Springer Series in Statistics

Advisors:

P. Bickel, P. Diggle, S. Fienberg, U. Gather,
I. Olkin, S. Zeger

The French edition of this work that is the basis of this expanded edition was translated by Vladimir Zaiats.

For other titles published in this series, go to
http://www.springer.com/series/692

Alexandre B. Tsybakov

Introduction to Nonparametric Estimation

Springer

Alexandre B. Tsybakov
Laboratoire de Statistique of CREST
3, av. Pierre Larousse
92240 Malakoff
France

and

LPMA
University of Paris 6
4, Place Jussieu
75252 Paris
France
alexandre.tsybakov@upmc.fr

ISBN: 978-0-387-79051-0 e-ISBN: 978-0-387-79052-7
DOI 10.1007/978-0-387-79052-7

Library of Congress Control Number: 2008939894

Mathematics Subject Classification: 62G05, 62G07, 62G20

Printed on acid-free paper

springer.com

Preface to the English Edition

This is a revised and extended version of the French book. The main changes are in Chapter 1 where the former Section 1.3 is removed and the rest of the material is substantially revised. Sections 1.2.4, 1.3, 1.9, and 2.7.3 are new. Each chapter now has the bibliographic notes and contains the exercises section. I would like to thank Cristina Butucea, Alexander Goldenshluger, Stephan Huckenmann, Yuri Ingster, Iain Johnstone, Vladimir Koltchinskii, Alexander Korostelev, Oleg Lepski, Karim Lounici, Axel Munk, Boaz Nadler, Alexander Nazin, Philippe Rigollet, Angelika Rohde, and Jon Wellner for their valuable remarks that helped to improve the text. I am grateful to Centre de Recherche en Economie et Statistique (CREST) and to Isaac Newton Institute for Mathematical Sciences which provided an excellent environment for finishing the work on the book. My thanks also go to Vladimir Zaiats for his highly competent translation of the French original into English and to John Kimmel for being a very supportive and patient editor.

Alexandre Tsybakov
Paris, June 2008

Preface to the French Edition

The tradition of considering the problem of statistical estimation as that of estimation of a finite number of parameters goes back to Fisher. However, parametric models provide only an approximation, often imprecise, of the underlying statistical structure. Statistical models that explain the data in a more consistent way are often more complex: Unknown elements in these models are, in general, some functions having certain properties of smoothness. The problem of nonparametric estimation consists in estimation, from the observations, of an unknown function belonging to a sufficiently large class of functions.

The theory of nonparametric estimation has been considerably developed during the last two decades focusing on the following fundamental topics:

(1) methods of construction of the estimators
(2) statistical properties of the estimators (convergence, rates of convergence)
(3) study of optimality of the estimators
(4) adaptive estimation.

Basic topics (1) and (2) will be discussed in Chapter 1, though we mainly focus on topics (3) and (4), which are placed at the core of this book. We will first construct estimators having optimal rates of convergence in a minimax sense for different classes of functions and different distances defining the risk. Next, we will study optimal estimators in the exact minimax sense presenting, in particular, a proof of Pinsker's theorem. Finally, we will analyze the problem of adaptive estimation in the Gaussian sequence model. A link between Stein's phenomenon and adaptivity will be discussed.

This book is an introduction to the theory of nonparametric estimation. It does not aim at giving an encyclopedic covering of the existing theory or an initiation in applications. It rather treats some simple models and examples in order to present basic ideas and tools of nonparametric estimation. We prove, in a detailed and relatively elementary way, a number of classical results that are well-known to experts but whose original proofs are sometimes

neither explicit nor easily accessible. We consider models with independent observations only; the case of dependent data adds nothing conceptually but introduces some technical difficulties.

This book is based on the courses taught at the MIEM (1991), the Katholieke Universiteit Leuven (1991–1993), the Université Pierre et Marie Curie (1993–2002) and the Institut Henri Poincaré (2001), as well as on mini-courses given at the Humboldt University of Berlin (1994), the Heidelberg University (1995) and the Seminar Paris–Berlin (Garchy, 1996). The contents of the courses have been considerably modified since the earlier versions. The structure and the size of the book (except for Sections 1.3, 1.4, 1.5, and 2.7) correspond essentially to the graduate course that I taught for many years at the Université Pierre et Marie Curie. I would like to thank my students, colleagues, and all those who attended this course for their questions and remarks that helped to improve the presentation.

I also thank Karine Bertin, Gérard Biau, Cristina Butucea, Laurent Cavalier, Arnak Dalalyan, Yuri Golubev, Alexander Gushchin, Gérard Kerkyacharian, Béatrice Laurent, Oleg Lepski, Pascal Massart, Alexander Nazin, and Dominique Picard for their remarks on different versions of the book. My special thanks go to Lucien Birgé and Xavier Guyon for numerous improvements that they have suggested. I am also grateful to Josette Saman for her help in typing of a preliminary version of the text.

Alexandre Tsybakov
Paris, April 2003

Notation

$\lfloor x \rfloor$	greatest integer strictly less than the real number x
$\lceil x \rceil$	smallest integer strictly larger than the real number x
x_+	$\max(x, 0)$
log	natural logarithm
$I(A)$	indicator of the set A
Card A	cardinality of the set A
$\triangleq$	equals by definition
$\lambda_{\min}(B)$	smallest eigenvalue of the symmetric matrix B
a^T, B^T	transpose of the vector a or of the matrix B
$\|\cdot\|_p$	$L_p([0,1], dx)$-norm or $L_p(\mathbf{R}, dx)$-norm for $1 \leq p \leq \infty$ depending on the context
$\|\cdot\|$	$\ell^2(\mathbf{N})$-norm or the Euclidean norm in $\mathbf{R}^d$, depending on the context
$\mathcal{N}(a, \sigma^2)$	normal distribution on $\mathbf{R}$ with mean a and variance σ^2
$\mathcal{N}_d(0, I)$	standard normal distribution in $\mathbf{R}^d$
$\varphi(\cdot)$	density of the distribution $\mathcal{N}(0, 1)$
$P \ll Q$	the measure P is absolutely continuous with respect to the measure Q

dP/dQ	the Radon–Nikodym derivative of the measure P with respect to the measure Q
$a_n \asymp b_n$	$0 < \liminf_{n\to\infty}(a_n/b_n) \le \limsup_{n\to\infty}(a_n/b_n) < \infty$
$h^* = \arg\min_{h\in H} F(h)$	means that $F(h^*) = \min_{h\in H} F(h)$
MSE	mean squared risk at a point (p. 4, p. 37)
MISE	mean integrated squared error (p. 12, p. 51)
$\Sigma(\beta, L)$	Hölder class of functions (p. 5)
$\mathcal{H}(\beta, L)$	Nikol'ski class of functions (p. 13)
$\mathcal{P}(\beta, L)$	Hölder class of densities (p. 6)
$\mathcal{P}_{\mathcal{H}}(\beta, L)$	Nikol'ski class of densities (p. 13)
$\mathcal{S}(\beta, L)$	Sobolev class of functions on $\mathbf{R}$ (p. 13)
$\mathcal{P}_{\mathcal{S}}(\beta, L)$	Sobolev class of densities (p. 25)
$W(\beta, L)$	Sobolev class of functions on $[0, 1]$ (p. 49)
$W^{per}(\beta, L)$	periodic Sobolev class (p. 49)
$\tilde{W}(\beta, L)$	Sobolev class based on an ellipsoid (p. 50)
$\Theta(\beta, Q)$	Sobolev ellipsoid (p. 50)
$H(P, Q)$	Hellinger distance between the measures P and Q (p. 83)
$V(P, Q)$	total variation distance between the measures P and Q (p. 83)
$K(P, Q)$	Kullback divergence between the measures P and Q (p. 84)
$\chi^2(P, Q)$	χ^2 divergence between the measures P and Q (p. 86)
ψ_n	optimal rate of convergence (p. 78)
$p_{e,M}$	minimax probability of error (p. 80)
$\overline{p}_{e,M}$	average probability of error (p. 111)
C^*	the Pinsker constant (p. 138)
$R(\lambda, \theta)$	integrated squared risk of the linear estimator (p. 67)

Assumption (A)	p. 51
Assumption (B)	p. 91
Assumption (C)	p. 174
Assumptions (LP)	p. 37

Contents

1

Nonparametric estimators

1.1 Examples of nonparametric models and problems

1. Estimation of a probability density

Let $X_1, \dots, X_n$ be identically distributed real valued random variables whose common distribution is absolutely continuous with respect to the Lebesgue measure on $\mathbf{R}$. The density of this distribution, denoted by p, is a function from $\mathbf{R}$ to $[0, +\infty)$ supposed to be unknown. The problem is to estimate p. An estimator of p is a function $x \mapsto p_n(x) = p_n(x, X_1, \dots, X_n)$ measurable with respect to the observation $\mathbf{X} = (X_1, \dots, X_n)$. If we know a priori that p belongs to a parametric family $\{g(x, \theta) : \theta \in \Theta\}$, where $g(\cdot, \cdot)$ is a given function, and Θ is a subset of $\mathbf{R}^k$ with a fixed dimension k independent of n, then estimation of p is equivalent to estimation of the finite-dimensional parameter θ. This is a *parametric* problem of estimation. On the contrary, if such a prior information about p is not available we deal with a *nonparametric* problem. In nonparametric estimation it is usually assumed that p belongs to some "massive" class $\mathcal{P}$ of densities. For example, $\mathcal{P}$ can be the set of all the continuous probability densities on $\mathbf{R}$ or the set of all the Lipschitz continuous probability densities on $\mathbf{R}$. Classes of such type will be called nonparametric classes of functions.

2. Nonparametric regression

Assume that we have n independent pairs of random variables $(X_1, Y_1), \dots, (X_n, Y_n)$ such that

$$Y_i = f(X_i) + \xi_i, \quad X_i \in [0, 1], \tag{1.1}$$

where the random variables ξ_i satisfy $\mathbf{E}(\xi_i) = 0$ for all i and where the function f from $[0, 1]$ to $\mathbf{R}$ (called the regression function) is unknown. The problem of nonparametric regression is to estimate f given a priori that this function belongs to a nonparametric class of functions $\mathcal{F}$. For example, $\mathcal{F}$ can be the set of all the continuous functions on $[0, 1]$ or the set of

A. B. Tsybakov, *Introduction to Nonparametric Estimation*,
DOI 10.1007/978-0-387-79052-7_1,

all the convex functions, etc. An estimator of f is a function $x \mapsto f_n(x) = f_n(x, \mathbf{X})$ defined on $[0,1]$ and measurable with respect to the observation $\mathbf{X} = (X_1, \ldots, X_n, Y_1, \ldots, Y_n)$. In what follows, we will mainly focus on the particular case $X_i = i/n$.

3. Gaussian white noise model

This is an idealized model that provides an approximation to the nonparametric regression (1.1). Consider the following stochastic differential equation:

$$dY(t) = f(t)dt + \frac{1}{\sqrt{n}}\, dW(t), \quad t \in [0,1],$$

where W is a standard Wiener process on $[0,1]$, the function f is an unknown function on $[0,1]$, and n is an integer. We assume that a sample path $\mathbf{X} = \{Y(t), 0 \le t \le 1\}$ of the process Y is observed. The statistical problem is to estimate the unknown function f. In the nonparametric case it is only known a priori that $f \in \mathcal{F}$ where $\mathcal{F}$ is a given nonparametric class of functions. An estimator of f is a function $x \mapsto f_n(x) = f_n(x, \mathbf{X})$ defined on $[0,1]$ and measurable with respect to the observation $\mathbf{X}$.

In either of the three above cases, we are interested in the asymptotic behavior of estimators as $n \to \infty$.

1.2 Kernel density estimators

We start with the first of the three problems described in Section 1.1. Let $X_1, \ldots, X_n$ be independent identically distributed (i.i.d.) random variables that have a probability density p with respect to the Lebesgue measure on $\mathbf{R}$. The corresponding distribution function is $F(x) = \int_{-\infty}^{x} p(t)dt$. Consider the empirical distribution function

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \le x),$$

where $I(\cdot)$ denotes the indicator function. By the strong law of large numbers, we have

$$F_n(x) \to F(x), \quad \forall\, x \in \mathbf{R},$$

almost surely as $n \to \infty$. Therefore, $F_n(x)$ is a consistent estimator of $F(x)$ for every $x \in \mathbf{R}$. How can we estimate the density p? One of the first intuitive solutions is based on the following argument. For sufficiently small $h > 0$ we can write an approximation

$$p(x) \approx \frac{F(x+h) - F(x-h)}{2h}\,.$$

Replacing F by the estimate F_n we define

$$\hat{p}_n^R(x) = \frac{F_n(x+h) - F_n(x-h)}{2h}\,.$$

The function $\hat{p}_n^R$ is an estimator of p called the *Rosenblatt estimator*. We can rewrite it in the form:

$$\hat{p}_n^R(x) = \frac{1}{2nh}\sum_{i=1}^{n} I(x-h < X_i \le x+h) = \frac{1}{nh}\sum_{i=1}^{n} K_0\left(\frac{X_i - x}{h}\right),$$

where $K_0(u) = \frac{1}{2}\, I(-1 < u \le 1)$. A simple generalization of the Rosenblatt estimator is given by

$$\hat{p}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right), \tag{1.2}$$

where $K : \mathbf{R} \to \mathbf{R}$ is an integrable function satisfying $\int K(u)du = 1$. Such a function K is called a *kernel* and the parameter h is called a *bandwidth* of the estimator (1.2). The function $x \mapsto \hat{p}_n(x)$ is called the *kernel density estimator* or the *Parzen–Rosenblatt estimator*.

In the asymptotic framework, as $n \to \infty$, we will consider a bandwidth h that depends on n, denoting it by h_n, and we will suppose that the sequence $(h_n)_{n\ge1}$ tends to 0 as $n \to \infty$. The notation h without index n will also be used for brevity whenever this causes no ambiguity.

Some classical examples of kernels are the following:

$K(u) = \frac{1}{2}\, I(|u| \le 1)$ (the rectangular kernel),
$K(u) = (1-|u|)I(|u| \le 1)$ (the triangular kernel),
$K(u) = \frac{3}{4}\,(1-u^2)I(|u| \le 1)$ (the parabolic kernel, or the Epanechnikov kernel),
$K(u) = \frac{15}{16}\,(1-u^2)^2 I(|u| \le 1)$ (the biweight kernel),
$K(u) = \frac{1}{\sqrt{2\pi}}\,\exp(-u^2/2)$ (the Gaussian kernel),
$K(u) = \frac{1}{2}\,\exp(-|u|/\sqrt{2})\sin(|u|/\sqrt{2} + \pi/4)$ (the Silverman kernel).

Note that if the kernel K takes only nonnegative values and if $X_1, \ldots, X_n$ are fixed, then the function $x \mapsto \hat{p}_n(x)$ is a probability density.

The Parzen–Rosenblatt estimator can be generalized to the multidimensional case. For example, we can define a kernel density estimator in two dimensions as follows. Suppose that we observe n pairs of random variables $(X_1, Y_1), \ldots, (X_n, Y_n)$ such that (X_i, Y_i) are i.i.d. with a density $p(x, y)$ in $\mathbf{R}^2$. A kernel estimator of $p(x, y)$ is then given by the formula

$$\hat{p}_n(x,y) = \frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\ K\left(\frac{Y_i - y}{h}\right) \tag{1.3}$$

where $K : \mathbf{R} \to \mathbf{R}$ is a kernel defined as above and $h > 0$ is a bandwidth.

1.2.1 Mean squared error of kernel estimators

A basic measure of the accuracy of estimator $\hat{p}_n$ is its *mean squared risk* (or *mean squared error*) at an arbitrary fixed point $x_0 \in \mathbf{R}$:

$$\mathrm{MSE} = \mathrm{MSE}(x_0) \stackrel{\triangle}{=} \mathbf{E}_p\Big[(\hat{p}_n(x_0) - p(x_0))^2\Big].$$

Here, MSE stands for "mean squared error" and $\mathbf{E}_p$ denotes the expectation with respect to the distribution of $(X_1, \ldots, X_n)$:

$$\mathbf{E}_p\Big[(\hat{p}_n(x_0) - p(x_0))^2\Big] \stackrel{\triangle}{=} \int \ldots \int (\hat{p}_n(x_0, x_1, \ldots, x_n) - p(x_0))^2 \prod_{i=1}^{n} [p(x_i)dx_i]\,.$$

We have

$$\mathrm{MSE} = b^2(x_0) + \sigma^2(x_0) \tag{1.4}$$

where

$$b(x_0) = \mathbf{E}_p[\hat{p}_n(x_0)] - p(x_0)$$

and

$$\sigma^2(x_0) = \mathbf{E}_p\Big[\Big(\hat{p}_n(x_0) - \mathbf{E}_p[\hat{p}_n(x_0)]\Big)^2\Big].$$

Definition 1.1 *The quantities $b(x_0)$ and $\sigma^2(x_0)$ are called the* **bias** *and the* **variance** *of the estimator $\hat{p}_n$ at a point x_0, respectively.*

To evaluate the mean squared risk of $\hat{p}_n$ we will analyze separately its variance and bias.

Variance of the estimator $\hat{p}_n$

Proposition 1.1 *Suppose that the density p satisfies $p(x) \le p_{\max} < \infty$ for all $x \in \mathbf{R}$. Let $K : \mathbf{R} \to \mathbf{R}$ be a function such that*

$$\int K^2(u)du < \infty. \tag{1.5}$$

Then for any $x_0 \in \mathbf{R}, h > 0$, and $n \ge 1$ we have

$$\sigma^2(x_0) \le \frac{C_1}{nh}$$

where $C_1 = p_{\max} \int K^2(u)du$.

PROOF. Put

$$\eta_i(x_0) = K\left(\frac{X_i - x_0}{h}\right) - \mathbf{E}_p\left[K\left(\frac{X_i - x_0}{h}\right)\right].$$

The random variables $\eta_i(x_0), i = 1, \ldots, n$, are i.i.d. with zero mean and variance

$$\begin{aligned}\mathbf{E}_p\left[\eta_i^2(x_0)\right] &\le \mathbf{E}_p\left[K^2\left(\frac{X_i - x_0}{h}\right)\right] \\ &= \int K^2\left(\frac{z - x_0}{h}\right) p(z)dz \le p_{\max} h \int K^2(u)du.\end{aligned}$$

Then

$$\sigma^2(x_0) = \mathbf{E}_p\left[\left(\frac{1}{nh}\sum_{i=1}^{n}\eta_i(x_0)\right)^2\right] = \frac{1}{nh^2}\mathbf{E}_p\left[\eta_1^2(x_0)\right] \le \frac{C_1}{nh}. \qquad (1.6)$$

∎

We conclude that if the bandwidth $h = h_n$ is such that $nh \to \infty$ as $n \to \infty$, then the variance $\sigma^2(x_0)$ goes to 0 as $n \to \infty$.

Bias of the estimator $\hat{p}_n$

The bias of the kernel density estimator has the form

$$b(x_0) = \mathbf{E}_p[\hat{p}_n(x_0)] - p(x_0) = \frac{1}{h}\int K\left(\frac{z - x_0}{h}\right) p(z)dz - p(x_0).$$

We now analyze the behavior of $b(x_0)$ as a function of h under some regularity conditions on the density p and on the kernel K.

In what follows $\lfloor\beta\rfloor$ will denote the greatest integer strictly less than the real number β.

Definition 1.2 *Let T be an interval in $\mathbf{R}$ and let β and L be two positive numbers. The* **Hölder class** $\Sigma(\beta, L)$ **on** T *is defined as the set of $\ell = \lfloor\beta\rfloor$ times differentiable functions $f : T \to \mathbf{R}$ whose derivative $f^{(\ell)}$ satisfies*

$$|f^{(\ell)}(x) - f^{(\ell)}(x')| \le L|x - x'|^{\beta - \ell}, \quad \forall\, x, x' \in T.$$

Definition 1.3 *Let $\ell \ge 1$ be an integer. We say that $K : \mathbf{R} \to \mathbf{R}$ is a* **kernel of order** ℓ *if the functions $u \mapsto u^j K(u), j = 0, 1, \ldots, \ell$, are integrable and satisfy*

$$\int K(u)du = 1, \qquad \int u^j K(u)du = 0, \quad j = 1, \ldots, \ell.$$

Some examples of kernels of order ℓ will be given in Section 1.2.2. It is important to note that another definition of an order ℓ kernel is often used in the literature: a kernel K is said to be of order $\ell+1$ (with integer $\ell \geq 1$) if Definition 1.3 holds and $\int u^{\ell+1}K(u)du \neq 0$. Definition 1.3 is less restrictive and seems to be more natural, since there is no need to assume that $\int u^{\ell+1}K(u)du \neq 0$ for noninteger β. For example, Proposition 1.2 given below still holds if $\int u^{\ell+1}K(u)du = 0$ and even if this integral does not exist.

Suppose now that p belongs to the class of densities $\mathcal{P} = \mathcal{P}(\beta, L)$ defined as follows:

$$\mathcal{P}(\beta, L) = \left\{ p \;\middle|\; p \geq 0, \int p(x)dx = 1, \text{ and } p \in \Sigma(\beta, L) \text{ on } \mathbf{R} \right\}$$

and assume that K is a kernel of order ℓ. Then the following result holds.

Proposition 1.2 *Assume that $p \in \mathcal{P}(\beta, L)$ and let K be a kernel of order $\ell = \lfloor \beta \rfloor$ satisfying*

$$\int |u|^{\beta}|K(u)|du < \infty.$$

Then for all $x_0 \in \mathbf{R}$, $h > 0$ and $n \geq 1$ we have

$$|b(x_0)| \leq C_2 h^{\beta}$$

where

$$C_2 = \frac{L}{\ell\,!} \int |u|^{\beta}|K(u)|du.$$

Proof. We have

$$\begin{aligned} b(x_0) &= \frac{1}{h}\int K\left(\frac{z-x_0}{h}\right)p(z)dz - p(x_0) \\ &= \int K(u)\Big[p(x_0+uh) - p(x_0)\Big]du. \end{aligned}$$

Next,

$$p(x_0+uh) = p(x_0) + p'(x_0)uh + \cdots + \frac{(uh)^{\ell}}{\ell\,!}p^{(\ell)}(x_0+\tau uh), \tag{1.7}$$

where $0 \leq \tau \leq 1$. Since K has order $\ell = \lfloor \beta \rfloor$, we obtain

$$\begin{aligned} b(x_0) &= \int K(u)\frac{(uh)^{\ell}}{\ell\,!}p^{(\ell)}(x_0+\tau uh)du \\ &= \int K(u)\frac{(uh)^{\ell}}{\ell!}(p^{(\ell)}(x_0+\tau uh) - p^{(\ell)}(x_0))du \end{aligned}$$

and

$$|b(x_0)| \le \int |K(u)| \frac{|uh|^\ell}{\ell\,!} \left| p^{(\ell)}(x_0 + \tau uh) - p^{(\ell)}(x_0) \right| du$$

$$\le L \int |K(u)| \frac{|uh|^\ell}{\ell\,!} \, |\tau uh|^{\beta-\ell} \, du \le C_2 h^\beta.$$

■

Upper bound on the mean squared risk

From Propositions 1.1 and 1.2, we see that the upper bounds on the bias and variance behave in opposite ways as the bandwidth h varies. The variance decreases as h grows, whereas the bound on the bias increases (cf. Figure 1.1). The choice of a small h corresponding to a large variance is called an *un-*

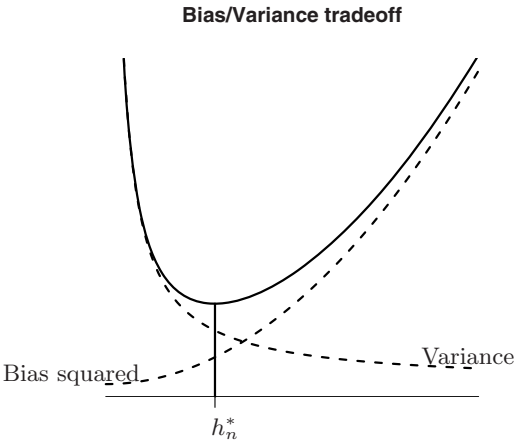

Figure 1.1. Squared bias, variance, and mean squared error (solid line) as functions of h.

dersmoothing. Alternatively, with a large h the bias cannot be reasonably controlled, which leads to *oversmoothing*. An optimal value of h that balances bias and variance is located between these two extremes. Figure 1.2 shows typical plots of the corresponding density estimators. To get an insight into the optimal choice of h, we can minimize in h the upper bound on the MSE obtained from the above results.

If p and K satisfy the assumptions of Propositions 1.1 and 1.2, we obtain

$$\mathrm{MSE} \le C_2^2 h^{2\beta} + \frac{C_1}{nh}\,. \tag{1.8}$$

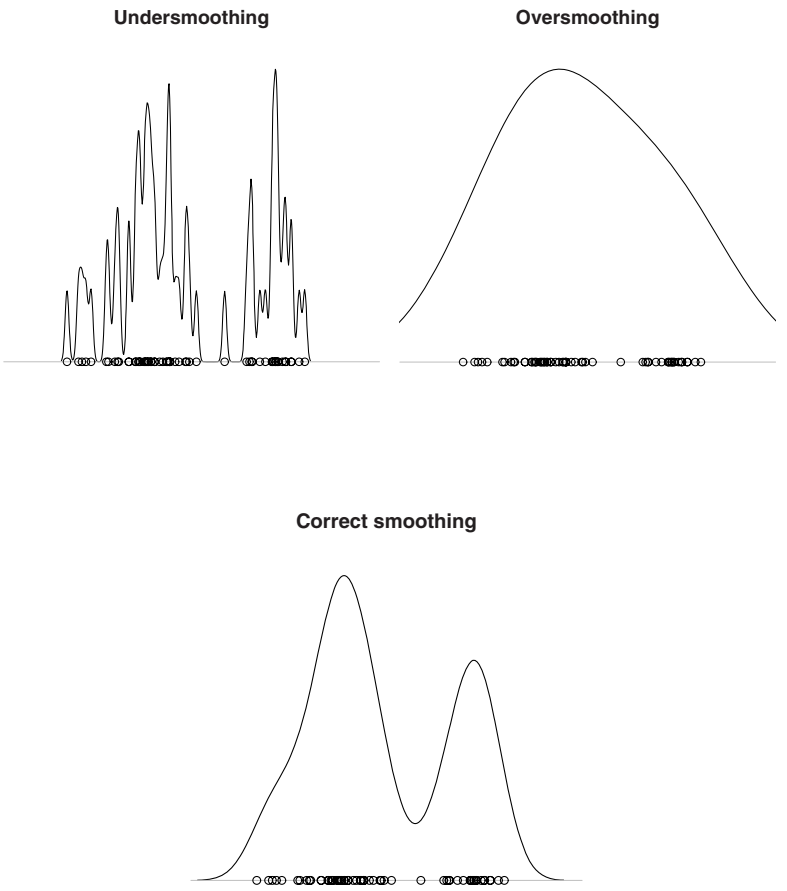

Figure 1.2. Undersmoothing, oversmoothing, and correct smoothing. The circles indicate the sample points X_i.

The minimum with respect to h of the right hand side of (1.8) is attained at

$$h_n^* = \left(\frac{C_1}{2\beta C_2^2} \right)^{\frac{1}{2\beta+1}} n^{-\frac{1}{2\beta+1}}.$$

Therefore, the choice $h = h_n^*$ gives

$$\mathrm{MSE}(x_0) = O\left(n^{-\frac{2\beta}{2\beta+1}} \right), \quad n \to \infty,$$

uniformly in x_0. We have the following result.

Theorem 1.1 *Assume that condition (1.5) holds and the assumptions of Proposition 1.2 are satisfied. Fix $\alpha > 0$ and take $h = \alpha n^{-\frac{1}{2\beta+1}}$. Then for $n \geq 1$ the kernel estimator $\hat{p}_n$ satisfies*

$$\sup_{x_0 \in \mathbf{R}} \sup_{p \in \mathcal{P}(\beta, L)} \mathbf{E}_p[(\hat{p}_n(x_0) - p(x_0))^2] \leq C n^{-\frac{2\beta}{2\beta+1}},$$

where $C > 0$ is a constant depending only on β, L, α and on the kernel K.

Proof. We apply (1.8) as shown above. To justify the application of Proposition 1.1, it remains to prove that there exists a constant $p_{\max} < \infty$ satisfying

$$\sup_{x \in \mathbf{R}} \sup_{p \in \mathcal{P}(\beta, L)} p(x) \leq p_{\max}. \tag{1.9}$$

To show (1.9), consider K^* which is a bounded kernel of order ℓ, not necessarily equal to K. Applying Proposition 1.2 with $h = 1$ we get that, for any $x_0 \in \mathbf{R}$ and any $p \in \mathcal{P}(\beta, L)$,

$$\left| \int K^*(z - x_0) p(z) dz - p(x_0) \right| \leq C_2^* \stackrel{\triangle}{=} \frac{L}{\ell!} \int |u|^\beta |K^*(u)| du.$$

Therefore, for any $x \in \mathbf{R}$ and any $p \in \mathcal{P}(\beta, L)$,

$$p(x) \leq C_2^* + \int |K^*(z - x)| p(z) dz \leq C_2^* + K_{\max}^*,$$

where $K_{\max}^* = \sup_{u \in \mathbf{R}} |K^*(u)|$. Thus, we get (1.9) with $p_{\max} = C_2^* + K_{\max}^*$. ∎

Under the assumptions of Theorem 1.1, the *rate of convergence* of the estimator $\hat{p}_n(x_0)$ is $\psi_n = n^{-\frac{\beta}{2\beta+1}}$, which means that for a finite constant C and for all $n \geq 1$ we have

$$\sup_{p \in \mathcal{P}(\beta, L)} \mathbf{E}_p \Big[(\hat{p}_n(x_0) - p(x_0))^2 \Big] \leq C \psi_n^2.$$

Now the following two questions arise. Can we improve the rate ψ_n by using other density estimators? What is the best possible rate of convergence? To answer these questions it is useful to consider the *minimax risk* R_n^* associated to the class $\mathcal{P}(\beta, L)$:

$$R_n^*(\mathcal{P}(\beta, L)) \stackrel{\triangle}{=} \inf_{T_n} \sup_{p \in \mathcal{P}(\beta, L)} \mathbf{E}_p \Big[(T_n(x_0) - p(x_0))^2 \Big],$$

where the infimum is over *all estimators*. One can prove a lower bound on the minimax risk of the form $R_n^*(\mathcal{P}(\beta, L)) \geq C' \psi_n^2 = C' n^{-\frac{2\beta}{2\beta+1}}$ with some constant $C' > 0$ (cf. Chapter 2, Exercise 2.8). This implies that under the assumptions of Theorem 1.1 the kernel estimator attains the optimal rate of convergence $n^{-\frac{\beta}{2\beta+1}}$ associated with the class of densities $\mathcal{P}(\beta, L)$. Exact definitions and discussions of the notion of optimal rate of convergence will be given in Chapter 2.

Positivity constraint

It follows easily from Definition 1.3 that kernels of order $\ell \geq 2$ must take negative values on a set of positive Lebesgue measure. The estimators $\hat{p}_n$ based on such kernels can also take negative values. This property is sometimes emphasized as a drawback of estimators with higher order kernels, since the density p itself is nonnegative. However, this remark is of minor importance because we can always use the positive part estimator

$$\hat{p}_n^+(x) \stackrel{\triangle}{=} \max\{0, \hat{p}_n(x)\}$$

whose risk is smaller than or equal to the risk of $\hat{p}_n$:

$$\mathbf{E}_p\Big[(\hat{p}_n^+(x_0) - p(x_0))^2\Big] \leq \mathbf{E}_p\Big[(\hat{p}_n(x_0) - p(x_0))^2\Big], \quad \forall\, x_0 \in \mathbf{R}. \tag{1.10}$$

In particular, Theorem 1.1 remains valid if we replace there $\hat{p}_n$ by $\hat{p}_n^+$. Thus, the estimator $\hat{p}_n^+$ is nonnegative and attains fast convergence rates associated with higher order kernels.

1.2.2 Construction of a kernel of order ℓ

Theorem 1.1 is based on the assumption that bounded kernels of order ℓ exist. In order to construct such kernels, one can proceed as follows.

Let $\{\varphi_m(\cdot)\}_{m=0}^{\infty}$ be the orthonormal basis of Legendre polynomials in $L_2([-1,1], dx)$ defined by the formulas

$$\varphi_0(x) \equiv \frac{1}{\sqrt{2}}, \quad \varphi_m(x) = \sqrt{\frac{2m+1}{2}}\ \frac{1}{2^m m!}\ \frac{d^m}{dx^m}\Big[(x^2-1)^m\Big], \quad m = 1, 2, \ldots,$$

for $x \in [-1, 1]$. Then

$$\int_{-1}^{1} \varphi_m(u)\varphi_k(u)du = \delta_{mk}, \tag{1.11}$$

where δ_{mk} is the Kronecker delta:

$$\delta_{mk} = \begin{cases} 1, \text{ if } m = k, \\ 0, \text{ if } m \neq k. \end{cases}$$

Proposition 1.3 *The function $K : \mathbf{R} \to \mathbf{R}$ defined by the formula*

$$K(u) = \sum_{m=0}^{\ell} \varphi_m(0)\varphi_m(u) I(|u| \leq 1) \tag{1.12}$$

is a kernel of order ℓ.

PROOF. Since φ_q is a polynomial of degree q, for all $j = 0, 1, \ldots, \ell$, there exist real numbers b_{qj} such that

$$u^j = \sum_{q=0}^{j} b_{qj}\varphi_q(u) \qquad \text{for all } u \in [-1, 1]. \tag{1.13}$$

Let K be the kernel given by (1.12). Then, by (1.11) and (1.13), we have

$$\int u^j K(u)du = \sum_{q=0}^{j}\sum_{m=0}^{\ell}\int_{-1}^{1} b_{qj}\varphi_q(u)\varphi_m(0)\varphi_m(u)du =$$
$$= \sum_{q=0}^{j} b_{qj}\varphi_q(0) = \begin{cases} 1, \text{ if } j = 0, \\ 0, \text{ if } j = 1, \ldots, \ell. \end{cases}$$

∎

A kernel K is called symmetric if $K(u) = K(-u)$ for all $u \in \mathbf{R}$. Observe that the kernel K defined by (1.12) is symmetric. Indeed, we have $\varphi_m(0) = 0$ for all odd m and the Legendre polynomials φ_m are symmetric functions for all even m. By symmetry, the kernel (1.12) is of order $\ell + 1$ for even ℓ. Moreover, the explicit form of kernels (1.12) uses the Legendre polynomials of even degrees only.

Example 1.1 The first two Legendre polynomials of even degrees are

$$\varphi_0(x) \equiv \sqrt{\frac{1}{2}}, \qquad \varphi_2(x) = \sqrt{\frac{5}{2}}\,\frac{(3x^2 - 1)}{2}.$$

Then Proposition 1.3 suggests the following kernel of order 2:

$$K(u) = \left(\frac{9}{8} - \frac{15}{8}\,u^2\right) I(|u| \le 1),$$

which is also a kernel of order 3 by the symmetry.

The construction of kernels suggested in Proposition 1.3 can be extended to bases of polynomials $\{\varphi_m\}_{m=0}^{\infty}$ that are orthonormal with weights. Indeed, a slight modification of the proof of Proposition 1.3 yields that a kernel of order ℓ can be defined in the following way:

$$K(u) = \sum_{m=0}^{\ell} \varphi_m(0)\varphi_m(u)\mu(u),$$

where μ is a positive weight function on $\mathbf{R}$ satisfying $\mu(0) = 1$, the function φ_m is a polynomial of degree m, and the basis $\{\varphi_m\}_{m=0}^{\infty}$ is orthonormal with weight μ:

$$\int \varphi_m(u)\varphi_k(u)\mu(u)du = \delta_{mk}.$$

This enables us to construct various kernels of order ℓ, in particular, those corresponding to the Hermite basis ($\mu(u) = e^{-u^2}$; the support of K is $(-\infty, +\infty)$) and to the Gegenbauer basis ($\mu(u) = (1-u^2)_+^\alpha$ with $\alpha > 0$; the support of K is $[-1, 1]$).

1.2.3 Integrated squared risk of kernel estimators

In Section 1.2.1 we have studied the behavior of the kernel density estimator $\hat{p}_n$ at an arbitrary fixed point x_0. It is also interesting to analyze the global risk of $\hat{p}_n$. An important global criterion is the *mean integrated squared error* (MISE):

$$\mathrm{MISE} \stackrel{\triangle}{=} \mathbf{E}_p \int (\hat{p}_n(x) - p(x))^2 dx.$$

By the Tonelli–Fubini theorem and by (1.4), we have

$$\mathrm{MISE} = \int \mathrm{MSE}(x) dx = \int b^2(x) dx + \int \sigma^2(x) dx. \tag{1.14}$$

Thus, the MISE is represented as a sum of the *bias term* $\int b^2(x)dx$ and the *variance term* $\int \sigma^2(x)dx$. To obtain bounds on these terms, we proceed in the same manner as for the analogous terms of the MSE (cf. Section 1.2.1). Let us study first the variance term.

Proposition 1.4 *Suppose that $K : \mathbf{R} \to \mathbf{R}$ is a function satisfying*

$$\int K^2(u) du < \infty.$$

Then for any $h > 0$, $n \geq 1$ and any probability density p we have

$$\int \sigma^2(x) dx \leq \frac{1}{nh} \int K^2(u) du.$$

Proof. As in the proof of Proposition 1.1 we obtain

$$\sigma^2(x) = \frac{1}{nh^2} \mathbf{E}_p[\eta_1^2(x)] \leq \frac{1}{nh^2} \mathbf{E}_p \left[K^2 \left(\frac{X_1 - x}{h} \right) \right]$$

for all $x \in \mathbf{R}$. Therefore

$$\begin{aligned} \int \sigma^2(x) dx &\leq \frac{1}{nh^2} \int \left[\int K^2 \left(\frac{z-x}{h} \right) p(z) dz \right] dx \qquad (1.15) \\ &= \frac{1}{nh^2} \int p(z) \left[\int K^2 \left(\frac{z-x}{h} \right) dx \right] dz \\ &= \frac{1}{nh} \int K^2(u) du. \end{aligned}$$

∎

The upper bound for the variance term in Proposition 1.4 does not require any condition on p: The result holds for any density. For the bias term in (1.14) the situation is different: We can only control it on a restricted subset of densities. As above, we specifically assume that p is smooth enough. Since the MISE is a risk corresponding to the $L_2(\mathbf{R})$-norm, it is natural to assume that p is smooth with respect to this norm. For example, we may assume that p belongs to a Nikol'ski class of functions defined as follows.

Definition 1.4 *Let $\beta > 0$ and $L > 0$. The* **Nikol'ski class** *$\mathcal{H}(\beta, L)$ is defined as the set of functions $f : \mathbf{R} \to \mathbf{R}$ whose derivatives $f^{(\ell)}$ of order $\ell = \lfloor \beta \rfloor$ exist and satisfy*

$$\left[\int \left(f^{(\ell)}(x+t) - f^{(\ell)}(x) \right)^2 dx \right]^{1/2} \leq L|t|^{\beta-\ell}, \quad \forall t \in \mathbf{R}. \tag{1.16}$$

Sobolev classes provide another popular way to describe smoothness in $L_2(\mathbf{R})$.

Definition 1.5 *Let $\beta \geq 1$ be an integer and $L > 0$. The* **Sobolev class** *$\mathcal{S}(\beta, L)$ is defined as the set of all $\beta - 1$ times differentiable functions $f : \mathbf{R} \to \mathbf{R}$ having absolutely continuous derivative $f^{(\beta-1)}$ and satisfying*

$$\int (f^{(\beta)}(x))^2 \, dx \leq L^2. \tag{1.17}$$

For integer β we have the inclusion $\mathcal{S}(\beta, L) \subset \mathcal{H}(\beta, L)$ that can be checked using the next lemma (cf. (1.21) below).

Lemma 1.1 (Generalized Minkowski inequality.) *For any Borel function g on $\mathbf{R} \times \mathbf{R}$, we have*

$$\int \left(\int g(u,x) \, du \right)^2 dx \leq \left[\int \left(\int g^2(u,x) \, dx \right)^{1/2} du \right]^2 .$$

A proof of this lemma is given in the Appendix (Lemma A.1).

We will now give an upper bound on the bias term $\int b^2(x)dx$ when p belongs to the class of probability densities that are smooth in the sense of Nikol'ski:

$$\mathcal{P}_{\mathcal{H}}(\beta, L) = \left\{ p \in \mathcal{H}(\beta, L) \;\middle|\; p \geq 0 \quad \text{and} \quad \int p(x)dx = 1 \right\}.$$

The bound will be a fortiori true for densities in the Sobolev class $\mathcal{S}(\beta, L)$.

Proposition 1.5 *Assume that $p \in \mathcal{P}_{\mathcal{H}}(\beta, L)$ and let K be a kernel of order $\ell = \lfloor \beta \rfloor$ satisfying*

$$\int |u|^{\beta} |K(u)| du < \infty.$$

Then, for any $h > 0$ and $n \geq 1$,

$$\int b^2(x)dx \le C_2^2 h^{2\beta},$$

where

$$C_2 = \frac{L}{\ell\,!}\int |u|^\beta |K(u)|du.$$

PROOF. Take any $x \in \mathbf{R}$, $u \in \mathbf{R}$, $h > 0$ and write the Taylor expansion

$$p(x+uh) = p(x) + p'(x)uh + \cdots + \frac{(uh)^\ell}{(\ell-1)\,!}\int_0^1 (1-\tau)^{\ell-1} p^{(\ell)}(x+\tau uh)d\tau.$$

Since the kernel K is of order $\ell = \lfloor\beta\rfloor$ we obtain

$$\begin{aligned} b(x) &= \int K(u)\frac{(uh)^\ell}{(\ell-1)\,!}\left[\int_0^1 (1-\tau)^{\ell-1} p^{(\ell)}(x+\tau uh)d\tau\right]du \qquad (1.18)\\ &= \int K(u)\frac{(uh)^\ell}{(\ell-1)!}\left[\int_0^1 (1-\tau)^{\ell-1}(p^{(\ell)}(x+\tau uh) - p^{(\ell)}(x))d\tau\right]du. \end{aligned}$$

Applying twice the generalized Minkowski inequality and using the fact that p belongs to the class $\mathcal{H}(\beta, L)$, we get the following upper bound for the bias term:

$$\begin{aligned} \int b^2(x)dx &\le \int\left(\int |K(u)|\frac{|uh|^\ell}{(\ell-1)\,!}\times \right. \qquad (1.19)\\ &\qquad \left.\int_0^1 (1-\tau)^{\ell-1}\left|p^{(\ell)}(x+\tau uh) - p^{(\ell)}(x)\right|d\tau du\right)^2 dx\\ &\le \left(\int |K(u)|\frac{|uh|^\ell}{(\ell-1)\,!}\times\right.\\ &\qquad \left.\left[\int\left(\int_0^1 (1-\tau)^{\ell-1}\left|p^{(\ell)}(x+\tau uh) - p^{(\ell)}(x)\right|d\tau\right)^2 dx\right]^{1/2} du\right)^2\\ &\le \left(\int |K(u)|\frac{|uh|^\ell}{(\ell-1)\,!}\times\right.\\ &\qquad \left.\left[\int_0^1 (1-\tau)^{\ell-1}\left[\int\left(p^{(\ell)}(x+\tau uh) - p^{(\ell)}(x)\right)^2 dx\right]^{1/2} d\tau\right] du\right)^2\\ &\le \left(\int |K(u)|\frac{|uh|^\ell}{(\ell-1)\,!}\left[\int_0^1 (1-\tau)^{\ell-1} L|uh|^{\beta-\ell}d\tau\right]du\right)^2\\ &= C_2^2 h^{2\beta}. \end{aligned}$$

∎

Under the assumptions of Propositions 1.4 and 1.5 we obtain

$$\text{MISE} \le C_2^2 h^{2\beta} + \frac{1}{nh}\int K^2(u)du,$$

and the minimizer $h = h_n^*$ of the right hand side is

$$h_n^* = \left(\frac{\int K^2(u)du}{2\beta C_2^2} \right)^{\frac{1}{2\beta+1}} n^{-\frac{1}{2\beta+1}}.$$

Taking $h = h_n^*$ we get

$$\text{MISE} = O\left(n^{-\frac{2\beta}{2\beta+1}}\right), \quad n \to \infty.$$

We see that the behavior of the MISE is analogous to that of the mean squared risk at a fixed point (MSE), cf. Section 1.2.1. We can summarize the above argument in the following way.

Theorem 1.2 *Suppose that the assumptions of Propositions 1.4 and 1.5 hold. Fix $\alpha > 0$ and take $h = \alpha n^{-\frac{1}{2\beta+1}}$. Then for any $n \geq 1$ the kernel estimator $\hat{p}_n$ satisfies*

$$\sup_{p \in \mathcal{P}_{\mathcal{H}}(\beta, L)} \mathbf{E}_p \int \left(\hat{p}_n(x) - p(x)\right)^2 dx \leq C n^{-\frac{2\beta}{2\beta+1}},$$

where $C > 0$ is a constant depending only on β, L, α and on the kernel K.

For densities in the Sobolev classes we get the following bound on the mean integrated squared risk.

Theorem 1.3 *Suppose that, for an integer $\beta \geq 1$:*

(i) the function K is a kernel of order $\beta - 1$ satisfying the conditions

$$\int K^2(u)du < \infty, \qquad \int |u|^\beta |K(u)|du < \infty;$$

(ii) the density p is $\beta - 1$ times differentiable, its derivative $p^{(\beta-1)}$ is absolutely continuous on $\mathbf{R}$ and

$$\int (p^{(\beta)}(x))^2 dx < \infty.$$

Then for all $n \geq 1$ and all $h > 0$ the mean integrated squared error of the kernel estimator $\hat{p}_n$ satisfies

$$\begin{aligned} \text{MISE} &\equiv \mathbf{E}_p \int (\hat{p}_n(x) - p(x))^2 dx \\ &\leq \frac{1}{nh} \int K^2(u)du + \frac{h^{2\beta}}{(\ell!)^2} \left(\int |u|^\beta |K(u)|du \right)^2 \int (p^{(\beta)}(x))^2 dx. \quad (1.20) \end{aligned}$$

Proof. We use (1.14) where we bound the variance term as in Proposition 1.4. For the bias term we apply (1.19) with $\ell = \lfloor \beta \rfloor = \beta - 1$, but we replace there L by $\left(\int (p^{(\beta)}(x))^2 dx\right)^{1/2}$ taking into account that, for all $t \in \mathbf{R}$,

$$\int \Big(p^{(\ell)}(x+t) - p^{(\ell)}(x)\Big)^2 dx \tag{1.21}$$
$$= \int \Big(t \int_0^1 p^{(\ell+1)}(x+\theta t) d\theta\Big)^2 dx$$
$$\le t^2 \Big(\int_0^1 \Big[\int \Big(p^{(\ell+1)}(x+\theta t)\Big)^2 dx\Big]^{1/2} d\theta\Big)^2$$
$$= t^2 \int (p^{(\beta)}(x))^2 dx$$

in view of the generalized Minkowski inequality. ∎

1.2.4 Lack of asymptotic optimality for fixed density

How to choose the kernel K and the bandwidth h for the kernel density estimators in an optimal way? An old and still popular approach is based on minimization in K and h of the asymptotic MISE *for fixed density* p. However, this does not lead to a consistent concept of optimality, as we are going to explain now. Other methods for choosing h are discussed in Section 1.4.

The following result on asymptotics for fixed p or its versions are often considered.

Proposition 1.6 *Assume that:*

(i) the function K is a kernel of order 1 satisfying the conditions

$$\int K^2(u)du < \infty, \qquad \int u^2|K(u)|du < \infty, \qquad S_K \stackrel{\triangle}{=} \int u^2 K(u)du \neq 0;$$

(ii) the density p is differentiable on $\mathbf{R}$, the first derivative p' is absolutely continuous on $\mathbf{R}$ and the second derivative satisfies

$$\int (p''(x))^2 dx < \infty.$$

Then for all $n \ge 1$ the mean integrated squared error of the kernel estimator $\hat{p}_n$ satisfies

$$\mathrm{MISE} \equiv \mathbf{E}_p \int (\hat{p}_n(x) - p(x))^2 dx$$
$$= \left[\frac{1}{nh}\int K^2(u)du + \frac{h^4}{4}\, S_K^2 \int (p''(x))^2 dx\right](1 + o(1)), \tag{1.22}$$

where the term $o(1)$ is independent of n (but depends on p) and tends to 0 as $h \to 0$.

A proof of this proposition is given in the Appendix (Proposition A.1).

The main term of the MISE in (1.22) is

$$\frac{1}{nh}\int K^2(u)du + \frac{h^4}{4}\, S_K^2 \int (p''(x))^2 dx. \tag{1.23}$$

Note that if K is a nonnegative kernel, expression (1.23) coincides with the nonasymptotic upper bound for the MISE which holds for all n and h (cf. Theorem 1.3 with $\beta = 2$).

The approach to optimality that we are going to criticize here starts from the expression (1.23). This expression is then minimized in h and in nonnegative kernels K, which yields the "optimal" bandwidth for given K:

$$h^{MISE}(K) = \left(\frac{\int K^2}{nS_K^2 \int (p'')^2}\right)^{1/5} \tag{1.24}$$

and the "optimal" nonnegative kernel:

$$K^*(u) = \frac{3}{4}(1-u^2)_+ \tag{1.25}$$

(the Epanechnikov kernel; cf. bibliographic notes in Section 1.11). In particular,

$$h^{MISE}(K^*) = \left(\frac{15}{n\int (p'')^2}\right)^{1/5}. \tag{1.26}$$

Note that the choices of h as in (1.24), (1.26) are not feasible since they depend on the second derivative of the unknown density p. Thus, the basic formula (1.2) with kernel $K = K^*$ and bandwidth $h = h^{MISE}(K^*)$ as in (1.26) does not define a valid estimator, but rather a random variable that can be qualified as a pseudo-estimator or *oracle* (for a more detailed discussion of oracles see Section 1.8 below). Denote this random variable by $p_n^E(x)$ and call it the *Epanechnikov oracle*. Proposition 1.6 implies that

$$\lim_{n\to\infty} n^{4/5}\mathbf{E}_p \int (p_n^E(x) - p(x))^2 dx = \frac{3^{4/5}}{5^{1/5}4}\left(\int (p''(x))^2 dx\right)^{1/5}. \tag{1.27}$$

This argument is often exhibited as a benchmark for the optimal choice of kernel K and bandwidth h, whereas (1.27) is claimed to be the best achievable MISE. The Epanechnikov oracle is declared optimal and its feasible analogs (for which the integral $\int (p'')^2$ in (1.26) is estimated from the data) are put forward. We now explain why such an approach to optimality is misleading. The following proposition is sufficiently eloquent.

Proposition 1.7 *Let assumption (ii) of Proposition 1.6 be satisfied and let K be a kernel of order 2 (thus, $S_K = 0$), such that*

$$\int K^2(u)du < \infty.$$

Then for any $\varepsilon > 0$ *the kernel estimator* $\hat{p}_n$ *with bandwidth*

$$h = n^{-1/5}\varepsilon^{-1} \int K^2(u)du$$

satisfies

$$\limsup_{n\to\infty} n^{4/5}\mathbf{E}_p \int (\hat{p}_n(x) - p(x))^2 dx \le \varepsilon. \tag{1.28}$$

The same is true for the positive part estimator $\hat{p}_n^+ = \max(0, \hat{p}_n)$*:*

$$\limsup_{n\to\infty} n^{4/5}\mathbf{E}_p \int (\hat{p}_n^+(x) - p(x))^2 dx \le \varepsilon. \tag{1.29}$$

A proof of this proposition is given in the Appendix (Proposition A.2).

We see that for all $\varepsilon > 0$ small enough the estimators $\hat{p}_n$ and $\hat{p}_n^+$ of Proposition 1.7 have smaller asymptotic MISE than the Epanechnikov oracle, *under the same assumptions on p*. Note that $\hat{p}_n$, $\hat{p}_n^+$ are true estimators, not oracles. So, if the performance of estimators is measured by their asymptotic MISE *for fixed p* there is a multitude of estimators that are strictly better than the Epanechnikov oracle. Furthermore, Proposition 1.7 implies:

$$\inf_{T_n} \limsup_{n\to\infty} n^{4/5}\mathbf{E}_p \int (T_n(x) - p(x))^2 dx = 0, \tag{1.30}$$

where $\inf_{T_n}$ is the infimum over all the kernel estimators or over all the positive part kernel estimators.

The positive part estimator $\hat{p}_n^+$ is included in Proposition 1.7 on purpose. In fact, it is often argued that one should use nonnegative kernels because the density itself is nonnegative. This would support the "optimality" of the Epanechnikov kernel because it is obtained from minimization of the asymptotic MISE over nonnegative kernels. Note, however, that non-negativity of density estimators is not necessarily achieved via non-negativity of kernels. Proposition 1.7 presents an estimator $\hat{p}_n^+$ which is nonnegative, asymptotically equivalent to the kernel estimator $\hat{p}_n$, and has smaller asymptotic MISE than the Epanechnikov oracle.

Proposition 1.7 plays the role of counterexample. The estimators $\hat{p}_n$ and $\hat{p}_n^+$ of Proposition 1.7 are by no means advocated as being good. They can be rather counterintuitive. Indeed, their bandwidth h contains an arbitrarily large constant factor ε^{-1}. This factor serves to diminish the variance term, whereas, for fixed density p, the condition $\int u^2 K(u)du = 0$ eliminates the main bias term if n is large enough, that is, if $n \ge n_0$, starting from some n_0 that depends on p. This elimination of the bias is possible for fixed p but not uniformly over p in the Sobolev class of smoothness $\beta = 2$. The message of

Proposition 1.7 is that even such counterintuitive estimators outperform the Epanechnikov oracle as soon as the asymptotics of the MISE *for fixed* p is taken as a criterion.

To summarize, the approach based on fixed p asymptotics does not lead to a consistent concept of optimality. In particular, saying that "the choice of h and K as in (1.24) – (1.26) is optimal" does not make much sense.

This explains why, instead of studying the asymptotics for fixed density p, in this book we focus on the *uniform* bounds on the risk over classes of densities (Hölder, Sobolev, Nikol'ski classes). We compare the behavior of estimators in a minimax sense on these classes. This leads to a valid concept of optimality (*among all estimators*) that we develop in detail in Chapters 2 and 3.

Remarks.

(1) Sometimes asymptotics of the MSE (risk at a fixed point) *for fixed* p is used to derive "optimal" h and K, leading to expressions similar to (1.24) – (1.26). This is yet another version of the inconsistent approach to optimality. The above critical remarks remain valid when the MISE is replaced by the MSE.

(2) The result of Proposition 1.7 can be enhanced. It can be shown that, under the same assumptions on p as in Propositions 1.6 and 1.7, one can construct an estimator $\tilde{p}_n$ such that

$$\lim_{n\to\infty} n^{4/5}\mathbf{E}_p \int (\tilde{p}_n(x) - p(x))^2 dx = 0 \tag{1.31}$$

(cf. Proposition 3.3 where we prove an analogous fact for the Gaussian sequence model). Furthermore, under mild additional assumptions, for example, if the support of p is bounded, the result of Proposition 1.7 holds for the estimator $p_n^+ / \int p_n^+$, which itself is a probability density.

1.3 Fourier analysis of kernel density estimators

In Section 1.2.3 we studied the MISE of kernel density estimators under classical but restrictive assumptions. Indeed, the results were valid only for densities p whose derivatives of given order satisfy certain conditions. In this section we will show that more general and elegant results can be obtained using Fourier analysis. In particular, we will be able to analyze the MISE of kernel estimators with kernels K that do not belong to $L_1(\mathbf{R})$, such as the *sinc kernel*

$$K(u) = \begin{cases} \frac{\sin u}{\pi u}, & \text{if } u \neq 0, \\ \frac{1}{\pi}, & \text{if } u = 0, \end{cases} \tag{1.32}$$

and will see that this kernel is better than the Epanechnikov kernel, the latter being inadmissible in the sense to be defined below.

Consider, as above, the kernel estimator

$$\hat{p}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$$

but now we only suppose that K belongs to $L_2(\mathbf{R})$, which allows us to cover, for example, the sinc kernel. We also assume throughout this section that K is symmetric, i.e., $K(u) = K(-u),\ \forall\ u \in \mathbf{R}$.

We first recall some facts related to the Fourier transform. Define the Fourier transform $\mathcal{F}[g]$ of a function $g \in L_1(\mathbf{R})$ by

$$\mathcal{F}[g](\omega) \stackrel{\triangle}{=} \int_{-\infty}^{\infty} e^{\mathrm{i}t\omega} g(t)dt, \quad \omega \in \mathbf{R},$$

where $\mathrm{i} = \sqrt{-1}$. The Plancherel theorem states that

$$\int_{-\infty}^{\infty} g^2(t)dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} |\mathcal{F}[g](\omega)|^2 d\omega \tag{1.33}$$

for any $g \in L_1(\mathbf{R}) \cap L_2(\mathbf{R})$. More generally, the Fourier transform is defined in a standard way for any $g \in L_2(\mathbf{R})$ using the fact that $L_1(\mathbf{R}) \cap L_2(\mathbf{R})$ is dense in $L_2(\mathbf{R})$. With this extension, (1.33) is true for any $g \in L_2(\mathbf{R})$.

For example, if K is the sinc kernel, a version of its Fourier transform has the form $\mathcal{F}[K](\omega) = I(|\omega| \leq 1)$. The Fourier transform of $g \in L_2(\mathbf{R})$ is defined up to an arbitrary modification on a set of Lebesgue measure zero. This will not be further recalled, in particular, all equalities between Fourier transforms will be understood in the almost everywhere sense.

For any $g \in L_2(\mathbf{R})$ we have

$$\mathcal{F}[g(\cdot/h)/h](\omega) = \mathcal{F}[g](h\omega), \quad \forall\ h > 0, \tag{1.34}$$

$$\mathcal{F}[g(t - \cdot)](\omega) = e^{\mathrm{i}t\omega} \mathcal{F}[g](-\omega), \quad \forall\ t \in \mathbf{R}. \tag{1.35}$$

Define the characteristic function associated to the density p by

$$\phi(\omega) = \int_{-\infty}^{\infty} e^{\mathrm{i}t\omega} p(t)dt = \int_{-\infty}^{\infty} e^{\mathrm{i}t\omega} dF(t), \quad \omega \in \mathbf{R},$$

and consider the empirical characteristic function

$$\phi_n(\omega) = \int_{-\infty}^{\infty} e^{\mathrm{i}t\omega} dF_n(t) = \frac{1}{n} \sum_{j=1}^{n} e^{\mathrm{i}X_j\omega}, \quad \omega \in \mathbf{R}.$$

Using (1.34) and (1.35) we may write the Fourier transform of the estimator $\hat{p}_n$, with kernel $K \in L_2(\mathbf{R})$, in the form

$$\mathcal{F}[\hat{p}_n](\omega) = \sum_{j=1}^{n} e^{\mathrm{i}X_j\omega} \mathcal{F}[h^{-1}K(\cdot/h)](-\omega) = \phi_n(\omega)\mathcal{F}[K](-h\omega).$$

If K is symmetric, $\mathcal{F}[K](-h\omega) = \mathcal{F}[K](h\omega)$. Therefore, writing for brevity

$$\widehat{K}(\omega) = \mathcal{F}[K](\omega),$$

for any symmetric kernel $K \in L_2(\mathbf{R})$ we get

$$\mathcal{F}[\hat{p}_n](\omega) = \phi_n(\omega)\widehat{K}(h\omega). \tag{1.36}$$

Lemma 1.2 *We have*

$$\mathbf{E}_p[\phi_n(\omega)] = \phi(\omega), \tag{1.37}$$

$$\mathbf{E}_p[\,|\phi_n(\omega)|^2] = \left(1 - \frac{1}{n}\right)|\phi(\omega)|^2 + \frac{1}{n}\,, \tag{1.38}$$

$$\mathbf{E}_p[\,|\phi_n(\omega) - \phi(\omega)|^2] = \frac{1}{n}\left(1 - |\phi(\omega)|^2\right). \tag{1.39}$$

Proof. Relation (1.37) is obvious, whereas (1.39) follows immediately from (1.37) and (1.38). To show (1.38), note that

$$\begin{aligned}
\mathbf{E}_p[\,|\phi_n(\omega)|^2] &= \mathbf{E}_p[\phi_n(\omega)\phi_n(-\omega)] \\
&= \mathbf{E}_p\Big[\frac{1}{n^2}\sum_{j,k\,:k\neq j} e^{\mathrm{i}(X_k - X_j)\omega}\Big] + \frac{1}{n} \\
&= \frac{n-1}{n}\phi(\omega)\phi(-\omega) + \frac{1}{n}\,.
\end{aligned}$$

∎

Assume now that both the kernel K and the density p belong to $L_2(\mathbf{R})$ and that K is symmetric. Using the Plancherel theorem and (1.36) we may write the MISE of kernel estimator $\hat{p}_n$ in the form

$$\begin{aligned}
\mathrm{MISE} &= \mathbf{E}_p \int (\hat{p}_n(x) - p(x))^2 dx \qquad (1.40)\\
&= \frac{1}{2\pi}\mathbf{E}_p \int \big|\mathcal{F}[\hat{p}_n](\omega) - \phi(\omega)\big|^2 d\omega \\
&= \frac{1}{2\pi}\mathbf{E}_p \int \big|\phi_n(\omega)\widehat{K}(h\omega) - \phi(\omega)\big|^2 d\omega.
\end{aligned}$$

The following theorem gives, under mild conditions, the exact MISE of $\hat{p}_n$ for any fixed n.

Theorem 1.4 *Let $p \in L_2(\mathbf{R})$ be a probability density, and let $K \in L_2(\mathbf{R})$ be symmetric. Then for all $n \geq 1$ and $h > 0$ the mean integrated squared error of the kernel estimator $\hat{p}_n$ has the form*

$$\begin{aligned}
\mathrm{MISE} &= \frac{1}{2\pi}\left[\int \big|1 - \widehat{K}(h\omega)\big|^2\big|\phi(\omega)\big|^2 d\omega + \frac{1}{n}\int \big|\widehat{K}(h\omega)\big|^2 d\omega\right] \qquad (1.41)\\
&\quad - \frac{1}{2\pi n}\int \big|\phi(\omega)\big|^2\big|\widehat{K}(h\omega)\big|^2 d\omega \\
&\triangleq J_n(K, h, \phi).
\end{aligned}$$

PROOF. Since $\phi \in L_2(\mathbf{R})$, $K \in L_2(\mathbf{R})$, and $|\phi(\omega)| \leq 1$ for all $\omega \in \mathbf{R}$, all the integrals in (1.41) are finite. To obtain (1.41) it suffices to develop the expression in the last line of (1.40):

$$\begin{aligned}
&\mathbf{E}_p \int \left|\phi_n(\omega)\widehat{K}(h\omega) - \phi(\omega)\right|^2 d\omega \\
&\quad = \mathbf{E}_p \int \left|(\phi_n(\omega) - \phi(\omega))\widehat{K}(h\omega) - (1 - \widehat{K}(h\omega))\phi(\omega)\right|^2 d\omega \\
&\quad = \int \left[\mathbf{E}_p\left[\ \left|\phi_n(\omega) - \phi(\omega)\right|^2\right]\left|\widehat{K}(h\omega)\right|^2 + \left|1 - \widehat{K}(h\omega)\right|^2\left|\phi(\omega)\right|^2\right] d\omega \\
&\quad = \int \left|1 - \widehat{K}(h\omega)\right|^2\left|\phi(\omega)\right|^2 d\omega + \frac{1}{n}\int \left(1 - \left|\phi(\omega)\right|^2\right)\left|\widehat{K}(h\omega)\right|^2 d\omega,
\end{aligned}$$

where we used (1.37) and (1.39). ∎

REMARKS.

(1) In Theorem 1.4 we assumed that the kernel K is symmetric, so its Fourier transform $\widehat{K}$ is real-valued.

(2) The expression in square brackets in (1.41) constitutes the main term of the MISE. It is similar to the expression obtained in Theorem 1.3 where we did not use Fourier analysis. In fact, by Plancherel's theorem and (1.34),

$$\frac{1}{2\pi n}\int \left|\widehat{K}(h\omega)\right|^2 d\omega = \frac{1}{nh}\int K^2(u)du, \tag{1.42}$$

which coincides with the upper bound on the variance term of the risk derived in Section 1.2.3. Note that the expression (1.41) based on Fourier analysis is somewhat more accurate because it contains a negative correction term

$$-\frac{1}{2\pi n}\int \left|\phi(\omega)\right|^2\left|\widehat{K}(h\omega)\right|^2 d\omega.$$

However, this term is typically of smaller order than (1.42). In fact, if $\widehat{K} \in L_\infty(\mathbf{R})$,

$$\begin{aligned}
\frac{1}{2\pi n}\int \left|\phi(\omega)\right|^2\left|\widehat{K}(h\omega)\right|^2 d\omega &\leq \frac{\|\widehat{K}\|_\infty^2}{2\pi n}\int \left|\phi(\omega)\right|^2 d\omega \\
&= \frac{\|\widehat{K}\|_\infty^2}{n}\int p^2(u)du
\end{aligned}$$

by Plancherel's theorem, where $\|\widehat{K}\|_\infty$ is the $L_\infty(\mathbf{R})$-norm of $\widehat{K}$. Thus, the correction term is of order $O(1/n)$, whereas the expression (1.42) is $O(1/(nh))$. So, for small h, the variance term is essentially given by (1.42) which is the same as the upper bound in Theorem 1.3. However, the bias term in (1.41) is different:

$$\frac{1}{2\pi}\int \left|1-\widehat{K}(h\omega)\right|^2\left|\phi(\omega)\right|^2 d\omega.$$

In contrast to Theorem 1.3, the bias term has this general form; it does not necessarily reduce to an expression involving a derivative of p.

(3) There is no condition $\int K = 1$ in Theorem 1.4; even more, K is not necessarily integrable. In addition, Theorem 1.4 applies to integrable K such that $\int K \neq 1$. This enlarges the class of possible kernels and, in principle, may lead to estimators with smaller MISE. We will see, however, that considering kernels with $\int K \neq 1$ makes no sense.

It is easy to see that a minimizer of the MISE (1.41) with respect to $\widehat{K}$ is given by the formula

$$\widehat{K}^*(h\omega) = \frac{\left|\phi(\omega)\right|^2}{\varepsilon^2(\omega)+\left|\phi(\omega)\right|^2}, \tag{1.43}$$

where $\varepsilon^2(\omega) = \big(1-|\phi(\omega)|^2\big)/n$. This is obtained by minimization of the expression under the integral in (1.41) for any fixed ω. Note that $\widehat{K}^*(0) = 1$, $0 \le \widehat{K}^*(\omega) \le 1$ for all $\omega \in \mathbf{R}$, and $\widehat{K}^* \in L_1(\mathbf{R}) \cap L_2(\mathbf{R})$. Clearly, $\widehat{K}^*$ cannot be used to construct estimators since it depends on the unknown characteristic function ϕ. The inverse Fourier transform of $\widehat{K}^*(h\omega)$ is an ideal (oracle) kernel that can be only regarded as a benchmark. Note that the right hand side of (1.43) does not depend on h, which implies that, to satisfy (1.43), the function $\widehat{K}^*(\cdot)$ itself should depend on h. Thus, the oracle does not correspond to a kernel estimator. The oracle risk (i.e., the MISE for $\widehat{K} = \widehat{K}^*$) is

$$\mathrm{MISE}^* = \frac{1}{2\pi}\int \frac{\varepsilon^2(\omega)|\phi(\omega)|^2}{\varepsilon^2(\omega)+|\phi(\omega)|^2} d\omega. \tag{1.44}$$

Theorem 1.4 allows us to compare the mean integrated squared risks $J_n(K,h,\phi)$ of different kernel estimators $\hat{p}_n$ nonasymptotically, for any fixed n. In particular, we can eliminate "bad" kernels using the following criterion.

Definition 1.6 *A symmetric kernel $K \in L_2(\mathbf{R})$ is called* **inadmissible** *if there exists another symmetric kernel $K_0 \in L_2(\mathbf{R})$ such that the following two conditions hold:*

(i) for all characteristic functions $\phi \in L_2(\mathbf{R})$

$$J_n(K_0,h,\phi) \le J_n(K,h,\phi), \quad \forall\ h>0,\ n\ge 1; \tag{1.45}$$

(ii) there exists a characteristic function $\phi_0 \in L_2(\mathbf{R})$ such that

$$J_n(K_0,h,\phi_0) < J_n(K,h,\phi_0), \quad \forall\ h>0,\ n\ge 1. \tag{1.46}$$

Otherwise, the kernel K is called **admissible**.

The problem of finding an admissible kernel is rather complex, and we will not discuss it here. We will only give a simple criterion allowing one to detect inadmissible kernels.

Proposition 1.8 *Let $K \in L_2(\mathbf{R})$ be symmetric. If*

$$\mathrm{Leb}(\,\omega :\, \widehat{K}(\omega) \notin [0,1]\,) > 0, \tag{1.47}$$

then K is inadmissible.

Proof. Denote by $\widehat{K}_0(\omega)$ the projection of $\widehat{K}(\omega)$ onto $[0,1]$, i.e., $\widehat{K}_0(\omega) = \min(1, \max(\widehat{K}(\omega), 0))$. Clearly,

$$\left|\widehat{K}_0(\omega)\right| \le \left|\widehat{K}(\omega)\right|, \quad \left|1 - \widehat{K}_0(\omega)\right| \le \left|1 - \widehat{K}(\omega)\right|, \qquad \forall\, \omega \in \mathbf{R}. \tag{1.48}$$

Since $\widehat{K} \in L_2(\mathbf{R})$, we get that $\widehat{K}_0 \in L_2(\mathbf{R})$. Therefore, there exists a function $K_0 \in L_2(\mathbf{R})$ with the Fourier transform $\widehat{K}_0$. Since K is symmetric, the Fourier transforms $\widehat{K}$ and $\widehat{K}_0$ are real-valued, so that K_0 is also symmetric.

Using (1.48) and the fact that $\left|\phi(\omega)\right| \le 1$ for any characteristic function ϕ, we get

$$\begin{aligned} &J_n(K,h,\phi) - J_n(K_0,h,\phi) \\ &\quad = \frac{1}{2\pi}\left[\int \left(\left|1 - \widehat{K}(h\omega)\right|^2 - \left|1 - \widehat{K}_0(h\omega)\right|^2\right)\left|\phi(\omega)\right|^2 d\omega \right. \\ &\qquad \left. + \frac{1}{n}\int (1 - \left|\phi(\omega)\right|^2)\left(\left|\widehat{K}(h\omega)\right|^2 - \left|\widehat{K}_0(h\omega)\right|^2\right) d\omega\right] \\ &\quad \ge 0. \end{aligned} \tag{1.49}$$

This proves (1.45). To check part (ii) of Definition 1.6 we use assumption (1.47). Let $\phi_0(\omega) = e^{-\omega^2/2}$ be the characteristic function of the standard normal distribution on $\mathbf{R}$. Since assumption (1.47) holds, at least one of the conditions $\mathrm{Leb}(\omega : \widehat{K}(\omega) < 0) > 0$ or $\mathrm{Leb}(\omega : \widehat{K}(\omega) > 1) > 0$ is satisfied.

Assume first that $\mathrm{Leb}(\omega : \widehat{K}(\omega) < 0) > 0$. Fix $h > 0$ and introduce the set $B_h^0 \stackrel{\triangle}{=} \{\omega : \widehat{K}(h\omega) < 0\} = \{\omega/h : \widehat{K}(\omega) < 0\}$. Note that $\mathrm{Leb}(B_h^0) > 0$. Indeed, B_h^0 is a dilation of the set $\{\omega : \widehat{K}(\omega) < 0\}$ of a positive Lebesgue measure. Then

$$\begin{aligned} &J_n(K,h,\phi_0) - J_n(K_0,h,\phi_0) \\ &\quad \ge \frac{1}{2\pi n}\int_{B_h^0} (1 - \left|\phi_0(\omega)\right|^2)\left(\left|\widehat{K}(h\omega)\right|^2 - \left|\widehat{K}_0(h\omega)\right|^2\right) d\omega \\ &\quad = \frac{1}{2\pi n}\int_{B_h^0} (1 - e^{-\omega^2})\left|\widehat{K}(h\omega)\right|^2 d\omega > 0 \end{aligned} \tag{1.50}$$

where the last inequality is due to the fact that $(1 - e^{-\omega^2})\left|\widehat{K}(h\omega)\right|^2 > 0$ almost everywhere on B_h^0.

