{P}\{Z \ge t\} \le e^{\varphi(t)} \quad \textit{for } t \ge np$$

and

$$\mathbb{P}\{Z \le t\} \le e^{\varphi(t)} \quad \textit{for } 0 < t \le np.$$

Proof. We only show the upper tail bounds and use the so-called Chernoff's bounding method. By Markov's inequality, for $\lambda \ge 0$,

$$\mathbb{P}\{Z \ge t\} \le \mathbb{E}[e^{\lambda Z - \lambda t}] = (1 - p + pe^{\lambda})^n e^{-\lambda t}.$$

Minimizing the upper bound with respect to λ yields the following equation for the optimal $\lambda^\star$:

$$\frac{npe^{\lambda^\star}}{1 - p + pe^{\lambda^\star}} = t, \quad \text{or} \quad e^{\lambda^\star} p(n-t) = (1-p)t.$$

The Chernoff's bound becomes

$$\left(\frac{npe^{\lambda^\star}}{t}\right)^n e^{-\lambda^\star t} = \left(\frac{np}{t}\right)^n \left(\frac{(1-p)t}{p(n-t)}\right)^{n-t} = \frac{n^n p^t (1-p)^{n-t}}{t^t (n-t)^{n-t}}.$$

This proves the first statement. The proof of (*ii*) is similar—just note that, in view of $1 + u \le e^u$, $u \in \mathbb{R}$,

$$\mathbb{P}\{Z \ge t\} \le e^{np(e^{\lambda}-1)-\lambda t},$$

and set $\lambda^\star = \log(t/np)$. The lower tail bound is obtained in the same way. □

The second set of inequalities of Theorem 20.5, though less sharp, are more tractable than the first ones. In any case, they are more than sufficient for the purpose of the book. The next corollary provides us with useful exponential bounds. It is but a special form of a more general result for sums of bounded random variables, due to Hoeffding (1963) (see Theorem 20.7 in the next section). Note, however, that our proof is interesting in its own right since it is tailored to binomial random variables. By construction, the bounds of Theorem 20.5(*i*) are sharper than the bounds of Corollary 20.1. We leave it as an exercise to prove that the bounds of Theorem 20.5(*ii*) are better whenever $p \le 1/4$.

Corollary 20.1. *Let Z be a binomial random variable with parameters $n \geq 1$ and $p \in [0,1]$. Then, for all $t > 0$,*

$$\mathbb{P}\{Z - np \geq t\} \leq e^{-2t^2/n}$$

and

$$\mathbb{P}\{Z - np \leq -t\} \leq e^{-2t^2/n}.$$

In particular,

$$\mathbb{P}\{|Z - np| \geq t\} \leq 2e^{-2t^2/n}.$$

Proof. We only show the upper tail bound, and assume that $p \in (0,1)$ and $t < n(1-p)$, for otherwise the proof is trivial. According to statement (i) of Theorem 20.5, for all $t > 0$,

$$\mathbb{P}\{Z - np \geq t\} \leq e^{-h(t)},$$

where

$$h(t) = (np + t)\log\left(1 + \frac{t}{np}\right) + (n(1-p) - t)\log\left(1 - \frac{t}{n(1-p)}\right).$$

Note that

$$h'(t) = \log\left(1 + \frac{t}{np}\right) - \log\left(1 - \frac{t}{n(1-p)}\right)$$

and

$$h''(t) = \frac{1}{np + t} + \frac{1}{n(1-p) - t}.$$

Since $h(0) = h'(0) = 0$, a Taylor series expansion shows that, for some $\theta \in [0,1]$,

$$h(t) = \frac{t^2}{2}\left(\frac{1}{np + t\theta} + \frac{1}{n(1-p) - t\theta}\right) = \frac{t^2}{2}\xi(t\theta),$$

where

$$\begin{aligned}\xi(s) &\stackrel{\text{def}}{=} \frac{1}{np + s} + \frac{1}{n(1-p) - s} = \frac{n}{(np + s)\,(n(1-p) - s)} \\ &= \frac{n}{n^2 p(1-p) + sn(1-2p) - s^2}, \quad s \geq 0.\end{aligned}$$

Since $t < n(1-p)$ and $\min_{0 \leq s \leq t} \xi(s) \geq 4/n$, we conclude $h(t) \geq 2t^2/n$. □

20.3.2 Gamma random variables

Theorem 20.6. *Let G_n be a gamma random variable with parameter $n > 0$. Then, for all $t > 0$,*

$$\mathbb{P}\{G_n \geq n(1+t)\} \leq \exp\big(-n\,[t - \log(1+t)]\big)$$

and, for all $t \in (0,1)$,

$$\mathbb{P}\{G_n \leq n(1-t)\} \leq \exp\big(-n\,[-t - \log(1-t)]\big).$$

Proof. We employ Chernoff's bounding method applied to the gamma density. For general n, we have, if $t > 0$ and $\lambda \in [0,1)$,

$$\mathbb{P}\{G_n \geq n(1+t)\} \leq \mathbb{E}\left[\exp(\lambda G_n - \lambda n - \lambda n t)\right] = \left(\frac{1}{1-\lambda}\right)^n e^{-\lambda n(1+t)}.$$

The upper bound is minimized for $\lambda = \frac{t}{1+t}$, and then

$$\mathbb{P}\{G_n \geq n(1+t)\} \leq (1+t)^n e^{-nt} = \exp\big(-n\,[t - \log(1+t)]\big).$$

Similarly, for $t \in (0,1)$, $\lambda \geq 0$,

$$\mathbb{P}\{G_n \leq n(1-t)\} \leq \mathbb{E}\left[\exp(-\lambda G_n + \lambda n - \lambda n t)\right] = \left(\frac{1}{1+\lambda}\right)^n e^{\lambda n(1-t)}.$$

Since the upper bound is minimized for $\lambda = \frac{t}{1-t}$, we obtain

$$\mathbb{P}\{G_n \leq n(1-t)\} \leq (1-t)^n e^{nt} = \exp\big(-n\,[-t - \log(1-t)]\big). \quad \square$$

Example 20.1. Let $U_{(1)}, \ldots, U_{(n)}$ be uniform $[0,1]$ order statistics, and let E be a standard exponential random variable. Clearly, $\mathbb{E}|\log(\frac{1}{E})| < \infty$ and, since $U_{(1)}$ is Beta$(1,n)$, $\mathbb{E}|\log(\frac{1}{nU_{(1)}})| < \infty$. Notice that $\gamma = \mathbb{E}\log(\frac{1}{E}) = -\int_0^\infty e^{-t}\log t\,\mathrm{d}t = 0.577215664901532\ldots$ is the Euler-Mascheroni constant. Similarly, for $q \in (0,1)$, $\mathbb{E}|\frac{1}{E}|^q < \infty$ and $\mathbb{E}|\frac{1}{nU_{(1)}}|^q < \infty$. Let us show, by way of example, that

$$\mathbb{E}\log\left(\frac{1}{nU_{(1)}}\right) \to \mathbb{E}\log\left(\frac{1}{E}\right) \tag{20.5}$$

and that, for $q \in (0,1)$,

$$\mathbb{E}\left|\left(\frac{1}{nU_{(1)}}\right)^q - \left(\frac{1}{E}\right)^q\right| \to 0. \tag{20.6}$$

(Note that, by Theorem 20.4, consistency (20.6) implies the uniform integrability of the sequence $\{(\frac{1}{nU_{(1)}})^q\}$.)

Recall (Corollary 1.1) that

$$U_{(1)} \stackrel{\mathscr{D}}{=} \frac{E_1}{G_{n+1}},$$

where $G_{n+1} = \sum_{i=1}^{n+1} E_i$ and $E_1, \ldots, E_{n+1}$ are independent standard exponential random variables. Therefore,

$$\frac{1}{nU_{(1)}} \stackrel{\mathscr{D}}{=} \frac{G_{n+1}}{nE_1}.$$

Thus, in view of

$$\log\left(\frac{G_{n+1}}{nE_1}\right) = \log\left(\frac{G_{n+1}}{n}\right) + \log\left(\frac{1}{E_1}\right),$$

identity (20.5) follows if $\mathbb{E}\log(\frac{G_{n+1}}{n}) \to 0$. On the one hand, by Jensen's inequality,

$$\limsup_{n\to\infty} \mathbb{E}\log\left(\frac{G_{n+1}}{n}\right) \le \limsup_{n\to\infty} \log \mathbb{E}\left(\frac{G_{n+1}}{n}\right) = 0.$$

So, we only need to show that

$$\limsup_{n\to\infty} \mathbb{E}\log\left(\frac{n}{G_{n+1}}\right) \le 0. \tag{20.7}$$

But, letting $G_n = \sum_{i=1}^{n} E_i$, we see that

$$\begin{aligned}
\mathbb{E}\log\left(\frac{n}{G_{n+1}}\right) &\le \mathbb{E}\log\left(\frac{n}{G_n}\right) \\
&\le \int_0^\infty \mathbb{P}\left\{\log\left(\frac{n}{G_n}\right) > t\right\} \mathrm{d}t \\
&= \int_0^\infty \mathbb{P}\left\{G_n < ne^{-t}\right\} \mathrm{d}t.
\end{aligned}$$

Recalling that G_n is Gamma(n) distributed, and evoking Theorem 20.6, we conclude that

$$\mathbb{E}\log\left(\frac{n}{G_{n+1}}\right) \le \int_0^\infty e^{-n(e^{-t}+t-1)} \mathrm{d}t.$$

Since $\int_0^\infty e^{-(e^{-t}+t-1)} \mathrm{d}t < \infty$, (20.7) follows by the Lebesgue dominated convergence theorem.

The proof of (20.6) starts from the observation that, by the c_r-inequality (Proposition 20.1),

$$\left(\frac{1}{E_1}\right)^q \left(\frac{S_n}{n}\right)^q \le \left(\frac{G_{n+1}}{nE_1}\right)^q \le \frac{1}{n^q} + \left(\frac{1}{E_1}\right)^q \left(\frac{S_n}{n}\right)^q,$$

where $S_n = \sum_{i=2}^{n+1} E_i$. Since S_n is independent of E_1, we only need to show that $\mathbb{E}|(\frac{S_n}{n})^q - 1| \to 0$. This is achieved by following what we did for the log. We leave it to the reader to play with the arguments above and prove, for example, that

$$\mathbb{E}\log^2(nU_{(1)}) \to \mathbb{E}\log^2 E \quad \text{and} \quad \mathbb{E}\log^2(nU_{(2)}) \to \mathbb{E}\log^2 G_2. \quad \square$$

20.4 Inequalities for independent random variables

In this section, we collect without proofs some of the classical inequalities for tail probabilities of sums of independent real-valued random variables. For more advanced material, the reader is referred to the textbooks by Massart (2007) and Boucheron et al. (2013).

