CWI Tracts

CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
Telephone + 31 - 20 592 9333
Telefax + 31 - 20 592 4199
WWW page http://www.cwi.nl

CWI is the nationally funded Dutch institute for research in Mathematics and Computer Science.

Entropy methods in statistical estimation

M.H. Wegkamp

1991 Mathematics Subject Classification: 60G15, 60G35, 60G50, 60E15, 62G07, 62G20
ISBN 90 6196 483 0
NUGI-code: 811

Editorial Preface

This CWI Tract contains the Ph.D. dissertation of Dr. Marten Wegkamp, as it has been defended at the University of Leiden on June 26, 1996. The research of this thesis has been performed within the framework of the Dutch Research School "Thomas Stieltjes Institute for Mathematics". The Stieltjes Institute has awarded Dr. Wegkamp's thesis with a prize, as being the best 1996 thesis (out of 12) written in this Research School. We congratulate Dr. Wegkamp with this prize, and appreciate that we have obtained his permission to include his thesis into the CWI Tract Series.

The editors

Contents

Part I

Preliminaries

Chapter 1

General introduction

This book is devoted to estimation problems in statistics. In particular we study the asymptotic performance of certain estimators - defined as minimizers of certain loss functions - under purely metric assumptions on the parameter space. The nature of the statistical problem may be nonparametric, i.e. the parameter space is allowed to be infinite dimensional. In the statistical analysis an important role is played by the metric structure of the parameter space as described by its metric entropy numbers. These numbers occur in classical approximation theory and modern probability theory. In the next chapter we give their formal definition and present some examples.

By now it is a well-known phenomenon in mathematical statistics that there exists a large class of estimators with good statistical properties for the unknown parameter of interest, provided it belongs to a "nice" (pseudo) metric space. At this point, "nice" should be understood in terms of small entropy numbers. We refer to the work of Le Cam [25], [26], Ibragimov & Has'minskii [22] and Birgé [4]. On the other hand it follows from the work of Birgé & Massart [5] and Van de Geer [45] that estimators, which seem reasonable at first sight- such as maximum likelihood estimators -, can have suboptimal rates of convergence or can even be inconsistent if the entropy numbers are too large.

Let us focus on the classical situation of estimating a finite dimensional parameter. This will be the topic of Chapter 3. Let $X_1, X_2, \ldots$ be a sequence of independent, identically distributed (i.i.d.) random

variables with common probability measure $P \in \mathcal{P}$. From the first n observations we construct the empirical measure P_n. This measure puts mass $1/n$ at each observation X_i, $i = 1, \ldots, n$. The signed measure $\sqrt{n}(P_n - P)$ will be denoted by E_n. We wish to estimate the true value θ_0 of a parameter $\theta \in \Theta \subset \mathbb{R}^k$; θ_0 is assumed to be an interior point of Θ which uniquely minimizes the deterministic quantity

$$M(\theta) = \int g(\cdot, \theta)\, dP, \tag{1.1}$$

over all $\theta \in \Theta$, where P is the true underlying probability measure. The M-estimator of θ_0, which we denote by $\hat{\theta}_n$, is implicitly defined as the minimizer of

$$M_n(\theta) = \int g(\cdot, \theta)\, dP_n = \frac{1}{n} \sum_{i=1}^{n} g(X_i, \theta), \tag{1.2}$$

the empirical counterpart of M. The function g is sometimes called the contrast function. In the classical situation $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$,

$$g(x, \theta) = -\log f_\theta(x), \quad -\infty < x < \infty$$

where f_θ is the density of P_θ with respect to the Lebesgue measure, and $\hat{\theta}_n$ is the maximum likelihood estimator. However, we also have models in mind where θ describes only some aspect of P rather than specifying P completely. Under the assumption that there exists a linear expansion of $g(\cdot, \theta)$ in a vicinity of θ_0, i.e.

$$g(\cdot, \theta) = g(\cdot, \theta_0) + (\theta - \theta_0)' \Delta(\cdot) + |\theta - \theta_0| r(\cdot, \theta) \tag{1.3}$$

such that

(i) $\int \Delta^2\, dP < \infty$

(ii) the empirical process $\{\int r(\cdot, \theta)\, dE_n : \theta \in \Theta\}$ is stochastically equicontinuous at θ_0,

Pollard proved asymptotic normality of the M-estimator $\hat{\theta}_n$. The stochastic equicontinuity assertion means that

$$\left| \int r(\cdot, \tilde{\theta}_n)\, dE_n - \int r(\cdot, \theta_0)\, dE_n \right| \xrightarrow{P} 0$$

for every sequence $\tilde{\theta}_n$ which converges to θ_0 in probability. Pollard showed that this condition is implied by certain entropy conditions on $\{r(\cdot,\theta) : \theta \in \Theta\}$ (cf. Pollard [35], p. 150, Lemma 15).

We obtain more general results if we allow the contrast function g to depend on the unknown probability measure P. This more complex case has been studied by for instance Stute [38]. He investigated the related problem where θ_0 is now given as the (unique) root of

$$\int g(\theta, F, x)\, dF(x) = 0, \tag{1.4}$$

where F is the distribution function associated with P. A natural estimator for θ_0 is any $\hat{\theta}_n \in \Theta$ which approximately solves

$$\int g(\theta, F_n, x)\, dF_n(x) = 0 \tag{1.5}$$

within $o(1/n)$. Here F_n is the empirical distribution function.

The initial goal of the present author was to give a new and shorter proof of the asymptotic normality of this estimator as considered by Stute, by using more recent probabilistic results of the theory of empirical processes. Moreover, it turned out that the formulation of the problem can be generalized. In fact we consider a function $\gamma : \Theta \times \mathcal{P} \to \mathbb{R}$ with the property that $\gamma(\theta, P)$ is minimal for $\theta = \theta_0$ with P being the true probability measure. We establish asymptotic normality of $\hat{\theta}_n$, which minimizes the empirical contrast

$$\gamma_n(\theta) = \gamma(\theta, P_n) \tag{1.6}$$

over $\theta \in \Theta$. We shall need a kind of stochastic differentiability of the empirical process $\sqrt{n}(\gamma(\cdot, P) - \gamma(\cdot, P_n))$ as the main condition for the proof of this result. We shall proceed by arguing that this condition can be checked by empirical process methods involving entropy assumptions.

In nonparametric statistical problems the "entropy approach" has turned out to be quite fruitful. Especially nonparametric maximum likelihood estimation has received much attention during the last decade. Again, let $X_1, X_2, \ldots$ be i.i.d. random variables taking values in a measurable space $(\mathcal{X}, \mathcal{A})$ with common probability measure P. The available

information we have of P is that it belongs to some family of probability measures $\mathcal{P}$ on $(\mathcal{X}, \mathcal{A})$. In addition, suppose that every $Q \in \mathcal{P}$ is dominated by some σ-finite measure μ. The induced densities are denoted by

$$f(x) = \frac{dQ}{d\mu}(x).$$

It can be shown (cf. Van de Geer [42]) that the following relation holds true

$$H^2(f_0, \hat{f}_n) \leq 2 \int_{f_0>0} [\sqrt{\frac{\hat{f}}{f_0}} - 1] \, d(P_n - P). \tag{1.7}$$

Here

$$H^2(f, g) = \int \left(f^{1/2} - g^{1/2}\right)^2 d\mu$$

is the squared Hellinger distance between two densities, $f_0 = dP/d\mu$ is the true density and $\hat{f}_n$ is the nonparametric maximum likelihood estimator which maximizes the likelihood

$$\prod_{i=1}^{n} f(X_i)$$

over $f = dQ/d\mu$, $Q \in \mathcal{P}$. As a consequence of the strong law of large numbers,

$$\int_{f_0>0} [\sqrt{\frac{f}{f_0}} - 1] \, d(P_n - P) \xrightarrow{a.s.} 0 \tag{1.8}$$

for any density f. If this convergence holds uniformly in $f = dQ/d\mu$, $Q \in \mathcal{P}$, it is easily seen from (1.7) that the nonparametric maximum likelihood estimator is Hellinger consistent, i.e $H(\hat{f}, f_0) \xrightarrow{P} 0$. Uniform convergence of $P_n - P$ over classes of functions has been studied extensively in the theory of empirical processes nowadays. As a result, we can formulate Hellinger consistency in terms of metric conditions on the class of densities. Second, we point out that the rate of convergence follows from a more precise analysis involving the modulus of oscillation of the empirical process $\int_{f_0>0} \left([f/f_0]^{1/2} - 1\right) d(P_n - P)$.

The last part of this study (Chapters 4, 5 and 6) is confined to a related topic, viz. nonparametric regression estimation. In this case we have n independent observations (X_i, Y_i), $i = 1, \ldots, n$. The Y_i are related to the X_i by a regression model

$$Y_i = g(X_i) + \varepsilon_i. \tag{1.9}$$

The random variables ε_i are viewed as disturbances in the model (1.9) and they are small in the sense that $\mathbb{E}\varepsilon_i = 0$ and $\mathbb{E}\varepsilon_i^2 < \infty$. Knowledge about the unknown regression function g is expressed by writing $g \in \mathcal{G}$, where $\mathcal{G}$ is some known class of functions. For instance we might assume that g is a linear, smooth or just a monotone function. At the moment we do not dwell on the various possible choices for X_i, $i = 1, \ldots, n$. They may be random or deterministic.

We shall consider the nonparametric least squares estimator. This estimator is defined implicitly as the minimizer of the sum of squares

$$\sum_{i=1}^{n} (Y_i - g(X_i))^2$$

and is denoted by $\hat{g}$. The statistical behavior of this estimator has been studied by Nemirovskii et al. (cf. [31], [32]). They restrict themselves to the classical Sobolev spaces. Van de Geer (cf. [39],[40]), however, considers spaces $\mathcal{G}$ which satisfy certain metric entropy conditions without using any analytical properties. These spaces include the Sobolev spaces as treated in [31],[32].

We shall follow Van de Geer's approach in this study. Parallel to inequality (1.7) we have in the regression setting

$$d_n^2(\hat{g}, g) \leq 2\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i(\hat{g} - g)(X_i), \tag{1.10}$$

where

$$d_n^2(f, g) = \frac{1}{n}\sum_{i=1}^{n} (f - g)^2(X_i)$$

is the squared $L^2(P_n)$ distance between two functions f and g. For the moment suppose $X_1, X_2, \ldots$ are i.i.d. with common probability measure

P. Again by virtue of the strong law of large numbers,

$$\int g\, d(P_n - P) \xrightarrow{a.s.} 0 \tag{1.11}$$

for $g \in L^1(P)$, and a uniform version of (1.11) would entail consistency of the least squares estimator (in the $L_2(P_n)$ metric). Giné & Zinn (cf. [14]) proved that the uniform strong law of large numbers

$$\sup_{g \in \mathcal{G}} \left| \int g\, d(P_n - P) \right| \xrightarrow{a.s.} 0 \tag{1.12}$$

is equivalent to the envelope condition

$$\int \sup_{g \in \mathcal{G}} |g|\, dP < \infty \tag{1.13}$$

and a certain entropy condition on $\mathcal{G}$. The envelope assumption (1.13) is unfortunately rather restrictive. It even excludes the classical linear model

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

with $\alpha, \beta \in \mathbb{R}$, $\mathbb{E}\varepsilon_i = 0$, $\mathbb{E}\varepsilon_i^2 < \infty$. Van de Geer established $L^2(P)$ consistency without assuming this restrictive envelope condition on $\mathcal{G}$. To formulate her result, we need some more notation. Define for each $g \in \mathcal{G}$,

$$f = f(g) = \frac{g}{1 + \|g\|_2},$$

with $\|g\|_2^2 = \int g^2\, dP$. Next, let $C > 0$ and

$$(f)_C = \begin{cases} C & \text{if } f > C \\ -C & \text{if } f < -C \\ f & \text{otherwise} \end{cases}$$

Under an entropy assumption on the class of rescaled functions $(f)_C$ and provided

$$\lim_{C \to \infty} \sup_{f \in \mathcal{F}} \int f^2 I\{|f| > C\}\, dP = 0, \tag{1.14}$$

the least squares estimator is $L^2(P)$ consistent.

We shall establish a similar result but under a different set of conditions. In particular, we do not need the uniform integrability assumption. Second, we are interested in the necessity of our metric assumptions. In other words, how good is this metric approach in fact?

The organization of the third part of the book is as follows. In Chapter 4 we consider consistency issues. More specifically, necessary and sufficient entropy conditions will be derived for consistency of the nonparametric least squares estimator.

Rates of convergence will be the topic in Chapter 5. We recall Van de Geer's result (cf. [41]) that more restrictive entropy conditions entail rates of convergence, provided the errors fulfill an exponential moment condition. We show that the rates of convergence are optimal for normally distributed disturbances ε_i. But situations where this restriction on the errors fails, will be discussed as well.

Finally, we present some asymptotic distribution theory for the least squares estimator in Chapter 6. In particular, we prove asymptotic normality of the squared empirical $L^2(P_n)$ norm of the least squares estimator.

But first we set out with an overview of some results of the theory of empirical processes, which will be needed in this work.

Chapter 2

Empirical process theory

We introduce the entropy numbers and discuss their role in the theory of empirical processes. The purpose of this chapter is to provide the reader with mathematical tools, which will be frequently used in the remainder of this work. We present some of the main results of the theory of empirical processes as developed by Dudley, Vapnik & Červonenkis, Giné & Zinn, Pollard and others. For an overview of this area, we refer to Van der Vaart & Wellner [49]. Another recent review with emphasis on the statistical applications is the forthcoming monograph by Van de Geer [46].

2.1 Covering numbers

Let (T, d) be a pseudo metric space. This means that d possesses the properties of a distance except that it does not necessarily distinguish between two different elements, so $d(s,t) = 0$ need not imply $s = t$. The diameter of T is denoted by $\Delta = \Delta(T)$, i.e.

$$\Delta = \sup \{d(s,t) : s, t \in T\},$$

possibly infinite. Let T_0 be a subset of T.

Definition 2.1 *Let $\delta > 0$. A set S_0 is called a* δ-covering net *for T_0 if and only if for every element $t \in T_0$ there exists an element $s \in S_0$ satisfying $d(s,t) \leq \delta$. Equivalently, $T_0 \subseteq \cup_{s \in S_0} B(s, \delta)$ with $B(s, \delta) = \{t \in T : d(s,t) \leq \delta\}$.*

We say a pseudo metric space is *totally bounded* if there exists a finite δ-covering net for every $\delta > 0$. (T_0, d) is compact if and only if it is both complete and totally bounded.

Definition 2.2 *Let $\delta > 0$. Denote by $N(\delta, d, T_0)$ the* δ-covering number *of T_0, defined as the smallest number of closed balls needed to cover T_0 or equivalently, the cardinality of the smallest δ-covering net. The* δ-entropy number *or metric entropy of T_0 is defined by*

$$H(\delta, d, T_0) = \log\left[N(\delta, d, T_0)\right].$$

Clearly these numbers are decreasing in δ. Metric entropy describes an important geometric feature of T_0. It can be viewed as a measure of how totally bounded (T_0, d) is. The smaller this number relative to its diameter, the more "airy" the space. Another way to measure the size or "thickness" of (T_0, d) is by means of its *δ-packing numbers.*

Definition 2.3 *Let $\delta > 0$. Denote by $D(\delta, d, T_0)$ the* δ-packing number *of T_0, defined as the largest number m for which there exist points $t_1, \ldots, t_m \in T_0$ with $d(t_i, t_j) > \delta$ for $i \neq j$. The* δ-capacity number *is given by*

$$C(\delta, d, T_0) = \log\left[D(\delta, d, T_0)\right].$$

These two concepts are essentially the same in view of the following relation between covering and packing numbers

$$N(\delta, d, T_0) \leq D(\delta, d, T_0) \leq N(\frac{\delta}{2}, d, T_0) \quad \delta > 0, \tag{2.1}$$

which has been proved by Kolmogorov & Tichomirov [23].

If there exists an ordering on T, a more refined way to describe the metric geometry of (T_0, d) is by the so-called *entropy with bracketing.* Define $T_0^B(\delta)$ as the smallest set for which each element $t \in T_0$ can be sandwiched between two δ-separated elements of $T_0^B(\delta)$, i.e. for each $t \in T_0$ there exist $t_L, t_U \in T_0^B(\delta)$ with $d(t_L, t_U) \leq \delta$ and $t_L \leq t \leq t_U$. Let $N_B(\delta, d, T_0)$ be the cardinality of $T_0^B(\delta)$; the logarithm of this number

is called the δ-entropy with bracketing number. Generally this entropy with bracketing number is larger than the ordinary metric entropy.

Let us end this section by presenting some examples of pseudo metric spaces. It is often rather difficult to compute entropy numbers, but for many interesting spaces good approximations of these numbers are available.

Example 2.1 (Sobolev spaces) For every $k = (k_1, \ldots, k_n) \in \mathbb{N}^n$, define the differential operator D^k by

$$D^k = \frac{\partial^{k_1+\ldots+k_n}}{\partial x_1^{k_1} \ldots \partial x_n^{k_n}}.$$

Let $\mathcal{F}$ be the class of real valued, continuous functions on the the unit cube S^n in $\mathbb{R}^n$ possessing uniformly bounded partial derivatives of order $k \leq p$, i.e. for some constant C_1 independent of f,

$$\max_{k_1+\ldots+k_n \leq p} \max_{x \in S_n} |D^k f(x)| \leq C_1.$$

Moreover the p-th order partial derivatives of each f satisfy a Lipschitz condition of order α $(0 < \alpha \leq 1)$, i.e. there exists a $C_2 > 0$ independent of f such that

$$\left|D^k f(x) - D^k f(y)\right| \leq C_2 \|x - y\|^\alpha$$

for all $x, y \in S^n$ and all $k \in \mathbb{N}^n$ with $k_1 + \ldots + k_n = p$. Under the uniform metric $\rho(f, g) = \max_{x \in S^n} |f(x) - g(x)|$, it is known that (cf. Kolmogorov & Tichomirov [23]) the entropy of this class is of the order

$$H(\delta, \rho, \mathcal{F}) \asymp \delta^{-n/q}, \; q = p + \alpha.$$

We use the convention $f \asymp g$ if there exist positive constants C_1 and C_2 such that $C_1|f| \leq |g| \leq C_2|f|$. For dimension $n = 1$, we define the smoothness of a function by

$$J_p(f) = \int_0^1 |f^{(p)}(x)|^2 dx.$$

One can show (cf. Birman & Solomjak [6]) that

$$H\left(\delta, \rho, \left\{f : [0,1] \to [-C, C] : J_p(f) \leq \tilde{C}\right\}\right) \asymp \delta^{-1/p}.$$

Example 2.2 Let f be a continuous function on $[0,1]$ and $\Delta(f,x,t)$ be the second difference of f at x with increment t, i.e.

$$\Delta(f,x,t) = f(x+2t) - 2f(x+t) + f(x).$$

Set

$$\omega(f,\delta) = \max_{|t|\leq\delta,\ x\in[0,1-2t]} |\Delta(f,x,t)|.$$

For a strictly increasing function ψ, we define the set A_ψ of all continuous functions $f : [0,1] \to \mathbb{R}$ with $\max_{x\in[0,1]} |f(x)| \leq K$ and $\omega(f,\delta) \leq \psi(\delta)$. If $\log(1/\psi(\delta)) \leq 1/\delta$ and $M(\delta) = \sum_{i=0}^{\infty} \psi(2^{-i}\delta) < \infty$, then the order of $H(\delta,\rho,A_\psi)$ is at most $1/M^{-1}(\delta)$ under the uniform metric ρ. Moreover if ψ is concave and strictly increasing, $H(\delta,\rho,A_\psi)$ is of order at least $1/\psi^{-1}(\delta)$. See Clements [7].

Example 2.3 (VC-classes) Let $\mathcal{C}$ be a collection of measurable subsets of a measurable space $(S,\mathcal{S})$. We say that $\mathcal{C}$ is a polynomial class or *VC-class* (named after Vapnik-Červonenkis) if there exists a discriminating polynomial p such that for every $N \in \mathbb{N}$ and every subset $A \subset S$ with cardinality N, there are at most $p(N)$ distinct subsets of the form $A \cap C$ with $C \in \mathcal{C}$. Thus $\mathcal{C}$ picks out only $p(N)$ from the 2^N possible subsets.

A collection $\mathcal{C}$ *shatters* a finite set C_0 if every subset of C_0 takes the form $C_0 \cap C$ with $C \in \mathcal{C}$. If a collection $\mathcal{C}$ can not shatter every set of N points, then it can be shown (see e.g. Pollard [35]) that $\mathcal{C}$ is a VC-class and the degree of the discriminating polynomial is less than N.

The most familiar example of a VC-class is the collection of quadrants $(-\infty,t]$, $t \in \mathbb{R}^k$. The class of sets $\{g \geq 0\}$ with g ranging over a finite dimensional vector space is another example of a VC-class (cf. Pollard [35]).

VC-classes have many known properties, for instance the classes $\{C^c : C \in \mathcal{C}_1\}$, $\{C_1 \cap C_2 : C_1 \in \mathcal{C}_1, C_2 \in \mathcal{C}_2\}$, $\{h(C) : C \in \mathcal{C}_1\}$ for any VC-classes $\mathcal{C}_1, \mathcal{C}_2$ and any function h on S, still possess the VC-property. As a consequence quite large classes can be built from elementary VC-classes.

Let Q be a probability measure on the space $(S, \mathcal{S})$ and let $\rho_r(Q)$ be the $L^r(Q)$ pseudo norm on S. Then we have the following entropy bound for any VC-class $\mathcal{C}$,

$$N(\delta, \rho_r(Q), \mathcal{C}) \leq K_V \left(\frac{1}{\delta}\right)^{r(V-1)}, \tag{2.2}$$

where K_V is a constant only depending on V, the smallest integer for which no set of V points can be shattered by $\mathcal{C}$ (cf. Van der Vaart & Wellner [49]). This means that the δ-covering number of a VC-class has a polynomial growth in $1/\delta$.

The *graph of a function* $f : S \to \mathbb{R}$ is defined as the set

$$G(f) = \{(x,t) \mid 0 \leq t \leq f(x) \text{ or } f(x) \leq t \leq 0\}.$$

We call a collection of functions $\mathcal{F}$ a *VC-graph class* if the graphs $\{G(f) : f \in \mathcal{F}\}$ form a VC-class in $S \times \mathbb{R}$. The *natural envelope* F of $\mathcal{F}$ is defined as the pointwise supremum of $|f|$, $F(s) = \sup_{f \in \mathcal{F}} |f(s)|$. If $F \in L^r(Q)$, then we find an upper bound for VC-graph classes, similar to (2.2); the δ-covering number grows polynomially in $1/\delta$. See for instance Pollard [35] or Van der Vaart & Wellner [49].

Another interesting result is that given the bound

$$N(\delta \|F\|_2, \rho_2(Q), \mathcal{F}) \leq C \left(\frac{1}{\delta}\right)^V, \ 0 < \delta < 1$$

the entropy of the *sequentially closed convex hull* of $\mathcal{F}$ satisfies

$$H\left(\delta \|F\|_2, \rho_2(Q), \overline{\text{conv}(\mathcal{F})}\right) \leq K_{C,V} \left(\frac{1}{\delta}\right)^{2V/(V+2)}, \ 0 < \delta < 1,$$

where $\|F\|_2$ denotes the $L^2(Q)$ pseudo norm of the envelope F of $\mathcal{F}$.

As a consequence of this, by taking $\mathcal{F} = \{I_{[0,t)}, t \geq 0\}$, we find for the class $\mathcal{G}$ of functions of bounded variation, which are uniformly bounded by 1, the entropy bound

$$H(\delta, \rho_2(Q), \mathcal{G}) = \mathcal{O}(\delta^{-1}),$$

for every probability measure Q. We refer to Birman & Solomjak [6], Ball & Pajor [2] and Van de Geer [42].

2.2 Gaussian processes

Gaussian processes form a fundamental part of probability theory. For an arbitrary index set T, a random process $Z = \{Z_t : t \in T\}$ is called *Gaussian* if each finite dimensional projection $(Z_{t_1}, \ldots, Z_{t_m})$ is normally distributed. We call $\{Z_t : t \in T\}$ *centered* if $\mathbb{E}Z_t = 0$ for each $t \in T$. The covariance structure

$$\Sigma(s,t) = \mathbb{E}Z_s Z_t$$

completely specifies the distribution of a centered Gaussian process Z. This makes the L^2 pseudo distance, defined by

$$d_Z(s,t) = \left(\mathbb{E}(Z_s - Z_t)^2\right)^{1/2}, \quad s,t \in T$$

a natural metric on T.

We put $\|z(t)\|_T = \sup_{t\in T} |z(t)|$ for any family of numbers $z(t)$ indexed by T. In many situations we are interested in the behavior of $\sup_t |Z_t|$. The entropy numbers $N(\delta, d_Z, T)$ as a measure of the massiveness of T, are quite useful for this purpose. However the quantity $\|Z(t)\|_T$ need not be a random variable. We avoid these measurability problems by assuming T is countable. The following result is known in the literature as *Sudakov's minorization.*

Theorem 2.1 *Let $\{Z_t : t \in T\}$ be a centered Gaussian process. Then for some numerical constant $C > 0$, we have*

$$\mathbb{E}\|Z_t\|_T \geq C \sup_{\delta>0} \delta\sqrt{H(\delta, d_Z, T)}. \tag{2.3}$$

An upper bound for $\mathbb{E}\|Z_t\|_T$ is given next in the following theorem of Dudley [9].

Theorem 2.2 *Let $Z = \{Z_t : t \in T\}$ be a centered Gaussian process. Then for some numerical constant $C > 0$, we have*

$$\mathbb{E}\|Z_t\|_T \leq C \int_0^\Delta \sqrt{H(\delta, d_Z, T)}\, d\delta, \tag{2.4}$$

where $\Delta = \sup_{s,t\in T} d_Z(s,t)$. Moreover Z has a version with almost all sample paths $Z_t = Z(\omega, t)$ bounded and uniformly continuous on (T, d_Z), provided the entropy integral on the right in (2.4) is finite.

Proofs of Theorems 2.1 and 2.2 are given in e.g. Ledoux & Talagrand [27]. Later on we shall see that (2.4) holds for more general processes satisfying a certain Lipschitz condition.

2.3 Subgaussian processes

Let (T, d) be a pseudo metric space. A stochastic process $Z = \{Z_t : t \in T\}$ is called *subgaussian* if

$$\mathbb{P}\{|Z_s - Z_t| > x\} \leq 2\exp\left(-\frac{1}{2}x^2 \Big/ d^2(s,t)\right) \tag{2.5}$$

holds true for all $s, t \in T$ and $x > 0$. The constants 2 and $1/2$ in the definition are irrelevant; they may be replaced by different positive constants. In many applications one is interested in local suprema of random processes. Let $t_0 \in T$, $\delta > 0$ and $T(\delta) = \{t \in T \mid d(t, t_0) \leq \delta\}$. Under nice behavior of the entropy numbers of (T, d), the following theorem states that the tails of the probability distribution of the quantity $\|Z_t\|_{T(\delta)}$ decrease exponentially fast. The proof is almost the same as the proof of Theorem 3.3 in Van de Geer [41].

Theorem 2.3 *Let $Z = \{Z(t) : t \in T\}$ be a subgaussian process with continuous sample paths. Then for some numerical constant $\kappa > 0$, we have*

$$\mathbb{P}\left\{\|Z(t) - Z(t_0)\|_{T(\delta)} \geq x\delta^2\right\} \leq 2(1+\varepsilon)\exp\left(-\frac{1}{144}x^2\delta^2\right) \tag{2.6}$$

for

$$x\delta \geq \kappa \int_0^2 \sqrt{H(\delta w, d, T(\delta))}\, dw \vee 12\sqrt{\log\frac{1+\varepsilon}{\varepsilon}}. \tag{2.7}$$

Proof. Set $\delta_k = 2^{-k}\delta$, $T^{(0)} = \{t_0\}$ and let $T^{(k)}$ be a minimal δ_k-covering net for $T(\delta)$, $k = 1, 2, \ldots$. By continuity of the sample paths, we may write

$$Z(t) - Z(t_0) = \sum_{k=1}^{\infty}\left(Z(t^{(k)}) - Z(t^{(k-1)})\right),$$

where $t^{(k)} \in T^{(k)}$ with $d(t, t^{(k)}) \leq \delta_k$ for each $t \in T(\delta)$. Next, we define

$$\eta_k = \frac{1}{2} \max \left(\frac{12\sqrt{2}\sqrt{H(\delta_k, d, T(\delta))}}{2^k x \delta}, \frac{\sqrt{k}}{2^k E} \right), \quad k = 1, 2, \ldots$$

with $E = \sum_{k=1}^{\infty} 2^{-k}\sqrt{k}$. For $x \geq 12\sqrt{2}\delta^{-2} \sum_{k=1}^{\infty} \delta_k \sqrt{H(\delta_k, d, T(\delta))}$, the series $\sum_{k=1}^{\infty} \eta_k \leq 1$, which implies

$$\begin{aligned} &\mathbb{P}\left\{ \|Z(t) - Z(t_0)\|_{T(\delta)} > x\delta^2 \right\} \leq \\ &\leq \sum_{k=1}^{\infty} \mathbb{P}\left\{ \|Z(t^{(k)}) - Z(t^{(k-1)})\|_{T(\delta)} > \eta_k x \delta^2 \right\}. \end{aligned}$$

Observe what has happened: the supremum is now taken over only at most $N(\delta_k, d, T(\delta)) \cdot N(\delta_{k-1}, d, T(\delta)) \leq \{N(\delta_k, d, T(\delta))\}^2$ elements. Hence

$$\begin{aligned} &\mathbb{P}\left\{ \|Z(t) - Z(t_0)\|_{T(\delta)} > x\delta^2 \right\} \leq \\ &\leq 2 \sum_{k=1}^{\infty} \exp\left(2H(\delta_k, d, T(\delta)) - \frac{1}{18} 2^{2k} \eta_k^2 \delta^2 x^2 \right) \leq \\ &\leq 2 \sum_{k=1}^{\infty} \exp\left(-\frac{1}{36} 2^{2k} \eta_k^2 \delta^2 x^2 \right) \leq 2 \sum_{k=1}^{\infty} \exp\left(-\frac{1}{36} \delta^2 x^2 E^{-2} k \right), \end{aligned}$$

where we used the subgaussian property of each single Z_t, the definition of η_k and the fact that $d(t^{(k)}, t^{(k-1)}) \leq \delta_k + \delta_{k-1} = 3\delta_k$ by the triangle inequality. Using the crude bound $E \leq 2$, the property of the geometric series

$$\sum_{k=1}^{\infty} \exp\left(-\left(\frac{x\delta}{12}\right)^2 k \right) = \frac{\exp\left(-\left(\frac{x\delta}{12}\right)^2 \right)}{1 - \exp\left(-\left(\frac{x\delta}{12}\right)^2 \right)}$$

and the inequality

$$(1 - \exp(-z))^{-1} \leq 1 + \eta \Longleftrightarrow z \geq \log\left(\frac{1+\eta}{\eta} \right),$$

for all $z > 0, \eta > 0$, we obtain (2.6). Using the monotonicity property of the entropy numbers, we may replace the infinite series $\sum_{k=1}^{\infty} \delta_k \sqrt{H(\delta_k, d, T(\delta))}$ by the corresponding integral (cf. Pollard [37], p. 12). The proof is complete. □

Remark 2.1 If the entropy integral $\int_0^2 \sqrt{H(x\delta, d, T(\delta))}\, dx$ makes sense, then $(T(\delta), d)$ is totally bounded. Hence there exists a sequence of countable dense subsets converging to $T(\delta)$. Because of the exponential decrease for the differences $|Z(s) - Z(t)|$, we can apply the Borel-Cantelli lemma to show that $\{Z(t) : t \in T(\delta)\}$ has uniformly continuous sample paths with probability one.