Finally, if $\mathrm{Leb}(\omega : \widehat{K}(\omega) > 1) > 0$, we define $B_h^1 \stackrel{\triangle}{=} \{\omega : \widehat{K}(h\omega) > 1\}$ and reasoning in a similar way as above we obtain

$$\begin{aligned} & J_n(K,h,\phi_0) - J_n(K_0,h,\phi_0) \\ & \quad \geq \frac{1}{2\pi}\int_{B_h^1}\left(\left|1-\widehat{K}(h\omega)\right|^2 - \left|1-\widehat{K}_0(h\omega)\right|^2\right)\left|\phi_0(\omega)\right|^2 d\omega \\ & \quad = \frac{1}{2\pi}\int_{B_h^1}\left|1-\widehat{K}(h\omega)\right|^2 e^{-\omega^2} d\omega > 0. \end{aligned}$$

∎

Since the Fourier transform of an integrable function K is continuous and $\widehat{K}(0) = \int K(u)du$, Proposition 1.8 implies that any integrable symmetric kernel with $\int K(u)du > 1$ is inadmissible. This conclusion does not extend to kernels with $0 < \int K(u)du < 1$: Proposition 1.8 does not say that all of them are inadmissible. However, considering such kernels makes no sense. In fact, if $\widehat{K}(0) < 1$ and $\widehat{K}$ is continuous, there exist positive constants ε and δ such that $\inf_{|t|\leq\varepsilon}|1-\widehat{K}(t)| = \delta$. Thus, we get

$$\int\left|1-\widehat{K}(h\omega)\right|^2|\phi(\omega)|^2 d\omega \geq \delta^2\int_{|\omega|\leq\varepsilon/h}|\phi(\omega)|^2 d\omega \to \delta^2\int|\phi(\omega)|^2 d\omega > 0$$

as $h \to 0$. Therefore, the bias term in the MISE of such estimators (cf. (1.41)) does not tend to 0 as $h \to 0$.

Corollary 1.1 *The Epanechnikov kernel is inadmissible.*

Proof. The Fourier transform of the Epanechnikov kernel has the form

$$\widehat{K}(\omega) = \begin{cases} \frac{3}{\omega^3}(\sin\omega - \omega\cos\omega), & \text{if } \omega \neq 0, \\ 1, & \text{if } \omega = 0. \end{cases}$$

It is easy to see that the set $\{\omega : \widehat{K}(\omega) < 0\}$ is of positive Lebesgue measure, so that Proposition 1.8 applies. ∎

Suppose now that p belongs to a *Sobolev class of densities* defined as follows:

$$\mathcal{P}_{\mathcal{S}}(\beta,L) = \left\{p \,\middle|\, p \geq 0, \int p(x)dx = 1 \text{ and } \int|\omega|^{2\beta}\left|\phi(\omega)\right|^2 d\omega \leq 2\pi L^2\right\},$$

where $\beta > 0$ and $L > 0$ are constants and $\phi = \mathcal{F}[p]$ denotes, as before, the characteristic function associated to p. It can be shown that for integer β the class $\mathcal{P}_{\mathcal{S}}(\beta,L)$ coincides with the set of all the probability densities belonging to the Sobolev class $\mathcal{S}(\beta,L)$. Note that if β is an integer and if the derivative $p^{(\beta-1)}$ is absolutely continuous, the condition

$$\int\left(p^{(\beta)}(u)\right)^2 du \leq L^2 \tag{1.51}$$

implies

$$\int |\omega|^{2\beta} \big|\phi(\omega)\big|^2 d\omega \le 2\pi L^2. \tag{1.52}$$

Indeed, the Fourier transform of $p^{(\beta)}$ is $(-\mathrm{i}\omega)^\beta \phi(\omega)$, so that (1.52) follows from (1.51) by Plancherel's theorem. Passing to characteristic functions as in (1.52) adds flexibility; the notion of a Sobolev class is thus extended from integer β to all $\beta > 0$, i.e., to a continuous scale of smoothness.

Theorem 1.5 *Let $K \in L_2(\mathbf{R})$ be symmetric. Assume that for some $\beta > 0$ there exists a constant A such that*

$$\operatorname{ess\,sup}_{t\in\mathbf{R}\setminus\{0\}} \frac{\big|1-\widehat{K}(t)\big|}{|t|^\beta} \le A. \tag{1.53}$$

Fix $\alpha > 0$ and take $h = \alpha n^{-\frac{1}{2\beta+1}}$. Then for any $n \ge 1$ the kernel estimator $\hat{p}_n$ satisfies

$$\sup_{p\in\mathcal{P}_S(\beta,L)} \mathbf{E}_p \int \left(\hat{p}_n(x) - p(x)\right)^2 dx \le C n^{-\frac{2\beta}{2\beta+1}}$$

where $C > 0$ is a constant depending only on L, α, A and on the kernel K.

Proof. In view of (1.53) and of the definition of $\mathcal{P}_S(\beta, L)$ we have

$$\begin{aligned}\int \big|1-\widehat{K}(h\omega)\big|^2 \big|\phi(\omega)\big|^2 d\omega &\le A^2 h^{2\beta} \int |\omega|^{2\beta} \big|\phi(\omega)\big|^2 d\omega \\ &\le 2\pi A^2 L^2 h^{2\beta}.\end{aligned}$$

Plugging this into (1.41) and using (1.42) we get, for $h = \alpha n^{-\frac{1}{2\beta+1}}$,

$$\begin{aligned}\mathbf{E}_p \int \left(\hat{p}_n(x) - p(x)\right)^2 dx &\le A^2 L^2 h^{2\beta} + \frac{1}{nh} \int K^2(u) du \\ &\le C n^{-\frac{2\beta}{2\beta+1}}.\end{aligned}$$

∎

Condition (1.53) implies that there exists a version $\widehat{K}$ that is continuous at 0 and satisfies $\widehat{K}(0) = 1$. Note that $\widehat{K}(0) = 1$ can be viewed as an extension of the assumption $\int K = 1$ to nonintegrable K, such as the sinc kernel. Furthermore, under the assumptions of Theorem 1.5, condition (1.53) is equivalent to

$$\exists\ t_0, A_0 < \infty : \quad \operatorname{ess\,sup}_{0<|t|\le t_0} \frac{\big|1-\widehat{K}(t)\big|}{|t|^\beta} \le A_0. \tag{1.54}$$

So, in fact, (1.53) is a local condition on the behavior of $\widehat{K}$ in a neighborhood of 0, essentially a restriction on the moments of K. One can show that for integer β assumption (1.53) is satisfied if K is a kernel of order $\beta - 1$ and $\int |u|^\beta \big|K(u)\big| du < \infty$ (Exercise 1.6).

Note that if condition (1.53) is satisfied for some $\beta = \beta_0 > 0$, then it also holds for all $0 < \beta < \beta_0$. For all the kernels listed on p. 3, except for the Silverman kernel, condition (1.53) can be guaranteed only with $\beta \le 2$. On the other hand, the Fourier transform of the Silverman kernel is

$$\widehat{K}(\omega) = \frac{1}{1+\omega^4}\,,$$

so that we have (1.53) with $\beta = 4$.

Kernels satisfying (1.53) exist for any given $\beta > 0$. Two important examples are given by kernels with the Fourier transforms

$$\widehat{K}(\omega) = \frac{1}{1+|\omega|^\beta} \qquad \text{(spline type kernel)}, \tag{1.55}$$

$$\widehat{K}(\omega) = (1-|\omega|^\beta)_+ \quad \text{(Pinsker kernel)}. \tag{1.56}$$

It can be shown that, for $\beta = 2m$, where m is an integer, kernel estimators with $\widehat{K}$ satisfying (1.55) are close to spline estimators (cf. Exercise 1.11 that treats the case $m = 2$). The kernel (1.56) is related to Pinsker's theory discussed in Chapter 3. The inverse Fourier transforms of (1.55) and (1.56) can be written explicitly for integer β. Thus, for $\beta = 2$ the Pinsker kernel has the form

$$K(u) = \begin{cases} \frac{2}{\pi u^3}(\sin u - u\cos u), & \text{if } u \neq 0, \\ \frac{2}{3\pi}, & \text{if } u = 0. \end{cases}$$

Finally, there exist *superkernels*, or *infinite power kernels*, i.e., kernels that satisfy (1.53) simultaneously for all $\beta > 0$. An example is the sinc kernel (1.32). Note that the sinc kernel can be successfully used not only in the context of Theorem 1.5 but also for other classes of densities, such as those with exponentially decreasing characteristic functions (cf. Exercises 1.7, 1.8). Thus, the sinc kernel is more flexible than its competitors discussed above: Those are associated to some prescribed number of derivatives of a density and cannot take advantage of higher smoothness.

1.4 Unbiased risk estimation. Cross-validation density estimators

In this section we suppose that the kernel K is fixed and we are interested in choosing the bandwidth h. Write $\mathrm{MISE} = \mathrm{MISE}(h)$ to indicate that the mean integrated squared error is a function of bandwidth and define the ideal value of h by

$$h_{\mathrm{id}} = \arg\min_{h>0} \mathrm{MISE}(h). \tag{1.57}$$

Unfortunately, this value remains purely theoretical since $\mathrm{MISE}(h)$ depends on the unknown density p. The results in the previous sections do not allow

us to construct an estimator approaching this ideal value. Therefore other methods should be applied. In this context, a common idea is to use unbiased estimation of the risk. Instead of minimizing MISE(h) in (1.57), it is suggested to minimize an unbiased or approximately unbiased estimator of MISE(h).

We now describe a popular implementation of this idea given by the cross-validation. First, note that

$$\mathrm{MISE}(h) = \mathbf{E}_p \int (\hat{p}_n - p)^2 = \mathbf{E}_p \left[\int \hat{p}_n^2 - 2 \int \hat{p}_n p \right] + \int p^2.$$

Here and often in the rest of this section we will write for brevity $\int(\dots)$ instead of $\int(\dots)dx$. Since the integral $\int p^2$ does not depend on h, the minimizer h_{id} of MISE(h) as defined in (1.57) also minimizes the function

$$J(h) \triangleq \mathbf{E}_p \left[\int \hat{p}_n^2 - 2 \int \hat{p}_n p \right].$$

We now look for an unbiased estimator of $J(h)$. For this purpose it is sufficient to find an unbiased estimator for each of the quantities $\mathbf{E}_p \left[\int \hat{p}_n^2 \right]$ and $\mathbf{E}_p \left[\int \hat{p}_n p \right]$. There exists a trivial unbiased estimator $\int \hat{p}_n^2$ of the quantity $\mathbf{E}_p \left[\int \hat{p}_n^2 \right]$. Therefore it remains to find an unbiased estimator of $\mathbf{E}_p \left[\int \hat{p}_n p \right]$. Write

$$\hat{p}_{n,-i}(x) = \frac{1}{(n-1)h} \sum_{j \neq i} K\left(\frac{X_j - x}{h} \right).$$

Let us show that an unbiased estimator of $G = \mathbf{E}_p \left[\int \hat{p}_n p \right]$ is given by

$$\hat{G} = \frac{1}{n} \sum_{i=1}^{n} \hat{p}_{n,-i}(X_i).$$

Indeed, since X_i are i.i.d., we have

$$\begin{aligned}
\mathbf{E}_p(\hat{G}) &= \mathbf{E}_p \left[\hat{p}_{n,-1}(X_1) \right] \\
&= \mathbf{E}_p \left[\frac{1}{(n-1)h} \sum_{j \neq 1} \int K\left(\frac{X_j - z}{h} \right) p(z)\, dz \right] \\
&= \frac{1}{h} \int p(x) \int K\left(\frac{x - z}{h} \right) p(z)\, dz\, dx
\end{aligned}$$

provided that the last expression is finite. On the other hand,

$$\begin{aligned}
G &= \mathbf{E}_p \left[\int \hat{p}_n p \right] \\
&= \mathbf{E}_p \left[\frac{1}{nh} \sum_{i=1}^{n} \int K\left(\frac{X_i - z}{h} \right) p(z)\, dz \right] \\
&= \frac{1}{h} \int p(x) \int K\left(\frac{x - z}{h} \right) p(z)\, dz\, dx,
\end{aligned}$$

implying that $G = \mathbf{E}_p(\hat{G})$.

Summarizing our argument, an unbiased estimator of $J(h)$ can be written as follows:

$$CV(h) = \int \hat{p}_n^2 - \frac{2}{n} \sum_{i=1}^{n} \hat{p}_{n,-i}(X_i)$$

where CV stands for "cross-validation." The function $CV(\cdot)$ is called the *leave-one-out cross-validation criterion* or simply the *cross-validation criterion*. Thus we have proved the following result.

Proposition 1.9 *Assume that for a function $K : \mathbf{R} \to \mathbf{R}$, for a probability density p satisfying $\int p^2 < \infty$ and $h > 0$ we have*

$$\int \int p(x) \left| K\left(\frac{x-z}{h}\right) \right| p(z)\, dz\, dx < \infty.$$

Then

$$\mathbf{E}_p[CV(h)] = \mathrm{MISE}(h) - \int p^2.$$

Thus, $CV(h)$ yields an unbiased estimator of $\mathrm{MISE}(h)$, up to a shift $\int p^2$ which is independent of h. This means that the functions $h \mapsto \mathrm{MISE}(h)$ and $h \mapsto \mathbf{E}_p[CV(h)]$ have the same minimizers. In turn, the minimizers of $\mathbf{E}_p[CV(h)]$ can be approximated by those of the function $CV(\cdot)$ which can be computed from the observations $X_1, \ldots, X_n$:

$$h_{CV} = \arg\min_{h>0} CV(h)$$

whenever the minimum is attained (cf. Figure 1.3). Finally, we define the *cross-validation estimator* $\hat{p}_{n,CV}$ of the density p in the following way:

$$\hat{p}_{n,CV}(x) = \frac{1}{nh_{CV}} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h_{CV}}\right).$$

This is a kernel estimator with random bandwidth h_{CV} depending on the sample $X_1, \ldots, X_n$. It can be proved that under appropriate conditions the integrated squared error of the estimator $\hat{p}_{n,CV}$ is asymptotically equivalent to that of the ideal kernel pseudo-estimator (oracle) which has the bandwidth h_{id} defined in (1.57). Similar results for another estimation problem are discussed in Chapter 3.

Cross-validation is not the only way to construct unbiased risk estimators. Other methods exist: for example, we can do this using the Fourier analysis of density estimators, in particular, formula (1.41). Let K be a symmetric kernel such that its (real-valued) Fourier transform $\widehat{K}$ belongs to $L_1(\mathbf{R}) \cap L_2(\mathbf{R})$. Consider the function $\tilde{J}(\cdot)$ defined by

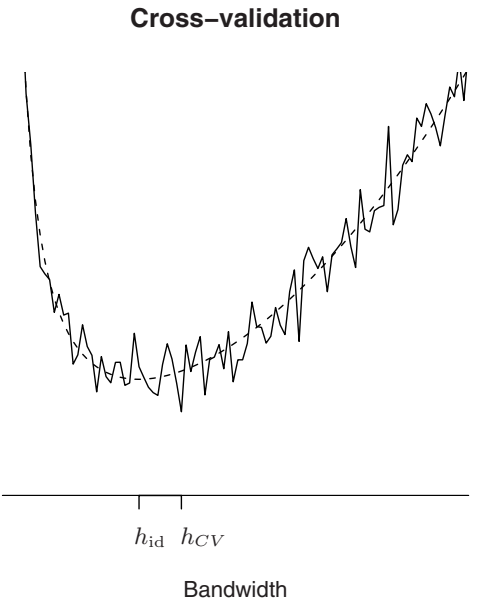

Figure 1.3. The functions $CV(h)$ (solid line), $\mathrm{MISE}(h) - \int p^2$ (dashed line) and their minimizers h_{CV}, $h_{\rm id}$.

$$\begin{aligned}\tilde{J}(h) &\triangleq \int \left(-2\widehat{K}(h\omega) + \widehat{K}^2(h\omega)\Big(1-\frac{1}{n}\Big)\right)\left|\phi_n(\omega)\right|^2 d\omega \\ &\quad + \frac{2}{n}\int \widehat{K}(h\omega)d\omega \\ &= \int \left(-2\widehat{K}(h\omega) + \widehat{K}^2(h\omega)\Big(1-\frac{1}{n}\Big)\right)\left|\phi_n(\omega)\right|^2 d\omega + \frac{4\pi K(0)}{nh}\,,\end{aligned} \tag{1.58}$$

where ϕ_n is the empirical characteristic function and we have used that, by the inverse Fourier transform, $\int \widehat{K}(\omega)d\omega = 2\pi K(0)$. From (1.38) and Theorem 1.4 we get

$$\begin{aligned}\mathbf{E}_p(\hat{J}(h)) &= \int \left(-2\widehat{K}(h\omega) + \widehat{K}^2(h\omega)\Big(1-\frac{1}{n}\Big)\right)\Big(1-\frac{1}{n}\Big)\left|\phi(\omega)\right|^2 d\omega \\ &\quad + \frac{1}{n}\Big(1-\frac{1}{n}\Big)\int \widehat{K}^2(h\omega)d\omega \\ &= \Big(1-\frac{1}{n}\Big)\left[\int \Big(1-\widehat{K}(h\omega)\Big)^2\left|\phi(\omega)\right|^2 d\omega - \int \left|\phi(\omega)\right|^2 d\omega\right. \\ &\quad \left. + \frac{1}{n}\int (1-\left|\phi(\omega)\right|^2)\widehat{K}^2(h\omega)d\omega\right]\end{aligned} \tag{1.59}$$

$$= 2\pi\Big(1 - \frac{1}{n}\Big)\left[\ \mathrm{MISE}(h) - \int p^2\ \right].$$

Therefore, the functions $h \mapsto \mathbf{E}_p(\tilde{J}(h))$ and $h \mapsto \mathrm{MISE}(h)$ have the same minimizers. In the same spirit as above we now approximate the unknown minimizers of $\mathrm{MISE}(\cdot)$ by

$$\tilde{h} = \arg\min_{h>0} \tilde{J}(h).$$

This is a data-driven bandwidth obtained from an unbiased risk estimation but different from the cross-validation bandwidth h_{CV}. The corresponding density estimator is given by

$$\tilde{p}_n(x) = \frac{1}{n\tilde{h}} \sum_{i=1}^{n} K\left(\frac{X_i - x}{\tilde{h}}\right).$$

It can be proved that, under appropriate conditions, the estimator $\tilde{p}_n$ behaves itself analogously to $\hat{p}_{n,CV}$: the MISE of $\tilde{p}_n$ is asymptotically equivalent to that of the ideal kernel pseudo-estimator (oracle) that has the bandwidth h_{id} defined in (1.57). The proof of this property is beyond the scope of the book but similar results for another estimation problem are discussed in Chapter 3.

1.5 Nonparametric regression. The Nadaraya–Watson estimator

The following two basic models are usually considered in nonparametric regression.

1. Nonparametric regression with random design

Let (X, Y) be a pair of real-valued random variables such that $\mathbf{E}|Y| < \infty$. The function $f : \mathbf{R} \to \mathbf{R}$ defined by

$$f(x) = \mathbf{E}(Y|X = x)$$

is called the regression function of Y on X. Suppose that we have a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of n i.i.d. pairs of random variables having the same distribution as (X, Y). We would like to estimate the function f from the data $(X_1, Y_1), \ldots, (X_n, Y_n)$. The nonparametric approach only assumes that $f \in \mathcal{F}$, where $\mathcal{F}$ is a given nonparametric class. The set of values $\{X_1, \ldots, X_n\}$ is called the *design*. Here the design is random.

The conditional residual $\xi \stackrel{\triangle}{=} Y - \mathbf{E}(Y|X)$ has mean zero, $\mathbf{E}(\xi) = 0$, and we may write

$$Y_i = f(X_i) + \xi_i, \quad i = 1, \ldots, n, \tag{1.60}$$

where ξ_i are i.i.d. random variables with the same distribution as ξ. In particular, $\mathbf{E}(\xi_i) = 0$. The variables ξ_i can therefore be interpreted as a "noise."

2. Nonparametric regression with fixed design

This model is also defined by (1.60) but now $X_i \in \mathbf{R}$ are fixed and deterministic instead of random and i.i.d.

Example 1.1 *Nonparametric regression model with regular design.*

Suppose that $X_i = i/n$. Assume that f is a function from $[0,1]$ to $\mathbf{R}$ and that the observations Y_i are given by

$$Y_i = f(i/n) + \xi_i, \quad i = 1, 2, \ldots, n,$$

where ξ_i are i.i.d. with mean zero ($\mathbf{E}(\xi_i) = 0$). In what follows, we will mainly focus on this model.

Given a kernel K and a bandwidth h, one can construct kernel estimators for nonparametric regression similar to those for density estimation. There exist different types of kernel estimators of the regression function f. The most celebrated one is the Nadaraya–Watson estimator defined as follows:

$$f_n^{NW}(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)}, \quad \text{if } \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \neq 0,$$

and $f_n^{NW}(x) = 0$, otherwise.

Example 1.2 *The Nadaraya–Watson estimator with rectangular kernel.*

If we choose $K(u) = \frac{1}{2}\, I(|u| \leq 1)$, then $f_n^{NW}(x)$ is the average of such Y_i that $X_i \in [x-h, x+h]$. *For fixed* n, the two extreme cases for the bandwidth are:

(i) $h \to \infty$. Then $f_n^{NW}(x)$ tends to $n^{-1}\sum_{i=1}^{n} Y_i$ which is a constant independent of x. The systematic error (bias) can be too large. This is a situation of *oversmoothing*.

(ii) $h \to 0$. Then $f_n^{NW}(X_i) = Y_i$ whenever $h < \min_{i,j} |X_i - X_j|$ and

$$\lim_{h \to 0} f_n^{NW}(x) = 0, \quad \text{if } x \neq X_i.$$

The estimator f_n^{NW} is therefore too oscillating: it reproduces the data Y_i at the points X_i and vanishes elsewhere. This makes the stochastic error (variance) too large. In other words, *undersmoothing* occurs.

An optimal bandwidth h yielding a balance between bias and variance is situated between these two extremes.

The Nadaraya–Watson estimator can be represented as a weighted sum of the Y_i:

$$f_n^{NW}(x) = \sum_{i=1}^{n} Y_i W_{ni}^{NW}(x)$$

where the weights are

$$W_{ni}^{NW}(x) = \frac{K\left(\dfrac{X_i - x}{h}\right)}{\sum\limits_{j=1}^{n} K\left(\dfrac{X_j - x}{h}\right)} \; I\left(\sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right) \neq 0\right).$$

Definition 1.7 *An estimator $\hat{f}_n(x)$ of $f(x)$ is called a* **linear nonparametric regression estimator** *if it can be written in the form*

$$\hat{f}_n(x) = \sum_{i=1}^{n} Y_i \, W_{ni}(x)$$

where the weights $W_{ni}(x) = W_{ni}(x, X_1, \ldots, X_n)$ depend only on n, i, x and the values $X_1, \ldots, X_n$.

Typically, the weights $W_{ni}(x)$ of linear regression estimators satisfy the equality

$$\sum_{i=1}^{n} W_{ni}(x) = 1$$

for all x (or for almost all x with respect to the Lebesgue measure).

An intuitive motivation of f_n^{NW} is clear. Suppose that the distribution of (X, Y) has density $p(x, y)$ with respect to the Lebesgue measure and $p(x) = \int p(x, y)dy > 0$. Then

$$f(x) = \mathbf{E}(Y|X = x) = \frac{\int yp(x, y)dy}{\int p(x, y)dy} = \frac{\int yp(x, y)dy}{p(x)}.$$

If we replace here $p(x, y)$ by the estimator $\hat{p}_n(x, y)$ of the density of (X, Y) defined by (1.3) and use the kernel estimator $\hat{p}_n(x)$ instead of $p(x)$, we obtain f_n^{NW} in view of the following result.

Proposition 1.10 *Let $\hat{p}_n(x)$ and $\hat{p}_n(x, y)$ be the kernel density estimators defined in (1.2) and (1.3), respectively, with a kernel K of order 1. Then*

$$f_n^{NW}(x) = \frac{\int y\hat{p}_n(x, y)dy}{\hat{p}_n(x)} \tag{1.61}$$

if $\hat{p}_n(x) \neq 0$.

PROOF. By (1.3), we have

$$\int y\hat{p}_n(x,y)dy = \frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\int y\, K\left(\frac{Y_i - y}{h}\right)dy.$$

Since K has order 1, we also obtain

$$\begin{aligned}\frac{1}{h}\int y\, K\left(\frac{Y_i - y}{h}\right)dy &= \int \frac{y - Y_i}{h}\, K\left(\frac{Y_i - y}{h}\right)dy + \frac{Y_i}{h}\int K\left(\frac{Y_i - y}{h}\right)dy \\ &= -h\int uK(u)du + Y_i\int K(u)du = Y_i.\end{aligned}$$

∎

If the marginal density p of X_i is known we can use $p(x)$ instead of $\hat{p}_n(x)$ in (1.61). Then we get the following estimator which is slightly different from f_n^{NW}:

$$\bar{f}_{nh}(x) = \frac{\int y\hat{p}_n(x,y)dy}{p(x)} = \frac{1}{nhp(x)}\sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h}\right).$$

In particular, if p is the density of the uniform distribution on $[0,1]$, then

$$\bar{f}_{nh}(x) = \frac{1}{nh}\sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h}\right). \tag{1.62}$$

Though the above argument concerns the regression model with random design, the estimator (1.62) is also applicable for the regular fixed design $(X_i = i/n)$.

1.6 Local polynomial estimators

If the kernel K takes only nonnegative values, the Nadaraya–Watson estimator f_n^{NW} satisfies

$$f_n^{NW}(x) = \arg\min_{\theta\in\mathbf{R}} \sum_{i=1}^{n} (Y_i - \theta)^2 K\left(\frac{X_i - x}{h}\right). \tag{1.63}$$

Thus f_n^{NW} is obtained by a local constant least squares approximation of the outputs Y_i. The locality is determined by a kernel K that downweights all the X_i that are not close to x whereas θ plays the role of a local constant to be fitted. More generally, we may define a local polynomial least squares approximation, replacing in (1.63) the constant θ by a polynomial of given degree ℓ. If $f \in \Sigma(\beta, L)$, $\beta > 1$, $\ell = \lfloor\beta\rfloor$, then for z sufficiently close to x we may write

$$f(z) \approx f(x) + f'(x)(z-x) + \cdots + \frac{f^{(\ell)}(x)}{\ell!}(z-x)^\ell = \theta^T(x) U\left(\frac{z-x}{h}\right)$$

where

$$U(u) = \left(1, u, u^2/2!, \ldots, u^\ell/\ell!\right)^T,$$
$$\theta(x) = \left(f(x), f'(x)h, f''(x)h^2, \ldots, f^{(\ell)}(x)h^\ell\right)^T.$$

We can therefore generalize (1.63) in the following way.

Definition 1.8 *Let $K : \mathbf{R} \to \mathbf{R}$ be a kernel, $h > 0$ be a bandwidth, and $\ell \geq 0$ be an integer. A vector $\hat{\theta}_n(x) \in \mathbf{R}^{\ell+1}$ defined by*

$$\hat{\theta}_n(x) = \arg \min_{\theta \in \mathbf{R}^{\ell+1}} \sum_{i=1}^n \left[Y_i - \theta^T U\left(\frac{X_i - x}{h}\right)\right]^2 K\left(\frac{X_i - x}{h}\right) \tag{1.64}$$

is called a **local polynomial estimator of order ℓ of $\theta(x)$** *or* LP(ℓ) *estimator of $\theta(x)$ for short. The statistic*

$$\hat{f}_n(x) = U^T(0)\hat{\theta}_n(x)$$

is called a **local polynomial estimator of order ℓ of $f(x)$** *or* LP(ℓ) *estimator of $f(x)$ for short.*

Note that $\hat{f}_n(x)$ is simply the first coordinate of the vector $\hat{\theta}_n(x)$. Comparing (1.64) and (1.63) we see that the Nadaraya–Watson estimator f_n^{NW} with kernel $K \geq 0$ is the LP(0) estimator. Furthermore, properly normalized coordinates of $\hat{\theta}_n(x)$ provide estimators of the derivatives $f'(x), \ldots, f^{(\ell)}(x)$ (cf. Exercise 1.4).

For a fixed x the estimator (1.64) is a weighted least squares estimator. Indeed, we can write $\hat{\theta}_n(x)$ as follows:

$$\hat{\theta}_n(x) = \arg \min_{\theta \in \mathbf{R}^{\ell+1}} (-2\theta^T \mathbf{a}_{nx} + \theta^T \mathcal{B}_{nx} \theta), \tag{1.65}$$

where the matrix $\mathcal{B}_{nx}$ and the vector $\mathbf{a}_{nx}$ are defined by the formulas

$$\mathcal{B}_{nx} = \frac{1}{nh} \sum_{i=1}^n U\left(\frac{X_i - x}{h}\right) U^T\left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x}{h}\right),$$

$$\mathbf{a}_{nx} = \frac{1}{nh} \sum_{i=1}^n Y_i U\left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x}{h}\right).$$

A necessary condition for $\hat{\theta}_n(x)$ to satisfy (1.65) is that the following system of normal equations hold:

$$\mathcal{B}_{nx} \hat{\theta}_n(x) = \mathbf{a}_{nx}. \tag{1.66}$$

If the matrix $\mathcal{B}_{nx}$ is positive definite ($\mathcal{B}_{nx} > 0$), the LP(ℓ) estimator is unique and is given by $\hat{\theta}_n(x) = \mathcal{B}_{nx}^{-1}\mathbf{a}_{nx}$ (equation (1.66) is then a necessary and sufficient condition characterizing the minimizer in (1.65)). In this case

$$\hat{f}_n(x) = \sum_{i=1}^{n} Y_i\, W_{ni}^*(x) \tag{1.67}$$

where

$$W_{ni}^*(x) = \frac{1}{nh}\, U^T(0)\mathcal{B}_{nx}^{-1} U\left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x}{h}\right)$$

proving the following result.

Proposition 1.11 *If the matrix $\mathcal{B}_{nx}$ is positive definite, the local polynomial estimator $\hat{f}_n(x)$ of $f(x)$ is a linear estimator.*

The local polynomial estimator of order ℓ has a remarkable property: It reproduces polynomials of degree $\leq \ell$. This is shown in the next proposition.

Proposition 1.12 *Let x be a real number such that $\mathcal{B}_{nx} > 0$ and let Q be a polynomial of degree $\leq \ell$. Then the* LP(ℓ) *weights W_{ni}^* are such that*

$$\sum_{i=1}^{n} Q(X_i) W_{ni}^*(x) = Q(x)$$

for any sample $(X_1, \ldots, X_n)$. In particular,

$$\sum_{i=1}^{n} W_{ni}^*(x) = 1 \quad \textit{and} \quad \sum_{i=1}^{n} (X_i - x)^k W_{ni}^*(x) = 0 \quad \textit{for} \;\; k = 1, \ldots, \ell. \tag{1.68}$$

Proof. Since Q is a polynomial of degree $\leq \ell$, we have

$$\begin{aligned} Q(X_i) &= Q(x) + Q'(x)(X_i - x) + \cdots + \frac{Q^{(\ell)}(x)}{\ell\,!}(X_i - x)^\ell \\ &= q^T(x) U\left(\frac{X_i - x}{h}\right) \end{aligned}$$

where $q(x) = (Q(x), Q'(x)h, \ldots, Q^{(\ell)}(x)h^\ell)^T \in \mathbf{R}^{\ell+1}$. Set $Y_i = Q(X_i)$. Then the LP(ℓ) estimator satisfies

$$\begin{aligned} \hat{\theta}_n(x) &= \arg\min_{\theta \in \mathbf{R}^{\ell+1}} \sum_{i=1}^{n} \left(Q(X_i) - \theta^T U\left(\frac{X_i - x}{h}\right)\right)^2 K\left(\frac{X_i - x}{h}\right) \\ &= \arg\min_{\theta \in \mathbf{R}^{\ell+1}} \sum_{i=1}^{n} \left((q(x) - \theta)^T U\left(\frac{X_i - x}{h}\right)\right)^2 K\left(\frac{X_i - x}{h}\right) \\ &= \arg\min_{\theta \in \mathbf{R}^{\ell+1}} (q(x) - \theta)^T\, \mathcal{B}_{nx} (q(x) - \theta). \end{aligned}$$

Therefore, if $\mathcal{B}_{nx} > 0$, we have $\hat{\theta}_n(x) = q(x)$ and we obtain $\hat{f}_n(x) = Q(x)$, since $\hat{f}_n(x)$ is the coordinate of $\hat{\theta}_n(x)$. The required result follows immediately by taking $Y_i = Q(X_i)$ in (1.67). ∎

1.6.1 Pointwise and integrated risk of local polynomial estimators

In this section we study statistical properties of the LP(ℓ) estimator constructed from observations (X_i, Y_i), $i = 1, \ldots, n$, such that

$$Y_i = f(X_i) + \xi_i, \qquad i = 1, \ldots, n, \tag{1.69}$$

where ξ_i are independent zero mean random variables ($\mathbf{E}(\xi_i) = 0$), the X_i are deterministic values belonging to $[0, 1]$, and f is a function from $[0, 1]$ to $\mathbf{R}$.

Let $\hat{f}_n(x_0)$ be an LP(ℓ) estimator of $f(x_0)$ at point $x_0 \in [0, 1]$. The bias and the variance of $\hat{f}_n(x_0)$ are given by the formulas

$$b(x_0) = \mathbf{E}_f\left[\hat{f}_n(x_0)\right] - f(x_0), \qquad \sigma^2(x_0) = \mathbf{E}_f\left[\hat{f}_n^2(x_0)\right] - \left(\mathbf{E}_f\left[\hat{f}_n(x_0)\right]\right)^2,$$

respectively, where $\mathbf{E}_f$ denotes expectation with respect to the distribution of the random vector $(Y_1, \ldots, Y_n)$ whose coordinates satisfy (1.69). We will sometimes write for brevity $\mathbf{E}$ instead of $\mathbf{E}_f$. The mean squared risk of $\hat{f}_n(x_0)$ at a fixed point x_0 is

$$\mathrm{MSE} = \mathrm{MSE}(x_0) \stackrel{\triangle}{=} \mathbf{E}_f\left[(\hat{f}_n(x_0) - f(x_0))^2\right] = b^2(x_0) + \sigma^2(x_0).$$

We will study separately the bias and the variance terms in this representation of the risk. First, we introduce the following assumptions.

Assumptions (LP)

(**LP1**) *There exist a real number $\lambda_0 > 0$ and a positive integer n_0 such that the smallest eigenvalue $\lambda_{\min}(\mathcal{B}_{nx})$ of $\mathcal{B}_{nx}$ satisfies*

$$\lambda_{\min}(\mathcal{B}_{nx}) \geq \lambda_0$$

for all $n \geq n_0$ and any $x \in [0, 1]$.

(**LP2**) *There exists a real number $a_0 > 0$ such that for any interval $A \subseteq [0, 1]$ and all $n \geq 1$*

$$\frac{1}{n}\sum_{i=1}^{n} I(X_i \in A) \leq a_0 \max(\mathrm{Leb}(A), 1/n)$$

where $\mathrm{Leb}(A)$ denotes the Lebesgue measure of A.

(**LP3**) *The kernel K has compact support belonging to $[-1, 1]$ and there exists a number $K_{\max} < \infty$ such that $|K(u)| \leq K_{\max}$, $\forall\, u \in \mathbf{R}$.*

Assumption (LP1) is stronger than the condition $\mathcal{B}_{nx} > 0$ introduced in the previous section since it is uniform with respect to n and x. We will see that this assumption is natural in the case where the matrix $\mathcal{B}_{nx}$ converges to a limit as $n \to \infty$. Assumption (LP2) means that the points X_i are dense enough in the interval $[0,1]$. It holds for a sufficiently wide range of designs. An important example is given by the regular design: $X_i = i/n$, for which (LP2) is satisfied with $a_0 = 2$. Finally, assumption (LP3) is not restrictive since the choice of K belongs to the statistician.

Since the matrix $\mathcal{B}_{nx}$ is symmetric, assumption (LP1) implies that, for all $n \geq n_0$, $x \in [0,1]$, and $v \in \mathbf{R}^{\ell+1}$,

$$\|\mathcal{B}_{nx}^{-1} v\| \leq \|v\|/\lambda_0 \tag{1.70}$$

where $\|\cdot\|$ denotes the Euclidean norm in $\mathbf{R}^{\ell+1}$.

Lemma 1.3 *Under Assumptions (LP1) – (LP3), for all $n \geq n_0$, $h \geq 1/(2n)$, and $x \in [0,1]$, the weights W_{ni}^* of the* LP(ℓ) *estimator are such that:*

(i) $\displaystyle\sup_{i,x} |W_{ni}^*(x)| \leq \frac{C_*}{nh}$;

(ii) $\displaystyle\sum_{i=1}^{n} |W_{ni}^*(x)| \leq C_*$;

(iii) $W_{ni}^*(x) = 0 \quad$ *if* $\quad |X_i - x| > h$,

where the constant C_* depends only on λ_0, a_0, and $K_{\max}$.

Proof. (i) By (1.70) and by the fact that $\|U(0)\| = 1$, we obtain

$$\begin{aligned}
|W_{ni}^*(x)| &\leq \frac{1}{nh} \left\| \mathcal{B}_{nx}^{-1} U\left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x}{h}\right) \right\| \\
&\leq \frac{1}{nh\lambda_0} \left\| U\left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x}{h}\right) \right\| \\
&\leq \frac{K_{\max}}{nh\lambda_0} \left\| U\left(\frac{X_i - x}{h}\right) \right\| \; I\left(\left|\frac{X_i - x}{h}\right| \leq 1\right) \\
&\leq \frac{K_{\max}}{nh\lambda_0} \sqrt{1 + 1 + \frac{1}{(2!)^2} + \cdots + \frac{1}{(\ell!)^2}} \leq \frac{2K_{\max}}{nh\lambda_0}.
\end{aligned}$$

(ii) In a similar way, by (LP2), we have

$$\sum_{i=1}^{n} |W_{ni}^*(x)| \leq \frac{K_{\max}}{nh\lambda_0} \sum_{i=1}^{n} \left\| U\left(\frac{X_i - x}{h}\right) \right\| \; I\left(\left|\frac{X_i - x}{h}\right| \leq 1\right)$$

$$\leq \frac{2K_{\max}}{nh\lambda_0} \sum_{i=1}^{n} I(x-h \leq X_i \leq x+h)$$

$$\leq \frac{2K_{\max}a_0}{\lambda_0} \max\left(2, \frac{1}{nh}\right) \leq \frac{4K_{\max}a_0}{\lambda_0}.$$

To complete the proof, we take $C_* = \max\{2K_{\max}/\lambda_0, 4K_{\max}a_0/\lambda_0\}$ and observe that (iii) follows from the fact that the support of K is contained in $[-1,1]$. ∎

Proposition 1.13 *Suppose that f belongs to a Hölder class $\Sigma(\beta, L)$ on $[0,1]$, with $\beta > 0$ and $L > 0$. Let $\hat{f}_n$ be the* LP(ℓ) *estimator of f with $\ell = \lfloor \beta \rfloor$. Assume also that:*

(i) the design points $X_1, \ldots, X_n$ are deterministic;
(ii) Assumptions (LP1)–(LP3) hold;
(iii) the random variables ξ_i are independent and such that for all $i = 1, \ldots, n$,

$$\mathbf{E}(\xi_i) = 0, \qquad \mathbf{E}(\xi_i^2) \leq \sigma_{\max}^2 < \infty.$$

Then for all $x_0 \in [0,1]$, $n \geq n_0$, and $h \geq 1/(2n)$ the following upper bounds hold:

$$|b(x_0)| \leq q_1 \, h^\beta, \qquad \sigma^2(x_0) \leq \frac{q_2}{nh},$$

where $q_1 = C_* L/\ell!$ and $q_2 = \sigma_{\max}^2 C_*^2$.

Proof. Using (1.68) and the Taylor expansion of f we obtain that, for $f \in \Sigma(\beta, L)$,

$$b(x_0) = \mathbf{E}_f\left[\hat{f}_n(x_0)\right] - f(x_0) = \sum_{i=1}^{n} f(X_i) \, W_{ni}^*(x_0) - f(x_0)$$

$$= \sum_{i=1}^{n} (f(X_i) - f(x_0)) W_{ni}^*(x_0)$$

$$= \sum_{i=1}^{n} \frac{f^{(\ell)}(x_0 + \tau_i(X_i - x_0)) - f^{(\ell)}(x_0)}{\ell\,!} (X_i - x_0)^\ell \, W_{ni}^*(x_0),$$

where $0 \leq \tau_i \leq 1$. This representation and statements (ii) and (iii) of Lemma 1.3 imply that

$$|b(x_0)| \leq \sum_{i=1}^{n} \frac{L|X_i - x_0|^\beta}{\ell\,!} |W_{ni}^*(x_0)|$$

$$= L \sum_{i=1}^{n} \frac{|X_i - x_0|^\beta}{\ell\,!} |W_{ni}^*(x_0)| I(|X_i - x_0| \leq h)$$

$$\leq L \sum_{i=1}^{n} \frac{h^\beta}{\ell!} |W_{ni}^*(x_0)| \leq \frac{LC_*}{\ell\,!} \, h^\beta = q_1 \, h^\beta.$$

The variance satisfies

$$\sigma^2(x_0) = \mathbf{E}\left[\left(\sum_{i=1}^{n} \xi_i\, W_{ni}^*(x_0)\right)^2\right] = \sum_{i=1}^{n} (W_{ni}^*(x_0))^2 \mathbf{E}(\xi_i^2)$$

$$\leq \sigma_{\max}^2 \sup_{i,x} |W_{ni}^*(x)| \sum_{i=1}^{n} |W_{ni}^*(x_0)| \leq \frac{\sigma_{\max}^2 C_*^2}{nh} = \frac{q_2}{nh}. \qquad \blacksquare$$

Proposition 1.13 implies that

$$\mathrm{MSE} \leq q_1^2 h^{2\beta} + \frac{q_2}{nh}$$

and that the minimizer h_n^* with respect to h of this upper bound on the risk is given by

$$h_n^* = \left(\frac{q_2}{2\beta q_1^2}\right)^{\frac{1}{2\beta+1}} n^{-\frac{1}{2\beta+1}}.$$

Therefore we obtain the following result.

Theorem 1.6 *Under the assumptions of Proposition 1.13 and if the bandwidth is chosen to be* $h = h_n = \alpha n^{-\frac{1}{2\beta+1}}, \alpha > 0$, *the following upper bound holds:*

$$\limsup_{n\to\infty} \sup_{f\in\Sigma(\beta,L)} \sup_{x_0\in[0,1]} \mathbf{E}_f\left[\psi_n^{-2} |\hat{f}_n(x_0) - f(x_0)|^2\right] \leq C < \infty, \tag{1.71}$$

where $\psi_n = n^{-\frac{\beta}{2\beta+1}}$ *is the rate of convergence and* C *is a constant depending only on* $\beta, L, \lambda_0, a_0, \sigma_{\max}^2, K_{\max}$, *and* α.

Corollary 1.2 *Under the assumptions of Theorem 1.6 we have*

$$\limsup_{n\to\infty} \sup_{f\in\Sigma(\beta,L)} \mathbf{E}_f\left[\psi_n^{-2} \|\hat{f}_n - f\|_2^2\right] \leq C < \infty, \tag{1.72}$$

where $\|f\|_2^2 = \int_0^1 f^2(x)dx$, $\psi_n = n^{-\frac{\beta}{2\beta+1}}$ *and where* C *is a constant depending only on* $\beta, L, \lambda_0, a_0, \sigma_{\max}^2, K_{\max}$, *and* α.

We now discuss Assumption (LP1) in more detail. If the design is regular and n is large enough, $\mathcal{B}_{nx}$ is close to the matrix $\mathcal{B} = \int U(u)U^T(u)K(u)du$, which is independent of n and x. Therefore, for Assumption (LP1) to hold we only need to assure that $\mathcal{B}$ is positive definite. This is indeed true, except for pathological cases, as the following lemma states.

Lemma 1.4 *Let* $K : \mathbf{R} \to [0, +\infty)$ *be a function such that the Lebesgue measure* $\mathrm{Leb}(u : K(u) > 0) > 0$. *Then the matrix*

$$\mathcal{B} = \int U(u)U^T(u)K(u)du$$

is positive definite.

PROOF. It is sufficient to prove that for all $v \in \mathbf{R}^{\ell+1}$ satisfying $v \neq 0$ we have

$$v^T \mathcal{B} v > 0.$$

Clearly,

$$v^T \mathcal{B} v = \int (v^T U(u))^2 K(u) du \geq 0.$$

If there exists $v \neq 0$ such that $\int [v^T U(u)]^2 K(u)\, du = 0$, then $v^T U(u) = 0$ for almost all u on the set $\{u : K(u) > 0\}$, which has a positive Lebesgue measure by assumption of the lemma. But the function $u \mapsto v^T U(u)$ is a polynomial of degree $\leq \ell$ which cannot be equal to zero except for a finite number of points. Thus, we come to a contradiction showing that $\int [v^T U(u)]^2 K(u)\, du = 0$ is impossible for $v \neq 0$. ∎

Lemma 1.5 *Suppose that there exist $K_{\min} > 0$ and $\Delta > 0$ such that*

$$K(u) \geq K_{\min} I(|u| \leq \Delta), \quad \forall\ u \in \mathbf{R}, \tag{1.73}$$

and that $X_i = i/n$ for $i = 1, \ldots, n$. Let $h = h_n$ be a sequence satisfying

$$h_n \to 0, \qquad n h_n \to \infty \tag{1.74}$$

as $n \to \infty$. Then Assumption (LP1) holds.

PROOF. Let us show that

$$\inf_{\|v\|=1} v^T \mathcal{B}_{nx} v \geq \lambda_0$$

for sufficiently large n. By (1.73), we have

$$v^T \mathcal{B}_{nx} v \geq \frac{K_{\min}}{nh} \sum_{i=1}^{n} (v^T U(z_i))^2 I(|z_i| \leq \Delta) \tag{1.75}$$

where $z_i = (X_i - x)/h$. Observe that $z_i - z_{i-1} = (nh)^{-1}$ and

$$z_1 = \frac{1}{nh} - \frac{x}{h} \leq \frac{1}{nh}, \qquad z_n = \frac{1-x}{h} \geq 0.$$

If $x < 1 - h\Delta$, then $z_n > \Delta$ and the points z_i form a grid with step $(nh)^{-1}$ on an interval covering $[0, \Delta]$. Moreover, $nh \to \infty$ and therefore

$$\begin{aligned} \frac{1}{nh} \sum_{i=1}^{n} (v^T U(z_i))^2 I(|z_i| \leq \Delta) \ &\geq\ \frac{1}{nh} \sum_{i=1}^{n} (v^T U(z_i))^2 I(0 \leq z_i \leq \Delta) \\ &\to \int_0^{\Delta} (v^T U(z))^2 dz \ \text{ as } n \to \infty, \end{aligned} \tag{1.76}$$

since the Riemann sum converges to the integral.

If $x \geq 1 - h\Delta$, then (1.74) implies that $z_1 < -\Delta$ for sufficiently large n and that the points z_i form a grid with step $(nh)^{-1}$ on an interval covering $[-\Delta, 0]$. As before, we obtain

$$\frac{1}{nh}\sum_{i=1}^{n}(v^T U(z_i))^2 I(|z_i| \leq \Delta) \geq \frac{1}{nh}\sum_{i=1}^{n}(v^T U(z_i))^2 I(-\Delta \leq z_i \leq 0)$$
$$\to \int_{-\Delta}^{0}(v^T U(z))^2 dz \quad \text{as } n \to \infty. \qquad (1.77)$$

It is easy to see that convergence in (1.76) and (1.77) is uniform on $\{\|v\| = 1\}$. This remark and (1.75)–(1.77) imply that

$$\inf_{\|v\|=1} v^T \mathcal{B}_{nx} v \geq \frac{K_{\min}}{2} \min\left\{ \inf_{\|v\|=1}\int_0^{\Delta}(v^T U(z))^2 dz, \ \inf_{\|v\|=1}\int_{-\Delta}^{0}(v^T U(z))^2 dz \right\}$$

for sufficiently large n. To complete the proof, it remains to apply Lemma 1.4 for $K(u) = I(0 \leq u \leq \Delta)$ and $K(u) = I(-\Delta \leq u \leq 0)$, respectively. ∎

Using Theorem 1.6, Corollary 1.2, and Lemma 1.5 we obtain the following result.

Theorem 1.7 *Assume that f belongs to the Hölder class $\Sigma(\beta, L)$ on $[0,1]$ where $\beta > 0$ and $L > 0$. Let $\hat{f}_n$ be the* LP(ℓ) *estimator of f with $\ell = \lfloor \beta \rfloor$. Suppose also that:*

(i) $X_i = i/n$ for $i = 1, \ldots, n$;
(ii) the random variables ξ_i are independent and satisfy

$$\mathbf{E}(\xi_i) = 0, \qquad \mathbf{E}(\xi_i^2) \leq \sigma_{\max}^2 < \infty$$

for all $i = 1, \ldots, n$;
(iii) there exist constants $K_{\min} > 0$, $\Delta > 0$ and $K_{\max} < \infty$ such that

$$K_{\min} I(|u| \leq \Delta) \leq K(u) \leq K_{\max} I(|u| \leq 1), \quad \forall\ u \in \mathbf{R};$$

(iv) $h = h_n = \alpha n^{-\frac{1}{2\beta+1}}$ for some $\alpha > 0$.

Then the estimator $\hat{f}_n$ satisfies (1.71) and (1.72).

1.6.2 Convergence in the sup-norm

Define the L_∞-risk of the estimator $\hat{f}_n$ as $\mathbf{E}_f \|\hat{f}_n - f\|_\infty^2$ where

$$\|f\|_\infty = \sup_{t \in [0,1]} |f(t)|.$$

In this section we study the rate at which the L_∞-risk of the local polynomial estimator tends to zero. We will need the following preliminary results.

Lemma 1.6 *Let $\eta_1, \dots, \eta_M$ be random variables such that, for two constants $\alpha_0 > 0$ and $C_0 < \infty$, we have $\max_{1 \le j \le M} \mathbf{E}[\exp(\alpha_0 \eta_j^2)] \le C_0$. Then*

$$\mathbf{E}\left[\max_{1 \le j \le M} \eta_j^2\right] \le \frac{1}{\alpha_0} \log(C_0 M).$$

Proof. Using Jensen's inequality we obtain

$$\begin{aligned}
\mathbf{E}\Big[\max_j \eta_j^2\Big] &= \frac{1}{\alpha_0}\mathbf{E}\Big[\max_j \log\left(\exp(\alpha_0 \eta_j^2)\right)\Big] = \frac{1}{\alpha_0}\mathbf{E}\Big[\log\Big(\max_j \exp(\alpha_0\, \eta_j^2)\Big)\Big] \\
&\le \frac{1}{\alpha_0}\log \mathbf{E}\Big[\max_j \exp(\alpha_0\, \eta_j^2)\Big] \le \frac{1}{\alpha_0}\log \mathbf{E}\left[\sum_{j=1}^{M} \exp(\alpha_0\, \eta_j^2)\right] \\
&\le \frac{1}{\alpha_0}\log\left(M \max_j \mathbf{E}\Big[\exp(\alpha_0\, \eta_j^2)\Big]\right) \le \frac{1}{\alpha_0}\log(C_0 M). \qquad \blacksquare
\end{aligned}$$

Observe that Lemma 1.6 does not require the random variables η_j to be independent.

Corollary 1.3 *Suppose that $\eta_1, \dots, \eta_M$ are Gaussian random vectors on $\mathbf{R}^d$ such that $\mathbf{E}(\eta_j) = 0$ and $\max_{1 \le j \le M} \max_{1 \le k \le d} \mathbf{E}(\eta_{jk}^2) \le \sigma_{\max}^2 < \infty$ where η_{jk} is the kth component of the vector η_j. Then*

$$\mathbf{E}\left[\max_{1 \le j \le M} \|\eta_j\|^2\right] \le 4d\sigma_{\max}^2 \log(\sqrt{2} M d),$$

where $\|\cdot\|$ denotes the Euclidean norm on $\mathbf{R}^d$.

Proof. We have

$$\mathbf{E}\left[\max_{1 \le j \le M} \|\eta_j\|^2\right] \le d\ \mathbf{E}\left[\max_{1 \le j \le M} \max_{1 \le k \le d} \eta_{jk}^2\right],$$

The random variables η_{jk} are Gaussian, have zero means and variances $\sigma_{jk}^2 = \mathbf{E}(\eta_{jk}^2) \le \sigma_{\max}^2$. Therefore

$$\mathbf{E}\left[\exp(\alpha_0 \eta_{jk}^2)\right] \le \frac{1}{\sqrt{2\pi}\sigma_{jk}} \int \exp\left(-\frac{x^2}{4\sigma_{jk}^2}\right) dx = \sqrt{2}$$

for $\alpha_0 = 1/(4\sigma_{\max}^2)$. To complete the proof it remains to apply Lemma 1.6 with $C_0 = \sqrt{2}$. $\blacksquare$

The following theorem establishes an upper bound on the L_∞-risk of local polynomial estimators.

Theorem 1.8 *Suppose that f belongs to a Hölder class $\Sigma(\beta, L)$ on $[0,1]$ where $\beta > 0$ and $L > 0$. Let $\hat{f}_n$ be the* LP(ℓ) *estimator of order $\ell = \lfloor \beta \rfloor$ with bandwidth*

$$h_n = \alpha \left(\frac{\log n}{n} \right)^{\frac{1}{2\beta+1}} \tag{1.78}$$

for some $\alpha > 0$. Suppose also that:

(i) the design points $X_1, \ldots, X_n$ are deterministic;
(ii) Assumptions (LP1)–(LP3) hold;
(iii) the random variables ξ_i are i.i.d. Gaussian $\mathcal{N}(0, \sigma_\xi^2)$ with $0 < \sigma_\xi^2 < \infty$;
(iv) K is a Lipschitz kernel: $K \in \Sigma(1, L_K)$ on $\mathbf{R}$ with $0 < L_K < \infty$.

Then there exists a constant $C < \infty$ such that

$$\limsup_{n\to\infty} \sup_{f\in\Sigma(\beta,L)} \mathbf{E}_f \left[\psi_n^{-2} \|\hat{f}_n - f\|_\infty^2 \right] \le C,$$

where

$$\psi_n = \left(\frac{\log n}{n} \right)^{\frac{\beta}{2\beta+1}}. \tag{1.79}$$

Proof. Using Proposition 1.13 and writing for brevity $\mathbf{E} = \mathbf{E}_f$ we get

$$\begin{aligned}
\mathbf{E}\|\hat{f}_n - f\|_\infty^2 &\le \mathbf{E}\left[\|\hat{f}_n - \mathbf{E}\hat{f}_n\|_\infty + \|\mathbf{E}\hat{f}_n - f\|_\infty \right]^2 \\
&\le 2\mathbf{E}\|\hat{f}_n - \mathbf{E}\hat{f}_n\|_\infty^2 + 2\Big(\sup_{x\in[0,1]} |b(x)| \Big)^2 \\
&\le 2\mathbf{E}\|\hat{f}_n - \mathbf{E}\hat{f}_n\|_\infty^2 + 2q_1^2 h_n^{2\beta}. \qquad (1.80)
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\mathbf{E}\|\hat{f}_n - \mathbf{E}\hat{f}_n\|_\infty^2 &= \mathbf{E}\left[\sup_{x\in[0,1]} \left| \hat{f}_n(x) - \mathbf{E}\left[\hat{f}_n(x) \right] \right|^2 \right] \\
&= \mathbf{E}\left[\sup_{x\in[0,1]} \left| \sum_{i=1}^n \xi_i \, W_{ni}^*(x) \right|^2 \right], \qquad (1.81)
\end{aligned}$$

where

$$\begin{aligned}
W_{ni}^*(x) &= \frac{1}{nh} U^T(0) \mathcal{B}_{nx}^{-1} U\left(\frac{X_i - x}{h} \right) K\left(\frac{X_i - x}{h} \right) \\
&= \frac{1}{nh} U^T(0) \mathcal{B}_{nx}^{-1} S_i(x)
\end{aligned}$$

and

$$S_i(x) = U\left(\frac{X_i - x}{h} \right) K\left(\frac{X_i - x}{h} \right).$$

In view of (1.70), we have

$$\left|\sum_{i=1}^{n}\xi_i\, W^*_{ni}(x)\right| \le \frac{1}{nh}\left\|\mathcal{B}^{-1}_{nx}\sum_{i=1}^{n}\xi_i S_i(x)\right\| \le \frac{1}{\lambda_0 nh}\left\|\sum_{i=1}^{n}\xi_i S_i(x)\right\|,$$

where $\|\cdot\|$ denotes the Euclidean norm. Set $M = n^2$ and $x_j = j/M$ for $j = 1, \ldots, M$. Then

$$\begin{aligned} A \stackrel{\triangle}{=} \sup_{x\in[0,1]}\left|\sum_{i=1}^{n}\xi_i\, W^*_{ni}(x)\right| &\le \frac{1}{\lambda_0 nh}\sup_{x\in[0,1]}\left\|\sum_{i=1}^{n}\xi_i S_i(x)\right\| \\ &\le \frac{1}{\lambda_0 nh}\left(\max_{1\le j\le M}\left\|\sum_{i=1}^{n}\xi_i\ S_i(x_j)\right\|\right. \\ &\quad \left. + \sup_{x,x':\ |x-x'|\le 1/M}\left\|\sum_{i=1}^{n}\xi_i\ (S_i(x)-S_i(x'))\right\|\right). \end{aligned}$$

Since $K \in \Sigma(1, L_K)$ and the support of the kernel K belongs to $[-1, 1]$, and since $U(\cdot)$ is a vector function with polynomial coordinates, there exists a constant $\bar{L}$ such that $\|U(u)K(u) - U(u')K(u')\| \le \bar{L}|u - u'|,\ \forall\ u, u' \in \mathbf{R}$. Thus

$$\begin{aligned} A^2 &\le \left(\frac{1}{\lambda_0 nh}\right)^2\left(\max_{1\le j\le M}\left\|\sum_{i=1}^{n}\xi_i S_i(x_j)\right\| + \frac{\bar{L}}{Mh}\sum_{i=1}^{n}|\xi_i|\right)^2 \\ &\le \frac{2}{\lambda_0^2 nh}\left[\max_{1\le j\le M}\|\eta_j\|^2\right] + \frac{2\bar{L}^2}{\lambda_0^2 n^2h^4M^2}\left(\sum_{i=1}^{n}|\xi_i|\right)^2, \end{aligned}$$

where the random vectors η_j are given by

$$\eta_j = \frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\xi_i\, S_i(x_j).$$

Therefore we have

$$\mathbf{E}(A^2) \le \frac{2}{\lambda_0^2 nh}\mathbf{E}\left[\max_{1\le j\le M}\|\eta_j\|^2\right] + \frac{2\bar{L}^2}{\lambda_0^2 n^2h^4M^2}\,\mathbf{E}\left[\left(\sum_{i=1}^{n}|\xi_i|\right)^2\right]. \tag{1.82}$$

Further,

$$\frac{1}{M^2n^2h^4}\mathbf{E}\left[\left(\sum_{i=1}^{n}|\xi_i|\right)^2\right] \le \frac{\mathbf{E}(\xi_1^2)}{M^2h^4} = \frac{\sigma_\xi^2}{(nh)^4} = o\left(\frac{1}{nh}\right). \tag{1.83}$$

Since η_j are zero mean Gaussian vectors, we repeat the argument of the proof of Lemma 1.3 to obtain

$$\mathbf{E}[\|\eta_j\|^2] = \frac{1}{nh}\sum_{i=1}^{n}\sigma_\xi^2 \left\|U\left(\frac{X_i - x_j}{h}\right)\right\|^2 K^2\left(\frac{X_i - x_j}{h}\right) \tag{1.84}$$

$$\leq \frac{4K_{\max}^2\sigma_\xi^2}{nh}\sum_{i=1}^{n} I\left(|X_i - x_j| \leq h\right)$$

$$\leq 4K_{\max}^2\sigma_\xi^2 a_0 \max\left(2, \frac{1}{nh}\right).$$

Then, by Corollary 1.3, we have

$$\mathbf{E}\left[\max_{1\leq j\leq M}\|\eta_j\|^2\right] = O(\log M) = O(\log n) \quad \text{as } n \to \infty. \tag{1.85}$$

From (1.81)–(1.85) we get

$$\mathbf{E}\|\hat{f}_n - \mathbf{E}\hat{f}_n\|_\infty^2 \leq \frac{q_3 \log n}{nh},$$

where $q_3 > 0$ is a constant independent of f and n. This upper bound combined with (1.80) implies that

$$\mathbf{E}\|\hat{f}_n - f\|_\infty^2 \leq \frac{q_3 \log n}{nh} + 2q_1^2 h^{2\beta}.$$

Choose the bandwidth according to (1.78) to complete the proof. ∎

Theorem 1.8 states that the rate ψ_n given by (1.79) is a uniform convergence rate of $\hat{f}_n$ with respect to the L_∞-norm on the class $\Sigma(\beta, L)$. In contrast to the rate of convergence at a fixed point x_0 or in the L_2-norm, an additional logarithmic factor appears, slowing down the convergence. We will prove in Chapter 2 that (1.79) is the optimal rate of convergence in the L_∞-norm on the class $\Sigma(\beta, L)$.

1.7 Projection estimators

Here we continue to consider the nonparametric regression model

$$Y_i = f(X_i) + \xi_i, \qquad i = 1, \ldots, n,$$

where ξ_i are independent random variables, $\mathbf{E}(\xi_i) = 0$, the values $X_i \in [0, 1]$ are deterministic and $f : [0, 1] \to \mathbf{R}$. We will mainly focus on a particular case, $X_i = i/n$.

Suppose that $f \in L_2[0, 1]$. Let θ_j be the Fourier coefficients of f with respect to an orthonormal basis $\{\varphi_j\}_{j=1}^{\infty}$ of $L_2[0, 1]$:

$$\theta_j = \int_0^1 f(x)\varphi_j(x)dx.$$

Assume that f can be represented as

$$f(x) = \sum_{j=1}^{\infty} \theta_j \varphi_j(x), \tag{1.86}$$

where the series converges for all $x \in [0,1]$.

Projection estimation of f is based on a simple idea: approximate f by its projection $\sum_{j=1}^{N} \theta_j \varphi_j$ on the linear span of the first N functions of the basis $\varphi_1, \ldots, \varphi_N$ and replace θ_j by their estimators. Observe that if X_i are scattered over $[0,1]$ in a sufficiently uniform way, which happens, e.g., in the case $X_i = i/n$, the coefficients θ_j are well approximated by the sums $n^{-1} \sum_{i=1}^{n} f(X_i)\varphi_j(X_i)$. Replacing in these sums the unknown quantities $f(X_i)$ by the observations Y_i we obtain the following estimators of θ_j:

$$\hat{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} Y_i \varphi_j(X_i). \tag{1.87}$$

Definition 1.9 *Let $N \geq 1$ be an integer. The statistic*

$$\hat{f}_{nN}(x) = \sum_{j=1}^{N} \hat{\theta}_j \varphi_j(x)$$

is called a **projection estimator** *(or an* **orthogonal series estimator***) of the regression function f at the point x.*

Let us emphasize that this definition only makes sense if the points X_i are scattered over $[0,1]$ in a sufficiently uniform way, e.g., if $X_i = i/n$ or X_i are i.i.d. uniformly distributed on $[0,1]$. A generalization to arbitrary X_i is given, for example, by the nonparametric least squares estimator discussed in Section 1.7.3.

The parameter N (called the *order* of the projection estimator) plays the same role as the bandwidth h for kernel estimators: similarly to h it is a *smoothing parameter*, i.e., a parameter whose choice is crucial for establishing the balance between bias and variance. The choice of very large N leads to undersmoothing, whereas for small values of N oversmoothing occurs. These effects can be understood through the results of Section 1.7.2 below.

Note that $\hat{f}_{nN}$ is a linear estimator, since we may write it in the form

$$\hat{f}_{nN}(x) = \sum_{i=1}^{n} Y_i W_{ni}^{**}(x)$$

with

$$W_{ni}^{**}(x) = \frac{1}{n} \sum_{j=1}^{N} \varphi_j(X_i)\varphi_j(x). \tag{1.88}$$

The bases $\{\varphi_j\}$ that are most frequently used in projection estimation are the trigonometric basis and the wavelet bases.

Example 1.3 *Trigonometric basis.*

This is the orthonormal basis in $L_2[0,1]$ defined by

$$\begin{aligned}\varphi_1(x) &\equiv 1,\\ \varphi_{2k}(x) &= \sqrt{2}\cos(2\pi kx),\\ \varphi_{2k+1}(x) &= \sqrt{2}\sin(2\pi kx), \quad k = 1, 2, \ldots,\end{aligned}$$

where $x \in [0,1]$.

Example 1.4 *Wavelet bases.*

Let $\psi : \mathbf{R} \to \mathbf{R}$ be a sufficiently smooth function with a compact support. Define an infinite set of functions as follows:

$$\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k), \quad j, k \in \mathbb{Z}. \tag{1.89}$$

It can be shown that, under certain assumptions on ψ, the system (1.89) is an orthonormal basis in $L_2(\mathbf{R})$ and, for all $f \in L_2(\mathbf{R})$,

$$f = \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} \theta_{jk}\psi_{jk}, \qquad \theta_{jk} = \int f\psi_{jk},$$

where the series converges in $L_2(\mathbf{R})$. We can view this expansion as a particular case of (1.86) if we switch from the double index at θ_{jk} and ψ_{jk} to a single one. Basis (1.89) is called a wavelet basis. There exists a similar construction for $L_2[0,1]$ instead of $L_2(\mathbf{R})$ where the functions ψ_{jk} are corrected at the extremes of the interval $[0,1]$ in order to preserve orthonormality.

The main difference between the trigonometric basis and wavelet bases consists in the fact that the trigonometric basis "localizes" the function f in the frequency domain only, while the wavelet bases "localize" it both in the frequency domain and time domain if we interpret x as a time variable (the index j corresponds to frequency and k characterizes position in time).

Projection estimators of a probability density are defined in a similar way. Let $X_1, \ldots, X_n$ be i.i.d. random variables with Lebesgue density $p \in L_2(A)$ where $A \subseteq \mathbf{R}$ is a given interval. Consider the Fourier coefficients $c_j = \int p\varphi_j$ of p with respect to an orthonormal basis $\{\varphi_j\}_{j=1}^{\infty}$ of $L_2(A)$. Introduce the following estimators of c_j:

$$\hat{c}_j = \frac{1}{n}\sum_{i=1}^{n} \varphi_j(X_i).$$

Definition 1.10 *Let $N \geq 1$ be an integer. The statistic*

$$\hat{p}_{nN}(x) = \sum_{j=1}^{N} \hat{c}_j \varphi_j(x)$$

is called a **projection estimator** *(or an* **orthogonal series estimator***) of the probability density p at the point x.*

It is straightforward to see that $\hat{c}_j$ is an unbiased estimator of c_j whatever are the interval A and the orthonormal basis $\{\varphi_j\}_{j=1}^{\infty}$ of $L_2(A)$. For the trigonometric basis, more detailed properties can be established (cf. Exercise 1.9).

In the rest of this section, we consider only projection estimators of a regression function f using the trigonometric basis and we study their convergence in the $L_2[0,1]$ norm.

1.7.1 Sobolev classes and ellipsoids

We will assume that the regression function f is sufficiently smooth, or more specifically, that it belongs to a Sobolev class of functions. Several definitions of Sobolev classes will be used below. First, we define the Sobolev class for integer smoothness β.

Definition 1.11 *Let $\beta \in \{1, 2, \ldots\}$ and $L > 0$. The Sobolev class $W(\beta, L)$ is defined by*

$$W(\beta, L) = \Big\{ f \in [0,1] \to \mathbf{R} : f^{(\beta-1)} \text{ is absolutely continuous and } \int_0^1 (f^{(\beta)}(x))^2 dx \leq L^2 \Big\}.$$

The periodic Sobolev class $W^{per}(\beta, L)$ is defined by

$$W^{per}(\beta, L) = \Big\{ f \in W(\beta, L) : f^{(j)}(0) = f^{(j)}(1), \quad j = 0, 1, \ldots, \beta - 1 \Big\}.$$

It is easy to see that for all $\beta \in \{1, 2, \ldots\}$ and all $L > 0$ the Sobolev class $W(\beta, L)$ contains the Hölder class $\Sigma(\beta, L)$ on the interval $[0, 1]$.

Recall that any function $f \in W^{per}(\beta, L)$ admits representation (1.86) where the sequence $\theta = \{\theta_j\}_{j=1}^{\infty}$ of its Fourier coefficients belongs to the space

$$\ell^2(\mathbf{N}) = \Big\{ \theta : \sum_{j=1}^{\infty} \theta_j^2 < \infty \Big\}$$

and $\{\varphi_j\}_{j=1}^{\infty}$ is the trigonometric basis defined in Example 1.3. We now give a necessary and sufficient condition on θ under which the function

$$f(x) = \theta_1\varphi_1(x) + \sum_{k=1}^{\infty}(\theta_{2k}\varphi_{2k}(x) + \theta_{2k+1}\,\varphi_{2k+1}(x))$$

belongs to the class $W^{per}(\beta, L)$. Define

$$a_j = \begin{cases} j^\beta, & \text{for even } j, \\ (j-1)^\beta, & \text{for odd } j. \end{cases} \tag{1.90}$$

Proposition 1.14 *Let $\beta \in \{1, 2, \ldots\}$, $L > 0$, and let $\{\varphi_j\}_{j=1}^{\infty}$ be the trigonometric basis. Then the function $f = \sum_{j=1}^{\infty}\theta_j\varphi_j$ belongs to $W^{per}(\beta, L)$ if and only if the vector θ of the Fourier coefficients of f belongs to an ellipsoid in $\ell^2(\mathbf{N})$ defined as follows:*

$$\Theta(\beta, Q) = \Big\{\theta \in \ell^2(\mathbf{N}) : \sum_{j=1}^{\infty} a_j^2\theta_j^2 \le Q\Big\} \tag{1.91}$$

where $Q = L^2/\pi^{2\beta}$ and a_j is given by (1.90).

A proof of this proposition is given in the Appendix (Lemma A.3).

The set $\Theta(\beta, Q)$ defined by (1.91) with $\beta > 0$ (not necessarily an integer), $Q > 0$, and a_j satisfying (1.90) is called a *Sobolev ellipsoid*. We mention the following properties of these ellipsoids.

(1) The monotonicity with respect to inclusion:

$$0 < \beta' \le \beta \Longrightarrow \Theta(\beta, Q) \subseteq \Theta(\beta', Q).$$

(2) If $\beta > 1/2$, the function $f = \sum_{j=1}^{\infty}\theta_j\varphi_j$ with the trigonometric basis $\{\varphi_j\}_{j=1}^{\infty}$ and $\theta \in \Theta(\beta, Q)$ is continuous (check this as an exercise). In what follows, we will basically consider this case.

(3) Since $a_1 = 0$, we can write

$$\Theta(\beta, Q) = \Big\{\theta \in \ell^2(\mathbf{N}) : \sum_{j=2}^{\infty} a_j^2\theta_j^2 \le Q\Big\}.$$

The ellipsoid $\Theta(\beta, Q)$ is well-defined for all $\beta > 0$. In this sense $\Theta(\beta, Q)$ is a more general object than the periodic Sobolev class $W^{per}(\beta, L)$, where β can only be an integer. Proposition 1.14 establishes an isomorphism between $\Theta(\beta, Q)$ and $W^{per}(\beta, L)$ for integer β. It can be extended to all $\beta > 0$ by generalizing the definition of $W^{per}(\beta, L)$ in the following way.

Definition 1.12 *For $\beta > 0$ and $L > 0$ the Sobolev class $\tilde{W}(\beta, L)$ is defined as follows:*

$$\tilde{W}(\beta, L) = \{f \in L_2[0,1] : \theta = \{\theta_j\} \in \Theta(\beta, Q)\}$$

where $\theta_j = \int_0^1 f\varphi_j$ and $\{\varphi_j\}_{j=1}^{\infty}$ is the trigonometric basis. Here $\Theta(\beta, Q)$ is the Sobolev ellipsoid defined by (1.91), where $Q = L^2/\pi^{2\beta}$ and the coefficients a_j are given in (1.90).

For all $\beta > 1/2$, the functions belonging to $\tilde{W}(\beta, L)$ are continuous. On the contrary, they are not always continuous for $\beta \le 1/2$; an example is given by the function $f(x) = \text{sign}(x - 1/2)$, whose Fourier coefficients θ_j are of order $1/j$.

1.7.2 Integrated squared risk of projection estimators

Let us now study the mean integrated squared error (MISE) of the projection estimator $\hat{f}_{nN}$:

$$\text{MISE} \triangleq \mathbf{E}_f \|\hat{f}_{nN} - f\|_2^2 = \mathbf{E}_f \int_0^1 (\hat{f}_{nN}(x) - f(x))^2 dx.$$

We will need the following assumption.

Assumption (A)

(i) We consider the nonparametric regression model

$$Y_i = f(X_i) + \xi_i, \ \ i = 1, \dots, n,$$

where f is a function from $[0,1]$ to $\mathbf{R}$. The random variables ξ_i are independent with

$$\mathbf{E}(\xi_i) = 0, \quad \mathbf{E}(\xi_i^2) = \sigma_\xi^2 < \infty$$

and $X_i = i/n$ for $i = 1, \dots, n$.

(ii) $\{\varphi_j\}_{j=1}^{\infty}$ is the trigonometric basis.

(iii) The Fourier coefficients $\theta_j = \int_0^1 f\varphi_j$ of f satisfy

$$\sum_{j=1}^{\infty} |\theta_j| < \infty.$$

It follows from parts (ii) and (iii) of Assumption (A) that the series $\sum_{j=1}^{\infty} \theta_j \varphi_j(x)$ is absolutely convergent for all $x \in [0,1]$, and thus the pointwise representation (1.86) holds.