Theorem 20.7 (Hoeffding, 1963). *Let $X_1, \ldots, X_n$ be independent real-valued random variables. Assume that each X_i takes its values in $[a_i, b_i]$ ($a_i < b_i$) with probability one, $1 \le i \le n$. Then, for all $t > 0$,*

$$\mathbb{P}\left\{\sum_{i=1}^n (X_i - \mathbb{E}X_i) \ge t\right\} \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

and

$$\mathbb{P}\left\{\sum_{i=1}^n (X_i - \mathbb{E}X_i) \le -t\right\} \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

In particular,

$$\mathbb{P}\left\{\left|\sum_{i=1}^n (X_i - \mathbb{E}X_i)\right| \ge t\right\} \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Consider the special case in which all X_i's take values in $[-c, c]$ ($c > 0$). Then Hoeffding's inequality states that

$$\mathbb{P}\left\{\left|\sum_{i=1}^n (X_i - \mathbb{E}X_i)\right| \ge t\right\} \le 2\exp\left(-\frac{t^2}{2nc^2}\right).$$

This bound, while useful for t larger than $c\sqrt{n}$, ignores variance information. When $\mathbb{E}X_i^2 \ll c^2$, it is indeed possible to outperform Hoeffding's inequality. Bennett's and Bernstein's inequalities provide such improvements.

Theorem 20.8 (Bennett, 1962; Bernstein, 1946). *Let $X_1, \ldots, X_n$ be independent real-valued random variables with finite variance. Assume that $X_i \le c$ with probability one for some $c > 0$, $1 \le i \le n$. Let*

$$s^2 = \sum_{i=1}^{n} \mathbb{E}X_i^2.$$

Then, for all $t > 0$,

$$\mathbb{P}\left\{ \sum_{i=1}^{n} (X_i - \mathbb{E}X_i) \ge t \right\} \le \exp\left(-\frac{t}{c} \left[\left(1 + \frac{s^2}{ct} \right) \log\left(1 + \frac{ct}{s^2} \right) - 1 \right] \right)$$

(Bennett, 1962), and

$$\mathbb{P}\left\{ \sum_{i=1}^{n} (X_i - \mathbb{E}X_i) \ge t \right\} \le \exp\left(-\frac{t^2}{2(s^2 + ct/3)} \right)$$

(Bernstein, 1946).

For $t \gg s^2/c$, Bernstein's inequality losses a logarithmic factor in the exponent with respect to Bennett's inequality. On the other hand, if s^2 is the dominant term in the denominator of the exponent, Bennett's and Bernstein's inequalities are almost equivalent. We also note that Bernstein's inequality is typically better than Hoeffding's inequality when $s^2 \ll nc^2$.

The next result, due to McDiarmid (1989), is called the bounded difference inequality. It generalizes Hoeffding's inequality to functions of independent random variables that are more complicated than simple sums, and that are relatively robust to individual changes in the values of the variables. It has found many applications in combinatorics as well as in nonparametric statistics (see, e.g., Devroye, 1991a, for a survey).

Theorem 20.9 (McDiarmid, 1989). *Let $X_1, \ldots, X_n$ be independent real-valued random variables taking values in a set A. Assume that $g : A^n \to \mathbb{R}$ is Borel measurable and satisfies*

$$\sup_{\substack{(x_1,\ldots,x_n)\in A^n \\ x_i' \in A}} \left| g(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n) \right| \le c_i, \quad 1 \le i \le n,$$

for some positive constants $c_1, \ldots, c_n$. Then, for all $t > 0$,

$$\mathbb{P}\left\{ g(X_1, \ldots, X_n) - \mathbb{E}g(X_1, \ldots, X_n) \ge t \right\} \le e^{-2t^2 / \sum_{i=1}^{n} c_i^2}$$

and

$$\mathbb{P}\{g(X_1,\ldots,X_n) - \mathbb{E}g(X_1,\ldots,X_n) \le -t\} \le e^{-2t^2/\sum_{i=1}^n c_i^2}.$$

In particular,

$$\mathbb{P}\{|g(X_1,\ldots,X_n) - \mathbb{E}g(X_1,\ldots,X_n)| \ge t\} \le 2e^{-2t^2/\sum_{i=1}^n c_i^2}.$$

We end this section with the Efron-Stein inequality (Efron and Stein, 1981; Steele, 1986), a powerful tool for deriving a bound for the variance of a general square-integrable function of independent random variables.

Theorem 20.10 (Efron and Stein, 1981; Steele, 1986). *Let $X_1,\ldots,X_n$ be independent random variables, and let $g(X_1,\ldots,X_n)$ be a square-integrable function of $X_1,\ldots,X_n$. Then, if $X_1',\ldots,X_n'$ are independent copies of $X_1,\ldots,X_n$,*

$$\mathbb{V}g(X_1,\ldots,X_n) \le \frac{1}{2}\sum_{i=1}^n \mathbb{E}\left|g(X_1,\ldots,X_i,\ldots,X_n) - g(X_1,\ldots,X_i',\ldots,X_n)\right|^2.$$

Also,

$$\mathbb{V}g(X_1,\ldots,X_n) \le \inf_{Z_i}\sum_{i=1}^n \mathbb{E}\left|g(X_1,\ldots,X_n) - Z_i\right|^2,$$

where the infimum is taken over the class of all $(X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n)$-measurable and square-integrable random variables Z_i, $1 \le i \le n$.

20.5 Some useful inequalities

Proposition 20.1 (c_r-inequality). *Let $r > 0$ and let $a_1,\ldots,a_p$ be real numbers. Then*

$$\left|\sum_{i=1}^p a_i\right|^r \le c_r \sum_{i=1}^p |a_i|^r,$$

where $c_r = p^{r-1}$ for $r \ge 1$ and $c_r = 1$ for $0 < r < 1$.

Proof. For $r \ge 1$, the inequality is implied by the convexity of the function $x \mapsto |x|^r$ on $\mathbb{R}$. For $0 < r < 1$, note that for $x, y \ge 0$,

$$|x+y|^r - |x|^r = \int_x^{x+y} rt^{r-1}\mathrm{d}t = \int_0^y r(x+s)^{r-1}\mathrm{d}s$$
$$\le \int_0^y rs^{r-1}\mathrm{d}s = |y|^r.$$

Since $|x+y|^r \le \big||x|+|y|\big|^r$, the inequality still holds for x and y of arbitrary sign. The conclusion follows by induction on p. □

Chebyshev's inequality states that if X is a real-valued random variable with $\mathbb{E}X^2 < \infty$, then, for all $t > 0$,

$$\mathbb{P}\{|X - \mathbb{E}X| \geq t\} \leq \frac{\mathbb{V}X}{t^2}.$$

The following theorem is a one-sided improved version of this inequality:

Theorem 20.11 (Chebyshev-Cantelli inequality). *Let X be a real-valued random variable such that $\mathbb{E}X^2 < \infty$. Then, for all $t > 0$,*

$$\mathbb{P}\{X - \mathbb{E}X \geq t\} \leq \frac{\mathbb{V}X}{\mathbb{V}X + t^2}.$$

Proof. Assume, without loss of generality, that $\mathbb{E}X = 0$. Write $t = \mathbb{E}[t - X] \leq \mathbb{E}[(t - X)\mathbb{1}_{[X<t]}]$, and apply the Cauchy-Schwarz inequality. □

Theorem 20.12 (Jensen's inequality). *Let X be a real-valued random variable such that $\mathbb{E}|X| < \infty$, and let $g : \mathbb{R} \to \mathbb{R}$ be a convex function such that $\mathbb{E}|g(X)| < \infty$. Then*

$$g(\mathbb{E}X) \leq \mathbb{E}g(X).$$

Proof. By convexity, there exists $a \in \mathbb{R}$ such that, for all x,

$$g(x) \geq a(x - \mathbb{E}X) + g(\mathbb{E}X).$$

Thus,

$$\mathbb{E}g(X) \geq \mathbb{E}\left[a(X - \mathbb{E}X) + g(\mathbb{E}X)\right] = g(\mathbb{E}X). \quad \square$$

Theorem 20.13 (Marcinkiewicz and Zygmund, 1937). *Let $X_1, \ldots, X_n$ be independent zero-mean real-valued random variables, and let $p \geq 1$. Assume that $\mathbb{E}|X_i|^p < \infty$, $1 \leq i \leq n$. Then*

$$\mathbb{E}\left|\sum_{i=1}^{n} X_i\right|^p \leq C_p\, \mathbb{E}\left|\sum_{i=1}^{n} X_i^2\right|^{p/2},$$

where C_p is a positive constant depending only upon p.

Proof. We recall that $\cosh(x) \leq \exp(x^2/2)$, which follows easily by a comparison of Taylor series. Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent Rademacher random variables, each taking the values $+1$ and -1 with equal probability. Then, for $x \in \mathbb{R}$,

$$\mathbb{E}e^{x\varepsilon_1} = \frac{1}{2}\left(e^x + e^{-x}\right) = \cosh(x) \leq e^{x^2/2}.$$

Thus, for $y_1, \ldots, y_n \geq 0$ fixed and $\lambda > 0$,

$$\mathbb{E}\left[e^{\lambda \sum_{i=1}^n \varepsilon_i y_i}\right] = \mathbb{E}\left[\prod_{i=1}^n e^{\lambda \varepsilon_i y_i}\right] = \prod_{i=1}^n \mathbb{E} e^{\lambda \varepsilon_i y_i}$$
$$\leq \prod_{i=1}^n e^{\lambda^2 y_i^2/2} = e^{\lambda^2 \sum_{i=1}^n y_i^2/2}.$$

For $x \geq 0$, we have by an easy maximization

$$x^p e^{-\lambda x} \leq \left(\frac{p}{\lambda e}\right)^p.$$

Therefore, if one of the y_i's is not 0,

$$\begin{aligned}
\mathbb{E}\left|\sum_{i=1}^n \varepsilon_i y_i\right|^p &\leq \left(\frac{p}{\lambda e}\right)^p \mathbb{E}\left[e^{\lambda |\sum_{i=1}^n \varepsilon_i y_i|}\right] \\
&\leq \left(\frac{p}{\lambda e}\right)^p \left(\mathbb{E}\left[e^{\lambda \sum_{i=1}^n \varepsilon_i y_i}\right] + \mathbb{E}\left[e^{-\lambda \sum_{i=1}^n \varepsilon_i y_i}\right]\right) \\
&\leq 2\left(\frac{p}{\lambda e}\right)^p e^{\lambda^2 \sum_{i=1}^n y_i^2/2} \\
&= 2\left(\frac{p}{e} \sum_{i=1}^n y_i^2\right)^{p/2}
\end{aligned}$$

by choosing $\lambda = \sqrt{p/\sum_{i=1}^n y_i^2}$. Clearly, the inequality remains true if all y_i's are equal to 0.