Suppose the stochastic process Z fulfills a first order Lipschitz condition

$$|Z(s) - Z(t)| \leq \alpha d(s,t), \quad s,t \in T \quad \text{(a.s.)} \tag{2.8}$$

Then the chaining can be stopped after finitely many steps and hence the entropy integral condition can be slightly weakened. Problems for the entropy numbers $H(\delta, d, T)$ typically arise in the neighborhood of $\delta = 0$.

Corollary 2.1 *Let $Z = \{Z(t) : t \in T\}$ be a subgaussian process with continuous sample paths and let (2.8) hold true for some $\alpha > 0$. Then for some numerical constants $\kappa_1, \kappa_2, \kappa_3 > 0$, we have*

$$\mathbb{P}\left\{\|Z(t) - Z(t_0)\|_{T(\delta)} \geq x\delta^2\right\} \leq 2(1+\varepsilon)\exp(-\kappa_1 x^2 \delta^2) \tag{2.9}$$

for

$$x\delta \geq \kappa_2 \int_{x\delta/4\alpha}^2 \sqrt{H(\delta w, d, T(\delta))}\, dw \vee \kappa_3 \sqrt{\log\left(\frac{1+\varepsilon}{\varepsilon}\right)}. \tag{2.10}$$

Proof. Define $\delta_k = 2^{-k}\delta$ as before and set

$$L = \inf\left\{k : \delta_k \leq \frac{x\delta^2}{2\alpha}\right\}.$$

Then by the triangle inequality we have with probability one

$$\begin{aligned} |Z(t) - Z(t_0)| &\leq \left|Z(t^{(L)}) - Z(t)\right| + \sum_{k=1}^{L} \left|Z(t^{(k)}) - Z(t^{(k-1)})\right| \\ &\leq \frac{x\delta^2}{2} + \sum_{k=1}^{L} \left|Z(t^{(k)}) - Z(t^{(k-1)})\right|. \end{aligned}$$

Consequently we have

$$\mathbb{P}\left\{\|Z(t) - z(t_0)\|_{T(\delta)} \geq x\delta^2\right\} \leq$$
$$\leq \sum_{k=1}^{L} \mathbb{P}\left\{\left\|Z(t^{(k)}) - Z(t^{(k-1)})\right\|_{T(\delta)} \geq \eta_k \frac{x\delta^2}{2}\right\}.$$

for all sequences η_k satisfying $\sum_{k=1}^{L} \eta_k \leq 1$. Define η_k as in the proof of the previous theorem and proceed in the same way. The claim follows. □

This result is included in Van de Geer [45] although no proof of the corollary is provided. Because both results of this section will be of importance in this book, e.g. to derive rates of convergence for the least squares estimator in nonparametric regression, we have included their full proofs.

2.4 Orlicz norms

Let $Z = \{Z(t) : t \in T\}$ be a random process indexed by some pseudo metric space (T, d). We have already seen in Section 2.2 that for Gaussian processes almost sure boundedness and continuity of the sample paths $Z(\omega, t)$ in (T, d) could be stated in terms of metric entropy conditions. Also the interesting quantity $\mathbb{E}\sup_{t\in T}|Z(t)|$ can be estimated using the metric entropy of (T, d). For the same reasons as in Section 2.2, we restrict our attention to countable T. Here, we shall generalize these results to stochastic processes satisfying a Lipschitz condition in some Orlicz space. Let us therefore recall the definition of such spaces.

Definition 2.4 (Orlicz space) *Let ψ be a convex, increasing function on $\mathbb{R}^+$ with $\psi(0) = 0$ and $\lim_{x\to\infty}\psi(x) = \infty$, a so-called* Young function.

An Orlicz space $L_\psi = L_\psi(\Omega, \mathcal{A}, \mathbb{P})$ *is the vector space of all random variables X on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$, for which*

$$\mathbb{E}\psi\left(\frac{|X|}{c}\right) < \infty, \quad \textit{holds, for some } c > 0.$$

Remark 2.2 We mention some properties of these spaces.

- With respect to the norm

$$\|X\|_\psi = \inf\left[c > 0 : \mathbb{E}\psi\left(\frac{|X|}{c}\right) < 1\right], \tag{2.11}$$

  L_ψ forms a Banach space.

- Since a Young function ψ is convex, we have as a result of Jensen's inequality

$$\psi\left(\frac{\mathbb{E}|X|}{c}\right) \leq \mathbb{E}\psi\left(\frac{|X|}{c}\right).$$

  It is now not difficult to see that $L_\psi \subset L^1$.

- If we take $\psi(x) = x^p$, $1 \leq p < \infty$, the Orlicz space L_ψ is just the usual L^p. If $\psi(x) = \exp(x^p) - 1$, then a bound in the corresponding Orlicz norm is stronger than in the L^p-norm.

In the next theorem, we study random processes $\{Z(t) : t \in T\}$ satisfying a Lipschitz condition in L_ψ, i.e. for some $C > 0$

$$\|Z(s) - Z(t)\|_\psi \leq Cd(s,t), \tag{2.12}$$

where the norm is defined in (2.11). Obviously, for any sequence random variables $X_1, \ldots, X_n$,

$$\left\|\max_{1\leq i\leq n} X_i\right\|_p \leq n^{1/p} \max_{1\leq i\leq n} \|X_i\|_p$$

holds and similarly we have a bound for general Orlicz norms. Under some regularity on ψ, i.e.

$$\limsup_{x,y\to\infty} \psi(x)\psi(y)/\psi(xy) < \infty, \tag{2.13}$$

the following is true (cf. Van der Vaart & Wellner [49]):

$$\left\|\max_{1\leq i\leq n} |X_i|\right\|_\psi \leq K_\psi \psi^{-1}(n) \max_{1\leq i\leq n} \|X_i\|_\psi,$$

where K_ψ is a constant depending only on ψ. This and a chaining argument are the main ingredients of the following theorem. In contrast to Theorem 2.2, we estimate the L_ψ norm of $\sup_{t\in T} |Z(t)|$ rather than its L^1 norm.

Theorem 2.4 *Let $X = \{X_t : t \in T\}$ be a process with continuous sample paths and increments controlled by (2.12). Let ψ fulfill condition (2.13). Then for some constant C_ψ*

$$\left\| \sup_{t \in T} |X_t| \right\|_\psi \leq C_\psi \int_0^\Delta \psi^{-1}(N(\delta, d, T)) d\delta, \tag{2.14}$$

where Δ is the diameter of the set T.

Proof. See for instance Pollard [37], Ledoux & Talagrand [27], Van der Vaart & Wellner [49]. □

With the aid of Theorem 2.4 and the basic Markov inequalities, it is an easy matter to derive probability bounds for suprema of general stochastic processes satisfying the Lipschitz condition (2.12) in the Orlicz space L_ψ. For any random variable Z in L_ψ, we have

$$\mathbb{P}\{|Z| > z\} \leq \frac{\mathbb{E}\psi(|Z|/C)}{\psi(|z|/C)} \leq \frac{1}{\psi(|z|/C)} \tag{2.15}$$

for $C > \|Z\|_\psi$, since ψ is increasing. In particular, if ψ is an exponential function, the tail of Z decreases exponentially fast. It can be shown, essentially by an argument based on Fubini's theorem, that the converse of this is also valid.

Lemma 2.1 *Let X be a random variable satisfying*

$$\mathbb{P}(|X| > x) \leq K \exp(-Cx^p) \quad \forall x > 0$$

and some constants $K, C > 0$ and $p \geq 1$. Then

$$\|X\|_\psi \leq \left(\frac{1+K}{C}\right)^{1/p}.$$

Proof. See Van der Vaart & Wellner [49]. □

Gaussian as well as subgaussian random processes are included in Theorem 2.4 as will be illustrated in the following examples.

Example 2.4 (Gaussian processes) Let $\psi(x) = \exp(x^2) - 1$. Let $\{X_t : t \in T\}$ be a centered Gaussian process, parametrized by (T, d), where

$$d(s,t) = d_X(s,t) = \left(\mathbb{E}|X_s - X_t|^2\right)^{1/2}.$$

Write for simplicity $Z = X_s - X_t$, $\sigma^2 = \sigma^2(Z)$. Then clearly Z is $N(0, \sigma^2)$ distributed. As

$$\begin{aligned} \mathbb{E}\psi\left(\frac{|Z|}{c}\right) + 1 &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}z^2\left(-2\left(\frac{1}{c^2} - \frac{1}{2\sigma^2}\right)\right)\right) dz \\ &= \frac{c}{\sqrt{c^2 - 2\sigma^2}}, \end{aligned}$$

we have

$$\|X_s - X_t\|_\psi \leq \sqrt{\frac{8}{3}}\, d_X(s,t),$$

so that (2.12) is satisfied.

Example 2.5 (Subgaussian processes) Let X_t be subgaussian random variables, i.e.

$$\mathbb{P}\{|X_s - X_t| > x\} \leq 2\exp\left(-\frac{1}{2}x^2 \Big/ d^2(s,t)\right).$$

Then (2.12) is again satisfied. To see this, apply Lemma 2.1 with the constants C, K and p chosen as follows: $1/C = 2d^2(s,t)$, $K = 2$, $p = 2$. It follows that

$$\|X_s - X_t\|_\psi \leq \sqrt{6}\, d(s,t).$$

2.5 Empirical processes

Let $X_1, \ldots, X_n$ be independent, identically distributed random variables defined on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with values in $(S, \mathcal{S}, P)$. The empirical measure P_n is usually defined by

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}, \quad n \in \mathbb{N},$$

where δ_a is the Dirac measure at point a. Let $\mathcal{F}$ be a collection of measurable, real valued, P-integrable functions on $(S, \mathcal{S})$. We are interested in the *empirical process*

$$f \mapsto \sqrt{n}(P_n - P)(f) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f(X_i) - \mathbb{E}_P f(X_i)), \quad f \in \mathcal{F}, \ n \geq 1.$$

We agree $\mathcal{F}$ will be permissible in the sense of Pollard [35] in order to cope with measurability problems.

2.5.1 Maximal inequalities

We are often interested in the (local) behavior of the empirical process $\{(P_n - P)(f) : f \in \mathcal{F}\}$. For instance, in probability theory it is the main ingredient of the proofs of uniform laws of large numbers and functional central limit theorems. As a result, it plays an important part in nonparametric statistics.

Let $\mathcal{F}_n \subset \mathcal{F}$. We are interested in obtaining sharp bounds for

$$\mathbb{P}\left\{ \sup_{f \in \mathcal{F}_n} \left| \int f \, d(P_n - P) \right| \geq \delta \right\}. \tag{2.16}$$

In general, the theory in Sections 2.2 - 2.4 can not be applied but similar results hold true under conditions on $\mathcal{F}_n$.

It has become a standard argument in the theory of empirical processes to symmetrize the process $P_n - P$, then to introduce additional randomness and finally to study the new process conditionally on the old randomness. First we present the Symmetrization lemma as given in Pollard [35]. Slightly different versions can be found in e.g. Giné & Zinn [15].

Lemma 2.2 (Symmetrization lemma) *Let $\{Z(t) : t \in T\}$ and $\{\tilde{Z}(t) : t \in T\}$ be independent stochastic processes sharing an index set T. Suppose there exist constants $\beta > 0$ and $\alpha > 0$ such that $\mathbb{P}\{|\tilde{Z}(t)| \leq \alpha\} \geq \beta$ for every $t \in T$. Then*

$$\mathbb{P}\left\{ \sup_t |Z(t)| > \varepsilon \right\} \leq \beta^{-1} \mathbb{P}\left\{ \sup_t |Z(t) - \tilde{Z}(t)| > \varepsilon - \alpha \right\}.$$

Proof. See Pollard [35]. □

Using this lemma we find, provided $\mathrm{Var}(f(X_1)) \leq n\delta^2/8$ for all $f \in \mathcal{F}_n$,

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}_n}\left|\int f\,d(P_n - P)\right| > \delta\right\} \leq 4\mathbb{P}\left\{\sup_{f\in\mathcal{F}_n}\left|\int f\,dP_n^0\right| > \frac{\delta}{4}\right\} \quad (2.17)$$

with

$$P_n^0 = \frac{1}{n}\sum_{i=1}^{n}\sigma_i\delta_{X_i}, \quad (2.18)$$

where the signs σ_i are independent random variables with $\mathbb{P}\{\sigma_i = -1\} = \mathbb{P}\{\sigma_i = 1\} = 1/2$ and are independently chosen of the observations $X_1, \ldots, X_n$. In e.g. Pollard [35] this has been worked out in full detail. The sequence σ_i is called a *Rademacher* sequence. Note that conditionally on the stochastic vector $(X_1, \ldots, X_n)$, the symmetrized empirical process $\{\sqrt{n}P_n^0(f) : f \in \mathcal{F}_n\}$ is subgaussian with respect to the $L^2(P_n)$ pseudo norm by Hoeffding's inequality:

Lemma 2.3 (Hoeffding's inequality) *Let Z_i be independent random variables with $a_i \leq Z_i \leq b_i$. Then, for all $\lambda > 0$,*

$$\mathbb{P}\left\{\left|\sum_{i=1}^{n}(Z_i - \mathbb{E}Z_i)\right| > \lambda\right\} \leq 2\exp\left(-2\lambda^2/\sum_{i=1}^{n}(b_i - a_i)^2\right).$$

Proof. See Hoeffding [21]. □

Moreover $\sqrt{n}P_n^0$ has continuous sample paths in the $L^2(P_n)$ pseudo norm. This follows from the Cauchy-Schwarz inequality

$$\left|\int (f-g)\,dP_n^0\right| \leq \left(\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - g(X_i))^2\right)^{1/2}. \quad (2.19)$$

This has important consequences because it allows us to apply the theory developed in Sections 2.3 and 2.4. We summarize these results in the following lemmas.

Lemma 2.4 *The symmetrized empirical process $\{\sqrt{n}P_n^0(f) : f \in \mathcal{F}\}$ with P_n^0 given by (2.18) is, conditionally on $X_1, \ldots, X_n$, subgaussian with respect to the $L^2(P_n)$ pseudo norm and almost all sample paths are continuous.*

Proof. The subgaussian property is implied by Lemma 2.3 and the continuity of the sample paths follows from (2.19). □

Lemma 2.5 *Let $\mathcal{F}_n \subset \mathcal{F}$ and $V_n^2 = \sup_{f\in\mathcal{F}_n} \mathrm{Var}(f(X_1))$. Then, for all $\delta > V_n 2\sqrt{2}/\sqrt{n}$, and for some constant $C > 0$, independent of n,*

$$\begin{aligned} &\mathbb{P}\left\{ \sup_{f\in\mathcal{F}_n} \left| \int f\, d(P_n - P) \right| \geq \delta \right\} \leq \\ &\leq 4\mathbb{P}\left\{ \sup_{f\in\mathcal{F}_n} \left| \int f\, dP_n^0 \right| \geq \frac{\delta}{4} \right\} \\ &\leq 8\mathbb{E}_X \exp\left(- \frac{Cn\delta^2}{\left[\int_0^{\Delta_n} \sqrt{H(x, d_n, \mathcal{F}_n)}\, dx\right]^2} \right), \end{aligned} \tag{2.20}$$

where $\mathbb{E}_X$ is the expectation with respect to $X = (X_1, \ldots, X_n)$, d_n is the $L^2(P_n)$ pseudo norm and Δ_n is the diameter of $\mathcal{F}_n$ with respect to d_n.

In the special case where $\mathcal{F}_n = \{f \in \mathcal{F} \mid d_n(f, f_0) \leq \delta\}$ with $f_0 \in \mathcal{F}$, we have for some numerical constants $\kappa_1, \kappa_2, \kappa_3 > 0$,

$$\begin{aligned} &\mathbb{P}\left\{ \sup_{f\in\mathcal{F}_n} \left| \int (f - f_0)\, d\sqrt{n}P_n^0 \right| \geq x\sqrt{n}\delta^2 \;\middle|\; X \right\} \leq \\ &\leq 2(1+\varepsilon)\exp(-\kappa_1 x^2 n\delta^2) \end{aligned} \tag{2.21}$$

for

$$x\sqrt{n}\delta \geq \kappa_2 \int_{x\delta/4\sqrt{6}}^2 \sqrt{H(\delta t, d_n, \mathcal{F}_n)}\, dt \vee \kappa_3 \sqrt{\log\left(\frac{1+\varepsilon}{\varepsilon}\right)}.$$

Proof. Because $V_n^2 < n\delta^2/8$, we can apply Lemma 2.2. Consequently, (2.17) is true. This implies the symmetrization step in (2.20). By Lemma 2.4, the conditions of Theorem 2.4 hold true for the index set $\mathcal{F}_n$, pseudo metric d_n, conditional empirical process $\int f\, d\sqrt{n}P_n^0$, and Young function $\psi(x) = \exp(x^2) - 1$. Hence the conclusion of Theorem 2.4 holds. Applying the Markov inequality (2.15), we obtain (2.20).

The second assertion (2.21) is a direct consequence of Corollary 2.1. □

Another possible strategy to obtain upper bounds for (2.16) employs Bernstein's inequality and entropy with bracketing instead of Hoeffding's

inequality and random entropy. We refer to the work of Alexander [1], Birgé & Massart [5], Van de Geer [44], Ossiander [33] for details.

2.5.2 Glivenko Cantelli classes

We call a class $\mathcal{F} \subset L^1(P)$ a *Glivenko-Cantelli class* (or a P-Glivenko-Cantelli class) if the empirical measure P_n tends to the theoretical probability measure P with probability one, uniformly in $\mathcal{F}$, i.e

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - \mathbb{E}_P f(X_i)) \right| \xrightarrow{a.s.} 0. \tag{2.22}$$

Necessary for this convergence is the existence of an P-integrable envelope F for $\mathcal{F}$. Recall that the envelope is defined by

$$F(s) = \sup_{f\in\mathcal{F}} |f(s)|.$$

Theorem 2.5 *Suppose* $\sup_{f\in\mathcal{F}} \int |f|\, dP < \infty$. *Necessary and sufficient conditions for* $\mathcal{F}$ *being a Glivenko-Cantelli class are*

$$\int F\, dP < \infty \quad \text{and} \quad \frac{H(\delta, d_{n,1}, \mathcal{F})}{n} \xrightarrow{P} 0 \text{ for all } \delta > 0, \tag{2.23}$$

where

$$d_{n,1}(f,g) = \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - g(X_i)|$$

is the $L^1(P_n)$ *pseudo distance between two functions.*

Proof. See Giné & Zinn [14]. □

2.5.3 Donsker classes

Assume $\mathcal{F} \subset L^2(P)$ and $\sup_{f\in\mathcal{F}} |f(s) - \int f\, dP| < \infty$ for every $s \in S$. The empirical process $E_n = \sqrt{n}(P_n - P)$ is a mapping from Ω into $\ell^\infty(\mathcal{F})$, the space of all bounded, real valued functions on $\mathcal{F}$, equipped with the supremum norm $\|\cdot\|_{\mathcal{F}}$. By definition, a *P-Donsker class* $\mathcal{F}$ fulfills the Functional central limit theorem, i.e. there exists a centered Gaussian process G such that

$$\mathbb{E}^* \phi(E_n) \to \mathbb{E}\phi(G)$$

for every bounded, continuous, real valued function ϕ on $\ell^\infty(\mathcal{F})$. Here $\mathbb{E}^*$ denotes the outer expectation. By the Finite dimensional central limit theorem, the covariance matrix of the limiting Gaussian process $G = G(P)$ is given by

$$\mathbb{E}G(f)G(g) = \int fg\,dP - \int f\,dP \cdot \int g\,dP, \quad f, g \in \mathcal{F}.$$

A P-Donsker class is characterized by an asymptotic equicontinuity condition.

Theorem 2.6 *Assume* $\sup_{f\in\mathcal{F}} \int |f|\,dP < \infty$. *Then* $\mathcal{F}$ *is* P*-Donsker if and only if* $(\mathcal{F}, d_2)$ *is totally bounded and for each* $\varepsilon > 0$ *there is some* $\eta > 0$ *such that*

$$\limsup_{n\to\infty} \mathbb{P}\left\{ \sup_{f\in\mathcal{F}_\eta} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i f(X_i) \right| > \varepsilon \right\} < \varepsilon, \tag{2.24}$$

where $\mathcal{F}_\eta = \{f - g \mid f, g \in \mathcal{F},\ d_2(f,g) < \eta\}$ *is a class of differences,* ξ_i *is a Rademacher sequence, and* d_2 *the* $L^2(P)$ *pseudo distance, i.e.*

$$d_2(f,g) = \left(\int (f-g)^2\,dP \right)^{1/2}.$$

Proof. See Dudley [9], [10]. □

The fact that we assumed $\sup_{f\in\mathcal{F}} \int |f|\,dP < \infty$ allows us to take the simpler $L^2(P)$ pseudo distance instead of the pseudo metric $\tau_P(f,g) = d_2(f - \int f\,dP,\ g - \int g\,dP)$. Using the techniques outlined in Section 2.5.1, sufficient conditions can be formulated in terms of the random entropy numbers in $L^2(P_n)$. Usually an uniform entropy integral condition is given to ensure the asymptotic tightness condition (2.24).

Part II

Finite dimensional problems

Chapter 3

Estimating a parameter in Euclidean spaces

In this chapter we investigate the asymptotic behavior of the root of a general parametric random equation and of the statistic defined by minimization of a stochastic process. We shall mainly use techniques inherited from the theory of empirical processes. Another important tool will be the notion of Hadamard differentiability of functionals. We apply our results to M- and R-estimation. Minimum distance estimators will also be discussed.

3.1 Introduction

Let $\mathcal{P}$ be a collection probability measures defined on a measurable space $(S, \mathcal{S})$. We have independent, identically distributed (i.i.d.) observations $X_1, \ldots, X_n$ with a common distribution $P \in \mathcal{P}$ at our disposal. We are interested, not so much in this entire unknown probability measure, but only in the true value θ_0 of some finite dimensional parameter θ. It is not necessary that P is completely specified by θ_0. For instance, we have in mind the case that θ is a location parameter.

In this chapter we investigate the asymptotic behavior of estimators $\hat{\theta}_n$ for θ_0. Two closely related situations will be considered. In the first, developed in the next section, one assumes that the parameter of interest is implicitly given as the solution of some equation $M(\theta) = 0$. Obviously

the function M will depend on P and we emphasize this dependence by writing $M(\theta) = \gamma(\theta; P)$. The empirical probability measure is defined by $P_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i}$. We obtain the empirical counterpart of M by plugging in P_n for P, i.e. $M_n(\theta) = \gamma(\theta; P_n)$. One then searches an estimator $\hat{\theta}_n$ which solves the equation $M_n(\theta) = 0$. In Section 3.3, θ_0 is given as the value at which some function M reaches its minimum.

There is a considerable literature available for special cases, especially the case where M has the simple form

$$M(\theta) = \gamma(\theta; P) = \int g(\cdot, \theta)\, dP$$

(M-estimation). See for instance Heesterman & Gill [20] and Pollard [35]. More general situations, including L- and R-estimation and Cramer-von Mises estimation, are treated - among others - in Fernholz [12] and Stute [38]. Infinite dimensional parameter spaces in the context of M-estimation are discussed in a recent paper of Van der Vaart [48].

The aim of this chapter is to study the asymptotic behavior of the statistical error $\hat{\theta}_n - \theta_0$ under simple conditions on M and M_n. The notion of stochastic equicontinuity will be appropriate in Section 3.2, whereas stochastic differentiability will be the key in Section 3.3. Examples will be given to clarify how empirical process theory can be used for checking these properties.

3.2 Estimation equations

Let $X_1, \ldots, X_n$ be an i.i.d. sequence with probability measure $P \in \mathcal{P}$ and let $\theta = \theta(P)$ be a parameter taking values in $\Theta \subset \mathbb{R}^k$. By θ_0 we denote the true value of the parameter. We consider some function $\gamma : \Theta \times \mathcal{P} \to \mathbb{R}^k$ and abbreviate $\gamma(\theta; P)$ by $M(\theta)$. We impose the following conditions on M:

(A1) $M(\theta) = 0$ if and only if $\theta = \theta_0$;

(A2) M is a local homeorphism at θ_0;

(A3) M is differentiable at θ_0 with a non-singular derivative M' at θ_0.

Let P_n be the empirical probability measure based on $X_1, \ldots, X_n$ and $\gamma(\theta; P_n) = M_n(\theta)$. We consider sequences $\hat{\theta}_n$ for which

$$M_n(\hat{\theta}_n) = o_P(n^{-\frac{1}{2}}), \tag{3.1}$$

and assume

(A4) There actually exist solutions $\hat{\theta}_n$ such that (3.1) holds true.

We consider $\sqrt{n}(M_n - M)$ as a random process in $l^\infty(\Theta)$, the space of all real valued bounded functions on Θ. This space is equipped with the uniform metric $\|\cdot\|$. Before we give restrictions on $\sqrt{n}(M_n - M)$, we recall the concept of stochastic equicontinuity.

Definition 3.1 (stochastic equicontinuity) *Let (T, d) be a pseudo-metric space and let $\{Z_n(t) : t \in T\}$ be a stochastic process, indexed by T. A sequence $\{Z_n\}$ is called* stochastically equicontinuous *at $t_0 \in T$, iff $\forall \eta > 0$ and $\forall \varepsilon > 0$ there exists a neighborhood V of t_0 for which*

$$\limsup_{n\to\infty} \mathbb{P}\left\{\sup_{t\in V} |Z_n(t) - Z_n(t_0)| > \eta\right\} \leq \varepsilon.$$

Equivalently $\{Z_n\}$ is called stochastically equicontinuous at $t_0 \in T$, if for any $\tau_n \xrightarrow{P} t_0$, we have $|Z_n(\tau_n) - Z_n(t_0)| \xrightarrow{P} 0$.

We make the following assumptions on $\alpha_n = \sqrt{n}(M_n - M)$:

(A5) $\|n^{-1/2}\alpha_n\| \to 0$ in probability;

(A6) α_n is stochastically equicontinuous at θ_0;

(A7) $\alpha_n(\theta_0)$ converges in distribution to some probability measure $L = L(\theta_0; P)$.

Theorem 3.1 *Under the conditions A1,...,A7, we have*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} -(M'(\theta_0))^{-1} \cdot L. \tag{3.2}$$

Proof. We first prove consistency of $\hat{\theta}_n$ for which $M_n(\hat{\theta}_n) = o_P(n^{-\frac{1}{2}})$. Using the uniform convergence A5, we have

$$M(\hat{\theta}_n) = M_n(\hat{\theta}_n) - (M_n - M)(\hat{\theta}_n) = o_P(1). \tag{3.3}$$

Since the function M is a local homeorphism at the solution θ_0 of $M(\theta) = 0$, we have

$$\hat{\theta}_n = M^{-1}(M(\hat{\theta}_n)) \xrightarrow{P} M^{-1}(0) = \theta_0.$$

The remainder of the proof is due to Van der Vaart [48]. As a consequence of A3, we have

$$M(\hat{\theta}_n) = M'(\theta_0)(\hat{\theta}_n - \theta_0) + o_P(|\hat{\theta}_n - \theta_0|). \tag{3.4}$$

Because α_n is stochastically equicontinuous at θ_0 and $\hat{\theta}_n \xrightarrow{P} \theta_0$, we have $\alpha_n(\hat{\theta}_n) = \alpha_n(\theta_0) + o_P(1) = \mathcal{O}_P(1)$. This and (3.3) imply $M(\hat{\theta}_n) = \mathcal{O}_P(n^{-\frac{1}{2}})$ and therefore we derive $|\hat{\theta}_n - \theta_0| = \mathcal{O}_P(n^{-\frac{1}{2}})$. Hence we only need to investigate $\sqrt{n}M(\hat{\theta}_n)$.

$$\begin{aligned} M(\hat{\theta}_n) &= -(M_n - M)(\hat{\theta}_n) + M_n(\hat{\theta}_n) \\ &= -(M_n - M)(\hat{\theta}_n) + (M_n - M)(\theta_0) - (M_n - M)(\theta_0) + M_n(\hat{\theta}_n) \\ &= -(M_n - M)(\theta_0) + o_P(n^{-\frac{1}{2}}), \end{aligned}$$

where we used in the last step that α_n is stochastically equicontinuous at θ_0 (condition A6.). Finally we use the weak convergence A7 to complete our proof. □

Remark 3.1 In full generality it is difficult to make a statement about the existence of a solution $\hat{\theta}_n$ of the estimation equation $M_n(\theta) = 0$. Let $C_b(\Theta)$ be the space of all bounded, continuous functions on Θ, equipped with the uniform metric $\|\cdot\|$. In case both M and M_n are in $C_b(\Theta)$, there exists a solution $\hat{\theta}_n$ with probability tending to one by the following reasoning.

Let $m : \Theta \to \mathbb{R}^k$ be continuous and inside the ball $B(M, \varepsilon)$ for some $\varepsilon > 0$. Consider the mapping $x \mapsto x - m \circ M^{-1}(x)$, which maps

the Euclidean ball $B(0,\varepsilon)$ continuously into itself. By Brouwer's fixed point theorem, there exists at least one $x_m \in B(0,\varepsilon)$ such that $x_m = x_m - m \circ M^{-1}(x_m)$. Since $M_n \in C_b(\Theta)$ and $\mathbb{P}\{\|M_n - M\| > \varepsilon\} \to 0$ for every $\varepsilon > 0$,

$$\mathbb{P}\left\{M_n(\hat{\theta}_n) = 0\right\} \to 1, \text{ as } n \to \infty.$$

This argument can be found in Heesterman & Gill [20].