We will use the following property of the trigonometric basis.

Lemma 1.7 *Let $\{\varphi_j\}_{j=1}^{\infty}$ be the trigonometric basis. Then*

$$\frac{1}{n}\sum_{s=1}^{n}\varphi_j(s/n)\varphi_k(s/n)=\delta_{jk},\qquad 1\le j,k\le n-1, \tag{1.92}$$

where δ_{jk} is the Kronecker delta.

Proof. For brevity we consider only the case $\varphi_j(x)=\sqrt{2}\cos(2\pi mx)$, $\varphi_k(x)=\sqrt{2}\sin(2\pi lx)$ where $j=2m$, $k=2l+1$, $j\le n-1$, $k\le n-1$, $n\ge 2$ and $m\ge 1$, $l\ge 1$ are integers. Other cases can be studied along similar lines. Put

$$a\triangleq \exp\{\mathrm{i}2\pi m/n\},\qquad b\triangleq \exp\{\mathrm{i}2\pi l/n\}.$$

Then

$$\begin{aligned}S&\triangleq \frac{1}{n}\sum_{s=1}^{n}\varphi_j(s/n)\varphi_k(s/n)=\frac{2}{n}\sum_{s=1}^{n}\frac{(a^s+a^{-s})(b^s-b^{-s})}{4\mathrm{i}}\\&=\frac{1}{2\mathrm{i}n}\sum_{s=1}^{n}\Big[(ab)^s-(a/b)^s+(b/a)^s-(ab)^{-s}\Big].\end{aligned}$$

Since $ab\ne 1$ and $(ab)^n=1$, we have

$$\sum_{s=1}^{n}(ab)^s=ab\frac{(ab)^n-1}{ab-1}=0.$$

By the same argument, $\sum_{s=1}^{n}(ab)^{-s}=0$. If $m\ne l$, then $\sum_{s=1}^{n}(a/b)^s=\sum_{s=1}^{n}(b/a)^s=0$, whereas for $m=l$ we have $\sum_{s=1}^{n}(a/b)^s=\sum_{s=1}^{n}(b/a)^s=n$. Thus, $S=0$. ∎

Lemma 1.7 implies that the projection estimator f_{nN} with the trigonometric basis $\{\varphi_j\}_{j=1}^{\infty}$ has the property of reproduction of polynomials similar to that of the local polynomial estimator (cf. Proposition 1.12). However, here we deal with trigonometric, rather than algebraic, polynomials of degree $\le N$, i.e., with functions of the form

$$Q(x)=\sum_{k=1}^{N}b_k\varphi_k(x)$$

where $\{\varphi_j\}_{j=1}^{\infty}$ is the trigonometric basis and b_k are some coefficients. In fact, the following proposition holds.

Proposition 1.15 *Let $N\le n-1$ and let $X_i=i/n$ for $i=1,\dots,n$. If Q is a trigonometric polynomial of degree $\le N$, we have*

$$\sum_{i=1}^{n} Q(X_i) W_{ni}^{**}(x) = Q(x)$$

for all $x \in [0,1]$.

PROOF follows immediately from Lemma 1.7 and from the definition of W_{ni}^{**}. ∎

The next result gives the bias and the squared risk of the estimators $\hat{\theta}_j$.

Proposition 1.16 *Under Assumption (A) the estimators* $\hat{\theta}_j$ *defined in (1.87) satisfy*

(i) $\mathbf{E}(\hat{\theta}_j) = \theta_j + \alpha_j$,

(ii) $\mathbf{E}[(\hat{\theta}_j - \theta_j)^2] = \sigma_\xi^2/n + \alpha_j^2, \quad 1 \le j \le n-1,$

where

$$\alpha_j = \frac{1}{n} \sum_{i=1}^{n} f(i/n)\varphi_j(i/n) - \int_0^1 f(x)\varphi_j(x)dx.$$

PROOF. We have

$$\hat{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} Y_i \, \varphi_j(i/n) = \frac{1}{n} \left(\sum_{i=1}^{n} f(i/n)\varphi_j(i/n) + \sum_{i=1}^{n} \xi_i \, \varphi_j(i/n) \right).$$

Therefore

$$\mathbf{E}(\hat{\theta}_j) = \frac{1}{n} \sum_{i=1}^{n} f(i/n)\varphi_j(i/n) = \alpha_j + \theta_j.$$

Then

$$\mathbf{E}[(\hat{\theta}_j - \theta_j)^2] = \mathbf{E}[(\hat{\theta}_j - \mathbf{E}(\hat{\theta}_j))^2] + (\mathbf{E}(\hat{\theta}_j) - \theta_j)^2 = \mathbf{E}[(\hat{\theta}_j - \mathbf{E}(\hat{\theta}_j))^2] + \alpha_j^2.$$

Moreover,

$$\hat{\theta}_j - \mathbf{E}(\hat{\theta}_j) = \frac{1}{n} \sum_{i=1}^{n} \xi_i \varphi_j(i/n),$$

and, by Lemma 1.7,

$$\mathbf{E}[(\hat{\theta}_j - \mathbf{E}(\hat{\theta}_j))^2] = \frac{1}{n^2} \sum_{i=1}^{n} \varphi_j^2(i/n)\sigma_\xi^2 = \frac{\sigma_\xi^2}{n}.$$

∎

The quantities α_j in Proposition 1.16 are the residuals coming from the approximation of sums by integrals. We will see in the sequel that the contribution of these residuals is negligible with respect to the main terms of the squared risk on the Sobolev classes if n is large. Let us first give some bounds for α_j.

Lemma 1.8 *For the trigonometric basis* $\{\varphi_j\}_{j=1}^{\infty}$ *the residuals* α_j *are such that:*

(i) if $\sum_{j=1}^{\infty} |\theta_j| < \infty$, *then* $\max_{1\le j\le n-1} |\alpha_j| \le 2 \sum_{m=n}^{\infty} |\theta_m|$, *for all* $n \ge 2$;

(ii) if $\theta \in \Theta(\beta, Q)$, $\beta > 1/2$, *then* $\max_{1\le j\le n-1} |\alpha_j| \le C_{\beta,Q} n^{-\beta+1/2}$ *for all* $n \ge 2$ *and for a constant* $C_{\beta,Q} < \infty$ *depending only on* β *and* Q.

Proof. Using Lemma 1.7 we obtain, for $1 \le j \le n-1$,

$$\begin{aligned}
\alpha_j &= \frac{1}{n}\sum_{i=1}^{n} f(i/n)\varphi_j(i/n) - \theta_j \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{m=1}^{\infty} \theta_m \varphi_m(i/n)\right)\varphi_j(i/n) - \theta_j \\
&= \sum_{m=1}^{n-1} \theta_m \frac{1}{n}\sum_{i=1}^{n} \varphi_m(i/n)\varphi_j(i/n) - \theta_j \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\sum_{m=n}^{\infty} \theta_m \varphi_m(i/n)\varphi_j(i/n) \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{m=n}^{\infty} \theta_m \varphi_m(i/n)\varphi_j(i/n).
\end{aligned}$$

Thus,

$$|\alpha_j| = \left|\sum_{m=n}^{\infty} \theta_m \left(\frac{1}{n}\sum_{i=1}^{n} \varphi_m(i/n)\varphi_j(i/n)\right)\right| \le 2\sum_{m=n}^{\infty} |\theta_m|.$$

Assume now that $\theta \in \Theta(\beta, Q)$. Then

$$\begin{aligned}
\sum_{m=n}^{\infty} |\theta_m| &= \sum_{m=1}^{\infty} |\theta_m| I(m \ge n) \\
&\le \left(\sum_{m=1}^{\infty} a_m^2 \theta_m^2\right)^{1/2} \left(\sum_{m=n}^{\infty} a_m^{-2}\right)^{1/2} \\
&\le Q^{1/2}\Big(\sum_{m=n}^{\infty} (m-1)^{-2\beta}\Big)^{1/2} \le C_{\beta,Q} n^{-\beta+1/2}.
\end{aligned}$$

∎

Proposition 1.17 *Under Assumption (A) the risk of the projection estimator* $\hat{f}_{nN}$ *has the form*

$$\mathrm{MISE} = \mathbf{E}\|\hat{f}_{nN} - f\|_2^2 = \mathcal{A}_{nN} + \sum_{j=1}^{N} \alpha_j^2,$$

where

$$\mathcal{A}_{nN} = \frac{\sigma_\xi^2 N}{n} + \rho_N \quad \textit{with} \quad \rho_N = \sum_{j=N+1}^{\infty} \theta_j^2.$$

PROOF. Using the expansions $\hat{f}_{nN} = \sum_{j=1}^{N} \hat{\theta}_j \varphi_j$, $f = \sum_{j=1}^{\infty} \theta_j \varphi_j$ and part (ii) of Proposition 1.16 we obtain:

$$\begin{aligned}
\mathbf{E}\|\hat{f}_{nN} - f\|_2^2 &= \mathbf{E} \int_0^1 (\hat{f}_{nN}(x) - f(x))^2 dx \\
&= \mathbf{E} \int_0^1 \left(\sum_{j=1}^{N} (\hat{\theta}_j - \theta_j)\varphi_j(x) - \sum_{j=N+1}^{\infty} \theta_j \varphi_j(x) \right)^2 dx \\
&= \sum_{j=1}^{N} \mathbf{E}[(\hat{\theta}_j - \theta_j)^2] + \sum_{j=N+1}^{\infty} \theta_j^2 = \mathcal{A}_{nN} + \sum_{j=1}^{N} \alpha_j^2. \qquad \blacksquare
\end{aligned}$$

Theorem 1.9 *Suppose that Assumption (A) holds, $\beta \geq 1$, and $L > 0$. For $\alpha > 0$, define an integer as follows:*

$$N = \lfloor \alpha n^{\frac{1}{2\beta+1}} \rfloor.$$

Then the projection estimator $\hat{f}_{nN}$ satisfies:

$$\limsup_{n\to\infty} \sup_{f \in \tilde{W}(\beta,L)} \mathbf{E}_f \left[n^{\frac{2\beta}{2\beta+1}} \|\hat{f}_{nN} - f\|_2^2 \right] \leq C$$

where $C < \infty$ is a constant depending only on β, L, and α.

PROOF. By Proposition 1.17,

$$\mathbf{E}_f \|\hat{f}_{nN} - f\|_2^2 = \mathcal{A}_{nN} + \sum_{j=1}^{N} \alpha_j^2. \tag{1.93}$$

Assume that n is sufficiently large to satisfy $1 \leq N \leq n-1$. By Proposition 1.14, Lemma 1.8 and by the fact that $\beta \geq 1$, we obtain

$$\begin{aligned}
\sum_{j=1}^{N} \alpha_j^2 &\leq N \max_{1\leq j\leq n-1} \alpha_j^2 \leq C_{\beta,Q}^2 N n^{1-2\beta} \\
&= O\left(n^{\frac{1}{2\beta+1} - 2\beta + 1} \right) = O\left(n^{-\frac{2\beta}{2\beta+1}} \right),
\end{aligned} \tag{1.94}$$

where the $O(\cdot)$ terms are uniform in $f \in \tilde{W}(\beta, L)$. Therefore

$$\mathcal{A}_{nN} \leq \sigma_\xi^2 \alpha n^{-\frac{2\beta}{2\beta+1}} + \rho_N. \tag{1.95}$$

Finally, since the sequence a_j is monotone, we have

$$\rho_N = \sum_{j=N+1}^{\infty} \theta_j^2 \le \frac{1}{a_{N+1}^2} \sum_{j=1}^{\infty} a_j^2 \theta_j^2 \le \frac{Q}{a_{N+1}^2} = O\left(n^{-\frac{2\beta}{2\beta+1}}\right), \tag{1.96}$$

where the $O(\cdot)$ term is uniform in $f \in \tilde{W}(\beta, L)$. The theorem follows from (1.93)–(1.96). ∎

Remarks.

(1) It is easy to see that, for $\beta > 1$, formula (1.94) can be improved to $\sum_{j=1}^{N} \alpha_j^2 = o\left(n^{-\frac{2\beta}{2\beta+1}}\right)$. Thus, the residual term $\sum_{j=1}^{N} \alpha_j^2$ is negligible with respect to the upper bound on $\mathcal{A}_{nN}$ in Theorem 1.9. More accurate but technical calculations show that this is also true for $\beta = 1$ and for a much more general choice of N than that in Theorem 1.9. Therefore, the quantity $\mathcal{A}_{nN}$ constitutes the leading part of the MISE of the projection estimator $\hat{f}_{nN}$. The terms $\sigma_\xi^2 N/n$ and ρ_N appearing in the definition of $\mathcal{A}_{nN}$ are approximately the *variance term* and the *bias term*, respectively, in the L_2 risk of the estimator $\hat{f}_{nN}$. From the inequalities in (1.96) we obtain $\sup_{f \in \tilde{W}(\beta,L)} \rho_N \le C N^{-2\beta}$ for a constant C and any $N \ge 1$. Therefore, the choice $N \asymp n^{1/(2\beta+1)}$ used in Theorem 1.9 comes from minimization with respect to N of the upper bound on the maximum risk of $\hat{f}_{nN}$ on the class of functions $\tilde{W}(\beta, L)$.

(2) Theorem 1.9 states that if N is chosen optimally, the rate of convergence of the projection estimator $\hat{f}_{nN}$ in the L_2-norm over the Sobolev class $\tilde{W}(\beta, L)$ is

$$\psi_n = n^{-\frac{\beta}{2\beta+1}}.$$

So, we have again the same rate of convergence as for the Hölder class. Moreover, an analogous result is obtained if we replace $\tilde{W}(\beta, L)$ by $W(\beta, L)$ and choose a basis $\{\varphi_j\}$ different from the trigonometric one. We do not study this case here since it requires somewhat different tools.

(3) The random sequence $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_N, 0, 0, \dots)$ is an estimator of $\theta = (\theta_1, \theta_2, \dots) \in \ell^2(\mathbf{N})$. If we denote the norm of $\ell^2(\mathbf{N})$ by $\|\cdot\|$, then Theorem 1.9, Proposition 1.14, and the isometry between $\ell^2(\mathbf{N})$ and L_2 imply

$$\limsup_{n\to\infty} \sup_{\theta \in \Theta(\beta,Q)} \mathbf{E}\left[n^{\frac{2\beta}{2\beta+1}} \|\hat{\theta} - \theta\|^2\right] \le C < \infty.$$

1.7.3 Generalizations

We now briefly discuss some generalizations of the projection estimators $\hat{f}_{nN}$.

1. Nonparametric least squares estimators

So far we have studied a particular regression model with the regular design $X_i = i/n$, and the projection estimators have been constructed using the trigonometric basis. Suppose now that the values $X_i \in [0,1]$ are arbitrary and $\{\varphi_j\}$ is an arbitrary orthonormal basis in $L_2[0,1]$. Introduce the vectors $\theta = (\theta_1, \dots, \theta_N)^T$ and $\varphi(x) = (\varphi_1(x), \dots, \varphi_N(x))^T$, $x \in [0,1]$. The least squares estimator $\hat{\theta}^{LS}$ of the vector θ is defined as follows:

$$\hat{\theta}^{LS} = \arg\min_{\theta \in \mathbf{R}^N} \sum_{i=1}^{n} (Y_i - \theta^T \varphi(X_i))^2.$$

If the matrix

$$B = n^{-1} \sum_{i=1}^{n} \varphi(X_i)\, \varphi^T(X_i) \tag{1.97}$$

is invertible, we can write

$$\hat{\theta}^{LS} = B^{-1} \Big(\frac{1}{n} \sum_{i=1}^{n} Y_i \varphi(X_i) \Big).$$

Then the nonparametric least squares estimator of $f(x)$ is given by the formula:

$$\hat{f}^{LS}_{nN}(x) = \varphi^T(x) \hat{\theta}^{LS}.$$

If $\{\varphi_j\}_{j=1}^{\infty}$ is the trigonometric basis, $N \le n-1$ and $X_i = i/n$, then B reduces to the identity matrix of size N in view of Lemma 1.7. In this particular case the projection estimators and the nonparametric least squares estimators coincide: $\hat{f}^{LS}_{nN} = \hat{f}_{nN}$.

2. Weighted projection estimators

For a sequence of coefficients $\lambda = \{\lambda_j\}_{j=1}^{\infty} \in \ell^2(\mathbf{N})$ define the *weighted projection estimator* in the following way:

$$f_{n,\lambda}(x) = \sum_{j=1}^{\infty} \lambda_j\, \hat{\theta}_j\, \varphi_j(x). \tag{1.98}$$

Here, as before,

$$\hat{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} Y_i\, \varphi_j(X_i)$$

and the random series in (1.98) is interpreted in the sense of mean square convergence. The projection estimator $\hat{f}_{nN}$ studied so far is a particular example of $f_{n,\lambda}$ corresponding to the weights $\lambda_j = I(j \le N)$. From now on, we will call $\hat{f}_{nN}$ the *simple* projection estimator. Another example is given by the Pinsker-type weights that we will consider in Chapter 3:

$$\lambda_j = (1 - \kappa j^{\beta})_+,$$

where $\kappa > 0$, $\beta > 0$, and $a_+ = \max(a, 0)$. In these two examples, we have $\lambda_j \neq 0$ for a finite number of integers j only. If $\lambda_j \neq 0$ for all j, the estimator $f_{n,\lambda}$ cannot be computed from (1.98). We may then consider truncating the sum at sufficiently large values of j, for example, at $j = n$, and introduce the finite approximation

$$f_{n,\lambda}(x) = \sum_{j=1}^{n} \lambda_j \, \hat{\theta}_j \, \varphi_j(x). \tag{1.99}$$

Since the class of weighted projection estimators is wider than that of simple projection estimators, one can expect to have a smaller value of the mean integrated squared error for $f_{n,\lambda}$ (with an appropriate choice of λ) than for simple projection estimators (cf. Exercise 1.10 below).

The mean integrated squared error of estimator (1.99) has the following form:

$$\begin{aligned} \mathrm{MISE} &= \mathbf{E}_f \int_0^1 \left(\sum_{j=1}^{n} (\lambda_j \hat{\theta}_j - \theta_j) \varphi_j(x) - \sum_{j=n+1}^{\infty} \theta_j \varphi_j(x) \right)^2 dx \quad (1.100) \\ &= \mathbf{E}_f \Big[\sum_{j=1}^{n} (\lambda_j \hat{\theta}_j - \theta_j)^2 \Big] + \rho_n. \end{aligned}$$

The last expectation typically constitutes the leading term of the MISE, whereas ρ_n is asymptotically negligible. For example, if $f \in W^{per}(\beta, L)$, $\beta \ge 1$, we have

$$\rho_n = \sum_{j=n+1}^{\infty} \theta_j^2 = O(n^{-2\beta}) = O(n^{-2}).$$

3. Penalized least squares estimators

Penalized least squares (PLS) estimators provide a generalization of both nonparametric least squares and weighted projection estimators. A popular version of the PLS is given by the Tikhonov regularization. The coefficients $\hat{\theta}^{TR}$ of the *Tikhonov regularization estimators* are defined as a solution of the minimization problem:

$$\hat{\theta}^{TR} = \arg \min_{\theta \in \mathbf{R}^N} \Big\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - \theta^T \varphi(X_i))^2 + \sum_{j=1}^{N} b_j \theta_j^2 \Big\}$$

where b_j are some positive constants. Equivalently,

$$\hat{\theta}^{TR} = \Big(B + \mathrm{diag}(b_1, \dots, b_N)\Big)^{-1} \Big(\frac{1}{n} \sum_{i=1}^{n} Y_i \varphi(X_i)\Big)$$

where the matrix B is defined in (1.97) and $\mathrm{diag}(b_1, \dots, b_N)$ is the diagonal $N \times N$ matrix whose diagonal elements are $b_1, \dots, b_N$. Then the Tikhonov regularization estimator of the value of the regression function $f(x)$ is given by

$$\hat{f}_{nN}^{TR}(x) = \varphi^T(x) \hat{\theta}^{TR}.$$

If B is the identity matrix and $N = n$, the components of vector $\hat{\theta}^{TR}$ take the form

$$\hat{\theta}_j^{TR} = \frac{\hat{\theta}_j}{1 + b_j} = \frac{1}{n(1 + b_j)} \sum_{i=1}^{n} Y_i \, \varphi_j(X_i),$$

and $\hat{f}_{nN}^{TR}$ reduces to a weighted projection estimator

$$\hat{f}_{nN}^{TR}(x) = \sum_{j=1}^{N} \frac{\hat{\theta}_j \varphi_j(x)}{1 + b_j} \, .$$

In particular, if $b_j \sim j^{2\beta}$ for an integer β, this estimator is approximately equivalent to the spline estimator (cf. Exercise 1.11, which considers the case $\beta = 2$).

Another important member of the PLS family is the ℓ^1-penalized least squares, or the *Lasso estimator*. Its coefficients are defined as a solution of the minimization problem:

$$\hat{\theta}^L = \arg \min_{\theta \in \mathbf{R}^N} \Big\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - \theta^T \varphi(X_i))^2 + \sum_{j=1}^{N} b_j |\theta_j| \Big\} \, .$$

For large N, the computation of Tikhonov estimators becomes problematic, since it involves inversion of an $N \times N$ matrix. On the other hand, the Lasso estimator remains numerically feasible for dimensions N that are much larger than the sample size n.

1.8 Oracles

Several examples of oracles have been already discussed in this chapter. Our aim now is to give a general definition that we will use in Chapter 3.

We start by considering the projection estimator of regression $\hat{f}_{nN}$. Recall that $\hat{f}_{nN}$ is entirely determined by the integer tuning parameter N. Therefore, it is interesting to choose N in an optimal way. Since we study $\hat{f}_{nN}$ under the L_2-risk, the optimal N is naturally defined by

$$N_n^* = \arg\min_{N\geq 1} \mathbf{E}_f \|\hat{f}_{nN} - f\|_2^2.$$

Unfortunately, the value $N_n^* = N_n^*(f)$ depends on the unknown function f, and thus it is not accessible. For the same reason $\hat{f}_{nN_n^*}$ is not an estimator: it depends on the unknown function f. We will call $\hat{f}_{nN_n^*}$ the *oracle*. This is the "best forecast" of f, which is, however, inaccessible: in order to construct it, we would need an "oracle" that knows f. Since we deal with projection estimators, we call $\hat{f}_{nN_n^*}$ more specifically the *projection oracle*. In the same way we can define oracles for other classes of nonparametric estimators: we have already done this above (cf. (1.57)). Let us now give a general definition of the oracle.

Assume that we would like to estimate a parameter θ in a statistical model $\{P_\theta, \theta \in \Theta\}$ where Θ is an arbitrary set and P_θ is a probability measure indexed by $\theta \in \Theta$. For example, θ may be the regression function f, Θ may be a Sobolev class, and P_θ may be the distribution of the vector $(Y_1, \ldots, Y_n)$ in the regression model (1.69). Suppose also that we have a family of estimators $\hat{\theta}_\tau$ of θ indexed by $\tau \in \mathcal{T}$:

$$\mathcal{K} = \{\hat{\theta}_\tau, \tau \in \mathcal{T}\}$$

where $\mathcal{T}$ is an arbitrary set and $\hat{\theta}_\tau$ takes values in a set Θ' such that $\Theta \subseteq \Theta'$. Usually τ is interpreted as a smoothing parameter and $\mathcal{T}$ as the set of possible values of τ. For example, $\hat{\theta}_\tau$ may be the kernel estimator with a fixed kernel and bandwidth $\tau = h$. Then it is natural to take $\mathcal{T} = \{h : h > 0\}$. Another example is given by the projection estimator; in this case we have $\tau = N$ and $\mathcal{T} = \{1, 2, \ldots\}$.

Introduce a risk function $r : \Theta' \times \Theta \to [0, \infty)$ such that $r(\hat{\theta}_\tau, \theta)$ characterizes the error of estimation of θ by $\hat{\theta}_\tau$. Two typical examples of $r(\cdot, \cdot)$ are the mean squared error MSE and the mean integrated squared error MISE.

Assume that for any $\theta \in \Theta$ there exists an optimal value $\tau^*(\theta)$ of the parameter τ such that

$$r(\hat{\theta}_{\tau^*(\theta)}, \theta) = \min_{\tau\in\mathcal{T}} r(\hat{\theta}_\tau, \theta). \tag{1.101}$$

Observe that $\hat{\theta}_{\tau^*(\theta)}$ is not a statistic since it depends on the unknown parameter θ.

Definition 1.13 *Assume that the class of estimators $\mathcal{K}$ is such that for any $\theta \in \Theta$ there exists a value $\tau^*(\theta) \in \mathcal{T}$ satisfying (1.101). Then the random function $\theta \mapsto \hat{\theta}_{\tau^*(\theta)}$ is called the* **oracle for $\mathcal{K}$ with respect to the risk** $r(\cdot, \cdot)$.

Let us emphasize that the oracle is determined not only by the class of estimators under consideration, but also by the choice of the risk (MSE or MISE, for example).

Instead of minimizing the exact risk as in (1.101), it is sometimes convenient to minimize an asymptotic approximation of the risk, as the sample size n tends to infinity. For example, Proposition 1.17 and Remark (1) after

Theorem 1.9 suggest that for the simple projection estimator the value $\mathcal{A}_{nN}$ constitutes the leading term of the risk

$$r(\hat{f}_{nN}, f) \stackrel{\triangle}{=} \mathbf{E}_f \|\hat{f}_{nN} - f\|_2^2$$

as $n \to \infty$. Therefore, instead of the exact oracle N_n^*, it makes sense to consider an approximate oracle that minimizes $\mathcal{A}_{nN}$. Since $\mathcal{A}_{nN} \to \infty$ as $N \to \infty$ for any fixed n, there always exists a minimizer of $\mathcal{A}_{nN}$:

$$\tilde{N}_n = \arg\min_{N \geq 1} \mathcal{A}_{nN}.$$

Then an approximate oracle can be defined as $\hat{f}_{n\tilde{N}_n}$.

An important question is the following: Can we construct an estimator f_n^* such that

$$\mathbf{E}_f \|f_n^* - f\|_2^2 \leq \mathbf{E}_f \|\hat{f}_{nN_n^*} - f\|_2^2 (1 + o(1)), \quad n \to \infty, \tag{1.102}$$

for any f in a sufficiently large class of functions? In other words, can we conceive a true estimator that mimics the asymptotic behavior of the oracle $\hat{f}_{nN_n^*}$? We will see in Chapter 3 that the answer to this question is positive for a model that is close to the regression model considered here. Such estimators f_n^* will be called adaptive to the oracle, in a precise sense defined in Chapter 3. Inequalities of the form (1.102) are known under the name of *oracle inequalities*. Construction of adaptive estimators is often based on the idea of unbiased risk estimation. The next section explains how to apply this idea in the problem of nonparametric regression.

1.9 Unbiased risk estimation for regression

In Section 1.4 we used unbiased estimation of the risk to obtain data-driven bandwidth selectors for the kernel density estimator. Similar methods exist for regression estimators, and we are going to describe some of them in this section. For example, they can be used to select the bandwidth h of the local polynomial estimator or the order N of the projection estimator. However, for the regression model, only approximately unbiased estimators of the MISE are, in general, available, with an approximation error due to the discreteness of the design. On the other hand, we can get exactly unbiased estimators of a discretized version of the MISE.

Consider the regression model (1.69). Let $\{f_\tau, \tau \in \mathcal{T}\}$ be a family of estimators based on the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ and depending on a parameter $\tau \in \mathcal{T}$. The dependence of f_τ on n is skipped for brevity. We assume that f_τ is entirely determined by $(X_1, Y_1), \ldots, (X_n, Y_n)$ and τ. Define a discretized version of the MISE by

$$r_{n,\tau}^D(f) = \mathbf{E}_f \|f_\tau - f\|_{2,n}^2$$

where

$$\|f_\tau - f\|_{2,n} \stackrel{\triangle}{=} \left(\frac{1}{n}\sum_{i=1}^{n}(f_\tau(X_i) - f(X_i))^2\right)^{1/2}.$$

Let f_τ be a linear nonparametric regression estimator indexed by τ, i.e.,

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

where the weights $W_{ni}(x,\tau) = W_{ni}(x,\tau,X_1,\ldots,X_n)$ depend only on n, i, τ, x and on the observations $X_1,\ldots,X_n$.

Throughout this section we will assume that

$$\mathbf{E}_f(\xi_i|X_1,\ldots,X_n) = 0 \quad \text{and} \quad \mathbf{E}_f(\xi_i\xi_k|X_1,\ldots,X_n) = \sigma^2\delta_{jk} \tag{1.103}$$

for $i,k = 1,\ldots,n$, where $\xi_i = Y_i - f(X_i)$. Note that

$$r^D_{n,\tau}(f) = \mathbf{E}_f\left[\|f_\tau\|^2_{2,n} - \frac{2}{n}\sum_{i=1}^{n} f_\tau(X_i)f(X_i)\right] + \mathbf{E}_f\left[\,\|f\|^2_{2,n}\right].$$

Since the value $\|f\|^2_{2,n}$ does not depend on τ, the minimizer of $r^D_{n,\tau}(f)$ in $\tau \in \mathcal{T}$ also minimizes the function

$$J(\tau) \stackrel{\triangle}{=} \mathbf{E}_f\left[\|f_\tau\|^2_{2,n} - \frac{2}{n}\sum_{i=1}^{n} f_\tau(X_i)f(X_i)\right].$$

We now look for an unbiased estimator of $J(\tau)$. A trivial unbiased estimator of $\mathbf{E}_f\left[\,\|f_\tau\|^2_{2,n}\right]$ being $\|f_\tau\|^2_{2,n}$, it remains to find an unbiased estimator of

$$\mathbf{E}_f\left[\frac{2}{n}\sum_{i=1}^{n} f_\tau(X_i)f(X_i)\right].$$

Such an estimator can be obtained in the form

$$\hat{G} = \frac{2}{n}\sum_{i=1}^{n} Y_i f_\tau(X_i) - \frac{2\sigma^2}{n}\sum_{i=1}^{n} W_{ni}(X_i,\tau).$$

Indeed, conditioning on $X_1,\ldots,X_n$ we find

$$\begin{aligned}
\mathbf{E}_f\left[\sum_{i=1}^{n} Y_i f_\tau(X_i)\Big|X_1,\ldots,X_n\right] &- \sum_{i=1}^{n} f_\tau(X_i)f(X_i) \\
&= \mathbf{E}_f\left[\sum_{i=1}^{n}\xi_i f_\tau(X_i)\Big|X_1,\ldots,X_n\right] \\
&= \mathbf{E}_f\left[\sum_{i=1}^{n}\xi_i\sum_{k=1}^{n}\xi_k W_{nk}(X_i,\tau)\Big|X_1,\ldots,X_n\right] \\
&= \sigma^2\sum_{i=1}^{n} W_{ni}(X_i,\tau)
\end{aligned}$$

and therefore, after taking expectations with respect to $X_1, \ldots, X_n$ we find

$$\mathbf{E}_f(\hat{G}) = \mathbf{E}_f \left[\frac{2}{n} \sum_{i=1}^{n} f_\tau(X_i) f(X_i) \right] .$$

Consequently,

$$\hat{J}(\tau) = \|f_\tau\|^2_{2,n} - \frac{2}{n} \sum_{i=1}^{n} Y_i f_\tau(X_i) + \frac{2\sigma^2}{n} \sum_{i=1}^{n} W_{ni}(X_i, \tau)$$

is an unbiased estimator of $J(\tau)$. Define now the C_p-*criterion*:

$$C_p(\tau) \stackrel{\triangle}{=} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_\tau(X_i))^2 + \frac{2\sigma^2}{n} \sum_{i=1}^{n} W_{ni}(X_i, \tau) .$$

Using the relation $\mathbf{E}_f[\hat{J}(\tau)] = J(\tau)$ and (1.103) we get

$$\mathbf{E}_f[C_p(\tau)] = r^D_{n,\tau}(f) + \sigma^2. \tag{1.104}$$

Thus, $C_p(\tau)$ yields an unbiased estimator of the discretized MISE $r^D_{n,\tau}$, up to a shift σ^2, which does not depend on τ. This suggests to approximate the minimizers of $r^D_{n,\tau}$ by those of the C_p-criterion:

$$\hat{\tau} = \arg\min_{\tau \in \mathcal{T}} C_p(\tau),$$

which provides a data-driven choice of parameter τ, since the function $C_p(\cdot)$ can be computed from the data. The C_p-estimator of the regression function is then defined as $f_{\hat{\tau}}$.

Consider now some examples. For the orthogonal series (projection) regression estimators $\hat{f}_{nN}$, we take $\tau = N$ and define the weights $W_{ni}(x, \tau)$ by the formula (cf. (1.88)):

$$W_{ni}(x, \tau) = \frac{1}{n} \sum_{j=1}^{N} \varphi_j(X_i) \varphi_j(x).$$

Then

$$\sum_{i=1}^{n} W_{ni}(X_i, \tau) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{N} \varphi_j^2(X_i),$$

so that the C_p-criterion for the projection regression estimators takes the form

$$C_p(N) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{f}_{nN}(X_i))^2 + \frac{2\sigma^2}{n} \sum_{j=1}^{N} \|\varphi_j\|^2_{2,n} .$$

If $\{\varphi_j\}$ is the trigonometric basis and $X_i = i/n$, for $N \le n-1$, we have $\|\varphi_j\|_{2,n}^2 = 1$ (cf. Lemma 1.7), and the C_p-criterion can be written in a particularly simple form:

$$C_p(N) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_{nN}(X_i))^2 + \frac{2\sigma^2 N}{n}\,. \tag{1.105}$$

As a second example, consider the kernel regression estimator $\bar{f}_{nh}$ defined in (1.62). Then $\tau = h$, the weights $W_{ni}(x,\tau)$ are given by

$$W_{ni}(x,\tau) = \frac{1}{nh}K\left(\frac{X_i - x}{h}\right),$$

and the C_p-criterion takes the form

$$C_p(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{f}_{nh}(X_i))^2 + \frac{2\sigma^2 K(0)}{nh}\,.$$

We finally discuss the cross-validation techniques. The leave-one-out cross-validation criterion for regression is defined by

$$CV_*(\tau) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f_{\tau,-i}(X_i))^2.$$

Here $f_{\tau,-i}$ is the estimator of the same form as f_τ based on the sample $(X_1,Y_1),\dots,(X_{i-1},Y_{i-1}),(X_{i+1},Y_{i+1}),\dots,(X_n,Y_n)$, with the observation (X_i,Y_i) left out. Assume that

$$\int f_{\tau,-i}^2(x)P_X(dx) < \infty, \tag{1.106}$$

where P_X is the marginal distribution of X. Then, under the assumptions (1.103) we easily get that $\mathbf{E}_f[\xi_i f_{\tau,-i}(X_i)] = 0$, and

$$\begin{aligned}\mathbf{E}_f\big[(Y_i - f_{\tau,-i}(X_i))^2\big] &= \mathbf{E}_f\big[(f_{\tau,-i}(X_i) - f(X_i))^2\big] \\ &\quad + 2\mathbf{E}_f\big[\xi_i(f_{\tau,-i}(X_i) - f(X_i))\big] + \sigma^2 \\ &= \mathbf{E}_f\big[(f_{\tau,-i}(X_i) - f(X_i))^2\big] + \sigma^2,\end{aligned} \tag{1.107}$$

so that

$$\mathbf{E}_f[CV_*(\tau)] = \mathbf{E}_f\left[\frac{1}{n}\sum_{i=1}^{n}(f_{\tau,-i}(X_i) - f(X_i))^2\right] + \sigma^2.$$

We see that the cross-validation criterion does not provide an unbiased estimator even for $r_{n,\tau}^D$ (the discretized version of the MISE). In order to justify that CV_* is a meaningful criterion, we would need to show that

$$\mathbf{E}_f\left[\frac{1}{n}\sum_{i=1}^{n}(f_{\tau,-i}(X_i)-f(X_i))^2\right]\approx \mathbf{E}_f\|f_\tau - f\|^2_{2,n}\,,$$

where the approximation is understood in a suitable sense. This would require more conditions and could be achieved only in specific contexts. A more general result is obtained if we modify the risk by passing to a weighted MISE,

$$r_{n-1,\tau}(f)\stackrel{\triangle}{=}\mathbf{E}_f\int(f_{\tau,-i}(x)-f(x))^2P_X(dx),$$

and assume that the pairs (X_i, Y_i) are i.i.d. and that $f_{\tau,-i}(x)$ has the same distribution as $f_{\tau,-1}(x)$ for all i, x. This assumption is satisfied for some examples. Then from (1.107) we get

$$\begin{aligned}\mathbf{E}_f\left[(Y_i-f_{\tau,-i}(X_i))^2\right] &= \mathbf{E}_f[(f_{\tau,-1}(X_1)-f(X_1))^2]+\sigma^2\\ &= r_{n-1,\tau}(f)+\sigma^2,\end{aligned}$$

so that

$$\mathbf{E}_f[CV_*(\tau)] = r_{n-1,\tau}(f)+\sigma^2. \tag{1.108}$$

Therefore, for the regression model with random design (i.i.d. observations) the cross-validation criterion $CV_*(\tau)$ yields an unbiased estimator of the risk $r_{n-1,\tau}(f)$, up to a constant shift σ^2. Note that this result is valid for estimators f_τ that are not necessarily linear, but such that $f_{\tau,-i}$ has the same distribution as $f_{\tau,-1}$. On the other hand, the pairs (X_i, Y_i) should be i.i.d., which is not a necessary requirement for the unbiased estimation of the discretized MISE via the C_p-criterion.

1.10 Three Gaussian models

In this chapter we have studied only two statistical models: the model of density estimation and that of nonparametric regression. Recall that in Section 1.1 we also introduced the third one, namely the Gaussian white noise (GWN) model. It is often defined in a slightly more general form than in Section 1.1:

$$dY(t) = f(t)dt+\varepsilon dW(t),\quad t\in[0,1]. \tag{1.109}$$

Here $0<\varepsilon<1$, $f:[0,1]\to\mathbf{R}$ and $W(\cdot)$ is the standard Wiener process on [0,1]. We mentioned in Section 1.1 that for $\varepsilon=1/\sqrt{n}$ this is an "ideal" model that gives a suitable approximation of nonparametric regression. Our aim here is to explain this remark and to go a bit further. More specifically, we will argue that the following three Gaussian models are closely related to each other: the Gaussian white noise model, the Gaussian nonparametric regression and the Gaussian sequence model. We will see that the study of these models

is essentially the same, up to a control of asymptotically negligible residual terms. For this reason we will consider in Chapter 3 only the two technically simplest models: the Gaussian sequence model and the GWN one. This will allow us to reduce the technicalities and to focus on the main ideas. The results of Chapter 3, with suitable modifications, are also valid for the regression model but this material is left beyond the scope of the book.

1. Connection between Gaussian white noise model and nonparametric regression

Suppose that we observe the process Y in the Gaussian white noise model (1.109). Let us now discretize (1.109) as follows. Integrating over $[t, t+\Delta]$ where $\Delta > 0$ we get

$$\frac{Y(t+\Delta) - Y(t)}{\Delta} = \frac{1}{\Delta}\int_t^{t+\Delta} f(s)ds + \frac{\varepsilon}{\Delta}(W(t+\Delta) - W(t)).$$

Define

$$y(t) \stackrel{\triangle}{=} \frac{Y(t+\Delta) - Y(t)}{\Delta}, \quad \xi(t) \stackrel{\triangle}{=} \frac{\varepsilon}{\Delta}(W(t+\Delta) - W(t)).$$

For any $t \in [0,1]$ the random variable $\xi(t)$ is Gaussian with mean zero and variance

$$\mathbf{E}(\xi^2(t)) = \frac{\varepsilon^2}{\Delta^2}\mathbf{E}[(W(t+\Delta) - W(t))^2] = \frac{\varepsilon^2}{\Delta}.$$

Take now $\varepsilon = 1/\sqrt{n}$ and $\Delta = 1/n$. Then for all t we have $\xi(t) \sim \mathcal{N}(0,1)$ and

$$y(t) \approx f(t) + \xi(t),$$

where the symbol $\approx$ denotes equality up to the deterministic residual

$$\frac{1}{\Delta}\int_t^{t+\Delta} f(s)ds - f(t),$$

which is small for sufficiently small Δ and sufficiently smooth f. In particular, for $Y_i = y(i/n)$ and $\xi_i = \xi(i/n)$ we have

$$Y_i \approx f(i/n) + \xi_i.$$

We recognize the nonparametric regression model with regular design and i.i.d. errors ξ_i distributed according to $\mathcal{N}(0,1)$. Thus, the two models under consideration are closely related to each other. We used here only heuristic arguments but they can be turned into a rigorous proof.

2. Connection between Gaussian white noise model and Gaussian sequence model

Suppose again that we observe the process Y in the Gaussian white noise model. Let $\{\varphi_j\}_{j=1}^{\infty}$ be an orthonormal basis in $L_2[0,1]$. Then (1.109) implies that

$$\int_0^1 \varphi_j(t)dY(t) = \theta_j + \varepsilon \int_0^1 \varphi_j(t)dW(t) \quad \text{with} \quad \theta_j = \int_0^1 f(t)\varphi_j(t)dt.$$

Define

$$y_j \triangleq \int_0^1 \varphi_j(t)dY(t), \quad \xi_j \triangleq \int_0^1 \varphi_j(t)dW(t).$$

Since the functions φ_j are orthonormal in $L_2[0,1]$, the variables ξ_j are i.i.d. with distribution $\mathcal{N}(0,1)$. Therefore, observing a continuous process Y in the Gaussian white noise model (1.109) the statistician has access to the following infinite sequence of Gaussian observations:

$$y_j = \theta_j + \varepsilon\xi_j, \quad j = 1, 2, \ldots. \tag{1.110}$$

Formula (1.110) defines the *Gaussian sequence model*.

Estimation of $f \in L_2[0,1]$ in Gaussian white noise model (1.109) is equivalent to estimation of the sequence $\{\theta_j\}_{j=1}^{\infty}$ of its Fourier coefficients. Thus, it is sufficient to consider estimation of θ_j in the model (1.110). In particular, y_j is an unbiased estimator of θ_j. One can consider y_j as an analog of the unbiased estimator $\hat{\theta}_j$ of θ_j in the regression model. In the spirit of (1.98), we can define the weighted projection estimator of f (called the *linear estimator* of f):

$$f_{\varepsilon,\lambda}(x) = \sum_{j=1}^{\infty} \lambda_j y_j \varphi_j(x), \tag{1.111}$$

where $\lambda = \{\lambda_j\}_{j=1}^{\infty}$ is a sequence belonging to $\ell^2(\mathbf{N})$; the series in (1.111) is interpreted in the sense of mean square convergence. The statistic $\lambda_j y_j$ is a linear estimator of θ_j.

The mean squared risk of $f_{\varepsilon,\lambda}$ is

$$\begin{aligned} \mathrm{MISE} = \mathbf{E}_f \|f_{\varepsilon,\lambda} - f\|_2^2 &= \sum_{j=1}^{\infty} \mathbf{E}_f \left[(\lambda_j y_j - \theta_j)^2\right] \\ &= \sum_{j=1}^{\infty} [(1-\lambda_j)^2\theta_j^2 + \varepsilon^2\lambda_j^2] \triangleq R(\lambda,\theta). \end{aligned} \tag{1.112}$$

Minimizing this expression with respect to the weights λ_j we obtain

$$\min_{\lambda \in \ell^2(\mathbf{N})} R(\lambda,\theta) = R(\lambda^*,\theta) = \sum_{j=1}^{\infty} \frac{\varepsilon^2\theta_j^2}{\varepsilon^2 + \theta_j^2}\,, \tag{1.113}$$

with the optimal weights $\lambda^* = \{\lambda_j^*\}_{j=1}^{\infty}$ given by

$$\lambda_j^* = \frac{\theta_j^2}{\varepsilon^2 + \theta_j^2} \ . \tag{1.114}$$

Finally, $f_{\varepsilon,\lambda^*}$ is the corresponding oracle called the *linear oracle*. Note that the expressions for the oracle risk (1.113) and oracle weights (1.114) can be viewed as analogs of those obtained in (1.44) and (1.43), respectively, for the problem of density estimation.

3. Connection between nonparametric regression and Gaussian sequence model

Suppose now that we observe $Y_1, \dots, Y_n$ in the nonparametric regression model

$$Y_i = f(i/n) + \xi_i, \qquad i = 1, \dots, n, \tag{1.115}$$

where ξ_i are i.i.d. random variables distributed according to $\mathcal{N}(0,1)$. Let $\{\varphi_j\}_{j=1}^{\infty}$ be the trigonometric basis or any other basis satisfying (1.92). Set

$$\hat{\theta}_j = n^{-1} \sum_{i=1}^{n} Y_i \varphi_j(i/n),$$
$$f_j = n^{-1} \sum_{i=1}^{n} f(i/n) \varphi_j(i/n),$$
$$\eta_j = \sum_{i=1}^{n} \xi_i \varphi_j(i/n) / \sqrt{n},$$

and $\varepsilon = 1/\sqrt{n}$. Then (1.115) implies

$$\hat{\theta}_j = f_j + \varepsilon \eta_j, \qquad j = 1, \dots, n,$$

which is close to the Gaussian sequence model (1.110) since the random variables η_j are i.i.d. with distribution $\mathcal{N}(0,1)$. A difference from (1.110) is in the fact that here we deal with a finite sequence $\{f_j\}_{j=1}^{n}$ of dimension n and f_j are not the true Fourier coefficients but rather their approximations. However, there is no significant asymptotic difference from the Gaussian sequence model as $n \to \infty$. For example, if $\{\varphi_j\}_{j=1}^{\infty}$ is the trigonometric basis, Lemma 1.8 yields that the residuals $\alpha_j = f_j - \theta_j$ are sufficiently small, so that we approximately have

$$\hat{\theta}_j \approx \theta_j + \varepsilon \eta_j, \qquad j = 1, \dots, n,$$

where $\theta_j = \int_0^1 f \varphi_j$. If we set here $y_j = \hat{\theta}_j$ we get a truncated version of model (1.110), up to small residual terms.

Similarly to (1.111), a linear estimator of regression function f can now be defined in the form

$$f_{n,\lambda}(x) = \sum_{j=1}^{n} \lambda_j \hat{\theta}_j \varphi_j(x),$$

where $\{\lambda_j\}_{j=1}^{\infty} \in \ell^2(\mathbf{N})$. This is exactly the weighted projection estimator (1.99).

1.11 Notes

The literature on kernel density estimation is very extensive. Some basic ideas can be traced back to Fix and Hodges (1951) and Akaike (1954). Influential papers of Rosenblatt (1956) and Parzen (1962) initiated the mathematical theory and stimulated further interest to the subject. For an overview of the literature on kernel density estimation we refer to the books of Devroye and Györfi (1985), Silverman (1986), Devroye (1987), Scott (1992), Wand and Jones (1995), Hart (1997), and Devroye and Lugosi (2000).

A detailed account on orthogonal polynomials is given by Szegö (1975). The derivation of the Epanechnikov kernel from optimization arguments is due to Bartlett (1963) and Epanechnikov (1969). Hodges and Lehmann (1956) did it even earlier, although not in the context of density estimation. A short proof implying that the Epanechnikov kernel minimizes (1.23) in $K \geq 0$ is given, e.g., by Devroye and Györfi (1985), Lemma 18 of Chapter 5. The approach to optimality based on asymptotics of the risk for fixed density dates back to Bartlett (1963) and Epanechnikov (1969). The inconsistency of this approach was brought to light as late as in the 1990s (cf. Brown et al. (1997), Johnstone (1998)).

The notions of Fourier analysis used in Section 1.3 can be found in standard textbooks, for instance, in Katznelson (2004) or Folland (1999). Fourier analysis of kernel density estimators was used already by Parzen (1962). The formula for the exact MISE (1.41) is due to Watson and Leadbetter (1963). They also obtained the expressions (1.43) and (1.44) for the kernel oracle and its risk. Admissibility has been studied by Cline (1988) within a more general class of kernels than in Definition 1.6. In particular, he showed that asymmetric and multimodal kernels are inadmissible. For the equivalence of conditions (1.51) and (1.52) when β is an integer see Malliavin (1995), Section 3.5, or Folland (1999), Section 9.3. The sinc kernel density estimator dates back to Konakov (1972); see also Davis (1975), who calls it the Fourier integral estimator. Various examples of superkernels are given in Chapter 5 of Devroye and Györfi (1985) and in Devroye (1987).

Cross-validation in the form considered in Section 1.4 was first suggested by Rudemo (1982). Stone (1984) proved that the integrated squared error of the estimator $\hat{p}_{n,CV}$ is asymptotically equivalent to that of the kernel oracle

with bandwidth h_{id} defined in (1.57). A similar property is established in the form of oracle inequality by Dalelane (2005). Analogous results hold for the data-driven kernel estimator whose bandwidth minimizes the Fourier-based unbiased criterion (1.58) (cf. Golubev (1992)).

The Nadaraya-Watson estimator is proposed by Nadaraya (1964) and Watson (1964). An overview of the literature on this estimator and on its modifications can be found, for example, in the books of Härdle (1990), Wand and Jones (1995), Hart (1997), and Györfi et al. (2002).

Local polynomial fitting has a long history: It was used in the analysis of time series as early as in the 1930s. Stone (1977) was the first to invoke local polynomials in the context of nonparametric regression. He considered local linear estimators with nearest neighbor weights. The now common Definition 1.8 of local polynomial estimator appeared in Katkovnik (1979). Stone (1980, 1982) established rates of convergence of LP(ℓ) estimators with rectangular kernel for regression with random design. For general LP(ℓ) estimators and their robust versions, asymptotics of the MSE and rates of convergence on the Hölder classes were obtained in Tsybakov (1986); see also Korostelev and Tsybakov (1993). Local polynomial estimators are discussed in the books by Wand and Jones (1995), Fan and Gijbels (1996), Loader (1999), and Györfi et al. (2002).

The idea of projection (orthogonal series) estimation belongs to Čencov (1962), who introduced the orthogonal series estimators of a probability density and studied their rates of convergence in L_2. Orthogonal series density estimation is discussed in detail in the books by Čencov (1972), Devroye and Györfi (1985), Efromovich (1999), and Massart (2007). Projection estimators of nonparametric regression started receiving attention only from the 1980s. Important early references are Shibata (1981) and Rice (1984). The model in Rice (1984) is the same as in Section 1.7.2: regression under regular design and (weighted) projection estimators with the trigonometric basis. Projection estimators in regression and in the Gaussian white noise model are discussed in the books by Eubank (1988), Efromovich (1999), Wasserman (2006), and Massart (2007). The literature on projection estimators has been rapidly growing since the 1990s, boosted by the invention of wavelets by Meyer (cf. Meyer (1990)). For a detailed account on wavelet bases we refer to the books by Hernández and Weiss (1996) and Härdle et al. (1998). Modifying the function ψ leads to wavelet bases with different approximation properties. An overview and references on statistical properties of wavelet estimators can be found in Johnstone (1998), Härdle et al. (1998), and in Chapter 18 of Györfi et al. (2002).

A more general version of the material of Section 1.7.2 (cf. Remark (1) after Theorem 1.9) is given in Polyak and Tsybakov (1990). A key technical fact is that Lemma 1.7 extends to $j, k \geq n$ modulo small correction terms.

Nemirovskii et al. (1983, 1984, 1985) studied the convergence rates of nonparametric least squares estimators on the L_p Sobolev classes of functions. A survey of more recent work on nonparametric least squares estimators can

be found, for example, in van de Geer (2000), Györfi et al. (2002), and Baraud (2002). These estimators have nice MISE properties for the regression model with random design where the study of local polynomial estimators is more involved and needs additional assumptions.

For the connection between Tikhonov regularization and spline smoothing we refer to the books by Eubank (1988) and Wahba (1990). An analysis of the convergence rates of spline estimators can be found, for example, in Speckman (1985), and Golubev and Nussbaum (1992).

Rates of convergence and oracle inequalities for the ℓ^1-penalized least squares are given by Bickel et al. (2007), Bunea et al. (2007a,b), Koltchinskii (2008), and van de Geer (2008).

The words "oracle" and "oracle inequalities" were brought into use by Donoho and Johnstone in the 1990s (cf. Johnstone (1998)).

The idea of unbiased risk estimation can be traced back to Akaike (1969) and Mallows (1973), who both considered the choice of integer τ (the dimension) in parametric models. Stein (1981) developed a method of unbiased estimation of the risk for a rather general class of estimators in Gaussian shift models (cf. Section 3.4). The C_p-criterion is due to Mallows (1973). There is a whole family of closely related criteria. Akaike's information criterion (AIC) in its general form is applicable for any parametric model where the number N of parameters is to be estimated (cf. Akaike (1974)). The AIC is defined as follows: Choose N to minimize $-2(\mathcal{L}_N - N)$ where $\mathcal{L}_N$ is the maximal value of the log-likelihood for the model with N parameters. We mention here two particular cases of the AIC. In the first case, the log-likelihood is computed for the Gaussian linear regression model with N parameters and unknown variance of the noise. Then the AIC reduces to minimization in N of the residual sum of squares multiplied by $\exp(2N/n)$. In the context of Section 1.9, this version of the AIC leads to the choice of N that minimizes

$$\mathrm{AIC}(N) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_{nN}(X_i))^2 \exp(2N/n). \tag{1.116}$$

The second example of the AIC is obtained if we consider the log-likelihood of the Gaussian linear model with known variance of the noise σ^2. Then the AIC coincides with the C_p-criterion. Note that the paper of Akaike (1974) does not mention this fact. Moreover, Akaike (1974) criticizes the C_p of Mallows because it requires the knowledge of σ^2.

More generally, we can consider a family of criteria

$$C(N) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_{nN}(X_i))^2 \nu(2N/n)$$

where $\nu(\cdot)$ is a monotone increasing function on $[0, \infty)$ such that $\nu(0) = 1$ and $\lim_{t\to 0}(\nu(t) - 1)/t = 1$ (cf. Polyak and Tsybakov (1992)). For $\nu(t) = \exp(t)$ we get the AIC. Other famous examples are $\nu(t) = 1 + t$, yielding

Shibata's criterion (cf. Shibata (1981)); $\nu(t) = 1/(1-t/2)^2$, corresponding to the GCV (*Generalized cross-validation criterion*, Craven and Wahba (1979)); and $\nu(t) = (1+t/2)/(1-t/2)$, corresponding to the FPE (*Final prediction error criterion*, Akaike (1969)). They can be compared with the C_p-criterion (1.105). For instance, Shibata's criterion can be viewed as an analog of (1.105) where the unknown σ^2 is estimated by the residual sum of squares $\frac{1}{n}\sum_{i=1}^n (Y_i - \hat{f}_{nN}(X_i))^2$. These criteria can be extended to general linear estimators of regression. For example, in the notation of Section 1.9, the GCV criterion for an arbitrary linear estimator f_τ is defined in the following form: choose τ that minimizes

$$\mathrm{GCV}(\tau) = \frac{1}{n}\sum_{i=1}^n (Y_i - f_\tau(X_i))^2 \left(1 - \frac{1}{n}\sum_{i=1}^n W_{ni}(X_i,\tau)\right)^{-2}.$$

More details about these and some other related criteria are given, for example, in the books by McQuarrie and Tsai (1998) and Ruppert et al. (2003).

The Gaussian white noise model and the Gaussian sequence model were first introduced in the context of nonparametric estimation by Ibragimov and Has'minskii in the 1970s (cf. Ibragimov and Has'minskii (1977, 1981)). The importance of these models is motivated by the equivalence arguments that were, however, not properly formalized until the late 1990s. Section 1.10 gives a sketch of such arguments. They reflect, in a very heuristic manner, the property of equivalence of experiments in the sense of Le Cam (cf. Le Cam and Yang (2000)). Brown and Low (1996) give a rigorous proof of the equivalence of nonparametric regression and Gaussian white noise models. An extension covering the multivariate case and random design regression was recently obtained by Reiss (2008). Nussbaum (1996) showed that, under suitable conditions, the density estimation model is equivalent to a Gaussian diffusion model, which is somewhat different from (1.109). More recent references on the equivalence of experiments are Brown et al. (2004) and Grama and Neumann (2006).

1.12 Exercises

Exercise 1.1 *Prove that any symmetric kernel K is a kernel of order 1 whenever the function $u \mapsto uK(u)$ is integrable. Find the maximum order of the Silverman kernel.* Hint: *Apply the Fourier transform and write the Silverman kernel as*

$$K(u) = \int_{-\infty}^{\infty} \frac{\cos(2\pi t u)}{1+(2\pi t)^4}\, dt.$$

Exercise 1.2 *Kernel estimator of the sth derivative $p^{(s)}$ of a density $p \in \mathcal{P}(\beta, L)$, $s < \beta$, can be defined as follows:*

$$\hat{p}_{n,s}(x) = \frac{1}{nh^{s+1}} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right).$$

Here $h > 0$ is a bandwidth and $K : \mathbf{R} \to \mathbf{R}$ is a bounded kernel with support $[-1, 1]$ satisfying for $\ell = \lfloor \beta \rfloor$:

$$\int u^j K(u) du = 0, \quad j = 0, 1, \ldots, s-1, s+1, \ldots, \ell, \tag{1.117}$$

$$\int u^s K(u) du = s! \tag{1.118}$$

(1) Prove that, uniformly over the class $\mathcal{P}(\beta, L)$, the bias of $\hat{p}_{n,s}(x_0)$ is bounded by $ch^{\beta-s}$ and the variance of $\hat{p}_{n,s}(x_0)$ is bounded by $c'(nh^{2s+1})^{-1}$ where $c > 0$ and $c' > 0$ are appropriate constants and x_0 is a given point in $\mathbf{R}$.

(2) Prove that the maximum of the MSE of $\hat{p}_{n,s}(x_0)$ over $\mathcal{P}(\beta, L)$ is of order $O\left(n^{-\frac{2(\beta-s)}{2\beta+1}}\right)$ as $n \to \infty$ if the bandwidth $h = h_n$ is chosen optimally.

(3) Let $\{\varphi_m\}_{m=0}^{\infty}$ be the orthonormal Legendre basis on $[-1, 1]$. Show that the kernel

$$K(u) = \sum_{m=0}^{\ell} \varphi_m^{(s)}(0) \varphi_m(u) I(|u| \le 1)$$

satisfies conditions (1.117) and (1.118).

Exercise 1.3 *Consider the estimator $\hat{p}_n$ defined in (1.3). Assume that the density $p(\cdot, \cdot)$ belongs to the class of all the probability densities on $\mathbf{R}^2$ satisfying*

$$|p(x, y) - p(x', y')| \le L(|x - x'|^\beta + |y - y'|^\beta), \quad \forall (x, y), (x', y') \in \mathbf{R}^2,$$

with given constants $0 < \beta \le 1$ and $L > 0$. Let (x_0, y_0) be a fixed point in $\mathbf{R}^2$. Derive upper bounds for the bias and the variance of $\hat{p}_n(x_0, y_0)$ and an upper bound on the mean squared risk at (x_0, y_0). Find the minimizer $h = h_n^*$ of the upper bound on the risk and the corresponding rate of convergence.

Exercise 1.4 *Define the $\mathrm{LP}(\ell)$ estimators of the derivatives $f^{(s)}(x)$, $s = 1, \ldots, \ell$, by*

$$\hat{f}_{ns}(x) = (U^{(s)}(0))^T \hat{\theta}_n(x) h^{-s}$$

where $U^{(s)}(u)$ is the vector whose coordinates are the sth derivatives of the corresponding coordinates of $U(u)$.

(1) Prove that if $\mathcal{B}_{nx} > 0$, then the estimator $\hat{f}_{ns}(x)$ is linear and it reproduces polynomials of degree $\le \ell - s$.

(2) Prove that, under the assumptions of Proposition 1.13, the maximum of the MSE of $\hat{f}_{ns}(x)$ over $\Sigma(\beta, L)$ is of order $O\left(n^{-\frac{2(\beta-s)}{2\beta+1}}\right)$ as $n \to \infty$ if the bandwidth $h = h_n$ is chosen optimally.

Exercise 1.5 *Show that the rectangular and the biweight kernels are inadmissible.*

Exercise 1.6 *Let $K \in L_2(\mathbf{R})$ be symmetric and such that $\widehat{K} \in L_\infty(\mathbf{R})$. Show that:*

(1) condition (1.53) is equivalent to (1.54),

(2) for integer β assumption (1.53) is satisfied if K is a kernel of order $\beta - 1$ and $\int |u|^\beta \big|K(u)\big| du < \infty$.

Exercise 1.7 *Let $\mathcal{P}$ be the class of all probability densities p on $\mathbf{R}$ such that*

$$\int \exp\left(\alpha|\omega|^r\right)\big|\phi(\omega)\big|^2 d\omega \le L^2,$$

where $\alpha > 0$, $r > 0$, $L > 0$ are given constants and $\phi = \mathcal{F}[p]$. Show that for any $n \ge 1$ the kernel density estimator $\hat{p}_n$ with the sinc kernel and appropriately chosen bandwidth $h = h_n$ satisfies

$$\sup_{p\in\mathcal{P}} \mathbf{E}_p \int \left(\hat{p}_n(x) - p(x)\right)^2 dx \le C\frac{(\log n)^{1/r}}{n},$$

where $C > 0$ is a constant depending only on r, L and α.

Exercise 1.8 *Let $\mathcal{P}_a$, where $a > 0$, be the class of all probability densities p on $\mathbf{R}$ such that the support of the characteristic function $\phi = \mathcal{F}[p]$ is included in a given interval $[-a, a]$. Show that for any $n \ge 1$ the kernel density estimator $\hat{p}_n$ with the sinc kernel and appropriately chosen bandwidth h satisfies*

$$\sup_{p\in\mathcal{P}_a} \mathbf{E}_p \int \left(\hat{p}_n(x) - p(x)\right)^2 dx \le \frac{a}{\pi n}.$$

This example, due to Ibragimov and Has'minskii (1983b), shows that it is possible to estimate the density with the $\sqrt{n}$ rate on sufficiently small nonparametric classes of functions.

Exercise 1.9 *Let $(X_1, \ldots, X_n)$ be an i.i.d. sample from a density $p \in L_2[0,1]$. Consider the projection estimator $\hat{p}_{nN}$ of p given in Definition 1.10.*

(1) Show that $\hat{c}_j$ are unbiased estimators of the Fourier coefficients $c_j = \int_0^1 p(x)\varphi_j(x)dx$ and find the variance of $\hat{c}_j$.

(2) Express the mean integrated squared error (MISE) of the estimator $\hat{p}_{nN}$ as a function of $p(\cdot)$ and $\{\varphi_j\}_{j=1}^\infty$. Denote it by $\mathrm{MISE}(N)$.

(3) Derive an unbiased risk estimation method. Show that

$$\mathbf{E}_p(\hat{J}(N)) = \mathrm{MISE}(N) - \int p^2,$$

where

$$\hat{J}(N) = \frac{1}{n-1}\sum_{j=1}^{N}\left[\frac{2}{n}\sum_{i=1}^{n}\varphi_j^2(X_i) - (n+1)\hat{c}_j^2\right].$$

Propose a data-driven selector of N.

(4) Suppose now that $\{\varphi_j\}_{j=1}^{\infty}$ is the trigonometric basis. Show that the MISE of $\hat{p}_{nN}$ is bounded by

$$\frac{N+1}{n} + \rho_N$$

where $\rho_N = \sum_{j=N+1}^{\infty} c_j^2$. Use this bound to prove that uniformly over the class of all the densities p belonging to $W^{per}(\beta, L)$, $\beta > 0$, and $L > 0$, the MISE of $\hat{p}_{nN}$ is of order $O\left(n^{-\frac{2\beta}{2\beta+1}}\right)$ for an appropriate choice of $N = N_n$.

Exercise 1.10 *Consider the nonparametric regression model under Assumption (A) and suppose that f belongs to the class $W^{per}(\beta, L)$ with $\beta \geq 2$. The aim of this exercise is to study the weighted projection estimator*

$$f_{n,\lambda}(x) = \sum_{j=1}^{n} \lambda_j \, \hat{\theta}_j \, \varphi_j(x).$$

(1) Prove that the risk MISE of $f_{n,\lambda}$ is minimized with respect to $\{\lambda_j\}_{j=1}^{n}$ at

$$\lambda_j^* = \frac{\theta_j(\theta_j + \alpha_j)}{\varepsilon^2 + (\theta_j + \alpha_j)^2} \ , \qquad j = 1, \ldots, n,$$

where $\varepsilon^2 = \sigma_\xi^2/n$ (λ_j^* are the weights corresponding to the weighted projection oracle).

(2) Check that the corresponding value of the risk is

$$\mathrm{MISE}(\{\lambda_j^*\}) = \sum_{j=1}^{n} \frac{\varepsilon^2 \theta_j^2}{\varepsilon^2 + (\theta_j + \alpha_j)^2} + \rho_n.$$

(3) Prove that

$$\sum_{j=1}^{n} \frac{\varepsilon^2 \theta_j^2}{\varepsilon^2 + (\theta_j + \alpha_j)^2} = (1 + o(1)) \sum_{j=1}^{n} \frac{\varepsilon^2 \theta_j^2}{\varepsilon^2 + \theta_j^2} \ .$$

(4) Prove that

$$\rho_n = (1 + o(1)) \sum_{j=n+1}^{\infty} \frac{\varepsilon^2 \theta_j^2}{\varepsilon^2 + \theta_j^2} \ .$$

(5) Deduce from the above results that

$$\mathrm{MISE}(\{\lambda_j^*\}) = \mathcal{A}_n^*(1 + o(1)), \quad n \to \infty,$$

where

$$\mathcal{A}_n^* = \sum_{j=1}^{\infty} \frac{\varepsilon^2 \theta_j^2}{\varepsilon^2 + \theta_j^2} \ .$$

(6) Check that

$$\mathcal{A}_n^* < \min_{N \geq 1} \mathcal{A}_{nN}.$$

Exercise 1.11 *(Equivalence between different types of estimators.)*
Consider the nonparametric regression model under Assumption (A). The smoothing spline estimator $f_n^{sp}(x)$ is defined as a solution of the following minimization problem (cf. Wahba (1990), Eubank (1988)):

$$f_n^{sp} = \arg\min_{f \in W} \left[\frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \kappa \int_0^1 (f''(x))^2 dx \right], \tag{1.119}$$

where $\kappa > 0$ is a smoothing parameter and W is one of the sets of functions defined below.

(1) First suppose that W is the set of all the functions $f : [0,1] \to \mathbf{R}$ such that f' is absolutely continuous. Prove that the estimator f_n^{sp} reproduces polynomials of degree ≤ 1 if $n \geq 2$.

(2) Suppose next that W is the set of all the functions $f : [0,1] \to \mathbf{R}$ such that (i) f' is absolutely continuous and (ii) the periodicity condition is satisfied: $f(0) = f(1)$, $f'(0) = f'(1)$. Prove that the minimization problem (1.119) is equivalent to:

$$\min_{\{b_j\}} \sum_{j=1}^{\infty} \Big(-2\hat{\theta}_j b_j + b_j^2 (\kappa \pi^4 a_j^2 + 1)[1 + O(n^{-1})] \Big), \tag{1.120}$$

where b_j are the Fourier coefficients of f, the term $O(n^{-1})$ is uniform in $\{b_j\}$, and a_j are defined according to (1.90).

(3) Assume now that the term $O(n^{-1})$ in (1.120) is negligible. Formally replacing it by 0, find the solution of (1.120) and conclude that the periodic spline estimator is approximately equal to a weighted projection estimator:

$$f_n^{sp}(x) \approx \sum_{j=1}^{\infty} \lambda_j^* \, \hat{\theta}_j \, \varphi_j(x)$$

with the weights λ_j^* written explicitly.

(4) Use (3) to show that for sufficiently small κ the spline estimator f_n^{sp} is approximated by the kernel estimator (1.62):

$$f_n(x) = \frac{1}{nh} \sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h} \right),$$

where $h = \kappa^{1/4}$ and K is the Silverman kernel (cf. Exercise 1.1).

2

Lower bounds on the minimax risk

2.1 Introduction

The examples of models studied in Chapter 1 show that the problem of nonparametric estimation is characterized by the following three ingredients:

- A nonparametric class of functions Θ containing the function θ that we want to estimate, for example, $\Theta = \Sigma(\beta, L)$ (the Hölder class) or $\Theta = W(\beta, L)$ (the Sobolev class).
- A family $\{P_\theta, \theta \in \Theta\}$ of probability measures, indexed by Θ, on a measurable space $(\mathcal{X}, \mathcal{A})$ associated with the data. For example, in the density model, P_θ is the probability measure associated with a sample $\mathbf{X} = (X_1, \dots, X_n)$ of size n when the density of X_i is $p(\cdot) = \theta$. For brevity, we do not indicate in our notation that P_θ, $\mathcal{X}$, and $\mathcal{A}$ depend on the number of observations n.
- A distance (or, more generally, a semi-distance) d on Θ used to define the risk.

We will call the *semi-distance* on Θ any function $d : \Theta \times \Theta \to [0, +\infty)$ satisfying $d(\theta, \theta') = d(\theta', \theta)$, $d(\theta, \theta') + d(\theta', \theta'') \geq d(\theta, \theta'')$ and $d(\theta, \theta) = 0$. In Chapter 1 we considered the following examples of semi-distances:

$$d(f, g) = \begin{cases} |f(x_0) - g(x_0)| \text{ for some fixed } x_0, \\ \|f - g\|_2, \\ \|f - g\|_\infty. \end{cases}$$

Throughout this chapter we will also suppose that the function $d(\cdot, \cdot)$ is a semi-distance. However, this assumption will often be redundant since the general results are valid for functions $d(\cdot, \cdot)$ satisfying only the triangle inequality.

A. B. Tsybakov, *Introduction to Nonparametric Estimation*,
DOI 10.1007/978-0-387-79052-7_2,

Given a semi-distance d, the performance of an estimator $\hat{\theta}_n$ of θ is measured by the *maximum risk* of this estimator on Θ :

$$r(\hat{\theta}_n) \stackrel{\triangle}{=} \sup_{\theta \in \Theta} \mathbf{E}_\theta \Big[d^2(\hat{\theta}_n, \theta) \Big],$$

where $\mathbf{E}_\theta$ denotes expectation with respect to P_θ. In Chapter 1 we established upper bounds on the maximum risk, that is, inequalities of the form

$$\sup_{\theta \in \Theta} \mathbf{E}_\theta \Big[d^2(\hat{\theta}_n, \theta) \Big] \leq C \psi_n^2$$

for certain estimators $\hat{\theta}_n$, certain positive sequences $\psi_n \to 0$, and constants $C < \infty$. The aim of this chapter is to complement these upper bounds by the corresponding lower bounds:

$$\forall\ \hat{\theta}_n : \quad \sup_{\theta \in \Theta} \mathbf{E}_\theta \Big[d^2(\hat{\theta}_n, \theta) \Big] \geq c\, \psi_n^2$$

(for sufficiently large n) where c is a positive constant. In this context, it is useful to define the *minimax risk* associated with a statistical model $\{P_\theta, \theta \in \Theta\}$ and with a semi-distance d:

$$\mathcal{R}_n^* \stackrel{\triangle}{=} \inf_{\hat{\theta}_n} \sup_{\theta \in \Theta} \mathbf{E}_\theta \Big[d^2(\hat{\theta}_n, \theta) \Big],$$

where the infimum is over all estimators. The upper bounds established in Chapter 1 imply that there exists a constant $C < \infty$ such that

$$\limsup_{n \to \infty} \psi_n^{-2} \mathcal{R}_n^* \leq C \tag{2.1}$$

for a sequence ψ_n converging to zero. The corresponding lower bounds claim that there exists a constant $c > 0$ such that, for the same sequence ψ_n,

$$\liminf_{n \to \infty} \psi_n^{-2} \mathcal{R}_n^* \geq c. \tag{2.2}$$

Definition 2.1 *A positive sequence $\{\psi_n\}_{n=1}^\infty$ is called an* **optimal rate of convergence** *of estimators on (Θ, d) if (2.1) and (2.2) hold. An estimator θ_n^* satisfying*

$$\sup_{\theta \in \Theta} \mathbf{E}_\theta \Big[d^2(\theta_n^*, \theta) \Big] \leq C'\, \psi_n^2,$$

where $\{\psi_n\}_{n=1}^\infty$ is the optimal rate of convergence and $C' < \infty$ is a constant, is called a **rate optimal estimator** *on (Θ, d).*

Definition 2.2 *An estimator θ_n^* is called* **asymptotically efficient** *on (Θ, d) if*

$$\lim_{n \to \infty} \frac{r(\theta_n^*)}{\mathcal{R}_n^*} = 1.$$

Remarks.

(1) Optimal rates of convergence are defined to within a multiplicative constant (or up to a bounded factor dependent on n). Indeed, if ψ_n is an optimal rate of convergence, then any sequence ψ'_n satisfying

$$0 < \liminf_{n\to\infty}(\psi_n/\psi'_n) \le \limsup_{n\to\infty}(\psi_n/\psi'_n) < \infty$$

is again an optimal rate of convergence. Sequences ψ_n and ψ'_n satisfying the above relation are said to have equivalent orders of magnitude. Any sequence belonging to the class of equivalent sequences can be taken as an optimal rate. Traditionally, the power sequences are convenient for use, e.g., $n^{-1/3}, n^{-2/5}$, in some cases (where appropriate) with an extra logarithmic factor, e.g., $(n/\log n)^{-1/3}, (n/\log n)^{-2/5}$.