Now, let $X_1', \ldots, X_n'$ be independent of, and distributed as $X_1, \ldots, X_n$. Assume that $\varepsilon_1, \ldots, \varepsilon_n$ are independent of $X_1, X_1', \ldots, X_n, X_n'$. Then,

$$\begin{aligned}
\mathbb{E}\left|\sum_{i=1}^n X_i\right|^p &= \mathbb{E}\left|\mathbb{E}\left[\sum_{i=1}^n (X_i - X_i') \,\middle|\, X_1, \ldots, X_n\right]\right|^p \\
&\leq \mathbb{E}\left|\sum_{i=1}^n (X_i - X_i')\right|^p \\
&\quad \text{(by Jensen's inequality, valid for } p \geq 1) \\
&= \mathbb{E}\left|\sum_{i=1}^n \varepsilon_i Y_i\right|^p \\
&\quad \text{(where } Y_i = |X_i - X_i'|) \\
&\leq 2\left(\frac{p}{e}\right)^{p/2} \mathbb{E}\left|\sum_{i=1}^n Y_i^2\right|^{p/2}.
\end{aligned}$$

However,

$$\begin{aligned}\mathbb{E}\left|\sum_{i=1}^n Y_i^2\right|^{p/2} &\le \mathbb{E}\left|2\sum_{i=1}^n X_i^2 + 2\sum_{i=1}^n X_i'^2\right|^{p/2} \\ &\le c_p\left(\mathbb{E}\left|2\sum_{i=1}^n X_i^2\right|^{p/2} + \mathbb{E}\left|2\sum_{i=1}^n X_i'^2\right|^{p/2}\right) \\ &= 2^{1+\frac{p}{2}} c_p\, \mathbb{E}\left|\sum_{i=1}^n X_i^2\right|^{p/2},\end{aligned}$$

where we used the c_r-inequality (Proposition 20.1), with $c_p = 2^{\frac{p}{2}-1}$ if $p \ge 2$ and $c_p = 1$ if $1 \le p < 2$. Therefore, for $p \ge 1$,

$$\mathbb{E}\left|\sum_{i=1}^n X_i\right|^p \le 2^{2+\frac{p}{2}} c_p \left(\frac{p}{e}\right)^{p/2} \mathbb{E}\left|\sum_{i=1}^n X_i^2\right|^{p/2}.$$

The desired result follows with

$$C_p = 2^{2+\frac{p}{2}} c_p \left(\frac{p}{e}\right)^{p/2}. \quad \square$$

Observing that, by Jensen's inequality, for $p \ge 2$,

$$\left|\sum_{i=1}^n X_i^2\right|^{p/2} \le n^{\frac{p}{2}-1} \sum_{i=1}^n |X_i|^p,$$

we deduce the following corollary:

Corollary 20.2. *Let $X_1, \ldots, X_n$ be i.i.d. zero-mean real-valued random variables, and let $p \ge 2$. Assume that $\mathbb{E}|X_i|^p < \infty$, $1 \le i \le n$. Then*

$$\mathbb{E}\left|\sum_{i=1}^n X_i\right|^p \le C_p\, n^{p/2} \mathbb{E}|X_1|^p.$$

Theorem 20.14. *The gamma function*

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} \mathrm{d}t$$

satisfies

$$1 < \frac{\Gamma(x)}{\left(\frac{x}{e}\right)^x \sqrt{\frac{2\pi}{x}}} \le e, \quad x \ge 1.$$

Furthermore, for $x \geq y \geq 1$,

$$\frac{e^{y-x}x^x}{y^y}\left(\frac{y}{x}\right) \leq \frac{\Gamma(x)}{\Gamma(y)} \leq \frac{e^{y-x}x^x}{y^y}\sqrt{\frac{y}{x}}.$$

Proof. The first inequality is due to Mortici and Chen (2011), while the second one is due to Kečkić and Vasić (1971). □

20.6 Equivalence inequalities for weights

Lemma 20.3. *Let* $(v_{n1}, \ldots, v_{nn})$ *be a probability weight vector. The following conditions are equivalent:*

(*I*) *There exists a sequence of integers* $\{k\} = \{k_n\}$ *such that*

$$(i)\ \ k \to \infty \quad \text{and} \quad k/n \to 0;$$
$$(ii)\ \ \sum_{i>k} v_{ni} \to 0.$$

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

$$\sum_{i>\varepsilon n} v_{ni} \to 0.$$

Proof. (*I*) implies (*II*) since for each $\varepsilon > 0$, and all n large enough, $k \leq \varepsilon n$. (*II*) implies (*I*) by construction. Let $\{n_j\}_{j\geq 1}$ be a strictly increasing sequence of integers such that $n_1 = 1, j/n_j \to 0$ as $j \to \infty$, and

$$\sum_{i>n/j} v_{ni} < \frac{1}{j} \quad \text{for all } n \geq n_j.$$

Let $k = n/j$ on $[n_j, n_{j+1})$. Clearly, $k \to \infty$ and $k/n \to 0$. Also, $\sum_{i>k} v_{ni} \to 0$ as $n \to \infty$. □

Lemma 20.4. *Let* $(v_{n1}, \ldots, v_{nn})$ *be a probability weight vector. The following conditions are equivalent:*

(*I*) *There exists a sequence of integers* $\{k\} = \{k_n\}$ *such that*

$$(i)\ \ k \to \infty \quad \text{and} \quad k/n \to 0;$$
$$(ii)\ \ \sum_{i>k} v_{ni} \to 0;$$
$$(iii)\ \ \sup_n (k \max_i v_{ni}) < \infty.$$

(*II*) *There exists a positive constant* α *such that*

$$\begin{aligned}&(iv)\ \textstyle\sum_{i>\alpha/\max_i v_{ni}} \to 0;\\ &(v)\ \ \textstyle\sum_{i>\varepsilon n} v_{ni} \to 0, \quad \textit{all } \varepsilon > 0;\\ &(vi)\ \max_i v_{ni} \to 0.\end{aligned}$$

Moreover, the same result is true when condition (i) *is replaced by*

$$(i)\ k/\log n \to \infty \quad \textit{and} \quad k/n \to 0$$

and condition (vi) *is replaced by*

$$(vi)\ \sup_n ((\log n)\max_i v_{ni}) \to 0.$$

Proof. Let us first assume that (I) is true. Then, clearly, (i) and (iii) imply (vi). Besides, (i) and (ii) imply (v) since for each $\varepsilon > 0$, and all n large enough, $k \le \varepsilon n$. Finally, denoting by α a positive constant in (iii) such that $k \le \alpha/\max_i v_{ni}$, we see that (ii) implies (iv).

Conversely, assume that (II) holds, and set $k = \lfloor \alpha/\max_i v_{ni} \rfloor$. Then (iii) is true and, clearly, (vi) implies $k \to \infty$ and (iv) implies (ii). The second statement of (I) is valid because for all $\varepsilon > 0$, $\varepsilon n \max_i v_{ni} \ge \sum_{i \le \varepsilon n} v_{ni} \to 1$ as $n \to \infty$.

The last assertion is proved in a similar fashion. □

20.7 Covering $\mathbb{R}^d$ with cones

A cone $\mathscr{C}(\mathbf{z}, \theta)$ for $\mathbf{z} \in \mathbb{R}^d - \{\mathbf{0}\}$, $0 < \theta \le \pi/2$, is defined by

$$\mathscr{C}(\mathbf{z}, \theta) = \{\mathbf{y} \in \mathbb{R}^d : \mathbf{y} = \mathbf{0} \text{ or } \operatorname{angle}(\mathbf{z}, \mathbf{y}) \le \theta\},$$

where

$$\operatorname{angle}(\mathbf{z}, \mathbf{y}) = \arccos\left(\frac{\sum_{j=1}^d z_j y_j}{\|\mathbf{z}\|\,\|\mathbf{y}\|}\right),$$

$\|\mathbf{z}\|^2 = \sum_{j=1}^d z_j^2$, $\|\mathbf{y}\|^2 = \sum_{j=1}^d y_j^2$. Equivalently, in vector notation,

$$\frac{\mathbf{z}^\top \mathbf{y}}{\|\mathbf{z}\|\,\|\mathbf{y}\|} \ge \cos\theta,$$

where $\top$ denotes transposition and vectors are in column format. The set $\mathbf{x} + \mathscr{C}(\mathbf{z}, \theta)$ is the translation of $\mathscr{C}(\mathbf{z}, \theta)$ by $\mathbf{x}$ (change of origin).

Let

$$C_\theta = \min\left\{n \ge 1 : \exists\, \mathbf{z}_1, \ldots, \mathbf{z}_n \in \mathbb{R}^d - \{\mathbf{0}\} \text{ such that } \bigcup_{i=1}^n \mathscr{C}(\mathbf{z}_i, \theta) = \mathbb{R}^d\right\}$$

be the minimal number of cones needed to cover $\mathbb{R}^d$. For $d = 1$, we can take $z_1 = 1$, $z_2 = -1$ for any $0 < \theta \le \pi/2$, and cover $\mathbb{R}$, so $C_\theta = 2$. In $\mathbb{R}^2$, we can take

$$\mathbf{z}_i = \big(\cos(2\theta i), \sin(2\theta i)\big), \quad 1 \le i \le \lceil \pi/\theta \rceil,$$

and verify that

$$\bigcup_{i=1}^{\lceil \pi/\theta \rceil} \mathscr{C}(\mathbf{z}_i, \theta) = \mathbb{R}^2.$$

In particular, $C_\theta \le \lceil \pi/\theta \rceil$, and $C_{\pi/6} \le 6$. In fact, one can easily see that $C_\theta \ge \pi/\theta$, and thus, we have $C_\theta = \lceil \pi/\theta \rceil$.

For general d, a simple covering argument of a compact set permits one to show that $C_\theta < \infty$ for all $0 < \theta \le \pi/2$. We first give a simple but suboptimal bound.