M is a local homeorphism at θ_0 if, by definition, M is continuous on a small neighborhood of θ_0 and has a continuous inverse on a small neighborhood of $M(\theta_0) = 0 \in \mathbb{R}^k$. A sufficient condition is that M is continuous and one-to-one and the domain Θ is compact. Unfortunately in many cases $M_n \notin C(\Theta)$ and it is not so obvious that the parameter space to be considered is a priori compact. Often, but not always, see e.g. Example 3.1 and 3.2, this is implied by assumption A5.

Remark 3.2 The proof of Theorem 3.1 is in the spirit of Stute [38], although in that paper the more explicit integral type of estimation equation is considered, viz.

$$\tilde{\gamma}(\theta, F) = \int \psi(x, \theta, F)\, dF(x)$$

where F is the probability distribution function associated with P. Stute assumes a smoothness property (Fréchet differentiability) of ψ in its third variable F. As we shall soon see, this kind of differentiability can be replaced by the weaker form of Hadamard differentiability.

Remark 3.3 For a different approach to the asymptotic behavior of $\hat{\theta}_n - \theta_0$, we refer to e.g Fernholz [12], Heesterman & Gill [20], Van der Vaart & Wellner [49], who make use of the concept of Hadamard differentiability.

Definition 3.2 (Hadamard differentiability) *Let X and Y be normed, linear*
spaces. A function $f : X_0 \subset X \to Y$ is called Hadamard differentiable *at*

$x \in X$, tangentially at $H \subset X$, iff there exists a linear, bounded function $f'_x : X \to Y$ such that

$$\frac{f(x + t_n h_n) - f(x)}{t_n} \to f'_x(h) \tag{3.5}$$

for all $t_n \to 0$ and all $h_n \to h$ with $h \in H$ and $x + t_n h_n \in X_0$ for all n.

Define $\mathcal{Z} = \mathcal{Z}(\Theta, \mathbb{R}^k) \subset l^\infty(\Theta, \mathbb{R}^k)$ as the space of all bounded maps from Θ into $\mathbb{R}^k$ containing at least one zero. On this space we consider the functional $\phi : \mathcal{Z} \to \Theta$ which supplies a solution $\theta_z = \phi(z)$ to a given $z \in \mathcal{Z}$, i.e. $z(\phi(z)) = 0$ for every $z \in \mathcal{Z}$. It can be shown under A1, A2 and A3 that this functional ϕ is differentiable at M, in the sense of Hadamard, tangentially to the subspace $\mathcal{Z}_0$ consisting of all functionals in $\mathcal{Z}$ which are continuous at θ_0. The derivative is given by $d\phi(M) \cdot h = -(M'(\theta_0))^{-1} \cdot h(\theta_0)$. A proof can be found in Heesterman & Gill [20]. Suppose α_n converges weakly in the sense of Hoffmann-Jørgensen to L in $l^\infty(\Theta)$ and in addition assume that L has continuous sample paths at θ_0, then an application of the generalized delta method for linear spaces yields the same conclusion as Theorem 3.1.

Remark 3.4 Often the limit L is a normal distribution. Let $\mathcal{F}$ be a class of functions such that $\mathcal{F} \subset L^2(P)$ for all $P \in \mathcal{P}$. On this space $\mathcal{F}$ we define the variance pseudo norm (denoted by τ_P) under $P \in \mathcal{P}$. Let $l^\infty(\mathcal{F})$ be the space of all bounded, real valued functions on $\mathcal{F}$ (equipped with the sup-norm) and let $UC(\mathcal{F}, \tau_P)$ be the subspace of all uniformly continuous (with respect to τ_P), real valued functions on $\mathcal{F}$. Suppose now that $\gamma(\theta, \cdot) : \mathcal{P} \to \mathbb{R}$ is Hadamard differentiable at $P \in \mathcal{P}$, tangentially to $UC(\mathcal{F}, \tau_P)$ and uniformly in θ. In particular we have

$$\gamma(\theta, P_n) = \gamma(\theta, P) + d\gamma(\theta, P) \cdot (P_n - P) + R_n(\theta; P),$$

uniformly in θ, where

$$\sup_{\theta \in \Theta} |R_n(\theta, P)| = o_P\left(n^{-1/2}\right)$$

and $d\gamma(\theta, P)$ is a bounded linear operator on $UC(\mathcal{F}, \tau_P)$. In this situation, condition A5 is fulfilled if $\sup_{\theta \in \Theta} |d\gamma(\theta, P)(P_n - P)| \xrightarrow{P} 0$ and for

assumption A6 we need stochastic equicontinuity of $d\gamma(\theta, P)E_n$ at θ_0, with $E_n = \sqrt{n}(P_n - P)$. If $\mathcal{F}$ is P-Donsker, then E_n converges weakly to a P-Brownian bridge E in $l^\infty(\mathcal{F})$ with $\mathbb{P}\{E \in UC(\mathcal{F}, \tau_P)\} = 1$. Because $d\gamma(\theta_0, P)$ is a bounded linear operator on $UC(\mathcal{F}, \tau_P)$, condition A7 holds true.

We shall illustrate the results obtained so far by checking A5 - A7 in the following examples.

Example 3.1 (M-estimation) Consider $\gamma(\theta, P) = \int \psi(\theta, \cdot)\, dP$. We define $g(\cdot) = \psi(\theta, \cdot)$ and in particular we set $g_0(\cdot) = \psi(\theta_0, \cdot)$. Endow the set $\mathcal{G} = \{\psi(\theta, \cdot) : \theta \in \Theta\}$ with the $L^2(P)$ pseudo metric. We give sufficient entropy conditions on $\mathcal{G}$ which imply assumptions A5 - A7.

Condition A5 is fulfilled if and only if $\mathcal{G}$ is a Glivenko-Cantelli class. Next we check that $\sqrt{n}(M_n - M)$ is stochastically equicontinuous at θ_0, which is assumption A6. For this matter, we define

$$\mathcal{G}(\delta) = \{g - g_0 : \|g - g_0\|_2 \leq \delta\},\ \mathcal{G}_n(\delta) = \{g - g_0 : \|g - g_0\|_{n,2} \leq \delta\},$$

where $\|\cdot\|_2$ denotes the $L^2(P)$ pseudo metric and $\|\cdot\|_{n,2}$ the $L^2(P_n)$ pseudo metric. Then after the usual symmetrization tricks, described in Section 2.5.1, one gets

$$\begin{aligned} &\mathbb{P}\left\{\sup_{g \in \mathcal{G}(\delta)} \left|\int g\, dE_n\right| > \varepsilon\right\} \leq \qquad (3.6)\\ &\leq 4\mathbb{P}\left\{\sup_{g \in \mathcal{G}_n(2\delta)} \left|\int g\, dE_n^0\right| > \frac{\varepsilon}{4}\right\} + \mathbb{P}\left\{\mathcal{G}(\delta) \not\subseteq \mathcal{G}_n(2\delta)\right\}, \end{aligned}$$

where $E_n = \sqrt{n}(P_n - P)$, $E_n^0 = \sqrt{n}P_n^0$, and P_n^0 is the signed empirical measure as defined in Section 2.5.1. For some constant $C_\psi > 0$, we obtain after an application of Lemma 2.5,

$$\begin{aligned} &\mathbb{P}\left\{\sup_{g \in \mathcal{G}_n(2\delta)} \left|\int g\, dE_n^0\right| > \varepsilon\right\} \leq \qquad (3.7)\\ &\leq 2\mathbb{E}_X \exp\left(-\left(\varepsilon / C_\psi \int_0^{4\delta} \sqrt{H_2(x, P_n, \mathcal{G}_n(2\delta))}\, dx\right)^2\right), \end{aligned}$$

where $\mathbb{E}_X$ denotes the expectation with respect to $(X_1, \ldots, X_n)$, and $H_2(\delta, P_n, \mathcal{G})$ the δ-entropy with respect to $\|\cdot\|_{n,2}$ of the class $\mathcal{G}$. If

$$\delta \int_0^1 \sqrt{H_2(4\delta x, P_n, \mathcal{G}_n(2\delta))}\, dx \downarrow 0 \quad \text{for } \delta \downarrow 0,$$

uniformly in n, the right-hand side in (3.7) tends to zero. Hence the first term on the right in (3.6) can be made arbitrarily small.

Set $\mathcal{H} = \{f - g : f, g \in \mathcal{G}\}$. We want the second term on the right in (3.6) to be asymptotically negligible, i.e. $\mathbb{P}\{\mathcal{G}(\delta) \subseteq \mathcal{G}_n(2\delta)\} \to 1$. For this purpose, it is sufficient to show that $\mathcal{H}^2$ is a Glivenko-Cantelli class. Assume $G = \sup_{g \in \mathcal{G}} |g|$ is P-square integrable and note that

$$N_1(\delta \|2G\|_{n,2}^2, P_n, \mathcal{H}^2) \leq N_2(\delta \|G\|_{n,2}, P_n, \mathcal{H}) \leq \left(N_2(\frac{\delta}{2}\|G\|_{n,2}, P_n, \mathcal{G})\right)^2$$

and $\|G\|_{n,2} \xrightarrow{a.s.} \|G\|_2 < \infty$. Hence $\mathcal{H}^2$ is a Glivenko-Cantelli class if $H_2(\delta, P_n, \mathcal{G})/n \xrightarrow{P} 0$ for all $\delta > 0$.

Condition A7 is a mere application of the classical central limit theorem.

Example 3.2 (R-estimation) Let Θ be a subset in $\mathbb{R}$ and let $D = D[-\infty, \infty]$ be the space all right continuous functions with left-hand limits on $\overline{\mathbb{R}}$. As usual, we equip $\mathbb{R}$ with the Euclidean and D with the uniform metric. Let $X_1, \ldots, X_n$ be i.i.d. real valued random variables with a continuous, strictly increasing distribution function F satisfying

$$F(x) = 1 - F(2\theta_0 - x), \; x \in \mathbb{R}.$$

Define $\tilde{\gamma} : \Theta \times D_0 \to \mathbb{R}$ by

$$\tilde{\gamma}(\theta, G) = \int_{\mathbb{R}} J\left(\frac{G(x) + 1 - G(2\theta - x)}{2}\right) dG(x) \tag{3.8}$$

where $J : \mathbb{R} \to \mathbb{R}$ is a strictly increasing, continuously differentiable score function and the domain $D_0 \subset D$ is given by $D_0 = \{G \in D : \int |\, dG| \leq 3\}$. Note that the assumption of $\tilde{\gamma}(\theta_0, F) = 0$ entails that $\int_0^1 J(x)\, dx = 0$ (for example $J(x) = 2x - 1$). We have that $M(\theta) =$

$\tilde{\gamma}(\theta, F)$ is a local homeorphism at θ_0. In different notation, the mapping (3.8) becomes

$$\gamma(\theta, P) = \int J\left(\int g_{\theta,x}(y)\, dP(y)\right) dP(x),$$

where $g_{\theta,x}(y) = (I_{(-\infty,x]}(y) + I_{(2\theta-x,\infty)}(y))/2$. Now, put $\mathcal{G} = \{g_{\theta,x} : \theta \in \Theta, x \in \mathbb{R}\}$ and consider the class $\mathcal{P}$ as a subset of $l^\infty(\mathcal{G})$. We define the $\mathcal{G}$-indexed empirical process as $E_n = \sqrt{n}(P_n - P)$. Since $\mathcal{G}$ is a Donsker class, we have E_n converges weakly to a Gaussian process E in $l^\infty(\mathcal{G})$. We have $M(\theta) = \int J(\int g_{\theta,\cdot}(y)\, dP(y))\, dP$ and $M_n(\theta) = \int J(\int g_{\theta,\cdot}(y)\, dP_n(y))\, dP_n$. We show in Lemma 3.1 that the following linear expansion holds

$$M_n(\theta) = M(\theta) + n^{-\frac{1}{2}} d\gamma(\theta; P) \cdot E_n + o_P(n^{-\frac{1}{2}}), \tag{3.9}$$

uniformly in θ with

$$\begin{aligned} d\gamma(\theta; P) \cdot E_n &= \frac{1}{2}\int E_n g_\theta \cdot J'(Pg_{\theta,\cdot})\, dP + \int J(Pg_{\theta,\cdot})\, dE_n \\ &= L_{n,1}(\theta) + L_{n,2}(\theta) \quad \text{(say)}. \end{aligned}$$

This representation is according to Stute [38], p.229. Fernholz [12] treats a slightly different functional for which she proves compact differentiability at (θ_0, F).

We are now in a position to check assumptions A5 - A7 of Theorem 3.1.

Since $E_n \Longrightarrow E$ in $l^\infty(\mathcal{G})$, we have by the continuous mapping theorem $\|E_n\|_{\mathcal{G}} \Longrightarrow \|E\|_{\mathcal{G}}$ as well and therefore we have by Slutsky's lemma that $\|(P_n - P)\|_{\mathcal{G}} \xrightarrow{P} 0$. If the derivative J' is bounded, we obtain after a dominated convergence argument that $\|n^{-1/2} L_{n,1}\|_\Theta \xrightarrow{P} 0$.

Consider the following transformation

$$h_\theta(x) = J\left(\int g_{\theta,x}(y)\, dP(y)\right) = J([F(x) + 1 - F(2\theta - x)]/2)$$

and define $\mathcal{H} = \{h_\theta : \theta \in \Theta\}$. Notice that h_θ is bounded and is an element of D_0. In the literature (cf. Birman & Solomjak [6] and Example 2.3) the upper bound

$$H_2(\delta, P_n, \mathcal{H}) = \mathcal{O}(1/\delta) \tag{3.10}$$

is available for the class of such functions. As a result of (3.10) and the fact that $\mathcal{H}$ is uniformly bounded, $\mathcal{H}$ satisfies the Glivenko-Cantelli property so $\|n^{-1/2}L_{n,2}\|_\Theta \xrightarrow{P} 0$, too. Hence $|d\gamma(\theta, P)(P_n - P)| \xrightarrow{P} 0$, uniformly in θ, i.e. condition A5 is fulfilled.

Let us check condition A6. See e.g. Stute [38], where a *chaining* argument is used. However, the compactness assumption on Θ used in that paper is unnecessary. One may prove this as follows. Observe that

$$\|J(Pg_{\theta,.}) - J(Pg_{\theta_0,.})\|_2 = \mathcal{O}(|\theta - \theta_0|),$$

provided F is differentiable. As a result, the map $\theta \mapsto h_\theta(x)$ is continuous. By the entropy bound (3.10),

$$\int_0^1 \sqrt{H_2(4\delta x, P_n, \mathcal{H}_n(2\delta))}\, dx = \mathcal{O}\left(1/\sqrt{\delta}\right).$$

Invoking Lemma 2.5, we can conclude that $L_{n,2}(\theta)$ is stochastically equicontinuous at θ_0.

Stochastic equicontinuity of $\tilde{E}_n(2\theta - x) = E_n I_{(-\infty, 2\theta - x]}$ and continuity of F and J' guarantee that

$$L_{n,1}(\theta) = \frac{1}{2}\int_{\mathbb{R}} [\tilde{E}_n(x) - \tilde{E}_n(2\theta - x)] J'\left(\frac{F(x) + 1 - F(2\theta - x)}{2}\right) dF(x)$$

is stochastically equicontinuous as well. Assumption A6 now follows from (3.9).

As E_n converges weakly to a Gaussian process and $d\gamma(\theta_0; P)$ is a bounded, linear transformation, condition A7 follows immediately from the classical CLT. Therefore

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -(1/M'(\theta_0)) d\tilde{\gamma}(\theta_0, F)(E_n) + o_P(1)$$

has in the limit a normal distribution with zero mean.

We still have to prove (3.9).

Lemma 3.1 *Let $J : [-2, 2] \to \mathbb{R}$ be an increasing map with a continuous, bounded derivative J'. Then $\tilde{\gamma}$ as given in (3.8) is Hadamard*

differentiable at F, uniformly in θ. The derivative is given by

$$d\tilde{\gamma}(\theta, F) \cdot h = \tag{3.11}$$
$$\frac{1}{2}\int_{\mathbb{R}} [h(x) - h(2\theta - x)] J'\left(\frac{F(x) + 1 - F(2\theta - x)}{2}\right) dF(x) +$$
$$+ \int_{\mathbb{R}} J\left(\frac{F(x) + 1 - F(2\theta - x)}{2}\right) dh(x),$$

where the integral with respect to h is defined via integration by parts if h is not of bounded variation. Recall that the domain was given by $D_0 = \{F \in D : \int |dF| \leq 3\}$ and the tangent space is $\{h \in D : h$ is uniformly continuous and bounded$\}$.

Proof. Let $h_n \xrightarrow{\|\cdot\|} h$ and $t_n \to 0$ as $n \to \infty$. Set

$$y_n(x, \theta) = \frac{1}{2}\left[(F + t_n h_n)(x) + 1 - (F + t_n h_n)(2\theta - x)\right], \text{ and}$$
$$z(x, \theta) = \frac{1}{2}\left[F(x) + 1 - F(2\theta - x)\right].$$

Since $h_n \xrightarrow{\|\cdot\|} h$ and $t_n \to 0$, we have $y_n \to z$, uniformly in both x and θ. We define $F_n = F + t_n h_n$ and let us consider only perturbations $F_n \in D_0$. In particular we have $\|z\| \leq 1$ and $\|y_n\| \leq 2$. Check that

$$\frac{\tilde{\gamma}(\theta, F + t_n h_n) - \tilde{\gamma}(\theta, F)}{t_n} - \frac{1}{2}\int [h_n - h_n(2\theta - \cdot)] J'\left(z(\cdot, \theta)\right) dF -$$
$$- \int J\left(z(x, \theta)\right) dh_n(x) = I + II,$$

where

$$I = \int \left[\frac{J(y_n(\cdot, \theta)) - J(z(\cdot, \theta))}{t_n} - \frac{h_n - h_n(2\theta - \cdot)}{2} J'(z(\cdot, \theta))\right] dF;$$
$$II = \int \left[J(y_n(x, \theta)) - J(z(x, \theta))\right] dh_n(x).$$

Since J is continuously differentiable, we have

$$J(y_n(x, \theta)) = J(z(x, \theta)) + \frac{t_n}{2}(h_n(x) - h_n(2\theta - x)) J'(\tilde{y}_n(x, \theta)),$$

with $\tilde{y}_n(x, \theta)$ between $y_n(x, \theta)$ and $z(x, \theta)$. Therefore $\tilde{y}_n(x, \theta) \to z(x, \theta)$, uniformly in x and θ. Moreover J' is uniformly continuous and bounded

on the compact interval $[-2,2]$ so that $J'(\tilde{y}_n(x,\theta)) \to J'(z(x,\theta))$ as $n \to \infty$, uniformly in x and θ.

We have

$$\begin{aligned} |I| &= \frac{1}{2}\left|\int (h_n(x) - h_n(2\theta - x)) \cdot (J'(\tilde{y}_n(x,\theta)) - J'(z(x,\theta))\ dF(x)\right| \\ &\leq 2\|h_n - h\|\|J'\| \int |\, dF| + \|h\| \cdot \|J'(\tilde{y}_n) - J'(z)\| \cdot \int |\, dF| \\ &\to 0. \end{aligned}$$

Next, we show that $|II| \to 0$ as $n \to \infty$. We have

$$\begin{aligned} |II| &= \left|\int [J(y_n(x,\theta)) - J(z(x,\theta)]\ dh_n(x)\right| \\ &= \frac{1}{2}\left|\int [h_n(x) - h_n(2\theta - x)]\, J'(\tilde{y}_n(x,\theta))\, d(F_n - F)(x)\right|. \end{aligned}$$

After two applications of the triangle inequality, we see it is enough to prove that

$$\left|\int h(x)J'(z(x,\theta))\, d(F_n - F)\right| \to 0.$$

Now, since $h \in D$, there exists for any $\varepsilon > 0$ a step function $\tilde{h}$ with a finite number of jumps (say L) and $\|\tilde{h} - h\| \leq \varepsilon$. Also $F \in D$, and we approximate this function by a step function $\tilde{F}$ with $L' < \infty$ jumps such that $\|\tilde{F} - F\| \leq \delta$. We abbreviate $[\tilde{F}(x) + 1 - \tilde{F}(2\theta - x)]/2$ by $\tilde{z}(x,\theta)$. Notice that $\tilde{J}' = J'(\tilde{z})$ is again a step function and for every $\theta \in \Theta$, we can use the same approximation function $\tilde{F}$. We have as a result of the triangle inequality

$$\begin{aligned} &\left|\int h(x)J'(z(x,\theta))\, d(F_n - F)(x)\right| \leq \qquad (3.12) \\ &\leq \left|\int [h(x) - \tilde{h}(x)]J'(\tilde{z}(x,\theta))\, d(F_n - F)(x)\right| + \\ &+ \left|\int h(x)[J'(z(x,\theta)) - J'(\tilde{z}(x,\theta))]\, d(F_n - F)(x)\right| + \\ &+ \left|\int \tilde{h}(x)J'(\tilde{z}(x,\theta))\, d(F_n - F)(x)\right|. \end{aligned}$$

Observe that

$$\begin{aligned} &\left|\int h(x)[J'(z(x,\theta)) - J'(\tilde{z}(x,\theta))]\, d(F_n - F)(x)\right| \\ &\leq \|h\| \cdot \|J'(z) - J'(\tilde{z})\| \cdot \int |\, d(F_n - F)| \end{aligned}$$

and

$$\left|\int [h-\tilde{h}]J'(\tilde{z}(\cdot,\theta))\,d(F_n-F)\right| \leq \|h-\tilde{h}\|\cdot\|J'\|\cdot\int|\,d(F_n-F)|.$$

The last term on the right of (3.12) tends to zero by using arguments similar to those used by Gill [13], p. 110,111, using partial integration

$$\left|\int \tilde{h}J'(\tilde{z}(\cdot,\theta))\,d(F_n-F)\right| \leq 2\|\tilde{h}\|\cdot\|\tilde{J'}\|\cdot\|F_n-F\|+$$
$$+\|F_n-F)\|\cdot\left\{\|\tilde{J'}\|\cdot 2L\|\tilde{h}\|+\|\tilde{h}\|\cdot 2L'\|\tilde{J'}\|\right\}.$$

It can be shown that the integral $\int h(2\theta-x)J'(\tilde{y}_n(x,\theta))\,d(F_n-F)$ tends to zero as well for $n\to\infty$, by repeating the same arguments. This completes our proof. □

3.3 Minimization problems

We assume that the functional $\gamma(\cdot,P)$ is now uniquely minimized by some θ_0, an interior point of our parameter space Θ. As before, we abbreviate $\gamma(\cdot,P)$ and $\gamma(\cdot,P_n)$ by $M(\cdot)$ and $M_n(\cdot)$ respectively.

We impose the following conditions on M:

(C1) $\theta_0=\arg\min_{\theta\in\Theta}M(\theta)$ lies in the interior of Θ;

(C2) M has a unique minimum at θ_0;

(C3) M is twice differentiable at θ_0 with a non-singular second derivative V at θ_0.

Whereas in Section 3.2 the key to the solution lay in the notion of stochastic equicontinuity, we shall need stochastic differentiability here.

Definition 3.3 (stochastic differentiability) *Let $\{Z_n(t):t\in T\}$ be a stochastic process, indexed by $T\subset\mathbb{R}^k$. A sequence Z_n is called* stochastically differentiable *at $t_0\in T$ with derivative W_n iff $\forall\eta>0$ and $\forall\varepsilon>0$ there exists a neighborhood V of t_0 for which*

$$\limsup_{n\to\infty}\mathbb{P}\left\{\sup_{t\in V}\left|\frac{Z_n(t)-Z_n(t_0)-(t-t_0)'W_n}{|t-t_0|}\right|>\eta\right\}\leq\varepsilon.$$

Equivalently, a sequence Z_n is called stochastically differentiable with derivative W_n at $t_0 \in T$ iff for any $\tau_n \xrightarrow{P} t_0$, we have $|Z_n(\tau_n) - Z_n(t_0) - (\tau_n - t_0)'W_n| = o_P(|\tau_n - t_0|)$.

We impose the following conditions on the stochastic part:

(C4) $M_n(\hat{\theta}_n) \leq \inf_\theta M_n(\theta) + o_P(1/n)$;

(C5) $\|M_n - M\| \to 0$ in probability;

(C6) $\alpha_n = \sqrt{n}(M_n - M)$ is stochastically differentiable with derivative W_n at θ_0, i.e.

$$\alpha_n(\theta) = \alpha_n(\theta_0) + (\theta - \theta_0)'W_n + o_P(|\theta - \theta_0|) \text{ near } \theta_0;$$

(C7) $W_n \xrightarrow{\mathcal{D}} W = W(\theta_0; P)$.

Theorem 3.2 *Under conditions C1,...,C7, we have* $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} -V^{-1}W(\theta_0)$.

Proof. Consistency can be proved in the following way. Using the minimization properties of θ_0 and $\hat{\theta}_n$, we have

$$M_n(\hat{\theta}_n) \leq M_n(\theta_0) + o_P(n^{-1}) \xrightarrow{P} M(\theta_0) \leq M(\hat{\theta}_n). \tag{3.13}$$

By condition C5, $|M_n(\hat{\theta}_n) - M(\hat{\theta}_n)| \xrightarrow{P} 0$. Invoke (3.13) and conditions C1 and C2 to see that $\hat{\theta}_n \xrightarrow{P} \theta_0$.

By the stochastic differentiability assumption C6, we have

$$\begin{aligned} &\gamma(\theta, P_n) - \gamma(\theta, P) = n^{-\frac{1}{2}}\alpha_n(\theta) \\ &n^{-\frac{1}{2}}\alpha_n(\theta_0) + n^{-\frac{1}{2}}(\theta - \theta_0)'W_n + o_P(n^{-\frac{1}{2}}|\theta - \theta_0|), \end{aligned} \tag{3.14}$$

and from the deterministic differentiability C3 assumption we obtain

$$\gamma(\theta, P) - \gamma(\theta_0, P) = \frac{1}{2}(\theta - \theta_0)'V(\theta - \theta_0) + o(|\theta - \theta_0|^2). \tag{3.15}$$

Consequently we have

$$\begin{aligned} &\gamma(\theta, P_n) - \gamma(\theta_0, P_n) = n^{-1/2}(\theta - \theta_0)'W_n + \frac{1}{2}(\theta - \theta_0)'V(\theta - \theta_0) + \\ &+o_P(n^{-1/2}|\theta - \theta_0| + |\theta - \theta_0|^2). \end{aligned} \tag{3.16}$$

With a fairly straightforward generalization of the proof of Theorem 5 in Pollard [35], p. 141, the asymptotic distribution follows. For the sake of completeness we provide the full proof. First we reparametrize in such a way that θ_0 equals zero and V equals the identity matrix in the new parametrization. Continue the proof by noting that C4 and (3.16) imply

$$\begin{aligned} o_P(n^{-1}) &\geq \gamma(\hat{\theta}_n, P_n) - \gamma(0, P_n) \\ &= n^{-\frac{1}{2}}\hat{\theta}_n' W_n + \frac{1}{2}|\hat{\theta}_n|^2 + o_P(n^{-\frac{1}{2}}|\hat{\theta}_n| + |\hat{\theta}_n|^2). \quad (3.17) \end{aligned}$$

The random vector W_n is of order $O_P(1)$ since it converges weakly to the random vector W. Conclude that $\hat{\theta}_n = O_P(n^{-\frac{1}{2}})$. We can therefore write representation (3.16) as

$$\begin{aligned} \gamma(\hat{\theta}_n, P_n) - \gamma(0, P_n) &= n^{-\frac{1}{2}}\hat{\theta}_n' W_n + \frac{1}{2}|\hat{\theta}_n|^2 + o_P(n^{-1}) \quad (3.18) \\ &= \frac{1}{2}\left|\hat{\theta}_n + n^{-\frac{1}{2}}W_n\right|^2 - \frac{1}{2}n^{-1}|W_n|^2 + o_P(n^{-1}). \end{aligned}$$

The same simplification holds true for any sequence in Θ with values of order $O_P(n^{-\frac{1}{2}})$. In particular, by replacing $\hat{\theta}_n$ by $-n^{-1/2}W_n$, we find

$$\gamma(-n^{-\frac{1}{2}}W_n, P_n) - \gamma(0, P_n) = -\frac{1}{2}n^{-1}|W_n|^2 + o_P(n^{-1}), \quad (3.19)$$

since, with probability tending to one, $-n^{-\frac{1}{2}}W_n$ is a point of Θ as θ_0 is an interior point. Subtracting (3.19) from (3.18) and using C4, we obtain

$$\frac{1}{2}\left|\hat{\theta}_n + n^{-\frac{1}{2}}W_n\right|^2 = o_P(n^{-1}).$$

As a direct consequence of this, we have $n^{\frac{1}{2}}\hat{\theta}_n = -W_n + o_P(1)$. Finally, we transform back to the old parametrization and obtain

$$\begin{aligned} \sqrt{n}(\hat{\theta}_n - \theta_0) &= -V^{-1}W_n + o_P(1) \\ &\xrightarrow{\mathcal{D}} -V^{-1}W. \end{aligned}$$

The theorem is proved. □

Remark 3.5 This theorem generalizes the result about M-estimation obtained by Pollard (cf. [35]), where the function $M(\theta) = \int g(\theta, \cdot)\, dP$. Pollard assumes that g has the following expansion near θ_0

$$g(\theta, \cdot) = g(\theta_0, \cdot) + (\theta - \theta_0)\Delta(\cdot) + |\theta - \theta_0| R(\theta, \cdot).$$

Observe that stochastic differentiability of α_n is guaranteed under suitable entropy conditions on $\{R(\theta, \cdot) : \theta \in \Theta\}$ which imply stochastic equicontinuity of the process $\sqrt{n} \int R(\theta, \cdot)\, d(P_n - P)$ at θ_0. This is the same requirement as in [35].

Let us finish with a corollary concerning asymptotic normality of the normalized sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$. We assume that the consistency $\hat{\theta}_n \xrightarrow{P} \theta_0$ has already been established, so that we may drop the requirement $M_n \xrightarrow{P} M$ in $l^\infty(\Theta)$. Apart from this, the following corollary is an obvious consequence of (3.16).

Corollary 3.1 *Let $\hat{\theta}_n$ be a random sequence in Θ converging in probability to θ_0 at which $M(\cdot)$ has its minimum. Suppose the following conditions are satisfied.*

- *θ_0 is an interior point of Θ;*
- *$\gamma(\hat{\theta}_n, P_n) \leq \inf_\theta \gamma(\theta, P_n) + o_P(1/n)$;*
- *γ is Hadamard differentiable at P, in particular we have*

  $$\gamma(\theta, P_n) = \gamma(\theta, P) + n^{-\frac{1}{2}} d\gamma(\theta; P) \cdot E_n + n^{-\frac{1}{2}} R_n(\theta),$$

  where $R_n(\theta) = o_P(1)$ for every $\theta \in \Theta$;
- *$d\gamma(\theta; P) \cdot E_n$ is stochastically differentiable at θ_0;*
- *$R_n(\theta) - R_n(\theta_0) = o_P(|\theta - \theta_0|)$ near θ_0;*
- *$\gamma(\theta, P)$ is twice differentiable at θ_0 with non-singular second derivative V.*

Then the random sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal.