(2) We can consider a more general framework where the maximum risk is defined as follows:

$$r_w(\hat\theta_n) = \sup_{\theta\in\Theta} \mathbf{E}_\theta\Big[w(\psi_n^{-1}d(\hat\theta_n,\theta))\Big]$$

with a *loss function* w such that

$$w : [0,\infty) \to [0,\infty) \text{ is monotone increasing, } w(0) = 0, \text{ and } w \not\equiv 0. \tag{2.3}$$

Some classical examples of loss functions are:

$$w(u) = u^p,\ p > 0, \qquad\qquad w(u) = I(u \ge A),\ A > 0$$

(in the latter case, the risk represents the probability to overshoot the fixed level A). In this general framework, lower bounds are formulated as inequalities of the following form:

$$\liminf_{n\to\infty} \inf_{\hat\theta_n} \sup_{\theta\in\Theta} \mathbf{E}_\theta\Big[w(\psi_n^{-1}\, d(\hat\theta_n,\theta))\Big] \ge c > 0. \tag{2.4}$$

2.2 A general reduction scheme

A general scheme for obtaining lower bounds is based on the following three remarks:

(a) *Reduction to bounds in probability.* Observe that it is sufficient to consider the loss function $w_0(u) = I(u \ge A)$ since, by the Markov inequality, for any loss function w and any $A > 0$ satisfying $w(A) > 0$ we have

$$\begin{aligned}\mathbf{E}_\theta\Big[w(\psi_n^{-1}d(\hat\theta_n,\theta))\Big] &\ge w(A)P_\theta(\psi_n^{-1}d(\hat\theta_n,\theta) \ge A) \\ &= w(A)P_\theta(d(\hat\theta_n,\theta) \ge s)\end{aligned} \tag{2.5}$$

with $s = s_n = A\psi_n$. Therefore, instead of searching for a lower bound on the minimax risk $\mathcal{R}_n^*$, it is sufficient to find a lower bound on the *minimax probabilities* of the form

$$\inf_{\hat{\theta}_n} \sup_{\theta \in \Theta} P_\theta(d(\hat{\theta}_n, \theta) \geq s).$$

This is a first simplification.

(b) *Reduction to a finite number of hypotheses.* It is clear that

$$\inf_{\hat{\theta}_n} \sup_{\theta \in \Theta} P_\theta(d(\hat{\theta}_n, \theta) \geq s) \geq \inf_{\hat{\theta}_n} \max_{\theta \in \{\theta_0, \ldots, \theta_M\}} P_\theta(d(\hat{\theta}_n, \theta) \geq s) \tag{2.6}$$

for any finite set $\{\theta_0, \ldots, \theta_M\}$ contained in Θ. In the examples, we will select $M \geq 1$ and $\theta_0, \ldots, \theta_M$ in an appropriate way. We will call *hypotheses* the $M+1$ elements $\theta_0, \theta_1, \ldots, \theta_M$ of Θ chosen to obtain lower bounds on the minimax risk. We will call a *test* any $\mathcal{A}$-measurable function $\psi : \mathcal{X} \to \{0, 1, \ldots, M\}$.

(c) *Choice of* $2s$*-separated hypotheses.* If

$$d(\theta_j, \theta_k) \geq 2s, \quad \forall\, k, j : \quad k \neq j, \tag{2.7}$$

then for any estimator $\hat{\theta}_n$

$$P_{\theta_j}(d(\hat{\theta}_n, \theta_j) \geq s) \geq P_{\theta_j}(\psi^* \neq j), \quad j = 0, 1, \ldots, M, \tag{2.8}$$

where $\psi^* : \mathcal{X} \to \{0, 1, \ldots, M\}$ is the *minimum distance test* defined by

$$\psi^* = \arg \min_{0 \leq k \leq M} d(\hat{\theta}_n, \theta_k).$$

Inequality (2.8) follows immediately from (2.7) and the triangle inequality.

It follows from (2.8) and (2.6) that if we can construct $M + 1$ hypotheses satisfying (2.7), then

$$\inf_{\hat{\theta}_n} \sup_{\theta \in \Theta} P_\theta(d(\hat{\theta}_n, \theta) \geq s) \geq \inf_{\hat{\theta}_n} \max_{\theta \in \{\theta_0, \ldots, \theta_M\}} P_\theta(d(\hat{\theta}_n, \theta) \geq s) \geq p_{e,M}, \tag{2.9}$$

where

$$p_{e,M} \stackrel{\triangle}{=} \inf_{\psi} \max_{0 \leq j \leq M} P_j(\psi \neq j), \quad P_j \stackrel{\triangle}{=} P_{\theta_j}$$

and $\inf_\psi$ denotes the infimum over all tests.

Conclusion: In order to obtain lower bounds as in (2.2) and (2.4), it is sufficient to check that

$$p_{e,M} \stackrel{\triangle}{=} \inf_{\psi} \max_{0 \leq j \leq M} P_j(\psi \neq j) \geq c', \tag{2.10}$$

where the hypotheses θ_j satisfy (2.7) with $s = A\psi_n$ and where the constant $c' > 0$ is independent of n. The quantity $p_{e,M}$ is called the *minimax probability of error* for the problem of testing $M + 1$ hypotheses $\theta_0, \theta_1, \ldots, \theta_M$.

2.3 Lower bounds based on two hypotheses

Consider first the simplest case, $M = 1$. This means that we take only two hypotheses θ_0 and θ_1 belonging to Θ. We will write for brevity $P_0 = P_{\theta_0}$, $P_1 = P_{\theta_1}$, $\hat{\theta} = \hat{\theta}_n$. We will first find lower bounds for the minimax probability of error $p_{e,1}$ and then for the minimax risk

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} P_\theta(d(\hat{\theta}, \theta) \geq s)$$

with $s > 0$. Consider the decomposition $P_0 = P_0^a + P_0^s$ where P_0^a and P_0^s denote the absolutely continuous component and the singular component of the measure P_0 with respect to the measure P_1. When no ambiguity is caused, we will use a short notation $\dfrac{dP_0^a}{dP_1}$ for the Radon–Nikodym derivative $\dfrac{dP_0^a}{dP_1}(\mathbf{X})$.

Proposition 2.1

$$p_{e,1} \geq \sup_{\tau > 0} \left\{ \frac{\tau}{1+\tau} P_1 \left(\frac{dP_0^a}{dP_1} \geq \tau \right) \right\}.$$

Proof. Fix $\tau > 0$. For any test $\psi : \mathcal{X} \to \{0, 1\}$,

$$\begin{aligned} P_0(\psi \neq 0) = P_0(\psi = 1) &\geq P_0^a(\psi = 1) \\ &= \int I(\psi = 1) \frac{dP_0^a}{dP_1} dP_1 \\ &\geq \tau \int I\left(\{\psi = 1\} \cap \left\{ \frac{dP_0^a}{dP_1} \geq \tau \right\} \right) dP_1 \geq \tau(p - \alpha_1), \end{aligned}$$

where $p = P_1(\psi = 1)$ and $\alpha_1 = P_1\left(\dfrac{dP_0^a}{dP_1} < \tau \right)$. Then

$$p_{e,1} = \inf_{\psi} \max_{j=0,1} P_j(\psi \neq j) \geq \min_{0 \leq p \leq 1} \max\{\tau(p - \alpha_1), 1 - p\} = \frac{\tau(1 - \alpha_1)}{1 + \tau}. \quad \blacksquare$$

We see that, in order to obtain a lower bound for the minimax probability of error $p_{e,1}$, it is sufficient to find constants $\tau > 0$ and $0 < \alpha < 1$ independent of n and satisfying

$$P_1 \left(\frac{dP_0^a}{dP_1} \geq \tau \right) \geq 1 - \alpha. \tag{2.11}$$

Proposition 2.1 implies the following lower bound on the minimax risk.

Theorem 2.1 *Assume that Θ contains two elements θ_0 and θ_1 satisfying $d(\theta_0, \theta_1) \geq 2s > 0$. Then*

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} P_\theta(d(\hat{\theta}, \theta) \geq s) \geq \sup_{\tau > 0} \left\{ \frac{\tau}{1+\tau} P_1 \left(\frac{dP_0^a}{dP_1} \geq \tau \right) \right\}.$$

PROOF: Straightforward in view of Proposition 2.1 and (2.9). ∎

REMARKS.

(1) Let $P_0 \ll P_1$ (then $P_0^a = P_0$). In this case, the random variable $\dfrac{dP_0}{dP_1}(\mathbf{X})$ is called the *likelihood ratio*.

(2) Condition (2.11) means that two probabilities P_0 and P_1 are not "very far" from each other. In other words, the closer P_0 is to P_1, the greater is the lower bound given in Theorem 2.1. If $P_0 = P_1$, condition (2.11) holds for $\tau = 1,\ \alpha = 0$, and the best lower bound that we can obtain using Proposition 2.1 is $p_{e,1} \geq 1/2$. Observe that this lower bound is not always sharp. Indeed, since $P_0 = P_1$, we have

$$p_{e,1} = \inf_{\psi} \max\{P_0(\psi = 1), P_0(\psi = 0)\},$$

and we can make the right hand side as close to 1 as we like by taking P_0 to be a suitably chosen Bernoulli distribution. In another extreme case, the measures P_0 and P_1 are mutually singular and Theorem 2.1 is trivial since the bound is equal to zero. Moreover, in this case we have $p_{e,1} = 0$ and the minimum with respect to ψ of the minimax probability of error is attained at the test taking value 1 on the support of P_1 and value 0 on the support of P_0.

(3) Even if $P_0 = P_1$, which may seem the most favorable case for obtaining lower bounds, the hypotheses θ_0 and θ_1 can be such that Theorem 2.1 would not give good results. The choice of the hypotheses is indeed very important, as illustrated by the following example.

Example 2.1 *A bad choice of the hypotheses θ_0 and θ_1.*

Consider the regression model

$$Y_i = f(i/n) + \xi_i, \qquad i = 1, \ldots, n,$$

where $f \in \Sigma(1,1)$ and where we would like to obtain a lower bound on the minimax risk over $\Theta = \Sigma(1,1)$. Assume that we have chosen the hypotheses

$$\theta_0 = f_0(\cdot) \equiv 0 \quad \text{and} \quad \theta_1 = f_1(\cdot),$$

where $f_1(x) = (2\pi n)^{-1} \sin(2\pi n x)$. Then $f_0(i/n) = f_1(i/n)$ for all i. It follows that the observations $(Y_1, \ldots, Y_n)$ are the same for $f = f_0$ and $f = f_1$. Then $P_0 = P_1$ and, by Proposition 2.1, we have $p_{e,1} \geq 1/2$ for any random errors ξ_i. Take the distance $d(f,g) = \|f - g\|_\infty$. Then $d(f_0, f_1) = (2\pi n)^{-1}$ and, since $f_0, f_1 \in \Sigma(1,1)$, we can use Theorem 2.1 and (2.5) with $s = (4\pi n)^{-1}$ to obtain inequality (2.2) for the class $\Theta = \Sigma(1,1)$ with rate $\psi_n \asymp 1/n$. This result is not satisfactory since $1/n$ is much smaller than the rate $(\log n/n)^{1/3}$ given by the upper

bound in Theorem 1.8. Indeed, we will see later (cf. Corollary 2.5) that $(\log n/n)^{1/3}$, and not $1/n$, is the optimal rate of convergence on $(\Sigma(1,1), \|\cdot\|_\infty)$.

2.4 Distances between probability measures

Let $(\mathcal{X}, \mathcal{A})$ be a measurable space and let P and Q be two probability measures on $(\mathcal{X}, \mathcal{A})$. Suppose that ν is a σ-finite measure on $(\mathcal{X}, \mathcal{A})$ satisfying $P \ll \nu$ and $Q \ll \nu$. Define $p = dP/d\nu,\ q = dQ/d\nu$. Observe that such a measure ν always exists since we can take, for example, $\nu = P + Q$.

Definition 2.3 *The* **Hellinger distance** *between P and Q is defined as follows:*

$$H(P,Q) = \left(\int (\sqrt{p} - \sqrt{q})^2 d\nu \right)^{1/2} \triangleq \left(\int \left[\sqrt{dP} - \sqrt{dQ} \right]^2 \right)^{1/2}. \tag{2.12}$$

It is easy to see that $H(P,Q)$ does not depend on the choice of the dominating measure ν. This explains the symbolic notation on the right hand side of (2.12). The following properties are straightforward.

Properties of the Hellinger distance

(i) $H(P,Q)$ satisfies the axioms of distance.

(ii) $0 \le H^2(P,Q) \le 2$.

(iii) $H^2(P,Q) = 2\left(1 - \int \sqrt{pq}\, d\nu\right) \triangleq 2\left(1 - \int \sqrt{dPdQ}\right)$.

(iv) If P and Q are product measures, $P = \otimes_{i=1}^n P_i, \quad Q = \otimes_{i=1}^n Q_i$, then

$$H^2(P,Q) = 2\left(1 - \prod_{i=1}^n \left(1 - \frac{H^2(P_i,Q_i)}{2}\right)\right).$$

We now introduce another distance between probability measures that will be useful in the sequel.

Definition 2.4 *The* **total variation distance** *between P and Q is defined as follows:*

$$V(P,Q) = \sup_{A \in \mathcal{A}} |P(A) - Q(A)| = \sup_{A \in \mathcal{A}} \left| \int_A (p - q) d\nu \right|.$$

The following two properties of the total variation distance are easy to prove.

Properties of the total variation distance

(i) $V(P,Q)$ satisfies the axioms of distance.

(ii) $0 \le V(P,Q) \le 1$.

Indeed, these properties follow from the next lemma. Write

$$\int \min(dP, dQ) \triangleq \int \min(p,q)d\nu.$$

Lemma 2.1 (Scheffé's theorem).

$$V(P,Q) = \frac{1}{2}\int |p-q|d\nu = 1 - \int \min(dP, dQ).$$

Proof. Observe that $A_0 = \{x \in \mathcal{X} : q(x) \ge p(x)\}$. Then

$$\int |p-q|d\nu = 2\int_{A_0} (q-p)d\nu$$

and

$$V(P,Q) \ge Q(A_0) - P(A_0) = \frac{1}{2}\int |p-q|d\nu = 1 - \int \min(p,q)d\nu.$$

On the other hand, for all $A \in \mathcal{A}$,

$$\begin{aligned}\left|\int_A (q-p)d\nu\right| &= \left|\int_{A\cap A_0} (q-p)d\nu + \int_{A\cap A_0^c} (q-p)d\nu\right| \\ &\le \max\left\{\int_{A_0}(q-p)d\nu, \int_{A_0^c}(p-q)d\nu\right\} = \frac{1}{2}\int |p-q|d\nu\end{aligned}$$

where A_0^c is the complement of A_0. Then

$$V(P,Q) = Q(A_0) - P(A_0) \tag{2.13}$$

implying the required result. ∎

Definition 2.5 *The* **Kullback divergence** *between P and Q is defined by*

$$K(P,Q) = \begin{cases} \int \log\dfrac{dP}{dQ}dP, & \text{if } P \ll Q, \\ +\infty, & \text{otherwise.}\end{cases}$$

The following lemma shows that this definition always makes sense, that is, the integral $\int \log\dfrac{dP}{dQ}dP$ is well-defined (it can be equal to $+\infty$) if $P \ll Q$.

Lemma 2.2 *If $P \ll Q$, then*

$$\int \left(\log \frac{dP}{dQ}\right)_- dP \le V(P,Q)$$

where $a_- = \max\{0, -a\}$.

Proof. If $P \ll Q$, we have $\{q > 0\} \supseteq \{p > 0\}$, $\{pq > 0\} = \{p > 0\}$. Therefore we can write

$$\int \left(\log \frac{dP}{dQ}\right)_- dP = \int_{pq>0} p \left(\log \frac{p}{q}\right)_- d\nu.$$

Take $A_1 = \{q \ge p > 0\} = A_0 \cap \{p > 0\}$. We have

$$\begin{aligned}\int_{pq>0} p \left(\log \frac{p}{q}\right)_- d\nu &= \int_{A_1} p \log \frac{q}{p} d\nu \le \int_{A_1} (q-p) d\nu \\ &= Q(A_1) - P(A_1) \le V(P,Q).\end{aligned}$$

∎

Thus we see that if $P \ll Q$, the Kullback divergence can be written as

$$\begin{aligned}K(P,Q) &= \int_{pq>0} p \log \frac{p}{q} d\nu \\ &= \int_{pq>0} p \left(\log \frac{p}{q}\right)_+ d\nu - \int_{pq>0} p \left(\log \frac{p}{q}\right)_- d\nu\end{aligned} \tag{2.14}$$

where $a_+ = \max\{a, 0\}$ and where the second integral on the right hand side is always finite.

Properties of the Kullback divergence

(i) $K(P,Q) \ge 0$. Indeed, it is sufficient to consider the case where all the integrals in (2.14) are finite. Then, by Jensen's inequality,

$$\int_{pq>0} p \log \frac{p}{q} d\nu = -\int_{pq>0} p \log \frac{q}{p} d\nu \ge -\log \left(\int_{p>0} q d\nu\right) \ge 0.$$

(ii) $K(P,Q)$ is not a distance (for example, it is not symmetric). One can also prove that its symmetrized version

$$K_*(P,Q) = K(P,Q) + K(Q,P),$$

defined for $P \sim Q$, that is for $P \ll Q$ and $Q \ll P$, is not a distance either.

(iii) If P and Q are product measures, $P = \otimes_{i=1}^n P_i$, $\quad Q = \otimes_{i=1}^n Q_i$, then

$$K(P,Q) = \sum_{i=1}^n K(P_i, Q_i).$$

The functions $V(\cdot,\cdot)$, $H^2(\cdot,\cdot)$, and the Kullback divergence are particular cases of the Csizsár f-divergences defined for $P \ll Q$ in the following way:

$$D(P,Q) = \int f\left(\frac{dP}{dQ}\right) dQ,$$

where f is a convex function on $(0,+\infty)$ satisfying certain conditions. Indeed, $V(\cdot,\cdot)$ and $H^2(\cdot,\cdot)$ correspond to $f(x) = |x-1|/2$ and $f(x) = (\sqrt{x}-1)^2$, while the Kullback divergence $K(P,Q)$ (if it is finite) is obtained for $f(x) = x\log x$. Among other f-divergences, the most famous is the χ^2 *divergence* defined as follows:

$$\chi^2(P,Q) = \begin{cases} \displaystyle\int \left(\frac{dP}{dQ} - 1\right)^2 dQ, \text{ if } P \ll Q, \\ +\infty, \qquad\qquad \text{otherwise.} \end{cases}$$

This is a particular case of $D(P,Q)$ corresponding to $f(x) = (x-1)^2$. It is often misnamed as the χ^2 "distance," whereas $\chi^2(\cdot,\cdot)$ is not a distance; it is sufficient to observe that it is not symmetric.

Properties of the χ^2 divergence.

(i) If $P \ll Q$, then

$$\chi^2(P,Q) = \int \left(\frac{dP}{dQ}\right)^2 dQ - 1 = \int_{pq>0} \frac{p^2}{q}\, d\nu - 1. \tag{2.15}$$

(ii) If P and Q are two product measures, $P = \otimes_{i=1}^n P_i$ and $Q = \otimes_{i=1}^n Q_i$, then

$$\chi^2(P,Q) = \prod_{i=1}^n \left(1 + \chi^2(P_i,Q_i)\right) - 1.$$

2.4.1 Inequalities for distances

In this subsection, we will often write for brevity $\int(\ldots)$ instead of $\int(\ldots)d\nu$. The following lemma establishes a link between the total variation distance and the Hellinger distance.

Lemma 2.3 (Le Cam's inequalities).

$$\int \min(dP, dQ) \geq \frac{1}{2}\left(\int \sqrt{dP\,dQ}\right)^2 = \frac{1}{2}\left(1 - \frac{H^2(P,Q)}{2}\right)^2, \tag{2.16}$$

$$\frac{1}{2}H^2(P,Q) \leq V(P,Q) \leq H(P,Q)\sqrt{1 - \frac{H^2(P,Q)}{4}}\,. \tag{2.17}$$

PROOF. Since $\int \max(p,q) + \int \min(p,q) = 2$, we obtain

$$\left(\int \sqrt{pq}\right)^2 = \left(\int \sqrt{\min(p,q)\max(p,q)}\right)^2 \leq \int \min(p,q) \int \max(p,q)$$
$$= \int \min(p,q) \left[2 - \int \min(p,q)\right], \tag{2.18}$$

proving the inequality in (2.16). The equality in (2.16) is nothing other than property (iii) of the Hellinger distance. The first inequality in (2.17) follows from Lemma 2.1 and property (iii) of the Hellinger distance. Indeed,

$$V(P,Q) = 1 - \int \min(p,q) \geq 1 - \int \sqrt{pq} = H^2(P,Q)/2.$$

In order to prove the second inequality in (2.17), observe that (2.18) can be written as

$$\left(1 - \frac{H^2(P,Q)}{2}\right)^2 \leq (1 - V(P,Q))(1 + V(P,Q)) = 1 - V^2(P,Q). \qquad \blacksquare$$

The next lemma links the Hellinger distance to the Kullback divergence.

Lemma 2.4

$$H^2(P,Q) \leq K(P,Q). \tag{2.19}$$

PROOF. It is sufficient to assume that $K(P,Q) < +\infty$ (and therefore $P \ll Q$). Since $-\log(x+1) \geq -x$ if $x > -1$, we have

$$K(P,Q) = \int_{pq>0} p \left(\log \frac{p}{q}\right) = 2 \int_{pq>0} p \left(\log \sqrt{\frac{p}{q}}\right)$$
$$= -2 \int_{pq>0} p \log \left(\left[\sqrt{\frac{q}{p}} - 1\right] + 1\right)$$
$$\geq -2 \int_{pq>0} p \left[\sqrt{\frac{q}{p}} - 1\right]$$
$$= -2 \left(\int \sqrt{pq} - 1\right) = H^2(P,Q). \qquad \blacksquare$$

Corollary 2.1 *Let φ be the density of the standard normal distribution $\mathcal{N}(0,1)$. Then*

(i) $\displaystyle\int \log \frac{\varphi(x)}{\varphi(x+t)}\, \varphi(x)dx = \frac{t^2}{2}\ , \quad \forall\ t \in \mathbf{R},$

(ii) $\displaystyle\int \left(\sqrt{\varphi(x)} - \sqrt{\varphi(x+t)}\right)^2 dx \leq \frac{t^2}{2}\ , \quad \forall\ t \in \mathbf{R}.$

Combining the right hand inequality in (2.17) and Lemma 2.4 we can link the total variation distance to the Kullback divergence:

$$V(P,Q) \le H(P,Q) \le \sqrt{K(P,Q)}. \tag{2.20}$$

However, (2.20) does not give the most accurate inequality between $V(P,Q)$ and $K(P,Q)$. It can be improved as stated in the following lemma.

Lemma 2.5 (Pinsker's inequalities).
(i) First Pinsker's inequality.

$$V(P,Q) \le \sqrt{K(P,Q)/2}.$$

(ii) Second Pinsker's inequality. If $P \ll Q$, then

$$\int \left|\log \frac{dP}{dQ}\right| dP \triangleq \int_{pq>0} p \left|\log \frac{p}{q}\right| d\nu \le K(P,Q) + \sqrt{2K(P,Q)}\,, \tag{2.21}$$

and

$$\int \left(\log \frac{dP}{dQ}\right)_+ dP \le K(P,Q) + \sqrt{K(P,Q)/2}. \tag{2.22}$$

Proof. (i) Introduce the function

$$\psi(x) = x \log x - x + 1, \quad x \ge 0,$$

where $0 \log 0 \triangleq 0$. Observe that $\psi(0) = 1$, $\psi(1) = 0$, $\psi'(1) = 0$, $\psi''(x) = 1/x \ge 0$, and $\psi(x) \ge 0$, $\forall x \ge 0$. Moreover,

$$\left(\frac{4}{3} + \frac{2}{3}x\right) \psi(x) \ge (x-1)^2, \quad x \ge 0. \tag{2.23}$$

Indeed, this inequality is clear for $x = 0$. If $x > 0$, the function

$$g(x) = (x-1)^2 - \left(\frac{4}{3} + \frac{2}{3}x\right) \psi(x)$$

satisfies

$$g(1) = 0, \quad g'(1) = 0, \quad g''(x) = -\frac{4\psi(x)}{3x} \le 0.$$

Thus, for ξ satisfying $|\xi - 1| < |x - 1|$ we have

$$g(x) = g(1) + g'(1)(x-1) + \frac{g''(\xi)}{2}(x-1)^2 = -\frac{4\psi(\xi)}{6\xi}(x-1)^2 \le 0,$$

proving (2.23). From (2.23), we obtain that if $P \ll Q$, then

$$
\begin{aligned}
V(P,Q) &= \frac{1}{2}\int |p-q| = \frac{1}{2}\int_{q>0}\left|\frac{p}{q}-1\right| q \\
&\le \frac{1}{2}\int_{q>0} q\sqrt{\left(\frac{4}{3}+\frac{2p}{3q}\right)\psi\left(\frac{p}{q}\right)} \\
&\le \frac{1}{2}\sqrt{\int\left(\frac{4q}{3}+\frac{2p}{3}\right)}\sqrt{\int_{q>0} q\psi\left(\frac{p}{q}\right)} \qquad \text{(Cauchy–Schwarz)} \\
&= \sqrt{\frac{1}{2}\int_{pq>0} p\log\frac{p}{q}} = \sqrt{K(P,Q)/2}.
\end{aligned}
$$

If $P \not\ll Q$, the inequality is straightforward.
(ii) Equality (2.14), Lemma 2.2, and the first Pinsker inequality imply that

$$
\begin{aligned}
\int_{pq>0} p\left|\log\frac{p}{q}\right| &= \int_{pq>0} p\left(\log\frac{p}{q}\right)_+ + \int_{pq>0} p\left(\log\frac{p}{q}\right)_- \\
&= K(P,Q) + 2\int_{pq>0} p\left(\log\frac{p}{q}\right)_- \\
&\le K(P,Q) + 2V(P,Q) \le K(P,Q) + \sqrt{2K(P,Q)}.
\end{aligned}
$$

This yields (2.21). Inequality (2.22) is obtained similarly. ∎

The first Pinsker inequality is exact in the sense that there exist probability measures P and Q for which it becomes equality. However, it is nontrivial only if $K(P,Q) \le 2$ since we always have $V(P,Q) \le 1$. A nontrivial extension to larger Kullback divergences is obtained using the following lemma.

Lemma 2.6

$$
\int \min(dP, dQ) \ge \frac{1}{2}\exp(-K(P,Q)). \tag{2.24}
$$

Proof. It is sufficient to assume that $K(P,Q) < +\infty$ (and therefore $P \ll Q$). Using the Jensen inequality we get

$$
\begin{aligned}
\left(\int\sqrt{pq}\right)^2 &= \exp\left(2\log\int_{pq>0}\sqrt{pq}\right) = \exp\left(2\log\int_{pq>0} p\sqrt{\frac{q}{p}}\right) \\
&\ge \exp\left(2\int_{pq>0} p\log\sqrt{\frac{q}{p}}\right) = \exp(-K(P,Q)).
\end{aligned}
$$

By comparing this result to inequality (2.16) we obtain (2.24). ∎

From Lemmas 2.1 and 2.6 we get

$$
V(P,Q) \le 1 - \frac{1}{2}\exp(-K(P,Q)). \tag{2.25}
$$

We finally establish a link between the Kullback and the χ^2 divergences.

Lemma 2.7

$$K(P,Q) \le \log(1+\chi^2(P,Q)) \le \chi^2(P,Q). \tag{2.26}$$

PROOF: Straightforward in view of (2.15) and Jensen's inequality. ∎

From (2.20) and (2.26) we get the following chain of inequalities:

$$V(P,Q) \le H(P,Q) \le \sqrt{K(P,Q)} \le \sqrt{\chi^2(P,Q)}. \tag{2.27}$$

These inequalities are clearly not the sharpest obtainable from the results stated above. However, they are quite instructive since they reveal the hierarchy existing between the divergences V, H, K, and χ^2.

2.4.2 Bounds based on distances

In order to apply Theorem 2.1 and Proposition 2.1 we need the condition (2.11) dealing directly with the distribution of the likelihood ratio of P_0 and P_1. This condition is quite general but not always easy to check. Therefore, other bounds on the minimax probability of error for two hypotheses are often used, based on the distances or divergences between P_0 and P_1. Some of them are given in the following theorem.

Theorem 2.2 *Let P_0 and P_1 be two probability measures on $(\mathcal{X},\mathcal{A})$.*

(i) If $V(P_1,P_0) \le \alpha < 1$, then

$$p_{e,1} \ge \frac{1-\alpha}{2} \quad \textbf{(total variation version)}.$$

(ii) If $H^2(P_1,P_0) \le \alpha < 2$, then

$$p_{e,1} \ge \frac{1}{2}\left(1-\sqrt{\alpha(1-\alpha/4)}\right) \quad \textbf{(Hellinger version)}.$$

(iii) If $K(P_1,P_0) \le \alpha < \infty$ (or $\chi^2(P_1,P_0) \le \alpha < \infty$), then

$$p_{e,1} \ge \max\left(\frac{1}{4}\exp(-\alpha), \frac{1-\sqrt{\alpha/2}}{2}\right) \quad \textbf{(Kullback/}\chi^2\textbf{ version)}.$$

PROOF.

$$\begin{aligned} p_{e,1} &= \inf_{\psi} \max_{j=0,1} P_j(\psi \ne j) \ge \frac{1}{2}\inf_{\psi}(P_0(\psi \ne 0) + P_1(\psi \ne 1)) \\ &= \frac{1}{2}(P_0(\psi^* \ne 0) + P_1(\psi^* \ne 1)) \end{aligned} \tag{2.28}$$

where ψ^* is the maximum likelihood test:

$$\psi^* = \begin{cases} 0, \text{ if } p_0 \geq p_1, \\ 1, \text{ otherwise,} \end{cases}$$

and where p_0 and p_1 are the densities of P_0 and P_1 with respect to ν. Next, Lemma 2.1 gives

$$\frac{1}{2}(P_0(\psi^* \neq 0) + P_1(\psi^* \neq 1)) = \frac{1}{2}\int \min(dP_0, dP_1) = (1 - V(P_0, P_1))/2.$$

This result combined with (2.28) implies part (i) of the theorem. From (i) and Lemma 2.3 we obtain part (ii). Finally, to prove part (iii) it suffices to bound $V(P_0, P_1)$ using inequality (2.24) or the first Pinsker inequality and then to apply (2.26). ∎

The idea of the proof of Theorem 2.2 differs from that of Theorem 2.1 since we bound the minimax probability of error from below by the average error. The average error is always less than or equal to 1/2 and therefore the bound also satisfies this restriction.

Theorem 2.2 sometimes enables us to obtain lower bounds that are technically more convenient than those based on Theorem 2.1. It is often easier to check the condition on the Kullback divergence than (2.11) or the assumptions involving other distances. However, the Kullback divergence is not finite for all probability measures. That is why the Hellinger version is more convenient in certain cases. An example is given in Exercise 2.7. Finally, there exist statistical models where the Kullback and the χ^2 divergences are not well-defined, the Hellinger and the total variation distances are difficult to handle, while the likelihood ratio version of Theorem 2.1 is effectively applicable.

2.5 Lower bounds on the risk of regression estimators at a point

We now apply the technique based on two hypotheses to obtain lower bounds in the nonparametric regression model. Assume that the following conditions are satisfied.

Assumption (B)

(i) The statistical model is that of nonparametric regression:

$$Y_i = f(X_i) + \xi_i, \ i = 1, \ldots, n,$$

where $f : [0, 1] \to \mathbf{R}$.

(ii) The random variables ξ_i *are i.i.d. having a density* $p_\xi(\cdot)$ *with respect to the Lebesgue measure on* $\mathbf{R}$ *such that*

$$\exists p_* > 0, v_0 > 0 : \quad \int p_\xi(u) \log \frac{p_\xi(u)}{p_\xi(u+v)} du \leq p_* v^2 \tag{2.29}$$

for all $|v| \le v_0$.
(iii) The variables $X_i \in [0,1]$ *are deterministic.*

By Corollary 2.1, condition (ii) in Assumption (B) holds if, for example, $p_\xi(\cdot)$ is the density of the normal distribution $\mathcal{N}(0,\sigma^2)$, $\sigma^2 > 0$.

We will also suppose in this section that Assumption (LP2) of Chapter 1 holds.

Our aim is to obtain a lower bound for the minimax risk on (Θ, d) where Θ is a Hölder class:

$$\Theta = \Sigma(\beta, L),\ \beta > 0, L > 0,$$

and where d is a distance at a fixed point $x_0 \in [0,1]$:

$$d(f,g) = |f(x_0) - g(x_0)|.$$

The rate that we would like to obtain is

$$\psi_n = n^{-\frac{\beta}{2\beta+1}}. \tag{2.30}$$

Indeed, this is the same rate as in the upper bounds of Chapter 1 which will enable us to conclude that (2.30) is optimal on (Θ, d).

By the general scheme of Section 2.2 it is sufficient to prove that

$$\inf_{\hat\theta_n} \max_{\theta \in \{\theta_0,\dots,\theta_M\}} P_\theta(d(\hat\theta_n, \theta) \ge s) \ge c' > 0,$$

where $s = A\psi_n$, with a constant $A > 0$. Using the notation of this section and taking $M = 1$ (two hypotheses) we can write the last display as follows:

$$\inf_{T_n} \max_{f \in \{f_{0n}, f_{1n}\}} P_f(|T_n(x_0) - f(x_0)| \ge A\psi_n) \ge c' > 0 \tag{2.31}$$

where $f_{0n}(\cdot) = \theta_0$ and $f_{1n}(\cdot) = \theta_1$ are two hypotheses, $A > 0$, and $\inf_{T_n}$ denotes the infimum over all estimators.

In order to obtain (2.31), we apply the Kullback version of Theorem 2.2 and (2.9). We choose the hypotheses $\theta_0 = f_{0n}(\cdot)$ and $\theta_1 = f_{1n}(\cdot)$ in the following way:

$$f_{0n}(x) \equiv 0, \quad f_{1n}(x) = L h_n^\beta K\left(\frac{x - x_0}{h_n}\right), \qquad x \in [0,1],$$

where

$$h_n = c_0 n^{-\frac{1}{2\beta+1}}, \quad c_0 > 0, \tag{2.32}$$

and where the function $K : \mathbf{R} \to [0, +\infty)$ satisfies

$$K \in \Sigma(\beta, 1/2) \cap C^\infty(\mathbf{R}) \quad \text{and} \quad K(u) > 0 \iff u \in (-1/2, 1/2). \tag{2.33}$$

There exist functions K satisfying this condition. For example, for a sufficiently small $a > 0$ we can take

$$K(u) = aK_0(2u), \quad \text{where} \quad K_0(u) = \exp\left(-\frac{1}{1-u^2}\right) I(|u| \le 1). \tag{2.34}$$

In order to apply Theorem 2.2 and (2.9), we need to check the following three conditions:

(a) $f_{jn} \in \Sigma(\beta, L), j = 0, 1,$

(b) $d(f_{1n}, f_{0n}) \ge 2s,$

(c) $K(P_0, P_1) \le \alpha < \infty.$

We now show that these conditions hold for sufficiently small c_0 and sufficiently large n.

(a) *The condition* $f_{jn} \in \Sigma(\beta, L)$, $j = 0, 1$.
For $\ell = \lfloor \beta \rfloor$, the ℓth order derivative of f_{1n} is

$$f_{1n}^{(\ell)}(x) = Lh_n^{\beta-\ell} K^{(\ell)}\left(\frac{x - x_0}{h_n}\right).$$

Then, by (2.33),

$$\begin{aligned} |f_{1n}^{(\ell)}(x) - f_{1n}^{(\ell)}(x')| &= Lh_n^{\beta-\ell} |K^{(\ell)}(u) - K^{(\ell)}(u')| \\ &\le Lh_n^{\beta-\ell} |u - u'|^{\beta-\ell}/2 = L|x - x'|^{\beta-\ell}/2 \end{aligned} \tag{2.35}$$

with $u = (x - x_0)/h_n$, $u' = (x' - x_0)/h_n$, and $x, x' \in \mathbf{R}$. This means that f_{1n} belongs to the class $\Sigma(\beta, L)$ on $\mathbf{R}$. Then it is clear that f_{1n} restricted to $[0, 1]$ belongs to the class $\Sigma(\beta, L)$ on $[0, 1]$.

(b) *The condition* $d(f_{1n}, f_{0n}) \ge 2s$.
We have

$$d(f_{1n}, f_{0n}) = |f_{1n}(x_0)| = Lh_n^\beta K(0) = Lc_0^\beta K(0) n^{-\frac{\beta}{2\beta+1}}.$$

Then the condition $d(f_{1n}, f_{0n}) \ge 2s$ holds with

$$s = s_n = \frac{1}{2} Lc_0^\beta K(0) n^{-\frac{\beta}{2\beta+1}} \triangleq An^{-\frac{\beta}{2\beta+1}} = A\psi_n.$$

(c) *The condition* $K(P_0, P_1) \le \alpha$.
Observe that P_j (the distribution of $Y_1, \dots, Y_n$ for $f = f_{jn}$) admits the following density with respect to the Lebesgue measure on $\mathbf{R}^n$:

$$p_j(u_1, \dots, u_n) = \prod_{i=1}^n p_\xi(u_i - f_{jn}(X_i)), \quad j = 0, 1.$$

There exists an integer n_0 depending only on $c_0, L, \beta, K_{\max}, v_0$ such that for all $n > n_0$ we have $nh_n \geq 1$ and $Lh_n^\beta K_{\max} \leq v_0$ where $K_{\max} = \max_u K(u)$. Then, by (2.29) and by Assumption (LP2) of Chapter 1, we obtain for $n > n_0$

$$
\begin{aligned}
K(P_0, P_1) &= \int \log \frac{dP_0}{dP_1} dP_0 \qquad (2.36)\\
&= \int \dots \int \log \prod_{i=1}^n \frac{p_\xi(u_i)}{p_\xi(u_i - f_{1n}(X_i))} \prod_{i=1}^n [p_\xi(u_i) du_i] \\
&= \sum_{i=1}^n \int \log \frac{p_\xi(y)}{p_\xi(y - f_{1n}(X_i))} p_\xi(y) dy \leq p_* \sum_{i=1}^n f_{1n}^2(X_i) \\
&= p_* L^2 h_n^{2\beta} \sum_{i=1}^n K^2 \left(\frac{X_i - x_0}{h_n} \right) \\
&\leq p_* L^2 h_n^{2\beta} K_{\max}^2 \sum_{i=1}^n I \left(\left| \frac{X_i - x_0}{h_n} \right| \leq \frac{1}{2} \right) \\
&\leq p_* a_0 L^2 K_{\max}^2 h_n^{2\beta} \max(nh_n, 1) \\
&= p_* a_0 L^2 K_{\max}^2 n h_n^{2\beta+1},
\end{aligned}
$$

where a_0 is the constant appearing in Assumption (LP2). If we choose

$$
c_0 = \left(\frac{\alpha}{p_* a_0 L^2 K_{\max}^2} \right)^{\frac{1}{2\beta+1}},
$$

then, by (2.32), we obtain $K(P_0, P_1) \leq \alpha$.

By part (iii) of Theorem 2.2, the above argument implies that, for any $n > n_0$ and for any estimator T_n,

$$
\begin{aligned}
\sup_{f \in \Sigma(\beta, L)} P_f(|T_n(x_0) - f(x_0)| \geq s_n) &\geq \max_{j=0,1} P_j(|T_n(x_0) - f_j(x_0)| \geq s_n) \\
&\geq \max \left(\frac{1}{4} \exp(-\alpha), \frac{1 - \sqrt{\alpha/2}}{2} \right) \\
&\triangleq V_0(\alpha).
\end{aligned}
$$

This yields the following result.

Theorem 2.3 *Suppose that $\beta > 0$ and $L > 0$. Under Assumption (B) and Assumption (LP2) of Chapter 1 we have, for all $x_0 \in [0, 1]$, $t > 0$,*

$$
\liminf_{n \to \infty} \inf_{T_n} \sup_{f \in \Sigma(\beta, L)} P_f \left(n^{\frac{\beta}{2\beta+1}} |T_n(x_0) - f(x_0)| \geq t^{\frac{\beta}{2\beta+1}} \right) \geq V_0(ct), \qquad (2.37)
$$

where $\inf_{T_n}$ *denotes the infimum over all estimators and* $c > 0$ *depends only on* β,L, p_*, *and* a_0. *Moreover,*

$$\liminf_{n\to\infty} \inf_{T_n} \sup_{f\in\Sigma(\beta,L)} \mathbf{E}_f \left[n^{\frac{2\beta}{2\beta+1}} (T_n(x_0) - f(x_0))^2 \right] \geq c_1, \tag{2.38}$$

where $c_1 > 0$ *depends only on* β,L, p_*, *and* a_0.

Corollary 2.2 *Consider the nonparametric regression model under the following conditions:*

(i) $X_i = i/n$ *for* $i = 1, \ldots, n$;
(ii) the random variables ξ_i *are i.i.d. with density* p_ξ *satisfying (2.29) and such that*

$$\mathbf{E}(\xi_i) = 0, \qquad \mathbf{E}(\xi_i^2) < \infty.$$

Then, for $\beta > 0$ *and* $L > 0$, *the rate of convergence* $\psi_n = n^{-\frac{\beta}{2\beta+1}}$ *is optimal on* $(\Sigma(\beta, L), d_0)$ *where* d_0 *is the distance at a fixed point* $x_0 \in [0, 1]$.

Moreover, if $\ell = \lfloor \beta \rfloor$, *the local polynomial estimator* $LP(\ell)$, *with the kernel* K *and the bandwidth* h_n *satisfying assumptions (iii) and (iv) of Theorem 1.7, is rate optimal on* $(\Sigma(\beta, L), d_0)$.

Remarks.

(1) It follows from (2.37) that

$$\liminf_{a\to 0} \liminf_{n\to\infty} \inf_{T_n} \sup_{f\in\Sigma(\beta,L)} P_f \left(n^{\frac{\beta}{2\beta+1}} |T_n(x_0) - f(x_0)| \geq a \right) \geq \frac{1}{2}\,. \tag{2.39}$$

Here the constant $1/2$ appears again; this is the maximum value that can be obtained for the lower bounds based on two hypotheses. However, using the techniques of M hypotheses with $M \to \infty$, inequality (2.39) can be improved to make the asymptotic constant equal to 1, see Exercise 2.9.

(2) Since V_0 does not depend on x_0, we have in fact proved a stronger inequality than (2.37), with a uniform bound in x_0:

$$\liminf_{n\to\infty} \inf_{T_n} \inf_{x_0\in[0,1]} \sup_{f\in\Sigma(\beta,L)} P_f \left(n^{\frac{\beta}{2\beta+1}} |T_n(x_0) - f(x_0)| \geq t^{\frac{\beta}{2\beta+1}} \right) \geq V_0(ct). \tag{2.40}$$

The techniques described in this section can be used to obtain a bound similar to that of Theorem 2.3 for the problem of estimation of a probability density (cf. Exercise 2.8).

2.6 Lower bounds based on many hypotheses

The lower bounds based on two hypotheses turn out to be inconvenient when we deal with estimation in L_p distances. Consider, for example, the L_2 distance:

$$d(f,g) = \|f-g\|_2 = \left(\int_0^1 (f(x)-g(x))^2 dx\right)^{1/2}$$

and suppose that Assumption (B) and Assumption (LP2) of Chapter 1 hold. Let us try to apply the technique of two hypotheses with f_{0n} and f_{1n} defined as in the previous section (taking $x_0 = 1/2$ as an example):

$$f_{0n}(x) \equiv 0,$$
$$f_{1n}(x) = Lh_n^\beta K\left(\frac{x-1/2}{h_n}\right).$$

Here, $h_n > 0$ and $K(\cdot)$ is a function satisfying (2.33). Apply now the Kullback version of Theorem 2.2. The condition $K(P_0, P_1) \le \alpha < \infty$ and inequality (2.36) impose the following restriction on h_n:

$$\limsup_{n\to\infty} nh_n^{2\beta+1} < \infty.$$

In other words, we obtain $h_n = O\left(n^{-\frac{1}{2\beta+1}}\right)$, as in the previous section. Now,

$$\begin{aligned} d(f_{0n}, f_{1n}) = \|f_{0n} - f_{1n}\|_2 &= \left(\int_0^1 f_{1n}^2(x)dx\right)^{1/2} \\ &= Lh_n^\beta \left(\int_0^1 K^2\left(\frac{x-1/2}{h_n}\right)dx\right)^{1/2} \\ &= Lh_n^{\beta+\frac{1}{2}} \left(\int K^2(u)du\right)^{1/2} \end{aligned}$$

for sufficiently large n. Therefore $d(f_{0n}, f_{1n}) \asymp h_n^{\beta+\frac{1}{2}} = O(n^{-1/2})$ implying that (2.9) can only be used for $s \le d(f_{0n}, f_{1n})/2 = O(n^{-1/2})$. To conclude, the technique based on two hypotheses gives a lower bound with the rate $n^{-1/2}$, which is not satisfactory because it is much smaller than $n^{-\frac{\beta}{2\beta+1}}$ appearing in the upper bound on the L_2-risk on $\Sigma(\beta, L)$ (cf. Corollary 1.2). This problem can be fixed by switching to M hypotheses with M tending to infinity as $n \to \infty$.

Proposition 2.2 *Let $P_0, P_1, \dots, P_M$ be probability measures on $(\mathcal{X}, \mathcal{A})$. Then*

$$p_{e,M} \ge \sup_{\tau>0} \frac{\tau M}{1+\tau M}\left[\frac{1}{M}\sum_{j=1}^M P_j\left(\frac{dP_{0,j}^a}{dP_j} \ge \tau\right)\right],$$

where $P_{0,j}^a$ is the absolutely continuous component of the measure P_0 with respect to P_j.

Proof. Let ψ be a test taking values in $\{0, 1, \dots, M\}$. Then

$$\bigcup_{j=1}^{M} \{\psi = j\} = \{\psi \neq 0\}$$

and

$$\{\psi = j\} \cap \{\psi = k\} = \emptyset \quad \text{for } k \neq j.$$

Introducing the random event $A_j = \left\{ \dfrac{dP_{0,j}^a}{dP_j} \geq \tau \right\}$ we can write

$$\begin{aligned} P_0(\psi \neq 0) &= \sum_{j=1}^{M} P_0(\psi = j) \geq \sum_{j=1}^{M} P_{0,j}^a(\psi = j) \\ &\geq \sum_{j=1}^{M} \tau P_j(\{\psi = j\} \cap A_j) \\ &\geq \tau M \left(\frac{1}{M} \sum_{j=1}^{M} P_j(\psi = j) \right) - \tau \sum_{j=1}^{M} P_j(A_j^c) \\ &= \tau M(p_0 - \alpha), \end{aligned}$$

where A_j^c is the complement of A_j and

$$p_0 = \frac{1}{M} \sum_{j=1}^{M} P_j(\psi = j), \qquad \alpha = \frac{1}{M} \sum_{j=1}^{M} P_j \left(\frac{dP_{0,j}^a}{dP_j} < \tau \right).$$

Then

$$\begin{aligned} \max_{0 \leq j \leq M} P_j(\psi \neq j) &= \max \left\{ P_0(\psi \neq 0), \ \max_{1 \leq j \leq M} P_j(\psi \neq j) \right\} \\ &\geq \max \left\{ \tau M(p_0 - \alpha), \ \frac{1}{M} \sum_{j=1}^{M} P_j(\psi \neq j) \right\} \\ &= \max\{\tau M(p_0 - \alpha), 1 - p_0\} \\ &\geq \min_{0 \leq p \leq 1} \max\{\tau M(p - \alpha), 1 - p\} \\ &= \frac{\tau M(1 - \alpha)}{1 + \tau M}. \end{aligned}$$

∎

Theorem 2.4 (Main theorem on lower bounds for the risk). *Assume that Θ contains elements $\theta_0, \theta_1, \ldots, \theta_M$ such that:*

(i) $d(\theta_j, \theta_k) \geq 2s > 0, \quad \forall\ 0 \leq j < k \leq M;$

(ii) there exist $\tau > 0$ *and* $0 < \alpha < 1$ *satisfying*

$$\frac{1}{M}\sum_{j=1}^{M} P_j\left(\frac{dP_{0,j}^a}{dP_j} \geq \tau\right) \geq 1-\alpha, \tag{2.41}$$

where $P_{0,j}^a$ is the absolutely continuous component of the measure $P_0 = P_{\theta_0}$ with respect to $P_j = P_{\theta_j}$. Then

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta} P_\theta(d(\hat{\theta},\theta) \geq s) \geq \frac{\tau M}{1+\tau M}(1-\alpha). \tag{2.42}$$

The proof of this theorem follows immediately from Proposition 2.2 and (2.9).

For $M = 1$, Proposition 2.2 and Theorem 2.4 coincide with Proposition 2.1 and Theorem 2.1, respectively. We now derive analogs of Theorem 2.4 where we replace condition (2.41) by appropriate assumptions on the Kullback or the χ^2 divergences between the measures P_j and P_0. We first obtain the following modification of Proposition 2.2 using the Kullback divergence.

Proposition 2.3 *Let $P_0, P_1, \ldots, P_M$ be probability measures on $(\mathcal{X}, \mathcal{A})$ satisfying*

$$\frac{1}{M}\sum_{j=1}^{M} K(P_j, P_0) \leq \alpha_* \tag{2.43}$$

with $0 < \alpha_* < \infty$. Then

$$p_{e,M} \geq \sup_{0<\tau<1}\left[\frac{\tau M}{1+\tau M}\left(1+\frac{\alpha_* + \sqrt{\alpha_*/2}}{\log\tau}\right)\right]. \tag{2.44}$$

Proof. We apply Proposition 2.2. It is sufficient to check that, for all $0 < \tau < 1$,

$$\frac{1}{M}\sum_{j=1}^{M} P_j\left(\frac{dP_{0,j}^a}{dP_j} \geq \tau\right) \geq 1-\alpha'$$

with

$$\alpha' = -\frac{\alpha_* + \sqrt{\alpha_*/2}}{\log\tau}\ .$$

By (2.43), we have $P_j \ll P_0$ and $dP_j/dP_0 = dP_j/dP_{0,j}^a$ everywhere except for a set having P_j-measure zero. Then we obtain

$$\begin{aligned} P_j\left(\frac{dP_{0,j}^a}{dP_j} \geq \tau\right) &= P_j\left(\frac{dP_j}{dP_0} \leq \frac{1}{\tau}\right) = 1 - P_j\left(\log\frac{dP_j}{dP_0} > \log\frac{1}{\tau}\right) \\ &\geq 1 - \frac{1}{\log(1/\tau)}\int\left(\log\frac{dP_j}{dP_0}\right)_+ dP_j \quad \text{(Markov's inequality)} \\ &\geq 1 - \frac{1}{\log(1/\tau)}\left[K(P_j,P_0) + \sqrt{K(P_j,P_0)/2}\ \right] \\ &\qquad \text{(2nd Pinsker's inequality).} \end{aligned}$$

By the Jensen inequality and by (2.43),

$$\frac{1}{M}\sum_{j=1}^{M}\sqrt{K(P_j,P_0)} \le \left(\frac{1}{M}\sum_{j=1}^{M}K(P_j,P_0)\right)^{1/2} \le \sqrt{\alpha_*}.$$

Then

$$\frac{1}{M}\sum_{j=1}^{M}P_j\left(\frac{dP_{0,j}^a}{dP_j}\ge\tau\right) \ge 1-\frac{\alpha_*+\sqrt{\alpha_*/2}}{\log(1/\tau)} = 1-\alpha'. \qquad \blacksquare$$

We are now in a position to obtain the following analog of Theorem 2.4 based on Kullback divergences.

Theorem 2.5 (Kullback version of the main theorem). *Assume that $M \ge 2$ and suppose that Θ contains elements $\theta_0, \theta_1, \ldots, \theta_M$ such that:*

(i) $d(\theta_j,\theta_k) \ge 2s > 0, \quad \forall\ 0 \le j < k \le M$;

(ii) $P_j \ll P_0, \quad \forall\ j = 1,\ldots,M$, *and*

$$\frac{1}{M}\sum_{j=1}^{M}K(P_j,P_0) \le \alpha\log M \tag{2.45}$$

with $0 < \alpha < 1/8$ *and* $P_j = P_{\theta_j}, j = 0,1,\ldots,M$. *Then*

$$\inf_{\hat\theta}\sup_{\theta\in\Theta}P_\theta(d(\hat\theta,\theta)\ge s) \ge \frac{\sqrt{M}}{1+\sqrt{M}}\left(1-2\alpha-\sqrt{\frac{2\alpha}{\log M}}\right) > 0. \tag{2.46}$$

Proof. We apply Proposition 2.3 where we set $\alpha_* = \alpha\log M$ and bound from below the supremum over τ on the right hand side of (2.44) by the term with $\tau = 1/\sqrt{M}$. This yields

$$\begin{aligned} p_{e,M} &\ge \frac{\sqrt{M}}{1+\sqrt{M}}\left(1-2\alpha-\sqrt{\frac{2\alpha}{\log M}}\right) \\ &\ge \frac{\sqrt{M}}{1+\sqrt{M}}\left(1-2\alpha-\sqrt{\frac{2\alpha}{\log 2}}\right) > 0 \end{aligned}$$

for $0 < \alpha < 1/8$, giving (2.46) in view of (2.9). $\blacksquare$

We now consider the χ^2 versions of Proposition 2.2 and Theorem 2.4.

Proposition 2.4 *Let $P_0, P_1, \ldots, P_M$ be probability measures on $(\mathcal{X},\mathcal{A})$ satisfying*

$$\frac{1}{M}\sum_{j=1}^{M}\chi^2(P_j,P_0) \le \alpha_* \tag{2.47}$$

with $0 < \alpha_* < \infty$. *Then*

$$p_{e,M} \geq \sup_{0<\tau<1} \left[\frac{\tau M}{1+\tau M} \Big(1 - \tau(\alpha_* + 1) \Big) \right]. \tag{2.48}$$

PROOF. Again, we apply Proposition 2.2. It is sufficient to check that under assumption (2.47), for all $0 < \tau < 1$,

$$\frac{1}{M} \sum_{j=1}^{M} P_j \left(\frac{dP_{0,j}^a}{dP_j} \geq \tau \right) \geq 1 - \tau(\alpha_* + 1). \tag{2.49}$$

As in the proof of Proposition 2.3, we find

$$\begin{aligned} P_j \left(\frac{dP_{0,j}^a}{dP_j} \geq \tau \right) &= P_j \left(\frac{dP_j}{dP_0} \leq \frac{1}{\tau} \right) = 1 - P_j \left(\frac{dP_j}{dP_0} > \frac{1}{\tau} \right) \\ &= 1 - \int \frac{dP_j}{dP_0} \, I \left(\frac{dP_j}{dP_0} > \frac{1}{\tau} \right) dP_0 \\ &\geq 1 - \tau \int \left(\frac{dP_j}{dP_0} \right)^2 dP_0 \quad \text{(Markov's inequality)} \\ &= 1 - \tau \left(\chi^2(P_j, P_0) + 1 \right), \end{aligned}$$

which, together with (2.47), yields (2.49). ∎

Theorem 2.6 (χ^2 version of the main theorem). *Assume that* $M \geq 2$ *and suppose that* Θ *contains elements* $\theta_0, \theta_1, \ldots, \theta_M$ *such that:*

(i) $d(\theta_j, \theta_k) \geq 2s > 0, \quad \forall\ 0 \leq j < k \leq M$;

(ii) $P_j \ll P_0, \quad \forall\ j = 1, \ldots, M$, *and*

$$\frac{1}{M} \sum_{j=1}^{M} \chi^2(P_j, P_0) \leq \alpha M \tag{2.50}$$

with $0 < \alpha < 1/2$ *and* $P_j = P_{\theta_j}, j = 0, 1, \ldots, M$. *Then*

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} P_\theta(d(\hat{\theta}, \theta) \geq s) \geq \frac{1}{2} \left(1 - \alpha - \frac{1}{M} \right) > 0. \tag{2.51}$$

PROOF. Use (2.9) and Proposition 2.4 setting there $\alpha_* = \alpha M$ and bounding from below the supremum over τ on the right hand side of (2.48) by the term with $\tau = 1/M$. ∎

Comparison of (2.45) and (2.50) shows that, to derive valid lower bounds, we can allow the χ^2 divergences between P_j and P_0 to be of much larger order than the Kullback ones, as $M \to \infty$.

The results of this section are valid for $M \geq 2$. Combining them with Theorem 2.2 that treats the case $M = 1$ and considering general loss functions, which is an easy extension (cf. (2.5)), we get the following theorem.

Theorem 2.7 *Let w be a loss function satisfying (2.3), and let $A > 0$ be such that $w(A) > 0$. Assume that Θ contains elements $\theta_0, \theta_1, \ldots, \theta_M$, $M \geq 1$, such that:*

(i) $d(\theta_j, \theta_k) \geq 2s > 0, \quad \forall\ 0 \leq j < k \leq M$;

(ii) $P_j \ll P_0, \quad \forall\ j = 1, \ldots, M$, *and*

$$\frac{1}{M}\sum_{j=1}^{M} K(P_j, P_0) \leq \alpha \log M \quad \text{or} \quad \frac{1}{M}\sum_{j=1}^{M} \chi^2(P_j, P_0) \leq \alpha M, \tag{2.52}$$

with $0 < \alpha < 1/8$ and $P_j = P_{\theta_j}, j = 0, 1, \ldots, M$.

Then for $\psi = s/A$ we have

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} E_\theta \Big[w(\psi^{-1} d(\hat{\theta}, \theta)) \Big] \geq c(\alpha) w(A),$$

where $\inf_{\hat{\theta}}$ denotes the infimum over all estimators and $c(\alpha) > 0$ is a constant depending only on α.

Proof. Combine (2.5), (2.9) and Theorems 2.2, 2.5, 2.6. ∎

Remarks.

(1) In the sequel we will use the bounds (2.42), (2.46), and (2.51) with $M = M_n$ depending on n such that $M_n \to \infty$ as $n \to \infty$. Note that the right hand side of (2.46) becomes arbitrarily close to 1 as $M \to \infty$ and $\alpha \to 0$. Moreover, it follows from the proof of Theorem 2.5 that

$$\liminf_{M \to \infty} p_{e,M} \geq 1 - 2\alpha. \tag{2.53}$$

In other words, the right hand side of (2.44) with $\alpha_* = \alpha \log M$ can be arbitrarily close to 1 for sufficiently large M and small α, in contrast to the bounds based on two hypotheses obtained in Sections 2.3 and 2.4.2. An example of application of this property is given in Exercise 2.9.

(2) For finite M, the constants in (2.46) and (2.51) are not optimal. They can be improved, for example, by direct computation of the maximum over $0 < \tau < 1$ in (2.44), (2.48) (Exercise 2.6) or by taking $\tau = M^{-\gamma}$ with $0 <$

$\gamma < 1$ and maximizing with respect to γ. More accurate evaluations in the Kullback case can be obtained using Fano's lemma (see Section 2.7.1). These modifications are not of a great importance in the context of this chapter, since here we are only interested in the rates of convergence. From the very beginning, we follow the general scheme of Section 2.2 based on rather rough inequalities. Therefore, improving bounds for $p_{e,M}$ will still leave the final result inaccurate in what concerns the constants. Recall that the scheme of Section 2.2 is quite general and can be applied to any estimation problem. The reader will not be surprised by the fact that the corresponding bounds are not the most accurate: this is a price to pay for generality. More refined methods should be applied, case by case, if we would like to optimize not only the rate of convergence but also the constants. This is only available for some remarkable problems; an example will be given in Chapter 3.

2.6.1 Lower bounds in L_2

Theorems 2.4 and 2.5 enable us to obtain lower bounds for the L_p risk with optimal rates. To illustrate this, consider the nonparametric regression model under Assumption (B) and let us focus on the L_2 risk. Then

$$d(f,g) = \|f-g\|_2 = \Big(\int_0^1 (f(x)-g(x))^2 dx \Big)^{1/2}. \tag{2.54}$$

Our first aim is to prove the lower bound (2.2) on the minimax risk for the Hölder class $\Theta = \Sigma(\beta, L)$ and the L_2 distance (2.54), with the rate

$$\psi_n = n^{-\frac{\beta}{2\beta+1}}.$$

Let M be an integer to be specified later on. Consider the following hypotheses:

$$\theta_j = f_{jn}(\cdot),\ j = 0, \dots, M,$$

where $f_{jn} \in \Sigma(\beta, L)$. By the general scheme of Section 2.2, it is sufficient to prove that

$$\inf_{\hat\theta_n} \max_{\theta \in \{\theta_0, \dots, \theta_M\}} P_\theta(d(\hat\theta_n, \theta) \ge s) \ge c' > 0,$$

where $s = A\psi_n$ and $A > 0$. If $\Theta = \Sigma(\beta, L)$ and if d is the L_2-distance, this inequality becomes

$$\inf_{T_n} \max_{f \in \{f_{0n}, \dots, f_{Mn}\}} P_f(\|T_n - f\|_2 \ge A\psi_n) \ge c' > 0, \tag{2.55}$$

where $\inf_{T_n}$ denotes the infimum over all estimators T_n. We will apply Theorem 2.5 to obtain (2.55). First, we define the functions f_{jn} that will be used in the proof.

Construction of the hypotheses f_{jn}

Take a real number $c_0 > 0$ and an integer $m \geq 1$. Define

$$m = \lceil c_0 n^{\frac{1}{2\beta+1}} \rceil, \quad h_n = \frac{1}{m}, \quad x_k = \frac{k - 1/2}{m},$$

$$\varphi_k(x) = L h_n^{\beta} K\left(\frac{x - x_k}{h_n}\right), \quad k = 1, \ldots, m, \quad x \in [0,1], \tag{2.56}$$

where $K : \mathbf{R} \to [0, +\infty)$ is a function satisfying (2.33). In what follows we denote by $\lceil x \rceil$ the smallest integer which is strictly greater than $x \in \mathbf{R}$. In view of (2.35), all the functions φ_k belong to $\Sigma(\beta, L/2)$. Consider the set of all binary sequences of length m:

$$\Omega = \big\{\omega = (\omega_1, \ldots, \omega_m),\ \omega_i \in \{0,1\}\big\} = \{0,1\}^m.$$

The hypotheses f_{jn} will be chosen in the collection of functions

$$\mathcal{E} = \Big\{ f_\omega(x) = \sum_{k=1}^m \omega_k \varphi_k(x),\ \omega \in \Omega \Big\}.$$

For all $\omega, \omega' \in \Omega$, we have

$$\begin{aligned} d(f_\omega, f_{\omega'}) &= \Big[\int_0^1 (f_\omega(x) - f_{\omega'}(x))^2 dx \Big]^{1/2} \\ &= \Big[\sum_{k=1}^m (\omega_k - \omega'_k)^2 \int_{\Delta_k} \varphi_k^2(x) dx \Big]^{1/2} \\ &= L h_n^{\beta+\frac{1}{2}} \|K\|_2 \Big[\sum_{k=1}^m (\omega_k - \omega'_k)^2 \Big]^{1/2} \\ &= L h_n^{\beta+\frac{1}{2}} \|K\|_2 \sqrt{\rho(\omega, \omega')}, \end{aligned} \tag{2.57}$$

where $\rho(\omega, \omega') = \sum_{k=1}^m I(\omega_k \neq \omega'_k)$ is the *Hamming distance* between the binary sequences $\omega = (\omega_1, \ldots, \omega_m)$ and $\omega' = (\omega'_1, \ldots, \omega'_m)$, and where Δ_k are the intervals

$$\Delta_1 = [0, 1/m], \qquad \Delta_k = ((k-1)/m, k/m], \quad k = 2, \ldots, m. \tag{2.58}$$

The set $\{f_{jn}, j = 0, \ldots, M\}$ will be composed of certain functions f_ω selected in $\mathcal{E}$. In order to apply Theorem 2.5, we need that any two functions $f_\omega, f_{\omega'}$ belonging to the selected set $\{f_{jn}, j = 0, \ldots, M\}$ satisfy the property $d(f_\omega, f_{\omega'}) \geq 2s_n \asymp n^{-\frac{\beta}{2\beta+1}}$. Therefore, it suffices to choose ω, ω' such that $\sqrt{\rho(\omega, \omega')} \asymp h_n^{-1/2}$, which is equivalent to $\rho(\omega, \omega') \asymp m$. Then the following

question arises: How massive can be the set of all binary sequences ω with pairwise separation by the Hamming distance of at least $\sim m$? A lower bound for the cardinality of this set is given by a result in information theory known under the name of the *Varshamov–Gilbert bound*. In order to prove this bound, we first introduce an exponential inequality for sums of independent bounded random variables.

Lemma 2.8 (Hoeffding's inequality). *Let $Z_1, \dots, Z_m$ be independent random variables such that $a_i \le Z_i \le b_i$. Then for all $t > 0$*

$$\mathbf{P}\left(\sum_{i=1}^{m}(Z_i - \mathbf{E}(Z_i)) \ge t\right) \le \exp\left(-\frac{2t^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right).$$

The proof of this lemma is given in Appendix (Lemma A.4).

Lemma 2.9 (Varshamov–Gilbert bound). *Let $m \ge 8$. Then there exists a subset $\{\omega^{(0)}, \dots, \omega^{(M)}\}$ of Ω such that $\omega^{(0)} = (0, \dots, 0)$,*

$$\rho(\omega^{(j)}, \omega^{(k)}) \ge \frac{m}{8}, \quad \forall\ 0 \le j < k \le M, \tag{2.59}$$

and

$$M \ge 2^{m/8}. \tag{2.60}$$

Proof. It is clear that Card $\Omega = 2^m$. Take $\omega^{(0)} = (0, \dots, 0)$ and exclude all $\omega \in \Omega$ belonging to the D-neighborhood of $\omega^{(0)}$, that is, such that $\rho(\omega, \omega^{(0)}) \le D \triangleq \lfloor m/8 \rfloor$. Set

$$\Omega_1 = \{\omega \in \Omega : \rho(\omega, \omega^{(0)}) > D\}.$$

Take as $\omega^{(1)}$ an arbitrary element of Ω_1. Then exclude all $\omega \in \Omega_1$ such that $\rho(\omega, \omega^{(1)}) \le D$, etc. In this way, we recurrently define subsets Ω_j of Ω:

$$\Omega_j = \{\omega \in \Omega_{j-1} : \rho(\omega, \omega^{(j-1)}) > D\}, \qquad j = 1, \dots, M,$$

where $\Omega_0 \triangleq \Omega$, $\omega^{(j)}$ is an arbitrary element of Ω_j and M is the smallest integer satisfying $\Omega_{M+1} = \emptyset$. Let n_j be the number of vectors ω excluded from the D-neighborhood of $\omega^{(j)}$ at the jth step of this procedure, that is, $n_j = \text{Card } A_j$ where

$$A_j = \{\omega \in \Omega_j : \rho(\omega, \omega^{(j)}) \le D\}, \qquad j = 0, \dots, M.$$

From the definition of the Hamming distance, we obtain the bound

$$n_j \le \sum_{i=0}^{D}\binom{m}{i}, \quad j = 0, \dots, M.$$

Since $A_0, \ldots, A_M$ are disjoint sets forming a partition of Ω, we have

$$n_0 + n_1 + \cdots + n_M = \text{Card } \Omega = 2^m.$$

Therefore,

$$(M+1) \sum_{i=0}^{D} \binom{m}{i} \geq 2^m. \tag{2.61}$$

Moreover, $\rho(\omega^{(j)}, \omega^{(k)}) \geq D + 1 = \lfloor m/8 \rfloor + 1 \geq m/8, \forall j \neq k$, by construction of the sequence $\omega^{(j)}$. We can write (2.61) as follows:

$$M + 1 \geq \frac{1}{p^*},$$

where p^* is the binomial probability

$$p^* = \sum_{i=0}^{D} 2^{-m} \binom{m}{i} = \mathbf{P}(Bi(m, 1/2) \leq \lfloor m/8 \rfloor),$$

$Bi(m, 1/2) = \sum_{i=1}^{m} Z_i$ and Z_i are i.i.d. Bernoulli random variables with parameter $1/2$. Since $0 \leq Z_i \leq 1$ and $\mathbf{E}(Z_i) = 1/2$, the Hoeffding inequality implies that

$$p^* \leq \exp(-9m/32) < 2^{-m/4}.$$

Therefore $M + 1 \geq 2^{m/4} \geq 2^{m/8} + 1$ for $m \geq 8$. ∎

Finally, we define

$$f_{jn}(x) = f_{\omega^{(j)}}(x), \quad j = 0, \ldots, M,$$

where $\{\omega^{(0)}, \ldots, \omega^{(M)}\}$ is a subset of Ω satisfying the assumptions of Lemma 2.9.

Application of Theorem 2.5

Fix $\alpha \in (0, 1/8)$. In order to apply Theorem 2.5 we need to check the following three conditions:

(a) $f_{jn} \in \Sigma(\beta, L),\ j = 0, \ldots, M,$

(b) $d(\theta_j, \theta_k) = \|f_{jn} - f_{kn}\|_2 \geq 2s > 0,\ 0 \leq j < k \leq M,$

(c) $\dfrac{1}{M} \sum_{j=1}^{M} K(P_j, P_0) \leq \alpha \log M.$

We will now show that these conditions are satisfied for all sufficiently large n.

(a) *The condition* $f_{jn} \in \Sigma(\beta, L)$.

Since $\varphi_k \in \Sigma(\beta, L/2)$, $|\omega_i| \le 1$ and the functions φ_k have disjoint supports, we have $f_\omega \in \Sigma(\beta, L)$ for all $\omega \in \Omega$.

(b) *The condition* $\|f_{jn} - f_{kn}\|_2 \ge 2s$.

By (2.57) and (2.59), we obtain

$$\begin{aligned}\|f_{jn} - f_{kn}\|_2 &= \|f_{\omega^{(j)}} - f_{\omega^{(k)}}\|_2 \\ &= Lh_n^{\beta+1/2}\|K\|_2\sqrt{\rho(\omega^{(j)}, \omega^{(k)})} \\ &\ge Lh_n^{\beta+\frac{1}{2}}\|K\|_2\sqrt{\frac{m}{16}} \\ &= \frac{L}{4}\|K\|_2 h_n^\beta = \frac{L}{4}\|K\|_2 m^{-\beta},\end{aligned}$$

whenever $m \ge 8$. Suppose that $n \ge n_*$ where $n_* = (7/c_0)^{2\beta+1}$. Then $m \ge 8$ and $m^\beta \le (1 + 1/7)^\beta c_0^\beta n^{\frac{\beta}{2\beta+1}} \le (2c_0)^\beta n^{\frac{\beta}{2\beta+1}}$, implying

$$\|f_{jn} - f_{kn}\|_2 \ge 2s$$

with

$$s = An^{-\frac{\beta}{2\beta+1}} = A\psi_n, \qquad A = \frac{L}{8}\|K\|_2(2c_0)^{-\beta}.$$

(c) *The condition* $\dfrac{1}{M}\sum_{j=1}^{M} K(P_j, P_0) \le \alpha \log M$.

As in (2.36) we have, for all $n \ge n_*$,

$$\begin{aligned}K(P_j, P_0) &\le p_* \sum_{i=1}^{n} f_{jn}^2(X_i) \le p_* \sum_{k=1}^{m} \sum_{i: X_i \in \Delta_k} \varphi_k^2(X_i) \\ &\le p_* L^2 K_{\max}^2 h_n^{2\beta} \sum_{k=1}^{m} \mathrm{Card}\{i : X_i \in \Delta_k\} \\ &= p_* L^2 K_{\max}^2 n h_n^{2\beta} \le p_* L^2 K_{\max}^2 c_0^{-(2\beta+1)} m.\end{aligned}$$

By (2.60), $m \le 8 \log M / \log 2$. Therefore if we choose

$$c_0 = \left(\frac{8p_* L^2 K_{\max}^2}{\alpha \log 2}\right)^{\frac{1}{2\beta+1}},$$

then $K(P_j, P_0) < \alpha \log M$, $j = 1, \dots, M$.

We conclude that the assumptions of Theorem 2.5 are satisfied. Therefore, for any estimator T_n,

$$\max_{f\in\{f_{0n},\ldots,f_{Mn}\}} P_f(\|T_n - f\|_2 \geq A\psi_n) \geq \frac{\sqrt{M}}{1+\sqrt{M}}\left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log M}}\right),$$

implying the following result.

Theorem 2.8 *Let $\beta > 0$ and $L > 0$. Under Assumption (B) we have*

$$\liminf_{n\to\infty} \inf_{T_n} \sup_{f\in\Sigma(\beta,L)} \mathbf{E}_f\left[n^{\frac{2\beta}{2\beta+1}}\|T_n - f\|_2^2\right] \geq c \tag{2.62}$$

where $\inf_{T_n}$ denotes the infimum over all estimators and where the constant $c > 0$ depends only on β, L and p_*.

This theorem and Theorem 1.7 imply the following corollary.