Theorem 20.15. *For all dimensions $d \ge 1$ and all $0 < \theta \le \pi/2$,*

$$C_\theta \le \left(2 \left\lceil \frac{\sqrt{d}}{1 - \cos\theta} \right\rceil + 1 \right)^d.$$

Proof. Consider the grid $\mathscr{G} = \{j/N : -N \le j \le N\}^d$, where $N \ge \sqrt{d}$ will be chosen later. We claim that $\bigcup_{\mathbf{z} \in \mathscr{G} \setminus \{\mathbf{0}\}} \mathscr{C}(\mathbf{z}, \theta) = \mathbb{R}^d$. To see this, take $\mathbf{y} = (y_1, \ldots, y_d) \in \mathbb{R}^d - \{\mathbf{0}\}$, and let $\|\mathbf{y}\|$ denote its Euclidean norm. Similarly, we write $\mathbf{z} = (z_1, \ldots, z_d)$ and $\|\mathbf{z}\|$. We associate with $y_j \ge 0$ the value $\frac{1}{N} \lfloor \frac{y_j}{\|\mathbf{y}\|} N \rfloor$, and with $y_j < 0$ the value $-\frac{1}{N} \lfloor \frac{y_j}{\|\mathbf{y}\|} N \rfloor$, where the values are clearly in $\mathscr{G}$. Call the vector of these values $\mathbf{z}$, and observe that $0 < \|\mathbf{z}\| \le 1$ by the truncations towards the origin. Next,

$$\left| \sum_{j=1}^d z_j y_j - \sum_{j=1}^d \frac{y_j^2}{\|\mathbf{y}\|} \right| \le \frac{1}{N} \sum_{j=1}^d |y_j| \le \frac{\sqrt{d}}{N} \|\mathbf{y}\|$$

by the Cauchy-Schwarz inequality. Therefore, since $z_j y_j \ge 0$, for $\mathbf{z} \ne \mathbf{0}$,

$$\begin{aligned} \cos\left(\text{angle}(\mathbf{z}, \mathbf{y})\right) = \frac{\sum_{j=1}^d z_j y_j}{\|\mathbf{z}\| \, \|\mathbf{y}\|} &\ge \frac{1}{\|\mathbf{y}\|} \sum_{j=1}^d z_j y_j \\ &\ge \sum_{j=1}^d \frac{y_j^2}{\|\mathbf{y}\|^2} - \frac{\sqrt{d}}{N} \\ &= 1 - \frac{\sqrt{d}}{N}. \end{aligned}$$

We have $\cos(\text{angle}(\mathbf{z}, \mathbf{y})) \geq \cos\theta$, and thus $\text{angle}(\mathbf{z}, \mathbf{y}) \leq \theta$, if $1 - \frac{\sqrt{d}}{N} \geq \cos\theta$. Therefore, it suffices to take

$$N \geq \frac{\sqrt{d}}{1 - \cos\theta}.$$

The proof follows since $|\mathscr{G}| = (2N + 1)^d$. □

Böröczky, Jr. and Wintsche (2003) (see also Böröczky, Jr., 2004) showed the following:

Theorem 20.16. *For all dimensions $d \geq 2$ and all $0 < \theta \leq \arccos(1/\sqrt{3})$,*

$$C_\theta \leq \alpha \times \frac{\cos\theta}{\sin^d\theta} \times d^{3/2} \log(1 + d\cos^2\theta),$$

where α is a positive universal constant independent of d.

This theorem implies immediately that for every $\varepsilon > 0$, there exists a universal constant $\alpha(\varepsilon)$ such that, for all $d \geq 2$,

$$C_{\pi/6} \leq \alpha(\varepsilon)(2 + \varepsilon)^d.$$

Our next lemma states an interesting geometrical property of cones of angle less than or equal to $\pi/6$ (see Figure 20.1 for an illustration in dimension 2).

Lemma 20.5. *Let $\mathscr{C}(\mathbf{z}, \theta)$ be a cone of angle $0 < \theta \leq \pi/6$. If $\theta < \pi/6$, then for $\mathbf{y}_1, \mathbf{y}_2 \in \mathscr{C}(\mathbf{z}, \pi/6)$, $\|\mathbf{y}_1\| > 0$ and $\|\mathbf{y}_1\| \leq \|\mathbf{y}_2\|$, one has $\|\mathbf{y}_1 - \mathbf{y}_2\| < \|\mathbf{y}_2\|$. If $\theta = \pi/6$, then for $\mathbf{y}_1, \mathbf{y}_2 \in \mathscr{C}(\mathbf{z}, \pi/6)$, $\|\mathbf{y}_1\| > 0$ and $\|\mathbf{y}_1\| < \|\mathbf{y}_2\|$, one has $\|\mathbf{y}_1 - \mathbf{y}_2\| < \|\mathbf{y}_2\|$.*

Proof. Take $\theta = \pi/6$ and note that

$$\begin{aligned}
\|\mathbf{y}_1 - \mathbf{y}_2\|^2 &= \|\mathbf{y}_1\|^2 + \|\mathbf{y}_2\|^2 - 2\|\mathbf{y}_1\|\,\|\mathbf{y}_2\| \frac{\mathbf{y}_1{}^\top \mathbf{y}_2}{\|\mathbf{y}_1\|\,\|\mathbf{y}_2\|} \\
&\leq \|\mathbf{y}_1\|^2 + \|\mathbf{y}_2\|^2 - 2\|\mathbf{y}_1\|\,\|\mathbf{y}_2\| \cos(2\theta) \\
&= \|\mathbf{y}_2\|^2 \left(1 + \frac{\|\mathbf{y}_1\|^2}{\|\mathbf{y}_2\|^2} - \frac{\|\mathbf{y}_1\|}{\|\mathbf{y}_2\|}\right) \\
&< \|\mathbf{y}_2\|^2.
\end{aligned}$$

The proof is similar for $0 < \theta < \pi/6$. □

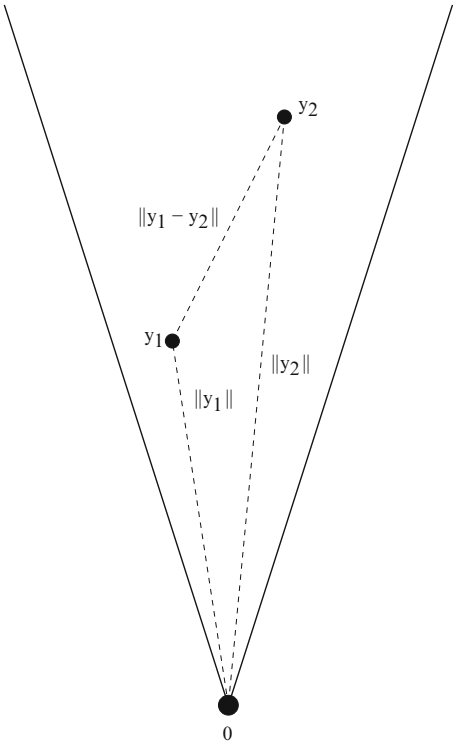

Fig. 20.1 The geometrical property of a cone of angle $0 < \theta \leq \pi/6$ (in dimension 2).

An elegant combinatorial implication of Lemma 20.5 is the following one:

Lemma 20.6. *Let* $\mathbf{x}, \mathbf{x}_1, \ldots, \mathbf{x}_n$ *be distinct vectors of* $\mathbb{R}^d$. *Then, for all* $0 < \theta < \pi/6$,

$$\sum_{i=1}^{n} \mathbb{1}_{[\mathbf{x} \text{ is the nearest neighbor of } \mathbf{x}_i \text{ in } \{\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_n\}]} \leq C_\theta$$

(distance ties are broken by comparing indices). In addition, if all distances $\|\mathbf{x}_i - \mathbf{x}\|$, $1 \leq i \leq n$, *are different, then*

$$\sum_{i=1}^{n} \mathbb{1}_{[\mathbf{x} \text{ is the nearest neighbor of } \mathbf{x}_i \text{ in } \{\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_n\}]} \leq C_{\pi/6}.$$

Proof. Fix $0 < \theta < \pi/6$ and cover $\mathbb{R}^d$ by C_θ cones $\mathbf{x} + \mathscr{C}(\mathbf{z}_\ell, \theta)$, $1 \leq \ell \leq C_\theta$. In each cone, mark the $\mathbf{x}_i$ nearest to $\mathbf{x}$, if such an $\mathbf{x}_i$ exists. If $\mathbf{x}_i$ belongs to $\mathbf{x} + \mathscr{C}(\mathbf{z}_\ell, \theta)$ and is not marked, then, by the first statement of Lemma 20.5, $\mathbf{x}$ cannot be the nearest neighbor of $\mathbf{x}_i$ in $\{\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_n\}$. This shows the first inequality. The case $\theta = \pi/6$ is proved by a similar argument, via the second statement of Lemma 20.5. □

An interesting geometrical consequence of this lemma is that if $\mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n$ are random variables drawn according to a common absolutely continuous distribution, then, with probability one, $\mathbf{X}$ can be the nearest neighbor of at most $C_{\pi/6}$ points.

20.8 Some results from real analysis

Theorem 20.17 (A denseness result). *Let $p \geq 1$. For any probability measure μ on $\mathbb{R}^d$, the set of continuous functions of bounded support is dense in $L^p(\mu)$. In other words, for any $g \in L^p(\mu)$ and $\varepsilon > 0$, there is a continuous function g_ε with compact support such that*

$$\int_{\mathbb{R}^d} |g(\mathbf{x}) - g_\varepsilon(\mathbf{x})|^p \, \mu(\mathrm{d}\mathbf{x}) \leq \varepsilon.$$

Proof. See, e.g., Györfi et al. (2002, Theorem A.1). □

We provide in the remainder of this section some results concerning differentiation of integrals. Good general references are Stein (1970), de Guzmán (1975), Wheeden and Zygmund (1977), and Györfi et al. (2002, Chapter 24). In the sequel, notation $B_\rho(\mathbf{x})$, $\rho \geq 0$, means indifferently the family of cubes with center $\mathbf{x}$ and edge length ρ, the family of closed balls centered at $\mathbf{x}$ of radius ρ, or the family of open balls centered at $\mathbf{x}$ of radius ρ. As always, λ denotes the Lebesgue measure on $\mathbb{R}^d$.