In the remainder of this chapter, we present some examples to illustrate the results. We shall mainly check the uniform convergence condition C5 and the stochastic differentiability assumption C6.

Example 3.3 (minimum distance estimation) Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a collection probability measures on the real line and let $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$ be the set of distribution functions associated with $\mathcal{P}$. F_n is the empirical distribution function. Consider an estimator $\hat{\theta}_n$, which minimizes the Cramer-von Mises distance $d(F_n, F_{\hat{\theta}_n})$ between F_n and $F_{\hat{\theta}_n}$, where

$$d^2(F, G) = \int (F - G)^2 \, dF.$$

Notice that in $d^2(F_n, F_{\hat{\theta}_n})$ the integration is with respect to F_n and not with respect to $F_{\hat{\theta}_n}$, which is perhaps more common. See the discussion in Stute [38] and the references given in that paper for the advantage of this method. We are interested in θ_0, the minimizer of $d^2(F, F_\theta)$, and assume that the true underlying F belongs to $\mathcal{F}$, i.e. $F = F_{\theta_0}$.

Set $\tilde{\gamma}(\theta, G) = \int (G - F_\theta)^2 \, dG$ and define the "derivative" $d\tilde{\gamma}(\theta, F)$ at F in the direction $F_n - F$ by

$$d\tilde{\gamma}(\theta, F)(F_n - F) = 2 \int (F - F_\theta) \cdot (F_n - F) \, dF + \int (F - F_\theta)^2 \, d(F_n - F).$$

Notice that $\tilde{\gamma}(\theta, F_n)$ can be written as

$$\begin{aligned} \tilde{\gamma}(\theta, F_n) &= \int (F_n - F_\theta)^2 \, dF_n \\ &= \int [(F_n - F) + (F - F_\theta)]^2 \, d[F + (F_n - F)] \\ &= \tilde{\gamma}(\theta, F) + d\tilde{\gamma}(\theta, F)(F_n - F) + R_n(\theta), \end{aligned}$$

where the remainder $R_n(\theta)$ is given by

$$R_n(\theta) = \int [F_n - F]^2 \, dF_n + 2 \int (F - F_\theta)(F_n - F) \, d(F_n - F).$$

As in Example 3.2, we abbreviate the ordinary empirical process by $\tilde{E}_n = \sqrt{n}(F_n - F)$. Since $\mathcal{F}$ is a uniformly bounded Glivenko-Cantelli class, $\|d\tilde{\gamma}(\cdot, F)(F_n - F)\| \xrightarrow{P} 0$ and $\|R_n(\cdot)\| \xrightarrow{P} 0$. Hence condition C5 holds true.

Let us check the stochastic differentiability condition C6. For this purpose, it will be convenient to assume that the distribution functions F_θ are uniformly differentiable,

$$\|F(\cdot,\theta) - F(\cdot,\theta_0) - (\theta-\theta_0)\Delta(\cdot)\| = o(|\theta-\theta_0|), \quad \text{near } \theta_0,$$

for some fixed function $\Delta \in D[-\infty,\infty]$. Then it follows straightforwardly that

$$d\tilde{\gamma}(\theta,F)\cdot\tilde{E}_n - d\tilde{\gamma}(\theta_0,F)\cdot\tilde{E}_n = -2(\theta-\theta_0)\int \Delta\tilde{E}_n\,dF + o_P(|\theta-\theta_0|)$$

and

$$\begin{aligned}
&|R_n(\theta) - R_n(\theta_0)| = \left|-2n^{-\frac{1}{2}}\int (F(\cdot,\theta)-F(\cdot,\theta_0))\tilde{E}_n\,d(F_n-F)\right| = \\
&= \left|-2n^{-\frac{1}{2}}(\theta-\theta_0)\int \Delta\tilde{E}_n\,d(F_n-F)\right| + o_P(n^{-\frac{1}{2}}|\theta-\theta_0|) \\
&= o_P(n^{-\frac{1}{2}}|\theta-\theta_0|).
\end{aligned}$$

Hence $\alpha_n(\theta) = \sqrt{n}[\tilde{\gamma}(\theta,F_n) - \tilde{\gamma}(\theta,F)]$ is stochastically differentiable with derivative

$$W_n = -2\int \Delta\cdot\tilde{E}_n\,dF,$$

and $V = 2\int\Delta^2\,dF$. Consequently, $\sqrt{n}(\hat{\theta}_n-\theta_0)$ is asymptotically normal.

Example 3.4 A more general and complicated case is obtained by introducing a continuously differentiable weight function J. We now consider

$$M(\theta) = \tilde{\gamma}((\theta,F) = \int J(F)\cdot(F-F_\theta)^2\,dF. \tag{3.20}$$

Define

$$\begin{aligned}
d\tilde{\gamma}(\theta;F)\cdot(F_n-F) \;=\; &\int (F_n-F)\cdot J'(F)(F-F_\theta)^2\,dF + \\
&+2\int (F_n-F)\cdot J(F)(F-F_\theta)\,dF + \\
&+\int J(F)(F-F_\theta)^2\,d(F_n-F).
\end{aligned} \tag{3.21}$$

The empirical function M_n is given by

$$\begin{aligned}\tilde{\gamma}(\theta, F_n) &= \int J(F_n)(F_n - F_\theta)^2\, dF_n \\ &= \int J(F_n)([F - F_\theta] + [F_n - F])^2\, d(F + [F_n - F]) \\ &= \int J(F_n)(F - F_\theta)^2\, dF + 2\int J(F_n)(F_n - F)(F - F_\theta)\, dF \\ &\quad + \int J(F_n)(F_n - F)^2\, dF_n + \int J(F_n)(F - F_\theta)^2\, d(F_n - F) \\ &\quad + 2\int J(F_n)(F_n - F)(F - F_\theta)\, d(F_n - F)\end{aligned}$$

and the remainder $R_n(\theta)$ by

$$\begin{aligned}&\int [J(F_n) - J(F) - J'(F)(F_n - F)]\,(F - F_\theta)^2\, dF + \\ &2\int [J(F_n) - J(F)](F_n - F)(F - F_\theta)\, dF + \int J(F_n)(F_n - F)^2\, dF_n + \\ &\int [J(F_n) - J(F)](F - F_\theta)^2\, d(F_n - F) + \\ &2\int J(F_n)(F_n - F)(F - F_\theta)\, d(F_n - F).\end{aligned}$$

We shall briefly verify conditions C5 and C6. Since the class $\mathcal{F}^2 = \{F^2 : F \in \mathcal{F}\}$ is a Glivenko-Cantelli class, and provided both $J(F)$ and $J'(F)$ are P-integrable functions, we have $d\tilde{\gamma}((\theta, F) \cdot (F_n - F) \xrightarrow{P} 0$ uniformly in θ. Also $\|R_n\| \xrightarrow{P} 0$, whence $\|M_n - M\| \xrightarrow{P} 0$. If $F(\cdot, \theta)$ is uniformly differentiable as before, we easily obtain $R_n(\theta) - R_n(\theta_0) = o_P(|\theta - \theta_0|)$. Again under the assumption that $F_{\theta_0} = F$, we find

$$\begin{aligned}W_n &= -2\int J(F) \cdot \Delta \cdot \tilde{E}_n\, dF; \\ V &= 2\int J(F) \cdot \Delta^2\, dF.\end{aligned}$$

Remark 3.6 (Non i.i.d. formulations) Theorems 3.1 and 3.2 can easily be extended to non-i.i.d. situations. This is mainly due to the fact that the uniform law of large numbers, stochastic equicontinuity and - differentiability allow a more general formulation. See for instance Pollard [37].

As an example we discuss parametric, non-linear regression. Let x_i be elements in some space S and let ε_i be independent random variables with zero means and finite variances σ_i^2, $i = 1, \ldots, n$. Let $\mathcal{G} = \{g_\theta : S \to \mathbb{R} : \theta \in \Theta\}$ be a class of functions, indexed by Θ, an open subset in $\mathbb{R}^k$. We observe $Y_i = g_{\theta_0}(x_i) + \varepsilon_i$, with $\theta_0 \in \Theta$. In addition, we assume that the following local linear approximation at θ_0 exists,

$$g_\theta(\cdot) = g_{\theta_0}(\cdot) + (\theta - \theta_0)'\Delta(\cdot) + |\theta - \theta_0| r_\theta(\cdot). \tag{3.22}$$

The parameter θ_0 can be consistently estimated by the least squares estimator $\hat{\theta}_n$, which minimizes

$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} (Y_i - g_\theta(x_i))^2 .$$

Let d_n be the empirical $L^2(P_n)$ pseudo metric on $\mathcal{G}$, based on $x_1, \ldots, x_n$, i.e.

$$d_n^2(g, h) = \frac{1}{n}\sum_{i=1}^{n} [g(x_i) - h(x_i)]^2 .$$

We use the short-hand notation $d_n(\theta, \tilde{\theta}) = d_n(g_\theta, g_{\tilde{\theta}})$. Suppose there exist $k_1, k_2 > 0$ such that for every $\theta, \tilde{\theta} \in \Theta$,

$$k_1(\theta - \tilde{\theta})^2 \leq d_n^2(\theta, \tilde{\theta}) \leq k_2(\theta - \tilde{\theta})^2.$$

The same condition is used in Wu [57]. It entails that (cf. (3.15))

$$\mathbb{E}[M_n(\theta) - M_n(\theta_0)] = d_n^2(\theta, \theta_0) = \mathcal{O}(\theta - \theta_0)^2.$$

Let us verify condition C5.

$$\begin{aligned} &(M_n - \mathbb{E}M_n)(\theta) = \\ &= \frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i^2 - \sigma_i^2) + 2\frac{1}{n}\sum_{i=1}^{n}(g_{\theta_0} - g_\theta)(x_i) \cdot \varepsilon_i. \end{aligned} \tag{3.23}$$

The first term on the right in (3.23) is asymptotically negligible under a uniform moment condition on the sequence ε_i. The second term on the right in (3.23) tends to zero, uniformly in θ, under the entropy condition

$$H(\delta, d_n, \mathcal{G}_n(\rho))/n \to 0 \quad \forall \delta > 0 \; \forall \rho > 0,$$

with $\mathcal{G}_n(\rho) = \{g \in \mathcal{G} : d_n(g, g_0) \leq \rho\}$. This will be proved in the next chapter, Corollary 4.3, under some additional assumptions on ε_i.

Condition C6 is verified if the stochastic process $n^{-\frac{1}{2}} \sum_{i=1}^n \varepsilon_i r_\theta(x_i)$ is stochastically equicontinuous at θ_0. Define $\mathcal{R} = \{r_\theta : \theta \in \Theta\}$ and $\mathcal{R}(\rho) = \{r_\theta \in \mathcal{R} : |\theta - \theta_0| < \rho\}$ and suppose $\mathcal{R}$ is uniformly bounded. Then we have by means of Theorem 2.4

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{|\theta-\theta_0|<\delta}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i r_\theta(x_i)\right| > \eta\right\} \leq \\
&\mathbb{P}\left\{\sup_{|\theta-\theta_0|<\delta}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i I\{|\varepsilon_i| > C\} r_\theta(x_i)\right| > \frac{\eta}{2}\right\} + \\
&+\mathbb{P}\left\{\sup_{|\theta-\theta_0|<\delta}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i I\{|\varepsilon_i| \leq C\} r_\theta(x_i)\right| > \frac{\eta}{2}\right\} \leq \\
&c_0 \sum_{i=1}^n \frac{\mathbb{E}\varepsilon_i^2 I\{|\varepsilon_i| > C\}}{n\eta^2} + 2\exp\left(-\left[\frac{\eta}{c_1 \delta \int_0^2 \sqrt{H_2(x\delta, P_n, \mathcal{R}(\delta))}\, dx}\right]^2\right).
\end{aligned}$$

where $c_0 = c_0(\mathcal{R}), c_1 = c_1(C)$ are some finite, positive constants. Provided ε_i are uniformly square integrable and under entropy conditions on $\mathcal{R}(\rho)$, property C6 is fulfilled.

For Condition C7 we have to verify the Lindeberg condition for $n^{-\frac{1}{2}} \sum_{i=1}^n \varepsilon_i \Delta(x_i)$.

Part III

Regression analysis

Chapter 4

Consistency

In the three remaining chapters we consider nonparametric regression. Suppose we have n independent observations satisfying the regression model

$$Y_i = g_0(x_i) + \varepsilon_i, \quad (i = 1, \ldots, n) \tag{4.1}$$

where the errors satisfy

$$\mathbb{E}\varepsilon_i = 0, \quad \mathbb{E}\varepsilon_i^2 = \sigma_i^2, \quad (i = 1, \ldots, n) \tag{4.2}$$

the design $\mathcal{X}_n = \{x_1, \ldots, x_n\}$ is a subset of $\mathbb{R}^{n \times k}$ and the regression function $g_0 : \mathbb{R}^k \to \mathbb{R}$ is unknown and to be estimated from the data.

If the regression function is known up to a finite dimensional parameter, the method of least squares is usually employed to estimate this unknown quantity. In a nonparametric setting of smooth functions, kernel estimators are a popular choice for estimation of g_0, owing to their relatively easy implementation and minimax properties.

We express our a priori knowledge of the regression function by

$$g_0 \in \mathcal{G}, \tag{4.3}$$

where $\mathcal{G}$ is a known class of functions.

The estimator of g_0 will be any random variable $\hat{g}$ that minimizes the sum of squares over $\mathcal{G}$ up to a constant $\eta_n \to 0$, i.e.

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{g}(x_i))^2 \leq \inf_{g \in \mathcal{G}} \frac{1}{n}\sum_{i=1}^{n}(Y_i - g(x_i))^2 + \eta_n. \tag{4.4}$$

With some abuse of terminology, we call $\hat{g}$ a least squares estimator (LSE). If $\eta_n = o(1/n)$, we can ignore η_n without affecting our results. Henceforth we assume that the infimum in (4.4) is attained and $\eta_n = 0$ for all n.

It should be stressed that in cases where only little information about g_0 is available, in other words where $\mathcal{G}$ is too large, $\hat{g}$ may simply interpolate between the y_i, and we obtain inconsistent estimators. In such cases other estimation procedures should be considered. In the context of least squares, one could think about sieved and penalized least squares estimators.

Model (4.1) has been studied extensively. The least squares estimator has also been investigated for particular choices of the design, i.i.d. errors and Sobolev classes $\mathcal{G}$ (cf. Nemirovskii et al. [31] and [32]). Van de Geer was the first to put the problem in a more general perspective by imposing entropy conditions on the set $\mathcal{G}$ rather than requiring smoothness properties. Here, we follow the same approach. To avoid making assumptions concerning the design $\mathcal{X}_n$, we shall consider metrics based on the design and formulate asymptotic properties like consistency and rates of convergence in these metrics.

At this point, let us introduce some notation used in the remainder of the book. We define the empirical measure P_n by

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i},$$

where δ_x is the Dirac measure at point $x \in \mathbb{R}^k$. I.e. this measure puts mass $1/n$ at each element of $\mathcal{X}_n$. The $L^p(P_n)$ pseudo norm on $\mathbb{R}^k$ is denoted by $\|\cdot\|_{n,p}$, thus

$$\|f\|_{n,p} = \begin{cases} \left(\int |f(x)|^p \, dP_n(x)\right)^{1/p} = \left(\frac{1}{n}\sum_{i=1}^{n} |f(x_i)|^p\right)^{1/p} & \text{if } 1 \le p < \infty; \\ \max_{1 \le i \le n} |f(x_i)| & \text{if } p = \infty. \end{cases}$$

Let $d_{n,p}(f,g) = \|f - g\|_{n,p}$ be its induced $L^p(P_n)$ pseudo metric, with $1 \le p \le \infty$ for any function f and g. Of particular interest is the case $p = 2$ as we shall formulate our results in their final form in the $L^2(P_n)$ metric. This metric has been chosen for mathematical convenience. For

the sake of brevity, if not specified differently, $\|\cdot\|_n$ and $d_n(f,g)$ will always denote the $L^2(P_n)$ pseudo norm and distance respectively.

For any family of numbers $z(g)$ indexed by $\mathcal{G}$, we put $\|z(g)\|_{\mathcal{G}} = \sup_{g\in\mathcal{G}} |z(g)|$. The canonical envelope of $\mathcal{G}$ will be denoted by G, i.e.

$$G(x) = \sup_{g\in\mathcal{G}} |g(x)|, \ x \in \mathbb{R}^k.$$

The δ-covering numbers in $L^p(P_n)$ are denoted by $N_p(\delta, P_n, \mathcal{G}) = N(\delta, d_{n,p}, \mathcal{G})$ and the associated entropies are given by $H_p(\delta, P_n, \mathcal{G}) = \log[N_p(\delta, P_n, \mathcal{G})]$ for each $\delta > 0$.

In addition to deterministic design, we shall allow randomness in the selection of the design points too (cf. Section 4.2 and Section 5.4). In this case the observation points X_i are assumed to be i.i.d. random variables with common probability distribution P, all independent of the disturbances ε_i. Moreover, we shall require $\int g_0^2\, dP < \infty$, i.e. $g_0 \in L^2(P)$. Note that P_n is now a random measure, to wit the empirical distribution, associated with the theoretical distribution P. The randomness of P_n has consequences for the covering numbers; they will be random as well. The $L^2(P)$ pseudo norm and metric will be denoted by $\|\cdot\|_p$ and $d_p(\cdot,\cdot)$ respectively. In order to avoid measurability problems - which may arise for uncountable classes $\mathcal{G}$ - we make the blanket assumption that $\mathcal{G}$ is permissible in the sense of Pollard [35].

The asymptotic behavior of the least squares estimator can be investigated using techniques which are described in Chapter 2. To see this, note that

$$\sum_{i=1}^n (Y_i - \hat{g}(x_i))^2 \le \sum_{i=1}^n (Y_i - g_0(x_i))^2 = \sum_{i=1}^n \varepsilon_i^2, \tag{4.5}$$

or equivalently,

$$\sum_{i=1}^n (\hat{g}(x_i) - g_0(x_i))^2 \le 2\sum_{i=1}^n \varepsilon_i\, (\hat{g}(x_i) - g_0(x_i))\,, \tag{4.6}$$

under the assumption that the true regression function g_0 does indeed belong to the set $\mathcal{G}$. From the last inequality, it is immediately clear

that asymptotic behavior of the difference $\|\hat{g} - g_0\|_n$ is governed by the empirical process

$$\left\{ n^{-1/2} \sum_{i=1}^{n} \varepsilon_i (g - g_0)(x_i) : g \in \mathcal{G} \right\}.$$

The probabilistic results concerning empirical processes which are given in Chapter 2, turn out to be very useful in the proofs.

In Chapters 4, 5 and 6 of this book, it is our aim to provide a systematic and unified study of the general regression problem (4.1) under the constraint (4.3). In particular, we are interested in the connection between asymptotic properties of the least squares estimator and entropy numbers (of subsets) of $\mathcal{G}$.

The remainder of this chapter is concerned with consistency issues. There is a considerable literature for cases where $\mathcal{G}$ is a subset of a Sobolev space. Van de Geer (cf. [39]) introduced the concept of (random) entropy numbers in $L^2(P_n)$ in the regression model (4.1) and gave sufficient conditions for establishing consistency. Our main concern will be to investigate the necessity of these entropy conditions. It will turn that in an appropriate setting, consistency is completely characterized by local entropy numbers. This clearly demonstrates the crucial role of entropy considerations in regression problems.

Most articles about least squares estimation discuss sufficient conditions for consistency and only few authors have dealt with their necessity. Wu (cf. [57]) considered the case of non-linear (parametric) regression, where $g(\cdot) = g(\theta, \cdot)$ is parametrized by $\theta \in \Theta \subset \mathbb{R}^N$. There it is shown that under the assumption of i.i.d. disturbances ε_i with an a.e. positive and absolutely continuous density and with finite Fisher information, the existence of a consistent estimator $\hat{\theta}(Y_1, \ldots, Y_n)$ of θ for all $\theta \in \Theta$ implies that $n d_n^2(\theta, \tilde{\theta}) \to \infty$ as $n \to \infty$ for all $\theta \neq \tilde{\theta}$, where $d_n(\theta, \tilde{\theta})$ denotes the $L^2(P_n)$ pseudo distance between $g(\cdot, \theta)$ and $g(\cdot, \tilde{\theta})$. Under assumptions on the model, this turns out to be a sufficient condition too.

Once consistency has been established, the next question is how fast this convergence turns out to be, i.e. the rate of convergence will be

examined. Under the hypothesis that the disturbances ε_i are almost Gaussian, and the entropy numbers of $\mathcal{G}$ intersected with shrinking balls in $L^2(P_n)$ with common center g_0 and radius proportional to $\delta_n \to 0$ behave like $n\delta_n^2$, Van de Geer proved that the rate of convergence will also be δ_n, i.e. $\|\hat{g} - g_0\|_n = \mathcal{O}_P(\delta_n)$. In many cases, this is the best rate one can achieve.

In Chapter 5, we give a detailed account of the rates of convergence. Upper and lower bounds will be derived and also the role of the error distributions will be discussed. Related work has been done by Birgé & Massart (cf. [5]) and Shen & Wong (cf. [55], [56]). The main difference is that their approach is based on entropy with bracketing.

Finally some asymptotic distribution theory concerning the least squares estimator $\hat{g}$ will be presented in chapter 6.

4.1 The envelope case

We start with introducing an i.i.d. regression model with stochastic design.

Model 1.

$$Y_i = g_0(X_i) + \sigma\varepsilon_i \quad (i = 1, \ldots, n). \tag{4.7}$$

- X_i are i.i.d. with probability distribution P on $\mathbb{R}^k$;
- ε_i are i.i.d. with probability distribution K on $\mathbb{R}$ and $\mathbb{E}\varepsilon_1 = 0$ and $\mathbb{E}\varepsilon_1^2 = 1$;
- $X_1, \ldots, X_n$, $\varepsilon_1, \ldots, \varepsilon_n$ are independent;
- $g_0 \in \mathcal{G} \subset L^2(P)$.

Note that $\sigma^2 \geq 0$ is the variance of the error $e = \sigma \cdot \varepsilon$.

Consistency is the weakest requirement for any reasonable estimator. In the case of least squares estimation, a natural way to measure the distance between the least squares estimator $\hat{g}$ and g_0 is by means of the $L^2(P_n)$ pseudo norm. We shall now define some concepts of consistency.

Definition 4.1 (Consistency)

- *A sequence of estimators* $\{\hat{g}_n\}$ *of* g_0 *is called* $L^2(P_n)$*-consistent if* $\|\hat{g}_n - g_0\|_{n,2} \to 0$ *in probability;*
- *A sequence of estimators* $\{\hat{g}_n\}$ *of* g_0 *is called strongly* $L^2(P_n)$*-consistent if* $\|\hat{g}_n - g_0\|_{n,2} \to 0$ *almost surely.*

$L^2(P)$-consistency and strong $L^2(P)$-consistency are defined in a similar way.

For finite $\mathcal{G}$, $L^2(P_n)$-consistency is easy to establish, and more generally we can show that if $\mathcal{G}$ is essentially not too large, $\hat{g}$ is a $L^2(P_n)$ consistent estimator of g_0. In Theorem 4.1 we shall make precise what is meant by "essentially not too large". Notice that g_0 minimizes $S(g) = \mathbb{E}\{Y - g(X)\}^2$ and that $\hat{g}$ minimizes $S_n(g) = n^{-1}\sum_{i=1}^n \{Y_i - g(X_i)\}^2$, the empirical counterpart of $S(g)$. By the strong law of large numbers, $S_n(g) \xrightarrow{a.s.} S(g)$, for any fixed $g \in L^2(P)$. If this convergence is uniform in $\mathcal{G}$ then $L^2(P_n)$-consistency is not hard to prove. We state a set of sufficient conditions which can be found in Van de Geer [39].

Proposition 4.1 *Consider regression model 1. Suppose the following conditions are satisfied:*

$$\int G^2 \, dP < \infty \quad \text{(envelope condition)}, \tag{4.8}$$

where $G = \sup_{g \in \mathcal{G}} |g|$ *(pointwise) is the canonical envelope of class* $\mathcal{G}$*, and*

$$\frac{1}{n} H_2(\delta, P_n, \mathcal{G}) \xrightarrow{P} 0 \text{ for all } \delta > 0 \quad \text{(entropy condition)}. \tag{4.9}$$

Then $\hat{g}$ *is both strongly* $L^2(P_n)$ *and strongly* $L^2(P)$ *consistent.*

The link with the theory of empirical processes will have become clear by now, since almost sure convergence of empirical processes uniformly over general classes $\mathcal{G}$ is one of the main topics in this field of

probability theory. Indeed the entropy and envelope conditions (4.8) and (4.9) ensure that $\mathcal{G}$ is a Glivenko-Cantelli class, i.e.

$$\left\| \int g \, d(P_n - P) \right\|_{\mathcal{G}} \xrightarrow{a.s.} 0. \tag{4.10}$$

A natural question is whether the converse of Proposition 4.1 holds true. The answer is negative because parametric linear regression is a counterexample: (4.8) is not satisfied, yet the least squares estimator is consistent. Example 4.1 below shows that even if $G \in L^2(P)$, condition (4.9) is not necessary for consistency.

Example 4.1 Let ε_1 be a Rademacher variable, i.e. $\mathbb{P}\{\varepsilon_1 = -1\} = \mathbb{P}\{\varepsilon_1 = 1\} = 1/2$. Indeed this variable fulfills the required properties $\mathbb{E}\varepsilon_1 = 0$ and $\mathbb{E}\varepsilon_1^2 = 1$. Suppose $g_0 \equiv 0$ and that $\mathcal{G} = \{I_A : A \in \mathcal{B}\}$, where $\mathcal{B}$ is the collection of all Borel sets. Then (4.8) is met with $G \equiv 1$, but (4.9) fails.

For $\sigma = 1$ the consistency fails because straightforward computation yields

$$d_n^2(\hat{g}, g_0) = n^{-1} \sum_{i=1}^{n} I\{\varepsilon_i = 1\} \xrightarrow{a.s.} 1/2.$$

Notice however that for $0 < \sigma < 1/2$, we have $\hat{g} \equiv g_0$, so $\hat{g}$ is certainly consistent. Hence condition (4.9) is sufficient but not necessary for $L^2(P_n)$ consistent least squares estimators.

We learn from Example 4.1 that there actually exist situations where consistency holds true only for some special values of σ. This phenomenon is undesirable. If $G \in L^2(P)$ and if we require consistency for all σ, necessary and sufficient entropy conditions can be established relatively easily. However, as already pointed out by Van de Geer [39], this envelope assumption is far too stringent in most cases.

Theorem 4.1 *Consider regression model 1. Assume $G \in L^2(P)$. The following two statements are equivalent:*

$$d_n(\hat{g}, g_0) \xrightarrow{a.s.} 0 \text{ for all } \sigma \in \mathbb{R} \tag{4.11}$$

$$n^{-1} H_2(\delta, P_n, \mathcal{G}) \xrightarrow{P} 0 \text{ for all } \delta > 0. \tag{4.12}$$

Before we prove this result, let us introduce some notation. Define for all functions $g \in \mathcal{G}$,

$$m_n(g) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i \left(g(X_i) - g_0(X_i) \right),$$
$$L_n(g;\sigma) = 2\sigma n^{-1/2} m_n(g) - d_n^2(g, g_0).$$

The least squares estimator $\hat{g}$ has the following property:

$$L_n(\hat{g};\sigma) = \sup_{g \in \mathcal{G}} L_n(g;\sigma) \tag{4.13}$$

because minimizing $S_n(g)$ is evidently the same as maximizing $L_n(g;\sigma)$ over $g \in \mathcal{G}$.

Proof of Theorem 4.1. The implication (4.12) $\Longrightarrow$ (4.11) has been proved in Van de Geer [39]. Therefore we only have to prove the necessity part (4.11) $\Longrightarrow$ (4.12). We first show that

$$\sup_{g \in \mathcal{G}} |n^{-1/2} m_n(g)| \xrightarrow{a.s.} 0. \tag{4.14}$$

We begin by noting that the joint distribution of $\{m_n(g) : g \in \mathcal{G}, n \in \mathbb{N}\}$ is independent of σ. Also, regardless of whether σ is positive or negative,

$$\sup_{g \in \mathcal{G}} \left(\sigma n^{-1/2} m_n(g) \right) \geq \sigma n^{-1/2} m_n(g_0) = 0. \tag{4.15}$$

As $d_n(\hat{g}, g_0) \xrightarrow{a.s.} 0$, the Cauchy-Schwarz inequality implies that

$$|n^{-1/2} m_n(\hat{g})| \leq d_n(\hat{g}, g_0) \left(\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 \right)^{1/2} \xrightarrow{a.s.} 0.$$

Hence, by the definition of the LSE we have

$$\sup_{g \in \mathcal{G}} L_n(g;\sigma) \xrightarrow{a.s.} 0. \tag{4.16}$$

Clearly,

$$\sup_{g \in \mathcal{G}} \left(2\sigma n^{-1/2} m_n(g) \right) - \sup_{g \in \mathcal{G}} d_n^2(g, g_0) \leq \sup_{g \in \mathcal{G}} L_n(g;\sigma) \xrightarrow{a.s.} 0,$$

and

$$\sup_{g\in\mathcal{G}} d_n^2(g,g_0) \le \frac{4}{n}\sum_{i=1}^{n} G^2(X_i) \xrightarrow{a.s.} 4\mathbb{E}G^2(X_1).$$

Combining this with (4.15) we find that for every $\sigma \in \mathbb{R}$,

$$0 \le \sup_{g\in\mathcal{G}} \left(2\sigma n^{-1/2} m_n(g)\right) \le \frac{4}{n}\sum_{i=1}^{n} G^2(X_i) + \sup_{g\in\mathcal{G}} L_n(g;\sigma) \xrightarrow{a.s.} 4\mathbb{E}G^2(X_1),$$

and (4.14) follows, to wit $\mathcal{H} = \{\varepsilon(g-g_0) \mid g \in \mathcal{G}\}$ is a Glivenko-Cantelli class. This collection has a square integrable envelope $H = 2|\varepsilon|G$.

Let Q be the product measure $P \times K$ and let Q_n be the empirical measure based on (X_i, ε_i), $i = 1, \ldots, n$. We shall show that $\mathcal{H}$ is a Q-Glivenko-Cantelli class implies that $\mathcal{G}$ is a P-Glivenko-Cantelli class. Because $\mathbb{E}\varepsilon^2 = 1$, there exists a number $0 < \eta < \infty$ for which $\pi_0 := \mathbb{P}\{|\varepsilon| > \eta\} > 0$.