Corollary 2.3 *Consider the nonparametric regression model under the following conditions:*

(i) $X_i = i/n$ *for* $i = 1, \ldots, n$;
(ii) the random variables ξ_i are i.i.d. with density p_ξ satisfying (2.29) and such that

$$\mathbf{E}(\xi_i) = 0, \qquad \mathbf{E}(\xi_i^2) < \infty.$$

Then, for all $\beta > 0$ and $L > 0$, the rate of convergence $\psi_n = n^{-\frac{\beta}{2\beta+1}}$ is optimal on $(\Sigma(\beta, L), \|\cdot\|_2)$.

Moreover, for $\ell = \lfloor\beta\rfloor$ the local polynomial estimator $LP(\ell)$ with kernel K and bandwidth h_n satisfying assumptions (iii) and (iv) of Theorem 1.7 is rate optimal on $(\Sigma(\beta, L), \|\cdot\|_2)$.

Sobolev classes

The construction described in this section can also be used to obtain a lower bound for the minimax risk on $(W^{per}(\beta, L), \|\cdot\|_2)$ and therefore a fortiori on $(W(\beta, L), \|\cdot\|_2)$ where $\beta \in \{1, 2, \ldots\}$, $L > 0$.

Indeed, if $K(\cdot)$ is defined by (2.34), the functions f_ω as well as all their derivatives are periodic on $[0, 1]$. Moreover, $f_\omega \in W(\beta, L)$ since $f_\omega \in \Sigma(\beta, L)$ and $\Sigma(\beta, L) \subset W(\beta, L)$. Therefore the functions $f_{0n}, \ldots, f_{Mn}$ introduced above belong to $W^{per}(\beta, L)$ and the argument of this section leads to the following result.

Theorem 2.9 *Let $\beta \in \{1, 2, \ldots\}$ and $L > 0$. Under Assumption (B) we have*

$$\liminf_{n\to\infty} \inf_{T_n} \sup_{f\in W^{per}(\beta,L)} \mathbf{E}_f\left[n^{\frac{2\beta}{2\beta+1}}\|T_n - f\|_2^2\right] \geq c$$

where $\inf_{T_n}$ denotes the infimum over all estimators and where the constant $c > 0$ depends only on β, L, and p_*.

This theorem and Theorem 1.9 imply the following corollary.

Corollary 2.4 *Consider the nonparametric regression model under the following conditions:*

(i) $X_i = i/n$ *for* $i = 1, \dots, n$;
(ii) the random variables ξ_i *are i.i.d. with density* p_ξ *satisfying (2.29) and such that*

$$\mathbf{E}(\xi_i) = 0, \qquad \mathbf{E}(\xi_i^2) < \infty.$$

Then, for $\beta \in \{1, 2, \dots\}$ *and* $L > 0$, *the rate of convergence* $\psi_n = n^{-\frac{\beta}{2\beta+1}}$ *is optimal on* $(W^{per}(\beta, L), \|\cdot\|_2)$.

Moreover, the simple projection estimator satisfying the assumptions of Theorem 1.9 is rate optimal on $(W^{per}(\beta, L), \|\cdot\|_2)$.

Finally note that the techniques of this section can be used to establish lower bounds, similar to those of Theorem 2.8, for the problem of estimation of a probability density (cf. Exercise 2.10).

2.6.2 Lower bounds in the sup-norm

We remain here in the framework of nonparametric regression under Assumption (B). However, we suppose now that the semi-distance $d(\cdot, \cdot)$ is defined as follows:

$$d(f, g) = \|f - g\|_\infty = \sup_{x \in [0,1]} |f(x) - g(x)|.$$

Our aim is to obtain the lower bound (2.2) for $(\Theta, d) = (\Sigma(\beta, L), \|\cdot\|_\infty)$ with the rate

$$\psi_n = \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+1}}.$$

For this purpose we apply again Theorem 2.5. Define the hypotheses:

$$\theta_0 = f_{0n}(\cdot) \equiv 0,$$
$$\theta_j = f_{jn}(\cdot), \quad j = 1, \dots, M,$$

with

$$f_{jn}(x) = L h_n^\beta K\left(\frac{x - x_j}{h_n}\right), \quad x_j = \frac{j - 1/2}{M}, \quad h_n = 1/M,$$

where $K : \mathbf{R} \to [0, +\infty)$ is a function satisfying (2.33) and $M > 1$ is an integer.

Fix $\alpha \in (0, 1/8)$. In order to apply Theorem 2.5, we have to check the following conditions:

(a) $f_{jn} \in \Sigma(\beta, L),\ j = 1, \dots, M,$

(b) $d(f_{jn}, f_{kn}) \geq 2s > 0,\ \forall k \neq j,$

(c) $\frac{1}{M}\sum_{j=1}^{M} K(P_j, P_0) \le \alpha \log M.$

Let us show that these conditions hold if n is sufficiently large.

(a) *The condition* $f_{jn} \in \Sigma(\beta, L)$: It holds in view of (2.35).

(b) *The condition* $d(f_{jn}, f_{kn}) \ge 2s$. We have

$$d(f_{jn}, f_{kn}) = \|f_{jn} - f_{kn}\|_\infty \ge L h_n^\beta K(0) \triangleq 2s,$$

where

$$s = \frac{L h_n^\beta K(0)}{2} .$$

We need to have $s \asymp \psi_n = \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+1}}$. Therefore we choose $h_n \asymp \left(\frac{\log n}{n}\right)^{\frac{1}{2\beta+1}}$. To be more explicit, we define $h_n = 1/M$ with

$$M = \left\lceil c_0 \left(\frac{n}{\log n}\right)^{\frac{1}{2\beta+1}} \right\rceil ,$$

where $c_0 > 0$ is a constant to be chosen later.

(c) *Condition* $\frac{1}{M}\sum_{j=1}^{M} K(P_j, P_0) \le \alpha \log M.$

By (2.36), we obtain

$$\begin{aligned}
\frac{1}{M}\sum_{j=1}^{M} K(P_j, P_0) &\le \frac{1}{M}\sum_{j=1}^{M} p_* \sum_{i=1}^{n} f_{jn}^2(X_i) \\
&\le p_* L^2 K_{\max}^2 h_n^{2\beta} \frac{1}{M}\sum_{j=1}^{M} \mathrm{Card}\{i : X_i \in \mathrm{supp}(f_{jn})\} \\
&= p_* L^2 K_{\max}^2 h_n^{2\beta} n/M = p_* L^2 K_{\max}^2 M^{-(2\beta+1)} n \\
&\le p_* L^2 K_{\max}^2 c_0^{-(2\beta+1)} \log n
\end{aligned}$$

where $\mathrm{supp}(f_{jn})$ denotes the support of the function f_{jn}. We have

$$\log M \ge \log\left(c_0 \left(\frac{n}{\log n}\right)^{\frac{1}{2\beta+1}}\right) = \frac{\log n}{2\beta+1}(1 + o(1)) \ge \frac{\log n}{2\beta+2}$$

for sufficiently large n. We conclude by choosing c_0 sufficiently large.

We have therefore proved the following theorem.

Theorem 2.10 *Let $\beta > 0$ and $L > 0$. Under Assumption (B), we have:*

$$\liminf_{n\to\infty} \inf_{T_n} \sup_{f\in\Sigma(\beta,L)} \left(\frac{n}{\log n}\right)^{\frac{2\beta}{2\beta+1}} \mathbf{E}_f \|T_n - f\|_\infty^2 \geq c,$$

where $\inf_{T_n}$ denotes the infimum over all estimators and where the constant $c > 0$ depends only on β, L, and p_*.

This theorem and Theorem 1.8 imply the following corollary.

Corollary 2.5 *Consider the nonparametric regression model under the following assumptions:*

(i) $X_i = i/n$ for $i = 1, \ldots, n$;
(ii) the random variables ξ_i are i.i.d. Gaussian with distribution $\mathcal{N}(0, \sigma_\xi^2)$ where $0 < \sigma_\xi^2 < \infty$.

Then for $\beta > 0$ and $L > 0$ the rate of convergence

$$\psi_n = \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+1}}$$

is optimal on $(\Sigma(\beta, L), \|\cdot\|_\infty)$.

Moreover, the local polynomial estimator $LP(\ell)$ for $\ell = \lfloor\beta\rfloor$, with kernel K and bandwidth h_n satisfying the assumptions of Theorem 1.8, is rate optimal on $(\Sigma(\beta, L), \|\cdot\|_\infty)$.

Observe that, by Corollaries 2.3 and 2.4, optimal rates of convergence in the L_2-norm on the Sobolev classes are the same as those on the Hölder classes. It is interesting to note that the situation becomes different for estimation in the L_∞-norm; here optimal rates on the Sobolev classes are substantially slower (cf. Exercise 2.11).

2.7 Other tools for minimax lower bounds

We are going to present now some more techniques for proving lower bounds on the minimax risk. This material can be omitted in the first reading.

2.7.1 Fano's lemma

The general scheme of Section 2.2 suggests a way to prove minimax lower bounds by switching to the minimax probability of error $p_{e,M}$. Our main efforts in this chapter have been devoted to the construction of lower bounds for $p_{e,M}$. Fano's lemma allows us to obtain similar results in a different way: by switching to a smaller quantity which is the average probability of error. Note

that, for the case of two hypotheses, bounds based on the average probability of error already appeared in Section 2.4.2.

Let $P_0, P_1, \ldots, P_M$ be probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$. For a test $\psi : \mathcal{X} \to \{0, 1, \ldots, M\}$, define the *average probability of error* and the minimum average probability of error by

$$\overline{p}_{e,M}(\psi) = \frac{1}{M+1} \sum_{j=0}^{M} P_j(\psi \neq j)$$

and

$$\overline{p}_{e,M} = \inf_{\psi} \overline{p}_{e,M}(\psi),$$

respectively. Introduce a probability measure $\overline{P}$ on $(\mathcal{X}, \mathcal{A})$ in the following way:

$$\overline{P} = \frac{1}{M+1} \sum_{j=0}^{M} P_j \ .$$

Lemma 2.10 (Fano's lemma). *Let $P_0, P_1, \ldots, P_M$ be probability measures on $(\mathcal{X}, \mathcal{A})$, $M \geq 1$. Then $\overline{p}_{e,M} \leq M/(M+1)$ and*

$$g(\overline{p}_{e,M}) \geq \log(M+1) - \frac{1}{M+1} \sum_{j=0}^{M} K(P_j, \overline{P}) \tag{2.63}$$

where, for $0 \leq x \leq 1$,

$$g(x) = x \log M + \mathcal{H}(x)$$

with $\mathcal{H}(x) = -x \log x - (1-x) \log(1-x)$ and $0 \log 0 \stackrel{\triangle}{=} 0$.

Proof. We have

$$\overline{p}_{e,M}(\psi) = \frac{1}{M+1} \mathbf{E}_{\overline{P}} \left[\sum_{j=0}^{M} I(A_j) \frac{dP_j}{d\overline{P}} \right] = \mathbf{E}_{\overline{P}} \Big[\sum_{j=0}^{M} b_j p_j \Big] \tag{2.64}$$

with $A_j = \{\psi \neq j\}$, $b_j = I(A_j)$,

$$p_j = \frac{dP_j}{(M+1) d\overline{P}} \ ,$$

where $\mathbf{E}_{\overline{P}}$ denotes the expectation with respect to the measure $\overline{P}$. The random variables b_j and p_j satisfy $\overline{P}$-almost surely the following conditions:

$$\sum_{j=0}^{M} b_j = M, \quad b_j \in \{0, 1\}, \qquad \text{and} \qquad \sum_{j=0}^{M} p_j = 1, \quad p_j \geq 0.$$

Then we have that, $\overline{P}$-almost surely,

$$\sum_{j=0}^{M} b_j p_j = \sum_{j \neq j_0} p_j \tag{2.65}$$

where j_0 is a random number, $0 \leq j_0 \leq M$. Apply now the following lemma, which will be proved later on.

Lemma 2.11 *For all $j_0 \in \{0, 1, \ldots, M\}$ and all real numbers $p_0, p_1 \ldots, p_M$, such that $\sum_{j=0}^{M} p_j = 1, \quad p_j \geq 0$, we have*

$$g\Big(\sum_{j \neq j_0} p_j \Big) \geq - \sum_{j=0}^{M} p_j \log p_j \tag{2.66}$$

where $0 \log 0 \stackrel{\triangle}{=} 0$.

The function $g(x) = x \log M + \mathcal{H}(x)$ is concave for $0 \leq x \leq 1$. Using (2.64), the Jensen inequality, and formulas (2.65) and (2.66) we obtain that, for any test ψ,

$$\begin{aligned} g(\overline{p}_{e,M}(\psi)) &= g\Big(\mathbf{E}_{\overline{P}} \Big[\sum_{j=0}^{M} b_j p_j \Big] \Big) \geq \mathbf{E}_{\overline{P}} g\Big(\sum_{j=0}^{M} b_j p_j \Big) \\ &\geq \mathbf{E}_{\overline{P}} \Big[- \sum_{j=0}^{M} p_j \log p_j \Big] \\ &= - \mathbf{E}_{\overline{P}} \left[\sum_{j=0}^{M} \frac{dP_j}{(M+1) d\overline{P}} \log \frac{dP_j}{(M+1) d\overline{P}} \right] \\ &= \log(M+1) - \frac{1}{M+1} \sum_{j=0}^{M} K(P_j, \overline{P}). \end{aligned}$$

Since there exists a sequence of tests $\{\psi^k\}_{k=0}^{\infty}$ such that $\overline{p}_{e,M}(\psi^k) \to \overline{p}_{e,M}$ as $k \to \infty$, we obtain by continuity of g

$$g(\overline{p}_{e,M}) = \lim_{k \to \infty} g(\overline{p}_{e,M}(\psi^k)) \geq \log(M+1) - \frac{1}{M+1} \sum_{j=0}^{M} K(P_j, \overline{P}).$$

It remains to prove that $\overline{p}_{e,M} \leq M/(M+1)$. For this purpose, we define a degenerate test $\psi_* \equiv 1$ and observe that

$$\inf_{\psi} \overline{p}_{e,M}(\psi) \leq \overline{p}_{e,M}(\psi_*) = \frac{1}{M+1} \sum_{j=0}^{M} P_j(j \neq 1) = \frac{M}{M+1} \ . \qquad \blacksquare$$

Proof of Lemma 2.11. It is sufficient to prove the result under the assumption $\sum_{j \neq j_0} p_j \neq 0$ since otherwise inequality (2.66) is clear. We have

$$\sum_{j=0}^{M} p_j \log p_j = p_{j_0} \log p_{j_0} + \Big(\sum_{j \neq j_0} p_j \Big) \log \Big(\sum_{j \neq j_0} p_j \Big) \tag{2.67}$$

$$+ \sum_{j \neq j_0} p_j \log \frac{p_j}{\sum_{i \neq j_0} p_i}$$

$$= -\mathcal{H}\Big(\sum_{j \neq j_0} p_j \Big) + \Big(\sum_{j \neq j_0} p_j \Big) \Big(\sum_{j \neq j_0} q_j \log q_j \Big)$$

with

$$q_j = \frac{p_j}{\sum_{i \neq j_0} p_i}\ , \quad \sum_{j \neq j_0} q_j = 1, \quad q_j \geq 0.$$

Suppose that $q_j > 0$; the case of $q_j = 0$ requires a trivial modification. Since the function $-\log x$ is convex for $x > 0$, we obtain by the Jensen inequality

$$\sum_{j \neq j_0} q_j \log q_j = -\sum_{j \neq j_0} q_j \log(1/q_j) \geq -\log M.$$

Lemma 2.11 follows from this inequality and (2.67). ∎

Using Fano's lemma we can bound from below the minimax probability of error $p_{e,M}$ in the following way:

$$p_{e,M} = \inf_{\psi} \max_{0 \leq j \leq M} P_j(\psi \neq j) \geq \inf_{\psi} \overline{p}_{e,M}(\psi) = \overline{p}_{e,M}$$

$$\geq g^{-1} \left(\log(M+1) - \frac{1}{M+1} \sum_{j=0}^{M} K(P_j, \overline{P}) \right), \tag{2.68}$$

where $g^{-1}(t) \stackrel{\triangle}{=} 0$ for $t < 0$ and, for $0 < t < \log(M+1)$, $g^{-1}(t)$ is a solution of the equation $g(x) = t$ with respect to $x \in [0, M/(M+1)]$; this solution exists since the function g is continuous and increasing on $[0, M/(M+1)]$ and $g(0) = 0$, $g(M/(M+1)) = \log(M+1)$. Then lower bounds on the minimax risk can be obtained following the general scheme of Section 2.2 and using inequality (2.68). It is sufficient to assure that the quantity

$$\log(M+1) - \frac{1}{M+1} \sum_{j=0}^{M} K(P_j, \overline{P})$$

is positive. We can check this fact in two ways. The first method is due to Ibragimov and Has'minskii who introduced Fano's lemma in the context of nonparametric estimation. Suppose that the measures P_j are mutually absolutely continuous; then one can readily see that

$$\frac{1}{M+1} \sum_{j=0}^{M} K(P_j, \overline{P}) \leq \frac{1}{(M+1)^2} \sum_{j=0}^{M} \sum_{k=0}^{M} K(P_j, P_k).$$

Thus, in order to obtain a nontrivial lower bound, it is sufficient to choose measures P_j satisfying $\max_{0\le j,k\le M} K(P_j,P_k)\le \alpha\log(M+1)$ with $0<\alpha<1$. The second method (which is more general since it does not require all the measures P_j to be mutually absolutely continuous) is based on the elementary equality

$$\frac{1}{M+1}\sum_{j=0}^{M} K(P_j,P_0)=\frac{1}{M+1}\sum_{j=0}^{M} K(P_j,\overline{P})+K(\overline{P},P_0). \tag{2.69}$$

Since $K(\overline{P},P_0)\ge 0$, inequalities (2.63) and (2.69) imply that

$$g(\overline{p}_{e,M})\ge \log(M+1)-\frac{1}{M+1}\sum_{j=1}^{M} K(P_j,P_0) \tag{2.70}$$

giving

$$\overline{p}_{e,M}\ge g^{-1}\Big(\log(M+1)-\alpha\log M\Big) \tag{2.71}$$

whenever $(M+1)^{-1}\sum_{j=1}^{M} K(P_j,P_0)\le \alpha\log M$ with $0<\alpha<1$ and $M\ge 2$. Unfortunately, inequality (2.71) is not explicit enough, since it contains the inverse function of g. A more explicit solution can be obtained if we simplify (2.71) in the following way.

Corollary 2.6 *Let $P_0,P_1,\dots,P_M$ be probability measures on $(\mathcal{X},\mathcal{A})$, $M\ge 2$. If*

$$\frac{1}{M+1}\sum_{j=1}^{M} K(P_j,P_0)\le \alpha\log M$$

with $0<\alpha<1$, then

$$p_{e,M}\ge \overline{p}_{e,M}\ge \frac{\log(M+1)-\log 2}{\log M}-\alpha. \tag{2.72}$$

Proof. It is sufficient to use the inequality $p_{e,M}\ge\overline{p}_{e,M}$, formula (2.70) and the fact that $\mathcal{H}(x)\le\log 2$ for $0\le x\le 1$. ∎

For $M=1$, inequality (2.70) gives

$$p_{e,1}\ge\overline{p}_{e,1}\ge\mathcal{H}^{-1}(\log 2-\alpha/2) \tag{2.73}$$

whenever $K(P_1,P_0)\le\alpha<\infty$ where

$$\mathcal{H}^{-1}(t)=\min\{p\in[0,1/2]:\mathcal{H}(p)\ge t\}.$$

Note that the bound (2.73) is coarser than the following one, obtained from part (iii) of Theorem 2.2 under the same conditions:

$$p_{e,1} \geq \overline{p}_{e,1} \geq \max\left(\frac{1}{4}\exp(-\alpha), \frac{1-\sqrt{\alpha/2}}{2}\right). \tag{2.74}$$

Indeed, the bound (2.74) is nontrivial for all $\alpha > 0$ while the term on the right hand side of (2.73) is positive for $\alpha < 2\log 2$ only. Moreover, for α sufficiently close to 0, which is the most interesting case in our context, the bound (2.73) is less accurate than (2.74).

Remarks.

(1) By taking the limit in (2.72) as $M \to \infty$, we come again to (2.53); in fact, we obtain a slightly stronger inequality:

$$\liminf_{M\to\infty} \overline{p}_{e,M} \geq 1 - \alpha. \tag{2.75}$$

(2) Corollary 2.6 is essentially of the same type as Proposition 2.3, except that it holds for the minimum average probability $\overline{p}_{e,M}$ and not only for the minimax probability $p_{e,M}$. This property is useful in certain applications, especially in obtaining lower bounds on the minimax risk in the nonparametric regression model with arbitrary design $X_1, \ldots, X_n$. Indeed, assume that we deal with the following framework.

Assumption (B1)
Conditions (i) and (ii) of Assumption (B) are satisfied and X_i are arbitrary random variables taking values in $[0,1]$ such that $(X_1, \ldots, X_n)$ is independent of $(\xi_1, \ldots, \xi_n)$.

Using (2.72) we obtain the following result.

Theorem 2.11 *Let $\beta > 0$ and $L > 0$. Under Assumption (B1), for $p = 2$ or $p = \infty$, and*

$$\psi_{n,2} = n^{-\frac{\beta}{2\beta+1}}, \quad \psi_{n,\infty} = \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+1}},$$

we have

$$\liminf_{n\to\infty} \inf_{T_n} \sup_{f\in\Sigma(\beta,L)} \mathbf{E}_f\left[\psi_{n,p}^{-2}\|T_n - f\|_p^2\right] \geq c$$

where $\inf_{T_n}$ denotes the infimum over all estimators and where the constant $c > 0$ depends only on β, L and p_*.

Proof. Let $f_{0n}, \ldots, f_{Mn}$ be the functions defined, for $p = 2$, in the proof of Theorem 2.8 and, for $p = \infty$, in the proof of Theorem 2.10. By construction, $\|f_{jn} - f_{kn}\|_p \geq 2s$, $j \neq k$, with $s = A\psi_{n,p}$ and $A > 0$. Denote by $E_{X_1,\ldots,X_n}$ the expectation with respect to the joint distribution of $X_1, \ldots, X_n$ and put $P_j = P_{f_{jn}}$.

For any estimator T_n, we have the following sequence of inequalities:

$$
\begin{aligned}
&\sup_{f\in\Sigma(\beta,L)} \mathbf{E}_f\left[\psi_{n,p}^{-2}\|T_n-f\|_p^2\right] \\
&\quad\ge A^2 \max_{f\in\{f_{0n},\dots,f_{Mn}\}} P_f\Big(\|T_n-f\|_p \ge A\psi_{n,p}\Big) \\
&\quad\ge A^2\frac{1}{M+1}\sum_{j=0}^{M} E_{X_1,\dots,X_n}\Big[P_j\Big(\|T_n-f\|_p \ge s|X_1,\dots,X_n\Big)\Big] \\
&\quad= A^2 E_{X_1,\dots,X_n}\left[\frac{1}{M+1}\sum_{j=0}^{M} P_j\Big(\|T_n-f\|_p \ge s|X_1,\dots,X_n\Big)\right] \\
&\quad\ge A^2 E_{X_1,\dots,X_n}\left[\inf_{\psi}\frac{1}{M+1}\sum_{j=0}^{M} P_j\Big(\psi\ne j|X_1,\dots,X_n\Big)\right]
\end{aligned}
$$

where the last inequality follows from (2.8).

Fix $X_1,\dots,X_n$. The proofs of Theorems 2.8 and 2.10 imply that

$$
\frac{1}{M+1}\sum_{j=1}^{M} K(P_j,P_0) \le \alpha\log M
$$

with $0<\alpha<1/8$. Then, by (2.72), we have

$$
\begin{aligned}
\overline{p}_{e,M} &= \inf_{\psi}\frac{1}{M+1}\sum_{j=0}^{M} P_j\Big(\psi\ne j|X_1,\dots,X_n\Big) \\
&\ge \frac{\log(M+1)-\log 2}{\log M}-\alpha.
\end{aligned}
$$

Since the right hand side of the last inequality is independent of $X_1,\dots,X_n$, we obtain the required result. ∎

In view of the remarks preceding Theorem 2.9, the result of Theorem 2.11 remains valid for $p=2$ if we replace $\Sigma(\beta,L)$ by the Sobolev class $W(\beta,L)$ or by $W^{per}(\beta,L)$.

2.7.2 Assouad's lemma

The construction known as *Assouad's lemma* deals with a particular case where the hypotheses constitute a cube, i.e., $\{P_0,P_1,\dots,P_M\}=\{P_\omega,\omega\in\Omega\}$ with $\Omega=\{0,1\}^m$ for some integer m. Assouad's lemma reduces the problem of obtaining a lower bound on the minimax risk to m problems of testing two hypotheses, in contrast to the methods presented above where the reduction has been made to *one* problem of testing $M+1$ hypotheses.

Lemma 2.12 (Assouad's lemma). *Let $\Omega = \{0,1\}^m$ be the set of all binary sequences of length m. Let $\{P_\omega, \omega \in \Omega\}$ be a set of 2^m probability measures on $(\mathcal{X}, \mathcal{A})$ and let the corresponding expectations be denoted by E_ω. Then*

$$\inf_{\hat\omega} \max_{\omega\in\Omega} E_\omega \rho(\hat\omega, \omega) \ge \frac{m}{2} \min_{\omega,\omega':\rho(\omega,\omega')=1} \inf_{\psi} \Big(P_\omega(\psi \ne 0) + P_{\omega'}(\psi \ne 1) \Big) \quad (2.76)$$

where $\rho(\omega, \omega')$ is the Hamming distance between ω and ω', $\inf_{\hat\omega}$ denotes the infimum over all estimators $\hat\omega$ taking values in Ω and where $\inf_\psi$ denotes the infimum over all tests ψ taking values in $\{0,1\}$.

Proof. Define

$$\hat\omega = (\hat\omega_1, \dots, \hat\omega_m), \qquad \omega = (\omega_1, \dots, \omega_m),$$

where $\hat\omega_j, \omega_j \in \{0,1\}$. Then

$$\begin{aligned} \max_{\omega\in\Omega} E_\omega \rho(\hat\omega, \omega) &\ge \frac{1}{2^m} \sum_{\omega\in\Omega} E_\omega \rho(\hat\omega, \omega) = \frac{1}{2^m} \sum_{\omega\in\Omega} E_\omega \sum_{j=1}^m |\hat\omega_j - \omega_j| \\ &= \frac{1}{2^m} \sum_{j=1}^m \left(\sum_{\omega\in\Omega:\omega_j=1} + \sum_{\omega\in\Omega:\omega_j=0} \right) E_\omega |\hat\omega_j - \omega_j|. \quad (2.77) \end{aligned}$$

All the terms in the last sum over j in (2.77) are bounded from below in a similar way. Consider, for example, the mth term:

$$\begin{aligned} &\left(\sum_{\omega\in\Omega:\omega_m=1} + \sum_{\omega\in\Omega:\omega_m=0} \right) E_\omega |\hat\omega_m - \omega_m| \quad (2.78) \\ &= \sum_{(\omega_1,\dots,\omega_{m-1})\in\{0,1\}^{m-1}} \left(E_{(\omega_1,\dots,\omega_{m-1},1)} |\hat\omega_m - 1| + E_{(\omega_1,\dots,\omega_{m-1},0)} |\hat\omega_m| \right). \end{aligned}$$

Here

$$\begin{aligned} &E_{(\omega_1,\dots,\omega_{m-1},1)} |\hat\omega_m - 1| + E_{(\omega_1,\dots,\omega_{m-1},0)} |\hat\omega_m| \quad (2.79) \\ &= P_{(\omega_1,\dots,\omega_{m-1},1)}(\hat\omega_m = 0) + P_{(\omega_1,\dots,\omega_{m-1},0)}(\hat\omega_m = 1) \\ &\ge \inf_\psi \Big(P_{(\omega_1,\dots,\omega_{m-1},1)}(\psi = 0) + P_{(\omega_1,\dots,\omega_{m-1},0)}(\psi = 1) \Big) \\ &\ge \min_{\omega,\omega':\rho(\omega,\omega')=1} \inf_\psi \Big(P_{\omega'}(\psi \ne 1) + P_\omega(\psi \ne 0) \Big). \end{aligned}$$

Carrying out evaluations similar to (2.78)–(2.79) for all j we obtain

$$\left(\sum_{\omega\in\Omega:\omega_j=1}+\sum_{\omega\in\Omega:\omega_j=0}\right)E_\omega|\hat{\omega}_j-\omega_j| \tag{2.80}$$
$$\geq 2^{m-1}\min_{\omega,\omega':\rho(\omega,\omega')=1}\inf_{\psi}\Big(P_{\omega'}(\psi\neq 1)+P_\omega(\psi\neq 0)\Big).$$

We complete the proof by combining (2.77) and (2.80). ∎

Lemma 2.12 is an intermediate result that will be developed further before being used. The following two steps should still be accomplished:

(i) an explicit lower bound for the minimum on the right hand side of (2.76) should be given;
(ii) the initial minimax risk should be reduced to the form

$$\inf_{\hat{\omega}}\max_{\omega\in\Omega}E_\omega\rho(\hat{\omega},\omega).$$

The following theorem carries out the first task. The second one will be explained by an example below (cf. Example 2.2).

Theorem 2.12 *Let $\Omega=\{0,1\}^m$ be the set of binary sequences of length m. Let $\{P_\omega,\omega\in\Omega\}$ be a set of 2^m probability measures on $(\mathcal{X},\mathcal{A})$ and let E_ω denote the corresponding expectations.*

(i) If there exist $\tau>0$ and $0<\alpha<1$ such that

$$P_\omega\left(\frac{dP^a_{\omega'}}{dP_\omega}\geq\tau\right)\geq 1-\alpha,\quad\forall\,\omega,\omega'\in\Omega:\rho(\omega,\omega')=1,$$

where $P^a_{\omega'}$ is the absolutely continuous component of $P_{\omega'}$ with respect to P_ω, then

$$\inf_{\hat{\omega}}\max_{\omega\in\Omega}E_\omega\rho(\hat{\omega},\omega)\geq\frac{m}{2}(1-\alpha)\min(\tau,1) \tag{2.81}$$

(likelihood ratio version).

(ii) If $V(P_{\omega'},P_\omega)\leq\alpha<1,\quad\forall\,\omega,\omega'\in\Omega:\rho(\omega,\omega')=1$, then

$$\inf_{\hat{\omega}}\max_{\omega\in\Omega}E_\omega\rho(\hat{\omega},\omega)\geq\frac{m}{2}(1-\alpha) \tag{2.82}$$

(total variation version).

(iii) If $H^2(P_{\omega'},P_\omega)\leq\alpha<2,\quad\forall\,\omega,\omega'\in\Omega:\rho(\omega,\omega')=1$, then

$$\inf_{\hat{\omega}}\max_{\omega\in\Omega}E_\omega\rho(\hat{\omega},\omega)\geq\frac{m}{2}\Big(1-\sqrt{\alpha(1-\alpha/4)}\Big) \tag{2.83}$$

(Hellinger version).

(iv) If $K(P_{\omega'}, P_\omega) \le \alpha < \infty$ *or* $\chi^2(P_{\omega'}, P_\omega) \le \alpha < \infty$, $\quad \forall\ \omega, \omega' \in \Omega$: $\rho(\omega, \omega') = 1$, *then*

$$\inf_{\hat\omega} \max_{\omega \in \Omega} E_\omega \rho(\hat\omega, \omega) \ge \frac{m}{2} \max\left(\frac{1}{2}\exp(-\alpha), \left(1 - \sqrt{\alpha/2}\right)\right) \tag{2.84}$$

(Kullback/χ^2 version).

Proof. In order to prove (ii)–(iv), it is sufficient to observe that

$$\inf_{\psi} \Big(P_\omega(\psi \ne 0) + P_{\omega'}(\psi \ne 1)\Big) = \int \min(dP_\omega, dP_{\omega'})$$

in (2.76) and to apply the same argument as in the proof of Theorem 2.2.

We now prove (i). In the same way as in the proof of Proposition 2.1 we obtain

$$\inf_{\psi} \Big(P_\omega(\psi \ne 0) + P_{\omega'}(\psi \ne 1)\Big) \ge \min_{0 \le p \le 1} (\max\{0, \tau(p - \alpha)\} + 1 - p).$$

If $\tau > 1$, the minimum on the right hand side is attained at $p = \alpha$, while for $\tau \le 1$ it is attained at $p = 1$. Inequality (2.81) follows from this remark and from Lemma 2.12. ∎

Example 2.2 *A lower bound on the minimax risk in* L_2 *via Assouad's lemma.*

Consider the nonparametric regression model under Assumptions (B) and Assumption (LP2) of Chapter 1. We will use the notation introduced in Section 2.6.1. In particular, $\omega = (\omega_1, \dots, \omega_m) \in \Omega = \{0,1\}^m$ and $f_\omega(x) = \sum_{k=1}^m \omega_k \varphi_k(x)$. The L_2-risk of an estimator T_n is given by

$$E_\omega\left[\|T_n - f_\omega\|_2^2\right] = E_\omega \int_0^1 |T_n(x) - f_\omega(x)|^2 dx = \sum_{k=1}^m E_\omega d_k^2(T_n, \omega_k),$$

where

$$d_k(T_n, \omega_k) = \left(\int_{\Delta_k} |T_n(x) - \omega_k \varphi_k(x)|^2 dx\right)^{1/2}$$

and where the intervals Δ_k are as in (2.58). Define the statistic

$$\hat\omega_k = \arg\min_{t=0,1} d_k(T_n, t).$$

Then

$$d_k(T_n, \omega_k) \ge \frac{1}{2} d_k(\hat\omega_k, \omega_k) \triangleq \frac{1}{2} |\hat\omega_k - \omega_k| \|\varphi_k\|_2. \tag{2.85}$$

Indeed, by the definition of $\hat\omega_k$, we have $d_k(T_n, \hat\omega_k) \le d_k(T_n, \omega_k)$ and therefore

$$d_k(\hat{\omega}_k, \omega_k) = \left(\int_{\Delta_k} |(\hat{\omega}_k - \omega_k)\varphi_k(x)|^2 dx \right)^{1/2}$$
$$\leq d_k(T_n, \hat{\omega}_k) + d_k(T_n, \omega_k) \leq 2d_k(T_n, \omega_k).$$

By (2.85), we obtain for all $\omega \in \Omega$

$$E_\omega \left[\|T_n - f_\omega\|_2^2 \right] \geq \frac{1}{4} \sum_{k=1}^{m} E_\omega \left[(\hat{\omega}_k - \omega_k)^2 \right] \|\varphi_k\|_2^2$$
$$= \frac{1}{4} L^2 h_n^{2\beta+1} \|K\|_2^2 E_\omega \rho(\hat{\omega}, \omega)$$

where $\hat{\omega} = (\hat{\omega}_1, \ldots, \hat{\omega}_m)$. Since $h_n = 1/m$, we conclude that, for any estimator T_n,

$$\max_{\omega \in \Omega} E_\omega \left[\|T_n - f_\omega\|_2^2 \right] \geq \frac{1}{4} L^2 h_n^{2\beta+1} \|K\|_2^2 \inf_{\hat{\omega}} \max_{\omega \in \Omega} E_\omega \rho(\hat{\omega}, \omega).$$

A bound for the last expression is obtained using part (iv) of Theorem 2.12 where the condition on the Kullback divergence is checked in the same way as in (2.36). Observe that in this proof, in contrast to that in Section 2.6.1, we cannot drop Assumption (LP2).

Remarks.

(1) Switching from the initial minimax risk to a risk of the form

$$\inf_{\hat{\omega}} \max_{\omega \in \Omega} E_\omega \rho(\hat{\omega}, \omega)$$

is possible only for some particular loss functions w and semi-distances $d(\cdot, \cdot)$. The application of Assouad's lemma is therefore limited by these constraints. For example, it cannot be used if the initial risk is defined with the indicator loss function $w(u) = I(u \geq A)$ or the L_∞-distance.

(2) An advantage of Assouad's lemma consists in the fact that it admits the Hellinger version and the total variation version adapted to the case of multiple hypotheses ($M \geq 2$). Note that such versions are not available in the framework of Section 2.6. We can apply Assouad's lemma, for example, if the Kullback divergence is not defined or if it is difficult to verify the condition (2.41) on the likelihood ratios.

2.7.3 The van Trees inequality

All the methods that we discussed in this chapter started with bounding from below the maximum risk over a functional class by the maximum (or average) risk over a *finite* family of members of the class. The technique that we are going to consider now is somewhat different. The idea is to bound from below the maximum risk over a functional class by the Bayes risk over a parametric

subfamily indexed by a continuous parameter t, and then to use the van Trees inequality to bound this parametric Bayes risk.

In order to introduce the van Trees inequality we need some notation. Let $T = [t_1, t_2]$ be an interval in $\mathbf{R}$ such that $-\infty < t_1 < t_2 < \infty$. Let $\{\mathsf{P}_t, t \in T\}$ be a family of probability measures on $(\mathcal{X}, \mathcal{A})$. We will be interested in the case $\mathsf{P}_t = P_{\theta_t}$ where the parametric family $\{\theta_t, t \in T\}$ is a subset of our initial class Θ (cf. Section 2.1), though this assumption will not be needed for the proof of the van Trees inequality. The sample space $\mathcal{X}$, the σ-algebra $\mathcal{A}$, and the measure P_t typically depend on the sample size n but we do not indicate it in the notation for the sake of brevity.

Assume that there exists a σ-finite measure ν on $(\mathcal{X}, \mathcal{A})$ such that $\mathsf{P}_t \ll \nu$ for all $t \in T$. Denote by $p(\cdot, t)$ the density of P_t with respect to ν.

Introduce a probability distribution on T with a density $\mu(\cdot)$ with respect to Lebesgue measure. For an arbitrary estimator $\hat{t}(\mathbf{X})$ where $\mathbf{X}$ is distributed according to P_t we consider the Bayes risk with a prior density μ:

$$\mathcal{R}^B(\hat{t}) \stackrel{\triangle}{=} \int_T \mathsf{E}_t\big[(\hat{t}(\mathbf{X}) - t)^2\big]\mu(t)dt = \int\int (\hat{t}(x) - t)^2 p(x,t)\nu(dx)\mu(t)dt \quad (2.86)$$

where E_t denotes expectation with respect to P_t.

Theorem 2.13 (The van Trees inequality). *Assume that:*

(i) The density $p(x,t)$ is measurable in (x,t) and absolutely continuous in t for almost all x with respect to the measure ν.

(ii) The Fisher information

$$\mathcal{I}(t) = \int \left(\frac{p'(x,t)}{p(x,t)}\right)^2 p(x,t)\nu(dx)\,,$$

where $p'(x,t)$ denotes the derivative of $p(x,t)$ in t, is finite and integrable on T:

$$\int_T \mathcal{I}(t)dt < \infty. \quad (2.87)$$

(iii) The prior density μ is absolutely continuous on its support T, satisfies the condition $\mu(t_1) = \mu(t_2) = 0$, and has finite Fisher information

$$\mathcal{J}(\mu) = \int_T \frac{(\mu'(t))^2}{\mu(t)}dt\,.$$

Then, for any estimator $\hat{t}(\mathbf{X})$, the Bayes risk is bounded as follows:

$$\int_T \mathsf{E}_t\big[(\hat{t}(\mathbf{X}) - t)^2\big]\mu(t)dt \geq \frac{1}{\int \mathcal{I}(t)\mu(t)dt + \mathcal{J}(\mu)}\,. \quad (2.88)$$

Proof. It suffices to consider the case $\mathcal{R}^B(\hat{t}) < \infty$ because otherwise the result is trivial. Since $p(x,t)$ and $\mu(t)$ are absolutely continuous and $\mu(t_1) = \mu(t_2) = 0$, we have

$$\int (p(x,t)\mu(t))'dt = 0$$

for almost all x with respect to ν. Here $(p(x,t)\mu(t))'$ is the derivative of $p(x,t)\mu(t)$ with respect to t. For the same reasons, after integration by parts we get

$$\int t(p(x,t)\mu(t))'dt = -\int p(x,t)\mu(t)dt\,.$$

The last two equalities imply

$$\int\int (\hat{t}(x)-t)(p(x,t)\mu(t))'dt\,\nu(dx) = \int\int p(x,t)\mu(t)dt\,\nu(dx) = 1. \quad (2.89)$$

Let us show that the first integral in (2.89) can be considered as an integral over $B = \{(x,t) : p(x,t)\mu(t) \neq 0\}$. Fix x such that $t \mapsto p(x,t)$ is absolutely continuous and consider the function $f(\cdot) = p(x,\cdot)\mu(\cdot)$ on T. Note that there exists a set N_x of Lebesgue measure 0 such that

$$S \stackrel{\triangle}{=} \{t \in T :\ f(t) = 0\} \subseteq \{t \in T :\ f'(t) = 0\} \cup N_x.^1 \quad (2.90)$$

Now, (2.90) implies that inserting the indicator $I(B)$ under the integral over t on the right hand side of (2.89) does not change the value of this integral for almost all x with respect to ν. Thus,

$$\int\int (\hat{t}(x)-t)(p(x,t)\mu(t))'I(B)dt\,\nu(dx) = 1.$$

Applying the Cauchy–Schwarz inequality to the left hand side of this equation and using (2.86) we find

$$\int_T \mathsf{E}_t\big[(\hat{t}(\mathbf{X})-t)^2\big]\mu(t)dt \int\int \frac{\big((p(x,t)\mu(t))'\big)^2}{p(x,t)\mu(t)}\, I(B)dt\,\nu(dx) \geq 1 \quad (2.91)$$

Now,

$$\begin{aligned}
&\int\int \frac{\big((p(x,t)\mu(t))'\big)^2}{p(x,t)\mu(t)}\, I(B)dt\,\nu(dx)\\
&= \int\int \left(\frac{(p(x,t)\mu(t))'}{p(x,t)\mu(t)}\right)^2 p(x,t)\mu(t)dt\,\nu(dx)\\
&= \int \mathcal{I}(t)\mu(t)dt + \mathcal{J}(\mu) + 2\int\int p'(x,t)\mu'(t)dt\,\nu(dx). \qquad (2.92)
\end{aligned}$$

[1] In fact, since f is absolutely continuous, the set S is closed and the derivative f' exists almost everywhere on S. The set of isolated points of S is at most countable and thus has Lebesgue measure 0. Take any $t_0 \in S$ which is not an isolated point of S and such that $f'(t_0)$ exists. Take a sequence $\{t_k\}_{k\geq 1} \subseteq S$ such that $t_k \to t_0$. Then

$$f'(t_0) = \lim_{k\to\infty} \frac{f(t_k) - f(t_0)}{t_k - t_0} = 0,$$

proving (2.90).

Here the integral $\int \mathcal{I}(t)\mu(t)dt$ is finite since μ is bounded and (2.87) holds. Taking into account that $\int \mathcal{I}(t)\mu(t)dt < \infty$, $\mathcal{J}(\mu) < \infty$, and using the Cauchy–Schwarz inequality we easily obtain

$$\int\int |p'(x,t)\mu'(t)|dt\,\nu(dx) < \infty.$$

In view of (2.91), to complete the proof of the theorem it suffices to show that the last double integral in (2.92) vanishes. Write

$$\int\int p'(x,t)\mu'(t)dt\,\nu(dx) = \int g(t)\mu'(t)dt,$$

where $g(t) = \int p'(x,t)\nu(dx)$. Let us show that $g(t) = 0$ for almost all $t \in T$. In fact, by the Cauchy–Schwarz inequality and (2.87),

$$\int_T \int |p'(x,t)|\nu(dx)\,dt \le \int_T \sqrt{\mathcal{I}(t)}dt \le \left(\int_T \mathcal{I}(t)dt\right)^{1/2} \sqrt{t_2 - t_1} < \infty.$$

Therefore, we can apply the Fubini theorem, which yields

$$\begin{aligned}\int_a^b g(t)dt &= \int \left(\int_a^b p'(x,t)dt\right)\nu(dx)\\ &= \int (p(x,b) - p(x,a))\nu(dx) = 0, \quad \forall\ t_1 \le a < b \le t_2,\end{aligned}$$

because $p(\cdot,t)$ is a probability density with respect to ν for any $t \in T$. Since a and b are arbitrary, we obtain that $g(t) = 0$ for almost all $t \in T$. Therefore, the last double integral in (2.92) vanishes and (2.88) follows. ∎

The following choice of the prior density μ is often convenient:

$$\mu(t) = \frac{1}{s}\mu_0\Big(\frac{t - t_0}{s}\Big) \tag{2.93}$$

where t_0 is the center of the interval T, $s = (t_2 - t_1)/2$, and

$$\mu_0(t) = \cos^2(\pi t/2)I(|t| \le 1), \tag{2.94}$$

so that $\mathcal{J}(\mu_0) = \pi^2$. Clearly, the density (2.93) satisfies assumption (iii) of Theorem 2.13. Moreover, one can show that it has the smallest Fisher information $\mathcal{J}(\mu)$ among all the densities μ supported on T and satisfying this assumption.

Example 2.3 *A lower bound on the minimax risk at a fixed point via the van Trees inequality.*

Consider the nonparametric regression model under Assumption (B), and Assumption (LP2) introduced in Chapter 1. Assume in addition that the random variables ξ_i are normal with mean 0 and variance σ^2. Our aim is to obtain a lower bound for the minimax risk on (Θ, d) where Θ is a Hölder class:

$$\Theta = \Sigma(\beta, L),\ \beta > 0, L > 0,$$

and where d is the distance at a fixed point $x_0 \in [0,1]$:

$$d(f,g) = |f(x_0) - g(x_0)|.$$

Choose the interval $T = [-1,1]$ and define the following parametric family of functions on $[0,1]$ indexed by $t \in [-1,1]$:

$$f_t(x) = tLh_n^\beta K\left(\frac{x - x_0}{h_n}\right), \qquad x \in [0,1],$$

where $h_n = c_0 n^{-\frac{1}{2\beta+1}}$ with $c_0 > 0$ and K satisfies (2.33). Arguing as in Section 2.5 we easily find that $f_t \in \Sigma(\beta, L)$ for all $t \in [-1,1]$. Therefore, choosing, for example, the prior density $\mu = \mu_0$ as defined in (2.94) we obtain that, for any estimator T_n,

$$\begin{aligned}
\sup_{f\in\Sigma(\beta,L)} E_f\left[(T_n(x_0) - f(x_0))^2\right] &\geq \sup_{t\in[-1,1]} E_{f_t}\left[(T_n(x_0) - f_t(x_0))^2\right] \\
&\geq \int_{-1}^{1} E_{f_t}\left[(T_n(x_0) - f_t(x_0))^2\right]\mu_0(t)dt \\
&= (Lh_n^\beta K(0))^2 \int_{-1}^{1} E_{f_t}\left[(\hat{t}_n - t)^2\right]\mu_0(t)dt \\
&= n^{-\frac{2\beta}{2\beta+1}}(Lc_0^\beta K(0))^2 \int_{-1}^{1} E_{f_t}\left[(\hat{t}_n - t)^2\right]\mu_0(t)dt \qquad (2.95)
\end{aligned}$$

where $\hat{t}_n = (Lh_n^\beta K(0))^{-1} T_n(x_0)$, and we used that $f_t(x_0) = tLh_n^\beta K(0)$. Observe that to prove the desired lower bound (cf. (2.38)) it suffices to show that the last integral, i.e., the Bayes risk for the chosen parametric subfamily of $\Sigma(\beta, L)$, is bounded from below by a constant independent of n. This is proved using the van Trees inequality. Indeed, the Fisher information for the parametric regression model

$$Y_i = tLh_n^\beta K\left(\frac{X_i - x_0}{h_n}\right) + \xi_i, \quad i = 1, \dots, n,$$

is independent of the parameter t and has the form

$$\mathcal{I}(t) \equiv \sigma^2 \sum_{i=1}^{n} \left(Lh_n^\beta K\left(\frac{X_i - x_0}{h_n}\right)\right)^2, \qquad t \in [-1,1]. \qquad (2.96)$$

Arguing as in (2.36) we get that, under Assumption LP2,

$$\mathcal{I}(t) \le \sigma^2 a_0 L^2 K_{\max}^2 n h_n^{2\beta+1} = \sigma^2 a_0 L^2 K_{\max}^2 c_0^{2\beta+1}.$$

Therefore, using the van Trees inequality (2.88) and the fact that $\mathcal{J}(\mu_0) = \pi^2$, we obtain

$$\int_{-1}^{1} E_{f_t}\big[(\hat{t}_n - t)^2\big]\mu_0(t)dt \ge \frac{1}{\sigma^2 a_0 L^2 K_{\max}^2 c_0^{2\beta+1} + \pi^2}\,.$$

The expression on the right hand side of this inequality does not depend on n. Hence, combining it with (2.95), we obtain the desired lower bound

$$\inf_{T_n} \sup_{f\in\Sigma(\beta,L)} E_f\big[(T_n(x_0) - f(x_0))^2\big] \ge c\, n^{-\frac{2\beta}{2\beta+1}}$$

where $c > 0$ is a constant.

Note that the result that we obtain in Example 2.3 does not improve upon Theorem 2.3. In this example we consider only Gaussian noise. The argument can be extended to any noise with finite Fisher information. However, Theorem 2.3 holds under a slightly less restrictive assumption (part (ii) of Assumption (B)). Another limitation is that the van Trees inequality applies only to the squared loss function. An advantage of the van Trees technique seems to be its relative simplicity and the fact that it can lead in some cases to asymptotically optimal constants in the lower bounds.

2.7.4 The method of two fuzzy hypotheses

We consider now a generalization of the technique of two hypotheses (cf. Theorems 2.1 and 2.2). The results of this section can be used to obtain lower bounds on the minimax risk in the problem of estimation of functionals and in nonparametric testing. Though these problems remain beyond the scope of the book, the corresponding lower bounds can readily be established in the same spirit as above, and we discuss them here for completeness.

Let $F(\theta)$ be a functional defined on a measurable space $(\Theta,\mathcal{U})$ and taking values in $(\mathbf{R},\mathcal{B}(\mathbf{R}))$ where $\mathcal{B}(\mathbf{R})$ is the Borel σ-algebra on $\mathbf{R}$. We would like to estimate $F(\theta)$ from observations $\mathbf{X}$ associated with a statistical model $\{P_\theta, \theta \in \Theta\}$ where the probability measures P_θ are defined on $(\mathcal{X},\mathcal{A})$. Typically, $\mathbf{X}, P_\theta, \mathcal{X}$, and $\mathcal{A}$ depend on the sample size n, though we do not reflect this fact in our notation for the sake of brevity. Let $\hat{F} = \hat{F}_n$ be an estimator of $F(\theta)$. For a loss function w and a rate ψ_n, define the maximum risk of $\hat{F}_n$ as follows:

$$\sup_{\theta\in\Theta} \mathbf{E}_\theta \left[w(\psi_n^{-1}|\hat{F}_n - F(\theta)|)\right]. \tag{2.97}$$

Our aim here is to give a nontrivial lower bound on risk (2.97) for all estimators $\hat{F}_n$. First, by Markov's inequality, we obtain

$$\inf_{\hat{F}_n} \sup_{\theta\in\Theta} \mathbf{E}_\theta \left[w(\psi_n^{-1}|\hat{F}_n - F(\theta)|) \right] \geq w(A) \inf_{\hat{F}_n} \sup_{\theta\in\Theta} P_\theta(|\hat{F}_n - F(\theta)| \geq A\psi_n)$$

for all $A > 0$. In words, we switch to the minimax probabilities, as we did in the general scheme of Section 2.2. However, the next step is different. Instead of passing to a finite number of simple hypotheses, we introduce two probability measures μ_0 and μ_1 on $(\Theta, \mathcal{U})$ and apply the bound

$$\sup_{\theta\in\Theta} P_\theta(|\hat{F} - F(\theta)| \geq s) \geq \max \left\{ \int P_\theta(|\hat{F} - F(\theta)| \geq s)\mu_0(d\theta), \right.$$
$$\left. \int P_\theta(|\hat{F} - F(\theta)| \geq s)\mu_1(d\theta) \right\} \quad (2.98)$$

where $s > 0$ and where we write for brevity $\hat{F}$ instead of $\hat{F}_n$. The measures μ_0 and μ_1 will be called *fuzzy hypotheses*, since their masses can be spread all over the set Θ. If μ_0 and μ_1 are Dirac measures, we are back to the case of two simple hypotheses analyzed in Sections 2.3 and 2.4.2.

Define two "posterior" probability measures $I\!P_0$ and $I\!P_1$ on $(\mathcal{X}, \mathcal{A})$ as follows:

$$I\!P_j(S) = \int P_\theta(S)\mu_j(d\theta), \quad \forall\ S \in \mathcal{A}, \quad j = 0, 1.$$

Theorem 2.14 *Assume that:*

(i) There exist $c \in \mathbf{R}$, $s > 0$, $0 \leq \beta_0, \beta_1 < 1$ *such that*

$$\mu_0(\theta : F(\theta) \leq c) \geq 1 - \beta_0,$$
$$\mu_1(\theta : F(\theta) \geq c + 2s) \geq 1 - \beta_1.$$

(ii) There exist $\tau > 0$ *and* $0 < \alpha < 1$ *such that*

$$I\!P_1 \left(\frac{dI\!P_0^a}{dI\!P_1} \geq \tau \right) \geq 1 - \alpha,$$

where $I\!P_0^a$ *is the absolutely continuous component of* $I\!P_0$ *with respect to* $I\!P_1$.

Then, for any estimator $\hat{F}$,

$$\sup_{\theta\in\Theta} P_\theta(|\hat{F} - F(\theta)| \geq s) \geq \frac{\tau(1 - \alpha - \beta_1) - \beta_0}{1 + \tau}\ .$$

Proof. Observe that

$$\int P_\theta(|\hat{F} - F(\theta)| \geq s)\mu_0(d\theta) \quad (2.99)$$
$$\geq \int I(\hat{F} \geq c + s,\ F(\theta) \leq c) dP_\theta \mu_0(d\theta)$$

$$
\begin{aligned}
&\geq \int I(\hat{F} \geq c+s) dP_\theta \mu_0(d\theta) \\
&\quad - \int I(F(\theta) > c) dP_\theta \mu_0(d\theta) \\
&= I\!P_0(\hat{F} \geq c+s) - \mu_0(\theta : F(\theta) > c) \\
&\geq I\!P_0(\hat{F} \geq c+s) - \beta_0.
\end{aligned}
$$

In a similar way,

$$
\begin{aligned}
&\int P_\theta(|\hat{F} - F(\theta)| \geq s) \mu_1(d\theta) \qquad (2.100) \\
&\qquad \geq \int I(\hat{F} < c+s, F(\theta) \geq c+2s) dP_\theta \mu_1(d\theta) \\
&\qquad \geq I\!P_1(\hat{F} < c+s) - \beta_1.
\end{aligned}
$$

By (2.98)–(2.101), we obtain

$$
\begin{aligned}
&\sup_{\theta \in \Theta} P_\theta(|\hat{F} - F(\theta)| \geq s) \qquad (2.101) \\
&\qquad \geq \max\Big\{ I\!P_0(\hat{F} \geq c+s) - \beta_0, I\!P_1(\hat{F} < c+s) - \beta_1 \Big\} \\
&\qquad \geq \inf_\psi \max\Big\{ I\!P_0(\psi = 1) - \beta_0, I\!P_1(\psi = 0) - \beta_1 \Big\},
\end{aligned}
$$

where $\inf_\psi$ denotes the infimum over all tests ψ taking values in $\{0,1\}$. By assumption (ii), we obtain, as in the proof of Proposition 2.1,

$$
I\!P_0(\psi = 1) \geq \int \frac{dI\!P_0^a}{dI\!P_1} I(\psi = 1) dI\!P_1 \geq \tau(I\!P_1(\psi = 1) - \alpha).
$$

It follows that

$$
\begin{aligned}
&\sup_{\theta \in \Theta} P_\theta(|\hat{F} - F(\theta)| \geq s) \\
&\qquad \geq \inf_\psi \max\Big\{ \tau(I\!P_1(\psi = 1) - \alpha) - \beta_0, 1 - I\!P_1(\psi = 1) - \beta_1 \Big\} \\
&\qquad \geq \min_{0 \leq p \leq 1} \max\Big\{ \tau(p - \alpha) - \beta_0, 1 - p - \beta_1 \Big\} \\
&\qquad = \frac{\tau(1 - \alpha - \beta_1) - \beta_0}{1 + \tau}.
\end{aligned}
$$

∎

Note that if $\beta_0 = \beta_1 = 0$, the measures μ_0 and μ_1 have disjoint supports. Theorem 2.14 gives a lower bound under the condition (ii) which deals directly with the distribution of the likelihood ratio. Other versions, similar to those of Theorem 2.2, are now immediately obtained as corollaries.

Theorem 2.15 *Suppose that assumption (i) of Theorem 2.14 holds.*

(i) If $V(I\!P_1, I\!P_0) \le \alpha < 1$, *then*

$$\inf_{\hat{F}} \sup_{\theta\in\Theta} P_\theta(|\hat{F} - F(\theta)| \ge s) \ge \frac{1 - \alpha - \beta_0 - \beta_1}{2} \tag{2.102}$$

(total variation version).

(ii) If $H^2(I\!P_1, I\!P_0) \le \alpha < 2$, *then*

$$\inf_{\hat{F}} \sup_{\theta\in\Theta} P_\theta(|\hat{F} - F(\theta)| \ge s) \ge \frac{1 - \sqrt{\alpha(1-\alpha/4)}}{2} - \frac{\beta_0 + \beta_1}{2} \tag{2.103}$$

(Hellinger version).

(iii) If $K(I\!P_1, I\!P_0) \le \alpha < \infty$ *(or* $\chi^2(I\!P_1, I\!P_0) \le \alpha < \infty$*), then*

$$\begin{aligned}\inf_{\hat{F}} \sup_{\theta\in\Theta} & P_\theta(|\hat{F} - F(\theta)| \ge s) \\ & \ge \max\left(\frac{1}{4}\exp(-\alpha), \frac{1 - \sqrt{\alpha/2}}{2}\right) - \frac{\beta_0 + \beta_1}{2}\end{aligned} \tag{2.104}$$

(Kullback/χ^2 version).

Proof. By (2.101), we have

$$\begin{aligned}\sup_{\theta\in\Theta} P_\theta(|\hat{F} - F(\theta)| \ge s) &\ge \inf_{\psi} \frac{I\!P_0(\psi = 1) + I\!P_1(\psi = 0)}{2} - \frac{\beta_0 + \beta_1}{2} \\ &= \frac{1}{2}\int \min(dI\!P_0, dI\!P_1) - \frac{\beta_0 + \beta_1}{2}\,.\end{aligned}$$

The proof is completed as in Theorem 2.2. ∎

2.7.5 Lower bounds for estimators of a quadratic functional

We now apply the method of two fuzzy hypotheses to prove lower bounds for estimators of a quadratic functional. Consider the nonparametric regression model under Assumption (B) and Assumption (LP2). Suppose that the random variables ξ_i are i.i.d. with distribution $\mathcal{N}(0,1)$. Put $\theta = f(\cdot)$ and

$$F(\theta) = \int_0^1 f^2(x)dx.$$

Suppose also that the class of functions f we are dealing with is the Hölder class, $\Theta = \Sigma(\beta, L)$, $\beta > 0, L > 0$. To obtain a lower bound on the minimax risk in estimation of $F(\theta)$, we apply part (iii) (χ^2 version) of Theorem 2.15.

Let μ_0 be the Dirac measure concentrated on the function $f \equiv 0$ and let μ_1 be a discrete measure supported on a finite set of functions:

$$f_\omega(x) = \sum_{k=1}^{m} \omega_k \varphi_k(x) \quad \text{with} \ \ \omega_k \in \{-1, 1\},$$

where $\varphi_k(\cdot)$ are defined in (2.56) with

$$h_n = 1/m, \quad m = \lceil c_0 n^{\frac{2}{4\beta+1}} \rceil, \quad c_0 > 0.$$

Suppose that the random variables $\omega_1, \dots, \omega_m$ are i.i.d. with $\mu_1(\omega_j = 1) = \mu_1(\omega_j = -1) = 1/2$. It is easy to see that $f_\omega \in \Sigma(\beta, L)$ for all $\omega_j \in \{-1, 1\}$. Moreover, by the same argument as in (2.57) we obtain

$$\int_0^1 f_\omega^2(x)dx = \sum_{k=1}^{m} \int \varphi_k^2(x)dx = mL^2 h_n^{2\beta+1} \|K\|_2^2 = L^2 h_n^{2\beta} \|K\|_2^2.$$

Therefore assumption (i) of Theorem 2.14 holds with

$$c = 0, \quad \beta_0 = \beta_1 = 0, \quad s = L^2 h_n^{2\beta} \|K\|_2^2/2 \geq A n^{-\frac{4\beta}{4\beta+1}},$$

where $A > 0$ is a constant. Posterior measures $I\!P_0$ and $I\!P_1$ admit the following densities with respect to the Lebesgue measure on $\mathbf{R}^n$:

$$p_0(u_1, \dots, u_n) = \prod_{i=1}^{n} \varphi(u_i) = \prod_{k=1}^{m} \prod_{i: X_i \in \Delta_k} \varphi(u_i),$$

$$p_1(u_1, \dots, u_n) = \prod_{k=1}^{m} \frac{1}{2} \left(\prod_{i: X_i \in \Delta_k} \varphi(u_i - \varphi_k(X_i)) + \prod_{i: X_i \in \Delta_k} \varphi(u_i + \varphi_k(X_i)) \right),$$

respectively, where $\varphi(\cdot)$ is the density of $\mathcal{N}(0, 1)$. Recall that X_i are deterministic and the measures $I\!P_0$ and $I\!P_1$ are associated with the distribution of $(Y_1, \dots, Y_n)$. Setting for brevity

$$\prod_{i \in (k)} = \prod_{i: X_i \in \Delta_k}, \quad S_k = \sum_{i: X_i \in \Delta_k} \varphi_k^2(X_i), \quad V_k(u) = \sum_{i: X_i \in \Delta_k} u_i \varphi_k(X_i),$$

we can write

$$\begin{aligned} \frac{dI\!P_1}{dI\!P_0}(u_1, \dots, u_n) &= \prod_{k=1}^{m} \left\{ \frac{\prod_{i \in (k)} \varphi(u_i - \varphi_k(X_i)) + \prod_{i \in (k)} \varphi(u_i + \varphi_k(X_i))}{2 \prod_{i \in (k)} \varphi(u_i)} \right\} \\ &= \prod_{k=1}^{m} \left\{ \frac{1}{2} \exp\left(-\frac{S_k}{2}\right) \left[\exp\left(V_k(u)\right) + \exp\left(-V_k(u)\right) \right] \right\}. \end{aligned}$$

Then the χ^2 divergence between $I\!P_1$ and $I\!P_0$ is as follows:

$$\chi^2(I\!P_1, I\!P_0) = \int \left(\frac{dI\!P_1}{dI\!P_0}\right)^2 dI\!P_0 - 1 \tag{2.105}$$

where

$$\int \left(\frac{dI\!P_1}{dI\!P_0}\right)^2 dI\!P_0 = \prod_{k=1}^{m} \Big\{ \frac{1}{4} \exp\left(-S_k\right) \times \int [\exp(V_k(u)) + \exp(-V_k(u))]^2 \prod_{i \in (k)} \varphi(u_i) du_i \Big\}.$$

Since $\int \exp(vt)\varphi(v)dv = \exp(t^2/2)$ for all $t \in \mathbf{R}$, we obtain

$$\begin{aligned} \int \exp\left(2V_k(u)\right) \prod_{i \in (k)} \varphi(u_i) du_i &= \int \exp\left(-2V_k(u)\right) \prod_{i \in (k)} \varphi(u_i) du_i \\ &= \exp(2S_k). \end{aligned}$$

Therefore

$$\int \left(\frac{dI\!P_1}{dI\!P_0}\right)^2 dI\!P_0 = \prod_{k=1}^{m} \frac{\exp(S_k) + \exp(-S_k)}{2} \ . \tag{2.106}$$

Using Assumption (LP2) and following the lines of (2.36) we obtain

$$\begin{aligned} S_k &= \sum_{i: X_i \in \Delta_k} \varphi_k^2(X_i) \qquad (2.107) \\ &\le L^2 K_{\max}^2 h_n^{2\beta} \sum_{i=1}^{n} I\left(\left|\frac{X_i - x_k}{h_n}\right| \le 1/2\right) \\ &\le a_0 L^2 K_{\max}^2 n h_n^{2\beta+1}, \end{aligned}$$

if $nh_n \ge 1$, where a_0 is the constant appearing in Assumption (LP2). Since $h_n \asymp n^{-\frac{2}{4\beta+1}}$, there exists a constant $c_1 < \infty$ such that $|S_k| \le c_1$ for all $n \ge 1$ and all $k = 1, \ldots, m$. Thus, for $|x| \le c_1$ we have $|e^x - 1 - x| \le c_2 x^2$ where c_2 is a finite constant. Therefore

$$\frac{\exp(S_k) + \exp(-S_k)}{2} \le 1 + c_2 S_k^2 \le \exp(c_2 S_k^2).$$

From this result and (2.106), we obtain

$$\int \left(\frac{dI\!P_1}{dI\!P_0}\right)^2 dI\!P_0 \le \exp\left(c_2 \sum_{k=1}^{m} S_k^2\right). \tag{2.108}$$

By (2.107),

$$\sum_{k=1}^{m} S_k^2 \le a_0^2 L^4 K_{\max}^4 (n h_n^{2\beta+1})^2 m = a_0^2 L^4 K_{\max}^4 n^2 m^{-(4\beta+1)}.$$

In view of the definition of m, it follows that the last expression is bounded by a constant depending only on a_0, L, $K_{\max}$, and c_0. Using this remark, (2.105), and (2.108), we conclude that there exists a real number α such that $\chi^2(I\!P_1, I\!P_0) \le \alpha$ for all n. Thus, all the assumptions of part (iii) of Theorem 2.15 are satisfied, and we obtain the lower bound

$$\inf_{\hat{F}_n} \sup_{f \in \Sigma(\beta,L)} P_f \left(n^{4\beta/(4\beta+1)} \left| \hat{F}_n - \int_0^1 f^2 \right| \ge A \right) \ge c_3 > 0. \tag{2.109}$$

Moreover, the following additional bound can be proved:

$$\inf_{\hat{F}_n} \sup_{f \in \Sigma(\beta,L)} P_f \left(\sqrt{n} \left| \hat{F}_n - \int_0^1 f^2 \right| \ge 1 \right) \ge c_4 > 0. \tag{2.110}$$

This inequality follows in a simple way, by choosing μ_0 and μ_1 to be two Dirac measures concentrated on the constant functions $f_0(x) \equiv 1$ and $f_1(x) \equiv 1 + n^{-1/2}$, respectively. The details of the proof are left to the reader.

Finally, (2.109) and (2.110) imply that

$$\inf_{\hat{F}_n} \sup_{f \in \Sigma(\beta,L)} \mathbf{E}_f \left[\psi_n^{-2} \left| \hat{F}_n - \int_0^1 f^2 \right|^2 \right] \ge c_5 > 0 \tag{2.111}$$

with the rate $\psi_n = \max(n^{-4\beta/(4\beta+1)}, n^{-1/2})$ which is faster than the optimal rate $n^{-\beta/(2\beta+1)}$ typical for estimation of smooth functions. It can be proved that bound (2.111) is sharp in the sense that the rate $n^{-4\beta/(4\beta+1)}$ is optimal for estimation of the quadratic functional if $\beta < 1/4$, while the optimal rate for $\beta \ge 1/4$ is $n^{-1/2}$ (see the bibliographic notes below).

2.8 Notes

The first minimax lower bound for nonparametric estimators dates back to Čencov (1962) (see also Čencov (1972)). He considered the problem of density estimation with the L_2-risk and proved his result using the integrated Cramér–Rao bound, a technique close to the van Trees inequality. Another early paper is due to Farrell (1972) who established a lower bound for density estimation at a fixed point. Le Cam's (1973) paper dealing mainly with parametric problems introduced important tools such as the inequalities of Lemma 2.3 and the Hellinger/total variation versions of the bounds based on two hypotheses (parts (i) and (ii) of Theorem 2.2).

Ibragimov and Has'minskii (1977) and Has'minskii (1978) pioneered the technique of lower bounds based on many hypotheses as well as the statistical application of Fano's lemma and of the Varshamov–Gilbert bound. These two powerful tools are borrowed from information theory (Fano (1952), Gilbert (1952), Gallager (1968), Cover and Thomas (2006)).

Lower bounds based on deviations of the likelihood ratios (Theorems 2.1 and 2.4, Propositions 2.1 and 2.3) are due to Korostelev and Tsybakov (1993). This technique is sometimes more convenient than the distance-based bounds. For instance, it can be useful in statistics of random processes when it is difficult to evaluate the classical distances (cf. Hoffmann (1999)).

A detailed account on the theory of f-divergences (originally introduced by Csizsár (1967)) can be found in the book of Vajda (1986).

Lemma 2.1 is due to Scheffé (1947). Pinsker (1964) proved a weaker version of the inequalities of Lemma 2.5. He showed the existence of constants $c_1 > 0$ and $c_2 > 0$ such that $V(P,Q) \le c_1\sqrt{K(P,Q)}$ for $K(P,Q) \le c_2$ and proved (2.21) with an unspecified constant $c_3 > 0$ instead of $\sqrt{2}$ in the second term. The first Pinsker inequality in its final form, as stated in Lemma 2.5, was obtained independently by Kullback (1967), Csizsár (1967), and Kemperman (1969). It is therefore sometimes called Kullback–Csizsár–Kemperman inequality. The second Pinsker inequality is a simple corollary of the first one; (2.21) can be found, for example, in Barron (1986). Lemma 2.6 is due to Bretagnolle and Huber (1979).

Minimax lower bounds at a fixed point for density estimation (extending those of Farrell (1972)) were obtained by Ibragimov and Has'minskii (1981) and Stone (1980), for nonparametric regression with random design by Stone (1980), and for nonparametric regression with fixed design by Korostelev and Tsybakov (1993). Minimax lower bounds in $L_p, 1 \le p \le \infty$, for density estimation are due to Čencov (1962, 1972), Has'minskii (1978), Bretagnolle and Huber (1979), and Ibragimov and Has'minskii (1983a). For nonparametric regression and for the Gaussian white noise model such bounds were obtained by Ibragimov and Has'minskii (1981, 1982, 1984). Stone (1982) established independently similar results for regression with random design and for density estimation. All these works proved optimal rates of the form $n^{-\beta/(2\beta+1)}$ and $(n/\log n)^{-\beta/(2\beta+1)}$ (or their multivariate analogs $n^{-\beta/(2\beta+d)}$ and $(n/\log n)^{-\beta/(2\beta+d)}$ where d is the dimension of the observations X_i), and considered mainly the Hölder classes of functions, along with some examples of Sobolev or Nikol'ski classes in the cases where such rates are optimal. Nemirovskii et al. (1985) and Nemirovskii (1985), considering the nonparametric regression problem with the L_p Sobolev classes, showed that other rates of convergence are optimal when the norm defining the class was not "matched" to the distance $d(\cdot,\cdot)$ defining the risk. They also showed that optimal rates might not be attained on linear estimators. Nemirovski (1985) established a complete description of optimal rates of convergence for the multivariate regression model when $d(\cdot,\cdot)$ is the L_p distance, and the functional class is the L_q Sobolev class. The same optimal rates of convergence are established for the Besov classes of functions (cf. Kerkyacharian and Picard (1992), Donoho and Johnstone (1998), Johnstone et al. (1996), Delyon and Juditsky (1996), Lepski et al. (1997)); for an overview and further references see Härdle et al. (1998).

Birgé (1983) and Yang and Barron (1999) suggested general techniques for derivation of minimax rates of convergence in an abstract setting. Their lower bounds are based on Fano's lemma. Refinements of Fano's lemma can be found in the papers of Gushchin (2002) and Birgé (2005).

Assouad's lemma appeared in Assouad (1983). In a slightly less general form it is given in the paper of Bretagnolle and Huber (1979) which contains already the main idea of the construction.

Inequality (2.88) is due to Gill and Levit (1995), who suggested calling it van Trees' inequality. They pioneered its use in the problem of estimation of functionals. Van Trees (1968, p. 72) heuristically presented a related but different result:

$$\mathbf{E}\left[(\xi - \mathbf{E}(\xi|\eta))^2\right] \geq \frac{1}{\mathbf{E}\left[\{\partial/\partial\xi \,(\log f(\xi,\eta))\}^2\right]},$$

where $f(\cdot,\cdot)$ is the joint density of two random variables ξ and η. Rigorous derivation of (2.88) from this inequality requires an additional technical step but Gill and Levit (1995) do not give all the details of the proof. They refer at this point to Borovkov and Sakhanenko (1980) and Borovkov (1984) who, however, worked under more restrictive assumptions. Borovkov and Sakhanenko (1980) and Borovkov (1984) assumed differentiability rather than absolute continuity of $t \mapsto p(x,t)$, and obtained some weighted versions of the van Trees inequality excluding the choice of weights that leads to (2.88). Belitser and Levit (1995) showed that the Pinsker constant (cf. Chapter 3) can be obtained using the van Trees inequality.

Lower bounds based on two fuzzy hypotheses are systematically used in the literature on nonparametric testing (cf. Ingster and Suslina (2003)). Usually it is sufficient to consider the measures μ_0 and μ_1 with disjoint supports. This is also sufficient to obtain the correct lower bounds for estimators of smooth functionals, such as the quadratic functional considered above (cf. Ibragimov et al. (1987)). However, for some nondifferentiable functionals (cf. Lepski et al. (1999)), the lower bounds invoke measures μ_0 and μ_1 whose supports are not disjoint. The results of Section 2.7.4 are applicable in this general case.