Theorem 20.18 (The Lebesgue differentiation theorem). *Let $g : \mathbb{R}^d \to \mathbb{R}$ be a locally integrable function. Then, at λ-almost all $\mathbf{x} \in \mathbb{R}^d$,*

$$\frac{1}{\lambda\left(B_\rho(\mathbf{x})\right)} \int_{B_\rho(\mathbf{x})} |g(\mathbf{y}) - g(\mathbf{x})| \, \mathrm{d}\mathbf{y} \to 0 \quad \text{as } \rho \downarrow 0.$$

A point $\mathbf{x}$ at which this statement is valid is called a Lebesgue point of g. In particular,

$$\frac{1}{\lambda\left(B_\rho(\mathbf{x})\right)} \int_{B_\rho(\mathbf{x})} g(\mathbf{y}) \mathrm{d}\mathbf{y} \to g(\mathbf{x}) \quad \text{as } \rho \downarrow 0.$$

Remark 20.1. Theorem 20.18 does not hold when the family $\{B_\rho(\mathbf{x})\}$ shrinks to $\mathbf{x}$ without any restriction on the behavior of the sets. In this more general framework, the slight additional condition $\int_{\mathbb{R}^d} |g(\mathbf{y})| \log^+ |g(\mathbf{y})| \mathrm{d}\mathbf{y} < \infty$ is required (it is true, in particular, if $\int_{\mathbb{R}^d} |g(\mathbf{y})|^p \mathrm{d}\mathbf{y} < \infty$ for some $p > 1$). Valuable ideas and statistical comments are presented in Devroye and Krzyżak (2002) and Biau et al. (2015). □

The Lebesgue differentiation theorem generalizes to measures other than the Lebesgue measure on $\mathbb{R}^d$. Throughout, we let μ_1 and μ_2 be two σ-finite Borel measures that are finite on the bounded Borel sets of $\mathbb{R}^d$.

Theorem 20.19 (The generalized Lebesgue differentiation theorem). *Let $p \geq 1$, and let $g : \mathbb{R}^d \to \mathbb{R}$ be such that $|g|^p$ is locally integrable with respect to μ_1. Then, at μ_1-almost all $\mathbf{x} \in \mathbb{R}^d$,*

$$\frac{1}{\mu_1\left(B_\rho(\mathbf{x})\right)} \int_{B_\rho(\mathbf{x})} |g(\mathbf{y}) - g(\mathbf{x})|^p \, \mu_1(\mathrm{d}\mathbf{y}) \to 0 \quad \text{as } \rho \downarrow 0.$$

In particular,

$$\frac{1}{\mu_1\left(B_\rho(\mathbf{x})\right)} \int_{B_\rho(\mathbf{x})} g(\mathbf{y}) \mu_1(\mathrm{d}\mathbf{y}) \to g(\mathbf{x}) \quad \text{as } \rho \downarrow 0.$$

Remark 20.2. Theorem 20.19 is usually proved for $p = 1$. Since for $a, b \geq 0$ and $p \geq 1$, $|a - b|^p \leq |a^p - b^p|$, the result is true if $g \geq 0$. For general g, split g into its positive and negative parts, and note that $|g(\mathbf{y}) - g(\mathbf{x})|^p \leq 2^{p-1}(|g^+(\mathbf{y}) - g^+(\mathbf{x})|^p + |g^-(\mathbf{y}) - g^-(\mathbf{x})|^p)$. □

Whenever the measure μ_2 is absolutely continuous with respect to μ_1, with a density f, then Theorem 20.19 states that the ratio $\frac{\mu_2(B_\rho(\mathbf{x}))}{\mu_1(B_\rho(\mathbf{x}))}$ shrinks μ_1-almost everywhere towards the value of f at the point $\mathbf{x}$. The most general differentiation theorem clarifies the local behavior of $\frac{\mu_2(B_\rho(\mathbf{x}))}{\mu_1(B_\rho(\mathbf{x}))}$ without assuming that μ_2 is absolutely continuous with respect to μ_1. Before continuing, recall that the Lebesgue decomposition theorem states that there exists a unique decomposition of μ_2 into an absolutely continuous part and a singular part with respect to μ_1. For any Borel set $A \subseteq \mathbb{R}^d$,

$$\mu_2(A) = \int_A f(\mathbf{y}) \mu_1(\mathrm{d}\mathbf{y}) + \sigma(A),$$

where f is a nonnegative function, integrable with respect to μ_1, and the measure σ is supported on a set of μ_1-measure zero.

Theorem 20.20. *If $\mu_2(A) = \int_A f \mathrm{d}\mu_1 + \sigma(A)$ is the decomposition of μ_2 into parts that are absolutely continuous and singular with respect to μ_1, then, at μ_1-almost all $\mathbf{x} \in \mathbb{R}^d$,*

$$\frac{\mu_2\left(B_\rho(\mathbf{x})\right)}{\mu_1\left(B_\rho(\mathbf{x})\right)} \to f(\mathbf{x}) \quad \text{as } \rho \downarrow 0.$$

In particular, if μ_1 and μ_2 are mutually singular, then, at μ_1-almost all $\mathbf{x} \in \mathbb{R}^d$,

$$\frac{\mu_2\left(B_\rho(\mathbf{x})\right)}{\mu_1\left(B_\rho(\mathbf{x})\right)} \to 0 \quad \text{as } \rho \downarrow 0.$$

The useful lemma below estimates the size of the maximal ratio $\frac{\mu_2(B_\rho(\mathbf{x}))}{\mu_1(B_\rho(\mathbf{x}))}$.

Lemma 20.7 (Fefferman and Stein, 1971). *There exists a constant c, depending only upon d, such that, for all $t > 0$,*

$$\mu_1\left(\left\{\mathbf{x}\in\mathbb{R}^d : \sup_{\rho>0}\left(\frac{\mu_2\left(B_\rho(\mathbf{x})\right)}{\mu_1\left(B_\rho(\mathbf{x})\right)}\right) > t\right\}\right) \leq \frac{c}{t}\,\mu_2(\mathbb{R}^d).$$

Moreover, for any Borel set $A \subseteq \mathbb{R}^d$,

$$\mu_1\left(\left\{\mathbf{x}\in A : \sup_{\rho>0}\left(\frac{\mu_2\left(B_\rho(\mathbf{x})\right)}{\mu_1\left(B_\rho(\mathbf{x})\right)}\right) > t\right\}\right) \leq \frac{c}{t}\,\mu_2(A).$$

When μ_1 is the Lebesgue measure λ and μ_2 is a probability measure that is absolutely continuous with respect to λ, with a density f, then Lemma 20.7 states that

$$\lambda\left(\left\{\mathbf{x}\in\mathbb{R}^d : f^\star(\mathbf{x}) > t\right\}\right) \leq \frac{c}{t},$$

where

$$f^\star(\mathbf{x}) = \sup_{\rho>0}\left(\frac{1}{\lambda\left(B_\rho(\mathbf{x})\right)}\int_{B_\rho(\mathbf{x})} f(\mathbf{y})\mathrm{d}\mathbf{y}\right).$$

The function $f^\star$ is called the Hardy-Littlewood maximal function of f. It should be understood as a gauge of the size of the averages of f around $\mathbf{x}$ (note that $f(\mathbf{x}) \leq f^\star(\mathbf{x})$ at λ-almost all $\mathbf{x}$).

If $\int_{\mathbb{R}^d} f^p(\mathbf{x})\mathrm{d}\mathbf{x} < \infty$ for some $p > 1$, then $\int_{\mathbb{R}^d} f^{\star p}(\mathbf{x})\mathrm{d}\mathbf{x} < \infty$ as well. On the other hand, it is not true that $\int_{\mathbb{R}^d} f(\mathbf{x})\mathrm{d}\mathbf{x} < \infty$ implies $\int_{\mathbb{R}^d} f^\star(\mathbf{x})\mathrm{d}\mathbf{x} < \infty$ (just take for f the uniform density on $[0, 1]$ and note that $\int_{\mathbb{R}} f^\star(x)\mathrm{d}x = \infty$). In fact, $f^\star$ is never integrable on all $\mathbb{R}^d$. This can be seen by making the observation that $f^\star(\mathbf{x}) \geq c\|\mathbf{x}\|^{-d}$, for $\|\mathbf{x}\| \geq 1$. Moreover, even if we limit our considerations to any bounded subset of $\mathbb{R}^d$, then the integrability of $f^\star$ requires stronger conditions than the integrability of f. For example, $\int_{\mathbb{R}^d} f^\star(\mathbf{x})\mathrm{d}\mathbf{x} < \infty$ as soon as $\int_{\mathbb{R}^d} f(\mathbf{x})\log(f(\mathbf{x}) + 1)\mathrm{d}\mathbf{x} < \infty$ and f is supported on a compact set (Stein, 1970, Chapter 1). The following lemma states this more formally.

Lemma 20.8. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a density. Assume that f has compact support and that $\int_{\mathbb{R}^d} f(\mathbf{x})\log(f(\mathbf{x}) + 1)\mathrm{d}\mathbf{x} < \infty$. Then $\int_{\mathbb{R}^d} f(\mathbf{x})\log(f^\star(\mathbf{x}) + 1)\mathrm{d}\mathbf{x} < \infty$. Similarly, $\int_{\mathbb{R}^d} f(\mathbf{x})\log^2(f(\mathbf{x}) + 1)\mathrm{d}\mathbf{x} < \infty$ implies $\int_{\mathbb{R}^d} f(\mathbf{x})\log^2(f^\star(\mathbf{x}) + 1)\mathrm{d}\mathbf{x} < \infty$.*

Proof. We prove the first assertion only and leave the second one as a small exercise. Observe that

$$\begin{aligned}
&\int_{\mathbb{R}^d} f(\mathbf{x}) \log(f^\star(\mathbf{x}) + 1) \mathrm{d}\mathbf{x} \\
&\quad = \int_0^\infty \lambda\Big(\big\{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) \log(f^\star(\mathbf{x}) + 1) > t\big\}\Big) \mathrm{d}t \\
&\quad \le \int_0^\infty \left[\lambda\left(\left\{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) > \frac{t}{\log(t+1)}\right\}\right) + \lambda\Big(\big\{\mathbf{x} \in \mathbb{R}^d : f^\star(\mathbf{x}) > t\big\}\Big)\right] \mathrm{d}t.
\end{aligned}$$

Now,

$$\int_0^\infty \lambda\Big(\big\{\mathbf{x} \in \mathbb{R}^d : f^\star(\mathbf{x}) > t\big\}\Big) \mathrm{d}t = \int_0^\infty f^\star(\mathbf{x}) \mathrm{d}\mathbf{x},$$

and this integral is finite by Stein's result (Stein, 1970) mentioned above. Finally, one easily verifies that for t larger than some positive $t^\star$, the inequality $f(\mathbf{x}) \log(f(\mathbf{x}) + 1) \le t/2$ implies $f(\mathbf{x}) \le \frac{t}{\log(t+1)}$. So,

$$\begin{aligned}
&\int_0^\infty \lambda\left(\left\{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) > \frac{t}{\log(t+1)}\right\}\right) \mathrm{d}t \\
&\le \int_0^{t^\star} \lambda\left(\left\{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) > \frac{t}{\log(t+1)}\right\}\right) \mathrm{d}t \\
&\quad + \int_{t^\star}^\infty \lambda\Big(\big\{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) \log(f(\mathbf{x}) + 1) > t/2\big\}\Big) \mathrm{d}t \\
&\le \int_0^\infty \lambda\Big(\big\{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) \log(t^\star + 1) > t\big\}\Big) \mathrm{d}t \\
&\quad + 2 \int_0^\infty \lambda\Big(\big\{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) \log(f(\mathbf{x}) + 1) > t\big\}\Big) \mathrm{d}t \\
&= \log(t^\star + 1) \int_{\mathbb{R}^d} f(\mathbf{x}) \mathrm{d}\mathbf{x} + 2 \int_{\mathbb{R}^d} f(\mathbf{x}) \log(f(\mathbf{x}) + 1) \mathrm{d}\mathbf{x} < \infty. \quad \square
\end{aligned}$$

20.9 Some useful probability distributions

Bernoulli distribution

Definition. A random variable X has the Bernoulli distribution with parameter $p \in [0, 1]$ if $\mathbb{P}\{X = 1\} = p$ and $\mathbb{P}\{X = 0\} = 1 - p$.