Define the measure $\tilde{P}_n$ as a discrete measure, which assigns mass $1/n$ to X_i if and only if $|\varepsilon_i| > \eta$. The random variable $N_n = \sum_{i=1}^{n} I\{|\varepsilon_i| > \eta\}$ counts the values for which this holds true. Since

$$\int g^2 \, d\tilde{P}_n \le \eta^{-2} \int \varepsilon^2 g^2 \, dQ_n,$$

it follows that $n^{-1}H_2(\delta/\eta, \tilde{P}_n, \mathcal{G}) \xrightarrow{P} 0$ for all $\delta > 0$. Observe that $\tilde{P}_n$ and $(N_n/n)P_{N_n}$ have the same conditional distribution, given $\varepsilon_1, \ldots \varepsilon_n$. Moreover, by the strong law of large numbers, we have $N_n/n \xrightarrow{a.s.} \pi_0$. Consequently, $n^{-1}H_2(\sqrt{n/N_n}\delta/\eta, P_{N_n}, \mathcal{G}) \xrightarrow{P} 0$ for all $\delta > 0$, and as a result $n^{-1}H_2(\delta, P_n, \mathcal{G}) \xrightarrow{P} 0$ for every $\delta > 0$ as well. This proves the theorem. □

Remark 4.1 In case the ε_i form an orthogaussian sequence, the last part of the proof of Theorem 4.1 can be simplified by means of Theorem 2.1. Note that $m_n(g)$ now is a centered Gaussian process. As a consequence of Sudakov's lower bound (Theorem 2.1), we have

$$\mathbb{E}\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \left(g(X_i) - g_0(X_i)\right) \ge C_S \mathbb{E}\frac{\delta\sqrt{H_2(\delta, P_n, \mathcal{G})}}{\sqrt{n}} \text{ for all } \delta > 0,$$

where the expectation on the right is taken with respect to P^n, the n-fold product measure of P. Hence by Chebyshev's inequality, we find for any $\alpha > 0, \delta > 0$,

$$\mathbb{P}\left\{\sqrt{H_2(\delta, P_n, \mathcal{G})} > \alpha\sqrt{n}\right\} \leq (C_S\delta\alpha)^{-1}\mathbb{E}\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\left(g(X_i) - g_0(X_i)\right).$$

Application of the Cauchy-Schwarz inequality yields

$$\begin{aligned}\sup_n \mathbb{E}\left(\|n^{-1/2}m_n(g)\|_{\mathcal{G}}\right)^2 \leq \quad & \sup_n \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\right)\cdot\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(2G(X_i))^2\right) \\ & = 4\mathbb{E}G^2(X_1) < \infty.\end{aligned}$$

By elementary arguments (see e.g. Billingsley [3], p.348), $\|n^{-1/2}m_n(g)\|_{\mathcal{G}}$ is uniformly integrable. Hence the almost sure convergence (4.14) implies convergence in mean of $\|n^{-1/2}m_n(g)\|_{\mathcal{G}}$. The global entropy condition follows from Chebyshev's inequality above.

Remark 4.2 Since $G \in L^2(P)$, the entropy assumption (4.12) is equivalent with $H_1(\delta, P_n, \mathcal{G}) = o_P(n)$ for all $\delta > 0$ (cf. Theorem 2.5). As a result, Theorem 4.1 can also be stated in terms of $L^1(P_n)$ entropy numbers.

Remark 4.3 Example 4.1 shows that it is essential to require the consistency (4.11) for all $\sigma \in \mathbb{R}$.

It should be noted that negative values for σ are needed to conduct the proof of Theorem 4.1. Alternatively, we could assume symmetric errors ε_i, i.e. $\mathbb{P}\{\varepsilon_1 \in B\} = \mathbb{P}\{-\varepsilon_1 \in B\}$ for every Borel set B, and consider only positive σ.

It has become apparent that minimizing the sum of squares $S_n(\cdot)$ over a Glivenko-Cantelli class produces an $L^2(P_n)$-consistent estimator, whereas essentially larger classes will give inconsistency. There is one unpleasant detail: the assumption $G \in L^2(P)$ is very restrictive. It even rules out the familiar case of parametric linear regression. The following lemma reveals, however, that at least every subclass of $\mathcal{G}$ with a P-square integrable envelope should be a Glivenko-Cantelli class, as is the case for linear regression.

Lemma 4.1 *Consider model 1. Suppose* $d_n(\hat{g}, g_0) \xrightarrow{a.s.} 0$ *for all* $\sigma \in \mathbb{R}$. *Then for every subclass* $\mathcal{G}^* \subset \mathcal{G}$ *with envelope* $G^* \in L^2(P)$ *and* $g_0 \in \mathcal{G}^*$, *we have that* $\mathcal{G}^*$ *is a Glivenko-Cantelli class.*

Proof of Lemma 4.1. From the almost sure convergence (4.16), we have

$$\sup_{g \in \mathcal{G}^*} L_n(g) \xrightarrow{a.s.} 0$$

since $\mathcal{G}^* \subset \mathcal{G}$. Repeat the same arguments as in the proof of Theorem 4.1 with $\mathcal{G}$ and G replaced by $\mathcal{G}^*$ and G^* respectively. □

Remark 4.4 The uniform convergence $\| \int g \, d(P_n - P) \|_{\mathcal{G}} \xrightarrow{a.s.} 0$ is certainly not necessary for obtaining consistent least squares estimators. We only mention that by a result due to Van de Geer the conditions in Proposition 4.1 can be relaxed considerably. By introducing scaled versions $f = f(g) = g/(1 + \|g\|_2)$ of $g \in \mathcal{G}$, it is possible to circumvent the envelope restriction and strong $L^2(P)$ consistency is established. See Van de Geer [39], Theorem 1.2, p.590.

4.2 Main result

We would like to extend Theorem 4.1 and Lemma 4.1 by dropping the envelope assumption. Since the restriction that G is in $L^2(P)$ is a necessary condition for characterizing the Glivenko-Cantelli property of a class $\mathcal{G}$, we lose a powerful tool when using the empirical process approach. Nevertheless, it appears that such conditions are indeed unnecessary technical restrictions, although the standard results of the theory of empirical processes are no longer applicable. Moreover, the entropy conditions can also be weakened.

In contrast with the previous section, we consider the case of fixed design, in other words we assume that P_n is a deterministic measure. We emphasize this by using lower case characters $x_1, \ldots, x_n$ for the design. The stochastic counterpart where $X_1, X_2, \ldots$ are i.i.d. follows directly because no restrictions on the design are imposed apart from the entropy

assumptions.

Model 2.

$$Y_i = g_0(x_i) + \sigma\varepsilon_i \quad (i = 1, \ldots, n). \tag{4.17}$$

- $x_1, x_2, \ldots$ is a sequence in $\mathbb{R}^k$;
- ε_i are i.i.d. with probability distribution K on $\mathbb{R}$, $\mathbb{E}\varepsilon_1 = 0$ and $\mathbb{E}\varepsilon_1^2 = 1$;
- $g_0 \in \mathcal{G}$.

The first result of this section shows that under certain entropy conditions consistency of the least squares estimator follows. Instead of considering the entropy of the entire space $\mathcal{G}$, the entropy of the subset

$$\mathcal{G}_n(R) = \{g \in \mathcal{G} : d_n(g, g_0) \leq R\}$$

is what really counts. This is a consequence of the fact that $d_n(\hat{g}, g_0)$ is almost surely bounded for all n sufficiently large. Indeed from the minimizing property

$$S_n(\hat{g}) \leq S_n(g) \quad \text{for all } g \in \mathcal{G},$$

we have in particular $S_n(\hat{g}) \leq S_n(g_0)$ (cf. (4.5)). Rewriting this inequality gives $d_n^2(\hat{g}, g_0) \leq 2|\sigma n^{-1/2} m_n(\hat{g})|$ (cf. (4.6)). Application of the Cauchy-Schwarz inequality yields $d_n(\hat{g}, g_0) \leq 2|\sigma| \sqrt{n^{-1} \sum_{i=1}^n \varepsilon_i^2} \xrightarrow{a.s.} 2|\sigma|$.

Theorem 4.2 *Consider regression model 2. The entropy condition*

$$n^{-1} H_1(\delta, P_n, \mathcal{G}_n(R)) \to 0 \quad \forall \delta > 0, R > 0 \tag{4.18}$$

implies strong $L^2(P_n)$-consistency of the least squares estimator.

The necessity of (4.18) is captured in the following theorem. We need an additional assumption on the distribution of the disturbances $\varepsilon_1, \ldots, \varepsilon_n$.

Theorem 4.3 *Consider model 2. Assume that the error-distribution K contains no atoms. If we have*

$$d_n(\hat{g}, g_0) \xrightarrow{P} 0 \quad \forall \sigma \in \mathbb{R}, \tag{4.19}$$

then the local entropy numbers fulfill (4.18).

Combination of Theorem 4.2 and Theorem 4.3 obviously yields

Corollary 4.1 *Consider model 2. In addition, assume that the distribution of ε_1 contains no atoms. The following statements are equivalent:*

1. $d_n(\hat{g}, g_0) \xrightarrow{P} 0 \quad \forall \sigma \in \mathbb{R}$;
2. $d_n(\hat{g}, g_0) \xrightarrow{a.s.} 0 \quad \forall \sigma \in \mathbb{R}$;
3. $n^{-1} H_1(\delta, P_n, \mathcal{G}_n(R)) \to 0 \quad \forall \delta > 0, R > 0$.

Proof. The implication (3) $\to$ (2) is given Theorem 4.2, (2) $\to$ (1) is obvious and (1) $\to$ (3) follows from Theorem 4.3. □

The remainder of this section is devoted to the proofs of our results. We set out with a probabilistic result, concerning an exponential upper bound for the supremum of the empirical process $m_n(g)$ over the subspace $\mathcal{G}_n(R)$, $R > 0$.

Lemma 4.2 (exponential bound for bounded random variables) *Let $|\varepsilon_1|$ be almost surely bounded by $C > 0$. Then the local entropy condition (4.18) implies*

$$\mathbb{P}\left\{ \sup_{g \in \mathcal{G}_n(R)} |n^{-1/2} m_n(g)| > a \right\} \leq 2 \exp\left(-\frac{1}{4} \frac{na^2}{C^2 R^2} \right) \tag{4.20}$$

for $n \geq n_0(C, R, a)$.

Proof. From Hoeffding's inequality (Lemma 2.3), we have for each $g \in L^2(P_n)$

$$\mathbb{P}\left\{ |n^{-1/2} m_n(g)| \geq a \right\} \leq 2 \exp\left(-\frac{1}{2} \frac{na^2}{C^2 d_n^2(g, g_0)} \right). \tag{4.21}$$

Let $\{g_i\}_{i=1}^M$ be the minimal $a/(2C)$-covering net of $\mathcal{G}_n(R)$ with respect to the $L^1(P_n)$-distance, so $M = N_1((a/2C), P_n, \mathcal{G}_n(R))$ and for every $g \in \mathcal{G}_n(R)$ there exists a $g^* \in \{g_i\}$ such that $(1/n)\sum_{i=1}^n |g(x_i) - g^*(x_i)| \leq a/(2C)$. But then $n^{-\frac{1}{2}}|m_n(g) - m_n(g^*)| \leq (a/2)$ holds, since ε_i are bounded by C. By virtue of the triangle inequality, we have

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{g\in\mathcal{G}_n(R)} |n^{-1/2}m_n(g)| > a\right\} \\
&= \mathbb{P}\left\{\sup_{g\in\mathcal{G}_n(R)} |n^{-1/2}m_n(g) - n^{-1/2}m_n(g^*) + n^{-1/2}m_n(g^*)| > a\right\} \\
&\leq \mathbb{P}\left\{\max_{1\leq i\leq M} |n^{-1/2}m_n(g_i)| > \frac{a}{2}\right\} \\
&\leq 2\exp\left(H_1(\frac{a}{2C}, P_n, \mathcal{G}_n(R)) - \frac{1}{2}\frac{na^2}{C^2R^2}\right) \\
&\leq 2\exp\left(-\frac{1}{4}\frac{na^2}{C^2R^2}\right)
\end{aligned}$$

for $n \geq n(C, R, a)$. □

Proof of Theorem 4.2. We have to prove the consistency for all $\sigma \in \mathbb{R}$. Fix $\sigma \in \mathbb{R}$. From the inequality $L_n(\hat{g}; \sigma) \geq L_n(g_0; \sigma)$ and the Cauchy-Schwarz inequality,

$$d_n^2(\hat{g}, g_0) \leq 2|\sigma| \cdot |n^{-1/2}m_n(\hat{g})| \leq 2|\sigma| \cdot \left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i^2\right)^{1/2} \cdot d_n(\hat{g}, g_0). \quad (4.22)$$

This inequality and the almost sure convergence $n^{-1}\sum_{i=1}^n \varepsilon_i^2 \xrightarrow{a.s.} 1$ imply that $d_n(\hat{g}, g_0) \leq 4|\sigma|$ almost surely for all sufficiently large n, and that

$$d_n^2(\hat{g}, g_0) \leq \sup_{g\in\mathcal{G}_n(4|\sigma|)} 2|\sigma| \cdot |n^{-1/2}m_n(g)|.$$

Hence it is enough to show that $\sup_{g\in\mathcal{G}_n(R)} |n^{-1/2}m_n(g)| \xrightarrow{a.s.} 0$ for all $R > 0$.

Truncation device.

The error terms $\varepsilon_1, \ldots, \varepsilon_n$ are generally not bounded. Therefore we need a truncation device in order to use Lemma 4.2. In general, let C

be positive and define

$$(\varepsilon_i)_C = \varepsilon_i I\{-C \leq \varepsilon_i \leq C\} - \mathbb{E}\varepsilon_i I\{-C \leq \varepsilon_i \leq C\}, \quad i = 1, \ldots, n \tag{4.23}$$

Obviously $\mathbb{E}(\varepsilon_1)_C = 0$, and $\mathbb{E}(\varepsilon_1 - (\varepsilon_1)_C)^2$ can be made arbitrarily small by taking C sufficiently large.

On the set $B_n = \{n^{-1}\sum_{i=1}^n (\varepsilon_i - (\varepsilon_i)_C)^2 < (a/2R)^2\}$ we have

$$\sup_{g \in \mathcal{G}_n(R)} \left| \frac{1}{n} \sum_{i=1}^n (\varepsilon_i - (\varepsilon_i)_C)\,(g(x_i) - g_0(x_i)) \right| \leq \frac{a}{2}$$

by the Cauchy-Schwarz inequality. Notice that by Kolmogorov's strong law of large numbers

$$\frac{1}{n}\sum_{i=1}^n (\varepsilon_i - (\varepsilon_i)_C)^2 \xrightarrow{a.s.} \mathbb{E}\,(\varepsilon_1 - \mathbb{E}(\varepsilon_1)_C)^2\,,$$

and for C sufficiently large, we have

$$\mathbb{E}(\varepsilon_1 - (\varepsilon_1)_C)^2 < \frac{1}{2}\left(\frac{a}{2R}\right)^2.$$

Thus for fixed positive numbers a and R,

$$\mathbb{P}\{\limsup_{n\to\infty} B_n^c\} = 0.$$

Next, we derive after an application of the triangle inequality that

$$\begin{aligned} &\mathbb{P}\left\{\limsup_{n\to\infty} \sup_{g\in\mathcal{G}_n(R)} \left|n^{-1/2} m_n(g)\right| > a\right\} \qquad (4.24)\\ &\leq \mathbb{P}\left\{\limsup_{n\to\infty} \sup_{g\in\mathcal{G}_n(R)} \left|\frac{1}{n}\sum_{i=1}^n (\varepsilon_i)_C\,(g(x_i) - g_0(x_i))\right| > \frac{a}{2}\right\} + \\ &+\mathbb{P}\{\limsup_{n\to\infty} B_n^c\}. \end{aligned}$$

As a result of the exponential bound (4.20),

$$\sum_{n=1}^{\infty} \mathbb{P}\left\{\sup_{g\in\mathcal{G}_n(R)} \left|\frac{1}{n}\sum_{i=1}^n (\varepsilon_i)_C\,(g(x_i) - g_0(x_i))\right| > \frac{a}{2}\right\} < \infty.$$

Application of the Borel-Cantelli lemma completes the proof. □

Proof of Theorem 4.3. Fix $R > 0$. The Cauchy-Schwarz inequality yields

$$|L_n(g;\sigma)| \leq 2|\sigma| d_n(g,g_0) \cdot \left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i^2\right)^{1/2} + d_n^2(g,g_0).$$

Therefore, for all $\sigma \in \mathbb{R}$, we have $\sup_{g\in\mathcal{G}} L_n(g;\sigma) \xrightarrow{P} 0$, and since $\mathcal{G}_n(R) \subset \mathcal{G}$ also $\sup_{g\in\mathcal{G}_n(R)} L_n(g;\sigma) \xrightarrow{P} 0$. Next, notice that

$$\sup_{g\in\mathcal{G}_n(R)} \left(2\sigma n^{-1/2} m_n(g)\right) - R^2 \leq \sup_{g\in\mathcal{G}_n(R)} L_n(g;\sigma) \xrightarrow{P} 0.$$

Hence for every $\sigma \in \mathbb{R}$,

$$0 \leq 2\left(\sup_{g\in\mathcal{G}_n(R)} \sigma n^{-1/2} m_n(g)\right) \leq R^2 + L_n(\hat{g};\sigma) \xrightarrow{P} R^2.$$

It follows that $\sup_{\mathcal{G}_n(R)} |n^{-1/2} m_n(g)| \xrightarrow{P} 0$. By the Cauchy-Schwarz inequality,

$$\sup_n \mathbb{E}\left(\sup_{g\in\mathcal{G}_n(R)} |n^{-1/2} m_n(g)|\right)^2 \leq \sup_n \mathbb{E}\left(\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i^2\right)^{1/2} \cdot R\right)^2 = R^2.$$

This implies that $\sup_{g\in\mathcal{G}_n(R)} |n^{-1/2} m_n(g)|$ is uniformly integrable, and hence

$$\mathbb{E} \sup_{g\in\mathcal{G}_n(R)} |n^{-1/2} m_n(g)| \to 0. \tag{4.25}$$

Symmetrization device.

Let ε_i^* be independent copies of ε_i and let τ_i be a Rademacher sequence (cf. Section 2.5), independent of the sequences ε_i and ε_i^* $(i = 1,\ldots,n)$. Note that the probability distribution of the quantity $\sup_{g\in\mathcal{G}_n(R)} |n^{-1}\sum_{i=1}^n (\varepsilon_i - \varepsilon_i^*)(g(x_i) - g_0(x_i))|$ is the same as the one of $\sup_{g\in\mathcal{G}_n(R)} |n^{-1}\sum_{i=1}^n \tau_i(\varepsilon_i - \varepsilon_i^*)(g(x_i) - g_0(x_i))|$. Hence by (4.25) and the triangle inequality we obtain

$$\mathbb{E} \sup_{g\in\mathcal{G}_n(R)} \left|\frac{1}{n}\sum_{i=1}^n \tau_i\,(\varepsilon_i - \varepsilon_i^*)\,(g(x_i) - g_0(x_i))\right| \to 0, \tag{4.26}$$

and therefore, by Markov's inequality,

$$\mathbb{E}\left(\sup_{g\in\mathcal{G}_n(R)} \left| \frac{1}{n}\sum_{i=1}^{n} \tau_i \left(\varepsilon_i - \varepsilon_i^*\right)\left(g(x_i) - g_0(x_i)\right)\right| \,\middle|\, \varepsilon_i - \varepsilon_i^*, 1 \le i \le n \right)$$
$$\xrightarrow{P} 0. \tag{4.27}$$

Let $\tilde{Q}_n$ be the empirical probability measure based on $(x_i, \varepsilon_i - \varepsilon_i^*)$, i.e. it puts mass $1/n$ at each $(x_i, \varepsilon_i - \varepsilon_i^*)$. Set $f(x_i, \varepsilon_i - \varepsilon_i^*; g) = (\varepsilon_i - \varepsilon_i^*)\left(g(x_i) - g_0(x_i)\right)$ with $g \in \mathcal{G}_n(R)$, and $\mathcal{F}_n(R) = \{f(\cdot,\cdot; g) : g \in \mathcal{G}_n(R)\}$. By Corollary 4.14 in Ledoux & Talagrand [27], p.116, we find that (4.27) implies $n^{-1}H_2(\delta, \tilde{Q}_n, \mathcal{F}_n(R)) \xrightarrow{P} 0$ for all $\delta > 0$.

Let $\mathcal{G}^A = \{g_1, \ldots, g_D\}$ be a maximal set in $\mathcal{G}_n(R)$ (a priori possibly with infinite cardinality) such that the $L^1(P_n)$-distance between every pair in $\mathcal{G}^A$ is larger than 2δ, i.e.

$$\int |g - \tilde{g}|\, dP_n > 2\delta \quad \text{for each } g, \tilde{g} \in \mathcal{G}^A,\ g \neq \tilde{g}. \tag{4.28}$$

By the regularity condition on the error distribution, there exists an $\eta > 0$ such that $\mathbb{P}\{|\varepsilon - \varepsilon^*| \le \eta\} \le \delta^2/(4R^2)$. Then we have almost surely, for large n

$$\begin{aligned}
&\int |\varepsilon - \varepsilon^*||(g - \tilde{g})(x)|\, d\tilde{Q}_n(x, \varepsilon - \varepsilon^*) \ge \\
&\ge \eta \left[\int |g - \tilde{g}|\, dP_n - \int |(g - \tilde{g})(x)| I\{|\varepsilon - \varepsilon^*| \le \eta\}\, d\tilde{Q}_n(x, \varepsilon - \varepsilon^*) \right] \ge \\
&\ge \eta \left[\int |g - \tilde{g}|\, dP_n - 2R\left(\frac{1}{n}\sum_{i=1}^{n} I\{|\varepsilon_i - \varepsilon_i^*| \le \eta\} \right)^{1/2} \right] \ge \\
&\ge \eta \left[\int |g - \tilde{g}|\, dP_n - 2R\sqrt{2\mathbb{P}\{|\varepsilon - \varepsilon^*| \le \eta\}} \right] \ge \\
&\ge \eta\delta.
\end{aligned}$$

Consequently, by the relation (2.1) between packing and covering numbers and the maximality property of packing numbers, we obtain almost surely for n large

$$\begin{aligned}
&N_1(2\delta, P_n, \mathcal{G}_n(R)) \le D_1(2\delta, P_n, \mathcal{G}_n(R)) = |\mathcal{G}^A| \le \\
&D_1(\eta\delta, \tilde{Q}_n, \mathcal{F}_n(R)) \le N_2(\eta\delta/2, \tilde{Q}_n, \mathcal{F}_n(R)).
\end{aligned}$$

As a result, (4.18) holds true. This concludes the proof. □

Remark 4.5 The technical condition on the error distribution K (it should contain no atoms) can be replaced by assuming that ε_1 is symmetric around 0 in combination with $\mathbb{P}(\varepsilon_1 = 0) = 0$. This follows from the proof of Theorem 4.3 by noting that in this case ε_1 and $\tau\varepsilon_1$ have the same distribution, where τ is an independent Rademacher variable. As a result we can skip the symmetrization device and invoke the result of Ledoux & Talagrand directly. Moreover, it suffices to require the consistency (4.19) only for $\sigma > 0$ (cf. Remark 4.3).

It should be noted that Theorem 4.3 and Corollary 4.1 can be stated in terms of $L^2(P_n)$ entropy conditions as well. This observation parallels Remark 4.2.

Corollary 4.2 *The following statements are equivalent:*

$$H_1(\delta, P_n, \mathcal{G}_n(R)) = o(n) \quad \text{for all } \delta > 0, R > 0; \tag{4.29}$$

$$H_2(\delta, P_n, \mathcal{G}_n(R)) = o(n) \quad \text{for all } \delta > 0, R > 0. \tag{4.30}$$

Proof. The relation (4.30) $\Longrightarrow$ (4.29) follows from $d_{n,1}(f,g) \leq d_{n,2}(f,g)$.

As a result of Theorem 4.2, the $L^1(P_n)$ entropy condition (4.29) implies the consistency (4.19) of the least squares estimator in the regression problem. In case the error distribution in our regression problem is standard normal, we shall prove that the consistency (4.19) implies (4.30). For $K = \mathcal{N}(0,1)$, $m_n(\cdot)$ is a centered Gaussian process. This property makes it feasible to apply Sudakov's lower bound (Theorem 2.1), yielding

$$\mathbb{E} \sup_{g \in \mathcal{G}_n(R)} m_n(g) \geq C \sup_{\delta > 0} \delta \sqrt{H_2(\delta, P_n, \mathcal{G}_n(R))}, \tag{4.31}$$

for some numerical constant $C > 0$. The local entropy condition (4.30) in $L^2(P_n)$ now follows from (4.25). Thus we have proved that in the regression model with Gaussian errors, the entropy statements (4.29) and (4.30) are the same. Since these statements do not depend on ε_i, but solely on the metric structure of $\mathcal{G}_n(R)$, the result follows. □

4.3 Some extensions

We shall briefly discuss some possible extensions. First we note that in applications of nonparametric regression, the parameter space often depends on the number of observations n. Second, the design points may form a triangular array $x_{n1}, \ldots, x_{nn}$ rather than a sequence $x_1, x_2, \ldots$. Extension of our results in these directions is straightforward, due to the power of the methods borrowed from empirical process theory (cf. Van de Geer [40] and Pollard [37]). We omit further details.

4.3.1 The heteroscedastic case

Consider model 2 (deterministic design) where the errors are independent, but not necessarily identically distributed. In particular, we discuss the so-called heteroscedastic case.

Model 3.

$$Y_i = g_0(x_i) + \sigma\varepsilon_i \quad (i = 1, \ldots, n). \tag{4.32}$$

- $x_1, x_2, \ldots$ is a sequence in $\mathbb{R}^k$;
- ε_i are independent with probability distributions K_i on $\mathbb{R}$; $\mathbb{E}\varepsilon_i = 0$, $\mathbb{E}\varepsilon_i^2 = \sigma_i^2$, $(i = 1, \ldots, n)$, and there exists $m > 2$ such that

$$\sup_{i \geq 1} \mathbb{E}|\varepsilon_i|^m < \infty \tag{4.33}$$

- $g_0 \in \mathcal{G}$.

We wish to extend Corollary 4.1 to cover this model 3 also.

A closer look at the proof of Theorem 4.2 reveals that a crucial step is the almost sure convergence

$$\lim_{C \to \infty} \limsup_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} (\varepsilon_i - (\varepsilon_i)_C)^2 = 0 \quad \text{a.s.}, \tag{4.34}$$

where $(\varepsilon_i)_C$ is defined as in (4.23). This enables us to employ the truncation device as Lemma 4.2 holds true for any sequence of random variables, provided they are independent and uniformly bounded.

As far as the necessity of the entropy condition is concerned (cf. Theorem 4.3), only formula (4.29) needs modification. To be more precise, the averages

$$\frac{1}{n}\sum_{i=1}^{n} I\{|\varepsilon_i - \varepsilon_i^*| < \eta\}$$

need to be small for small $\eta > 0$ and for n sufficiently large. It is quite obvious that we need a fraction of the observations $\varepsilon_1, \ldots, \varepsilon_n$ larger than $\eta > 0$ to obtain knowledge about $\int |g(x)|\, dP_n(x)$ from $\int |g(x)\varepsilon|\, dQ_n(x, \varepsilon)$.

Corollary 4.3 *Consider regression model 3. Suppose that the sequence*

$$l_n(\eta) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\{|\varepsilon_i - \varepsilon_i^*| < \eta\} \tag{4.35}$$

is equicontinuous at $\eta = 0$. Then the following two statements are equivalent

$$d_n(\hat{g}, g_0) \xrightarrow{a.s.} 0 \quad \forall \sigma \in \mathbb{R}; \tag{4.36}$$

$$n^{-1} H_1(\delta, P_n, \mathcal{G}_n(R)) \to 0 \quad \forall \delta > 0, R > 0. \tag{4.37}$$

Proof. The proof will parallel that of the one given in full detail for the i.i.d. situation. For this reason we concentrate on the differences.
Sufficiency part. We now have almost surely $d_n(\hat{g}, g_0) \leq 4|\sigma| \max_{i \leq n} |\sigma_i|$ for n sufficiently large. Because the ε_i are uniformly square integrable in view of (4.33), there exists $C > 0$ such that

$$\sup_{i \geq 1} \mathbb{E}(\varepsilon_i - (\varepsilon_i)_C)^2 \leq \frac{1}{2}\left(\frac{a}{2R}\right)^2,$$

where $(\varepsilon_i)_C$ is given by formula (4.23). From e.g. Petrov [34], we see that (4.34) holds. Hence for the event $B_n = \{n^{-1}\sum_{i=1}^{n}(\varepsilon_i - (\varepsilon_i)_C)^2 \leq (a/2R)^2\}$, we have
$\mathbb{P}\{\limsup_{n\to\infty} B_n\} = 1$, and the truncation device can be performed successfully. Since Hoeffding's inequality (4.21) holds true for independent, not necessarily identically distributed random variables, the remainder of the proof parallels the proof of Theorem 4.2.
Necessity part. The reasoning for model 2 remains valid, apart from the

fact that $H_1(\delta, \tilde{Q}_n, \mathcal{F}_n(R)) = o(n)$ for all $R > 0$, $\delta > 0$ yields the same statement with $\tilde{Q}_n$ and $\mathcal{F}_n(R)$ replaced by P_n and $\mathcal{G}_n(R)$. But since we have

$$\left| \frac{1}{n} \sum_{i=1}^{n} I\{|\varepsilon_i - \varepsilon_i^*| < \eta\} - \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}\{|\varepsilon_i - \varepsilon_i^*| < \eta\} \right| \xrightarrow{a.s.} 0,$$

the result follows immediately from formula (4.29) and condition (4.35). □

4.3.2 Sieves

We now focus our attention on "sieved" least squares estimation which may be useful in situations where the class $\mathcal{G}$ is too large in the sense that the entropy condition (4.18) is not met. Minimizing the sum of squares

$$S_n(g) = n^{-1} \sum_{i=1}^{n} (Y_i - g(x_i))^2$$

over the entire class $\mathcal{G}$ may lead to an inconsistent estimate. This problem can be overcome by taking approximating spaces $\mathcal{G}^{(n)}$ which do satisfy the entropy condition (4.18). The least squares estimator obtained by this procedure will be written as $\tilde{g}$ to distinguish it from $\hat{g}$. Since g_0 is not necessarily an element of $\mathcal{G}^{(n)}$, we define $g_0^{(n)}$ as the projection of g_0 on $\mathcal{G}^{(n)}$; the approximating error will be denoted by $\alpha_n = \|g_0 - g_0^{(n)}\|_n$. Observe that by the triangle inequality we have $\|\tilde{g} - g_0\|_n \geq \|\tilde{g} - g_0^{(n)}\|_n - \alpha_n$. Due to the minimizing property of $\tilde{g}$, we find that, for each $\eta > 0$,

$$\mathbb{P}\left\{\|\tilde{g} - g_0^{(n)}\|_n > \eta\right\} \leq \mathbb{P}\left\{ \sup_{g \in \mathcal{G}^{(n)}, \|g - g_0^{(n)}\|_n \geq \eta} S_n(g_0^{(n)}) - S_n(g) \geq 0 \right\}$$

$$\leq \mathbb{P}\left\{ \sup_{g \in \mathcal{G}^{(n)}, \|g - g_0^{(n)}\|_n \geq \eta} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i (g - g_0^{(n)})(x_i) \geq \frac{1}{2}(\eta^2 - \alpha_n \eta) \right\}.$$

It is now easily seen that consistency can be established under a slight modification of the entropy condition (4.18), where we replace $\mathcal{G}$ by its approximating class $\mathcal{G}^{(n)}$, provided the sequence α_n converges to zero.