Optimal rates of estimation of the quadratic functional and of more general differentiable functionals were established by Ibragimov et al. (1987) and Nemirovskii (1990) for the Gaussian white noise model. Bickel and Ritov (1988) studied estimation of the quadratic functional for the density model. These papers discovered the elbow in the rates that occurs at $\beta = 1/4$. For a comprehensive account on estimation of functionals in the Gaussian white noise model see Nemirovskii (2000).

2.9 Exercises

Exercise 2.1 *Give an example of measures P_0 and P_1 such that $p_{e,1}$ is arbitrarily close to 1.* Hint: *Consider two discrete measures on $\{0,1\}$.*

Exercise 2.2 *Let P and Q be two probability mesures with densities p and q w.r.t. the Lebesgue measure on $[0,1]$ such that $0 < c_1 \le p(x), q(x) < c_2 < \infty$ for all $x \in [0,1]$. Show that the Kullback divergence $K(P,Q)$ is equivalent to the squared L_2 distance between the two densities, i.e.,*

$$k_1 \int (p(x) - q(x))^2 dx \le K(P,Q) \le k_2 \int (p(x) - q(x))^2 dx$$

where $k_1, k_2 > 0$ are constants. The same is true for the χ^2 divergence.

Exercise 2.3 *Prove that if the probability mesures P and Q are mutually absolutely continuous,*

$$K(P,Q) \le \chi^2(Q,P)/2.$$

Exercise 2.4 *Consider the nonparametric regression model*

$$Y_i = f(i/n) + \xi_i, \ i = 1, \dots, n,$$

where f is a function on $[0,1]$ with values in $\mathbf{R}$ and ξ_i are arbitrary random variables. Using the technique of two hypotheses show that

$$\liminf_{n\to\infty} \inf_{T_n} \sup_{f \in C[0,1]} E_f \|T_n - f\|_\infty = +\infty,$$

where $C[0,1]$ is the space of all continuous functions on $[0,1]$. In words, no rate of convergence can be attained uniformly on such a large functional class as $C[0,1]$.

Exercise 2.5 *Suppose that Assumptions (B) and (LP2) hold and assume that the random variables ξ_i are Gaussian. Prove (2.38) using Theorem 2.1.*

Exercise 2.6 *Improve the bound of Theorem 2.6 by computing the maximum on the right hand side of (2.48). Do we obtain that $p_{e,M}$ is arbitrarily close to 1 for $M \to \infty$ and $\alpha \to 0$, as in the Kullback case (cf. (2.53))?*

Exercise 2.7 *Consider the regression model with random design:*

$$Y_i = f(X_i) + \xi_i, \ i = 1, \dots, n,$$

where X_i are i.i.d. random variables with density $\mu(\cdot)$ on $[0,1]$ such that $\mu(x) \le \mu_0 < \infty, \forall\, x \in [0,1]$, the random variables ξ_i are i.i.d. with density p_ξ on $\mathbf{R}$, and the random vector $(X_1, \dots, X_n)$ is independent of $(\xi_1, \dots, \xi_n)$. Let $f \in \Sigma(\beta, L)$, $\beta > 0, L > 0$ and let $x_0 \in [0,1]$ be a fixed point.

(1) Suppose first that p_ξ satisfies

$$\int \left(\sqrt{p_\xi(y)} - \sqrt{p_\xi(y+t)} \right)^2 dy \le p_* t^2, \quad \forall\, t \in \mathbf{R},$$

where $0 < p_* < \infty$. Prove the bound

$$\liminf_{n\to\infty} \inf_{T_n} \sup_{f\in\Sigma(\beta,L)} \mathbf{E}_f \left[n^{\frac{2\beta}{2\beta+1}} |T_n(x_0) - f(x_0)|^2 \right] \geq c,$$

where $c > 0$ depends only on β, L, μ_0, p_*.

(2) Suppose now that the variables ξ_i are i.i.d. and uniformly distributed on $[-1, 1]$. Prove the bound

$$\liminf_{n\to\infty} \inf_{T_n} \sup_{f\in\Sigma(\beta,L)} \mathbf{E}_f \left[n^{\frac{2\beta}{\beta+1}} |T_n(x_0) - f(x_0)|^2 \right] \geq c',$$

where $c' > 0$ depends only on β, L, μ_0. Note that the rate here is $n^{-\frac{\beta}{\beta+1}}$, which is faster than the usual rate $n^{-\frac{\beta}{2\beta+1}}$. Furthermore, it can be proved that $\psi_n = n^{-\frac{\beta}{\beta+1}}$ is the optimal rate of convergence in the model with uniformly distributed errors.

Exercise 2.8 *Let $X_1, \ldots, X_n$ be i.i.d. random variables on $\mathbf{R}$ having density $p \in \mathcal{P}(\beta, L), \beta > 0, L > 0$. Show that*

$$\liminf_{n\to\infty} \inf_{T_n} \sup_{p\in\mathcal{P}(\beta,L)} \mathbf{E}_p \left[n^{\frac{2\beta}{2\beta+1}} |T_n(x_0) - p(x_0)|^2 \right] \geq c$$

for any $x_0 \in \mathbf{R}$ where $c > 0$ depends only on β and L.

Exercise 2.9 *Suppose that Assumptions (B) and (LP2) hold and let $x_0 \in [0, 1]$. Prove the bound (Stone, 1980):*

$$\lim_{a\to 0} \liminf_{n\to\infty} \inf_{T_n} \sup_{f\in\Sigma(\beta,L)} P_f \left(n^{\frac{\beta}{2\beta+1}} |T_n(x_0) - f(x_0)| \geq a \right) = 1. \tag{2.112}$$

Hint: *Introduce the hypotheses*

$$f_{0n}(x) \equiv 0, \quad f_{jn}(x) = \theta_j L h_n^\beta K\left(\frac{x - x_0}{h_n} \right),$$

with $\theta_j = j/M$, $j = 1, \ldots, M$.

Exercise 2.10 *Let $X_1, \ldots, X_n$ be i.i.d. random variables on $\mathbf{R}$ with density $p \in \mathcal{P}(\beta, L)$ where $\beta > 0$ and $L > 0$. Prove the bound*

$$\liminf_{n\to\infty} \inf_{T_n} \sup_{p\in\mathcal{P}(\beta,L)} \mathbf{E}_p \left[n^{\frac{2\beta}{2\beta+1}} \|T_n - p\|_2^2 \right] \geq c,$$

where $c > 0$ depends only on β and L.

Exercise 2.11 *Consider the nonparametric regression model*

$$Y_i = f(i/n) + \xi_i, \ i = 1, \ldots, n,$$

where the random variables ξ_i are i.i.d. with distribution $\mathcal{N}(0, 1)$ and where $f \in W^{per}(\beta, L)$, $L > 0$, and $\beta \in \{1, 2, \ldots\}$. Prove the bound

$$\liminf_{n\to\infty} \inf_{T_n} \sup_{f\in W^{per}(\beta,L)} \left(\frac{n}{\log n} \right)^{\frac{2\beta-1}{2\beta}} \mathbf{E}_f \|T_n - f\|_\infty^2 \geq c,$$

where $c > 0$ depends only on β and L.

3

Asymptotic efficiency and adaptation

3.1 Pinsker's theorem

In contrast to Chapters 1 and 2, here we will deal not only with the rates of convergence of estimators but with the exact asymptotic efficiency in the sense of Definition 2.2. More specifically, we will focus on exact asymptotic behavior of the minimax L_2-risk on the Sobolev ellipsoids (Pinsker's theorem).

Consider first the Gaussian white noise model defined in Chapter 1:

$$dY(t) = f(t)dt + \varepsilon dW(t), \quad t \in [0,1], \quad 0 < \varepsilon < 1. \tag{3.1}$$

We observe a sample path $\mathbf{X} = \{Y(t), 0 \le t \le 1\}$ of the process Y. In this chapter it will be mostly assumed that the function $f : [0,1] \to \mathbf{R}$ belongs to a Sobolev class. Recall that in Chapter 1 we defined several types of Sobolev classes. For $L > 0$ and integer β, the Sobolev classes $W(\beta, L)$ and $W^{per}(\beta, L)$ are given in Definition 1.11. Then the Sobolev classes $\tilde{W}(\beta, L)$ are introduced in Definition 1.12 as an extension of the periodic classes $W^{per}(\beta, L)$ to all $\beta > 0$. In this chapter we are going to deal mainly with classes $\tilde{W}(\beta, L)$. Recall their definition:

$$\tilde{W}(\beta, L) = \{f \in L_2[0,1] : \theta = \{\theta_j\} \in \Theta(\beta, Q)\}, \qquad Q = \frac{L^2}{\pi^{2\beta}},$$

where $\theta_j = \int_0^1 f\varphi_j$, $\{\varphi_j\}_{j=1}^{\infty}$ is the trigonometric basis defined in Example 1.3, and $\Theta(\beta, Q)$ is the ellipsoid

$$\Theta(\beta, Q) = \left\{\theta = \{\theta_j\} \in \ell^2(\mathbf{N}) : \sum_{j=1}^{\infty} a_j^2 \theta_j^2 \le Q\right\}$$

with

$$a_j = \begin{cases} j^\beta, & \text{if } j \text{ is even,} \\ (j-1)^\beta, & \text{if } j \text{ is odd.} \end{cases} \tag{3.2}$$

A. B. Tsybakov, *Introduction to Nonparametric Estimation*,
DOI 10.1007/978-0-387-79052-7_3,

If $\beta \geq 1$ is integer, we have $W^{per}(\beta, L) = \tilde{W}(\beta, L)$ (see Chapter 1).

Given the model (3.1), the following infinite sequence of Gaussian observations is available to the statistician:

$$y_j = \int_0^1 \varphi_j(t) dY(t) = \theta_j + \varepsilon\, \xi_j, \quad j = 1, 2, \ldots,$$

where $\xi_j = \int_0^1 \varphi_j(x) dW(x)$ are i.i.d. $\mathcal{N}(0,1)$ random variables. Define the following estimator of f:

$$\hat{f}_\varepsilon(x) = \sum_{j=1}^{\infty} \ell_j^* y_j \varphi_j(x) \tag{3.3}$$

where

$$\ell_j^* = (1 - \kappa^* a_j)_+, \tag{3.4}$$

$$\kappa^* = \left(\frac{\beta}{(2\beta+1)(\beta+1)Q} \right)^{\frac{\beta}{2\beta+1}} \varepsilon^{\frac{2\beta}{2\beta+1}}. \tag{3.5}$$

Observe that $\hat{f}_\varepsilon$ is a weighted projection estimator. The number of nonzero terms $N = \max\{j : \ell_j^* > 0\}$ in the sum (3.3) is finite, so that we can write

$$\hat{f}_\varepsilon(x) = \sum_{j=1}^{N} \ell_j^* y_j \varphi_j(x).$$

It is easy to see that $N = N_\varepsilon$ tends to infinity with the rate $\varepsilon^{-2/(2\beta+1)}$, as $\varepsilon \to 0$.

Theorem 3.1 (Pinsker's theorem). *Let $\beta > 0, L > 0$. Then*

$$\lim_{\varepsilon \to 0} \sup_{f \in \tilde{W}(\beta, L)} \varepsilon^{-\frac{4\beta}{2\beta+1}} \mathbf{E}_f \|\hat{f}_\varepsilon - f\|_2^2 = \lim_{\varepsilon \to 0} \inf_{T_\varepsilon} \sup_{f \in \tilde{W}(\beta, L)} \varepsilon^{-\frac{4\beta}{2\beta+1}} \mathbf{E}_f \|T_\varepsilon - f\|_2^2 = C^*,$$

where $\inf_{T_\varepsilon}$ denotes the infimum over all estimators, $\mathbf{E}_f$ stands for the expectation with respect to distribution of the observation $\mathbf{X}$ under the model (3.1), $\|\cdot\|_2$ is the $L_2([0,1], dx)$-norm, and

$$\begin{aligned} C^* &= L^{\frac{2}{2\beta+1}} (2\beta+1)^{\frac{1}{2\beta+1}} \left(\frac{\beta}{\pi(\beta+1)} \right)^{\frac{2\beta}{2\beta+1}} \\ &= [Q(2\beta+1)]^{\frac{1}{2\beta+1}} \left(\frac{\beta}{\beta+1} \right)^{\frac{2\beta}{2\beta+1}} \end{aligned} \tag{3.6}$$

with $Q = L^2/\pi^{2\beta}$.

The quantity C^* given by (3.6) is called the *Pinsker constant*. The proof of Theorem 3.1 is deferred to Section 3.3.

Theorem 3.1 implies that estimator (3.3) is asymptotically efficient on $(\tilde{W}(\beta, L), \|\cdot\|_2)$ in the sense of Definition 2.2:

$$\lim_{\varepsilon \to 0} \sup_{f \in \tilde{W}(\beta,L)} \frac{\mathbf{E}_f \|\hat{f}_\varepsilon - f\|_2^2}{\mathcal{R}_\varepsilon^*} = 1, \tag{3.7}$$

where $\mathcal{R}_\varepsilon^*$ is the minimax risk

$$\mathcal{R}_\varepsilon^* \stackrel{\triangle}{=} \inf_{T_\varepsilon} \sup_{f \in \tilde{W}(\beta,L)} \mathbf{E}_f \|T_\varepsilon - f\|_2^2.$$

Observe that we use here a slightly modified version of Definition 2.2, with the real-valued asymptotic parameter ε tending to zero instead of the integer-valued n tending to ∞.

A result similar to Theorem 3.1 holds for the nonparametric regression model

$$Y_i = f(i/n) + \xi_i, \quad i = 1, \ldots, n, \tag{3.8}$$

where ξ_i are i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables, $\sigma^2 > 0$. A direct correspondence can be obtained by simply putting $\varepsilon = \sigma/\sqrt{n}$ in Theorem 3.1, as it can be seen from the following theorem.

Theorem 3.2 *There exists an estimator $\hat{f}_n$ of f such that*

$$\lim_{n \to \infty} \sup_{f \in \mathcal{F}} \mathbf{E}_f \left(n^{\frac{2\beta}{2\beta+1}} \|\hat{f}_n - f\|_2^2 \right) = \lim_{n \to \infty} \inf_{T_n} \sup_{f \in \mathcal{F}} \mathbf{E}_f \left(n^{\frac{2\beta}{2\beta+1}} \|T_n - f\|_2^2 \right)$$
$$= C^* \sigma^{\frac{4\beta}{2\beta+1}},$$

where $\inf_{T_n}$ *denotes the infimum over all estimators,* $\mathbf{E}_f$ *stands for the expectation with respect to the distribution of* $(Y_1, \ldots, Y_n)$ *under the model (3.8) and* $\mathcal{F} = W(\beta, L)$, $\beta \in \{1, 2, \ldots\}, L > 0$, *or* $\mathcal{F} = \tilde{W}(\beta, L)$, $\beta \geq 1$, $L > 0$.

The proof of this theorem follows essentially the same lines as that of Theorem 3.1, up to some additional technicalities related to the discreteness of the design points and possible nonperiodicity of the underlying functions f. In order to focus on the main ideas, we will only give the proof of Theorem 3.1.

Consider the class of all *linear estimators*, that is, the estimators of the form

$$f_{\varepsilon,\lambda}(x) = \sum_{j=1}^{\infty} \lambda_j y_j \varphi_j(x), \tag{3.9}$$

where the weights $\lambda_j \in \mathbf{R}$ are such that the sequence

$$\lambda = (\lambda_1, \lambda_2, \ldots)$$

belongs to $\ell^2(\mathbf{N})$; equation (3.9) is understood in the sense that $f_{\varepsilon,\lambda}$ is the mean square limit of the random series on the right hand side.

Observe that $\hat{f}_\varepsilon$ defined by (3.3) is a linear estimator. Since $\hat{f}_\varepsilon$ is asymptotically efficient among all the estimators in the minimax sense (cf. (3.7)), it follows that $\hat{f}_\varepsilon$ is asymptotically efficient among the linear estimators, that is,

$$\lim_{\varepsilon\to 0} \frac{\sup_{f\in\tilde{W}(\beta,L)} \mathbf{E}_f \|\hat{f}_\varepsilon - f\|_2^2}{\inf_\lambda \sup_{f\in\tilde{W}(\beta,L)} \mathbf{E}_f \|f_{\varepsilon,\lambda} - f\|_2^2} = 1.$$

From now on, we will write $\inf_\lambda = \inf_{\lambda\in\ell^2(\mathbf{N})}$. Before proving Theorem 3.1, let us first check that this linear optimality holds.

3.2 Linear minimax lemma

In this section we deal with the Gaussian sequence model

$$y_j = \theta_j + \varepsilon\xi_j, \quad j = 1, 2, \ldots, \tag{3.10}$$

with $\theta = (\theta_1, \theta_2, \ldots) \in \ell^2(\mathbf{N})$ and $0 < \varepsilon < 1$ where ξ_j are i.i.d. $\mathcal{N}(0,1)$ random variables. We observe the random sequence

$$y = (y_1, y_2, \ldots).$$

Recall that we have an access to such a sequence of observations if we deal with the Gaussian white noise model (3.1): in this case we can take $y_j = \int_0^1 \varphi_j(t)dY(t)$ and $\theta_j = \int_0^1 \varphi_j(t)f(t)dt$, where $\{\varphi_j\}$ is the trigonometric basis (cf. Section 1.10). Put

$$\hat{\theta}_j(\lambda) = \lambda_j y_j, \quad j = 1, 2, \ldots,$$
$$\hat{\theta}(\lambda) = (\hat{\theta}_1(\lambda), \hat{\theta}_2(\lambda), \ldots).$$

By (1.112), the risk of the linear estimator $f_{\varepsilon,\lambda}$ is

$$\begin{aligned}\mathbf{E}_f \|f_{\varepsilon,\lambda} - f\|_2^2 &= \mathbf{E}_\theta \|\hat{\theta}(\lambda) - \theta\|^2 \\ &= \sum_{j=1}^\infty [(1-\lambda_j)^2\theta_j^2 + \varepsilon^2\lambda_j^2] \\ &\triangleq R(\lambda, \theta),\end{aligned}$$

where $\mathbf{E}_\theta$ denotes expectation with respect to the distribution of y in model (3.10). Therefore, the linear minimax risk in model (3.1) is equal to the linear minimax risk in model (3.10):

$$\inf_\lambda \sup_{f\in\tilde{W}(\beta,L)} \mathbf{E}_f \|f_{\varepsilon,\lambda} - f\|_2^2 = \inf_\lambda \sup_{\theta\in\Theta(\beta,Q)} \mathbf{E}_\theta \|\hat{\theta}(\lambda) - \theta\|^2. \tag{3.11}$$

The results of this section will lead us to the following asymptotics for the risk of linear estimators:

$$\inf_{\lambda} \sup_{\theta\in\Theta(\beta,Q)} \mathbf{E}_\theta \|\hat\theta(\lambda)-\theta\|^2 = C^* \varepsilon^{\frac{4\beta}{2\beta+1}}(1+o(1)), \quad \varepsilon\to 0, \tag{3.12}$$

where C^* is the Pinsker constant defined in (3.6).

Consider now a general ellipsoid (not necessarily a Sobolev one):

$$\Theta = \Big\{\theta=\{\theta_j\} : \sum_{j=1}^{\infty} a_j^2\theta_j^2 \le Q\Big\}, \tag{3.13}$$

where $a_j \ge 0$ are arbitrary coefficients and $Q>0$ is a finite constant.

Definition 3.1 *The* **linear minimax risk** *on the ellipsoid Θ is defined by*

$$R^L = \inf_{\lambda}\sup_{\theta\in\Theta} R(\lambda,\theta).$$

A linear estimator $\hat\theta(\lambda^*)$ with $\lambda^*\in\ell^2(\mathbf{N})$ is called a **linear minimax estimator** *if*

$$\sup_{\theta\in\Theta} R(\lambda^*,\theta) = R^L$$

or a **linear asymptotically minimax estimator** *if*

$$\lim_{\varepsilon\to 0} \frac{\sup_{\theta\in\Theta} R(\lambda^*,\theta)}{R^L} = 1.$$

It is easy to see that

$$\inf_{\lambda} R(\lambda,\theta) = \sum_{j=1}^{\infty} \frac{\varepsilon^2\theta_j^2}{\varepsilon^2+\theta_j^2}\ . \tag{3.14}$$

Now introduce an equation with respect to the variable κ,

$$\frac{\varepsilon^2}{\kappa}\sum_{j=1}^{\infty} a_j(1-\kappa a_j)_+ = Q \tag{3.15}$$

and let us show that solutions $\kappa=\kappa(\varepsilon)>0$ of (3.15) exist. This equation will play an important role in what follows.

Lemma 3.1 *If $a_j\ge 0$ is an increasing sequence and $a_j\to+\infty$, then there exists a unique solution of (3.15) given by*

$$\kappa = \frac{\varepsilon^2\sum_{m=1}^{N} a_m}{Q+\sum_{m=1}^{N}\varepsilon^2 a_m^2}\ , \tag{3.16}$$

with

$$N = \max\Big\{j : \varepsilon^2 \sum_{m=1}^{j} a_m(a_j - a_m) < Q\Big\} < +\infty.$$

PROOF. Observe that the sequence $\tilde{a}_j = \sum_{m=1}^{j} a_m(a_j - a_m)$ is increasing and $\tilde{a}_j \to +\infty$. Thus, the value N defined in the statement of the lemma is finite. For all $j \le N$, we have

$$\varepsilon^2 \sum_{m=1}^{N} a_m(a_N - a_m) \ge \varepsilon^2 \sum_{m=1}^{N} a_m(a_j - a_m).$$

By the definition of N, this implies that

$$\forall j \le N : \quad Q > \varepsilon^2 \sum_{m=1}^{N} a_m(a_j - a_m). \tag{3.17}$$

On the other hand, by the same definition, we obtain for all $j > N$

$$\begin{aligned} \varepsilon^2 \sum_{m=1}^{N} a_m(a_j - a_m) &\ge \varepsilon^2 \sum_{m=1}^{N} a_m(a_{N+1} - a_m) \\ &= \varepsilon^2 \sum_{m=1}^{N+1} a_m(a_{N+1} - a_m) \ge Q. \end{aligned} \tag{3.18}$$

By (3.16)–(3.18), we have $1 - \kappa a_j > 0$ for $j \le N$ and $1 - \kappa a_j \le 0$ for $j > N$. Then

$$N = \max\{j : a_j < 1/\kappa\}. \tag{3.19}$$

By (3.16) and (3.19),

$$\frac{\varepsilon^2}{\kappa} \sum_{j=1}^{\infty} a_j(1 - \kappa a_j)_+ = \frac{\varepsilon^2}{\kappa} \sum_{j=1}^{N} a_j(1 - \kappa a_j) = Q.$$

This means that the value κ defined by (3.16) is a solution of (3.15). This solution is unique since the function

$$\frac{\varepsilon^2}{t} \sum_{j=1}^{\infty} a_j(1 - t a_j)_+ = \varepsilon^2 \sum_{j=1}^{\infty} a_j(1/t - a_j)_+$$

is decreasing in t for $0 < t \le 1/\min\{a_j : a_j > 0\}$. In addition, each solution κ of (3.15) should necessarily satisfy $\kappa \le 1/\min\{a_j : a_j > 0\}$ since otherwise we have $\kappa a_j > 1$ for all j such that $a_j \ne 0$, and the left hand side of (3.15) becomes zero. ∎

Suppose now that there exists a solution κ of (3.15). This is the case, for example, when the assumptions of Lemma 3.1 are satisfied. For such a solution, put

$$\ell_j \stackrel{\triangle}{=} (1 - \kappa a_j)_+, \quad j = 1, 2, \ldots, \qquad \ell = (\ell_1, \ell_2, \ldots), \tag{3.20}$$

and

$$\mathcal{D}^* \stackrel{\triangle}{=} \varepsilon^2 \sum_{j=1}^{\infty} (1 - \kappa a_j)_+ = \varepsilon^2 \sum_{j=1}^{\infty} \ell_j,$$

assuming that the last sum is finite.

Lemma 3.2 (Linear minimax lemma.) *Suppose that Θ is a general ellipsoid (3.13) with $Q > 0$ and let the sequence $a_j \geq 0$ be such that* $\mathrm{Card}\{j : a_j = 0\} < \infty$. *Suppose also that there exists a solution κ of (3.15) and $\mathcal{D}^* < \infty$. Assume that ℓ is defined by (3.20). Then the risk $R(\lambda, \theta)$ satisfies*

$$\inf_{\lambda} \sup_{\theta \in \Theta} R(\lambda, \theta) = \sup_{\theta \in \Theta} \inf_{\lambda} R(\lambda, \theta) = \sup_{\theta \in \Theta} R(\ell, \theta) = \mathcal{D}^*. \tag{3.21}$$

Proof. Obviously,

$$\sup_{\theta \in \Theta} \inf_{\lambda} R(\lambda, \theta) \leq \inf_{\lambda} \sup_{\theta \in \Theta} R(\lambda, \theta) \leq \sup_{\theta \in \Theta} R(\ell, \theta).$$

Therefore it remains to show that

$$\sup_{\theta \in \Theta} R(\ell, \theta) \leq \mathcal{D}^* \tag{3.22}$$

and

$$\sup_{\theta \in \Theta} \inf_{\lambda} R(\lambda, \theta) \geq \mathcal{D}^*. \tag{3.23}$$

Proof of (3.22). For all $\theta \in \Theta$, we have

$$\begin{aligned}
R(\ell, \theta) &= \sum_{i=1}^{\infty} ((1 - \ell_i)^2 \theta_i^2 + \varepsilon^2 \ell_i^2) \\
&= \varepsilon^2 \sum_{i=1}^{\infty} \ell_i^2 + \sum_{i: a_i > 0}^{\infty} (1 - \ell_i)^2 a_i^{-2} a_i^2 \theta_i^2 \quad (\text{since } \ell_i = 1 \text{ for } a_i = 0) \\
&\leq \varepsilon^2 \sum_{i=1}^{\infty} \ell_i^2 + Q \sup_{i: a_i > 0} [(1 - \ell_i)^2 a_i^{-2}] \\
&\leq \varepsilon^2 \sum_{i=1}^{\infty} \ell_i^2 + Q\kappa^2 \quad (\text{since } 1 - \kappa a_i \leq \ell_i \leq 1) \\
&= \varepsilon^2 \sum_{i=1}^{\infty} \ell_i^2 + \varepsilon^2 \kappa \sum_{i=1}^{\infty} a_i \ell_i \quad (\text{by (3.15)})
\end{aligned}$$

$$= \varepsilon^2 \sum_{i=1}^{\infty} \ell_i(\ell_i + \kappa a_i)$$
$$= \varepsilon^2 \sum_{i:\ell_i \neq 0} \ell_i(\ell_i + \kappa a_i) = \varepsilon^2 \sum_{i:\ell_i \neq 0} \ell_i = \mathcal{D}^*. \tag{3.24}$$

Proof of (3.23). Denote by V the set of all sequences $v = (v_1, v_2, \ldots)$ such that $v_j \in \mathbf{R}$ (without any restriction) if $a_j = 0$, and

$$v_j^2 = \frac{\varepsilon^2(1 - \kappa a_j)_+}{\kappa a_j}, \quad \text{if } a_j > 0. \tag{3.25}$$

Then $V \subset \Theta$ by (3.15). Therefore

$$\begin{aligned}
\sup_{\theta \in \Theta} \inf_{\lambda} R(\lambda, \theta) &\geq \sup_{v \in V} \inf_{\lambda} \sum_{i=1}^{\infty} [(1 - \lambda_i)^2 v_i^2 + \varepsilon^2 \lambda_i^2] \\
&= \sup_{v \in V} \left[\sum_{i:a_i=0} \frac{v_i^2 \varepsilon^2}{v_i^2 + \varepsilon^2} + \sum_{i:a_i>0} \frac{v_i^2 \varepsilon^2}{v_i^2 + \varepsilon^2} \right] \\
&= \varepsilon^2 \mathrm{Card}\{i : a_i = 0\} + \sum_{i:a_i>0} \frac{\varepsilon^4 (1 - \kappa a_i)_+}{\varepsilon^2(\kappa a_i + (1 - \kappa a_i)_+)} \\
&= \varepsilon^2 \mathrm{Card}\{i : a_i = 0\} + \varepsilon^2 \sum_{i:a_i>0} (1 - \kappa a_i)_+ \\
&= \varepsilon^2 \sum_{i=1}^{\infty} (1 - \kappa a_i)_+ = \mathcal{D}^*. \qquad \blacksquare
\end{aligned}$$

The estimator $\hat{\theta}(\ell)$ with the weight sequence ℓ defined by (3.20) and (3.15) is called the *Pinsker estimator* for the (general) ellipsoid Θ. The weights ℓ in (3.20) are called the *Pinsker weights*. Lemma 3.2 then shows that the Pinsker estimator is a linear minimax estimator for a general ellipsoid Θ. Let us now study the case of the Sobolev ellipsoid $\Theta(\beta, Q)$ in more detail.

Lemma 3.3 *Consider the ellipsoid $\Theta = \Theta(\beta, Q)$ defined by (3.13) with $Q > 0$ and*

$$a_j = \begin{cases} j^\beta, & \text{for even } j, \\ (j-1)^\beta, & \text{for odd } j, \end{cases}$$

where $\beta > 0$. Then:

(i) there exists a solution κ of (3.15) which is unique and satisfies

$$\kappa = \kappa^*(1 + o(1)) \quad \text{as} \quad \varepsilon \to 0, \tag{3.26}$$

for κ^* defined in (3.5);

(ii)

$$\mathcal{D}^* = C^* \varepsilon^{\frac{4\beta}{2\beta+1}} (1 + o(1)) \quad \text{as} \quad \varepsilon \to 0 \tag{3.27}$$

where C^* is the Pinsker constant;

(iii)

$$\max_{j\geq 2} v_j^2 a_j^2 = O(\varepsilon^{\frac{2}{2\beta+1}}) \quad as \;\; \varepsilon \to 0 \tag{3.28}$$

where v_j^2 is defined in (3.25).

Proof. (i) We have

$$a_1 = 0, \quad a_{2m} = a_{2m+1} = (2m)^\beta, \quad m = 1, 2, \ldots .$$

Lemma 3.1 implies that there exists a unique solution of (3.15). Moreover, from (3.15) we get

$$\begin{aligned} Q &= \frac{\varepsilon^2}{\kappa} \sum_{j=2}^{\infty} a_j (1 - \kappa a_j)_+ \\ &= \frac{2\varepsilon^2}{\kappa} \sum_{m=1}^{\infty} (2m)^\beta (1 - \kappa (2m)^\beta)_+ = \frac{2\varepsilon^2}{\kappa} \sum_{m=1}^{M} (2m)^\beta (1 - \kappa (2m)^\beta) \end{aligned}$$

with $M = \lfloor (1/\kappa)^{1/\beta}/2 \rfloor$. Next, for $a > 0$,

$$\sum_{m=1}^{M} m^a = \frac{M^{a+1}}{a+1}(1 + o(1)) \qquad \text{as } M \to \infty$$

giving

$$Q = \frac{\varepsilon^2 \beta}{(2\beta+1)(\beta+1)\kappa^{(2\beta+1)/\beta}}(1 + o(1)) \qquad \text{as } \kappa \to 0.$$

This implies that the solution κ of (3.15) satisfies

$$\kappa = \left(\frac{\beta}{(2\beta+1)(\beta+1)Q} \right)^{\frac{\beta}{2\beta+1}} \varepsilon^{\frac{2\beta}{2\beta+1}} (1 + o(1)) \qquad \text{as } \varepsilon \to 0.$$

(ii) Using the argument as in (i) and invoking (3.26) we obtain

$$\begin{aligned} \mathcal{D}^* &= \varepsilon^2 \sum_{j=1}^{\infty} (1 - \kappa a_j)_+ = \varepsilon^2 + 2\varepsilon^2 \sum_{m=1}^{M} (1 - \kappa (2m)^\beta) \\ &= \varepsilon^2 + 2\varepsilon^2 \left[M - 2^\beta \kappa \frac{M^{\beta+1}}{\beta+1}(1 + o(1)) \right] \\ &= [Q\,(2\beta+1)]^{\frac{1}{2\beta+1}} \left(\frac{\beta}{\beta+1} \right)^{\frac{2\beta}{2\beta+1}} \varepsilon^{\frac{4\beta}{2\beta+1}} (1 + o(1)) \\ &= C^* \varepsilon^{\frac{4\beta}{2\beta+1}} (1 + o(1)). \end{aligned}$$

(iii) In order to prove (3.28), observe that $v_j^2 = 0$ for $j > N$, whereas $a_N < 1/\kappa$. Therefore

$$v_j^2 a_j^2 = \frac{\varepsilon^2 a_j (1 - \kappa a_j)_+}{\kappa} \leq \frac{\varepsilon^2 a_N}{\kappa} \leq \frac{\varepsilon^2}{\kappa^2} = O(\varepsilon^{\frac{2}{2\beta+1}}) \quad \text{as } \varepsilon \to 0. \qquad \blacksquare$$

Corollary 3.1 *Let $\hat{\theta}(\ell)$ be the Pinsker estimator on the ellipsoid $\Theta(\beta, Q)$ with $\beta > 0$ and $Q > 0$. Then*

$$\inf_{\lambda} \sup_{\theta \in \Theta(\beta,Q)} \mathbf{E}_\theta \|\hat{\theta}(\lambda) - \theta\|^2 = \sup_{\theta \in \Theta(\beta,Q)} \mathbf{E}_\theta \|\hat{\theta}(\ell) - \theta\|^2 \tag{3.29}$$
$$= C^* \varepsilon^{\frac{4\beta}{2\beta+1}} (1 + o(1))$$

as $\varepsilon \to 0$ where C^* is the Pinsker constant.

The proof follows immediately from (3.21) and (3.27).

3.3 Proof of Pinsker's theorem

The proof of Theorem 3.1 consists in establishing the *upper bound on the risk*:

$$\sup_{f \in \tilde{W}(\beta,L)} \mathbf{E}_f \|\hat{f}_\varepsilon - f\|_2^2 \le C^* \varepsilon^{\frac{4\beta}{2\beta+1}} (1 + o(1)), \quad \text{as} \quad \varepsilon \to 0, \tag{3.30}$$

and the *lower bound on the minimax risk*:

$$\mathcal{R}_\varepsilon^* \stackrel{\triangle}{=} \inf_{T_\varepsilon} \sup_{f \in \tilde{W}(\beta,L)} \mathbf{E}_f \|T_\varepsilon - f\|_2^2 \ge C^* \varepsilon^{\frac{4\beta}{2\beta+1}} (1 + o(1)), \tag{3.31}$$

as $\varepsilon \to 0$.

3.3.1 Upper bound on the risk

Since

$$\hat{f}_\varepsilon(x) = \sum_{j=1}^{\infty} \ell_j^* y_j \varphi_j(x) \quad \text{with} \quad \ell_j^* = (1 - \kappa^* a_j)_+ \ ,$$

we can write

$$\mathbf{E}_f \|\hat{f}_\varepsilon - f\|_2^2 = \mathbf{E}_\theta \|\hat{\theta}(\ell^*) - \theta\|^2 = R(\ell^*, \theta), \tag{3.32}$$

where θ is the sequence of Fourier coefficients of f and ℓ^* is the sequence of weights defined by (3.4):

$$\ell^* = (\ell_1^*, \ell_2^*, \ldots).$$

We now show that the maximum risk of $\hat{\theta}(\ell^*)$ on $\Theta(\beta, Q)$ asymptotically behaves in the same way as that of the Pinsker estimator $\hat{\theta}(\ell)$. Since the definition of $\hat{\theta}(\ell^*)$ is explicit and more simple than that of $\hat{\theta}(\ell)$, we will call $\hat{\theta}(\ell^*)$ the *simplified Pinsker estimator* and the weights ℓ^* the *simplified Pinsker weights*.

As in the proof of (3.22) we obtain, for all $\theta \in \Theta(\beta, Q)$,

$$R(\ell^*, \theta) \le \varepsilon^2 \sum_{j=1}^{\infty} (\ell_j^*)^2 + Q(\kappa^*)^2.$$

Define $M^* = \lfloor (1/\kappa^*)^{1/\beta}/2 \rfloor$, $M = \lfloor (1/\kappa)^{1/\beta}/2 \rfloor$ where κ is the solution of (3.15). Applying the same argument as that used to prove Lemma 3.3 and invoking (3.26) we find

$$\begin{aligned}
\varepsilon^2 \sum_{j=1}^{\infty} (\ell_j^*)^2 + Q(\kappa^*)^2 &= \varepsilon^2 + 2\varepsilon^2 \sum_{m=1}^{M^*} (1 - \kappa^*(2m)^\beta)^2 + Q\kappa^2(1 + o(1)) \\
&= \varepsilon^2 + 2\varepsilon^2 \left(M^* - 2^{\beta+1}\kappa^* \frac{(M^*)^{\beta+1}}{\beta+1} \right. \\
&\quad \left. + 4^\beta (\kappa^*)^2 \frac{(M^*)^{2\beta+1}}{2\beta+1} \right) (1 + o(1)) + Q\kappa^2(1 + o(1)) \\
&= \varepsilon^2 + 2\varepsilon^2 \left(M - 2^{\beta+1}\kappa \frac{M^{\beta+1}}{\beta+1} \right. \\
&\quad \left. + 4^\beta \kappa^2 \frac{M^{2\beta+1}}{2\beta+1} \right) (1 + o(1)) + Q\kappa^2(1 + o(1)) \\
&= \left[\varepsilon^2 + 2\varepsilon^2 \sum_{m=1}^{M} (1 - \kappa(2m)^\beta)^2 + Q\kappa^2 \right] (1 + o(1)) \\
&= \left[\varepsilon^2 \sum_{j=1}^{\infty} \ell_j^2 + Q\kappa^2 \right] (1 + o(1)) \\
&= \mathcal{D}^*(1 + o(1)) \quad \text{as } \varepsilon \to 0
\end{aligned}$$

where the last equality follows from (3.24). Therefore,

$$\sup_{\theta \in \Theta(\beta, Q)} R(\ell^*, \theta) \le \mathcal{D}^*(1 + o(1)) \quad \text{as } \varepsilon \to 0. \tag{3.33}$$

Upper bound (3.30) follows from (3.33) and (3.27) if we observe that

$$\sup_{f \in \tilde{W}(\beta, L)} \mathbf{E}_f \|\hat{f}_\varepsilon - f\|_2^2 = \sup_{\theta \in \Theta(\beta, Q)} R(\ell^*, \theta).$$

3.3.2 Lower bound on the minimax risk

Preliminaries: A Bayes problem in dimension 1

Consider a statistical model with a single Gaussian observation $x \in \mathbf{R}$:

$$x = a + \varepsilon\xi, \quad a \in \mathbf{R}, \quad \xi \sim \mathcal{N}(0, 1), \quad \varepsilon > 0. \tag{3.34}$$

For an estimator $\hat{a} = \hat{a}(x)$ of the parameter a, define its squared risk $\mathbf{E}\left[(\hat{a} - a)^2\right]$, as well as its *Bayes risk* with respect to the prior distribution $\mathcal{N}(0, s^2)$ with $s > 0$:

$$\mathcal{R}^B(\hat{a}) = \int \mathbf{E}\left[(\hat{a}-a)^2\right]\mu_s(a)da \tag{3.35}$$
$$= \int_{\mathbf{R}^2} (\hat{a}(x)-a)^2 \mu_\varepsilon(x-a)\mu_s(a)\,dx\,da,$$

where

$$\mu_s(u) = \frac{1}{s}\varphi\left(\frac{u}{s}\right)$$

and where $\varphi(\cdot)$ denotes the density of $\mathcal{N}(0,1)$. The *Bayes estimator* $\hat{a}^B$ is defined as the minimizer of the Bayes risk among all estimators:

$$\hat{a}^B = \arg\min_{\hat{a}} \mathcal{R}^B(\hat{a}).$$

The Bayes risk can be represented in the form

$$\mathcal{R}^B(\hat{a}) = I\!E\left[(\hat{a}(x)-a)^2\right],$$

where $I\!E$ denotes expectation with respect to the distribution of the Gaussian pair (x,a) such that $x = a+\varepsilon\xi$, where a is Gaussian with distribution $\mathcal{N}(0,s^2)$ and independent of ξ. By a classical argument, $\hat{a}^B$ and $\mathcal{R}^B(\hat{a})$ are equal to the conditional mean and variance:

$$\hat{a}^B = I\!E(a|x),$$
$$\mathcal{R}^B(\hat{a}^B) = \min_{\hat{a}} \mathcal{R}^B(\hat{a}) = I\!E\left[\mathrm{Var}(a|x)\right].$$

Since the pair (x,a) is Gaussian, the variance $\mathrm{Var}(a|x)$ is independent of x and we easily get the following lemma.

Lemma 3.4 *The Bayes estimator of parameter a in model (3.34) is*

$$\hat{a}^B = \frac{s^2}{\varepsilon^2+s^2}\,x$$

and the minimum value of the Bayes risk is

$$\mathcal{R}^B(\hat{a}^B) = \mathrm{Var}(a|x) \equiv \frac{s^2\varepsilon^2}{\varepsilon^2+s^2}\ .$$

We proceed now to the proof of the lower bound (3.31). It is divided into four steps.

Step 1. Reduction to a parametric family

Let $N = \max\{j : \ell_j > 0\}$ where ℓ_j are the Pinsker weights (3.20) and let

$$\Theta_N = \left\{\theta^N = (\theta_2,\dots,\theta_N) \in \mathbf{R}^{N-1} : \sum_{j=2}^{N} a_j^2\theta_j^2 \le Q\right\}$$

and

$$\mathcal{F}_N = \Big\{ f_{\theta^N}(x) = \sum_{j=2}^{N} \theta_j \varphi_j(x) : (\theta_2, \dots, \theta_N) \in \Theta_N \Big\}.$$

The set $\mathcal{F}_N$ is a parametric family of finite dimension $N-1$ and

$$\mathcal{F}_N \subset \tilde{W}(\beta, L).$$

Therefore

$$\mathcal{R}_\varepsilon^* \geq \inf_{T_\varepsilon} \sup_{f \in \mathcal{F}_N} \mathbf{E}_f \|T_\varepsilon - f\|_2^2.$$

For all $f \in \mathcal{F}_N$ and all T_ε, there exists a random vector $\hat{\theta}^N = (\hat{\theta}_2, \dots, \hat{\theta}_N) \in \Theta_N$ such that

$$\|T_\varepsilon - f\|_2 \geq \Big\| \sum_{j=2}^{N} \hat{\theta}_j \varphi_j - f \Big\|_2 \tag{3.36}$$

almost surely. In fact, if the realization Y is such that $T_\varepsilon \in L_2[0,1]$, it is sufficient to take as estimator $\sum_{j=2}^{N} \hat{\theta}_j \varphi_j$ the $L_2[0,1]$ projection of T_ε on $\mathcal{F}_N$ (indeed, the set $\mathcal{F}_N$ is convex and closed). If $T_\varepsilon \notin L_2[0,1]$, the left hand side of (3.36) equals $+\infty$ and inequality (3.36) is trivial for all $(\hat{\theta}_2, \dots, \hat{\theta}_N) \in \Theta_N$.

With the notation $\mathbf{E}_\theta \stackrel{\triangle}{=} \mathbf{E}_{f_{\theta^N}}$ and in view of (3.36), we obtain

$$\begin{aligned} \mathcal{R}_\varepsilon^* &\geq \inf_{\hat{\theta}^N \in \Theta_N} \sup_{\theta^N \in \Theta_N} \mathbf{E}_\theta \Big\| \sum_{j=2}^{N} (\hat{\theta}_j - \theta_j) \varphi_j \Big\|_2^2 \\ &= \inf_{\hat{\theta}^N \in \Theta_N} \sup_{\theta^N \in \Theta_N} \mathbf{E}_\theta \Big[\sum_{j=2}^{N} (\hat{\theta}_j - \theta_j)^2 \Big]. \end{aligned} \tag{3.37}$$

Step 2. From the minimax to the Bayes risk

Introduce the following probability density with respect to the Lebesgue measure on $\mathbf{R}^{N-1}$:

$$\mu(\theta^N) = \prod_{k=2}^{N} \mu_{s_k}(\theta_k), \quad \theta^N = (\theta_2, \dots, \theta_N),$$

where

$$s_k^2 = (1-\delta) v_k^2 \quad \text{with } 0 < \delta < 1$$

for v_k^2 defined by (3.25). The density μ is supported on $\mathbf{R}^{N-1}$. Now, by (3.37), we can bound the minimax risk $\mathcal{R}_\varepsilon^*$ from below by the Bayes risk, so that

$$\mathcal{R}_\varepsilon^* \geq \inf_{\hat{\theta}^N \in \Theta_N} \sum_{k=2}^{N} \int_{\Theta_N} \mathbf{E}_\theta \left[(\hat{\theta}_k - \theta_k)^2 \right] \mu(\theta^N) d\theta^N \geq I^* - r^*, \tag{3.38}$$

where the main term of the Bayes risk I^* and the residual term r^* are given by

$$I^* = \inf_{\hat\theta^N} \sum_{k=2}^{N} \int_{\mathbf{R}^{N-1}} \mathbf{E}_\theta \left[(\hat\theta_k - \theta_k)^2 \right] \mu(\theta^N) d\theta^N,$$

$$r^* = \sup_{\hat\theta^N \in \Theta_N} \sum_{k=2}^{N} \int_{\Theta_N^c} \mathbf{E}_\theta \left[(\hat\theta_k - \theta_k)^2 \right] \mu(\theta^N) d\theta^N$$

with $\Theta_N^c = \mathbf{R}^{N-1} \setminus \Theta_N$. In order to prove (3.31), it is sufficient to obtain the following lower bound for the main term of the Bayes risk:

$$I^* \geq C^* \varepsilon^{\frac{4\beta}{2\beta+1}} (1 + o(1)) \quad \text{as} \quad \varepsilon \to 0, \tag{3.39}$$

and to prove that the residual term r^* is negligible:

$$r^* = o(\varepsilon^{\frac{4\beta}{2\beta+1}}) \quad \text{as} \quad \varepsilon \to 0. \tag{3.40}$$

Indeed, (3.31) follows from (3.38)–(3.40).

Step 3. Lower bound for the main term of the Bayes risk

The main term of the Bayes risk I^* is a sum of $N-1$ terms, each of them depending on a single coordinate $\hat\theta_k$:

$$\begin{aligned} I^* &= \inf_{\hat\theta^N} \sum_{k=2}^{N} \int_{\mathbf{R}^{N-1}} \mathbf{E}_\theta \left[(\hat\theta_k - \theta_k)^2 \right] \mu(\theta^N) d\theta^N \\ &\geq \sum_{k=2}^{N} \inf_{\hat\theta_k} \int_{\mathbf{R}^{N-1}} \mathbf{E}_\theta \left[(\hat\theta_k - \theta_k)^2 \right] \mu(\theta^N) d\theta^N. \end{aligned} \tag{3.41}$$

Define $\mathbf{P}_\theta = \mathbf{P}_{f_{\theta^N}}$ and let $\mathbf{P}_f$ be the distribution of $\mathbf{X} = \{Y(t), t \in [0,1]\}$ in model (3.1). In particular, $\mathbf{P}_0$ is the distribution of $\{\varepsilon W(t), t \in [0,1]\}$, where W is a standard Wiener process. By Girsanov's theorem (Lemma A.5 in the Appendix) and by the definition of f_{θ^N}, the likelihood ratio can be written as follows:

$$\begin{aligned} \frac{d\mathbf{P}_\theta}{d\mathbf{P}_0}(\mathbf{X}) &= \exp\left(\varepsilon^{-2} \sum_{j=2}^{N} \theta_j y_j - \frac{\varepsilon^{-2}}{2} \sum_{j=2}^{N} \theta_j^2 \right) \\ &\triangleq S(y_2, \ldots, y_N, \theta^N) \end{aligned}$$

with

$$y_j = \int_0^1 \varphi_j(t) dY(t), \quad j = 2, \ldots, N.$$

Note that we can replace the infimum over arbitrary estimators $\hat{\theta}_k(\mathbf{X})$ by the infimum over estimators $\bar{\theta}_k(y_2, \ldots, y_N)$ depending only on the statistics $y_2, \ldots, y_N$. Indeed, using Jensen's inequality,

$$\begin{aligned}
&\mathbf{E}_\theta \left[(\hat{\theta}_k(\mathbf{X}) - \theta_k)^2 \right] \\
&\quad = \mathbf{E}_0 \left[\frac{d\mathbf{P}_\theta}{d\mathbf{P}_0}(\mathbf{X}) (\hat{\theta}_k(\mathbf{X}) - \theta_k)^2 \right] \\
&\quad = \mathbf{E}_0 \left[\mathbf{E}_0 \left[(\hat{\theta}_k(\mathbf{X}) - \theta_k)^2 \Big| \, y_2, \ldots, y_N \right] S(y_2, \ldots, y_N, \theta^N) \right] \\
&\quad \geq \mathbf{E}_0 \left[(\bar{\theta}_k(y_2, \ldots, y_N) - \theta_k)^2 S(y_2, \ldots, y_N, \theta^N) \right] \\
&\quad = \mathbf{E}_\theta \left[(\bar{\theta}_k(y_2, \ldots, y_N) - \theta_k)^2 \right],
\end{aligned}$$

where $\bar{\theta}_k(y_2, \ldots, y_N) = \mathbf{E}_0(\hat{\theta}_k(\mathbf{X}) | y_2, \ldots, y_N)$. Therefore

$$\begin{aligned}
&\inf_{\hat{\theta}_k} \int_{\mathbf{R}^{N-1}} \mathbf{E}_\theta \left[(\hat{\theta}_k - \theta_k)^2 \right] \mu(\theta^N) d\theta^N \qquad (3.42) \\
&\geq \inf_{\bar{\theta}_k(\cdot)} \int_{\mathbf{R}^{N-1}} \mathbf{E}_\theta \left[(\bar{\theta}_k(y_2, \ldots, y_N) - \theta_k)^2 \right] \mu(\theta^N) d\theta^N \\
&= \inf_{\bar{\theta}_k(\cdot)} \int_{\mathbf{R}^{N-1}} \int_{\mathbf{R}^{N-1}} (\bar{\theta}_k(u_2, \ldots, u_N) - \theta_k)^2 \prod_{j=2}^{N} \left[\mu_\varepsilon(u_j - \theta_j) \mu_{s_j}(\theta_j) du_j d\theta_j \right] \\
&\geq \int_{\mathbf{R}^{N-2}} \int_{\mathbf{R}^{N-2}} I_k(\{u_j\}_{j \neq k}) \prod_{j \neq k} \left[\mu_\varepsilon(u_j - \theta_j) \mu_{s_j}(\theta_j) du_j d\theta_j \right],
\end{aligned}$$

where $\inf_{\bar{\theta}_k(\cdot)}$ denotes the infimum over all the Borel functions $\bar{\theta}_k(\cdot)$ on $\mathbf{R}^{N-1}$, $\{u_j\}_{j \neq k} \stackrel{\triangle}{=} (u_2, \ldots, u_{k-1}, u_{k+1}, \ldots, u_N)$ and

$$I_k(\{u_j\}_{j \neq k}) \stackrel{\triangle}{=} \inf_{\bar{\theta}_k(\cdot)} \int_{\mathbf{R}^2} (\bar{\theta}_k(u_2, \ldots, u_N) - \theta_k)^2 \mu_\varepsilon(u_k - \theta_k) \mu_{s_k}(\theta_k) du_k d\theta_k.$$

For any fixed $\{u_j\}_{j \neq k}$ we obtain

$$\begin{aligned}
I_k(\{u_j\}_{j \neq k}) &\geq \inf_{\tilde{\theta}_k(\cdot)} \int_{\mathbf{R}^2} (\tilde{\theta}_k(u_k) - \theta_k)^2 \mu_\varepsilon(u_k - \theta_k) \mu_{s_k}(\theta_k) du_k d\theta_k \qquad (3.43) \\
&= \frac{s_k^2 \varepsilon^2}{\varepsilon^2 + s_k^2} \qquad \text{(by Lemma 3.4)},
\end{aligned}$$

where $\inf_{\tilde{\theta}_k(\cdot)}$ denotes the infimum over all Borel functions $\tilde{\theta}_k(\cdot)$ on $\mathbf{R}$. Inequality (3.41) combined with (3.42) and (3.43) implies

$$I^* \geq \sum_{k=2}^{N} \frac{s_k^2 \varepsilon^2}{\varepsilon^2 + s_k^2} = (1 - \delta) \sum_{k=2}^{N} \frac{\varepsilon^2 v_k^2}{\varepsilon^2 + (1 - \delta) v_k^2}$$

$$\geq (1-\delta)\sum_{k=2}^{N}\frac{\varepsilon^2 v_k^2}{\varepsilon^2+v_k^2} = (1-\delta)\varepsilon^2\sum_{k=2}^{\infty}(1-\kappa a_k)_+ \qquad \text{(by (3.23))}$$
$$= (1-\delta)(\mathcal{D}^*-\varepsilon^2) = (1-\delta)C^*\varepsilon^{\frac{4\beta}{2\beta+1}}(1+o(1)) \quad \text{as} \quad \varepsilon\to 0.$$

The proof is completed by making δ tend to 0.

Step 4. Negligibility of the residual term

We now prove (3.40), i.e., the fact that the residual term r^* is negligible, as compared to the main term I^* of the Bayes risk. Set $\|\theta^N\|^2 = \sum_{k=2}^{N}\theta_k^2$ and $d_N = \sup_{\theta^N\in\Theta_N}\|\theta^N\|$. We have

$$\begin{aligned} r^* &= \sup_{\hat\theta^N\in\Theta_N}\int_{\Theta_N^c}\mathbf{E}_\theta\|\hat\theta^N-\theta^N\|^2\mu(\theta^N)d\theta^N \\ &\leq 2\int_{\Theta_N^c}(d_N^2+\|\theta^N\|^2)\mu(\theta^N)d\theta^N \\ &\leq 2\left[d_N^2\mathbb{P}_\mu(\Theta_N^c)+\left(\mathbb{P}_\mu(\Theta_N^c)\mathbb{E}_\mu\|\theta^N\|^4\right)^{1/2}\right] \qquad \text{(Cauchy–Schwarz)}, \end{aligned}$$

where $\mathbb{P}_\mu$ and $\mathbb{E}_\mu$ denote the probability measure and the expectation associated with the density μ, respectively. On the other hand,

$$d_N^2 = \sup_{\theta^N\in\Theta_N}\sum_{k=2}^{N}\theta_k^2 \leq \frac{1}{a_2^2}\sup_{\theta^N\in\Theta_N}\sum_{k=2}^{N}a_k^2\theta_k^2 \leq \frac{Q}{a_2^2}.$$

Since θ_k and θ_j are independent, we have

$$\begin{aligned} \mathbb{E}_\mu\|\theta^N\|^4 &= \mathbb{E}_\mu\left[\Big(\sum_{k=2}^{N}\theta_k^2\Big)^2\right] = \sum_{k\neq j}\mathbb{E}_\mu(\theta_k^2)\mathbb{E}_\mu(\theta_j^2)+\sum_{k=2}^{N}\mathbb{E}_\mu(\theta_k^4) \\ &= \sum_{k\neq j}s_k^2s_j^2+3\sum_{k=2}^{N}s_k^4 \\ &\leq 3\left(\sum_{k=2}^{N}s_k^2\right)^2 \leq 3a_2^{-4}\left(\sum_{k=2}^{N}a_k^2s_k^2\right)^2 \leq 3a_2^{-4}Q^2, \end{aligned}$$

where the last inequality is obtained if we observe that, by the definition of s_k^2, (3.15), and (3.25), we have

$$\sum_{k=2}^{N}a_k^2s_k^2 = (1-\delta)\sum_{k=2}^{N}a_k^2v_k^2, \quad \text{and} \quad \sum_{k=2}^{N}a_k^2v_k^2 = Q. \tag{3.44}$$

The above calculations imply that

$$r^* \le 2a_2^{-2}Q\left(I\!P_\mu(\Theta_N^c) + \sqrt{3I\!P_\mu(\Theta_N^c)}\right) \le 6a_2^{-2}Q\sqrt{I\!P_\mu(\Theta_N^c)}. \tag{3.45}$$

Therefore, in order to obtain (3.40) it is sufficient to check that

$$I\!P_\mu(\Theta_N^c) = o\Big(\varepsilon^{\frac{8\beta}{2\beta+1}}\Big) \quad \text{as} \quad \varepsilon \to 0. \tag{3.46}$$

Using (3.44) and the fact that $I\!E_\mu(\theta_k^2) = s_k^2 = (1-\delta)v_k^2$ we obtain

$$\begin{aligned} I\!P_\mu(\Theta_N^c) &= I\!P_\mu\left(\sum_{k=2}^N a_k^2\theta_k^2 > Q\right) \\ &= I\!P_\mu\left(\sum_{k=2}^N a_k^2(\theta_k^2 - I\!E_\mu(\theta_k^2)) > Q - (1-\delta)\sum_{k=2}^N a_k^2 v_k^2\right) \\ &= I\!P_\mu\left(\sum_{k=2}^N a_k^2(\theta_k^2 - I\!E_\mu(\theta_k^2)) > \delta Q\right) \\ &= \mathbf{P}\left(\sum_{k=2}^N Z_k > \frac{\delta}{1-\delta}\sum_{k=2}^N b_k^2\right) \end{aligned} \tag{3.47}$$

with $b_k^2 = a_k^2 s_k^2$, $Z_k = (\xi_k^2 - 1)b_k^2$, and with the i.i.d. $\mathcal{N}(0,1)$ variables ξ_k. The last probability can be bounded from above as follows.

Lemma 3.5 *For all $0 < t < 1$ we have*

$$\mathbf{P}\left(\sum_{k=2}^N Z_k \ge t\sum_{k=2}^N b_k^2\right) \le \exp\left(-\frac{t^2\sum_{k=2}^N b_k^2}{8\max\limits_{2\le k\le N} b_k^2}\right).$$

Proof. Fix $x > 0$ and $\gamma > 0$. By the Markov inequality,

$$\mathbf{P}\left(\sum_{k=2}^N Z_k \ge x\right) \le \exp(-\gamma x)\prod_{k=2}^N \mathbf{E}\left[\exp(\gamma Z_k)\right].$$

Here

$$\begin{aligned} \mathbf{E}\left[\exp(\gamma Z_k)\right] &= \frac{1}{\sqrt{2\pi}}\int \exp\left(\gamma(\xi^2-1)b_k^2 - \frac{\xi^2}{2}\right)d\xi \\ &= \exp(-\gamma b_k^2)(1-2\gamma b_k^2)^{-1/2} \le \exp(2(\gamma b_k^2)^2) \end{aligned}$$

whenever $\gamma b_k^2 < 1/4$. Indeed, $e^{-x}(1-2x)^{-1/2} \le e^{2x^2}$ if $0 < x < 1/4$. This implies that

$$\begin{aligned} \mathbf{P}\left(\sum_{k=2}^N Z_k \ge x\right) &\le \exp\left(-\gamma x + 2\gamma^2\sum_{k=2}^N b_k^4\right) \\ &\le \exp\left(-\gamma x + 2\gamma^2\max_{2\le k\le N} b_k^2\sum_{k=2}^N b_k^2\right) \end{aligned}$$

whenever $0 < \gamma < \dfrac{1}{4 \max\limits_{2\le k\le N} b_k^2}$. The proof is completed by taking

$$x = t\sum_{k=2}^{N} b_k^2, \qquad \gamma = \frac{t}{4 \max\limits_{2\le k\le N} b_k^2}$$

with $0 < t < 1$. ∎

Using (3.47) and Lemma 3.5 we get that, for $0 < \delta < 1/2$,

$$I\!P_\mu(\Theta_N^c) \le \exp\left(-\frac{\delta^2}{8(1-\delta)^2}\frac{\sum_{k=2}^N a_k^2 s_k^2}{\max\limits_{2\le k\le N} a_k^2 s_k^2}\right). \tag{3.48}$$

In addition, from (3.44) we have $\sum_{k=2}^N a_k^2 s_k^2 = (1-\delta)Q$ and, by (3.28), $\max_{2\le k\le N} a_k^2 s_k^2 = O(\varepsilon^{\frac{2}{2\beta+1}})$. Hence, for a constant $C > 0$,

$$I\!P_\mu(\Theta_N^c) \le \exp\left(-C\varepsilon^{-\frac{2}{2\beta+1}}\right),$$

implying (3.46) and (3.40). This completes the proof of the lower bound (3.31). ∎

Remarks.

(1) The proofs of this section also yield an analog of Theorem 3.1 for the Gaussian sequence model (3.10), i.e., the following result:

$$\begin{aligned}\inf_{\hat\theta_\varepsilon}\sup_{\theta\in\Theta(\beta,Q)} \mathbf{E}_\theta\|\hat\theta_\varepsilon - \theta\|^2 &= \sup_{\theta\in\Theta(\beta,Q)} \mathbf{E}_\theta\|\hat\theta(\ell^*) - \theta\|^2 \\ &= C^*\varepsilon^{\frac{4\beta}{2\beta+1}}(1+o(1)), \quad \varepsilon\to 0,\end{aligned} \tag{3.49}$$

where the infimum is over all estimators. This result holds under the same conditions as in Theorem 3.1.

(2) Theorem 3.1 and (3.49) remain valid if we replace the weights ℓ_j^* (the simplified Pinsker weights) in the definition of the estimator by the minimax linear weights ℓ_j given by (3.20) (the Pinsker weights) with a_j as in (3.2). To check this fact it is sufficient to compare (3.29) and (3.49).

(3) In the definition of the prior density μ_k the value of δ is fixed. This is not the only possibility. We can also take $\delta = \delta_\varepsilon$ depending on ε and converging to 0 slowly enough as $\varepsilon \to 0$, for example, $\delta_\varepsilon = (\log 1/\varepsilon)^{-1}$. It is easy to see that in this case (3.48) still implies (3.46).

(4) An argument similar to that in the proof of Lemma 3.5 shows that

$$I\!P_\mu\left((1-2\delta)Q \le \sum_{k=2}^N a_k^2\theta_k^2 \le Q\right) \to 1 \quad \text{as} \quad \varepsilon\to 0$$

at an exponential rate. Similarly to (3.46), this relation remains valid if $\delta = \delta_\varepsilon$ depends on ε and converges to 0 slowly enough as $\varepsilon \to 0$. This means that almost all the mass of the prior distribution $I\!P_\mu$ is concentrated in a small (asymptotically shrinking) neighborhood of the boundary $\{\theta : \sum_k a_k^2 \theta_k^2 = Q\}$ of the ellipsoid $\Theta(\beta, Q)$. The values θ in this neighborhood can be viewed as being the least favorable, i.e., the hardest to estimate. Since the neighborhood depends on ε, the least favorable values θ are different for different ε. Even more, one can show that there exist no fixed (that is, independent of ε) θ^* belonging to the ellipsoid $\Theta(\beta, Q)$ and such that

$$\mathbf{E}_{\theta^*} \|\hat{\theta}(\ell^*) - \theta^*\|^2 = C^* \varepsilon^{\frac{4\beta}{2\beta+1}} (1 + o(1)), \quad \varepsilon \to 0.$$

We will come back to this property in Section 3.8.

3.4 Stein's phenomenon

In this section we temporarily switch to the parametric Gaussian models, and discuss some notions related to Stein's phenomenon. This material plays an auxiliary role. It will be helpful for further constructions in the chapter. Consider the following two Gaussian models.

Model 1

This is a truncated version of the Gaussian sequence model:

$$y_j = \theta_j + \varepsilon \xi_j, \quad j = 1, \dots, d,$$

where $\varepsilon > 0$ and ξ_j are i.i.d. $\mathcal{N}(0, 1)$ random variables. In this section we will denote by y, θ, and ξ the following d-dimensional vectors:

$$y = (y_1, \dots, y_d), \quad \theta = (\theta_1, \dots, \theta_d), \quad \xi = (\xi_1, \dots, \xi_d) \sim \mathcal{N}_d(0, I),$$

where $\mathcal{N}_d(0, I)$ stands for the standard d-dimensional normal distribution. Then we can write

$$y = \theta + \varepsilon \xi, \quad \xi \sim \mathcal{N}_d(0, I). \tag{3.50}$$

The statistical problem is to estimate the unknown parameter $\theta \in \mathbf{R}^d$.

Model 2

We observe random vectors $X_1, \dots, X_n$ satisfying

$$X_i = \theta + \eta_i, \ i = 1, \dots, n,$$

with $\theta \in \mathbf{R}^d$ where η_i are i.i.d. Gaussian vectors with distribution $\mathcal{N}_d(0, I)$. The statistical problem is to estimate θ. The vector $\bar{X} = n^{-1} \sum_{i=1}^n X_i$ is a sufficient statistic in this model. We can write

$$\bar{X} = \theta + \frac{1}{\sqrt{n}}\,\xi = \theta + \varepsilon\xi$$

with

$$\varepsilon = \frac{1}{\sqrt{n}} \quad \text{and} \quad \xi = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \eta_i \sim \mathcal{N}_d(0, I).$$

Throughout this section $\mathbf{E}_\theta$ will denote the expectation with respect to the distribution y in Model 1 or with respect to the distribution of $\bar{X}$ in Model 2, and $\|\cdot\|$ will denote the Euclidean norm in $\mathbf{R}^d$. In what follows, we will write $\|\theta\|$ to denote either the $\ell^2(\mathbf{N})$-norm or the Euclidean norm on $\mathbf{R}^d$ of the vector θ according to whether $\theta \in \ell^2(\mathbf{N})$ or $\theta \in \mathbf{R}^d$.

Model 1 with $\varepsilon = 1/\sqrt{n}$ is equivalent to Model 2 in the following sense: for any Borel function $\hat{\theta} : \mathbf{R}^d \to \mathbf{R}^d$ the squared risk $\mathbf{E}_\theta\|\hat{\theta}(y) - \theta\|^2$ of the estimator $\hat{\theta}(y)$ in Model 1 with $\varepsilon = 1/\sqrt{n}$ is equal to the risk $\mathbf{E}_\theta\|\hat{\theta}(\bar{X}) - \theta\|^2$ of the estimator $\hat{\theta}(\bar{X})$ in Model 2.

Model 1 is a useful building block in the context of nonparametric estimation, as we will see later. On the other hand, Model 2 is classical for parametric statistics. In this section proofs of the results are only given for Model 1. In view of the equivalence, analogous results for Model 2 are obtained as an immediate by-product.

Definition 3.2 *An estimator θ^* of the parameter θ is called* **inadmissible** *on $\Theta \subseteq \mathbf{R}^d$ with respect to the squared risk if there exists another estimator $\hat{\theta}$ such that*

$$\mathbf{E}_\theta\|\hat{\theta} - \theta\|^2 \le \mathbf{E}_\theta\|\theta^* - \theta\|^2 \quad \textit{for all } \theta \in \Theta,$$

and there exists $\theta_0 \in \Theta$ such that

$$\mathbf{E}_{\theta_0}\|\hat{\theta} - \theta_0\|^2 < \mathbf{E}_{\theta_0}\|\theta^* - \theta_0\|^2.$$

Otherwise, the estimator θ^* is called **admissible**.

The squared risk of the estimator $\bar{X}$ in Model 2 is given by

$$\mathbf{E}_\theta\|\bar{X} - \theta\|^2 = \frac{d}{n} = d\varepsilon^2, \quad \forall\ \theta \in \mathbf{R}^d.$$

This risk is therefore constant as a function of θ.

Stein (1956) considered Model 2 and showed that if $d \ge 3$, then the estimator $\bar{X}$ is inadmissible. This property is known as *Stein's phenomenon*. Moreover, Stein proposed an estimator whose risk is smaller than that of $\bar{X}$ everywhere on $\mathbf{R}^d$ if $d \ge 3$. This improved estimator is based on a shrinkage of $\bar{X}$ towards the origin with a shrinkage factor that depends on $\|\bar{X}\|$.

3.4.1 Stein's shrinkage and the James–Stein estimator

We now explain the idea of Stein's shrinkage for Model 1. The argument for Model 2 is analogous and we omit it. We start with two preliminary lemmas.

Lemma 3.6 (Stein's lemma). *Suppose that a function $f : \mathbf{R}^d \to \mathbf{R}$ satisfies:*

(i) $f(u_1, \dots, u_d)$ is absolutely continuous in each coordinate u_i for almost all values (with respect to the Lebesgue measure on $\mathbf{R}^{d-1}$) of other coordinates $(u_j, j \neq i)$,

(ii)

$$\mathbf{E}_\theta \left| \frac{\partial f(y)}{\partial y_i} \right| < \infty, \qquad i = 1, \dots, d.$$

Then

$$\mathbf{E}_\theta \left[(\theta_i - y_i) f(y) \right] = -\varepsilon^2 \mathbf{E}_\theta \left[\frac{\partial f}{\partial y_i}(y) \right], \qquad i = 1, \dots, d.$$

Proof. We will basically use integration by parts with a slight modification due to the fact that the function f is not differentiable in the standard sense.