Notation. Ber(p).

Moments. $\mathbb{E}X = p$ and $\mathbb{V}X = p(1 - p)$.

Binomial distribution

Definition. A random variable X has the binomial distribution with parameters $n \in \mathbb{N} - \{0\}$ and $p \in [0, 1]$ if

$$\mathbb{P}\{X = j\} = \binom{n}{j} p^j (1-p)^{n-j}, \quad 0 \le j \le n.$$

In particular, $X \stackrel{\mathscr{D}}{=} \sum_{i=1}^{n} X_i$, where $X_1, \ldots, X_n$ are independent $\mathrm{Ber}(p)$ random variables.

Notation. $\mathrm{Bin}(n, p)$.

Moments. $\mathbb{E}X = np$ and $\mathbb{V}X = np(1-p)$.

Poisson distribution

Definition. A random variable X has the Poisson distribution with parameter $\lambda > 0$ if

$$\mathbb{P}\{X = j\} = \frac{e^{-\lambda}\lambda^j}{j!}, \quad j \in \mathbb{N}.$$

Notation. $\mathrm{Poisson}(\lambda)$.

Moments. $\mathbb{E}X = \lambda$ and $\mathbb{V}X = \lambda$.

If $np_n \to \lambda > 0$ as $n \to \infty$, then $\mathrm{Bin}(n, p_n) \stackrel{\mathscr{D}}{\to} \mathrm{Poisson}(\lambda)$.

Rademacher distribution

Definition. A random variable X has the Rademacher distribution if $\mathbb{P}\{X = -1\} = \mathbb{P}\{X = +1\} = 1/2$.

Moments. $\mathbb{E}X = 0$ and $\mathbb{V}X = 1$.

Uniform distribution

Definition. A random variable X has the uniform distribution on $[0, 1]$ if it has density $f(x) = 1, 0 \le x \le 1$.

Moments. $\mathbb{E}X = 1/2$ and $\mathbb{V}X = 1/12$.

Normal distribution

Definition. A random variable X has the normal (or Gaussian) distribution with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ if it has density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

If $\mu = 0$ and $\sigma^2 = 1$, the distribution is called the standard normal distribution.

Moments. $\mathbb{E}X = \mu$ and $\mathbb{V}X = \sigma^2$.

Exponential distribution

Definition. A random variable X has the exponential distribution with parameter $\lambda > 0$ if it has density $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$. If $\lambda = 1$, the distribution is called the standard exponential distribution.

Moments. For all $q > -1$,

$$\mathbb{E}X^q = \frac{\Gamma(q+1)}{\lambda^q}.$$

In particular, $\mathbb{E}X = 1/\lambda$ and $\mathbb{V}X = 1/\lambda^2$. For $q \leq -1$, $\mathbb{E}X^q = \infty$.

Beta distribution

Definition. A random variable X has the beta distribution with (shape) parameters $\alpha > 0$ and $\beta > 0$ if it has density

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \quad 0 \leq x \leq 1,$$

where

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Notation. Beta(α,β).

Moments. For all $q > -\alpha$,

$$\mathbb{E}X^q = \frac{B(\alpha+q,\beta)}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+q)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha+\beta+q)}.$$

In particular,

$$\mathbb{E}X = \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \mathbb{V}X = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Assume that α and β are positive integers. Then, for all $p \in [0,1]$, $\mathbb{P}\{\text{Beta}(\alpha,\beta) \le p\} = \mathbb{P}\{\text{Bin}(\alpha+\beta-1,p) \ge \alpha\}$. This is called the binomial-beta duality (see, e.g., Terrell, 1999, Chapter 9).

Gamma distribution

Definition. A random variable X has the gamma distribution with parameter $\alpha > 0$ if it has density

$$f(x) = \frac{x^{\alpha-1}e^{-x}}{\Gamma(\alpha)}, \quad x \ge 0.$$

Notation. Gamma(α).

Moments. For all $q > -\alpha$,

$$\mathbb{E}X^q = \frac{\Gamma(\alpha+q)}{\Gamma(\alpha)}.$$

In particular, $\mathbb{E}X = \mathbb{V}X = \alpha$.

The Gamma(1) corresponds to a standard exponential. If N is standard normal, then $2\,\text{Gamma}(1/2) \stackrel{\mathscr{D}}{=} N^2$. The random variable N^2 is also called a chi-square with one degree of freedom. In general, $2\,\text{Gamma}(n/2) \stackrel{\mathscr{D}}{=} \sum_{i=1}^n N_i^2$, where $N_1,\ldots,N_n$ are independent standard normal random variables. It is called a chi-square with n degrees of freedom.

The following lemma is left as an exercise.

Lemma 20.9. *If G_α and G_β are independent* Gamma(α) *and* Gamma(β) *random variables, then $\frac{G_\alpha}{G_\alpha+G_\beta}$ and $G_\alpha+G_\beta$ are independent* Beta(α,β) *and* Gamma($\alpha+\beta$) *random variables.*

Thus, in particular, for $E_1,\ldots,E_n$ independent standard exponential random variables, the sum $\sum_{i=1}^n E_i \stackrel{\mathscr{D}}{=}$ Gamma(n). Et donc voilà.

References

H. Akaike, An approximation to the density function. Ann. Inst. Stat. Math. **6**, 127–132 (1954)

A. Antos, L. Devroye, L. Györfi, Lower bounds for Bayes error estimation. IEEE Trans. Pattern Anal. Mach. Intell. **21**, 643–645 (1999)

J.-Y. Audibert, A.B. Tsybakov, Fast learning rates for plug-in classifiers. Ann. Stat. **35**, 608–633 (2007)

T. Bailey, A. Jain, A note on distance-weighted k-nearest neighbor rules. IEEE Trans. Syst. Man Cybern. **8**, 311–313 (1978)

J. Beck, The exponential rate of convergence of error for k_n-NN nonparametric regression and decision. Probl. Control Inf. Theory **8**, 303–311 (1979)

J. Beirlant, E.J. Dudewicz, L. Györfi, E.C. van der Meulen, Nonparametric entropy estimation: an overview. Int. J. Math. Stat. Sci. **6**, 17–39 (1997)

G. Bennett, Probability inequalities for the sum of independent random variables. J. Am. Stat. Assoc. **57**, 33–45 (1962)

A. Berlinet, S. Levallois, Higher order analysis at Lebesgue points, in *Asymptotics in Statistics and Probability*, ed. by M.L. Puri. Papers in Honor of George Gregory Roussas (VSP, Utrecht, 2000), pp. 17–32.

S.N. Bernstein, *The Theory of Probabilities* (Gastehizdat Publishing House, Moscow, 1946)

A.C. Berry, The accuracy of the Gaussian approximation to the sum of independent variates. Trans. Am. Math. Soc. **49**, 122–136 (1941)

G. Biau, F. Cérou, A. Guyader, On the rate of convergence of the bagged nearest neighbor estimate. J. Mach. Learn. Res. **11**, 687–712 (2010)

G. Biau, F. Chazal, L. Devroye, D. Cohen-Steiner, C. Rodríguez, A weighted k-nearest neighbor density estimate for geometric inference. Electron. J. Stat. **5**, 204–237 (2011)

G. Biau, L. Devroye, V. Dujmović, A. Krzyżak, An affine invariant k-nearest neighbor regression estimate. J. Multivar. Anal. **112**, 24–34 (2012)

G. Biau, F. Cérou, A. Guyader, New insights into Approximate Bayesian Computation. Ann. Inst. Henri Poincaré (B) Probab. Stat. **51**, 376–403 (2015)

P.J. Bickel, L. Breiman, Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test. Ann. Probab. **11**, 185–214 (1983)

P.J. Bickel, Y. Ritov, Estimating integrated squared density derivatives: sharp best order of convergence estimates. Sankhyā A **50**, 381–393 (1988)

P. Billingsley, *Probability and Measure*, 3rd edn. (Wiley, New York, 1995)

N.H. Bingham, C.M. Goldie, J.L. Teugels, *Regular Variation* (Cambridge University Press, Cambridge, 1987)

L. Birgé, P. Massart, Estimation of integral functionals of a density. Ann. Stat. **23**, 11–29 (1995)

G. Biau, L. Devroye, *Lectures on the Nearest Neighbor Method*,
Springer Series in the Data Sciences, DOI 10.1007/978-3-319-25388-6