Observe the trade-off between the approximation error and the size of the class $\mathcal{G}^{(n)}$. For consistency issues, it is enough to assume that $\alpha_n = o(1)$. However, if one is interested in the rate of convergence, the matter becomes more delicate. One can prove that if $H_2(\delta_n, P_n, \mathcal{G}_n^{(n)}(R\delta_n)) \asymp n\delta_n^2$, the approximation error α_n should be of order $\mathcal{O}(\delta_n)$ if the desired rate is $\|\tilde{g} - g_0\|_n = \mathcal{O}_P(\delta_n)$. See Van de Geer [45] for details.

4.3.3 Uniform consistency

In this subsection we shall work within the regression framework with deterministic design, but - in contrast with the preceding sections - we employ a different notion of consistency. We shall state our results in terms of the more restrictive uniform $L^2(P_n)$ consistency. This is in the spirit of related work of Birgé [4] and Ibragimov & Has'minskii [22]. Let us formulate the definition of uniform consistency.

Definition 4.2 *Let (Θ, d) be a pseudo-metric space. A sequence of estimators $\hat{\theta}_n$ is called uniformly d-consistent if and only if for each $\varepsilon > 0$*

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta} P_\theta \left\{ d(\hat{\theta}_n, \theta) > \varepsilon \right\} = 0.$$

A sufficient entropy condition for uniform $L^2(P_n)$ consistency of the least squares estimator can be established relatively easily. For this matter we define

$$\mathcal{G}_n(g; R) = \{h \in \mathcal{G} : d_n(g, h) \leq R\}, \ g \in \mathcal{G}, \ R > 0.$$

Theorem 4.4 *Consider model 3 with $\sigma = 1$. Uniform $L^2(P_n)$ consistency of the least squares estimator is implied by the uniform entropy condition*

$$\sup_{g\in\mathcal{G}} H_1(\delta, P_n, \mathcal{G}_n(g; R)) = o(n), \ \forall \delta > 0, \ \forall R > 0. \tag{4.38}$$

Proof. We follow the steps of the proofs of Theorem 4.2 and Corollary 4.3. First note that we still have $d_n(\hat{g}, g) \leq 2\left(n^{-1}\sum_{i=1}^n \varepsilon_i^2\right)^{1/2}$, where the right-hand side is independent of g.

Second, the maximal inequality (4.20) can easily be extended in the uniform sense

$$\sup_{g\in\mathcal{G}} \mathbb{P}_g \left\{ \sup_{h\in\mathcal{G}_n(g;R)} \left| n^{-1/2} m_n(h) \right| > a \right\} \leq 2\exp\left(-\frac{1}{4}\frac{na^2}{C^2R^2} \right)$$

by exploiting the uniformity property of the entropy condition (4.38).

Finally, we remark that the truncation device as used in the proof of Theorem 4.2 can be implemented without change since it is solely based on the sequence $\varepsilon_1, \ldots, \varepsilon_n$.

With these modifications we can argue as in the proofs of Theorem 4.2 and Corollary 4.3. □

For some special regression models, the necessity of the entropy condition (4.38) follows from Fano's lemma (see e.g. Ibragimov & Has'minskii [22], Birgé [4], Devroye [8]).

Lemma 4.3 (Fano's Lemma) *Let $P^{(1)}, \ldots, P^{(J)}$ be J probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$. Let X be a random variable with probability measure $P \in \{P^{(1)}, \ldots, P^{(J)}\}$. Then for any decision rule $\psi : \mathcal{X} \to \{1, 2, \ldots, J\}$ we have*

$$\max_{1\leq i\leq J} \mathbb{P}_i \{\psi(X) \neq i\} \geq 1 - \frac{J^{-2}\sum\sum_{i,j} K(P^{(i)}, P^{(j)}) + \log 2}{\log(J-1)}, \quad (4.39)$$

where $K(P^{(i)}, P^{(j)})$ is the Kullback-Leibler information, defined by

$$K\left(P^{(i)}, P^{(j)}\right) = \begin{cases} \int \log\left(dP^{(i)}/dP^{(j)}\right) dP^{(i)} & \text{if } P^{(i)} \ll P^{(j)}, \\ +\infty & \text{otherwise.} \end{cases}$$

Consider model 2 with the ε_i normally $\mathcal{N}(0,1)$ distributed. Consequently, the vector $(Y_1, \ldots, Y_n)$ is normally distributed. We denote the probability measure of $(Y_1, \ldots, Y_n)$ by P_g, emphasizing the dependence on the regression function g. In this case we can compute the Kullback-Leibler number of P_g and P_h explicitly. We find

$$K(P_g, P_h) =$$
$$= \int \log \frac{\prod_{k=1}^n \varphi\left(y_k - g(x_k)/\sigma\right)}{\prod_{k=1}^n \varphi\left(y_k - h(x_k)/\sigma\right)} \prod_{k=1}^n \frac{1}{\sigma}\varphi\left(\frac{y_k - g(x_k)}{\sigma}\right) d(y_1, \ldots, y_n)$$

$$= \sum_{k=1}^{n} \int \log \frac{\varphi\left(y_k - g(x_k)/\sigma\right)}{\varphi\left(y_k - h(x_k)/\sigma\right)} \frac{1}{\sigma} \varphi\left(\frac{y_k - g(x_k)}{\sigma}\right) dy_k$$

$$= \sum_{k=1}^{n} \left(g(x_k) - h(x_k)\right)^2 / (2\sigma^2) = n d_n^2(g,h)/(2\sigma^2),$$

where φ denotes the standard normal density.

For $\delta > 0, R > 0$, let $\mathcal{G}_n^{(2\delta)}(g;R) = \{g_1, \ldots, g_m\}$ be a set of points in $\mathcal{G}_n(g;R)$ with $d_n(g_i, g_j) > 2\delta$, $i \neq j$, where $m = D_2(2\delta, P_n, \mathcal{G}_n(g;R))$ is the 2δ-packing number of $\mathcal{G}_n(g;R)$ with respect to d_n. Obviously,

$$\begin{aligned} \sup_{g\in\mathcal{G}} \mathbb{P}_g \left\{\|g - \hat{g}\|_n > \delta\right\} &= \sup_{g\in\mathcal{G}} \sup_{f\in\mathcal{G}_n(g;R)} \mathbb{P}_f \left\{\|f - \hat{f}\|_n > \delta\right\} \qquad (4.40) \\ &\geq \sup_{g\in\mathcal{G}, m(g)\geq 3} \max_{f\in\mathcal{G}_n^{(2\delta)}(g;R)} \mathbb{P}_f \left\{\|f - \hat{f}\|_n > \delta\right\}. \end{aligned}$$

For any estimator $\hat{f}$ of the unknown function f we define

$$\psi(Y_1, \ldots, Y_n) = \begin{cases} f^* \in \mathcal{G}_n^{(2\delta)}(g;R) & \text{if } \|\hat{f} - f^*\|_n \leq \delta; \\ \textit{arbitrary} & \text{otherwise.} \end{cases}$$

It now follows easily from (4.40) and Lemma 4.3 that

$$\begin{aligned} &\sup_{g\in\mathcal{G}} \mathbb{P}_g \left\{\|\hat{g} - g\|_n > \delta\right\} \geq \qquad (4.41) \\ &\geq \sup_{g\in\mathcal{G}, m(g)\geq 3} \max_{f\in\mathcal{G}_n^{(2\delta)}(g;R)} \mathbb{P}_f \left\{\psi(Y_1, \ldots, Y_n) \neq f\right\} \\ &\geq 1 - \inf_{g\in\mathcal{G}, m(g)\geq 3} \frac{2n(R/\sigma)^2 + \log 2}{\log\left(D_2(2\delta, P_n, \mathcal{G}_n(g;R)) - 1\right)} \\ &\geq 1 - \inf_{g\in\mathcal{G}, m(g)\geq 3} \frac{4n(R/\sigma)^2 + 2\log 2}{\log\left(D_2(2\delta, P_n, \mathcal{G}_n(g;R))\right)}. \end{aligned}$$

The relation (2.1) between covering and packing numbers yields

$$\sup_{g\in\mathcal{G}} \mathbb{P}_g \left\{\|\hat{g} - g\|_n > \delta\right\} \geq 1 - \inf_{g\in\mathcal{G}, m(g)\geq 3} \frac{4n(R/\sigma)^2 + 2\log 2}{H_2(2\delta, P_n, \mathcal{G}_n(g;R))}. \qquad (4.42)$$

These considerations yield the following result.

Theorem 4.5 *Consider model 2. Let ε_i be $\mathcal{N}(0,1)$ distributed. If for every $\delta > 0$,*

$$\limsup_{n\to\infty} \sup_{g\in\mathcal{G}} \mathbb{P}_g \left\{\|\hat{g} - g\|_n > \delta\right\} = 0 \quad \forall \sigma > 0, \qquad (4.43)$$

then the uniform entropy condition (4.38) must be satisfied.

Proof. If no $g \in \mathcal{G}$ exists with $m(g) \geq 3$, then (4.38) is trivially satisfied in view of (2.1). Otherwise, if the LSE is consistent in the strong sense that the left-hand side of (4.42) tends to zero, the entropy numbers should satisfy

$$\limsup_{n\to\infty} \sup_{g\in\mathcal{G}, m(g)\geq 3} \frac{H_2(2\delta, P_n, \mathcal{G}_n(g;R))}{nR^2} \leq \frac{4}{\sigma^2}. \tag{4.44}$$

Since (4.43) holds for all $\sigma > 0$ and the left-hand side in (4.44) is independent of σ, the result follows. □

Chapter 5

Rates of convergence

In the previous chapter we have dealt with consistency issues of the least squares estimator in various regression models. The natural continuation of our study is to investigate the rate of convergence of a consistent least squares estimator. In other words, the purpose is to find metric conditions on $\mathcal{G}$ which guarantee the existence of sequences $n^{-1/2} \leq \delta_n \to 0$ such that

$$\limsup_{R\to\infty} \limsup_{n\to\infty} \mathbb{P}\left\{d_n(\hat{g}, g_0) \geq R\delta_n\right\} = 0.$$

This chapter will be entirely devoted to this question.

5.1 Upper bounds

In the parametric case, i.e. the case where $\mathcal{G}$ can be parametrized by some subset Θ of the Euclidean space $\mathbb{R}^k$, the normalized sequence $\sqrt{n}(\hat{\theta} - \theta_0)$ typically has a limiting Gaussian distribution, where $\hat{g}(\cdot) = g(\hat{\theta}, \cdot)$ and $g_0(\cdot) = g(\theta_0, \cdot)$. Thus the parametric least squares estimator converges with the optimal rate $n^{-1/2}$ under certain regularity conditions. It is well-known that if we try to estimate an infinite dimensional parameter, the problem usually becomes harder and as a result the convergence is in general slower; the rate is of order $n^{-\alpha}$, $\alpha < 1/2$. In the nonparametric regression context, for instance, the rate is of order $n^{-1/3}$, if the regressors are monotone functions bounded in supremum norm by a constant. Another familiar example is the class of smooth

functions where the rate depends on the number of derivatives. More specifically, let $\mathcal{G}$ be defined as the set of all functions $g : [0,1] \to [0,1]$ with $J_m^2(g) = \int (g^{(m)}(x))^2 \, dx \leq J < \infty$, where $g^{(m)}$ denotes the m-th derivative of g. The entropy for this class is considered in Birman & Solomjak [6] and their result implies that the rate of convergence is of order $n^{-m/(2m+1)}$, which is the case for kernel estimators as well.

Whereas the local entropy numbers $H_2(\delta, P_n, \mathcal{G}_n(R))$ determine the consistency of the least squares estimator, the speed of convergence follows from the behavior of the entropy integral

$$\int_0^2 \sqrt{H_2(xR\delta_n, P_n, \mathcal{G}_n(R\delta_n))} \, dx, \quad R > 0$$

and $n^{-1/2} \leq \delta_n \to 0$. As a rule of thumb, $H_2(xR\delta_n, P_n, \mathcal{G}_n(R\delta_n)) \asymp n\delta_n^2$ implies $\|\hat{g} - g_0\|_n = \mathcal{O}_P(\delta_n)$ (cf. Van de Geer [41]). It should be noted that we restrict ourselves to subgaussian errors at this point. A rough argument is as follows. From inequality (4.6), we see it is sufficient to consider the behavior of $n^{-1/2} \sum_{i=1}^n \varepsilon_i(\hat{g}(x_i) - g_0(x_i))$. If the disturbances ε_i are subgaussian, it can be shown that the process

$$\left\{ n^{-1/2} \sum_{i=1}^n \varepsilon_i (g - g_0)(x_i), \; g \in \mathcal{G}_n(R\delta_n) \right\}$$

is subgaussian with respect to the $L^2(P_n)$ pseudo norm. This means that the tails of

$$\left\| n^{-1/2} \sum_{i=1}^n \varepsilon_i (g - g_0)(x_i) \right\|_{\mathcal{G}_n(R\delta_n)}$$

decrease exponentially fast by virtue of Theorem 2.3.

First we restate a modification of Van de Geer's result. We consider a deterministic design and heteroscedastic errors.

Theorem 5.1 (Van de Geer [41]) *Consider model 3 with $\sigma = 1$. Suppose the ε_i are uniformly subgaussian, i.e. there exists a constant $\lambda > 0$ such that*

$$\sup_{i \geq 1} \mathbb{E} \exp\left(\lambda \varepsilon_i^2\right) < \infty. \tag{5.1}$$

Furthermore, assume that there exist a sequence δ_n *with* $n^{-1/2} \leq \delta_n \downarrow 0$ *as* $n \to \infty$, *and an integer* n_0 *such that*

$$\limsup_{R\to\infty} \sup_{n\geq n_0} \int_{R\delta_n/(8s)}^{2} \frac{\sqrt{H_2(xR\delta_n, P_n, \mathcal{G}_n(R\delta_n))}\, dx}{R\sqrt{n}\delta_n} = 0, \tag{5.2}$$

with $s = \sup_{i\geq 1} |\sigma_i|$. *Then we have*

$$\|\hat{g} - g_0\|_n = \mathcal{O}_P(\delta_n).$$

In fact, for some constants $\kappa_i > 0$ $(i = 1, 2)$ *not depending on* n *and* R,

$$\mathbb{P}\{d_n(\hat{g}, g_0) \geq R\delta_n\} \leq \kappa_1 \exp(-\kappa_2 R^2 n\delta_n^2), \tag{5.3}$$

for $n \geq n_0$ *and* $R \geq R_0$.

At this point, several remarks are in order.

Remark 5.1 First of all, note that the region of integration in (5.2) is $[R\delta_n/(8s), 2]$, rather than $[0, 2]$ as in Van de Geer [41]. In Van de Geer [45], the integration is over the former interval, but no proof is given. Although in many interesting cases, the integration may be extended to $[0, 2]$ without influencing the result, this observation is relevant as there exist situations where the entropy integral diverges, due typically to the behavior of the entropy numbers in the vicinity of $x = 0$ (see Birgé & Massart [5], Van de Geer [45]).

Second, it follows from Birgé & Massart [5], Wong & Shen ([55], [56]) that (5.1) can be slightly relaxed by assuming

$$\sup_{i\geq 1} \mathbb{E} \exp(\lambda|\varepsilon_i|) < \infty, \tag{5.4}$$

at the price of stronger conditions on $\mathcal{G}$, viz. local entropy with bracketing and uniform boundedness restrictions on $\mathcal{G}$. However, it is clear that (5.1) and (5.4) are too strong in many cases, in particular if $\mathcal{G}$ can be approximated by finite dimensional sieves arbitrarily well. See Van de Geer [45] for details. In this context, we note that virtually no moment assumptions are needed in least deviation regression (cf. Birgé & Massart [5], Van de Geer [41]).

Finally, we emphasize that the result (5.3) holds uniformly in $g_0 \in \mathcal{G}$, provided the entropy condition (5.2) is valid uniformly in $g_0 \in \mathcal{G}$. Thus (5.1) and

$$\limsup_{R\to\infty} \sup_{n\geq n_0} \sup_{g\in\mathcal{G}} \int_{R\delta_n/(8\sigma)}^{2} \frac{\sqrt{H_2(xR\delta_n, P_n, \mathcal{G}_n(g; R\delta_n))}\, dx}{R\sqrt{n}\delta_n} = 0$$

imply that

$$\limsup_{R,n\to\infty} \sup_{g\in\mathcal{G}} \mathbb{P}_g \{\|\hat{g} - g\|_n \geq R\delta_n\} = 0.$$

Under normality assumptions on the errors, $\varepsilon_i \stackrel{\mathcal{D}}{=} \mathcal{N}(0, \sigma^2)$, and Fano's Lemma (see Section 4.3.3, Lemma 4.3) we get some insight in the necessity of these entropy conditions. More specifically, for each $R \geq L \geq 1$, we have

$$\sup_{g\in\mathcal{G}} \mathbb{P}_g \{\|\hat{g} - g\|_n \geq L\delta_n\} = \sup_{g\in\mathcal{G}} \max_{f\in\mathcal{G}_n(g;R\delta_n)} \mathbb{P}_f \{\|\hat{g} - f\|_n \geq L\delta_n\} \geq$$

$$\geq \sup_{g\in\mathcal{G}, m(g)\geq 3} \left[1 - \frac{4nR^2\delta_n^2\sigma^{-2} + 2\log 2}{H_2(2L\delta_n, P_n, \mathcal{G}_n(g; R\delta_n))}\right].$$

5.1.1 Proof of Theorem 5.1

For reasons of completeness, we prove Theorem 5.1. Moreover, the structure of the proof resembles that of Theorems 5.2 and 5.4.

Proof of Theorem 5.1 Take δ_n as defined in Theorem 5.1. From (4.6), we have for every integer $l > 1$ that

$$\mathbb{P}\left\{\|\hat{g} - g_0\|_n \geq 2^l\delta_n\right\} \leq$$

$$\leq \mathbb{P}\left\{\sup_g \frac{2}{n}\sum_{k=1}^{n} \varepsilon_k(g(x_k) - g_0(x_k)) - \|g - g_0\|_n^2 \geq 0\right\},$$

where the supremum is taken over all functions in $\mathcal{G}$ with $\|g - g_0\|_n > 2^l\delta_n$, and hence

$$\mathbb{P}\left\{\|\hat{g} - g_0\|_n \geq 2^l\delta_n\right\} \leq \sum_{j=l}^{\infty} \mathbb{P}\left\{\sup_g 2m_n(g) \geq 2^{2j}\sqrt{n}\delta_n^2\right\}$$

$$:= \sum_{j=l}^{\infty} P_j, \tag{5.5}$$

where the supremum is taken over all functions in $\mathcal{G}$ with $2^j\delta_n \leq \|g - g_0\|_n \leq 2^{j+1}\delta_n$, and $m_n(g)$ is defined in Section 4.1. From Propositions 5.1 and 5.2 below, it follows that for some $\rho \geq 1$,

$$\mathbb{P}\{|m_n(f) - m_n(g)| > \lambda\} \leq 2\exp\left(-\frac{\lambda^2}{\rho\|f-g\|_n^2}\right).$$

Hence $m_n(\cdot)$ is a subgaussian process with respect to d_n. Moreover, by the Cauchy-Schwarz inequality, we have a.s.

$$|m_n(f) - m_n(g)| \leq \sqrt{n}\|f-g\|_n \left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i^2\right)^{\frac{1}{2}}.$$

We can now apply the maximal inequality stated in Corollary 2.1 on the probabilities P_j, which yields (5.3). □

Remark 5.2 The main ingredient of the proof of Theorem 5.1 is the maximal inequality for the empirical process $m_n(\cdot)$ (Corollary 2.1). It should be emphasized that the perhaps more common maximal inequality in Theorem 2.4 leads to slightly more stringent entropy conditions. In fact, if we appeal to Theorem 2.4, the upper bound on the probabilities P_j becomes

$$2\exp\left(-\left[\frac{2^j\delta_n\sqrt{n}}{C_\psi \int_0^2 \sqrt{H_2(x2^{j+1}\delta_n, P_n, \mathcal{G}_n(2^{j+1}\delta_n))}\,dx}\right]^2\right).$$

To ensure that the series $\sum_{j=l}^\infty P_j$ is still convergent, we need a slightly stronger assumption on the local entropy numbers than in Theorem 5.1, viz.

$$\limsup_{R\to\infty}\sup_{n\geq n_0} \frac{\int_0^2 \sqrt{H_2(xR\delta_n, P_n, \mathcal{G}_n(R\delta_n))}\,dx}{R^\alpha\sqrt{n}\delta_n} < \infty$$

for some $\alpha < 1$.

Finally we show that under assumption (5.1) the empirical process $\{m_n(g) : g \in \mathcal{G}\}$ is subgaussian with respect to the $L^2(P_n)$ pseudo norm. A similar result is given in Kuelbs [24] with a more complicated proof.

Proposition 5.1 *Let X be a random variable with*

$$\mathbb{E}X = 0 \text{ and } \mathbb{E}\exp(\lambda X^2) \leq A$$

for some constants $\lambda > 0$ and $A \geq 1$. Then

$$\mathbb{E}\exp(\beta X) \leq \exp\left(\frac{2A\beta^2}{\lambda}\right)$$

holds for every $\beta > 0$.

Proof. Since for all $t > 0$, $\mathbb{P}(|X| > t) \leq A\exp(-\lambda t^2)$ holds, we have for all integers $m \geq 2$,

$$\begin{aligned}\mathbb{E}|X|^m &= \int_0^\infty \mathbb{P}\{|X|^m > t\}\, dt \leq A\int_0^\infty \exp\left(-\lambda t^{2/m}\right)\, dt \\ &= A\lambda^{-m/2}\Gamma\left(\frac{m}{2}+1\right).\end{aligned}$$

Note that $\Gamma^2(\frac{m}{2}+1) \leq \Gamma(m+1)$ by Cauchy-Schwarz. The following inequalities are now self-evident.

$$\begin{aligned}\mathbb{E}\exp(\beta X) &= 1 + \sum_{m=2}^\infty \frac{1}{m!}\beta^m \mathbb{E}X^m \leq 1 + A\sum_{m=2}^\infty \lambda^{-m/2}\beta^m \frac{\Gamma\left(\frac{m}{2}+1\right)}{\Gamma(m+1)} \\ &\leq 1 + A\sum_{m=2}^\infty \lambda^{-m/2}\beta^m \frac{1}{\Gamma\left(\frac{m}{2}+1\right)} \\ &= 1 + A\sum_{m=1}^\infty \left(\frac{\beta^2}{\lambda}\right)^m \frac{1}{\Gamma(m+1)} + \\ &\quad + A\sum_{m=1}^\infty \left(\frac{\beta^2}{\lambda}\right)^{m+\frac{1}{2}} \frac{1}{\Gamma\left(m+\frac{3}{2}\right)} \\ &\leq 1 + A\sum_{m=1}^\infty \left(\frac{\beta^2}{\lambda}\right)^m \left(1 + \left(\frac{\beta^2}{\lambda}\right)^{\frac{1}{2}}\right) \frac{1}{\Gamma(m+1)}.\end{aligned}$$

Finally, invoke the inequality $1 + (1+\sqrt{x})(\exp(x) - 1) \leq \exp(2x)$ for $x > 0$, to obtain the result. □

Proposition 5.2 *Let $X_1, \ldots, X_n$ be independent random variables with*

$$\mathbb{E}X_i = 0 \text{ and } \mathbb{E}\exp\left(\lambda X_i^2\right) \leq A,\ i = 1, \ldots, n$$

for some constants $\lambda > 0$ *and* $A \geq 1$. *Let* $a_1, \ldots, a_n \in \mathbb{R}$, *and write* $\|a\| = \{\sum_{i=1}^n a_i^2\}^{\frac{1}{2}}$. *Then*

$$\mathbb{P}\left\{\left|\sum_{i=1}^n a_i X_i\right| \geq t\right\} \leq 2\exp\left(-\frac{\lambda t^2}{8A\|a\|^2}\right).$$

Proof. Using the independence of the X_i and the previous proposition, one obtains

$$\mathbb{E}\exp\left(\beta\sum_{i=1}^n a_i X_i\right) = \prod_{i=1}^n \mathbb{E}\exp\left(\beta a_i X_i\right) \leq \exp\left(2\frac{A\beta^2}{\lambda}\|a\|^2\right).$$

Hence

$$\mathbb{P}\left\{\sum_{i=1}^n a_i X_i > t\right\} \leq \exp\left(2\frac{A\beta^2}{\lambda}\|a\|^2 - \beta t\right).$$

Choose $\beta = \lambda t/4A\|a\|^2$. □

5.2 Non subgaussian disturbances

The regression model with deterministic design and heteroscedastic errors (model 3) will be considered. In Section 5.1, the restriction (5.1) was imposed on the disturbances. We shall establish a trade-off between the information of the errors, given in terms of the number of moments, and the size of the class of regressors, given by local entropy conditions.

The subgaussian property of the process

$$\left\{n^{-1/2}\sum_{i=1}^n \varepsilon_i(g - g_0)(x_i),\ g \in \mathcal{G}_n(R\delta_n)\right\}$$

and therefore the exponential decrease of its tails in sup-norm, is no longer guaranteed if the errors ε_i have only finite moments $\mathbb{E}|\varepsilon_i|^k$ for $k \leq m$. Under more stringent entropy conditions than (5.2), we still obtain a rate of convergence, but generally this rate is inferior to the rate obtained under restriction (5.1).

Although the unconditional process $m_n(\cdot)$ may fail to be subgaussian, we can show that the symmetrized version is subgaussian conditionally on $\varepsilon_1, \ldots, \varepsilon_n$. This observation and a maximal inequality yield the following result.

Theorem 5.2 *Consider model 3 with $\sigma = 1$. Let $m > 2$ such that*

$$\sup_{i \geq 1} \mathbb{E}|\varepsilon_i|^m < \infty. \tag{5.6}$$

Suppose that there exist a sequence δ_n with $n^{-1/2} \leq \delta_n \downarrow 0$ as $n \to \infty$, and an integer n_0 such that

$$\limsup_{R \to \infty} \sup_{n \geq n_0} \int_0^2 \frac{x^{-1/p} \sqrt{H_2\left(xR\delta_n, P_n, \mathcal{G}_n(R\delta_n)\right)}\, dx}{\sqrt{n}(R\delta_n)^{1+\frac{1}{p}}} = 0 \tag{5.7}$$

for all $1 \leq p < m/2$. Moreover, let the class $\mathcal{G}$ be uniformly bounded. Then we have for all $\alpha > 0$,

$$\limsup_{R \to \infty} \limsup_{n \to \infty} \mathbb{P}\left\{ d_n(\hat{g}, g_0) > R\delta_n^{1-\alpha} \right\} = 0.$$

Proof. Let δ_n be defined as in Theorem 5.2 and set $v_n^2 = \max_{1 \leq i \leq n} \sigma_i^2$. For all $g \in \mathcal{G}$ with $d_n(g, g_0) \geq R\delta_n$, $R > 0$, we have

$$\operatorname{Var}\left(\frac{m_n(g)}{d_n^2(g, g_0)} \right) \leq \frac{v_n^2}{d_n^2(g, g_0)} \leq \left(\frac{v_n}{R\delta_n} \right)^2,$$

and hence by Chebyshev's inequality, we have

$$\mathbb{P}\left\{ |m_n(g)| > \sqrt{n}x \right\} \leq \frac{v_n^2}{(xR)^2 n\delta_n^2} \leq \frac{v_n^2}{(xR)^2} \leq \frac{1}{2}$$

for $xR > \sqrt{2}\sup_{i \geq 1} \sigma_i \geq \sqrt{2}v_n$. Let $\tilde{\varepsilon}_i$ be independent copies of ε_i and define

$$\tilde{m}_n(g) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\varepsilon}_i \left[g(x_i) - g_0(x_i) \right].$$

After an application of Lemma 2.2, we find that

$$\begin{aligned} &\mathbb{P}\left\{ \sup_{d_n(g,g_0) \geq R\delta_n} \frac{|m_n(g)|}{d_n^2(g, g_0)} \geq \frac{1}{2}\sqrt{n} \right\} \leq \\ &\leq 2\mathbb{P}\left\{ \sup_{d_n(g,g_0) \geq R\delta_n} \frac{|m_n(g) - \tilde{m}_n(g)|}{d_n^2(g, g_0)} \geq \frac{1}{4}\sqrt{n} \right\}. \end{aligned} \tag{5.8}$$

Let $\tau_1, \ldots, \tau_n$ be a Rademacher sequence, independent of $\varepsilon_1, \ldots, \varepsilon_n$. Observe that $\varepsilon_i - \tilde{\varepsilon}_i$ is equal in distribution to $\tau_i(\varepsilon_i - \tilde{\varepsilon}_i)$ by symmetry. Therefore we may bound the probability in (5.8) further by

$$\mathbb{P}\left\{ \sup_{d_n(g,g_0) \geq R\delta_n} \frac{|m_n(g)|}{d_n^2(g, g_0)} \geq \frac{1}{2}\sqrt{n} \right\} \leq 4\mathbb{P}\left\{ \sup_{d_n(g,g_0) \geq R\delta_n} \frac{|m_n^o(g)|}{d_n^2(g, g_0)} \geq \frac{1}{8}\sqrt{n} \right\},$$

where $m_n^o(g)$ is defined by

$$m_n^o(g) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tau_i \varepsilon_i \left[g(x_i) - g_0(x_i) \right].$$

Let Q_n be the discrete measure which puts mass $1/n$ at each pair (x_i, ε_i) and define $h(x_i, \varepsilon_i; g) = \varepsilon_i g(x_i)$ with $g \in \mathcal{G}$, and $\mathcal{H}_n(R\delta_n) = \{h(\cdot, \cdot; g) : g \in \mathcal{G}_n(R\delta_n)\}$, $R > 0$. An application of Lemma 2.3 yields

$$\begin{aligned} &\mathbb{P}\left\{ |m_n^o(f) - m_n^o(g)| > x \mid \varepsilon_1, \ldots, \varepsilon_n \right\} \leq \\ &\leq 2 \exp\left(- \frac{nx^2}{2 \sum_{i=1}^{n} [\varepsilon_i (f(x_i) - g(x_i))]^2} \right). \end{aligned}$$

Hence, conditionally on the vector $(\varepsilon_1, \ldots, \varepsilon_n)$, the symmetrized process $\{m_n^o(g) : g \in \mathcal{G}\}$ is subgaussian with respect to the $L^2(Q_n)$ pseudo norm. We can not apply the maximal inequality in Theorem 2.3 directly. However, for positive constants C, c, and $\tilde{c}$, we can derive from its proof that, conditionally on $\varepsilon_1, \ldots, \varepsilon_n$,

$$\begin{aligned} &\mathbb{P}\left\{ \sup_{g \in \mathcal{G}_n(R\delta_n)} |m_n^o(g)| > \lambda (R\delta_n)^2 \;\middle|\; \varepsilon_1, \ldots, \varepsilon_n \right\} \leq \\ &\leq C \exp\left(-c\lambda^2 n (R\delta_n)^2 \right), \end{aligned} \tag{5.9}$$

for

$$\lambda \sqrt{n} \geq \tilde{c} (R\delta_n)^{-2} \int_0^{\Delta_n} \sqrt{H_2(x, Q_n, \mathcal{H}_n(R\delta_n))}\, dx, \tag{5.10}$$

where Δ_n is the diameter of the set $\mathcal{H}_n(R\delta_n)$.