Observe first that it is sufficient to prove the lemma for $\theta = 0$ and $\varepsilon = 1$. Indeed, the random vector $\zeta = \varepsilon^{-1}(y - \theta)$ has distribution $\mathcal{N}_d(0, I)$. Hence, for $\tilde{f}(y) = f(\varepsilon y + \theta)$ we have

$$\mathbf{E}_\theta \left[\varepsilon^{-1} (\theta_i - y_i) f(y) \right] = -\mathbf{E} \left[\zeta_i \tilde{f}(\zeta) \right], \qquad \mathbf{E} \left[\frac{\partial f}{\partial \zeta_i}(\zeta) \right] = \varepsilon \mathbf{E} \left[\frac{\partial \tilde{f}}{\partial \zeta_i}(\zeta) \right],$$

where $\zeta_1, \dots, \zeta_d$ are the coordinates of ζ. It is clear that f satisfies assumption (ii) of the lemma if and only if $\tilde{f}$ satisfies the inequality

$$\mathbf{E} \left| \frac{\partial \tilde{f}(\zeta)}{\partial \zeta_i} \right| < \infty, \qquad i = 1, \dots, d, \tag{3.51}$$

where $\zeta \sim \mathcal{N}_d(0, I)$. Therefore it is sufficient to prove that for any function $\tilde{f}$ satisfying (3.51) and assumption (i) of the lemma we have

$$\mathbf{E}[\zeta_i \tilde{f}(\zeta)] = \mathbf{E} \left[\frac{\partial \tilde{f}}{\partial \zeta_i}(\zeta) \right], \qquad i = 1, \dots, d. \tag{3.52}$$

Without loss of generality, it is enough to prove (3.52) for $i = 1$ only. To do this, it suffices to show that, almost surely,

$$\mathbf{E} \left[\zeta_1 \tilde{f}(\zeta) | \zeta_2, \dots, \zeta_d \right] = \mathbf{E} \left[\frac{\partial \tilde{f}}{\partial \zeta_1}(\zeta) \Big| \ \zeta_2, \dots, \zeta_d \right]. \tag{3.53}$$

Since the variables ζ_j are mutually independent with distribution $\mathcal{N}(0, 1)$, equality (3.53) will be proved if we show that for almost all $\zeta_2, \dots, \zeta_d$ with respect to the Lebesgue measure on $\mathbf{R}^{d-1}$ we have

$$\int_{-\infty}^{\infty} u\tilde{f}(u,\zeta_2,\dots,\zeta_d)e^{-u^2/2}du = \int_{-\infty}^{\infty} \frac{\partial \tilde{f}}{\partial u}(u,\zeta_2,\dots,\zeta_d)e^{-u^2/2}du.$$

Put $h(u) = \tilde{f}(u,\zeta_2,\dots,\zeta_d)$. In order to complete the proof, it remains to show that for any absolutely continuous function $h : \mathbf{R} \to \mathbf{R}$ such that

$$\int_{-\infty}^{\infty} |h'(u)|e^{-u^2/2}du < \infty,$$

we have

$$\int_{-\infty}^{\infty} uh(u)e^{-u^2/2}du = \int_{-\infty}^{\infty} h'(u)e^{-u^2/2}du. \tag{3.54}$$

To show (3.54) note first that

$$e^{-u^2/2} = \begin{cases} \int_u^{\infty} ze^{-z^2/2}dz, & \text{if } u > 0, \\ -\int_{-\infty}^{u} ze^{-z^2/2}dz, & \text{if } u < 0. \end{cases}$$

Therefore,

$$\begin{aligned} \int_{-\infty}^{\infty} h'(u)e^{-u^2/2}du &= \int_0^{\infty} h'(u)\Big[\int_u^{\infty} ze^{-z^2/2}dz\Big]du \\ &\quad - \int_{-\infty}^{0} h'(u)\Big[\int_{-\infty}^{u} ze^{-z^2/2}dz\Big]du \\ &= \int_0^{\infty} ze^{-z^2/2}\Big[\int_0^{z} h'(u)du\Big]dz \\ &\quad - \int_{-\infty}^{0} ze^{-z^2/2}\Big[\int_z^{0} h'(u)du\Big]dz \\ &= \left(\int_0^{\infty} + \int_{-\infty}^{0}\right)\{ze^{-z^2/2}[h(z) - h(0)]\}dz \\ &= \int_{-\infty}^{\infty} zh(z)e^{-z^2/2}dz \end{aligned}$$

implying (3.54). ∎

Lemma 3.7 *Let $d \geq 3$. Then, for all $\theta \in \mathbf{R}^d$,*

$$0 < \mathbf{E}_\theta\left(\frac{1}{\|y\|^2}\right) < \infty.$$

Proof. By (3.50), we have

$$\mathbf{E}_\theta\left(\frac{1}{\|y\|^2}\right) = \frac{1}{\varepsilon^2}\mathbf{E}\left(\frac{1}{\|\varepsilon^{-1}\theta + \xi\|^2}\right),$$

where $\xi \sim \mathcal{N}_d(0, I)$ is a standard Gaussian d-dimensional vector. Since the distribution $\mathcal{N}_d(0, I)$ is spherically symmetric,

$$\forall\ v, v' \in \mathbf{R}^d :\ \|v\| = \|v'\| \implies \|\xi + v\| \stackrel{\mathcal{D}}{=} \|\xi + v'\|, \tag{3.55}$$

where $\stackrel{\mathcal{D}}{=}$ denotes equality in distribution. Indeed, since the norms of v and v' are equal, there exists an orthogonal matrix Γ such that $v' = \Gamma v$. Since $\Gamma\xi \stackrel{\mathcal{D}}{=} \xi$, we obtain (3.55). In particular,

$$\mathbf{E}\left(\frac{1}{\|\varepsilon^{-1}\theta + \xi\|^2}\right) = \mathbf{E}\left(\frac{1}{\|v_0 + \xi\|^2}\right)$$

with $v_0 = (\|\theta\|/\varepsilon, 0, \ldots, 0)$. On the other hand,

$$\begin{aligned}\mathbf{E}\left(\frac{1}{\|v_0 + \xi\|^2}\right) &= \frac{1}{(\sqrt{2\pi})^d}\int_{\mathbf{R}^d} \exp\left(-\frac{\|x\|^2}{2}\right)\|v_0 + x\|^{-2}dx \\ &= \frac{1}{(\sqrt{2\pi})^d}\exp\left(-\frac{\|\theta\|^2}{2\varepsilon^2}\right)\ \times \\ &\qquad \int_{\mathbf{R}^d}\exp\left(\frac{u_1\|\theta\|}{\varepsilon} - \frac{\|u\|^2}{2}\right)\|u\|^{-2}du\end{aligned}$$

with $u = (u_1, \ldots, u_d)$. Since $xy \le 3x^2 + y^2/3$ for $x \ge 0, y \ge 0$, we have $|u_1|\|\theta\|/\varepsilon \le 3\|\theta\|^2/\varepsilon^2 + \|u\|^2/3$. Then

$$\mathbf{E}\left(\frac{1}{\|v_0 + \xi\|^2}\right) \le \frac{1}{(\sqrt{2\pi})^d}\exp\left(\frac{5\|\theta\|^2}{2\varepsilon^2}\right)\int_{\mathbf{R}^d}\exp\left(-\frac{\|u\|^2}{6}\right)\|u\|^{-2}du.$$

We complete the proof by observing that if $d \ge 3$, there exists a constant $C > 0$ such that

$$\int_{\mathbf{R}^d}\exp\left(-\frac{\|u\|^2}{6}\right)\|u\|^{-2}du = C\int_0^\infty e^{-r^2/6}r^{d-3}dr < \infty. \qquad \blacksquare$$

Stein introduced the class of estimators of the form

$$\hat{\theta} = g(y)y, \tag{3.56}$$

where $g : \mathbf{R}^d \to \mathbf{R}$ is a function to be chosen. The coordinates of the vector $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_d)$ have the form

$$\hat{\theta}_j = g(y)y_j.$$

On the other hand, the random vector y is a natural estimator of θ, similar to the arithmetic mean $\bar{X}$ in Model 2. The risk of this estimator equals

$$\mathbf{E}_\theta\|y - \theta\|^2 = d\varepsilon^2.$$

Let us look for a function g such that the risk of the estimator $\hat{\theta} = g(y)y$ is smaller than that of y. We have

$$\begin{aligned}\mathbf{E}_\theta\|\hat{\theta}-\theta\|^2 &= \sum_{i=1}^{d}\mathbf{E}_\theta\left[(g(y)y_i-\theta_i)^2\right]\\ &= \sum_{i=1}^{d}\Big\{\mathbf{E}_\theta\left[(y_i-\theta_i)^2\right]+2\mathbf{E}_\theta\Big[(\theta_i-y_i)(1-g(y))y_i\Big]\\ &\qquad +\mathbf{E}_\theta\Big[y_i^2(1-g(y))^2\Big]\Big\}.\end{aligned}$$

Suppose now that the function g is such that the assumptions of Lemma 3.6 hold for the functions $f = f_i$ where $f_i(y) = (1-g(y))y_i$, $i = 1,\dots,d$. Then

$$\mathbf{E}_\theta[(\theta_i-y_i)(1-g(y))y_i] = -\varepsilon^2\mathbf{E}_\theta\left[1-g(y)-y_i\frac{\partial g}{\partial y_i}(y)\right],$$

and

$$\mathbf{E}_\theta\left[(\hat{\theta}_i-\theta_i)^2\right] = \varepsilon^2 - 2\varepsilon^2\mathbf{E}_\theta\left[1-g(y)-y_i\frac{\partial g}{\partial y_i}(y)\right]+\mathbf{E}_\theta\Big[y_i^2(1-g(y))^2\Big].$$

Summing over i gives

$$\mathbf{E}_\theta\|\hat{\theta}-\theta\|^2 = d\varepsilon^2+\mathbf{E}_\theta[W(y)] \tag{3.57}$$

with

$$W(y) = -2\varepsilon^2 d(1-g(y))+2\varepsilon^2\sum_{i=1}^{d}y_i\frac{\partial g}{\partial y_i}(y)+\|y\|^2(1-g(y))^2.$$

The above argument is summarized in the following way.

Lemma 3.8 (Stein's unbiased risk estimator). *Consider Model 1 with $d \geq 3$ and the estimator $\hat{\theta}$ defined in (3.56). Let the assumptions of Lemma 3.6 be fulfilled for the functions $f = f_i$ where $f_i(y) = (1-g(y))y_i$, $i = 1,\dots,d$. Then an unbiased estimator of the risk $\mathbf{E}_\theta\|\hat{\theta}-\theta\|^2$ is given by the formula*

$$\mathrm{SURE} = \varepsilon^2 d(2g(y)-1)+2\varepsilon^2\sum_{i=1}^{d}y_i\frac{\partial g}{\partial y_i}(y)+\|y\|^2(1-g(y))^2.$$

Here SURE stands for *Stein's unbiased risk estimator*. Note that the result of Lemma 3.8 is of the same type as those obtained in Section 1.4 for unbiased estimators of the risk of kernel density estimators.

The risk of $\hat{\theta}$ is smaller than that of y if we choose g such that

$$\mathbf{E}_\theta[W(y)] < 0.$$

In order to satisfy this inequality, Stein suggested to search for g among the functions of the form

$$g(y) = 1 - \frac{c}{\|y\|^2}$$

with an appropriately chosen constant $c > 0$. If g has this form, the functions f_i defined by $f_i(y) = (1 - g(y))y_i$ satisfy the assumptions of Lemma 3.6, and (3.57) holds with

$$\begin{aligned} W(y) &= -2\varepsilon^2 d\frac{c}{\|y\|^2} + 2\varepsilon^2 \sum_{i=1}^{d} y_i^2 \frac{2c}{\|y\|^4} + \frac{c^2}{\|y\|^2} \\ &= \frac{1}{\|y\|^2}\Big(-2dc\varepsilon^2 + 4\varepsilon^2 c + c^2\Big). \end{aligned} \tag{3.58}$$

The minimizer in c of (3.58) is equal to

$$c_{opt} = \varepsilon^2(d-2).$$

The function g and the estimator $\hat{\theta} = g(y)y$ associated to this choice of g are given by

$$g(y) = 1 - \frac{\varepsilon^2(d-2)}{\|y\|^2},$$

and

$$\hat{\theta}_{JS} = \left(1 - \frac{\varepsilon^2(d-2)}{\|y\|^2}\right) y, \tag{3.59}$$

respectively. The statistic $\hat{\theta}_{JS}$ is called the *James–Stein estimator* of θ. If the norm $\|y\|$ is sufficiently large, multiplication of y by $g(y)$ shrinks the value of y to 0. This is called the *Stein shrinkage*. If $c = c_{opt}$, then

$$W(y) = -\frac{\varepsilon^4(d-2)^2}{\|y\|^2}. \tag{3.60}$$

For this function W, Lemma 3.7 implies $-\infty < \mathbf{E}_\theta[W(y)] < 0$, provided that $d \geq 3$. Therefore, if $d \geq 3$, the risk of the James–Stein estimator satisfies

$$\mathbf{E}_\theta\|\hat{\theta}_{JS} - \theta\|^2 = d\varepsilon^2 - \mathbf{E}_\theta\left(\frac{\varepsilon^4(d-2)^2}{\|y\|^2}\right) < \mathbf{E}_\theta\|y - \theta\|^2$$

for all $\theta \in \mathbf{R}^d$.

Conclusion: If $d \geq 3$, the James–Stein estimator $\hat{\theta}_{JS}$ (which is biased) is better than the (unbiased) estimator y for all $\theta \in \mathbf{R}^d$ and therefore the estimator y is not admissible in Model 1.

The James–Stein estimator for Model 2 is obtained in a similar way; we just need to replace y by $\bar{X}$ and ε by $1/\sqrt{n}$ in (3.59):

$$\hat{\theta}_{JS} = \left(1 - \frac{d-2}{n\|\bar{X}\|^2}\right)\bar{X}. \tag{3.61}$$

Since Models 1 and 2 are equivalent, (3.61) is better than the estimator $\bar{X}$ for all $\theta \in \mathbf{R}^d$ when $d \geq 3$. Therefore we have proved the following result.

Theorem 3.3 (Stein's phenomenon). *Let $d \geq 3$. Then the estimator $\hat{\theta} = y$ is inadmissible on $\mathbf{R}^d$ in Model 1 and the estimator $\hat{\theta} = \bar{X}$ is inadmissible on $\mathbf{R}^d$ in Model 2.*

It is interesting to analyze the improvement given by $\hat{\theta}_{JS}$ with respect to y. For $\theta = 0$ the risk of the James–Stein estimator is

$$\mathbf{E}_0\|\hat{\theta}_{JS}\|^2 = d\varepsilon^2 - \varepsilon^4(d-2)^2\mathbf{E}\left(\frac{1}{\|\varepsilon\xi\|^2}\right) = 2\varepsilon^2,$$

since $\mathbf{E}\left(\|\xi\|^{-2}\right) = 1/(d-2)$ (check this as an exercise). Therefore, for $\theta = 0$ the improvement is characterized by the ratio

$$\frac{\mathbf{E}_0\|\hat{\theta}_{JS}\|^2}{\mathbf{E}_0\|y\|^2} = \frac{2}{d}, \tag{3.62}$$

which is a constant independent of ε. On the contrary, for all $\theta \neq 0$ the ratio of the squared risks of $\hat{\theta}_{JS}$ and y tends to 1 as $\varepsilon \to 0$ (cf. Lehmann and Casella (1998), p. 407) making the improvement asymptotically negligible.

3.4.2 Other shrinkage estimators

It follows from (3.58) that there exists a whole family of estimators that are better than y in Model 1 when the dimension d is large enough: It is sufficient to take the constant c in the definition of g so that $-2dc\varepsilon^2 + 4\varepsilon^2 c + c^2 < 0$. For example, if $c = \varepsilon^2 d$, we obtain the *Stein estimator* :

$$\hat{\theta}_S \triangleq \left(1 - \frac{\varepsilon^2 d}{\|y\|^2}\right) y.$$

This estimator is better than y for $d \geq 5$. However, it is worse than $\hat{\theta}_{JS}$ for $d \geq 3$.

Estimators performing even better correspond to nonnegative functions g:

$$g(y) = \left(1 - \frac{c}{\|y\|^2}\right)_+$$

with $c > 0$. For example, taking here $c = \varepsilon^2(d-2)$ and $c = \varepsilon^2 d$ we obtain the *positive part James–Stein estimator* and the *positive part Stein estimator*:

$$\hat{\theta}_{JS+} = \left(1 - \frac{\varepsilon^2(d-2)}{\|y\|^2}\right)_+ y,$$

and

$$\hat{\theta}_{S+} = \left(1 - \frac{\varepsilon^2 d}{\|y\|^2}\right)_+ y$$

respectively.

Lemma 3.9 *For all $d \geq 1$ and all $\theta \in \mathbf{R}^d$,*

$$\mathbf{E}_\theta \|\hat{\theta}_{JS+} - \theta\|^2 < \mathbf{E}_\theta \|\hat{\theta}_{JS} - \theta\|^2, \qquad \mathbf{E}_\theta \|\hat{\theta}_{S+} - \theta\|^2 < \mathbf{E}_\theta \|\hat{\theta}_S - \theta\|^2.$$

A proof of this lemma is given in the Appendix (Lemma A.6).

Thus, the estimators $\hat{\theta}_{JS+}$ and $\hat{\theta}_{S+}$ are better than $\hat{\theta}_{JS}$ and $\hat{\theta}_S$, respectively. Though the four estimators are better than y, they are all inadmissible (since $\hat{\theta}_{JS+}$ and $\hat{\theta}_{S+}$ are inadmissible; see, for example, Lehmann and Casella (1998), p. 357). However, it can be shown that the estimator $\hat{\theta}_{JS+}$ can be improved in the smaller order terms only, so that it is "quite close" to being admissible. We mention also that there exists an admissible estimator of θ, though its construction is more cumbersome than that of $\hat{\theta}_{JS+}$.

Lemma 3.10 *Let $\theta \in \mathbf{R}^d$. For all $d \geq 4$,*

$$\mathbf{E}_\theta \|\hat{\theta}_S - \theta\|^2 \leq \frac{d\varepsilon^2 \|\theta\|^2}{\|\theta\|^2 + d\varepsilon^2} + 4\varepsilon^2 \tag{3.63}$$

and, for all $d \geq 1$,

$$\mathbf{E}_\theta \|\hat{\theta}_{S+} - \theta\|^2 \leq \frac{d\varepsilon^2 \|\theta\|^2}{\|\theta\|^2 + d\varepsilon^2} + 4\varepsilon^2. \tag{3.64}$$

Proof. We first prove (3.63). From (3.57) and (3.58) with $c = \varepsilon^2 d$ we obtain

$$\begin{aligned}\mathbf{E}_\theta \|\hat{\theta}_S - \theta\|^2 &= d\varepsilon^2 + (-2dc\varepsilon^2 + 4\varepsilon^2 c + c^2)\mathbf{E}_\theta \left(\frac{1}{\|y\|^2}\right) \\ &= d\varepsilon^2 - (d^2 - 4d)\varepsilon^4 \mathbf{E}_\theta \left(\frac{1}{\|y\|^2}\right).\end{aligned}$$

By Jensen's inequality,

$$\mathbf{E}_\theta \left(\frac{1}{\|y\|^2}\right) \geq \frac{1}{\mathbf{E}_\theta \|y\|^2} = \frac{1}{\|\theta\|^2 + \varepsilon^2 d}\,.$$

Therefore

$$\mathbf{E}_\theta \|\hat{\theta}_S - \theta\|^2 \leq d\varepsilon^2 - \frac{\varepsilon^4 d(d-4)}{\|\theta\|^2 + \varepsilon^2 d} = \frac{d\varepsilon^2 \|\theta\|^2}{\|\theta\|^2 + \varepsilon^2 d} + \frac{4\varepsilon^4 d}{\|\theta\|^2 + \varepsilon^2 d}$$

implying (3.63).

We now prove (3.64). By Lemma 3.9 and (3.63), it is sufficient to show (3.64) for $d \le 3$. Observe that the function $f(y) = (1 - g(y))y_i$ satisfies the assumptions of Lemma 3.6 if $g(y) = (1 - \varepsilon^2 d/\|y\|^2)_+$. In particular,

$$\frac{\partial g(y)}{\partial y_i} = \frac{2\varepsilon^2 d y_i}{\|y\|^4} I(\|y\|^2 > \varepsilon^2 d).$$

Hence, by formula (3.57),

$$\mathbf{E}_\theta \|\hat\theta_{S+} - \theta\|^2 = d\varepsilon^2 + \mathbf{E}_\theta[W(y)],$$

where

$$\begin{aligned} W(y) &= \Big(\|y\|^2 - 2\varepsilon^2 d\Big) I(\|y\|^2 \le \varepsilon^2 d) + \frac{\varepsilon^4 d(4-d)}{\|y\|^2} I(\|y\|^2 > \varepsilon^2 d) \\ &\le \frac{\varepsilon^4 d(4-d)}{\|y\|^2} I(\|y\|^2 > \varepsilon^2 d). \end{aligned}$$

If $d \le 3$, the last expression is less than or equal to $\varepsilon^2(4-d)$. Therefore, for $d \le 3$,

$$\mathbf{E}_\theta \|\hat\theta_{S+} - \theta\|^2 \le 4\varepsilon^2,$$

implying (3.64). ∎

Two other important types of shrinkage are hard and soft thresholding. If we choose the shrinkage factor in the form $g(y) = I(\|y\| > \tau)$ with some $\tau > 0$, we obtain the *global hard thresholding estimator* of θ in Model 1:

$$\hat\theta_{GHT} = I(\|y\| > \tau) y.$$

At first sight, this thresholding seems very rough: We either keep or "kill" all the observations. Nevertheless, some important properties of the Stein shrinkage are preserved. In particular, if $\tau = c\varepsilon\sqrt{d}$ for a suitably chosen absolute constant $c > 0$, a result similar to Lemma 3.10 remains valid for $\hat\theta_{GHT}$, though with coarser constants (cf. Exercise 3.7). Analogous properties can be proved for the *global soft thresholding estimator*

$$\hat\theta_{GST} = \left(1 - \frac{\tau}{\|y\|}\right)_+ y.$$

One can also consider coordinate-wise rather than global shrinkage of y. The main examples are: the *hard thresholding estimator* $\hat\theta_{HT}$ whose components are equal to

$$\hat\theta_{j,HT} = I(|y_j| > \tilde\tau) y_j\,;$$

the *soft thresholding estimator* $\hat\theta_{ST}$ with the components

$$\hat\theta_{j,ST} = \operatorname{sign}(y_j)\big(|y_j| - \tilde\tau\big)_+ = \left(1 - \frac{\tilde\tau}{|y_j|}\right)_+ y_j\,;$$

and the *nonnegative garotte estimator* $\hat{\theta}_G$ with the components

$$\hat{\theta}_{j,G} = \left(1 - \frac{\tilde{\tau}^2}{y_j^2}\right)_+ y_j \,.$$

Here $\tilde{\tau} > 0$ is a threshold, which usually has the form $\tilde{\tau} = c\varepsilon\sqrt{\log(1/\varepsilon)}$, for a suitable absolute constant $c > 0$.

In either case, the coordinate-wise shrinkage keeps large observations (perhaps, slightly transforming them) and sets others equal to 0. Note that the nonnegative garotte is a particular case of the positive part Stein shrinkage corresponding to $d = 1$.

Finally, the coordinate-wise *linear shrinkage* is equivalent to the Tikhonov regularization:

$$\hat{\theta}_j^{TR} = \frac{y_j}{1 + b_j}$$

where $b_j > 0$ (cf. Section 1.7.3).

3.4.3 Superefficiency

The estimator $\bar{X}$ is asymptotically efficient on $(\mathbf{R}^d, \|\cdot\|)$ in Model 2 in the sense of Definition 2.2 and the estimator y is asymptotically efficient on $(\mathbf{R}^d, \|\cdot\|)$ in Model 1 for $\varepsilon = 1/\sqrt{n}$. In fact, these estimators are not only asymptotically efficient, but also minimax in the nonasymptotic sense for all fixed n (or ε) (cf. Lehmann and Casella (1998), p. 350). In particular, the minimax risk associated to Model 1 is equal to the maximal risk of y:

$$\inf_{\hat{\theta}_\varepsilon} \sup_{\theta \in \mathbf{R}^d} \mathbf{E}_\theta \|\hat{\theta}_\varepsilon - \theta\|^2 = \sup_{\theta \in \mathbf{R}^d} \mathbf{E}_\theta \|y - \theta\|^2 = d\varepsilon^2,$$

where the infimum is over all estimators. So, the maximal risk of any asymptotically efficient estimator in Model 1 is $d\varepsilon^2(1 + o(1))$ as $\varepsilon \to 0$. Estimators with smaller asymptotic risk can be called superefficient. More precisely, the following definition is used.

Definition 3.3 *We say that an estimator θ_ε^* is* **superefficient** *in Model 1 if*

$$\limsup_{\varepsilon \to 0} \frac{\mathbf{E}_\theta \|\theta_\varepsilon^* - \theta\|^2}{d\varepsilon^2} \le 1, \quad \forall\ \theta \in \mathbf{R}^d, \tag{3.65}$$

and if there exists $\theta = \overline{\theta} \in \mathbf{R}^d$ such that the inequality in (3.65) is strict. The points $\overline{\theta}$ satisfying the strict inequality are called **superefficiency points** *of θ_ε^*.*

The remarks after Theorem 3.3 imply that $\hat{\theta}_{JS}$ is superefficient with the only superefficiency point $\overline{\theta} = 0$ for $d \ge 3$. In a similar way, it can be shown that $\hat{\theta}_S$ is superefficient if $d \ge 5$. Using Lemma 3.9 and the remarks preceding it we obtain the following result.

Proposition 3.1 *The estimators $\hat{\theta}_{JS}$ and $\hat{\theta}_{JS+}$ are superefficient in Model 1 if $d \geq 3$. The estimators $\hat{\theta}_S$ and $\hat{\theta}_{S+}$ are superefficient in Model 1 if $d \geq 5$.*

Note that the concept of superefficiency is in some sense weaker than that of admissibility since superefficiency is an asymptotic property. However, there is no general relation between superefficiency and admissibility. For example, the estimators mentioned in Proposition 3.1 are not admissible; they are, however, superefficient. On the other hand, in dimension $d = 1$ the estimator y is admissible (see Lehmann and Casella (1998), p. 324) but it is not superefficient.

Observe also that superefficiency is not a consequence of Stein's phenomenon. Indeed, in dimension $d = 1$ the Stein phenomenon does not occur, but there exist superefficient estimators like the Hodges estimator (see, for example, Ibragimov and Has'minskii (1981), p. 91).

Le Cam (1953) proved that for any finite d (i.e., in the parametric case) the set of superefficiency points of an estimator has necessarily the Lebesgue measure zero. Therefore, roughly speaking, the superefficiency phenomenon is negligible when the model is parametric. We will see in Section 3.8 that the situation becomes completely different in nonparametric models: For the Gaussian sequence model (where $d = \infty$) there exist estimators which are superefficient everywhere on a "massive" set like, for example, the ellipsoid $\Theta(\beta, Q)$.

3.5 Unbiased estimation of the risk

We now return to the Gaussian sequence model

$$y_j = \theta_j + \varepsilon\xi_j, \quad j = 1, 2, \dots .$$

A linear estimator of the sequence $\theta = (\theta_1, \theta_2, \dots)$ is an estimator of the form

$$\hat{\theta}(\lambda) = (\hat{\theta}_1, \hat{\theta}_2, \dots) \quad \text{with} \quad \hat{\theta}_j = \lambda_j y_j,$$

where $\lambda = \{\lambda_j\}_{j=1}^{\infty} \in \ell^2(\mathbf{N})$ is a sequence of weights. The mean squared risk of $\hat{\theta}(\lambda)$ is

$$R(\lambda, \theta) = \mathbf{E}_\theta \|\hat{\theta}(\lambda) - \theta\|^2 = \sum_{j=1}^{\infty} \left[(1 - \lambda_j)^2 \theta_j^2 + \varepsilon^2 \lambda_j^2\right]$$

(cf. (1.112)). How to choose the sequence of weights λ in an optimal way? Suppose that λ belongs to a class of sequences Λ such that $\Lambda \subseteq \ell^2(\mathbf{N})$. Some examples of classes Λ that are interesting in the context of nonparametric estimation will be given below. A mean square optimal on Λ sequence λ is a solution of the following minimization problem:

$$\lambda^{oracle}(\Lambda, \theta) = \arg\min_{\lambda \in \Lambda} R(\lambda, \theta)$$

if such a solution exists. The mapping $\theta \mapsto \hat{\theta}(\lambda^{oracle}(\Lambda, \theta))$ is an oracle in the sense of Definition 1.13. It can be called the *linear oracle with weights in the class* Λ. Since the underlying θ is unknown, the oracle value $\hat{\theta}(\lambda^{oracle}(\Lambda, \theta))$ is not an estimator. When no ambiguity is caused, we will also attribute the name "oracle" to the sequence of weights $\lambda^{oracle}(\Lambda, \theta)$.

An important question in this context is the following: Can we construct an estimator whose risk would converge to the risk of the oracle, i.e., to $\min_{\lambda \in \Lambda} R(\lambda, \theta)$, as $\varepsilon \to 0$?

A general way to answer this question is based on the idea of *unbiased estimation of the risk* that was already discussed in Chapter 1. To develop this idea for our framework observe first that

$$\|\hat{\theta}(\lambda) - \theta\|^2 = \sum_j (\lambda_j^2 y_j^2 - 2\lambda_j y_j \theta_j + \theta_j^2)$$

for λ, θ, y such that the sum on right hand side is finite. Put

$$\mathcal{J}(\lambda) \stackrel{\triangle}{=} \sum_j (\lambda_j^2 y_j^2 - 2\lambda_j (y_j^2 - \varepsilon^2)).$$

Then

$$\mathbf{E}_\theta[\mathcal{J}(\lambda)] = \mathbf{E}_\theta \|\hat{\theta}(\lambda) - \theta\|^2 - \sum_j \theta_j^2 = R(\lambda, \theta) - \sum_j \theta_j^2.$$

In other words, $\mathcal{J}(\lambda)$ is an unbiased estimator of the risk $R(\lambda, \theta)$, up to the term $\sum_j \theta_j^2$ independent of λ. Therefore we can expect that the minimizer of $\mathcal{J}(\lambda)$ would be close to the minimizer in λ of $R(\lambda, \theta)$.

Define

$$\tilde{\lambda} = \tilde{\lambda}(\Lambda) = \arg\min_{\lambda \in \Lambda} \mathcal{J}(\lambda).$$

The sequence $\tilde{\lambda} = (\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots)$ is a random sequence whose elements $\tilde{\lambda}_j = \tilde{\lambda}_j(y)$ in general depend on all the data $y = (y_1, y_2, \ldots)$. Define an estimator with weights $\tilde{\lambda}$ as follows:

$$\tilde{\theta}(\Lambda) = \hat{\theta}(\tilde{\lambda}) = \{\tilde{\theta}_j\},$$

where

$$\tilde{\theta}_j = \tilde{\lambda}_j(y) y_j, \quad j = 1, 2, \ldots .$$

We will see in the examples given below that $\tilde{\theta}$ is a nonlinear estimator, i.e., the coefficients $\tilde{\lambda}_j(y)$ are not constant as functions of y.

The role of $\tilde{\theta}(\Lambda)$ is to mimic the behavior of the oracle $\hat{\theta}(\lambda^{oracle}(\Lambda, \theta))$: as we will see it in the next section, under fairly general conditions the risk of $\tilde{\theta}(\Lambda)$ is asymptotically smaller than or equal to that of the oracle. This property will allow us to interpret $\tilde{\theta}(\Lambda)$ as an adaptive estimator; it adapts to the unknown oracle.

Definition 3.4 *Let $\Theta \subseteq \ell^2(\mathbf{N})$ be a class of sequences and let $\Lambda \subseteq \ell^2(\mathbf{N})$ be a class of weights. An estimator θ^*_ε of θ in model (3.10) is called* **adaptive to the oracle** $\lambda^{oracle}(\Lambda, \cdot)$ *on Θ if there exists a constant $C < \infty$ such that*

$$\mathbf{E}_\theta \|\theta^*_\varepsilon - \theta\|^2 \le C \inf_{\lambda \in \Lambda} \mathbf{E}_\theta \|\hat\theta(\lambda) - \theta\|^2$$

for all $\theta \in \Theta$ and $0 < \varepsilon < 1$.

An estimator θ^*_ε of θ is called **adaptive to the oracle** $\lambda^{oracle}(\Lambda, \cdot)$ **in the exact sense** *on Θ if it satisfies*

$$\mathbf{E}_\theta \|\theta^*_\varepsilon - \theta\|^2 \le (1 + o(1)) \inf_{\lambda \in \Lambda} \mathbf{E}_\theta \|\hat\theta(\lambda) - \theta\|^2$$

for all $\theta \in \Theta$ where $o(1)$ tends to 0 as $\varepsilon \to 0$ uniformly in $\theta \in \Theta$.

Below we will consider some examples of classes Λ, of the corresponding oracles $\lambda^{oracle}(\Lambda, \theta)$ and estimators $\tilde\theta(\Lambda)$ obtained by minimization of $\mathcal{J}(\lambda)$. The following two remarks are important to design the classes Λ in a natural way.

(1) *It is sufficient to consider $\lambda_j \in [0, 1]$.* Indeed, the projection of $\lambda_j \notin [0, 1]$ on $[0, 1]$ only reduces the risk $R(\lambda, \theta)$ of a linear estimator $\hat\theta(\lambda)$.

(2) *Usually it is sufficient to set $\lambda_j = 0$ for $j > N_{\max}$ where*

$$N_{\max} = \lfloor 1/\varepsilon^2 \rfloor. \tag{3.66}$$

Indeed, we mainly deal here with $\theta \in \Theta(\beta, Q)$ for $\beta > 0$ and $Q > 0$. A typical situation is that θ corresponds to a continuous function, so that it makes sense to consider $\beta > 1/2$ (cf. remark at the end of Section 1.7.1). The squared risk of the linear estimator is

$$R(\lambda, \theta) = \sum_{j=1}^{N_{\max}} \left[(1 - \lambda_j)^2 \theta_j^2 + \varepsilon^2 \lambda_j^2 \right] + r_0(\varepsilon)$$

where the residual $r_0(\varepsilon)$ is controlled in the following way for $\theta \in \Theta(\beta, Q)$ and $\beta > 1/2$:

$$\begin{aligned} r_0(\varepsilon) &= \sum_{j > N_{\max}} \left[(1 - \lambda_j)^2 \theta_j^2 + \varepsilon^2 \lambda_j^2 \right] \\ &\le \sum_{j > N_{\max}} \theta_j^2 + o(\varepsilon^2) \qquad (\text{since } 0 \le \lambda_j \le 1,\ \lambda \in \ell^2(\mathbf{N}),\ N_{\max} \to \infty) \\ &\le N_{\max}^{-2\beta} \sum_{j > N_{\max}} (j-1)^{2\beta} \theta_j^2 + o(\varepsilon^2) = o(N_{\max}^{-2\beta} + \varepsilon^2) = o(\varepsilon^2) \end{aligned}$$

as $\varepsilon \to 0$.

Another reason for keeping only a finite number of coordinates $\hat{\theta}_i$ is that we would like to construct a computationally feasible estimator. In general, the cutoff $N_{\max}$ is taken to be finite even though it may differ from (3.66).

Example 3.1 *Estimators with constant weights in Model 1.*

Consider the finite-dimensional model

$$y_j = \theta_j + \varepsilon\xi_j, \quad j = 1, \dots, d,$$

where ξ_j are i.i.d. $\mathcal{N}(0,1)$ random variables (Model 1 of Section 3.4). Introduce the class Λ as follows:

$$\Lambda_{const} = \{\lambda \mid \lambda_j \equiv t,\ j = 1, \dots, d,\ t \in [0,1]\}.$$

The estimator with constant weights of the vector $\theta = (\theta_1, \dots, \theta_d)$ is defined by

$$\hat{\theta}(t) = ty = (ty_1, \dots, ty_d).$$

It is easy to see that the minimum of the squared risk among all estimators with constant weights is equal to

$$\min_t \mathbf{E}_\theta \|\hat{\theta}(t) - \theta\|^2 = \min_t \sum_{j=1}^{d} [(1-t)^2\theta_j^2 + \varepsilon^2 t^2] = \frac{d\varepsilon^2\|\theta\|^2}{d\varepsilon^2 + \|\theta\|^2}\ . \quad (3.67)$$

The value of $t = t^*$ that achieves this minimum,

$$t^* = \frac{\|\theta\|^2}{d\varepsilon^2 + \|\theta\|^2}\ ,$$

corresponds to the *oracle with constant weights* $\lambda^{oracle}(\Lambda_{const}, \theta) = (t^*, \dots, t^*)$. For weights $\lambda = (t, \dots, t)$ belonging to Λ_{const}, the function $\mathcal{J}(\lambda)$ in the unbiased estimator of the risk has the form

$$\mathcal{J}(\lambda) = \sum_{j=1}^{d} (t^2 y_j^2 - 2t(y_j^2 - \varepsilon^2)) = (t^2 - 2t)\|y\|^2 + 2td\varepsilon^2,$$

and the minimizer in $t \in [0,1]$ of this expression is

$$\tilde{t} = \left(1 - \frac{d\varepsilon^2}{\|y\|^2}\right)_+ .$$

The corresponding estimator $\tilde{\theta}$ is therefore the positive part Stein estimator

$$\tilde{\theta} = \tilde{\theta}(\Lambda_{const}) = \left(1 - \frac{d\varepsilon^2}{\|y\|^2}\right)_+ y = \hat{\theta}_{S+}.$$

By Lemma 3.10,

$$\mathbf{E}_\theta\|\tilde{\theta}-\theta\|^2 \le \frac{d\varepsilon^2\|\theta\|^2}{d\varepsilon^2+\|\theta\|^2}+4\varepsilon^2.$$

This result and (3.67) imply the following inequality, valid under Model 1, which we will refer to as the *first oracle inequality*:

$$\mathbf{E}_\theta\|\tilde{\theta}-\theta\|^2 \le \min_t \mathbf{E}_\theta\|\hat{\theta}(t)-\theta\|^2+4\varepsilon^2. \tag{3.68}$$

Note that in this example $\mathcal{J}(\lambda)$ is equal to SURE up to a summand that does not depend on t (cf. Lemma 3.8 with $g(y)\equiv t$). Thus, the positive part Stein estimator is the estimator whose weights are obtained by minimization of SURE in $t\in[0,1]$.

Example 3.2 *Projection estimators.*

Consider the class of weights

$$\Lambda_{proj}=\{\lambda \mid \lambda_j=I\{j\le N\},\ N\in\{1,2,\dots,N_{\max}\}\}.$$

The corresponding linear estimator is given by

$$\hat{\theta}_{j,N}=\begin{cases} y_j, \text{ if } 1\le j\le N,\\ 0, \ \text{ if } j>N.\end{cases}$$

This is a simple projection estimator similar to that studied in Chapter 1 for the nonparametric regression model. If $\lambda\in\Lambda_{proj}$, the function $\mathcal{J}(\lambda)$ in the unbiased estimator of the risk is as follows:

$$\mathcal{J}(\lambda)=\sum_{j\le N}(y_j^2-2(y_j^2-\varepsilon^2))=2N\varepsilon^2-\sum_{j\le N}y_j^2$$

and the weights $\tilde{\lambda}_j$ obtained by minimization of $\mathcal{J}(\lambda)$ are of the form

$$\tilde{\lambda}_j=I\{j\le\tilde{N}\} \tag{3.69}$$

with

$$\tilde{N}=\arg\min_{1\le N\le N_{\max}}\big(2N\varepsilon^2-\sum_{j\le N}y_j^2\big) \tag{3.70}$$

Note that we can write

$$\tilde{N}=\arg\min_{1\le N\le N_{\max}}\big(\sum_{j=1}^{N_{\max}}(y_j-\hat{\theta}_{j,N})^2+2N\varepsilon^2\big). \tag{3.71}$$

Thus, $\tilde{N}$ is a minimizer of the penalized residual sum of squares. The penalty is $2N\varepsilon^2$. This can be linked to the C_p-criterion for regression (1.105) using the standard correspondence $\varepsilon=\sigma/\sqrt{n}$ (cf. Section 1.10 for the equivalence argument). In other words, $\tilde{N}$ is a minimizer of the C_p-criterion for the Gaussian sequence model.

Example 3.3 *Spline-type estimators.*

By Exercise 1.11, the spline estimator is approximated by the weighted projection estimator with weights

$$\lambda_j = \frac{1}{1 + \kappa\pi^2 a_j^2},$$

where $\kappa > 0$ and

$$a_j = \begin{cases} j^\beta, & \text{for even } j, \\ (j-1)^\beta, & \text{for odd } j. \end{cases}$$

Following this, we can define a class of linear estimators that are close to spline estimators:

$$\Lambda_{spline} = \left\{ \lambda \;\Big|\; \lambda_j = \frac{1}{1 + s a_j^2}\ I\{j \le N_{\max}\},\ s \in S, \beta \in B \right\}$$

with appropriate sets $S \subseteq (0, \infty)$ and $B \subseteq (0, \infty)$, where the integer $N_{\max}$ is defined by (3.66). The corresponding nonlinear estimator $\tilde{\theta}$ has weights $\tilde{\lambda}(\Lambda_{spline})$ minimizing $\mathcal{J}(\lambda)$ over Λ_{spline}. A similar definition can be given for the class

$$\Lambda'_{spline} = \left\{ \lambda \;\Big|\; \lambda_j = \frac{1}{1 + s j^{2\beta}}\ I\{j \le N_{\max}\},\ s \in S, \beta \in B \right\}.$$

Example 3.4 *Pinsker-type estimators.*

Consider the class of weights

$$\Lambda_{Pinsk} = \{\lambda \mid \lambda_j = (1 - s a_j)_+\ I\{j \le N_{\max}\},\ s \in S, \beta \in B\},$$

where $S \subseteq (0, \infty)$ and $B \subseteq (0, \infty)$ are given sets and a_j are defined as in Example 3.3. A similar class is

$$\Lambda'_{Pinsk} = \{\lambda \mid \lambda_j = (1 - s j^\beta)_+\ I\{j \le N_{\max}\},\ s \in S, \beta \in B\}.$$

Pinsker weights (3.4) belong to Λ_{Pinsk} under a fairly general choice of the sets S and B. Observe also that, by definition (3.66) of $N_{\max}$, for $B \subset (1/2, \infty)$ and for a reasonable choice of S we have

$$(1 - s a_j)_+\ I(j \le N_{\max}) = (1 - s a_j)_+.$$

Minimization of the unbiased estimator of the risk over this class of sequences λ leads to a nonlinear estimator $\tilde{\theta}(\Lambda_{Pinsk})$.

The classes Λ defined in Examples 3.1–3.4 are important special classes. It will also be useful to introduce two "super-classes": the class of monotone weights and the class of blockwise constant weights. The class of monotone weights can be called a "super-class" since it contains all the classes defined in Examples 3.1–3.4 (indeed, in Examples 3.1–3.4 the weights λ_j are non-increasing functions of j), as well as many other interesting classes. The class of blockwise constant weights is important because it provides a sufficiently accurate approximation of the class of monotone weights, as we will see it in the next section.

Example 3.5 *Estimators with monotone weights.*

Define the following class of weights:

$$\Lambda_{mon} = \{\lambda \mid 1 \geq \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{N_{\max}} \geq 0,\ \lambda_j = 0,\ j > N_{\max}\}$$

and call the sequence

$$\lambda^{oracle}(\Lambda_{mon}, \theta) = \arg \min_{\lambda \in \Lambda_{mon}} R(\lambda, \theta)$$

the *monotone oracle*. The respective data-driven choice of weights is defined by

$$\tilde{\lambda} = \tilde{\lambda}(\Lambda_{mon}) = \arg \min_{\lambda \in \Lambda_{mon}} \mathcal{J}(\lambda).$$

Example 3.6 *Estimators with blockwise constant weights.*

Consider a partitioning of the set $\{1, 2, \ldots, N_{\max}\}$ in blocks $B_j,\ j = 1, \ldots, J$:

$$\bigcup_{j=1}^{J} B_j = \{1, 2, \ldots, N_{\max}\}, \qquad B_i \cap B_j = \emptyset,\ i \neq j.$$

Suppose also that $\min\{k :\ k \in B_j\} > \max\{k :\ k \in B_{j-1}\}$. The class of blockwise constant weights is defined as follows:

$$\Lambda^* = \Big\{\lambda \ \Big|\ \lambda_k = \sum_{j=1}^{J} t_j I(k \in B_j) :\ 0 \leq t_j \leq 1,\ j = 1, \ldots, J\Big\}.$$

The importance of this class is explained by the fact that one can approximate rather different weights by blockwise constant weights. Minimization of $\mathcal{J}(\lambda)$ over Λ^* is particularly simple and explicit. Indeed, the coordinates of the vector

$$\tilde{\lambda} = \arg \min_{\lambda \in \Lambda^*} \mathcal{J}(\lambda)$$

are blockwise constant:

$$\tilde{\lambda}_k = \sum_{j=1}^{J} \tilde{\lambda}_{(j)} I(k \in B_j),$$

where

$$\tilde{\lambda}_{(j)} = \arg \min_{t \in [0,1]} \sum_{k \in B_j} (t^2 y_k^2 - 2t(y_k^2 - \varepsilon^2)).$$

Note that

$$\arg \min_{t \in \mathbf{R}} \sum_{k \in B_j} (t^2 y_k^2 - 2t(y_k^2 - \varepsilon^2)) = 1 - \frac{\varepsilon^2 T_j}{\|y\|_{(j)}^2}, \tag{3.72}$$

where

$$\|y\|_{(j)}^2 \stackrel{\triangle}{=} \sum_{k \in B_j} y_k^2, \qquad T_j \stackrel{\triangle}{=} \text{Card } B_j.$$

The projection of (3.72) on $[0, 1]$ is

$$\tilde{\lambda}_{(j)} = \left(1 - \frac{\varepsilon^2 T_j}{\|y\|_{(j)}^2}\right)_+, \quad j = 1, \ldots, J. \tag{3.73}$$

Hence the adaptive estimator obtained by minimizing $\mathcal{J}(\lambda)$ over Λ^* has the following form:

$$\tilde{\theta}_k = \begin{cases} \tilde{\lambda}_{(j)} y_k, \text{ if } k \in B_j, \ j = 1, \ldots, J, \\ 0, \qquad \text{if } k > N_{\max} \end{cases} \tag{3.74}$$

with $\tilde{\lambda}_{(j)}$ defined in (3.73).

Conclusion: Minimization of $\mathcal{J}(\lambda)$ over Λ^* produces blockwise constant positive part Stein estimators.

Definition 3.5 *The estimator $\tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2, \ldots)$ where $\tilde{\theta}_k$ is defined by (3.74) is called the* **block Stein estimator**.

Remark.

The results in Section 3.4 show that Stein's shrinkage gives an improvement only if $d \geq 3$. Therefore the weights $\tilde{\lambda}_{(j)}$ in (3.74) can be replaced by 1 in blocks of size $T_j \leq 2$.

3.6 Oracle inequalities

The aim of this section is to establish some inequalities for the risk of the block Stein estimator.

Let $\tilde{\theta}$ be the block Stein estimator and let θ be any sequence in $\ell^2(\mathbf{N})$. Define the corresponding vectors $\theta_{(j)}, \tilde{\theta}_{(j)} \in \mathbf{R}^{T_j}$:

$$\theta_{(j)} = (\theta_k,\ k \in B_j), \qquad \tilde{\theta}_{(j)} = (\tilde{\theta}_k,\ k \in B_j), \quad j = 1, \dots, J.$$

By the first oracle inequality (3.68), for each block B_j we have

$$\mathbf{E}_\theta \|\tilde{\theta}_{(j)} - \theta_{(j)}\|^2_{(j)} \le \min_{t_j} \sum_{k \in B_j} [(1-t_j)^2 \theta_k^2 + \varepsilon^2 t_j^2] + 4\varepsilon^2, \quad j = 1, \dots, J.$$

Then

$$\begin{aligned}
\mathbf{E}_\theta \|\tilde{\theta} - \theta\|^2 &= \sum_{j=1}^{J} \mathbf{E}_\theta \|\tilde{\theta}_{(j)} - \theta_{(j)}\|^2_{(j)} + \sum_{k > N_{\max}} \theta_k^2 \\
&\le \sum_{j=1}^{J} \min_{t_j} \sum_{k \in B_j} [(1-t_j)^2 \theta_k^2 + \varepsilon^2 t_j^2] + \sum_{k > N_{\max}} \theta_k^2 + 4J\varepsilon^2 \\
&= \min_{\lambda \in \Lambda^*} R(\lambda, \theta) + 4J\varepsilon^2.
\end{aligned}$$

Hence the following result is proved.

Theorem 3.4 *Let $\tilde{\theta}$ be the block Stein estimator. Then, for all $\theta \in \ell^2(\mathbf{N})$,*

$$\mathbf{E}_\theta \|\tilde{\theta} - \theta\|^2 \le \min_{\lambda \in \Lambda^*} R(\lambda, \theta) + 4J\varepsilon^2. \tag{3.75}$$

In what follows, (3.75) will be referred to as the *second oracle inequality*. Like the first oracle inequality, it is nonasymptotic, that is, it holds for all ε. It says that, up to the residual term $4J\varepsilon^2$ independent of θ, the block Stein estimator $\tilde{\theta}$ mimics the *blockwise constant oracle*

$$\lambda^{oracle}(\Lambda^*, \theta) = \arg\min_{\lambda \in \Lambda^*} R(\lambda, \theta).$$

Let us now show that the blockwise constant oracle is almost as good as the monotone oracle. We will need the following assumption on the system of blocks.

Assumption (C)
There exists $\eta > 0$ such that

$$\max_{1 \le j \le J-1} \frac{T_{j+1}}{T_j} \le 1 + \eta.$$

Lemma 3.11 *If Assumption (C) holds then, for all $\theta \in \ell^2(\mathbf{N})$,*

$$\min_{\lambda \in \Lambda^*} R(\lambda, \theta) \le (1+\eta) \min_{\lambda \in \Lambda_{mon}} R(\lambda, \theta) + \varepsilon^2 T_1.$$

PROOF. It is sufficient to show that for any sequence $\lambda \in \Lambda_{mon}$ there exists a sequence $\bar{\lambda} \in \Lambda^*$ such that

$$R(\bar{\lambda}, \theta) \le (1+\eta) R(\lambda, \theta) + \varepsilon^2 T_1, \qquad \forall\ \theta \in \ell^2(\mathbf{N}). \tag{3.76}$$

We are going to prove that inequality (3.76) holds for a sequence $\bar{\lambda} = (\bar{\lambda}_1, \bar{\lambda}_2, \ldots)$ defined as follows:

$$\bar{\lambda}_k = \begin{cases} \bar{\lambda}_{(j)} \stackrel{\triangle}{=} \max_{m \in B_j} \lambda_m, \text{ if } k \in B_j,\ j = 1, \ldots, J, \\ 0, \qquad\qquad\qquad\quad \text{if } k > N_{\max}. \end{cases}$$

It is clear that $\bar{\lambda}_k \ge \lambda_k$ for $k = 1, 2, \ldots$. Hence,

$$R(\bar{\lambda}, \theta) = \sum_{k=1}^{\infty} [(1-\bar{\lambda}_k)^2 \theta_k^2 + \varepsilon^2 \bar{\lambda}_k^2] \le \sum_{k=1}^{\infty} [(1-\lambda_k)^2 \theta_k^2 + \varepsilon^2 \bar{\lambda}_k^2].$$

Since

$$R(\lambda, \theta) = \sum_{k=1}^{\infty} [(1-\lambda_k)^2 \theta_k^2 + \varepsilon^2 \lambda_k^2],$$

the proof will be complete if we show that

$$\varepsilon^2 \sum_{k=1}^{\infty} \bar{\lambda}_k^2 \le (1+\eta) \varepsilon^2 \sum_{k=1}^{\infty} \lambda_k^2 + \varepsilon^2 T_1. \tag{3.77}$$

But (3.77) follows from the chain of inequalities:

$$\begin{aligned} \sum_{k=1}^{\infty} \bar{\lambda}_k^2 &= \sum_{k \le N_{\max}} \bar{\lambda}_k^2 \\ &\le T_1 + \sum_{j=2}^{J} \sum_{k \in B_j} \bar{\lambda}_k^2 \qquad (\text{since } 0 \le \bar{\lambda}_1 \le 1) \\ &= T_1 + \sum_{j=2}^{J} T_j \bar{\lambda}_{(j)}^2 \\ &\le T_1 + (1+\eta) \sum_{j=2}^{J} T_{j-1} \bar{\lambda}_{(j)}^2 \qquad (\text{by Assumption (C)}) \\ &\le T_1 + (1+\eta) \sum_{j=2}^{J} \sum_{m \in B_{j-1}} \lambda_m^2 \qquad (\text{since } \bar{\lambda}_{(j)} \le \lambda_m,\ \forall\ m \in B_{j-1}) \end{aligned}$$

$$= T_1 + (1+\eta)\sum_{j=1}^{J-1}\sum_{m\in B_j}\lambda_m^2$$

$$\le T_1 + (1+\eta)\sum_{k=1}^{\infty}\lambda_k^2. \qquad \blacksquare$$

Theorem 3.5 *Suppose that the blocks satisfy Assumption (C). Then for all* $\theta \in \ell^2(\mathbf{N})$ *the risk of the block Stein estimator* $\tilde\theta$ *satisfies*

$$\mathbf{E}_\theta\|\tilde\theta - \theta\|^2 \le (1+\eta)\min_{\lambda\in\Lambda_{mon}} R(\lambda,\theta) + \varepsilon^2(T_1 + 4J). \tag{3.78}$$

The proof of this theorem is straightforward in view of Theorem 3.4 and Lemma 3.11.

Formula (3.78) will be called the *third oracle inequality*. Like the first two oracle inequalities, it is nonasymptotic, i.e., it holds for all ε. It says that if η is sufficiently small the block Stein estimator $\tilde\theta$ mimics the monotone oracle, up to the residual term $\varepsilon^2(T_1 + 4J)$ independent of θ.

The question arising now is as follows: How to construct good systems of blocks, i.e., systems $\{B_j\}$ such that η and the residual term $\varepsilon^2(T_1+4J)$ would be sufficiently small? Let us consider some examples.

Example 3.7 *Diadic blocks.*

Let $T_j = 2^j$ for $j = 1, \dots, J-1$. This assumption is standard in the context of wavelet estimation. Then $\eta = 1$ in Assumption (C), and the total number J of blocks $\{B_j\}$ satisfies $J \le \log_2(2 + 1/\varepsilon^2)$ by (3.66). Therefore, inequality (3.78) takes the following form:

$$\mathbf{E}_\theta\|\tilde\theta - \theta\|^2 \le 2\min_{\lambda\in\Lambda_{mon}} R(\lambda,\theta) + \varepsilon^2(2 + 4\log_2(2 + 1/\varepsilon^2)),$$

where $\tilde\theta$ is the Stein estimator with diadic blocks. Note that the residual term is small (of order $\varepsilon^2\log(1/\varepsilon)$) but the oracle risk on the right hand side is multiplied by 2. Therefore the inequality is rather rough; it does not guarantee that the risk of $\tilde\theta$ becomes close to that of the oracle, even asymptotically. This is explained by the fact that the lengths T_j of diadic blocks increase too fast; this system of blocks is not sufficiently flexible. A better performance is achieved by using another system of blocks described in the next example.

Example 3.8 *Weakly geometric blocks.*

This construction of blocks is entirely determined by a value $\rho_\varepsilon > 0$ such that $\rho_\varepsilon \to 0$ as $\varepsilon \to 0$. We will take

$$\rho_\varepsilon = (\log(1/\varepsilon))^{-1}, \tag{3.79}$$

though there exist other choices of ρ_ε leading to analogous results. The lengths of the blocks T_j are defined as follows:

$$\begin{aligned} T_1 &= \lceil \rho_\varepsilon^{-1} \rceil = \lceil \log(1/\varepsilon) \rceil, \\ T_2 &= \lfloor T_1(1+\rho_\varepsilon) \rfloor, \\ &\vdots \\ T_{J-1} &= \lfloor T_1(1+\rho_\varepsilon)^{J-2} \rfloor, \\ T_J &= N_{\max} - \sum_{j=1}^{J-1} T_j \end{aligned} \tag{3.80}$$

where

$$J = \min\{m : \ T_1 + \sum_{j=2}^{m} \lfloor T_1(1+\rho_\varepsilon)^{j-1} \rfloor \geq N_{\max}\}. \tag{3.81}$$

Observe that

$$T_J \leq \lfloor T_1(1+\rho_\varepsilon)^{J-1} \rfloor.$$

Definition 3.6 *The system of blocks* $\{B_j\}$ *defined by (3.79)–(3.81) with* $N_{\max}$ *defined by (3.66) is called a* **weakly geometric system of blocks**, *or a* **WGB system**. *The corresponding block Stein estimator is called the* **Stein WGB estimator**.

The quantities η and J for the WGB system are given in the following lemma.

Lemma 3.12 *Let* $\{B_j\}$ *be a WGB system. Then there exist* $0 < \varepsilon_0 < 1$ *and* $C > 0$ *such that:*

(i) $J \leq C \log^2(1/\varepsilon)$ *for any* $\varepsilon \in (0, \varepsilon_0)$,
(ii) Assumption (C) holds with $\eta = 3\rho_\varepsilon$ *for all* $\varepsilon \in (0, \varepsilon_0)$.

Proof. Suppose that ε is sufficiently small for the inequality $\rho_\varepsilon < 1$ to hold and observe that

$$\lfloor x \rfloor \geq x - 1 \geq x(1-\rho_\varepsilon), \qquad \forall x \geq \rho_\varepsilon^{-1}.$$

Then

$$\lfloor T_1(1+\rho_\varepsilon)^{j-1} \rfloor \geq T_1(1+\rho_\varepsilon)^{j-1}(1-\rho_\varepsilon). \tag{3.82}$$

Using (3.82) we obtain

$$\begin{aligned} T_1 + \sum_{j=2}^{J-1} \lfloor T_1(1+\rho_\varepsilon)^{j-1} \rfloor &\geq T_1 \Big(1 + \sum_{j=1}^{J-2} (1+\rho_\varepsilon)^j (1-\rho_\varepsilon) \Big) \\ &\geq \rho_\varepsilon^{-1} \left(1 + \rho_\varepsilon^{-1} [(1+\rho_\varepsilon)^{J-2} - 1](1-\rho_\varepsilon^2) \right). \end{aligned}$$

It follows from (3.81) that

$$\rho_\varepsilon^{-1}\left(1+\rho_\varepsilon^{-1}[(1+\rho_\varepsilon)^{J-2}-1](1-\rho_\varepsilon^2)\right) \le N_{\max} \le \varepsilon^{-2}.$$

Therefore, for a constant $C<\infty$ and all $\varepsilon>0$ small enough,

$$(1+\rho_\varepsilon)^{J-2} \le C\rho_\varepsilon^2\varepsilon^{-2},$$

implying (i). To prove (ii) observe that, by (3.82),

$$\begin{aligned}\frac{T_{j+1}}{T_j} &\le \frac{\lfloor T_1(1+\rho_\varepsilon)^j\rfloor}{\lfloor T_1(1+\rho_\varepsilon)^{j-1}\rfloor} \le \frac{(1+\rho_\varepsilon)^j}{(1+\rho_\varepsilon)^{j-1}(1-\rho_\varepsilon)}\\ &= \frac{1+\rho_\varepsilon}{1-\rho_\varepsilon} \le 1+3\rho_\varepsilon\end{aligned}$$

if $\rho_\varepsilon \le 1/3$. ∎

Corollary 3.2 *Let $\tilde\theta$ be a Stein WGB estimator. Then there exist constants $0<\varepsilon_0<1$ and $C<\infty$ such that*

$$\mathbf{E}_\theta\|\tilde\theta-\theta\|^2 \le (1+3\rho_\varepsilon)\min_{\lambda\in\Lambda_{mon}} R(\lambda,\theta) + C\varepsilon^2\log^2(1/\varepsilon) \tag{3.83}$$

for all $\theta\in\ell^2(\mathbf{N})$ and all $0<\varepsilon<\varepsilon_0$.

The proof is straightforward in view of Theorem 3.5 and Lemma 3.12.

Since $\rho_\varepsilon = o(1)$, the oracle inequality (3.83) is asymptotically exact. More specifically, it implies the following asymptotic result.

Corollary 3.3 *Let $\tilde\theta$ be a Stein WGB estimator and let $\theta\in\ell^2(\mathbf{N})$ be a sequence satisfying*

$$\frac{\min_{\lambda\in\Lambda} R(\lambda,\theta)}{\varepsilon^2\log^2(1/\varepsilon)} \;\to\; \infty \quad \text{as } \varepsilon\to 0$$

for a class $\Lambda\subseteq\Lambda_{mon}$. Then

$$\mathbf{E}_\theta\|\tilde\theta-\theta\|^2 \le (1+o(1))\min_{\lambda\in\Lambda} R(\lambda,\theta) \quad \text{as } \varepsilon\to 0. \tag{3.84}$$

Remark.

It is clear that inequality (3.83) remains valid if we replace there $\min_{\lambda\in\Lambda_{mon}}$ by $\min_{\lambda\in\Lambda}$ for a class $\Lambda\subset\Lambda_{mon}$. Therefore inequalities (3.83) and (3.84) can be applied to the classes $\Lambda=\Lambda_{proj}$, Λ_{spline}, Λ_{Pinsk}, etc. Thus, the Stein WGB estimator is asymptotically at least as good as, and in fact even better than, the oracles corresponding to these particular classes. In other words, the Stein WGB estimator is adaptive to the oracles $\lambda^{oracle}(\Lambda_{proj},\cdot)$, $\lambda^{oracle}(\Lambda_{spline},\cdot)$, $\lambda^{oracle}(\Lambda_{Pinsk},\cdot)$ in the exact sense.

3.7 Minimax adaptivity

Suppose that θ belongs to an ellipsoid $\Theta = \Theta(\beta, Q)$ with $\beta > 0$, $Q > 0$ and that we consider estimation of θ in the Gaussian sequence model (3.10). The definition of asymptotically efficient estimator for this model takes the following form (cf. Definition 2.2).

Definition 3.7 *An estimator θ_ε^* of θ in model (3.10) is called* **asymptotically efficient** *on the class Θ if*

$$\lim_{\varepsilon \to 0} \frac{\sup_{\theta \in \Theta} \mathbf{E}_\theta \|\theta_\varepsilon^* - \theta\|^2}{\inf_{\hat{\theta}_\varepsilon} \sup_{\theta \in \Theta} \mathbf{E}_\theta \|\hat{\theta}_\varepsilon - \theta\|^2} = 1,$$

where the infimum is over all estimators.

Corollary 3.1 and formula (3.49) imply that the simplified Pinsker estimator $\hat{\theta}(\ell^*)$, as well as the Pinsker estimator $\hat{\theta}(\ell)$ (where the sequences of optimal weights $\ell^* = (\ell_1^*, \ell_2^*, \ldots)$ and $\ell = (\ell_1, \ell_2, \ldots)$ are defined by (3.4) and (3.20), respectively) are asymptotically efficient on the class $\Theta(\beta, Q)$.

The main drawback of these two estimators is that they depend on the parameters β and Q which are unknown in practice.

Definition 3.8 *An estimator θ_ε^* of θ in model (3.10) is called* **adaptive in the exact minimax sense** *on the family of classes $\{\Theta(\beta, Q), \beta > 0, Q > 0\}$ if it is asymptotically efficient for all classes $\Theta(\beta, Q)$, $\beta > 0, Q > 0$, simultaneously.*

Clearly, an adaptive estimator cannot depend on the parameters β and Q of individual classes $\Theta(\beta, Q)$.

We now prove that the Stein WGB estimator $\tilde{\theta}$ is adaptive in the exact minimax sense on the family of classes $\{\Theta(\beta, Q), \beta > 0, Q > 0\}$. This property of $\tilde{\theta}$ follows from oracle inequality (3.83) and Lemma 3.2. Indeed, by taking the upper bound on both sides of (3.83) with respect to $\theta \in \Theta(\beta, Q)$, we obtain

$$\sup_{\theta \in \Theta(\beta, Q)} \mathbf{E}_\theta \|\tilde{\theta} - \theta\|^2 \le (1 + 3\rho_\varepsilon) \sup_{\theta \in \Theta(\beta, Q)} \min_{\lambda \in \Lambda_{mon}} R(\lambda, \theta) + C\varepsilon^2 \log^2(1/\varepsilon). \tag{3.85}$$

Next, note that the sequence of linear minimax weights ℓ belongs to Λ_{mon} for sufficiently small ε. Indeed, the coefficients $\ell_j = (1 - \kappa a_j)_+$ in (3.20) are decreasing in j, and $\ell_j = 0$ if $j > c\varepsilon^{-\frac{2}{2\beta+1}}$ for a constant $c > 0$. We have $\ell_j = 0$ if $j > N_{\max}$ for sufficiently small ε, since $N_{\max} \sim 1/\varepsilon^2$ by (3.66). It follows that for sufficiently small ε we have $\min_{\lambda \in \Lambda_{mon}} R(\lambda, \theta) \le R(\ell, \theta)$ for all $\theta \in \Theta(\beta, Q)$. By plugging this inequality into (3.85) we obtain

$$\begin{aligned}\sup_{\theta\in\Theta(\beta,Q)} \mathbf{E}_\theta\|\tilde\theta-\theta\|^2 &\le (1+3\rho_\varepsilon)\sup_{\theta\in\Theta(\beta,Q)} R(\ell,\theta)+C\varepsilon^2\log^2(1/\varepsilon)\\ &= (1+3\rho_\varepsilon)\mathcal{D}^*+C\varepsilon^2\log^2(1/\varepsilon) \quad \text{(by Lemma 3.2)}\\ &= (1+3\rho_\varepsilon)C^*\varepsilon^{\frac{4\beta}{2\beta+1}}(1+o(1))\\ &\quad + C\varepsilon^2\log^2(1/\varepsilon) \qquad\qquad \text{(by (3.27))}\\ &= C^*\varepsilon^{\frac{4\beta}{2\beta+1}}(1+o(1)), \quad \varepsilon\to 0.\end{aligned}$$

Therefore, we have proved the following result.

Theorem 3.6 *The Stein WGB estimator $\tilde\theta$ is adaptive in the exact minimax sense on the family of Sobolev ellipsoids $\{\Theta(\beta,Q),\beta>0,Q>0\}$.*

This theorem is our main result on adaptivity on the family of classes $\Theta(\beta,Q)$. It shows that the Stein WGB estimator $\tilde\theta$ is much more attractive than the Pinsker estimators $\hat\theta(\ell^*)$ and $\hat\theta(\ell)$: $\tilde\theta$ possesses a much stronger efficiency property and the construction of this estimator is independent of β and Q. We also see that there is no price to pay for adaptivity: one can switch from the Pinsker estimator to an estimator independent of β and Q without increasing the asymptotic risk. Finally we mention that the Stein WGB estimator is not the only adaptive estimator in the sense of Definition 3.8 on the family of classes $\{\Theta(\beta,Q),\beta>0,Q>0\}$ (see the bibliographic notes in Section 3.9 below).

There also exist estimators having a weaker adaptivity property, which manifests itself only in the rates of convergence. The following definition describes this property.

Definition 3.9 *An estimator θ^*_ε of θ in model (3.10) is called* **adaptive in the minimax sense** *on the family of classes $\{\Theta(\beta,Q),\beta>0,Q>0\}$ if*

$$\sup_{\theta\in\Theta(\beta,Q)} \mathbf{E}_\theta\|\theta^*_\varepsilon-\theta\|^2 \le C(\beta,Q)\psi^2_\varepsilon, \quad \forall\ \beta>0,Q>0,0<\varepsilon<1,$$

where $\psi_\varepsilon=\varepsilon^{\frac{2\beta}{2\beta+1}}$ and where $C(\beta,Q)$ is a finite constant depending only on β and Q.

For example, one can prove that the Mallows C_p estimator, i.e., the estimator with weights $\tilde\lambda_j$ defined by (3.69)–(3.70), is adaptive in the minimax sense on the family of classes $\{\Theta(\beta,Q),\beta>0,Q>0\}$, though it is not adaptive in the exact minimax sense.

3.8 Inadmissibility of the Pinsker estimator

We now consider another corollary of the oracle inequality (3.83). It consists in the fact that the adaptive estimator $\tilde\theta$ is uniformly better than the Pinsker

estimator on any ellipsoid that is strictly contained in $\Theta(\beta, Q)$, so that the Pinsker estimator is inadmissible. The notion of admissibility is understood here in the sense of Definition 3.2, where we consider Θ as a subset of $\ell^2(\mathbf{N})$ and $\|\cdot\|$ as the $\ell^2(\mathbf{N})$-norm.

Proposition 3.2 *Let $\hat\theta(\ell)$ be the Pinsker estimator for the ellipsoid $\Theta(\beta, Q)$ with $\beta > 0$ and $Q > 0$. Then for any $0 < Q' < Q$ there exists $\varepsilon_1 \in (0,1)$ such that the Stein WGB estimator $\tilde\theta$ satisfies*

$$\mathbf{E}_\theta\|\tilde\theta - \theta\|^2 < \mathbf{E}_\theta\|\hat\theta(\ell) - \theta\|^2 \tag{3.86}$$

for all $\theta \in \Theta(\beta, Q')$ and $\varepsilon \in (0, \varepsilon_1)$. Hence $\hat\theta(\ell)$ is inadmissible on $\Theta(\beta, Q')$ for all $\varepsilon \in (0, \varepsilon_1)$.

Proof. Let $\ell' = (\ell'_1, \ell'_2, \dots)$ be the sequence of weights of the simplified Pinsker estimator for the ellipsoid $\Theta(\beta, Q')$:

$$\ell'_j = (1 - \kappa' a_j)_+ \quad \text{with} \quad \kappa' = \left(\frac{\beta}{(2\beta+1)(\beta+1)Q'}\right)^{\frac{\beta}{2\beta+1}} \varepsilon^{\frac{2\beta}{2\beta+1}}.$$

Observe that $\ell' \in \Lambda_{mon}$ for sufficiently small ε.

In view of (3.83), for all $\theta \in \ell^2(\mathbf{N})$ and $0 < \varepsilon < \varepsilon_0$ we get

$$\begin{aligned}\mathbf{E}_\theta\|\tilde\theta - \theta\|^2 &\le (1 + 3\rho_\varepsilon) R(\ell', \theta) + C\varepsilon^2 \log^2(1/\varepsilon) \\ &= R(\ell, \theta) + 3\rho_\varepsilon R(\ell', \theta) + [R(\ell', \theta) - R(\ell, \theta)] \\ &\quad + C\varepsilon^2 \log^2(1/\varepsilon),\end{aligned} \tag{3.87}$$

where ℓ is a sequence of Pinsker weights for $\Theta(\beta, Q)$ defined by (3.20). By (3.24), we have $\mathcal{D}^* = \varepsilon^2 \sum_{j=1}^\infty \ell_j^2 + Q\kappa^2$ implying, by (3.26) and (3.27),

$$Q\kappa^2 = \frac{\mathcal{D}^*}{2\beta+1}(1 + o(1)), \qquad \varepsilon^2 \sum_{j=1}^\infty \ell_j^2 = \frac{2\beta\mathcal{D}^*}{2\beta+1}(1 + o(1)) \tag{3.88}$$

as $\varepsilon \to 0$.

Observe that $\ell'_j \le \ell_j$ for all j. In the same way as in (3.24) we obtain, for all $\theta \in \Theta(\beta, Q')$,

$$\begin{aligned}\sum_{j=1}^\infty [(1-\ell'_j)^2 - (1-\ell_j)^2]\theta_j^2 &\le Q' \sup_{j: a_j > 0} \left([(1-\ell'_j)^2 - (1-\ell_j)^2] a_j^{-2} \right) \\ &\le Q'[(\kappa')^2 - \kappa^2].\end{aligned}$$

This inequality combined with (3.26), (3.27), and (3.88) implies that, for all $\theta \in \Theta(\beta, Q')$,

$$
\begin{aligned}
R(\ell',\theta) - R(\ell,\theta) &= \sum_{j=1}^{\infty}[(1-\ell'_j)^2 - (1-\ell_j)^2]\theta_j^2 \qquad (3.89)\\
&\quad + \ \varepsilon^2 \sum_{j=1}^{\infty}[(\ell'_j)^2 - \ell_j^2]\\
&\le \left[\varepsilon^2 \sum_{j=1}^{\infty}(\ell'_j)^2 + Q(\kappa')^2\right] - \varepsilon^2 \sum_{j=1}^{\infty}\ell_j^2 - Q'\kappa^2\\
&= \mathcal{D}' - \frac{2\beta\mathcal{D}^*}{2\beta+1}(1+o(1))\\
&\quad - Q'\left(\frac{\beta}{(2\beta+1)(\beta+1)Q}\right)^{\frac{2\beta}{2\beta+1}} \varepsilon^{\frac{4\beta}{2\beta+1}}(1+o(1)),
\end{aligned}
$$

where

$$
\begin{aligned}
\mathcal{D}' &= \varepsilon^2 \sum_{j=1}^{\infty}(\ell'_j)^2 + Q(\kappa')^2 \qquad (3.90)\\
&= [Q'(2\beta+1)]^{\frac{1}{2\beta+1}}\left(\frac{\beta}{\beta+1}\right)^{\frac{2\beta}{2\beta+1}} \varepsilon^{\frac{4\beta}{2\beta+1}}(1+o(1)).
\end{aligned}
$$

By (3.89) and (3.90), for all $\theta \in \Theta(\beta, Q')$ and all sufficiently small ε,

$$
\begin{aligned}
&R(\ell',\theta) - R(\ell,\theta) \qquad (3.91)\\
&\quad \le AQ^{\frac{1}{2\beta+1}}\left(\frac{\beta}{(2\beta+1)(\beta+1)}\right)^{\frac{2\beta}{2\beta+1}} \varepsilon^{\frac{4\beta}{2\beta+1}}(1+o(1))\\
&\quad \le -c_1\varepsilon^{\frac{4\beta}{2\beta+1}},
\end{aligned}
$$

where

$$
A = (2\beta+1)\left(\frac{Q'}{Q}\right)^{\frac{1}{2\beta+1}} - 2\beta - \frac{Q'}{Q},
$$

$c_1 > 0$ is a constant depending only on β, Q, and Q', and where we have used the fact that $(2\beta+1)x^{\frac{1}{2\beta+1}} < 2\beta + x$ for $0 \le x < 1$. On the other hand, by Lemma 3.2, (3.26), and (3.27), we have

$$
R(\ell',\theta) \le \sup_{\theta\in\Theta(\beta,Q')} R(\ell',\theta) = \mathcal{D}'(1+o(1)) \le c_2\varepsilon^{\frac{4\beta}{2\beta+1}} \qquad (3.92)
$$

for a constant $c_2 > 0$ depending only on Q' and β. Substituting (3.91) and (3.92) into (3.87) we obtain

$$
\mathbf{E}_\theta\|\tilde{\theta} - \theta\|^2 \le R(\ell,\theta) + (3c_2\rho_\varepsilon - c_1)\varepsilon^{\frac{4\beta}{2\beta+1}} + C\varepsilon^2\log^2(1/\varepsilon)
$$

for all $\theta \in \Theta(\beta, Q')$ and all sufficiently small ε. To complete the proof, it is enough to note that $(3c_2\rho_\varepsilon - c_1)\varepsilon^{\frac{4\beta}{2\beta+1}} + C\varepsilon^2\log^2(1/\varepsilon) < 0$ for all sufficiently small ε. ∎

This argument does not give an answer to the question on whether inequality (3.86) can be extended to the boundary of $\Theta(\beta, Q)$ and therefore whether $\hat{\theta}(\ell)$ is inadmissible over the whole set $\Theta(\beta, Q)$. However, we have the following asymptotic result:

$$\limsup_{\varepsilon \to 0} \sup_{\theta \in \Theta(\beta,Q)} \frac{\mathbf{E}_\theta \|\tilde{\theta} - \theta\|^2}{\mathbf{E}_\theta \|\hat{\theta}(\ell) - \theta\|^2} \le 1, \quad \forall\ \beta > 0, Q > 0. \tag{3.93}$$

Indeed, using (3.88) we get, for all $\theta \in \ell^2(\mathbf{N})$,

$$\begin{aligned} \mathbf{E}_\theta \|\hat{\theta}(\ell) - \theta\|^2 &= \sum_{j=1}^{\infty} (1 - \ell_j)^2 \theta_j^2 + \varepsilon^2 \sum_{j=1}^{\infty} \ell_j^2 \ \ge\ \varepsilon^2 \sum_{j=1}^{\infty} \ell_j^2 \\ &= \frac{2\beta \mathcal{D}^*}{2\beta + 1} (1 + o(1)) \ge c_3 \varepsilon^{\frac{4\beta}{2\beta+1}} \end{aligned} \tag{3.94}$$

for sufficiently small ε where $c_3 > 0$ is a constant depending only on β and Q. On the other hand, (3.83) implies

$$\mathbf{E}_\theta \|\tilde{\theta} - \theta\|^2 \le (1 + 3\rho_\varepsilon) \mathbf{E}_\theta \|\hat{\theta}(\ell) - \theta\|^2 + C\varepsilon^2 \log^2(1/\varepsilon). \tag{3.95}$$

Inequality (3.93) follows directly from (3.94) and (3.95). Observe that (3.94) and (3.95) hold for any fixed θ in $\ell^2(\mathbf{N})$, implying in fact a stronger inequality than (3.93):

$$\limsup_{\varepsilon \to 0} \sup_{\theta \in \ell^2(\mathbf{N})} \frac{\mathbf{E}_\theta \|\tilde{\theta} - \theta\|^2}{\mathbf{E}_\theta \|\hat{\theta}(\ell) - \theta\|^2} \le 1. \tag{3.96}$$

The following nonuniform result can also be proved.