K. Böröczky, Jr., *Finite Packing and Covering* (Cambridge University Press, Cambridge, 2004)
K. Böröczky, Jr., G. Wintsche, Covering the sphere by equal balls, in *Discrete and Computational Geometry: The Goodman-Pollack Festschrift*, ed. by B. Aronov, S. Basu, J. Pach, M. Sharir (Springer, Berlin, 2003), pp. 235–251
S. Boucheron, G. Lugosi, P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence* (Oxford University Press, Oxford, 2013)
L. Breiman, W. Meisel, E. Purcell, Variable kernel estimates of multivariate densities. Technometrics **19**, 135–144 (1977)
T. Cacoullos, Estimation of a multivariate density. *Ann. Inst. Stat. Math.* **18**, 178–189 (1966)
F. Cérou, A. Guyader, Nearest neighbor classification in infinite dimension. ESAIM: Probab. Stat. **10**, 340–355 (2006)
P.E. Cheng, Strong consistency of nearest neighbor regression function estimators. J. Multivar. Anal. **15**, 63–72 (1984)
H. Chernoff, A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. Ann. Math. Stat. **23**, 493–507 (1952)
G. Collomb, Estimation de la régression par la méthode des k points les plus proches avec noyau: quelques propriétés de convergence ponctuelle, in *Statistique non Paramétrique Asymptotique*, ed. by J.-P. Raoult. Lecture Notes in Mathematics, vol. 821 (Springer, Berlin, 1980), pp. 159–175
G. Collomb, Estimation non paramétrique de la régression: revue bibliographique. Int. Stat. Rev. **49**, 75–93 (1981)
T.M. Cover, Estimation by the nearest neighbor rule. IEEE Trans. Inf. Theory **14**, 50–55 (1968)
T.M. Cover, P.E. Hart, Nearest neighbor pattern classification. IEEE Trans. Inf. Theory **13**, 21–27 (1967)
T.M. Cover, J.A. Thomas, *Elements of Information Theory*, 2nd edn. (Wiley, Hoboken, 2006)
T.M. Cover, J.M. Van Campenhout, On the possible orderings in the measurement selection problem. IEEE Trans. Syst. Man Cybern. **7**, 657–661 (1977)
S. Csibi, *Stochastic Processes with Learning Properties* (Springer, Wien, 1975)
B.V. Dasarathy, *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques* (IEEE Computer Society Press, Los Alamitos, 1991)
M. de Guzmán, *Differentiation of Integrals in* $\mathbb{R}^n$. Lecture Notes in Mathematics, vol. 481 (Springer, Berlin, 1975)
P.A. Devijver, A note on ties in voting with the k-NN rule. Pattern Recogn. **10**, 297–298 (1978)
P.A. Devijver, New error bounds with the nearest neighbor rule. IEEE Trans. Inf. Theory **25**, 749–753 (1979)
P.A. Devijver, An overview of asymptotic properties of nearest neighbor rules, in *Pattern Recognition in Practice*, ed. by E.S. Gelsema, L.N. Kanal (North-Holland, Amsterdam, 1980), pp. 343–350
L. Devroye, On the almost everywhere convergence of nonparametric regression function estimates. Ann. Stat. **9**, 1310–1319 (1981a)
L. Devroye, On the inequality of Cover and Hart in nearest neighbor discrimination. IEEE Trans. Pattern Anal. Mach. Intell. **3**, 75–78 (1981b)
L. Devroye, On the asymptotic probability of error in nonparametric discrimination. Ann. Stat. **9**, 1320–1327 (1981c)
L. Devroye, Necessary and sufficient conditions for the pointwise convergence of nearest neighbor regression function estimates. Z. Warhscheinlichkeitstheorie Verwandte Geb. **61**, 467–481 (1982)
L. Devroye, *Non-Uniform Random Variate Generation* (Springer, New York, 1986)
L. Devroye, *A Course in Density Estimation* (Birkhäuser, Boston, 1987)
L. Devroye, Automatic pattern recognition: a study of the probability of error. IEEE Trans. Pattern Anal. Mach. Intell. **10**, 530–543 (1988)
L. Devroye, Exponential inequalities in nonparametric estimation, in *Nonparametric Functional Estimation and Related Topics*, ed. by G. Roussas (Springer, Dordrecht, 1991a), pp. 31–44

L. Devroye, A universal k-nearest neighbor procedure in discrimination, in *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*, ed. by B.V. Dasarathy (IEEE Computer Society Press, Los Alamitos, 1991b), pp. 101–106
L. Devroye, L. Györfi, *Nonparametric Density Estimation: The L_1 View* (Wiley, New York, 1985)
L. Devroye, A. Krzyżak, New multivariate product density estimators. J. Multivar. Anal. **82**, 88–110 (2002)
L. Devroye, G. Lugosi, *Combinatorial Methods in Density Estimation* (Springer, New York, 2001)
L. Devroye, T.J. Wagner, Nearest neighbor methods in discrimination, in *Handbook of Statistics*, vol. 2, ed. by P.R. Krishnaiah, L.N. Kanal (North-Holland, Amsterdam, 1982), pp. 193–197
L. Devroye, L. Györfi, A. Krzyżak, G. Lugosi, On the strong universal consistency of nearest neighbor regression function estimates. Ann. Stat. **22**, 1371–1385 (1994)
L. Devroye, L. Györfi, G. Lugosi, *A Probabilistic Theory of Pattern Recognition* (Springer, New York, 1996)
L. Devroye, L. Györfi, D. Schäfer, H. Walk, The estimation problem of minimum mean squared error. Stat. Decis. **21**, 15–28 (2003)
L.P. Devroye, The uniform convergence of nearest neighbor regression function estimators and their application in optimization. IEEE Trans. Inf. Theory **2**, 142–151 (1978)
L.P. Devroye, T.J. Wagner, The strong uniform consistency of nearest neighbor density estimates. Ann. Stat. **5**, 536–540 (1977)
W. Doeblin, Exposé de la théorie des chaînes simples constantes de Markov à un nombre fini d'états. Rev. Math. Union Interbalkanique **2**, 77–105 (1937)
D. Donoho, One-sided inference about functionals of a density. Ann. Stat. **16**, 1390–1420 (1988)
B. Efron, C. Stein, The jackknife estimate of variance. Ann. Stat. **9**, 586–596 (1981)
C.-G. Esseen, On the Liapunoff limit of error in the theory of probability. Arkiv Matematik Astronomi Fysik **A28**, 1–19 (1942)
D. Evans, A.J. Jones, W.M. Schmidt, Asymptotic moments of near-neighbour distance distributions. Proc. R. Soc. A **458**, 2839–2849 (2002)
C. Fefferman, E.M. Stein, Some maximal inequalities. Am. J. Math. **93**, 107–115 (1971)
E. Fix, J.L. Hodges, *Discriminatory analysis – Nonparametric discrimination: consistency properties*. Project 21-49-004, Report Number 4 (USAF School of Aviation Medicine, Randolph Field, Texas, 1951), pp. 261–279
E. Fix, J.L. Hodges, *Discriminatory analysis – Nonparametric discrimination: small sample performance*. Project 21-49-004, Report Number 11 (USAF School of Aviation Medicine, Randolph Field, Texas, 1952), pp. 280–322
E. Fix, J.L. Hodges, Discriminatory analysis: nonparametric discrimination: consistency properties, in *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*, ed. by B.V. Dasarathy (IEEE Computer Society Press, Los Alamitos, 1991a), pp. 32–39
E. Fix, J.L. Hodges, Discriminatory analysis: nonparametric discrimination: small sample performance, in *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*, ed. by B.V. Dasarathy (IEEE Computer Society Press, Los Alamitos, 1991b), pp. 40–56
J. Fritz, Distribution-free exponential error bound for nearest neighbor pattern classification. IEEE Trans. Inf. Theory **21**, 552–557 (1975)
S. Gada, T. Klein, C. Marteau, *Classification with the nearest neighbor rule in general finite dimensional spaces*. Ann. Stat. arXiv:1411.0894 (2015)
J. Galambos, *The Asymptotic Theory of Extreme Order Statistics* (Wiley, New York, 1978)
M. Giaquinta, G. Modica, *Mathematical Analysis: An Introduction to Functions of Several Variables* (Birkhäuser, Boston, 2009)
N. Glick, Sample-based multinomial classification. Biometrics **29**, 241–256 (1973)
G.R. Grimmett, D.R. Stirzaker, *Probability and Random Processes*, 3rd edn. (Oxford University Press, Oxford, 2001)
B. Grünbaum, *Arrangements and Spreads* (American Mathematical Society, Providence, 1972)
I. Guyon, A. Elisseeff, An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)

L. Györfi, An upper bound of error probabilities for multihypothesis testing and its application in adaptive pattern recognition. Probl. Control Inf. Theory **5**, 449–457 (1976)
L. Györfi, On the rate of convergence of nearest neighbor rules. IEEE Trans. Inf. Theory **24**, 509–512 (1978)
L. Györfi, Z. Györfi, An upper bound on the asymptotic error probability of the k-nearest neighbor rule for multiple classes. IEEE Trans. Inf. Theory **24**, 512–514 (1978)
L. Györfi, M. Kohler, A. Krzyżak, H. Walk, *A Distribution-Free Theory of Nonparametric Regression* (Springer, New York, 2002)
T. Hagerup, C. Rüb, A guided tour of Chernoff bounds. Inf. Process. Lett. **33**, 305–308 (1990)
P. Hall, On near neighbour estimates of a multivariate density. J. Multivar. Anal. **13**, 24–39 (1983)
T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn. (Springer, New York, 2009)
W. Hoeffding, Probability inequalities for sums of bounded random variables. J. Am. Stat. Assoc. **58**, 13–30 (1963)
O. Kallenberg, *Foundations of Modern Probability*, 2nd edn. (Springer, New York, 2002)
R.M. Karp, *Probabilistic Analysis of Algorithms*. Class Notes (University of California, Berkeley, 1988)
E. Kaufmann, R.-D. Reiss, On conditional distributions of nearest neighbors. J. Multivar. Anal. **42**, 67–76 (1992)
J.D. Kečkić, P.M. Vasić, Some inequalities for the gamma function. Publ. Inst. Math. **11**, 107–114 (1971)
J. Kiefer, Iterated logarithm analogues for sample quantiles when $p_n \downarrow 0$, in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, ed. by L.M. Le Cam, J. Neyman, E.L. Scott. Theory of Statistics, vol. 1 (University of California Press, Berkeley, 1972), pp. 227–244
B.K. Kim, J. Van Ryzin, Uniform consistency of a histogram density estimator and modal estimation. Commun. Stat. **4**, 303–315 (1975)
R. Kohavi, G.H. John, Wrappers for feature subset selection. Artif. Intell. **97**, 273–324 (1997)
M. Kohler, A. Krzyżak, On the rate of convergence of local averaging plug-in classification rules under a margin condition. IEEE Trans. Inf. Theory **53**, 1735–1742 (2007)
M. Kohler, A. Krzyżak, H. Walk, Rates of convergence for partitioning and nearest neighbor regression estimates with unbounded data. J. Multivar. Anal. **97**, 311–323 (2006)
L.F. Kozachenko, N.N. Leonenko, Sample estimate of the entropy of a random vector. Probl. Inf. Transm. **23**, 95–101 (1987)
S.R. Kulkarni, S.E. Posner, Rates of convergence of nearest neighbor estimation under arbitrary sampling. IEEE Trans. Inf. Theory **41**, 1028–1039 (1995)
V. Kumar, S. Minz, Feature selection: a literature review. Smart Comput. Rev. **4**, 211–229 (2014)
S.L. Lai, *Large Sample Properties of k-Nearest Neighbor Procedures*. Ph.D. Thesis, University of California, Los Angeles, 1977
B. Laurent, Efficient estimation of integral functionals of a density. Ann. Stat. **24**, 659–681 (1996)
N. Leonenko, L. Pronzato, V. Savani, A class of Rényi information estimators for multidimensional densities. Ann. Stat. **36**, 2153–2182 (2008)
E. Liitiäinen, A. Lendasse, F. Corona, Non-parametric residual variance estimation in supervised learning, in *Computational and Ambient Intelligence: 9th International Work-Conference on Artificial Neural Networks*, ed. by F. Sandoval, A. Prieto, J. Cabestany, M. Graña (Springer, Berlin, 2007), pp. 63–71
E. Liitiäinen, A. Lendasse, F. Corona, Bounds on the mean power-weighted nearest neighbour distance. Proc. R. Soc. A **464**, 2293–2301 (2008a)
E. Liitiäinen, A. Lendasse, F. Corona, On nonparametric residual variance estimation. Neural Process. Lett. **28**, 155–167 (2008b)
E. Liitiäinen, F. Corona, A. Lendasse, Residual variance estimation using a nearest neighbor statistic. J. Multivar. Anal. **101**, 811–823 (2010)
J.W. Lindeberg, Über das Exponentialgesetz in der Wahrscheinlichkeitsrechnung. Ann. Acad. Sci. Fenn. **16**, 1–23 (1920)