Next, we replace the entropy of the set $\mathcal{H}_n(R\delta_n)$ in (5.10) by an entropy of $\mathcal{G}_n(R\delta_n)$. We may assume without loss of generality that $\sup_{g \in \mathcal{G}} \|g\|_\infty \leq 1$ since $\mathcal{G}$ is uniformly bounded. Observe that by Hölder's inequality for $1 < p < m/2$

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 g^2(x_i) &\leq \left(\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^{2p} \right)^{1/p} \cdot \left(\frac{1}{n} \sum_i g^{\frac{2p}{p-1}}(x_i) \right)^{1-1/p} \\ &\leq \left(\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^{2p} \right)^{1/p} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} g^2(x_i) \right)^{1-1/p} \quad \text{(a.s.)} \end{aligned}$$

Put $S_n(p) = \left(n^{-1}\sum_{i=1}^n \mathbb{E}\varepsilon_i^{2p}\right)^{1/p}$, $1 < p < m/2$. Thus we have almost surely for all $1 < p < m/2$ and large n,

$$H_2(\delta, Q_n, \mathcal{H}_n(R\delta_n)) \leq H_2((\delta/2S_n(p))^{\frac{p}{p-1}}, P_n, \mathcal{G}_n(R\delta_n))), \qquad (5.11)$$

since $n^{-1}\sum_{i=1}^n [\varepsilon_i^{2p} - \mathbb{E}\varepsilon_i^{2p}]$ converges almost surely to zero for all $1 < p < m/2$ under assumption (5.6). Combination of (5.7), (5.9), (5.10) and (5.11) entails by the same arguments as in the proof of Theorem 5.1 the desired rate of convergence. □

Remark 5.3 Observe the interplay between moment requirements on ε_i and entropy conditions on $\mathcal{G}$. The more moments $k = 2p$ we require ε_i to possess, the higher the speed of convergence of the least squares estimator. It is interesting to see that for $k = \infty$, when every moment of ε_i is finite and we are almost in the subgaussian case, the entropy conditions (5.2) and (5.7) are almost the same. We illustrate this by two examples.

Example 5.1 (monotone functions) Consider the class $\mathcal{G}$ of monotone functions on the real line, uniformly bounded by some finite constant. If we compute the entropy of this class with respect to the $L^2(P_n)$ pseudo norm, it turns out that

$$H_2(\delta, P_n, \mathcal{G}) \asymp 1/\delta.$$

In case the disturbances are subgaussian, it is well-known (see Van de Geer [41]) that the rate of convergence is

$$\|\hat{g} - g_0\|_n = \mathcal{O}_P\left(n^{-1/3}\right).$$

This also follows from Theorem 5.1. However, if we only know that the disturbances are i.i.d. with $\mathbb{E}|\varepsilon_1|^{2p} < \infty$, then Theorem 5.2 yields that

$$\|\hat{g} - g_0\|_n = \mathcal{O}_P\left(n^{-\frac{p}{3p+2}}\right).$$

Indeed, we see that for $p \to \infty$, the rate behaves like $n^{-1/3}$.

Example 5.2 (smooth functions) Next, we consider classes of functions $\mathcal{G}$ which satisfy the entropy bound

$$H_2(\delta, P_n, \mathcal{G}) \asymp \delta^{-\frac{1}{k}}.$$

This assumption is fulfilled, for instance, by the class of all k-times differentiable functions $g : [0,1] \to [0,1]$ with $\int \left[g^{(k)}(x)\right]^2 dx$ bounded above by a finite constant. By Theorem 5.2

$$\|\hat{g} - g_0\|_n = \mathcal{O}_P\left(n^{-\frac{kp}{2kp+2k+p}}\right)$$

for i.i.d. errors satisfying $\mathbb{E}|\varepsilon_1|^{2p} < \infty$. If condition (5.1) is met, however, we find that the $L^2(P_n)$ distance between $\hat{g}$ and g_0 is of order $n^{-k/(2k+1)}$. Note that $n^{-kp/(2pk+2k+p)}$ converges to the optimal rate $n^{-k/(2k+1)}$ as $p \to \infty$.

5.3 Lower bounds

So far, we have not discussed lower bounds for the rates of convergence. In this section we show that at least in some interesting situations, Theorem 5.1 yields optimal rates.

For $0 < \delta < 2R\delta_n$, $g \in \mathcal{G}$, we typically encounter the entropy behavior

$$H_2(\delta, P_n, \mathcal{G}_n(g; R\delta_n)) \asymp \delta^{-V}, \ 0 \le V < 2. \tag{5.12}$$

Then the rule of thumb $H_2(\delta_n, P_n, \mathcal{G}_n(g; R\delta_n)) \asymp n\delta_n^2$ is solved by $\delta_n = n^{-\frac{1}{2+V}}$. The following theorem states that this rate $n^{-\frac{1}{2+V}}$ is the best one can obtain. But first we have to make precise what is meant by an optimal rate. Recall that an estimator $\hat{g}$ converges with rate δ_n if

$$\limsup_{R\to\infty} \limsup_{n\to\infty} \mathbb{P}\left\{\|\hat{g} - g_0\|_n > L\delta_n\right\} = 0.$$

We call such a rate optimal if

$$\liminf_{n\to\infty} \mathbb{P}\left\{\|\hat{g} - g_0\|_n > \alpha_n\delta_n\right\} > 0$$

for every sequence $\alpha_n \downarrow 0$ as $n \to \infty$.

Theorem 5.3 *Consider model 3 with $\sigma = 1$. Suppose ε_i are Gaussian $\mathcal{N}(0, \sigma_i^2)$ random variables, where*

$$\inf_{i \geq 1} \sigma_i^2 > 0 \quad \textit{and} \quad \sup_{i \geq 1} \sigma_i^2 < \infty,$$

and assume (5.12) holds true. Then for all sequences $\alpha_n \downarrow 0$,

$$\liminf_{n \to \infty} \mathbb{P}\left\{ d_n(\hat{g}, g_0) > \alpha_n n^{-\frac{1}{2+V}} \right\} > 0. \tag{5.13}$$

Proof. Let $\delta_n = n^{-\frac{1}{2+V}}$. The entropy bound (5.12) and the Gaussian distribution of ε_i entail that $Z_n(R\delta_n) \xrightarrow{a.s.} 0$, with

$$Z_n(R) = \|n^{-1/2} m_n(g)\|_{\mathcal{G}_n(R)},$$

and since $Z_n(R)$ is uniformly integrable, we also have convergence in mean. Let us now prove that $\mathbb{E} Z_n(R\delta_n) \asymp R^{\frac{2-V}{2}} \delta_n^2$.

Using Sudakov's lower bound (Theorem 2.1), which is feasible by the normality assumption on the errors, we find by (5.12),

$$\begin{aligned} \mathbb{E} \sup_{g \in \mathcal{G}_n(R\delta_n)} m_n(g) &\geq C \sup_{x>0} x \sqrt{H_2(x, P_n, \mathcal{G}_n(R\delta_n))} \\ &\geq CR\delta_n \sqrt{H_2(R\delta_n, P_n, \mathcal{G}_n(R\delta_n))} \\ &= C^* R^{\frac{2-V}{2}} \sqrt{n} \delta_n^2. \end{aligned} \tag{5.14}$$

By Dudley's upper bound (Theorem 2.2)

$$\begin{aligned} \mathbb{E} \sup_{g \in \mathcal{G}_n(R\delta_n)} m_n(g) &\leq \tilde{C} \int_0^{2R\delta_n} \sqrt{H_2(x, P_n, \mathcal{G}_n(R\delta_n))}\, dx \\ &\leq \tilde{C} R\delta_n \int_0^2 \sqrt{H_2(xR\delta_n, P_n, \mathcal{G}_n(R\delta_n))}\, dx \\ &= \overline{C} R^{\frac{2-V}{2}} \sqrt{n} \delta_n^2. \end{aligned} \tag{5.15}$$

It now follows from (5.14) and (5.15) that indeed $\mathbb{E} Z_n(R\delta_n) \asymp R^{\frac{2-V}{2}} \delta_n^2$.

Consequently, with probability tending to one, we find that $\sup_{g \in \mathcal{G}_n(r\delta_n)} L_n(g) = \mathcal{O}(\delta_n^2)$ for $r \in (0, r_0)$ and some $r_0 > 0$. On the other hand we have, with probability tending to one, for each sequence $\alpha_n \downarrow 0$, that $\sup_{g \in \mathcal{G}_n(\alpha_n \delta_n)} L_n(g) = o(\delta_n^2)$. Since the LSE maximizes L_n, it is clearly impossible that $\hat{g}$ lies within the ball $\mathcal{G}_n(\alpha_n \delta_n)$ with probability tending to one. □

5.4 Stochastic design

In this section we shall consider the case of stochastic design. Let $X_1, \ldots, X_n$ be i.i.d. random variables with common probability measure P on $\mathbb{R}^k$. The results derived in the previous paragraph can easily be restated for the stochastic case. Since the sequence $X_1, X_2, \ldots$ is no longer deterministic, the entropies involved in Theorem 5.1 are stochastic. For this matter, we define the random variables

$$W_{n,R} = \frac{\int_0^2 \sqrt{H_2(xR\delta_n, P_n, \mathcal{G}_n(R\delta_n))}\, dx}{\sqrt{n}\delta_n R}, \quad R \geq 0, \quad n = 1, 2, \ldots \quad (5.16)$$

Recall that $\delta_n \downarrow 0$ and $n\delta_n^2 \geq 1$ as $n \to \infty$. If there exists a deterministic sequence $\{\alpha_R\}$, $\alpha_R \downarrow 0$ as $R \to \infty$ such that

$$\limsup_{n\to\infty} \mathbb{P}\left\{ \bigcup_{R=R_0}^{\infty} \{W_{n,R} > \alpha_R\} \right\} = 0, \text{ for some } R_0 > 0,$$

it follows from the proof of Theorem 5.1 that $\|\hat{g} - g_0\|_n = \mathcal{O}_P(\delta_n)$.

Next, since we are dealing with random measures, one might ask if the theoretical distances show similar behavior. To cope with this question, we need an adequate link between these related distances. We recall the following lemma which can be found in Pollard [35].

Lemma 5.1 *Let $\mathcal{F}$ be a permissible class of functions with $(\int f^2\, dP)^{\frac{1}{2}} \leq \delta$ and $\|f\|_\infty \leq 1$ for each f in $\mathcal{F}$. Then*

$$\mathbb{P}\left\{ \sup_{\mathcal{F}} \left(\int f^2\, dP_n \right)^{\frac{1}{2}} > 8\delta \right\} \leq 4\mathbb{E}[N_2(\delta, P_n, \mathcal{F}) \exp(-n\delta^2) \wedge 1]. \quad (5.17)$$

Proof. See Pollard [35], p. 31. □

We shall use a slight modification of this lemma to get rid of the mathematical expectation in (5.17). Define

$$\mathcal{G}(R) = \{g \in \mathcal{G} : \|g - g_0\|_2 \leq R\}, \quad R = 1, 2, \ldots$$

and

$$A_{n,j} = \left\{ H_2(j\delta_n, P_n, \mathcal{G}(j\delta_n)) \leq \frac{n\delta_n^2 j^2}{4} \right\}$$
$$A_n = \cap_{j \geq 1} A_{n,j}. \quad (5.18)$$

Lemma 5.2 *Suppose $\mathcal{G}$ is uniformly bounded by 1. Then we have for $\delta_n > 0$ with $n\delta_n^2 \geq 1$ and $\delta_n \downarrow 0$ as $n \to \infty$, $L \geq 1$ and constants $c, C > 0$,*

$$\mathbb{P}\left\{\sup_{\|g\|_2 > L\delta_n} \left(\frac{\int g^2\, dP_n}{\int g^2\, dP}\right)^{\frac{1}{2}} > 16\right\} \leq C\exp\left(-cL^2 n\delta_n^2\right) + \frac{4}{3}\mathbb{P}\left\{A_n^c\right\},$$

where A_n is defined previously in (5.18).

Proof. Let P_n' be an independent copy of P_n. Because we have

$$\mathbb{P}\left\{\frac{1}{\|g\|_2}\left(\int g^2\, dP_n'\right)^{\frac{1}{2}} \leq 2\right\} \geq 1 - \frac{\mathbb{E}\int g^2\, dP_n'/\|g\|_2^2}{4} = \frac{3}{4},$$

we can apply Lemma 2.2 with $\varepsilon = 14$, $\alpha = 2$ and $\beta = 3/4$, whence

$$\mathbb{P}\left\{\sup_{\|g\|_2 > L\delta_n} \left(\frac{\int g^2\, dP_n}{\int g^2\, dP}\right)^{\frac{1}{2}} > 16\right\} \leq \tag{5.19}$$

$$\frac{4}{3}\mathbb{P}\left\{\sup_{\|g\|_2 > L\delta_n} \frac{1}{(\int g^2\, dP)^{\frac{1}{2}}}\left|\left(\int g^2\, dP_n\right)^{\frac{1}{2}} - \left(\int g^2\, dP_n'\right)^{\frac{1}{2}}\right| > 14\right\}.$$

A closer look at the proof of Lemma 5.1 reveals that, using the same notation as in Lemma 5.1 where $\|f\|_\infty \leq 1$ and $\|f\|_2 \leq \delta$, the following bound holds true with $C_n = \left\{H_2(\delta, \frac{1}{2}[P_n + P_n'], \mathcal{F}) \leq \frac{n\delta^2}{2}\right\}$,

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}}\left|\left(\int f^2 dP_n\right)^{\frac{1}{2}} - \left(\int f^2 dP_n'\right)^{\frac{1}{2}}\right| > 6\delta, C_n\right\}$$

$$\leq 3\exp(-\frac{n\delta^2}{2}). \tag{5.20}$$

Use that the independent measures P_n and P_n' have the same distribution and invoke the relation

$$H_2(\sqrt{2}\delta, \frac{1}{2}(P_n + P_n'), \mathcal{F}) \leq H_2(\delta, P_n, \mathcal{F}) + H_2(\delta, P_n', \mathcal{F}).$$

If we peel the event $\{\|g\|_2 \geq L\delta_n\}$ into countably many small annuli

$$\{(j-1)\delta_n \leq \|g\|_2 \leq j\delta_n\}, \quad j = L+1, L+2, \ldots$$

and transform inequality (5.20) to the present setting with $\delta = 2^j\delta_n$, we see that (5.19) can be bounded further by

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{\|g\|_2>L\delta_n}\left(\frac{\int g^2\,dP_n}{\int g^2\,dP}\right)^{\frac{1}{2}} > 16\right\} \leq \\
&\frac{4}{3}\mathbb{P}\left\{\sup_{\|g\|_2\geq L\delta_n}\frac{1}{(\int g^2\,dP)^{\frac{1}{2}}}\left|\left(\int g^2\,dP_n\right)^{\frac{1}{2}} - \left(\int g^2\,dP_n'\right)^{\frac{1}{2}}\right| > 12, A_n\right\} + \\
&\quad + \frac{4}{3}\mathbb{P}\{A_n^c\} \leq \\
&\frac{4}{3}\sum_{j=j_0}^{\infty}\mathbb{P}\left\{\sup_{\|g\|_2\leq 2^j\delta_n}\left|\left(\int g^2\,dP_n\right)^{\frac{1}{2}} - \left(\int g^2\,dP_n'\right)^{\frac{1}{2}}\right| > 6(2^j\delta_n), A_{n,2^j}\right\} \\
&\quad + \frac{4}{3}\mathbb{P}\{A_n^c\} \leq \\
&\leq 4\sum_{j=j_0}^{\infty}\exp(-2^{2j-1}n\delta_n^2) + \frac{4}{3}\mathbb{P}\{A_n^c\},
\end{aligned}$$

whence the result follows. □

Now we are in a position to prove the following result.

Theorem 5.4 *Consider model 1 with σ fixed. Suppose that the sequence $\varepsilon_1, \varepsilon_2, \ldots$ is uniformly subgaussian, and that $\mathcal{G}$ is uniformly bounded by $B > 0$. Let $\delta_n > 0$ with $n^{-1/2} \leq \delta_n \downarrow 0$ as $n \to \infty$. Finally, assume that*

$$\limsup_{n\to\infty}\mathbb{P}\left\{\bigcup_{R=R_0}^{\infty}\left\{\frac{\int_0^2\sqrt{H_2(xR\delta_n, P_n, \mathcal{G})}\,dx}{\sqrt{n}\delta_n R} > \alpha_R\right\}\right\} = 0 \qquad (5.21)$$

for a deterministic sequence $\{\alpha_R\}$, $\alpha_R \downarrow 0$ as $R \to \infty$. Then we have

$$\|\hat{g} - g_0\|_n = \mathcal{O}_P(\delta_n) \quad \textit{as well as} \quad \|\hat{g} - g_0\|_2 = \mathcal{O}_P(\delta_n).$$

Proof. We only prove the second result, because the rate for $\|\hat{g} - g_0\|_n$ follows fairly straightforward from Theorem 5.1 as explained in the beginning of this section.

Without loss of generality we take the constants $\sigma = 1$ and $B = 1$.

Since $S_n(g_0) - S_n(g)$ is maximized over the set $\mathcal{G}$ by the LSE $\hat{g}$, we have for $L \geq R_0$

$$\mathbb{P}\{\|\hat{g} - g_0\|_2 > L\delta_n\} \leq P_L^{(1)} + P_L^{(2)},$$

with

$$P_L^{(1)} = \mathbb{P}\left\{ \sup_{\|g-g_0\|_2 > L\delta_n} \left| 2n^{-1/2} m_n(g) - \frac{1}{2}\|g-g_0\|_2^2 \right| \geq 0 \right\}$$

and

$$P_L^{(2)} = \mathbb{P}\left\{ \sup_{\|g-g_0\|_2 > L\delta_n} \left| \int (g-g_0)^2 \, d(P-P_n) - \frac{1}{2}\|g-g_0\|_2^2 \right| \geq 0 \right\}.$$

By the peeling device as used in the proofs of Theorem 5.1 and Lemma 5.2, we can write

$$P_L^{(1)} \leq \mathbb{P}\{B_n^c\} + \sum_{j=l}^{\infty} \tilde{P}_j^{(1)}$$

with

$$B_n = \left\{ \sup_{\|g\|_2 > L\delta_n} \left(\frac{\int g^2 \, dP_n}{\int g^2 \, dP} \right)^{\frac{1}{2}} \leq 16 \right\},$$

and

$$\begin{aligned}
\tilde{P}_j^{(1)} &= \mathbb{P}\left\{ \sup_{2^{j-1}\delta_n \leq \|g-g_0\|_2 \leq 2^j \delta_n} \left| 2n^{-1/2} m_n(g) - \frac{1}{2}\|g-g_0\|_2^2 \right| \geq 0, \ B_n \right\} \\
&\leq \mathbb{P}\left\{ \sup_{2^{j-1}\delta_n \leq \|g-g_0\|_2 \leq 2^j \delta_n} \left| 2n^{-1/2} m_n(g) \right| \geq \frac{1}{4}(2^j\delta_n)^2, \ B_n \right\} \\
&\leq \mathbb{P}\left\{ \sup_{\|g-g_0\|_n \leq 2^{4+j}\delta_n} \left| 2n^{-1/2} m_n(g) \right| \geq \frac{1}{4}(2^j\delta_n)^2 \right\}.
\end{aligned}$$

Arguing as in the proof of Theorem 5.1, we see that

$$\sum_{j=l}^{\infty} \tilde{P}_j^{(1)} \leq 3\exp\left(-\kappa l^2 n\delta_n^2\right)$$

for some $\kappa > 0$. An application of Lemma 5.2 implies that $\limsup_{n,L\to\infty} \mathbb{P}\{B_n^c\} = 0$. Therefore it remains to prove an adequate upper bound for the probabilities $P_L^{(2)}$.

Note first of all that, using $\|g\|_\infty \leq 1$,

$$\mathrm{Var}\left(\frac{\int (g-g_0)^2 \, dP_n}{\|g-g_0\|_2^2} \right) \leq 4n^{-1}\|g-g_0\|_2^{-2}.$$

Hence by Chebyshev's inequality

$$\mathbb{P}\left\{\frac{\int (g-g_0)^2\,dP_n}{\|g-g_0\|_2^2} > \frac{1}{4}\right\} \leq 2^{6-2L}(n\delta_n^2)^{-1} \leq 2^{-1},$$

for all $g \in \mathcal{G}$ satisfying $\|g-g_0\|_2 \geq L\delta_n$ and $L = 4, 5, \ldots$. By Lemma 2.2, it follows that

$$\begin{aligned} P_L^{(2)} &\leq 2\mathbb{P}\left\{\sup_{\|g-g_0\|_2 > L\delta_n} \frac{\int (g-g_0)^2\,d(P_n - P_n')}{\|g-g_0\|_2^2} \geq \frac{1}{4}\right\} \\ &\leq 4\mathbb{P}\left\{\sup_{\|g-g_0\|_2 > L\delta_n} \frac{\int (g-g_0)^2\,dP_n^0}{\|g-g_0\|_2^2} \geq \frac{1}{16}\right\}, \end{aligned}$$

where P_n' is an independent copy of P_n and P_n^0 is the symmetrized empirical measure.

Using the peeling argument once more we obtain

$$\begin{aligned} P_L^{(2)} &\leq 4\mathbb{P}\left\{\sup_{\|g-g_0\|_2 > L\delta_n} \frac{\int (g-g_0)^2\,dP_n^0}{\|g-g_0\|_2^2} \geq \frac{1}{16}\right\} \\ &\leq 4\sum_{j=l}^{\infty} \mathbb{P}\left\{\sup_{2^{j-1}\delta_n < \|g-g_0\|_2 < 2^j\delta_n} \int (g-g_0)^2\,dP_n^0 \geq \frac{(2^{j-1}\delta_n)^2}{16},\ B_n\right\} \\ &\quad +4\mathbb{P}\{B_n^c\} \\ &\leq 4\mathbb{P}\{B_n^c\} + \\ &\quad +4\sum_{j=l}^{\infty} \mathbb{P}\left\{\sup_{\|g-g_0\|_n < 2^{j+4}\delta_n} \int (g-g_0)^2\,dP_n^0 \geq \frac{1}{16}(2^{j-1}\delta_n)^2\right\} \\ &\leq 4\mathbb{P}\{B_n^c\} + \\ &\quad +4\sum_{j=l}^{\infty} \mathbb{P}\left\{\sup_{\|(g-g_0)^2\|_n < 2^{j+4}\delta_n} \int (g-g_0)^2\,dP_n^0 \geq \frac{1}{16}(2^{j-1}\delta_n)^2\right\}, \end{aligned}$$

where we used $\|g\|_\infty \leq 1$ in the last inequality. Observe that

$$H_2\left(\sqrt{12}\delta, P_n, \left\{(g-g_0)^2 :\ g \in \mathcal{G}\right\}\right) \leq H_2(\delta, P_n, \mathcal{G}) \quad \text{for all } \delta > 0.$$

Conclude from Lemma 2.5 and the entropy condition (5.21) that also $P_L^{(2)} \to 0$ for $n \to \infty$, $L \to \infty$. □

Remark 5.4 In Van de Geer [40] a special case of Theorem 5.4 is treated. The main difference is that a different kind of entropy (entropy with bracketing) is used to facilitate the change from the empirical $L^2(P_n)$ to the theoretical $L^2(P)$ distance.

In Van de Geer [42] and Wegkamp [52] similar techniques to switch between these related distances are used in the context of nonparametric likelihood estimation.

Chapter 6

Some asymptotic distribution theory

The lack of a stochastic expansion for the nonparametric least squares estimator, which is due to its implicit definition, makes it difficult to obtain pointwise asymptotic results, such as the pointwise convergence

$$\alpha_n \left(\hat{g}(x) - g_0(x)\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

where α_n are normalizing constants.

Central limit theorems (CLT's) for monotone regression functions are well-known. See e.g. Leurgans [28], where it is shown that a suitably normalized version of the least squares estimator converges in distribution to a normal law, pointwise. Groeneboom (cf. [16], [18]) proved asymptotic normality of the Grenander estimator pointwise and in L^1, which is related to our least squares estimating problems.

In this chapter we shall prove two different central limit theorems. The first one concerns the asymptotic distribution of the squared $L^2(P_n)$ norm of the least squares estimator in regression model 3, i.e.

$$\mathbb{P}\left\{\alpha_n \sum_{i=1}^{n} \left(\hat{g}^2(x_i) - g_0^2(x_i)\right) \leq x\right\} \to \Phi(x), \quad n \to \infty,$$

uniformly in x, with $\Phi(x)$ the standard normal distribution function, and α_n normalizing constants. We shall restrict attention to the classical Sobolev spaces.

In Section 6.2 we consider a partial linear model and prove asymptotic normality of the estimator of the parametric component. The design variables $X_i = (X_i^{(1)}, X_i^{(2)})$ are stochastic and take values in $[0,1] \times [0,1]$. The regression function evaluated at X_i takes the form

$$g(X_i) = \theta X_i^{(1)} + f(X_i^{(2)}).$$

We impose smoothness conditions on the function f. In general, one has to use a smoothness penalty on the sum of squares for consistent least squares estimation. However, under additional restrictions on f, this appears to be unnecessary.

6.1 A CLT for the empirical norm of the LSE

We focus again on the regression model with deterministic design and heteroscedastic disturbances, which agrees with model 3 with $\sigma = 1$. The aim is to prove that the distribution of

$$\sqrt{n} \int \left(\hat{g}^2(x) - g_0^2(x) \right) dP_n(x)$$

is Gaussian in the limit. We first present an informal discussion why this may be true.

Recall that the process $S_n(g)$, $g \in \mathcal{G}$ is given by

$$S_n(g) = \frac{1}{n} \sum_{i=1}^{n} [Y_i - g(x_i)]^2, \ g \in \mathcal{G}. \tag{6.1}$$

If $\mathcal{G}$ is an open subset of some vector space B, then we must have

$$\frac{d}{dt} S_n(\hat{g} + th) \Big|_{t=0} = 0 \quad \forall h \in B.$$

Computing this Gateaux derivative at g in the direction h gives us

$$\begin{aligned} \psi(g;h) &= \lim_{t \to 0} \frac{S_n(g+th) - S_n(g)}{t} \\ &= \lim_{t \to 0} \frac{1}{n} \sum_{i=1}^{n} \left[-2 \left(Y_i - g(x_i) \right) h(x_i) + t h^2(x_i) \right] \\ &= -2 \frac{1}{n} \sum_{i=1}^{n} \left[\varepsilon_i + (g_0(x_i) - g(x_i)) \right] h(x_i) \end{aligned}$$

for $\int h^2\, dP_n < \infty$. As a result, we obtain the following identity

$$\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i h(x_i) = \frac{1}{n}\sum_{i=1}^{n} \left[\hat{g}(x_i) - g_0(x_i)\right] h(x_i). \tag{6.2}$$

If we choose the direction $h = \hat{g} + g_0$, then we find

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \varepsilon_i \left[\hat{g}(x_i) + g_0(x_i)\right] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \left[\hat{g}^2(x_i) - g_0^2(x_i)\right]$$

The left-hand side can be rewritten as

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \varepsilon_i \left[\hat{g}(x_i) - g_0(x_i)\right] + \frac{2}{\sqrt{n}}\sum_{i=1}^{n} \varepsilon_i g_0(x_i) := Z_1 + Z_2.$$

If $n^{-1}\sum_{i=1}^{n} \sigma_i^2 g_0^2(x_i)$ converges, then, as a consequence of the central limit theorem, Z_2 is asymptotically Gaussian. We know under which circumstances $Z_1 \xrightarrow{P} 0$. Namely, we have to show that $\|\hat{g} - g_0\|_n \to 0$ in probability, and that the process $m_n(g) = n^{-1/2}\sum_{i=1}^{n} \varepsilon_i (g(x_i) - g_0(x_i))$ is stochastically equicontinuous at g_0 with respect to d_n. The latter can be derived by means of the maximal inequalities of Chapter 2.

Unfortunately, in regression problems, $\mathcal{G}$ is often not open and $\hat{g}$ may lie on the boundary of $\mathcal{G}$. It is reasonable to assume that g_0 is an interior point. For convex $\mathcal{G}$, one may hope that

$$(1-\alpha)\hat{g} + \alpha g_0 + t(\hat{g} + g_0) \in \mathcal{G} \tag{6.3}$$

for special small choices α and t. In a different context of maximum likelihood estimation, the same idea of taking a convex combination has been applied successfully earlier by Van de Geer (cf. [43]). If (6.3) holds, we can actually establish

$$\frac{1}{n}\sum_{i=1}^{n} \left[\hat{g}^2(x_i) - g_0^2(x_i)\right] = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \left(\hat{g}(x_i) + g_0(x_i)\right) + o_P(n^{-1/2}). \tag{6.4}$$

We shall consider a special class for which this convexity argument holds true. To be more specific, we take the following Sobolev space

$$\mathcal{G} = \{g : [0,1] \to \mathbb{R},\ \|g\|_\infty \le C_1,\ \|g\|_{TV} \le C_2\}, \tag{6.5}$$

where C_1 and C_2 are known positive constants and $\|\cdot\|_\infty$ and $\|\cdot\|_{TV}$ are the supremum and the total variation norm on $\mathcal{G}$ respectively. Recall that the total variation norm is defined by

$$\|g\|_{TV} = \sup\left\{ \sum_{i=1}^{n} |g(x_i) - g(x_{i-1})| \;\middle|\; x_0 < x_1 < \ldots < x_n,\, x_i \in [0,1] \right\}.$$

Our assumption that g_0 is an interior point of $\mathcal{G}$ should be understood in terms of the metrics involved, i.e. both $\|g_0\|_{TV} < C_2$ and $\|g_0\|_\infty < C_1$ hold true.