Proposition 3.3 *Let $\hat{\theta}(\ell)$ be the Pinsker estimator for the ellipsoid $\Theta(\beta, Q)$ where $\beta > 0$ and $Q > 0$. Then for all $\theta \in \Theta(\beta, Q)$ the Stein WGB estimator $\tilde{\theta}$ satisfies*

$$\lim_{\varepsilon \to 0} \frac{\mathbf{E}_\theta \|\tilde{\theta} - \theta\|^2}{\mathbf{E}_\theta \|\hat{\theta}(\ell) - \theta\|^2} = 0 \tag{3.97}$$

and

$$\lim_{\varepsilon \to 0} \varepsilon^{-\frac{4\beta}{2\beta+1}} \mathbf{E}_\theta \|\tilde{\theta} - \theta\|^2 = 0. \tag{3.98}$$

Proof. Since $\Lambda_{proj} \subset \Lambda_{mon}$, inequality (3.83) yields

$$\begin{aligned} \mathbf{E}_\theta \|\tilde{\theta} - \theta\|^2 &\le (1 + 3\rho_\varepsilon) \min_{\lambda \in \Lambda_{proj}} R(\lambda, \theta) + C\varepsilon^2 \log^2(1/\varepsilon) \\ &= (1 + 3\rho_\varepsilon) \min_{N \le N_{\max}} \Big(\sum_{i=N+1}^{\infty} \theta_i^2 + \varepsilon^2 N \Big) + C\varepsilon^2 \log^2(1/\varepsilon). \end{aligned}$$

Put $N_\varepsilon = \lceil \delta \varepsilon^{-\frac{2}{2\beta+1}} \rceil \geq \delta \varepsilon^{-\frac{2}{2\beta+1}}$ with $\delta > 0$. For ε small enough, we have $N_\varepsilon < N_{\max}$ by (3.66). Then

$$\begin{aligned}\min_{N \leq N_{\max}} \Big(\sum_{i=N+1}^{\infty} \theta_i^2 + \varepsilon^2 N \Big) &\leq \sum_{i=N_\varepsilon}^{\infty} \theta_i^2 + \varepsilon^2 N_\varepsilon \\ &\leq N_\varepsilon^{-2\beta} \sum_{i=N_\varepsilon}^{\infty} i^{2\beta} \theta_i^2 + \varepsilon^2 N_\varepsilon \\ &\leq \delta^{-2\beta} \varepsilon^{\frac{4\beta}{2\beta+1}} \alpha_\varepsilon + \varepsilon^2 (\delta \varepsilon^{-\frac{2}{2\beta+1}} + 1),\end{aligned}$$

where $\alpha_\varepsilon = \sum_{i=N_\varepsilon}^{\infty} i^{2\beta} \theta_i^2 = o(1)$ as $\varepsilon \to 0$ for all $\theta \in \Theta(\beta, Q)$. Hence

$$\mathbf{E}_\theta \|\tilde{\theta} - \theta\|^2 \varepsilon^{-\frac{4\beta}{2\beta+1}} \leq \delta^{-2\beta} \alpha_\varepsilon + \delta(1 + o(1)).$$

By taking the limit as $\varepsilon \to 0$, we obtain

$$\limsup_{\varepsilon \to 0} \mathbf{E}_\theta \|\tilde{\theta} - \theta\|^2 \varepsilon^{-\frac{4\beta}{2\beta+1}} \leq \delta. \tag{3.99}$$

Since $\delta > 0$ is arbitrary, this yields (3.98). Finally, (3.97) follows from (3.94) and (3.99). ∎

Remarks.

(1) Proposition 3.2 demonstrates the superiority of the Stein WGB estimator $\tilde{\theta}$ (an adaptive estimator) over the Pinsker estimator, which is not adaptive.

(2) At first sight, the result of Proposition 3.3 seems to be surprising: One can improve the Pinsker estimator everywhere on the ellipsoid where this estimator is minimax. Moreover, the rate of convergence is also improved. However, it would seem natural that at least in the most unfavorable case (i.e., when θ belongs to the boundary of the ellipsoid) the Pinsker estimator could not be improved. The explanation of this paradox is simple: Although the least favorable sequence θ belongs to the boundary of the ellipsoid, this sequence depends on ε (indeed, this is the sequence $\theta(\varepsilon) = \{v_j\}$ with v_j defined by (3.25)). On the other hand, in Proposition 3.3 we deal with a sequence $\theta \in \ell^2(\mathbf{N})$ which is *fixed and independent of* ε. The rate of convergence to 0 in (3.97) and (3.98) is not uniform in θ; it becomes slower and slower as θ approaches the boundary of the ellipsoid $\Theta(\beta, Q)$.

(3) The result (3.97) in Proposition 3.3 can be enhanced in the following way:

$$\lim_{\varepsilon \to 0} \frac{\mathbf{E}_\theta \|\tilde{\theta} - \theta'\|^2}{\inf_{\theta \in \ell^2(\mathbf{N})} \mathbf{E}_\theta \|\hat{\theta}(\ell) - \theta\|^2} = 0, \qquad \forall\, \theta' \in \Theta(\beta, Q). \tag{3.100}$$

Indeed, it is easy to see that we can insert $\inf_{\theta \in \ell^2(\mathbf{N})}$ in front of the expectation in (3.94).

Arguing analogously to the finite-dimensional case considered in Section 3.4, we can define the concept of superefficiency in the nonparametric problem that we study here. Definiton 3.3 of superefficiency is naturally modified in the following way: instead of the quantity $d\varepsilon^2$ representing the minimax risk in Model 1 (d-dimensional Gaussian model), we now introduce $C^*\varepsilon^{\frac{4\beta}{2\beta+1}}$, which is the asymptotic value of the minimax risk on the ellipsoid $\Theta(\beta, Q)$.

Definition 3.10 *We say that an estimator θ^*_ε is* **superefficient** *at a point $\theta \in \Theta(\beta, Q)$ if*

$$\limsup_{\varepsilon \to 0} \frac{\mathbf{E}_\theta \|\theta^*_\varepsilon - \theta\|^2}{C^* \varepsilon^{\frac{4\beta}{2\beta+1}}} < 1,$$

where C^* is the Pinsker constant.

Then the following corollary of Proposition 3.3 is immediate.

Corollary 3.4 *The Stein WGB estimator $\tilde{\theta}$ is superefficient at any point of $\Theta(\beta, Q)$ for all $\beta > 0$ and $Q > 0$.*

This result differs dramatically from its finite-dimensional analog in Section 3.4 (cf. Proposition 3.1). The Pinsker estimator is an asymptotically efficient estimator whose role is similar to that of the asymptotically efficient estimator y in Model 1 of Section 3.4. In turn, the Stein WGB estimator is an analog of the finite-dimensional Stein estimator in Section 3.4. Whereas in the finite-dimensional case superefficiency is possible only on a set of Lebesgue measure zero (note the remark at the end of Section 3.4), we see that in the nonparametric situation there exist estimators that are everywhere superefficient.

3.9 Notes

Theorems 3.1 and 3.2 are due to Pinsker (1980) and Nussbaum (1985), respectively. Pinsker (1980) established a more general result than Theorem 3.1, not necessarily restricted to the Sobolev ellipsoids. He imposed only very mild conditions on a_j. Another proof of Pinsker's lower bound for the Sobolev ellipsoids can be derived from the van Trees inequality (cf. Belitser and Levit (1995)). Linear minimax lemma in this form was first stated by Pinsker (1980). A similar result was proved earlier by Kuks and Olman (1971) for finite-dimensional regression models.

Lemmas 3.6 and 3.8 are due to Stein (1981). The estimator $\hat{\theta}_{JS}$ was introduced by James and Stein (1961). Strawderman (1971) constructed an admissible estimator of θ in Model 1. For a more detailed account on the subject of Section 3.4 we refer the reader to Lehmann and Casella (1998).

Mallows' C_p and related techniques were already discussed in Section 1.11. Birgé and Massart (2001) proposed some extensions of the C_p-criterion in

the Gaussian sequence model. They considered definition (3.71) with penalties close to $2N\varepsilon^2$ and proved oracle inequalities showing that the estimators with the corresponding weights $\tilde{\lambda}_j = I(j \le \tilde{N})$ mimic the projection oracle $\lambda^{oracle}(\Lambda_{proj}, \cdot)$.

Kneip (1994) studied adaptation to the oracle for several examples of monotone weights. Direct minimization of $\mathcal{J}(\lambda)$ on the class of all monotone weights Λ_{mon} was considered by Beran and Dümbgen (1998). Such a minimization is numerically feasible but the resulting estimator $\tilde{\lambda}(\Lambda_{mon})$ is not proved to have optimality properties as those obtained for the block Stein estimator.

The Stein WGB estimator is not the only estimator that has the advantage of being exact adaptive on the family of Sobolev ellipsoids. A large variety of other estimators share the same property; cf. Efroimovich and Pinsker (1984), Golubev (1987), Golubev and Nussbaum (1992), Nemirovski (2000), Cavalier and Tsybakov (2001), Efromovich (2004).

The block Stein estimator with diadic blocks (cf. Example 3.7) was suggested by Donoho and Johnstone (1995), who also showed that it satisfies the oracle inequality (3.75). Brown et al. (1997) and Cai (1999) obtained similar inequalities for modifications of the Stein estimator with diadic blocks. The block Stein estimator with arbitrary blocks is introduced in Cavalier and Tsybakov (2001, 2002), in a more general form than in (3.74):

$$\tilde{\theta}_k = \begin{cases} y_k, & \text{if } k \in B_j \text{ with } j \in \mathcal{J}_0, \\ y_k \left(1 - \frac{\varepsilon^2 T_j (1+p_j)}{\|y\|^2_{(j)}}\right)_+, & \text{if } k \in B_j \text{ with } j \notin \mathcal{J}_0, \\ 0, & k > N_{\max}, \end{cases} \tag{3.101}$$

where $0 \le p_j < 1$ and $\mathcal{J}_0$ is a set of indices that can be chosen, for example, as $\mathcal{J}_0 = \{j : T_j \le 4/(1-p_j)\}$ where $T_j = \text{Card } B_j$. Such an estimator is called a *penalized block Stein estimator*. Because of the penalizing factor $(1 + p_j)$, the estimator (3.101) has fewer nonzero coefficients $\tilde{\theta}_k$ than the simple block Stein estimator (3.74), in other words it is more sparse. A major penalty choice discussed in Cavalier and Tsybakov (2001) is $p_j \sim \left(\frac{\log T_j}{T_j}\right)^{1/2}$ and this is in some sense the smallest penalty, but one can consider, for example, $p_j \sim T_j^{-\gamma}$ with $0 < \gamma < 1/2$ or other similar choices. An intuitive motivation is the following. The ratio of standard deviation to expectation for the stochastic error term corresponding to jth block is of order $T_j^{-1/2}$. Hence, to control the variability of stochastic terms, one needs a penalty that is slightly larger than $T_j^{-1/2}$. As shown in Cavalier and Tsybakov (2001), the penalized block Stein estimator is: (i) adaptive in the exact minimax sense on any ellipsoid in ℓ_2 (cf. (3.13)) with monotone nondecreasing a_j; (ii) almost sharp asymptotically minimax on other bodies such as hyperrectangles, tail-classes, Besov classes with $p \ge 2$; and (iii) attains the optimal rate of convergence (up to a logarithmic factor) on the Besov classes with $p < 2$. Cavalier and Tsybakov (2002) prove similar

results for an extension of (3.101) to the heteroscedastic sequence space model $y_k = b_k\theta_k + \varepsilon\xi_k, k = 1, 2, \ldots$, where $b_k > 0$ are known constants. This corresponds to statistical inverse problems. Sections 3.6 and 3.7 present a simplified version of some results in Cavalier and Tsybakov (2001). Section 3.8 is new, though essentially in the spirit of Brown et al. (1997). Further developments on the block Stein estimators, a survey of more recent work, and numerical studies can be found in Rigollet (2006a,b).

3.10 Exercises

Exercise 3.1 *Consider an* exponential ellipsoid:

$$\Theta = \left\{\theta = \{\theta_j\}_{j=1}^{\infty} : \sum_{j=1}^{\infty} e^{2\alpha j}\theta_j^2 \le Q\right\} \tag{3.102}$$

where $\alpha > 0$ and $Q > 0$.

(1) Give an asymptotic expression, as $\varepsilon \to 0$, for the minimax linear risk on Θ.

(2) Prove that the simple projection estimator defined by

$$\hat{\theta}_k = y_k I(k \le N^*), \quad k = 1, 2, \ldots,$$

with an appropriately chosen integer $N^* = N^*(\varepsilon)$, is an asymptotically minimax linear estimator on the ellipsoid (3.102). Therefore it shares this property with the Pinsker estimator for the same ellipsoid.

Exercise 3.2 *Suppose that we observe*

$$y_j = \theta_j + \xi_j, \quad j = 1, \ldots, d, \tag{3.103}$$

where the random variables ξ_j are i.i.d. with distribution $\mathcal{N}(0, 1)$. Consider the estimation of parameter $\theta = (\theta_1, \ldots, \theta_d)$. Take $\Theta(Q) = \{\theta \in \mathbf{R}^d : \|\theta\|^2 \le Qd\}$ with some $Q > 0$, where $\|\cdot\|$ denotes the Euclidean norm on $\mathbf{R}^d$. Define the minimax risk

$$\mathcal{R}_d^*(\Theta(Q)) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta(Q)} \mathbf{E}_\theta \left[\frac{1}{d}\|\hat{\theta} - \theta\|^2\right],$$

where $\mathbf{E}_\theta$ is the expectation with respect to the joint distribution of $(y_1, \ldots, y_d)$ satisfying (3.103). Prove that

$$\lim_{d\to\infty} \mathcal{R}_d^*(\Theta(Q)) = \frac{Q}{Q+1}\,.$$

Hint: *To obtain the lower bound on the minimax risk, take $0 < \delta < 1$ and apply the scheme of Section 3.3.2 with the prior distribution $\mathcal{N}(0, \delta Q)$ on each of the coordinates of θ.*

Exercise 3.3 *Consider the setting of Exercise 3.2.*

(1) Prove that the Stein estimator

$$\hat{\theta}_S = \left(1 - \frac{d}{\|y\|^2}\right) y,$$

as well as the positive part Stein estimator

$$\hat{\theta}_{S+} = \left(1 - \frac{d}{\|y\|^2}\right)_+ y,$$

are adaptive in the exact minimax sense over the family of classes $\{\Theta(Q), Q > 0\}$, *that is, for all* $Q > 0$,

$$\limsup_{d\to\infty} \sup_{\theta\in\Theta(Q)} \mathbf{E}_\theta \left(\frac{1}{d}\|\hat{\theta} - \theta\|^2\right) \le \frac{Q}{Q+1}$$

with $\hat{\theta} = \hat{\theta}_S$ *or* $\hat{\theta} = \hat{\theta}_{S+}$. *(Here, we deal with adaptation at an unknown radius* Q *of the ball* $\Theta(Q)$.*)* Hint: *Apply Lemma 3.10.*

(2) Prove that the linear minimax estimator on $\Theta(Q)$ *(the Pinsker estimator) is inadmissible on any class* $\Theta(Q')$ *such that* $0 < Q' < Q$ *for all* $d > d_1$ *where* d_1 *depends only on* Q *and* Q'.

Exercise 3.4 *Consider Model 1 of Section 3.4. Let* $\tilde{\tau} > 0$.

(1) Show that the hard thresholding estimator $\hat{\theta}_{HT}$ *with the components*

$$\hat{\theta}_{j,HT} = I(|y_j| > \tilde{\tau}) y_j, \quad j = 1, \dots, d,$$

is a solution of the minimization problem

$$\min_{\theta\in\mathbf{R}^d} \left\{ \sum_{j=1}^d (y_j - \theta_j)^2 + \tilde{\tau}^2 \sum_{j=1}^d I(\theta_j \neq 0) \right\}.$$

(2) Show that the soft thresholding estimator $\hat{\theta}_{ST}$ *with the components*

$$\hat{\theta}_{j,ST} = \left(1 - \frac{\tilde{\tau}}{|y_j|}\right)_+ y_j, \quad j = 1, \dots, d,$$

is a solution of the minimization problem

$$\min_{\theta\in\mathbf{R}^d} \left\{ \sum_{j=1}^d (y_j - \theta_j)^2 + 2\tilde{\tau} \sum_{j=1}^d |\theta_j| \right\}.$$

Exercise 3.5 *Consider Model 1 of Section 3.4. Using Stein's lemma, show that the statistic*

$$\mathcal{J}_1(\tilde{\tau}) = \sum_{j=1}^{d} (2\varepsilon^2 + \tilde{\tau}^2 - y_j^2)\, I(|y_j| \geq \tilde{\tau})$$

is an unbiased estimator of the risk of the soft thresholding estimator $\hat{\theta}_{ST}$*, up to the additive term* $\|\theta\|^2$ *that does not depend on* $\tilde{\tau}$*:*

$$\mathbf{E}_\theta \left[\mathcal{J}_1(\tilde{\tau})\right] = \mathbf{E}_\theta \|\hat{\theta}_{ST} - \theta\|^2 - \|\theta\|^2.$$

Based on this, suggest a data-driven choice of the threshold $\tilde{\tau}$*.*

Exercise 3.6 *Consider Model 1 of Section 3.4. Let* $\tau > 0$*.*
(1) Show that the global hard thresholding estimator

$$\hat{\theta}_{GHT} = I(\|y\| > \tau)\, y$$

is a solution of the minimization problem

$$\min_{\theta \in \mathbf{R}^d} \left\{ \sum_{j=1}^{d} (y_j - \theta_j)^2 + \tau^2 I(\|\theta\| \neq 0) \right\}.$$

(2) Show that the global soft thresholding estimator

$$\hat{\theta}_{GST} = \left(1 - \frac{\tau}{\|y\|}\right)_+ y$$

is a solution of the minimization problem

$$\min_{\theta \in \mathbf{R}^d} \left\{ \sum_{j=1}^{d} (y_j - \theta_j)^2 + 2\tau \|\theta\| \right\}.$$

Exercise 3.7 *Consider first Model 1 of Section 3.4. Define a global hard thresholding estimator of the vector* $\theta = (\theta_1, \ldots, \theta_d)$ *as follows:*

$$\hat{\theta} = I(\|y\| > \tau)\, y,$$

where $\tau = 2\varepsilon\sqrt{d}$*.*
(1) Prove that for $\|\theta\|^2 \leq \varepsilon^2 d/4$ *we have*

$$\mathbf{P}_\theta(\hat{\theta} = y) \leq \exp(-c_0 d),$$

where $c_0 > 0$ *is an absolute constant.* Hint: *Use the following inequality (cf. Lemma 3.5):*

$$\mathbf{P}\left(\sum_{k=1}^{d} (\xi_k^2 - 1) \geq td\right) \leq \exp\left(-\frac{t^2 d}{8}\right), \quad \forall\, 0 < t \leq 1.$$

(2) Based on (1) prove that

$$\mathbf{E}_\theta\|\hat\theta-\theta\|^2 \le \|\theta\|^2 + c_1\varepsilon^2 d\exp(-c_0 d/2)$$

for $\|\theta\|^2 \le \varepsilon^2 d/4$ *with an absolute constant* $c_1>0$.

(3) Show that, for all $\theta\in\mathbf{R}^d$,

$$\mathbf{E}_\theta\|\hat\theta-\theta\|^2 \le 9\varepsilon^2 d.$$

(4) Combine (2) and (3) to prove the oracle inequality

$$\mathbf{E}_\theta\|\hat\theta-\theta\|^2 \le c_2\frac{d\varepsilon^2\|\theta\|^2}{d\varepsilon^2+\|\theta\|^2} + c_1\varepsilon^2 d\exp(-c_0 d/2), \quad \forall\ \theta\in\mathbf{R}^d,$$

where $c_2>0$ *is an absolute constant.* Hint: $\min(a,b)\le 2ab/(a+b)$ *for all* $a\ge 0, b>0$.

(5) We switch now to the Gaussian sequence model (3.10). Introduce the blocks B_j *of size* $\operatorname{card}(B_j)=j$ *and define the estimators*

$$\tilde\theta_k = I(\|y\|_{(j)}>\tau_j)y_k \quad \text{for}\ \ k\in B_j,\ \ j=1,2,\dots,J,$$

where $\tau_j=2\varepsilon\sqrt{j}$, $J\ge 1/\varepsilon^2$, *and* $\tilde\theta_k=0$ *for* $k>\sum_{j=1}^J \operatorname{card}(B_j)$. *Set* $\tilde\theta=(\tilde\theta_1,\tilde\theta_2,\dots)$. *Prove the oracle inequality*

$$\mathbf{E}_\theta\|\tilde\theta-\theta\|^2 \le c_3\min_{\lambda\in\Lambda_{mon}} R(\lambda,\theta) + c_4\varepsilon^2, \quad \forall\ \theta\in\ell^2(\mathbf{N}),$$

where $c_3>0$ *and* $c_4>0$ *are absolute constants.*

(6) Show that the estimator $\tilde\theta$ *defined in (5) is adaptive in the minimax sense on the family of classes* $\{\Theta(\beta,Q),\beta>0,Q>0\}$ *(cf. Definition 3.9).*

Appendix

This Appendix contains proofs of some auxiliary results used in Chapters 1–3. In order to make reading more feasible, we also reproduce here the statements of the results.

Lemma A.1 (Generalized Minkowski inequality). *For any Borel function g on $\mathbf{R} \times \mathbf{R}$, we have*

$$\int \Big(\int g(u,x)du \Big)^2 dx \le \bigg(\int \Big(\int g^2(u,x)dx \Big)^{1/2} du \bigg)^2 .$$

Proof. It suffices to assume that

$$\int \Big(\int g^2(u,x)dx \Big)^{1/2} du \triangleq C_g < \infty,$$

since otherwise the result of the lemma is trivial. Put

$$S(x) = \int |g(u,x)|du.$$

For all $f \in L_2(\mathbf{R})$,

$$\begin{aligned} \Big| \int S(x)f(x)dx \Big| &\le \int |f(x)| \int |g(u,x)|du \; dx \\ &= \int du \int |f(x)||g(u,x)|dx \quad \text{(Tonelli–Fubini)} \\ &\le \|f\|_2 \int \Big(\int g^2(u,x)dx \Big)^{1/2} du \quad \text{(Cauchy–Schwarz)} \\ &= C_g \|f\|_2 \end{aligned}$$

with $\|f\|_2 = (\int f^2(x)dx)^{1/2}$. This implies that the linear functional $f \mapsto \int S(x)f(x)dx$ is continuous on $L_2(\mathbf{R})$. Then $S \in L_2(\mathbf{R})$ and

$$\|S\|_2 = \sup_{f \neq 0} \frac{\int Sf}{\|f\|_2} \leq C_g$$

implying the required result. ∎

Lemma A.2 *If $f \in L_2(\mathbf{R})$, then*

$$\lim_{\delta \to 0} \sup_{|t| \leq \delta} \int (f(x+t) - f(x))^2 dx = 0.$$

Proof. Denote by Φ the Fourier transform of f. Then for $t \in \mathbf{R}$ the Fourier transform of $f(\cdot + t)$ is the function $\omega \mapsto \Phi(\omega) e^{\mathrm{i}t\omega}$. By the Plancherel theorem, for all $t \in \mathbf{R}$,

$$\begin{aligned} \int (f(x+t) - f(x))^2 dx &= \int |\Phi(\omega)|^2 |e^{\mathrm{i}t\omega} - 1|^2 d\omega \\ &= 4 \int |\Phi(\omega)|^2 \sin^2(\omega t/2) d\omega. \end{aligned}$$

Let $0 < \delta < \pi^2$ and $|t| \leq \delta$. Then $\sin^2(\omega t/2) \leq \sin^2(\sqrt{\delta}/2)$ whenever $|\omega| \leq t^{-1/2}$, and we get

$$\begin{aligned} \int (f(x+t) - f(x))^2 dx &\leq 4 \left[\int_{|\omega| \leq t^{-1/2}} |\Phi(\omega)|^2 \sin^2(\omega t/2) d\omega \right. \\ &\qquad \left. + \int_{|\omega| > t^{-1/2}} |\Phi(\omega)|^2 d\omega \right] \\ &\leq 4 \left[\sin^2(\sqrt{\delta}/2) \int |\Phi(\omega)|^2 d\omega \right. \\ &\qquad \left. + \int_{|\omega| > \delta^{-1/2}} |\Phi(\omega)|^2 d\omega \right] \\ &= o(1) \qquad \text{as } \delta \to 0, \end{aligned}$$

since $\Phi \in L_2(\mathbf{R})$. ∎

Proposition A.1 *Assume that:*

(i) the function K is a kernel of order 1 satisfying the conditions

$$\int K^2(u) du < \infty, \qquad \int u^2 |K(u)| du < \infty, \qquad S_K \triangleq \int u^2 K(u) du \neq 0;$$

(ii) the density p is differentiable on $\mathbf{R}$, the first derivative p' is absolutely continuous on $\mathbf{R}$ and the second derivative satisfies

$$\int (p''(x))^2 dx < \infty.$$

Then for all $n \geq 1$ the mean integrated squared error of the kernel estimator $\hat{p}_n$ satisfies

$$\begin{aligned}\mathrm{MISE} &\equiv \mathbf{E}_p \int (\hat{p}_n(x) - p(x))^2 dx \\ &= \left[\frac{1}{nh}\int K^2(u)du + \frac{h^4}{4}\, S_K^2 \int (p''(x))^2 dx\right](1 + o(1)),\end{aligned}$$

where $o(1)$ is independent of n and tends to 0 as $h \to 0$.

Proof. (i) First, consider the variance term $\int \sigma^2(x)dx$. Using (1.6), we obtain

$$\int \sigma^2(x)dx = \frac{1}{nh}\int K^2(u)du - \frac{1}{nh^2}\int \left(\int K\left(\frac{z-x}{h}\right)p(z)dz\right)^2 dx.$$

The assumptions of the proposition imply that the probability density p is uniformly bounded on $\mathbf{R}$. Therefore $p \in L_2(\mathbf{R})$. By the Cauchy–Schwarz inequality and the Tonelli–Fubini theorem, we obtain

$$\begin{aligned}&\int \left(\int K\left(\frac{z-x}{h}\right)p(z)dz\right)^2 dx \\ &\qquad \leq \int \left[\int \left|K\left(\frac{t-x}{h}\right)\right| dt\right] \int \left|K\left(\frac{z-x}{h}\right)\right| p^2(z)dzdx \\ &\qquad = h^2 \left(\int |K(u)|du\right)^2 \int p^2(z)dz.\end{aligned}$$

Therefore the variance term satisfies

$$\int \sigma^2(x)dx = \frac{1}{nh}\int K^2(u)du\ (1 + o(1)) \tag{A.1}$$

where $o(1)$ is independent of n and tends to 0 as $h \to 0$. (ii) We now study the bias term $\int b^2(x)dx$. From (1.18) with $\ell = 2$ we get

$$b(x) = h^2 \int u^2 K(u)\Big[\int_0^1 (1-\tau)p''(x + \tau uh)d\tau\Big]du. \tag{A.2}$$

Define

$$\begin{aligned}b^* &= \frac{h^4}{4}\Big(\int u^2 K(u)du\Big)^2 \int (p''(x))^2 dx \\ &= h^4 \int \Big[\int u^2 K(u)\Big(\int_0^1 (1-\tau)p''(x)d\tau\Big)du\Big]^2 dx\end{aligned}$$

and observe that

$$\Big|\int b^2(x)dx - b^*\Big| = h^4\Big|\int A_1(x)A_2(x)dx\Big| \tag{A.3}$$
$$\leq h^4\Big(\int A_1^2(x)dx\Big)^{1/2}\Big(\int A_2^2(x)dx\Big)^{1/2}$$

with

$$A_1(x) \stackrel{\triangle}{=} \int u^2K(u)\Big(\int_0^1 (p''(x+\tau uh) - p''(x))(1-\tau)d\tau\Big)du,$$

and

$$A_2(x) \stackrel{\triangle}{=} \int u^2K(u)\Big(\int_0^1 (p''(x+\tau uh) + p''(x))(1-\tau)d\tau\Big)du.$$

By a successive application of the generalized Minkowski inequality, the Cauchy–Schwarz inequality and the Tonelli–Fubini theorem, we obtain

$$\int\left(\int u^2|K(u)|\Big[\int_0^1 |p''(x+\tau uh)|(1-\tau)d\tau\Big]du\right)^2 dx \tag{A.4}$$
$$\leq \left(\int u^2|K(u)|\left(\int\Big[\int_0^1 |p''(x+\tau uh)|(1-\tau)d\tau\Big]^2 dx\right)^{1/2} du\right)^2$$
$$\leq \left(\int u^2|K(u)|\times\right.$$
$$\left.\left(\int\int_0^1 (p''(x+\tau uh))^2(1-\tau)d\tau dx\int_0^1(1-\tau)d\tau\right)^{1/2} du\right)^2$$
$$= \frac{1}{4}\Big(\int u^2|K(u)|du\Big)^2\int (p''(x))^2dx < \infty.$$

This implies that the integral $\int A_2^2(x)dx$ is bounded by a constant independent of h. By the same argument as in (A.4) and by dividing the domain of integration into two parts, $|u| \leq h^{-1/2}$ and $|u| > h^{-1/2}$, we get

$$\int A_1^2(x)dx \tag{A.5}$$
$$\leq \left(\int u^2|K(u)|\left(\int\Big[\int_0^1 |p''(x+\tau uh) - p''(x)|d\tau\Big]^2 dx\right)^{1/2} du\right)^2$$
$$\leq \left(\int u^2|K(u)|\left(\int\int_0^1 (p''(x+\tau uh) - p''(x))^2 d\tau dx\right)^{1/2} du\right)^2$$
$$\leq \left(\sup_{|u|\leq h^{-1/2}}\left[\int_0^1\int (p''(x+\tau uh) - p''(x))^2 dxd\tau\right]^{1/2}\int u^2|K(u)|du\right.$$
$$\left. + 2\Big[\int (p''(x))^2dx\Big]^{1/2}\int_{|u|>h^{-1/2}} u^2|K(u)|du\right)^2.$$

By Lemma A.2, we have

$$\sup_{|u|\le h^{-1/2}} \int_0^1 \int (p''(x+\tau uh) - p''(x))^2 dx d\tau \tag{A.6}$$
$$\le \sup_{|t|\le h^{1/2}} \int (p''(x+t) - p''(x))^2 dx = o(1)$$

as $h \to 0$. From (A.3)–(A.6) we finally obtain

$$\int b^2(x)dx = b^*(1+o(1)) \quad \text{as } h \to 0.$$

This relation combined with (A.1) proves the proposition. ∎

Proposition A.2 *Let assumption (ii) of Proposition A.1 be satisfied and let K be a kernel of order 2 such that*

$$\int K^2(u)du < \infty.$$

Then, for any $\varepsilon > 0$, the kernel estimator $\hat{p}_n$ with bandwidth

$$h = n^{-1/5}\varepsilon^{-1} \int K^2(u)du$$

satisfies

$$\limsup_{n\to\infty} n^{4/5} \mathbf{E}_p \int (\hat{p}_n(x) - p(x))^2 dx \le \varepsilon. \tag{A.7}$$

The same is true for the positive part estimator $\hat{p}_n^+ = \max(0, \hat{p}_n)$:

$$\limsup_{n\to\infty} n^{4/5} \mathbf{E}_p \int (\hat{p}_n^+(x) - p(x))^2 dx \le \varepsilon. \tag{A.8}$$

Proof. Since K is a kernel of order 2, we have $\int u^2 K(u)du = 0$. Under this assumption, following the proof of Proposition A.1 we get $b^* = o(h^4)$, and therefore $\int b^2(x)dx = o(h^4)$. Since the variance term satisfies (A.1), we obtain

$$\mathbf{E}_p \int (\hat{p}_n(x) - p(x))^2 dx = \frac{1}{nh} \int K^2(u)du \ (1+o(1)) + o(h^4).$$

This implies (A.7) in view of the choice of h. Finally, (A.8) follows from (A.7) and (1.10). ∎

Lemma A.3 *Let β be an integer, $\beta \geq 1$, $L > 0$, and let $\{\varphi_j\}_{j=1}^{\infty}$ be the trigonometric basis. Then the function $f = \sum_{j=1}^{\infty} \theta_j \varphi_j$ belongs to $W^{per}(\beta, L)$ if and only if the vector θ of the Fourier coefficients of f belongs to the ellipsoid in $\ell^2(\mathbf{N})$ defined by*

$$\Theta(\beta, Q) = \Big\{\theta \in \ell^2(\mathbf{N}) : \sum_{j=1}^{\infty} a_j^2 \theta_j^2 \leq Q\Big\},$$

where $Q = L^2/\pi^{2\beta}$ and a_j are given by (1.90).

Proof. *Necessity.* First, we prove that if $f \in W^{per}(\beta, L)$, then $\theta \in \Theta(\beta, Q)$. For $f \in W^{per}(\beta, L)$ and $j = 1, \ldots, \beta$, define the Fourier coefficients of $f^{(j)}$ with respect to the trigonometric basis:

$$\begin{aligned}
s_1(j) &\triangleq \int_0^1 f^{(j)}(t)dt = f^{(j-1)}(1) - f^{(j-1)}(0) = 0,\\
s_{2k}(j) &\triangleq \sqrt{2}\int_0^1 f^{(j)}(t)\cos(2\pi kt)dt,\\
s_{2k+1}(j) &\triangleq \sqrt{2}\int_0^1 f^{(j)}(t)\sin(2\pi kt)dt, \quad \text{for } k = 1, 2, \ldots,
\end{aligned}$$

and put $s_{2k}(0) \triangleq \theta_{2k}$, $s_{2k+1}(0) \triangleq \theta_{2k+1}$. Integrating by parts we obtain

$$\begin{aligned}
s_{2k}(\beta) &= \sqrt{2} f^{(\beta-1)}(t)\cos(2\pi kt)\Big|_0^1 \\
&\quad + (2\pi k)\sqrt{2}\int_0^1 f^{(\beta-1)}(t)\sin(2\pi kt)dt \\
&= (2\pi k)\sqrt{2}\int_0^1 f^{(\beta-1)}(t)\sin(2\pi kt)dt \\
&= (2\pi k)s_{2k+1}(\beta - 1)
\end{aligned} \tag{A.9}$$

and

$$\begin{aligned}
s_{2k+1}(\beta) &= -(2\pi k)\sqrt{2}\int_0^1 f^{(\beta-1)}(t)\cos(2\pi kt)dt \\
&= -(2\pi k)s_{2k}(\beta - 1).
\end{aligned} \tag{A.10}$$

In particular, $s_{2k}^2(\beta) + s_{2k+1}^2(\beta) = (2\pi k)^2(s_{2k}^2(\beta - 1) + s_{2k+1}^2(\beta - 1))$. By recurrence, we find

$$s_{2k}^2(\beta) + s_{2k+1}^2(\beta) = (2\pi k)^{2\beta}(\theta_{2k}^2 + \theta_{2k+1}^2), \quad \text{for } k = 1, 2, \ldots. \tag{A.11}$$

Next, note that

$$\sum_{k=1}^{\infty}(2\pi k)^{2\beta}(\theta_{2k}^2+\theta_{2k+1}^2)=\pi^{2\beta}\sum_{j=1}^{\infty}a_j^2\,\theta_j^2, \tag{A.12}$$

implying, by the Parseval equality,

$$\int_0^1 (f^{(\beta)}(t))^2dt=\sum_{k=1}^{\infty}(s_{2k}^2(\beta)+s_{2k+1}^2(\beta))=\pi^{2\beta}\sum_{j=1}^{\infty}a_j^2\,\theta_j^2.$$

Since $\int_0^1(f^{(\beta)}(t))^2dt\le L^2$, we obtain $\theta\in\Theta(\beta,Q)$.

Sufficiency. Suppose now that $\theta\in\Theta(\beta,Q)$ and let us prove that the function f with the sequence θ of Fourier coefficients satisfies $f\in W^{per}(\beta,L)$. Observe first that if $\theta\in\Theta(\beta,Q)$, we have for $j=0,1,\dots,\beta-1$,

$$\begin{aligned}\sum_{k=1}^{\infty}k^j(|\theta_{2k}|+|\theta_{2k+1}|)&\le\sum_{k=1}^{\infty}k^{\beta-1}(|\theta_{2k}|+|\theta_{2k+1}|)\\&\le\left(2\sum_{k=1}^{\infty}k^{2\beta}(\theta_{2k}^2+\theta_{2k+1}^2)\right)^{1/2}\left(\sum_{k=1}^{\infty}k^{-2}\right)^{1/2}<\infty.\end{aligned}$$

This implies that the series $f(x)=\sum_{j=1}^{\infty}\theta_j\varphi_j(x)$, as well as its derivatives

$$f^{(j)}(x)=\sum_{k=1}^{\infty}(2\pi k)^j(\theta_{2k}\tilde\varphi_{2k,j}(x)+\theta_{2k+1}\tilde\varphi_{2k+1,j}(x)),$$

for $j=1,\dots,\beta-1$, converge uniformly in $x\in[0,1]$. Here

$$\tilde\varphi_{2k,j}(x)=\sqrt{2}\frac{d^j}{du^j}(\cos u)\Big|_{u=2\pi kx},\quad \tilde\varphi_{2k+1,j}(x)=\sqrt{2}\frac{d^j}{du^j}(\sin u)\Big|_{u=2\pi kx}.$$

Since the functions $\tilde\varphi_{m,j}$ are periodic, we have $f^{(j)}(0)=f^{(j)}(1)$ for $j=0,1,\dots,\beta-1$.

Now let $\{s_m(\beta-1)\}_{m=1}^{\infty}$ be the Fourier coefficients of the function $f^{(\beta-1)}$. Define $\{s_m(\beta)\}_{m=1}^{\infty}$ from $\{s_m(\beta-1)\}_{m=1}^{\infty}$ by (A.9) and (A.10) if $m\ge 2$ and put $s_1(\beta)=0$. It follows from the Parseval equality and (A.11)–(A.12) that the function $g\in L_2[0,1]$ defined by the sequence of Fourier coefficients $\{s_m(\beta)\}_{m=1}^{\infty}$ satisfies

$$\int_0^1 g^2(x)dx=\sum_{m=1}^{\infty}s_m^2(\beta)\le\pi^{2\beta}Q=L^2$$

whenever $\theta\in\Theta(\beta,Q)$. Let us now show that g equals the derivative of the function $f^{(\beta-1)}$ almost everywhere. Indeed, since the Fourier series of any function in $L_2[0,1]$ is termwise integrable on any interval $[a,b]\subseteq[0,1]$, we can write

$$\begin{aligned}\int_a^b g(x)dx &\equiv \int_a^b \sum_{k=1}^{\infty}(s_{2k}(\beta)\sqrt{2}\cos(2\pi kx) + s_{2k+1}(\beta)\sqrt{2}\sin(2\pi kx))dx \\ &= \sum_{k=1}^{\infty}(2\pi k)^{-1}(s_{2k}(\beta)\sqrt{2}\sin(2\pi kx) - s_{2k+1}(\beta)\sqrt{2}\cos(2\pi kx))\Big|_a^b \\ &= \sum_{k=1}^{\infty}(s_{2k}(\beta-1)\sqrt{2}\sin(2\pi kx) + s_{2k+1}(\beta-1)\sqrt{2}\cos(2\pi kx))\Big|_a^b \\ &= f^{(\beta-1)}(b) - f^{(\beta-1)}(a).\end{aligned}$$

This proves that $f^{(\beta-1)}$ is absolutely continuous on $[0,1]$ and that its derivative $f^{(\beta)}$ is equal to g almost everywhere on $[0,1]$ with respect to the Lebesgue measure. Thus, $\int_0^1 (f^{(\beta)})^2 \le L^2$, completing the proof. ∎

Lemma A.4 (Hoeffding's inequality). *Let $Z_1, \ldots, Z_m$ be independent random variables such that $a_i \le Z_i \le b_i$. Then for all $t > 0$*

$$\mathbf{P}\left(\sum_{i=1}^{m}(Z_i - \mathbf{E}(Z_i)) \ge t\right) \le \exp\left(-\frac{2t^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right).$$

Proof. It is sufficient to study the case where $\mathbf{E}(Z_i) = 0,\ i = 1, \ldots, m$. By the Markov inequality, for all $v > 0$,

$$\begin{aligned}\mathbf{P}\left(\sum_{i=1}^{m} Z_i \ge t\right) &\le \exp(-vt)\mathbf{E}\left[\exp\left(v\sum_{i=1}^{m} Z_i\right)\right] \qquad (A.13)\\ &= e^{-vt}\prod_{i=1}^{m}\mathbf{E}\left[e^{vZ_i}\right].\end{aligned}$$

Note that

$$\mathbf{E}\left[e^{vZ_i}\right] \le \exp\left(\frac{v^2(b_i - a_i)^2}{8}\right). \qquad (A.14)$$

Indeed, since the exponential function is convex, we have

$$e^{vx} \le \frac{b_i - x}{b_i - a_i}e^{va_i} + \frac{x - a_i}{b_i - a_i}e^{vb_i}, \quad a_i \le x \le b_i.$$

Taking the expectations and using the fact that $\mathbf{E}(Z_i) = 0$, we obtain

$$\begin{aligned}\mathbf{E}\left[e^{vZ_i}\right] &\le \frac{b_i}{b_i - a_i}e^{va_i} - \frac{a_i}{b_i - a_i}e^{vb_i} \\ &= (1 - s + se^{v(b_i - a_i)})e^{-sv(b_i - a_i)} \triangleq e^{g(u)},\end{aligned}$$

where $u = v(b_i - a_i)$, $s = -a_i/(b_i - a_i)$ and $g(u) = -su + \log(1 - s + se^u)$. It is easy to see that $g(0) = g'(0) = 0$ and $g''(u) \le 1/4$ for all u. By expanding g in Taylor series, we obtain, for some $0 \le \tau \le 1$,

$$g(u) = u^2 g''(\tau u)/2 \le u^2/8 = v^2(b_i - a_i)^2/8$$

implying (A.14). From (A.13) and (A.14) we get

$$\mathbf{P}\left(\sum_{i=1}^m Z_i \ge t\right) \le e^{-vt} \prod_{i=1}^m \exp\left(\frac{v^2(b_i - a_i)^2}{8}\right)$$
$$= \exp\left(-\frac{2t^2}{\sum_{i=1}^m (b_i - a_i)^2}\right)$$

if we take $v = 4t/\sum_{i=1}^m (b_i - a_i)^2$. ∎

Denote by $\mathcal{U}$ the σ-algebra of subsets of $C[0,1]$ generated by cylindric sets $\{Y(t_1) \in B_1, \ldots, Y(t_m) \in B_m\}$, where B_j are Borel sets in $\mathbf{R}$. Let $\mathbf{P}_f$ be the probability measure on $(C[0,1], \mathcal{U})$ generated by the process $\mathbf{X} = \{Y(t), 0 \le t \le 1\}$ satisfying the Gaussian white noise model (3.1) for a function $f \in L_2[0,1]$. In particular, $\mathbf{P}_0$ is the measure corresponding to the function $f \equiv 0$. Denote by $\mathbf{E}_f$ and $\mathbf{E}_0$ the expectations with respect to $\mathbf{P}_f$ and $\mathbf{P}_0$.

Lemma A.5 (Girsanov's theorem). *The measure $\mathbf{P}_f$ is absolutely continuous with respect to $\mathbf{P}_0$ and the Radon–Nikodym derivative satisfies*

$$\frac{d\mathbf{P}_f}{d\mathbf{P}_0}(\mathbf{X}) = \exp\left(\varepsilon^{-2} \int_0^1 f(t)dY(t) - \frac{\varepsilon^{-2}}{2} \int_0^1 f^2(t)dt\right).$$

In particular, for any measurable function $F : (C[0,1], \mathcal{U}) \to (\mathbf{R}, \mathcal{B}(\mathbf{R}))$,

$$\mathbf{E}_f[F(\mathbf{X})] = \mathbf{E}_0\left[F(\mathbf{X}) \exp\left(\varepsilon^{-2} \int_0^1 f(t)dY(t) - \frac{\varepsilon^{-2}}{2} \int_0^1 f^2(t)dt\right)\right].$$

The proof of this result can be found, for example, in Ibragimov and Hasminskii (1981), Appendix 2.

Lemma A.6 *Consider Model 1 of Section 3.4. For a finite constant $c > 0$, consider the estimator*

$$\tilde{\theta} = g(y)y$$

with

$$g(y) = 1 - \frac{c}{\|y\|^2}$$

and the estimator

$$\tilde{\theta}_+ = \left(1 - \frac{c}{\|y\|^2}\right)_+ y.$$

Then, for all $\theta \in \mathbf{R}^d$,

$$\mathbf{E}_\theta \|\tilde{\theta}_+ - \theta\|^2 < \mathbf{E}_\theta \|\tilde{\theta} - \theta\|^2.$$

PROOF. Observe that $\mathbf{E}_\theta\|\tilde{\theta}_+ - \theta\|^2 < \infty$ for all $\theta \in \mathbf{R}^d$. It is sufficient to consider the case of $d \geq 3$, since for $d = 1$ and $d = 2$ we have $\mathbf{E}_\theta(\|y\|^{-2}) = +\infty$ (see the proof of Lemma 3.7) and $\mathbf{E}_\theta\|\tilde{\theta} - \theta\|^2 = +\infty,\ \forall\ \theta \in \mathbf{R}^d$. If $d \geq 3$, the expectation $\mathbf{E}_\theta\|\tilde{\theta} - \theta\|^2$ is finite by Lemma 3.7.

Set for brevity $g = g(y)$ and write

$$\mathbf{E}_\theta\|\tilde{\theta} - \theta\|^2 = \mathbf{E}_\theta\left[g^2\|y\|^2 - 2(\theta, y)g + \|\theta\|^2\right],$$

where (θ, y) is the standard scalar product of θ and y in $\mathbf{R}^d$. Then

$$\begin{aligned}\mathbf{E}_\theta\|\tilde{\theta}_+ - \theta\|^2 &= \mathbf{E}_\theta\|ygI(g > 0) - \theta\|^2 \\ &= \mathbf{E}_\theta\left[g^2\|y\|^2 I(g > 0) - 2(\theta, y)gI(g > 0) + \|\theta\|^2\right].\end{aligned}$$

Therefore

$$\mathbf{E}_\theta\|\tilde{\theta} - \theta\|^2 - \mathbf{E}_\theta\|\tilde{\theta}_+ - \theta\|^2 = \mathbf{E}_\theta\left[g^2\|y\|^2 I(g \leq 0) - 2(\theta, y)gI(g \leq 0)\right].$$

If $\theta = 0$, the lemma is proved since the right hand side is positive. Indeed, the definition of Model 1 implies that $\mathbf{E}_\theta\left[g^2\|y\|^2 I(g \leq 0)\right]$ is the integral of a positive function on a set of nonzero Lebesgue measure.

Let now $\theta \neq 0$. Without loss of generality suppose that $\theta_1 \neq 0$. To prove the lemma it is sufficient to show that

$$-\mathbf{E}_\theta[(\theta, y)gI(g \leq 0)] > 0. \tag{A.15}$$

This inequality will be proved if we show that

$$-\mathbf{E}_\theta[\theta_i y_i gI(g \leq 0)] > 0 \quad \text{for all } i \in \{1, \ldots, d\} \text{ such that } \theta_i \neq 0. \tag{A.16}$$

Our aim now is to show (A.16). It is sufficient to do this for $i = 1$. Apply conditional expectations to obtain

$$\begin{aligned}-\mathbf{E}_\theta[\theta_1 y_1 gI(g \leq 0)] &= -\mathbf{E}_\theta\Big[\theta_1 \mathbf{E}_\theta(y_1 gI(g \leq 0)|\, y_1^2)\Big] \\ &= \mathbf{E}_\theta\Big[\,\theta_1 \mathbf{E}_\theta(y_1|\, y_1^2)\, \mathbf{E}_\theta(|g|I(g \leq 0)|\, y_1^2)\Big]. \end{aligned} \tag{A.17}$$

Let us calculate $\mathbf{E}_\theta(y_1|\, y_1^2)$. It is easy to see that for all $a \geq 0$

$$\mathbf{E}_\theta(y_1|\, y_1^2 = a^2) = a\mathbf{E}_\theta\Big[\operatorname{sgn}(y_1)|\, |y_1| = a\Big],$$

where $\operatorname{sgn}(y_1) = I(y_1 \geq 0) - I(y_1 < 0)$. For all $\delta > 0$ and $a \geq 0$, put

$$\begin{aligned}E(\delta) &\stackrel{\triangle}{=} \mathbf{E}_\theta\Big[\operatorname{sgn}(y_1)I(a \leq |y_1| \leq a + \delta)\Big] \\ &= \int_a^{a+\delta} \mathbf{E}_\theta\Big[\operatorname{sgn}(y_1)|\ |y_1| = t\Big]p(t)dt,\end{aligned} \tag{A.18}$$

where $p(\cdot)$ is the density of $|y_1|$. Since $y_1 = \theta_1 + \varepsilon\xi_1$ with $\xi_1 \sim \mathcal{N}(0,1)$, we have

$$p(t) = \frac{1}{\varepsilon}\varphi\left(\frac{t-\theta_1}{\varepsilon}\right) + \frac{1}{\varepsilon}\varphi\left(\frac{-t-\theta_1}{\varepsilon}\right), \quad t \geq 0, \tag{A.19}$$

where φ is the density of $\mathcal{N}(0,1)$. On the other hand,

$$E(\delta) = P_\theta(a \leq y_1 \leq a+\delta) - P_\theta(-a-\delta \leq y_1 \leq -a).$$

Then

$$E'(0) = p_{y_1}(a) - p_{y_1}(-a), \tag{A.20}$$

where $p_{y_1}(\cdot)$ is the density of the distribution of y_1, that is

$$p_{y_1}(a) = \frac{1}{\varepsilon}\varphi\left(\frac{a-\theta_1}{\varepsilon}\right).$$

By differentiating (A.18) with respect to δ at the point $\delta = 0$, we obtain, in view of (A.19) and (A.20),

$$\begin{aligned}\mathbf{E}_\theta\Big[\operatorname{sgn}(y_1)|\ |y_1| = a\Big] &= \frac{E'(0)}{p(a)} \\ &= \frac{\varphi\left(\frac{a-\theta_1}{\varepsilon}\right) - \varphi\left(\frac{-a-\theta_1}{\varepsilon}\right)}{\varphi\left(\frac{a-\theta_1}{\varepsilon}\right) + \varphi\left(\frac{-a-\theta_1}{\varepsilon}\right)} = \tanh(a\theta_1\varepsilon^{-2}).\end{aligned}$$

Therefore, for all $a > 0$,

$$\theta_1 \mathbf{E}_\theta(y_1 | y_1^2 = a^2) = a\theta_1 \tanh(a\theta_1\varepsilon^{-2}) > 0, \tag{A.21}$$

since $u\tanh(u) > 0$ for all $u \neq 0$. By (A.17) and (A.21), we obtain

$$\begin{aligned}-\mathbf{E}_\theta[\theta_1 y_1 g I(g \leq 0)] &= \mathbf{E}_\theta\Big[\,|y_1|\theta_1 \tanh(|y_1|\theta_1\varepsilon^{-2})\,\mathbf{E}_\theta(|g|I(g \leq 0)|\, y_1^2)\Big] \\ &= \mathbf{E}_\theta\Big[\,I(|y_1| < \sqrt{c})|y_1|\theta_1 \tanh(|y_1|\theta_1\varepsilon^{-2})\,\mathbf{E}_\theta(|g|I(g \leq 0)|\, y_1^2)\Big]. \end{aligned} \tag{A.22}$$

Using the definition of Model 1 it is easy to show that for $0 < a < \sqrt{c}$ we have $\mathbf{E}_\theta(|g|I(g \leq 0)|\, y_1^2 = a^2) > 0$, since we integrate a positive function on a set of nonzero Lebesgue measure. This remark combined with formulas (A.21) and (A.22) implies (A.16) and thus proves the lemma. ∎

Bibliography

1. Akaike, H. (1954) An approximation to the density function. *Ann. Inst. Statist. Math.*, **6**, 127-132.
2. Akaike, H. (1969) Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.*, **21**, 243-247.
3. Akaike, H. (1974) A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, **19**, 716-723.
4. Assouad, P. (1983) Deux remarques sur l'estimation. *C.R. Acad. Sci. Paris, sér. I*, **296**, 1021-1024.
5. Baraud, Y. (2002) Model selection for regression on a random design. *ESAIM: Probability and Statistics*, **7**, 127-146.
6. Barron, A. (1986) Entropy and the central limit theorem. *Annals of Probability*, **14**, 336-342.
7. Bartlett, M.S. (1963) Statistical estimation of density functions. *Sankhyā, Ser. A*, **25**, 245-254.
8. Belitser, E.N. and Levit, B.Ya. (1995) On minimax filtering on ellipsoids. *Mathematical Methods of Statistics*, **4**, 259-273.
9. Beran, R. and Dümbgen, L. (1998) Modulation estimators and confidence sets. *Annals of Statistics*, **26**, 1826-1856.
10. Bickel, P.J. and Ritov,Y. (1988) Estimating integrated squared density derivatives: Sharp best order of convergence estimates. *Sankhyā, Ser. A*, **50**, 381-393.
11. Bickel, P.J., Ritov, Y. and Tsybakov, A.B. (2007) Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, to appear.
12. Birgé, L. (1983) Approximation dans les espaces métriques et théorie de l'estimation. *Z. für Wahrscheinlichkeitstheorie und verw. Geb.*, **65**, 181-238.
13. Birgé, L. (2005) A new lower bound for multiple hypothesis testing. *IEEE Trans. on Information Theory*, **51**, 1611-1615.
14. Birgé, L. and Massart, P. (2001) Gaussian model selection. *J. Eur. Math. Soc.*, **3**, 203-268.
15. Borovkov, A.A. (1984) *Mathematical Statistics.* Nauka, Moscow. English translation by Gordon and Breach, Singapore e.a., 1998.
16. Borovkov, A.A. and Sakhanenko, A.I. (1980) On estimates of the averaged squared risk. *Probability and Mathematical Statistics*, **1**, 185-195 (in Russian).
17. Bretagnolle, J. and Huber, C. (1979) Estimation des densités: risque minimax. *Z. für Wahrscheinlichkeitstheorie und verw. Geb.*, **47**, 199-137.

18. Brown, L.D. and Low, M.G. (1996) Asymptotic equivalence of nonparametric regression and white noise. *Annals of Statistics*, **24**, 2384-2398.
19. Brown, L.D., Low, M.G. and Zhao, L.H. (1997) Superefficiency in nonparametric function estimation. *Annals of Statistics*, **25**, 898-924.
20. Brown, L.D., Carter, A., Low, M.G. and Zhang, C.-H. (2004) Equivalence theory for density estimation, Poisson processes and Gaussian white noise with drift. *Annals of Statistics*, **32**, 2399-2430.
21. Bunea, F., Tsybakov, A.B. and Wegkamp, M.H. (2007a) Aggregation for Gaussian regression. *Annals of Statistics*, **35**, 1674-1697.
22. Bunea, F., Tsybakov, A.B. and Wegkamp, M.H. (2007b) Sparsity oracle inequalities for the Lasso. *Electronic J. of Statistics*, **1**, 169-194.
23. Cai, T. (1999) Adaptive wavelet estimation: A block thresholding and oracle inequality approach. *Annals of Statistics*, **27**, 2607-2625.
24. Cavalier, L. and Tsybakov, A.B. (2001) Penalized blockwise Stein's method, monotone oracles and sharp adaptive estimation. *Mathematical Methods of Statistics*, **10**, 247-282.
25. Cavalier, L. and Tsybakov, A.B. (2002) Sharp adaptation for inverse problems with random noise. *Probability Theory and Related Fields*, **123**, 323-354.
26. Čencov, N.N. (1962) Evaluation of an unknown distribution density from observations. *Soviet Mathematics. Doklady*, **3**, 1559-1562.
27. Čencov, N.N. (1972) *Statistical Decision Rules and Optimal Inference.* Nauka, Moscow. English translation in *Translations of Mathematical Monographs*, **53**, AMS, Providence, RI, 1982.
28. Cline, D.B.H. (1988) Admissible kernel estimators of a multivariate density. *Annals of Statistics*, **16**, 1421-1427.
29. Cover, T.M. and Thomas, J.A. (2006) *Elements of Information Theory.* Wiley, New York.
30. Craven, P. and Wahba, G. (1979) Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, **31**, 377-403.
31. Csiszár, I. (1967) Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungarica*, **2**, 299-318.
32. Dalelane, C. (2005) Exact oracle inequality for a sharp adaptive kernel density estimator. `arXiv:math/0504382v1`
33. Davis, K.B (1975) Mean square error properties of density estimates. *Annals of Statistics*, **3**, 1025-1030.
34. Delyon, B. and Juditsky, A. (1996) On minimax wavelet estimators. *Appl. Comput. Harmonic Anal.*, **3**, 215-228.
35. Devroye, L. (1987) *A Course in Density Estimation.* Birkhäuser, Boston.
36. Devroye, L. and Györfi, L. (1985) *Nonparametric Density Estimation: The L_1 View.* Springer, New York.
37. Devroye, L. and Lugosi, G. (2000) *Combinatorial Methods in Density Estimation.* Springer, New York.
38. Donoho, D.L. and Johnstone, I.M. (1995) Adapting to unknown smoothness via wavelet shrinkage. *J.Amer. Statist. Assoc.*, **90**, 1200-1224.
39. Donoho, D.L. and Johnstone, I.M. (1998) Minimax estimation via wavelet shrinkage. *Annals of Statistics*, **26**, 789-921.
40. Efroimovich, S.Yu. and Pinsker, M.S. (1984) Learning algorithm for nonparametric filtering. *Automation and Remote Control*, **11**, 1434-1440.

41. Efromovich, S. (1999) *Nonparametric Curve Estimation.* Springer, New York.
42. Efromovich, S. (2004) Oracle inequalities for Efromovich–Pinsker blockwise estimates. *Methodol. Comput. Appl. Probab.*, **6**, 303-322.
43. Epanechnikov, V.A. (1969) Non-parametric estimation of a multivariate probability density. *Theory of Probability and Its Applications*, **14**, 153-158.
44. Eubank, R.L. (1988) *Spline Smoothing and Nonparametric Regression.* Marcel–Dekker, New York.
45. Fano, R.M. (1952) Class Notes for Transmission of Information. Course 6.574, MIT, Cambridge, Massachusetts.
46. Farrel, R. (1972) On the best obtainable asymptotic rates of convergence in estimation of a density at a point. *Annals of Mathematical Statistics*, **43**, 170-180.
47. Fan, J. and Gijbels, I. (1996) *Local Polynomial Modelling and Its Applications.* Chapman and Hall, London.
48. Fix, E. and Hodges, J.L. (1951) Discriminatory analysis – Nonparametric discrimination: Consistency properties. Technical Report. USAF School of Aviation Medicine. Published in *Internat. Statist. Review*, **57**, 238-247 (1989).
49. Folland, G.B. (1999) *Real Analysis.* Wiley, New York.
50. Gallager, R.G. (1968) *Information Theory and Reliable Communication*, Wiley, New York.
51. Gilbert, E.N. (1952) A comparison of signalling alphabets. *Bell System Technical J.*, **31**, 504-522.
52. Green, P.J. and Silverman, B.W. (1994) *Nonparametric Regression and Generalized Linear Models.* Chapman and Hall, London.
53. Gill, R.D. and Levit, B.Y. (1995) Applications of the van Trees inequality: A Bayesian Cramér–Rao bound. *Bernoulli*, **1**, 59-79.
54. Golubev, G.K. (1987) Adaptive asymptotically minimax estimates of smooth signals. *Problems of Information Transmission*, **23**, 57-67.
55. Golubev, G.K. (1992) Nonparametric estimation of smooth densities of a distribution in L_2. *Problems of Information Transmission*, **28**, 44-54.
56. Golubev, G.K. and Nussbaum, M. (1992) Adaptive spline estimates in a nonparametric regression model. *Theory of Probability and Its Applications*, **37**, 521-529.
57. Grama, I.G. and Nussbaum, M. (2006) Asymptotic equivalence of nonparametric autoregression and nonparametric regression. *Annals of Statistics*, **34**, 1701-1732.
58. Gushchin, A.A. (2002) On Fano's lemma and similar inequalities for the minimax risk. *Probability Theory and Mathematical Statistics*, **67**, 26-37.
59. Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002) *A Distribution-Free Theory of Nonparametric Regression.* Springer, New York.
60. Has'minskii, R.Z. (1978) A lower bound on the risks of nonparametric estimates of densities in the uniform metric. *Theory of Probability and Its Applications*, **23**, 794-798.
61. Härdle, W. (1990) *Applied Nonparametric Regression.* Cambridge Univ. Press, Cambridge.
62. Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. (1998) *Wavelets, Approximation and Statistical Applications.* Lecture Notes in Statistics, v. 129. Springer, New York.
63. Hart, J.D. (1997) *Nonparametric Smoothing and Lack-of-Fit Tests.* Springer, New York.

64. Hernández, E. and Weiss, G. (1996) *A First Course on Wavelets.* CRC Press, Boca Raton, New York.
65. Hodges, J.L. and Lehmann, E.L. (1956) The efficiency of some nonparametric competitors to the t-test. *Annals of Mathematical Statistics*, **13**, 324-335.
66. Hoffmann, M. (1999) Adaptive estimation in diffusion processes. *Stochastic Processes and Their Applications*, **79**, 135-163.
67. Ibragimov, I.A. and Has'minskii, R.Z. (1977) On the estimation of an infinite-dimensional parameter in Gaussian white noise. *Soviet Mathematics. Doklady*, **18**, 1307-1309.
68. Ibragimov, I.A. and Has'minskii, R.Z. (1981) *Statistical Estimation: Asymptotic Theory.* Springer, New York.
69. Ibragimov, I.A. and Khas'minskii, R.Z. (1982) Bounds for the risks of nonparametric regression estimates. *Theory of Probability and Its Applications*, **27**, 84-99.
70. Ibragimov, I.A. and Has'minskii, R.Z. (1983a) Estimation of distribution density. *J. of Soviet Mathematics*, **25**, 40-57. (Originally published in Russian in 1980).
71. Ibragimov, I.A. and Has'minskii, R.Z. (1983b) Estimation of distribution density belonging to a class of entire functions. *Theory of Probability and Its Applications*, **27**, 551-562.
72. Ibragimov, I.A. and Has'minskii, R.Z. (1984) Asymptotic bounds on the quality of the nonparametric regression estimation in L_p. *J. of Soviet Mathematics*, **25**, 540-550. (Originally published in Russian in 1980).
73. Ibragimov, I.A., Nemirovskii, A.S. and Khas'minskii, R.Z. (1987) Some problems of nonparametric estimation in Gaussian white noise. *Theory of Probability and Its Applications*, **31**, 391-406.
74. Ingster, Yu.I. and Suslina, I.A. (2003) *Nonparametric Goodness-of-fit Testing under Gaussian Models.* Springer, New York.
75. James, W. and Stein, C. (1961) Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, **1**, 311-319.
76. Johnstone, I.M. (1998) *Function Estimation in Gaussian Noise: Sequence Models.* Draft of a monograph. `http://www-stat.stanford.edu/~imj`
77. Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1996) Density estimation by wavelet thresholding. *Annals of Statistics*, **24**, 508-539.
78. Katkovnik, V.Y. (1979) Linear and nonlinear methods of nonparametric regression analysis. *Soviet Automatic Control*, **5**, 35-46.
79. Katznelson, Y. (2004) *An Introduction to Harmonic Analysis.* Cambridge Univ. Press, Cambridge.
80. Kemperman, J. (1969) On the optimum rate of transmitting information. *Probability and Information Theory*, Lecture Notes in Mathematics, v. 89. Springer, Berlin, 126-169.
81. Kerkyacharian, G. and Picard, D. (1992) Density estimation in Besov spaces. *Statistics and Probablility Letters*, **13**, 15-24.
82. Kneip, A. (1994) Ordered linear smoothers. *Annals of Statistics*, **22**, 835-866.
83. Koltchinskii, V. (2008) Sparsity in penalized empirical risk minimization. *Annales de l'Insitut Henri Poincaré*, to appear.
84. Konakov, V.D. (1972) Non-parametric estimation of density functions. *Theory of Probability and Its Applications*, **17**, 361-362.
85. Korostelev, A.P. and Tsybakov, A.B. (1993) *Minimax Theory of Image Reconstruction.* Lecture Notes in Statistics, v. 82. Springer, New York.

86. Kuks, J.A. and Olman, V. (1971) A minimax linear estimator of regression coefficients. *Izv. Akad. Nauk Eston. SSR*, **20**, 480-482 (in Russian).
87. Kullback, S. (1967) A lower bound for discrimination information in terms of variation. *IEEE Trans. on Information Theory*, **13**, 126-127.
88. Le Cam, L. (1953) On some asymptotic properties of maximum likelihood and related Bayes estimates. *Univ. of California Publications in Statist.*, **1**, 277-330.
89. Le Cam, L. (1973) Convergence of estimates under dimensionality restrictions. *Annals of Statistics*, **1**, 38-53.
90. Le Cam, L. and Yang, G.L. (2000) *Asymptotics in Statistics. Some Basic Concepts.* Springer, New York.
91. Lehmann, E.L. and Casella, G. (1998) *Theory of Point Estimation.* Springer, New York.
92. Lepski, O.V. (1990) On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability and Its Applications*, **35**, 454-466.
93. Lepski, O., Mammen, E. and Spokoiny, V. (1997) Optimal spatial adaptation to inhomogeneous smoothness: An approach based on kernel estimators with variable bandwidth selectors. *Annals of Statistics*, **25**, 929-947.
94. Lepski, O., Nemirovski, A. and Spokoiny,V. (1999) On estimation of the L_r norm of a regression function. *Probability Theory and Related Fields*, **113**, 221-253.
95. Loader, C. (1999) *Local Regression and Likelihood.* Springer, New York.
96. Malliavin, P. (1995) *Integration and Probability.* Springer, New York.
97. Mallows, C.L. (1973) Some comments on C_p. *Technometrics*, **15**, 661-675.
98. Massart, P. (2007) *Concentration Inequalities and Model Selection.* Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003. Lecture Notes in Mathematics, v. 1879. Springer, New York.
99. McQuarrie, A.D.R., Tsai, C.-L. (1998) *Regression and Time Series Model Selection.* World Scientific, Singapore.
100. Meyer, Y. (1990) *Ondelettes et opérateurs.* Hermann, Paris.
101. Nadaraya, E.A. (1964) On estimating regression. *Theory of Probabilty and Its Applications*, **9**, 141-142.
102. Nemirovskii, A.S. (1985) Nonparametric estimation of smooth regression functions. *Soviet J. of Computer and Systems Sciences*, **23**, 1-11.
103. Nemirovskii, A.S. (1990) On necessary conditions for the efficient estimation of functionals of a nonparametric signal which is observed in white noise. *Theory of Probability and Its Applications*, **35**, 94-103.
104. Nemirovski, A. (2000) *Topics in Non-parametric Statistics.* Ecole d'Eté de Probabilités de Saint-Flour XXVIII - 1998. Lecture Notes in Mathematics, v. 1738. Springer, New York.
105. Nemirovskii A.S., Polyak B.T. and Tsybakov, A.B. (1983) Estimators of maximum likelihood type for nonparametric regression. *Soviet Mathematics. Doklady*, **28**, 788-792.
106. Nemirovskii A.S., Polyak B.T. and Tsybakov, A.B. (1984) Signal processing by the nonparametric maximum likelihood method. *Problems of Information Transmission*, **20**, 177-192.
107. Nemirovskii A.S., Polyak B.T. and Tsybakov, A.B. (1985) Rate of convergence of nonparametric estimators of maximum-likelihood type. *Problems of Information Transmission*, **21**, 258-272.
108. Nussbaum, M. (1985) Spline smoothing in regression models and asymptotic efficiency in L_2. *Annals of Statistics*, **13**, 984-997.

109. Nussbaum, M. (1996) Asymptotic equivalence of density estimation and Gaussian white noise. *Annals of Statistics*, **24**, 2399-2430.
110. Parzen, E. (1962) On the estimation of a probability density function and mode. *Annals of Mathematical Statistics*, **33**, 1065-1076.
111. Pinsker, M.S. (1964) *Information and Information Stability of Random Variables and Processes*. Holden-Day, San Francisco.
112. Pinsker, M.S. (1980) Optimal filtering of square integrable signals in Gaussian white noise. *Problems of Information Transmission*, **16**, 120-133.
113. Polyak, B.T. and Tsybakov, A.B. (1990) Asymptotic optimality of the C_p-test for the orthogonal series estimation of regression. *Theory of Probability and Its Applications*, **35**, 293-306.
114. Polyak, B.T. and Tsybakov, A.B. (1992) A family of asymptotically optimal methods for choosing the order of a projective regression estimate. *Theory of Probability and Its Applications*, **37**, 471-481.
115. Reiss, M. (2008) Asymptotic equivalence for nonparametric regression with multivariate and random design. *Annals of Statistics*, **36**, 1957-1982.
116. Rice, J. (1984) Bandwidth choice for nonparametric regression. *Annals of Statistics*, **12**, 1215-1230.
117. Rigollet, P. (2006a) Adaptive density estimation using the blockwise Stein method. *Bernoulli*, **12**, 351-370.
118. Rigollet, P. (2006b) Inégalités d'oracle, agrégation et adaptation. Thesis, Université Paris VI. `http://tel.archives-ouvertes.fr/tel-00115494`
119. Rosenblatt, M. (1956) Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, **27**, 832-837.
120. Rudemo, M. (1982) Empirical choice of histograms and kernel density estimators. *Scandinavian J. of Statistics*, **9**, 65-78.
121. Ruppert, D., Wand, M. and Carroll, R. (2003) *Semiparametric Regression.* Cambridge Univ. Press, Cambridge.
122. Scott, D. W. (1992) *Multivariate Density Estimation.* Wiley, New York.
123. Scheffé, H. (1947) A useful convergence theorem for probability distributions. *Annals of Mathematical Statistics*, **18**, 434-458.
124. Shibata, R. (1981) An optimal selection of regression variables. *Biometrika*, **68**, 45-54.
125. Silverman, B.W. (1984) Spline smoothing: The equivalent variable kernel method. *Annals of Statistics*, **12**, 898-916.
126. Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.
127. Speckman, P. (1985) Spline smoothing and optimal rates of convergence in nonparametric regression models. *Annals of Statistics*, **13**, 970-983.
128. Stein, C. (1956) Inadmissibility of the usual estimator of the mean of a multivariate distribution. *Proc. Third Berkeley Symp. Math. Statist. Prob.*, **1**, 197-206.
129. Stein, C.M. (1981) Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, **9**, 1135-1151.
130. Stone, C.J. (1977) Consistent nonparametric regression. *Annals of Statistics*, **5**, 595-645.
131. Stone, C.J. (1980) Optimal rates of convergence for nonparametric estimators. *Annals of Statistics*, **8**, 1348-1360.
132. Stone, C.J. (1982) Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, **10**, 1040-1053.

133. Stone, C.J. (1984) An asymptotically optimal window selection rule for kernel density estimates. *Annals of Statistics*, **12**, 1285-1297.
134. Strawderman, W.E. (1971) Proper Bayes minimax estimators of the multivariate normal mean. *Annals of Mathematical Statistics*, **42**, 385-388.
135. Szegö, G. (1975) *Orthogonal Polynomials.* AMS, Providence, Rhode Island.
136. Tsybakov, A.B. (1986) Robust reconstruction of functions by a local-approximation method. *Problems of Information Transmission*, **22**, 133-146.
137. Vajda, I. (1986) *Theory of Statistical Inference and Information.* Kluwer, Dordrecht.
138. van de Geer, S.A. (2000) *Applications of Empirical Processes Theory.* Cambridge Univ. Press, Cambridge.
139. van de Geer, S.A. (2008) High dimensional generalized linear models and the Lasso. *Annals of Statistics*, **36**, 614-645.
140. van Trees, H.L. (1968) *Detection, Estimation and Modulation Theory*, Part 1. Wiley, New York.
141. Wahba, G. (1990) *Spline Models for Observational Data.* SIAM, New York.
142. Wand, M.P. and Jones, M.C. (1995) *Kernel Smoothing.* Chapman and Hall, London.
143. Wasserman, L. (2006) *All of Nonparametric Statistics.* Springer, New York.
144. Watson, G.S. (1964) Smooth regression analysis. *Sankhyā, Ser. A*, **26**, 359-372.
145. Watson, G.S. and Leadbetter, M.R. (1963) On the estimation of the probability density, I. *Annals of Mathematical Statistics*, **34**, 480-491.
146. Yang, Y. and Barron, A. (1999) Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, **27**, 1564-1599.

Index