D.O. Loftsgaarden, C.P. Quesenberry, A nonparametric estimate of a multivariate density function. Ann. Math. Stat. **36**, 1049–1051 (1965)
Y.P. Mack, Asymptotic normality of multivariate k-NN density estimates. Sankhyā A **42**, 53–63 (1980)
Y.P. Mack, Local properties of k-NN regression estimates. SIAM J. Algorithms Discret. Meth. **2**, 311–323 (1981)
Y.P. Mack, Rate of strong uniform convergence of k-NN density estimates. J. Stati. Plann. Inference **8**, 185–192 (1983)
Y.P. Mack, M. Rosenblatt, Multivariate k-nearest neighbor density estimates. J. Multivar. Anal. **9**, 1–15 (1979)
J. Marcinkiewicz, A. Zygmund, Sur les fonctions indépendantes. Fundam. Math. **29**, 60–90 (1937)
A.W. Marshall, I. Olkin, *Inequalities: Theory of Majorization and Its Applications* (Academic Press, New York, 1979)
P. Massart, *Concentration Inequalities and Model Selection* (Springer, Berlin, 2007)
P. Massart, E. Nédélec, Risk bounds for statistical learning. Ann. Stat. **34**, 2326–2366 (2006)
J. Matoušek, *Lectures on Discrete Geometry* (Springer, New York, 2002)
C. McDiarmid, On the method of bounded differences, in *Surveys in Combinatorics*, ed. by J. Siemons. London Mathematical Society Lecture Note Series, vol. 141 (Cambridge University Press, Cambridge, 1989), pp. 148–188
J.V. Michalowicz, J.M. Nichols, F. Bucholtz, *Handbook of Differential Entropy* (CRC, Boca Raton, 2014)
K.S. Miller, *Multidimensional Gaussian Distributions* (Wiley, New York, 1964)
J.W. Milnor, On the Betti numbers of real algebraic varieties. Proc. Am. Math. Soc. **15**, 275–280 (1964)
D.S. Moore, E.G. Henrichon, Uniform consistency of some estimates of a density function. Ann. Math. Stat. **40**, 1499–1502 (1969)
D.S. Moore, J.W. Yackel, Large sample properties of nearest neighbor density function estimators, in *Statistical Decision Theory and Related Topics II: Proceedings of a Symposium Held at Purdue University, May 17–19, 1976*, ed. by S.S. Gupta, D.S. Moore (Academic Press, New York, 1977a), pp. 269–279
D.S. Moore, J.W. Yackel, Consistency properties of nearest neighbor density function estimators. Ann. Stat. **5**, 143–154 (1977b)
C. Mortici, C.-P. Chen, New sharp double inequalities for bounding the gamma and digamma function. Analele Universităţii de Vest din Timişoara, Seria Matematică-Informatică **49**, 69–75 (2011)
E.A. Nadaraya, On estimating regression. Theory Probab. Appl. **9**, 141–142 (1964)
E.A. Nadaraya, On nonparametric estimates of density functions and regression curves. Theory Probab. Appl. **10**, 186–190 (1965)
R. Olshen, Discussion on a paper by C.J. Stone. Ann. Stat. **5**, 632–633 (1977)
E. Parzen, On the estimation of a probability density function and the mode. Ann. Math. Stat. **33**, 1065–1076 (1962)
M.D. Penrose, J.E. Yukich, Laws of large numbers and nearest neighbor distances, in *Advances in Directional and Linear Statistics: A Festschrift for Sreenivasa Rao Jammalamadaka*, ed. by M.T. Wells, A. SenGupta (Physica, Heidelberg, 2011), pp. 189–199
V.V. Petrov, Sums of Independent Random Variables (Springer, Berlin, 1975)
I.G. Petrovskiĭ, O.A. Oleĭnik, On the topology of real algebraic surfaces. Am. Math. Soc. Translat. **70** (1952)
R. Pollack, M.-F. Roy, On the number of cells defined by a set of polynomials. Comp. R. Acad. Sci. Sér. 1: Math. **316**, 573–577 (1993)
S.T. Rachev, L. Rüschendorf, *Mass Transportation Problems. Volume I: Theory* (Springer, New York, 1998)
B.L.S. Prakasa Rao, *Nonparametric Functional Estimation* (Academic Press, Orlando, 1983)

A. Rényi, On measures of entropy and information, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Contributions to the Theory of Statistics, vol. 1 (University of California Press, Berkeley, 1961), pp. 547–561

A. Rényi, *Probability Theory* (North-Holland, Amsterdam, 1970)

C. Rodríguez, J. Van Ryzin, Large sample properties of maximum entropy histograms. IEEE Trans. Inf. Theory **32**, 751–759 (1986)

C.C. Rodríguez, On a new class of density estimators. Technical Report (Department of Mathematics and Statistics, University at Albany, Albany, 1986)

C.C. Rodríguez, Optimal recovery of local truth, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 19th International Workshop*, vol. 567, ed. by J.T. Rychert, G.J. Erickson, C.R. Smith (American Institute of Physics Conference Proceedings, Melville, 2001), pp. 89–115

C.C. Rodríguez, J. Van Ryzin, Maximum entropy histograms. Stat. Probab. Lett. **3**, 117–120 (1985)

M. Rosenblatt, Remarks on some nonparametric estimates of a density function. Ann. Math. Stat. **27**, 832–837 (1956)

R.M. Royall, A class of non-parametric estimates of a smooth regression function. Technical Report No. 14 (Department of Statistics, Stanford University, Stanford, 1966)

R.J. Samworth, Optimal weighted nearest neighbour classifiers. Ann. Stat. **40**, 2733–2763 (2012)

D.W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization* (Wiley, New York, 1992)

C.E. Shannon, A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423 (1948)

B.W. Silverman, *Density Estimation for Statistics and Data Analysis* (Chapman & Hall, London, 1986)

J.M. Steele, An Efron-Stein inequality for nonparametric statistics. Ann. Stat. **14**, 753–758 (1986)

E.M. Stein, *Singular Integrals and Differentiability Properties of Functions* (Princeton University Press, Princeton, 1970)

C.J. Stone, Consistent nonparametric regression (with discussion). Ann. Stat. **5**, 595–645 (1977)

C.J. Stone, Optimal rates of convergence for nonparametric estimators. Ann. Stat. **8**, 1348–1360 (1980)

C.J. Stone, Optimal global rates of convergence for nonparametric regression. Ann. Stat. **10**, 1040–1053 (1982)

W. Stute, Asymptotic normality of nearest neighbor regression function estimates. Ann. Stat. **12**, 917–926 (1984)

G.R. Terrell, *Mathematical Statistics: A Unified Introduction* (Springer, New York, 1999)

R. Thom, On the homology of real algebraic varieties, in *Differential and Combinatorial Topology*, ed. by S.S. Cairns (Princeton University Press, Princeton, 1965, in French)

Y.L. Tong, *Probability Inequalities in Multivariate Distributions* (Academic Press, New York, 1980)

A.B. Tsybakov, Optimal aggregation of classifiers in statistical learning. Ann. Stat. **32**, 135–166 (2004)

A.B. Tsybakov, *Introduction to Nonparametric Estimation* (Springer, New York, 2008)

A.B. Tsybakov, E.C. van der Meulen, Root-n consistent estimators of entropy for densities with unbounded support. Scand. J. Stat. **23**, 75–83 (1996)

L.R. Turner, Inverse of the Vandermonde matrix with applications. NASA Technical Note D-3547 (Washington, 1966)

A.W. van der Vaart, *Asymptotic Statistics* (Cambridge University Press, Cambridge, 1998)

J. Van Ryzin, Bayes risk consistency of classification procedures using density estimation. Sankhyā A **28**, 161–170 (1966)

V.N. Vapnik, A.Y. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities. Theory Probab. Appl. **16**, 264–280 (1971)

T.J. Wagner, Strong consistency of a nonparametric estimate of a density function. IEEE Trans. Syst. Man Cybern. **3**, 289–290 (1973)

H. Walk, A universal strong law of large numbers for conditional expectations via nearest neighbors. J. Multivar. Anal. **99**, 1035–1050 (2008)

H.E. Warren, Lower bounds for approximation by nonlinear manifolds. Trans. Am. Math. Soc. **133**, 167–178 (1968)
G.S. Watson, Smooth regression analysis. Sankhyā A **26**, 359–372 (1964)
G.S. Watson, M.R. Leadbetter, On the estimation of the probability density. Ann. Math. Stat. **34**, 480–491 (1963)
R.L. Wheeden, A. Zygmund, *Measure and Integral: An Introduction to Real Analysis* (Marcel Dekker, New York, 1977)
P. Whittle, On the smoothing of probability density functions. J. R. Stat. Soc. B **20**, 334–343 (1958)
C.T. Wolverton, T.J. Wagner, Asymptotically optimal discriminant functions for pattern classification. IEEE Trans. Inf. Theory **15**, 258–265 (1969)
L.C. Zhao, Exponential bounds of mean error for the nearest neighbor estimates of regression functions. J. Multivar. Anal. **21**, 168–178 (1987)

Index

G. Biau, L. Devroye, *Lectures on the Nearest Neighbor Method*,
Springer Series in the Data Sciences, DOI 10.1007/978-3-319-25388-6