Theorem 6.1 *Consider model 3 with $\sigma = 1$. Assume that g_0 is an interior point of the class $\mathcal{G}$ as defined in (6.5), and*

$$\liminf_{n\to\infty} \|g_0\|_n > 0,$$

and that the errors $\varepsilon_1, \varepsilon_2, \ldots$ fulfill condition (5.1). Then we have

$$\frac{1}{2S_n} \sum_{i=1}^{n} \left[\hat{g}^2(x_i) - g_0^2(x_i)\right] \xrightarrow{\mathcal{D}} \mathcal{N}(0,1), \tag{6.6}$$

with $S_n^2 = g_0^2(x_1)\sigma_1^2 + \ldots + g_0^2(x_n)\sigma_n^2$.

Remark 6.1 Theorem 6.1 is valid for many more classes than the one considered. For instance, the result holds true for the Sobolev space

$$\left\{ g : [0,1] \to \mathbb{R},\; \|g\|_\infty \le C_1,\; \int \left[g^{(m)}(x)\right]^2 dx \le C_2 \right\}.$$

We prove that (6.4) is satisfied and that the process $m_n(\cdot)$ is stochastically equicontinuous at g_0 with respect to d_n. The latter follows from suitable entropy conditions. To show (6.4), we need entropy conditions as well as technical conditions on $\mathcal{G}$ to ensure that (6.3) holds for α and t small enough.

Other consequences of the identity (6.2) are obtained by choosing different directions h. For instance, take $h = 1$ and suppose $n^{-1}\sum_{i=1}^n \sigma_i^2$ converges. As a consequence of (6.2), we find

$$\frac{1}{\sqrt{\sum_{i=1}^n \sigma_i^2}} \sum_{i=1}^{n} [\hat{g}(x_i) - g_0(x_i)] \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

which is equivalent with

$$\frac{1}{\sqrt{\sum_{i=1}^n \sigma_i^2}} \sum_{i=1}^n [\hat{\varepsilon}_i - \varepsilon_i] \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

where $\hat{\varepsilon}_i = y_i - \hat{g}(x_i)$ are the estimated residuals.

6.1.1 Proof of Theorem 6.1

We need two lemma's.

Lemma 6.1 (rates of convergence) *Under the conditions of Theorem 6.1, we have*

$$d_n(\hat{g}, g_0) = \mathcal{O}_P(n^{-1/3}) \quad \textit{and} \quad n^{-1/2} m_n(\hat{g}) = \mathcal{O}_P(n^{-2/3}).$$

Proof. Note that $H_2(\delta, P_n, \mathcal{G}) \asymp 1/\delta$ (cf. Example 2.3). The first assertion follows from Theorem 5.1. The second assertion follows from the relation

$$\begin{aligned} \mathbb{P}\left\{n^{-1/2} m_n(\hat{g}) > L\delta_n^2\right\} \quad \leq \quad & \mathbb{P}\left\{\sup_{\mathcal{G}_n(R\delta_n)} n^{-1/2} m_n(g) > L\delta_n^2\right\} + \\ & + \mathbb{P}\left\{\|\hat{g} - g_0\|_n > R\delta_n\right\}, \end{aligned}$$

the first assertion and Theorem 2.3. □

Lemma 6.2 *Under the conditions of Theorem 6.1, identity (6.4) is true.*

Proof. Consider for each element $g \in \mathcal{G}$ the following convex combination

$$g_\alpha = (1-\alpha)g + \alpha g_0, \ 0 \leq \alpha \leq 1.$$

For notational convenience, we write $h(g) = g + g_0$ and $g_{\alpha,t} = g_\alpha + th(g)$ with $t \in \mathbb{R}$. Also, $\hat{h} = h(\hat{g})$, $\hat{g}_\alpha = (1-\alpha)\hat{g} + \alpha g_0$, and $\hat{g}_{\alpha,t} = \hat{g}_\alpha + t\hat{h}$.

Without loss of generality we assume that the constants C_1 and C_2 appearing in the definition of the class $\mathcal{G}$ both equal one. Moreover, since the function g_0 is an interior point of $\mathcal{G}$, we assume $\|g_0\|_\infty \leq 1/2$. Note

that $\|\hat{g}_{\alpha,t}\|_\infty < 1$ for $\alpha = 4|t|$ sufficiently small. The same arguments can be repeated with the sup-norm replaced by the total variation norm to prove that $\|\hat{g}_{\alpha,t}\|_{TV} < 1$ for $\alpha = 4|t|$ sufficiently small.

We have

$$\begin{aligned} & S_n(g_{\alpha,t}) - S_n(g_\alpha) = \\ & = \frac{1}{n}\sum_{i=1}^n ((y_i - g_\alpha(x_i)) - th(g)(x_i))^2 - \frac{1}{n}\sum_{i=1}^n ((y_i - g_\alpha(x_i))^2 \\ & = -2t\frac{1}{n}\sum_{i=1}^n (y_i - g_\alpha(x_i))\, h(g)(x_i) + t^2\frac{1}{n}\sum_{i=1}^n h^2(g)(x_i). \end{aligned} \tag{6.7}$$

Choose

$$\hat{t}_n = \frac{\sum_{i=1}^n (y_i - \hat{g}_\alpha(x_i))\, \hat{h}(x_i)}{\sum_{i=1}^n \hat{h}^2(x_i)}.$$

Then (6.7) with t and g replaced by $\hat{t}_n$ and $\hat{g}$ reads

$$\begin{aligned} & S_n(\hat{g}_{\alpha,\hat{t}_n}) - S_n(\hat{g}_\alpha) = \\ & = -\left(\frac{1}{n}\sum_{i=1}^n (y_i - \hat{g}_\alpha(x_i))\, \hat{h}(x_i)\right)^2 \Big/ \frac{1}{n}\sum_{i=1}^n \hat{h}^2(x_i) \\ & = -\left(\frac{1}{n}\sum_{i=1}^n [\varepsilon_i + (1-\alpha)(g_0 - \hat{g})(x_i)]\, \hat{h}(x_i)\right)^2 \Big/ \frac{1}{n}\sum_{i=1}^n \hat{h}^2(x_i). \end{aligned} \tag{6.8}$$

Since $\hat{g}$ minimizes $S_n(g)$, we have

$$\begin{aligned} & S_n(\hat{g}_{\alpha,\hat{t}_n}) - S_n(\hat{g}_\alpha) \geq S_n(\hat{g}) - S_n(\hat{g}_\alpha) = \\ & = 2\alpha\frac{1}{n}\sum_{i=1}^n (y_i - \hat{g}(x_i))(g_0(x_i) - \hat{g}(x_i)) - \alpha^2\frac{1}{n}\sum_{i=1}^n (g_0(x_i) - \hat{g}(x_i))^2 \\ & = 2\alpha\frac{1}{n}\sum_{i=1}^n \varepsilon_i(g_0(x_i) - \hat{g}(x_i)) + (2\alpha - \alpha^2)\frac{1}{n}\sum_{i=1}^n (g_0(x_i) - \hat{g}(x_i))^2. \end{aligned}$$

Now Lemma 6.1 implies further that

$$\begin{aligned} S_n(\hat{g}_{\alpha,\hat{t}_n}) - S_n(\hat{g}_\alpha) \;\geq\; & -2\alpha n^{-1/2} m_n(\hat{g}) + (2\alpha - \alpha^2) d_n^2(\hat{g}, g_0) = \\ & = \mathcal{O}_P(\alpha\delta_n^2). \end{aligned} \tag{6.9}$$

Moreover, since

$$\frac{1}{n}\sum_{i=1}^n \hat{h}^2(x_i) = \frac{1}{n}\sum_{i=1}^n (2g_0(x_i) + (\hat{g} - g_0)(x_i))^2 = 4\|g_0\|_n^2 + \mathcal{O}_P(\delta_n^2),$$

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\hat{h}(x_i) = \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i g_0(x_i) + n^{-1/2}m_n(\hat{g}) = \mathcal{O}_P(n^{-1/2})$$

by the central limit theorem and Lemma 6.1, and

$$\frac{1}{n}\sum_{i=1}^{n}(g_0(x_i) - \hat{g}(x_i))\hat{h}(x_i) = \mathcal{O}_P(n^{-1/3})$$

by the Cauchy-Schwarz inequality, we have $\hat{t}_n = \mathcal{O}_P(n^{-1/3})$.

Set $\hat{\alpha}_n = 4|\hat{t}_n| = \mathcal{O}_P(n^{-1/3})$, then we find from (6.8) and (6.9)

$$\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\hat{h}(x_i) - (1-\hat{\alpha}_n)\frac{1}{n}\sum_{i=1}^{n}\left[\hat{g}^2(x_i) - g_0^2(x_i)\right]\right)^2 = \mathcal{O}_P(n^{-1}). \quad (6.10)$$

Because, as already noted $n^{-1}\sum_{i=1}^{n}\varepsilon_i\hat{h}(x_i) = \mathcal{O}_P(n^{-1/2})$, we obtain from (6.10) that also $n^{-1}\sum_i[\hat{g}^2(x_i) - g_0^2(x_i)] = \mathcal{O}_P(n^{-1/2})$. But then $\hat{t}_n = \mathcal{O}_P(n^{-1/2})$ and (6.10) becomes now

$$\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\hat{h}(x_i) - (1-\hat{\alpha}_n)\frac{1}{n}\sum_{i=1}^{n}\left[\hat{g}^2(x_i) - g_0^2(x_i)\right]\right)^2 = o_P(n^{-1}). \quad (6.11)$$

The proof of the lemma is complete. □

Now, we can prove Theorem 6.1, since by Lemma 6.1 we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{g}(x_i) + g_0(x_i)) &= \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{g}(x_i) - g_0(x_i)) + \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i g_0(x_i) \\ &= \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i g_0(x_i) + \mathcal{O}_P(n^{-2/3}).\end{aligned}$$

Invoke (6.4), proved in Lemma 6.2, and the CLT to conclude (6.6). □

6.2 Partial linear models

6.2.1 The model

We consider the following regression model

$$Y_i = g(X_i) + \varepsilon_i, \quad (i = 1, \ldots, n),$$

where

- the observation points X_i are i.i.d. two dimensional random variables with mass concentrated on the unit square, i.e. $\mathbb{P}\{X \in [0,1] \times [0,1]\} = 1$. We write $X = (X^{(1)}, X^{(2)})$ and assume that

$$\mathbb{E}\left(X^{(1)} - \mathbb{E}(X^{(1)} \mid X^{(2)})\right)^2 > 0 \tag{6.12}$$

- the disturbances ε_i are i.i.d. centered random variables with

$$\mathbb{E} \exp\left(\lambda \varepsilon_1^2\right) < \infty \tag{6.13}$$

for some constant $\lambda > 0$. Moreover we assume that X_i and ε_i are independent.

- the regression function g consists of a linear part and a smooth part,

$$g(x_1, x_2) = \theta x_1 + f(x_2), \quad x_1, x_2 \in [0,1]. \tag{6.14}$$

Here $\theta \in \mathbb{R}$ and $f \in \mathcal{F}$, with

$$\mathcal{F} = \{f : [0,1] \to \mathbb{R},\ \|f\|_\infty \leq C,\ J_m^2(f) \leq C'\},$$

where

$$J_m^2(f) = \int_0^1 |f^{(m)}(x)|^2 \, dx.$$

In the sequel we write

$$\mathcal{G} = \{g : [0,1] \times [0,1] \to \mathbb{R},\ g(x_1, x_2) = \theta x_1 + f(x_2) \ \mid \theta \in \mathbb{R},\ f \in \mathcal{F}\}.$$

Throughout we assume that the unknown regression function $g = g_0$ is an interior point of the parameter space $\mathcal{G}$, in the sense that $\|f_0\|_\infty < C$ and $J_m(f_0) < C'$.

This model is also investigated in Mammen & Van de Geer [30]. In that paper the unknown regression function g is estimated by the penalized least squares estimator, which means that

$$S_n(g) + \lambda_n^2 J_m^2(g)$$

is minimized over all $g \in \mathcal{G}$ where C' is unknown. However, if this constant is known, one can take $\lambda_n = 0$; the minimization has just become a least squares estimation problem and one does not have to cope with the technicalities involved in the choice of the smoothing parameter λ_n. Note that the class $\mathcal{G}$ is small enough in the sense that it fulfills the entropy conditions of Theorem 4.2. Consequently, the least squares estimator $\hat{g}(x_1, x_2) = \hat{\theta} x_1 + \hat{f}(x_2)$ is consistent.

Theorem 6.2 (consistency and rates) *Consider the partial linear model described above. We have* $\|\hat{g} - g_0\|_n \xrightarrow{a.s.} 0$ *and* $\|\hat{g} - g_0\|_2 \xrightarrow{a.s.} 0$. *Also* $\|\hat{f} - f_0\|_2 \xrightarrow{a.s.} 0$ *and* $|\hat{\theta} - \theta_0| \xrightarrow{a.s.} 0$. *In fact,* $\|\hat{g} - g_0\|_n = \mathcal{O}_P(n^{-m/(2m+1)})$.

Proof. In order to apply Theorem 4.2, we have to compute the entropy of the class $\mathcal{G}_n(R)$. By a result of Kolmogorov & Tichomirov [23],

$$H_\infty(\delta, P_n, \mathcal{F}) = \mathcal{O}(\delta^{-1/m}).$$

Since $n^{-1} \sum_{i=1}^n [X_i^{(1)}]^2 \xrightarrow{a.s.} \mathbb{E}[X^{(1)}]^2$ and $\|f\|_\infty \leq C$, there exists some $R' > 0$, depending on R and C, such that $|\theta| \leq R'$ almost surely for n large enough. Note that

$$H_2(\delta, P_n, \mathcal{G}_n(R)) \leq H_2(\delta/2, P_n, \mathcal{F}) + H_2(\delta/2, P_n, \{\theta X^{(1)}, |\theta| \leq R'\})$$

for every $\delta > 0$ and for large n. As a result we have $H_\infty(\delta, P_n, \mathcal{G}_n(R)) = \mathcal{O}(\delta^{-1/m})$. Theorem 4.2 asserts that $\|\hat{g} - g_0\|_n \xrightarrow{a.s.} 0$ and the rate $n^{-m/(2m+1)}$ follows from Theorem 5.1. Because the entropy bounds are valid with probability one, the fact that the design $X_1, \ldots, X_n$ is stochastic does not cause any problem. See the discussion in Section 5.4.

The $L^2(P)$-convergence $\|\hat{g} - g_0\|_2 \xrightarrow{a.s.} 0$ follows from Theorem 3.1.2 in Van de Geer [40].

Note that $g(X)$ can be decomposed into two orthogonal parts

$$g(X) = \left[f(X^{(2)}) + \theta \mathbb{E}\left(X^{(1)} \mid X^{(2)}\right)\right] + \left[\theta\left(X^{(1)} - \mathbb{E}\left(X^{(1)} \mid X^{(2)}\right)\right)\right].$$

Since $\|X^{(1)} - \mathbb{E}\left(X^{(1)} \mid X^{(2)}\right)\|_2 > 0$, we also have almost sure convergence of the components $|\hat{\theta} - \theta_0| \xrightarrow{a.s.} 0$ and $\|\hat{f} - f_0\|_2 \xrightarrow{a.s.} 0$ separately. □

6.2.2 Asymptotic normality

Having obtained rates of convergence, we proceed by showing asymptotic normality of $\sqrt{n}(\hat{\theta} - \theta_0)$. We shall employ the same arguments as in the proof of Lemma 6.2 in Section 6.1. We exploit the fact that $S_n(\cdot)$ is minimized at $\hat{g}$ and we consider a special direction $h(X) = X^{(1)} - \mathbb{E}\left(X^{(1)} \mid X^{(2)}\right)$.

Theorem 6.3 *Consider the model described in Subsection 6.2.1. In addition, assume* $J_m\left(\mathbb{E}\left(X^{(1)} \mid X^{(2)}\right)\right) < \infty$. *We have*

$$\hat{\theta} - \theta_0 = \frac{\sum_{i=1}^n \varepsilon_i h(X_i)}{\sum_{i=1}^n h^2(X_i)} + o_P\left(n^{-1/2}\right). \tag{6.15}$$

Proof. First we prove

$$\frac{1}{n}\sum_{i=1}^n \varepsilon_i h(X_i) = \frac{1}{n}\sum_{i=1}^n \left(\hat{g}(X_i) - g_0(X_i)\right) h(X_i) + o_P\left(n^{-1/2}\right). \tag{6.16}$$

Define g_α as in the proof of Lemma 6.2, i.e.

$$g_\alpha = (1-\alpha)g + \alpha g_0 \quad (0 \le \alpha \le 1)$$

and take $g_{\alpha,t} = g_\alpha + th$, with $t \in \mathbb{R}$, where the direction h is given by

$$h(X) = X^{(1)} - \mathbb{E}\left(X^{(1)} \mid X^{(2)}\right).$$

Now it is crucial that $\hat{g}_{\alpha,t} \in \mathcal{G}$ for sufficiently small α and t. Since g_0 lies in the interior of $\mathcal{G}$, we may assume that $\|f_0\|_\infty \le C/2$ and $J_m(f_0) \le C'/2$. Using the fact that $J_m\left(\mathbb{E}\left(X^{(1)} \mid X^{(2)}\right)\right) < \infty$ and $\left|\mathbb{E}\left(X^{(1)} \mid X^{(2)}\right)\right| \le 1$ with probability one, we conclude that

$$\begin{aligned}\hat{g}_{\alpha,t}(X) + th(X) = &\left[(1-\alpha)\hat{\theta} + t + \alpha\theta_0\right] X^{(1)} + \\ &+ \left[(1-\alpha)\hat{f}(X^{(2)}) + \alpha f_0(X^{(2)}) - t\mathbb{E}\left(X^{(1)} \mid X^{(2)}\right)\right] \in \mathcal{G}\end{aligned}$$

for $4|t| = \alpha \min\left[C, C', J_m\left(\mathbb{E}\left(X^{(1)} \mid X^{(2)}\right)\right)\right]$.

Next, notice that

$$\frac{1}{n}\sum_{i=1}^n h^2(X_i) \xrightarrow{a.s.} \|h\|_2^2, \quad 0 < \|h\|_2^2 \le 4$$

by the strong law of large numbers. Choose $\delta_n = n^{-m/(2m+1)}$. Since $\|\hat{g} - g_0\|_n = \mathcal{O}_P(\delta_n)$ and $\|m_n(g)\|_{\mathcal{G}_n(R\delta_n)} = \mathcal{O}_P\left(n^{1/2}\delta_n^2\right)$, it follows that

$$\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \left(\hat{g}(X_i) - g_0(X_i)\right) = \mathcal{O}_P\left(\delta_n^2\right).$$

Finally,

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i)\left(\hat{g}(X_i) - g_0(X_i)\right) = \mathcal{O}_P\left(\delta_n\right)$$

by the Cauchy-Schwarz inequality and

$$\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i h(X_i) = \mathcal{O}_P\left(n^{-1/2}\right)$$

by the CLT. The same arguments leading to (6.4) in the proof of Lemma 6.2, may be used to prove (6.16).

Set $l(X^{(2)};\theta;f) = \theta \mathbb{E}(X^{(1)} \mid X^{(2)}) + f(X^{(2)})$, and $k(X;\theta;f) = l(X^{(2)})h(X)$. Note that

$$g\left(X^{(1)}, X^{(2)}\right) = \theta X^{(1)} + f(X^{(2)}) = \theta h(X) + l(X^{(2)};\theta;f).$$

We can express (6.16) by

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i h(X_i) &= \left(\hat{\theta} - \theta_0\right)\frac{1}{n}\sum_{i=1}^{n} h^2(X_i) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\left(k(X_i;\hat{\theta};\hat{f}) - k(X_i;\theta_0;f_0)\right) + o_P\left(n^{-1/2}\right). \end{aligned} \tag{6.17}$$

We only have to prove that the second term on the right of (6.17) can be neglected in our analysis to obtain (6.15).

To see this, note that $\mathbb{E}k(X;\theta;f) = 0$ and $\|k(\cdot;\hat{\theta};\hat{f}) - k(\cdot;\theta_0;f_0)\|_2 \xrightarrow{a.s.} 0$, since both $|\hat{\theta} - \theta_0| \xrightarrow{a.s.} 0$ and $\|\hat{f} - f_0\|_2 \xrightarrow{a.s.} 0$ by Theorem 6.2. The entropy bound on smooth classes as used in the proof of Theorem 6.2 ensures that for all $\eta > 0$ there exists $\delta > 0$ such that

$$\mathbb{P}\left\{\sup_{\|k(\cdot;\theta;f) - k(\cdot;\theta_0;f_0)\|_2 \leq \delta} \left|\int (k(\cdot;\theta;f) - k(\cdot;\theta_0;f_0)\, d(P_n - P)\right| > \frac{\eta}{\sqrt{n}}\right\} \leq \eta.$$

This completes the proof of Theorem 6.3. □

Bibliography

[1] K.S. Alexander. Probability inequalities for empirical processes and a law of the iterated logarithm. *Ann. Prob.*, 12:1041–1067, 1984.

[2] K. Ball and A. Pajor. The entropy of convex bodies with "few" extreme points. In *Proceedings of the 1989 Conf. in Banach Spaces at Strobl. Austria.* Cambridge Univ. Press, 1990.

[3] P. Billingsley. *Probability and measure.* Wiley series in probability and mathematical statistics. John Wiley & Sons, 1985.

[4] L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahr. Verw. Geb.*, 65:181–237, 1983.

[5] L. Birgé and P. Massart. Rates of convergence for minimum contrast estimators. *Probab. Th. Rel. Fields*, 97:113–150, 1993.

[6] M.S. Birman and A. Solomjak. Piecewise-polynomial approximations of functions of the classes w_p^α. *Math. Sb.*, 73:295–317, 1967.

[7] G.F. Clements. Entropies of several sets of functions. *Pacific J. Math.*, 13:1085–1097, 1963.

[8] L. Devroye. *A course in density estimation*, volume 14 of *Progress in probability and statistics.* Birkhäuser, 1987.

[9] R.M. Dudley. Central limit theorems for empirical processes. *Ann. Prob.*, 6:899–929, 1978.

[10] R.M. Dudley. *A course on empirical processes*, volume 1097 of *Lect. Notes in Math.*, pages 2–142. Springer Verlag, 1984.

[11] R.M. Dudley. *Real analysis and probabilty.* The Wadsworth & Brooks/Cole Mathematics Series. Wadsworth, 1989.

[12] L.T. Fernholz. *Von Mises Calculus for Statistical Functionals, Lecture Notes in Statistics.* Springer Verlag, 1983.

[13] R.D. Gill. Non- and semi-parametric maximum likelihood estimators and the von mises method (part 1). *Scand. J. Statist.*, 16:97–128, 1989.

[14] E. Giné and J. Zinn. Some limit theorems for empirical processes. *Ann. Prob.*, 12:929–989, 1984.

[15] E. Giné and J. Zinn. *Lectures on the Central Limit Theorem for empirical processes*, volume 1221 of *Lecture Notes in Mathematics*, pages 50–113. Springer Verlag, 1985.

[16] P. Groeneboom. Estimating a monotone density. In L.M. Le Cam and R.A. Ohlsen, editors, *Proceedings of the Berkeley conferences in honor of Neyman, J. and Kiefer, J.*, pages 539–555, 1985.

[17] P. Groeneboom. Some current developments in density estimation. Technical Report MS-R8503, CWI, Amsterdam, 1985.

[18] P. Groeneboom. Brownian motion with a parabolic drift and airy functions. *Probab. Th. Rel. Fields*, 81:79–109, 1989.

[19] P. Groeneboom and J.A. Wellner. *Information bounds and nonparametric maximum likelihood estimation*, volume 19 of *BMW Seminar.* Birkhäuser, 1992.

[20] C.C. Heesterman and R.D. Gill. A central limit theorem for m-estimators by the von mises method. *Statistica Neerlandica*, 46(2):165–177, 1992.

[21] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Amer. Stat. Ass.*, 58:13–30, 1963.

[22] I.A. Ibragimov and R.Z. Has'minskii. *Statistical estimation: asymptotic theory.* Springer Verlag, 1981.

[23] A.N. Kolmogorov and V.M. Tihomirov. ε-entropy and ε-capacity of sets in functional spaces. *Amer. Math. Soc. Transl.*, 17:277–364, 1961.

[24] J. Kuelbs. Some exponential moments of sums of independent random variables. *Amer. Math. Soc. Transl.*, 240:145–162, 1978.

[25] L. Le Cam. Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, 1:38–53, 1973.

[26] L. Le Cam and G.L. Yang. *Asymptotics in statistics: some basic concepts.* Springer Verlag, 1991.

[27] M. Ledoux and M. Talagrand. *Probability in Banach spaces.* Springer Verlag, 1991.

[28] S. Leurgans. Asymptotic distributions of slope-of-greatest-convex-minorant estimators. *Ann. Statist.*, 10:287–296, 1982.

[29] E. Mammen and S.A. van de Geer. Locally adaptive regression splines. Technical report, Humboldt Universität, 1995.

[30] E. Mammen and S.A. van de Geer. Penalized quasi-likelihood estimation in partial linear models. Technical Report TW9504, University of Leiden, 1995.

[31] A.S. Nemirovskii, B.S. Polyak, and A.B. Tsybakov. Signal processing by the nonparametric maximum-likelihood method. *Probl. Inf. Transm.*, 20:177–192, 1984.

[32] A.S. Nemirovskii, B.T. Polyak, and A.B. Tsybakov. Rate of convergence of nonparametric estimates of maximum-likelihood type. *Probl. Inf. Transm.*, 21:258–272, 1985.

[33] M. Ossiander. A central limit theorem under metric entropy with l_2 bracketing. *Ann. Prob.*, 15:897–919, 1987.

[34] V.V. Petrov. *Sums of independent random variables.* Springer Verlag, 1975.

[35] D. Pollard. *Convergence of stochastic processes.* Springer Verlag, 1984.

[36] D. Pollard. Asymptotics via empirical processes (with discussion). *Statist. Sci.*, 4:341–366, 1989.

[37] D. Pollard. *Empirical processes: Theory and applications.* NSF-CBMS Regional Conference Series in Probability and Statistics 2, 1990.

[38] W. Stute. Parameter estimation in smooth empirical processes. *Stoch. Proc. Appl.*, 22:223–244, 1986.

[39] S.A. van de Geer. A new approach to least-squares estimation, with applications. *Ann. Statist.*, 15:587–602, 1987.

[40] S.A. van de Geer. *Regression analysis and empirical processes.* PhD thesis, University of Leiden, 1988.

[41] S.A. van de Geer. Estimating a regression function. *Ann. Statist.*, 18:907–924, 1990.

[42] S.A. van de Geer. Rates of convergence for the maximum likelihood estimator in mixture models. Technical report, University of Leiden, 1994.

[43] S.A. van de Geer. Asymptotic normality in mixture models. *ESAIM, Prob. Statist.*, 1:17–33, 1995. http://www.emath.fr/.

[44] S.A. van de Geer. A maximal inequality for empirical processes. Technical Report TW9505, University of Leiden, 1995.

[45] S.A. van de Geer. The method of sieves and minimum contrast estimators. *Math. Methods of Stat.*, 4:20–38, 1995.

[46] S.A. van de Geer. Applications of empirical process theory. Technical report, University of Leiden, 1996. Lecture notes for the "Landelijke AIO-cursus stochastiek".

[47] S.A. van de Geer and M.H. Wegkamp. Consistency for the least squares estimator in nonparametric regression. Technical Report TW9404, University of Leiden, 1994. To appear in *Ann. Statist.*

[48] A.W. van der Vaart. Efficiency of infinite dimensional m-estimators. *Statistica Neerlandica*, 49(1):9–30, 1995.

[49] A.W. van der Vaart and J.A. Wellner. Weak convergence and empirical processes. Unpublished manuscript, to appear by Springer Verlag, 1994.

[50] V.N. Vapnik and Y.A. Červonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Th. Prob. Appl.*, 16:264–280, 1971.

[51] V.N. Vapnik and Y.A. Červonenkis. Necessary and sufficient conditions for the uniform convergence of means to their expectattions. *Th. Prob. Appl.*, 26:532–553, 1981.

[52] M.H. Wegkamp. Exponential bounds with applications to nonparametric regression and nonparametric density estimation. Technical Report TW9408, University of Leiden, 1994.

[53] M.H. Wegkamp. Asymptotic results for parameter estimation in general empirical processes. Technical Report TW9504, University of Leiden, 1995.

[54] M.H. Wegkamp. An l^2 approach towards nonparametric regression. Technical Report TW9514, University of Leiden, 1995.

[55] W.H. Wong and X. Shen. Convergence rate of sieve estimators. *Ann. Statist.*, 22:580–615, 1994.

[56] W.H. Wong and X. Shen. Probability inequalities for likelihood ratios and convergence rates of sieve mles. *Ann. Statist.*, 23:339–362, 1995.

[57] C.F. Wu. Asymptotic theory of nonlinear least squares estimation. *Ann. Statist.*, 9:501–513, 1981.

MATHEMATICAL CENTRE TRACTS

1 T. van der Walt. *Fixed and almost fixed points.* 1963.

2 A.R. Bloemena. *Sampling from a graph.* 1964.

3 G. de Leve. *Generalized Markovian decision processes, part I: model and method.* 1964.

4 G. de Leve. *Generalized Markovian decision processes, part II: probabilistic background.* 1964.

5 G. de Leve, H.C. Tijms, P.J. Weeda. *Generalized Markovian decision processes, applications.* 1970.

6 M.A. Maurice. *Compact ordered spaces.* 1964.

7 W.R. van Zwet. *Convex transformations of random variables.* 1964.

8 J.A. Zonneveld. *Automatic numerical integration.* 1964.

9 P.C. Baayen. *Universal morphisms.* 1964.

10 E.M. de Jager. *Applications of distributions in mathematical physics.* 1964.

11 A.B. Paalman-de Miranda. *Topological semigroups.* 1964.

12 J.A.Th.M. van Berckel, H. Brandt Corstius, R.J. Mokken, A. van Wijngaarden. *Formal properties of newspaper Dutch.* 1965.

13 H.A. Lauwerier. *Asymptotic expansions.* 1966, out of print: replaced by MCT 54.

14 H.A. Lauwerier. *Calculus of variations in mathematical physics.* 1966.

15 R. Doornbos. *Slippage tests.* 1966.

16 J.W. de Bakker. *Formal definition of programming languages with an application to the definition of ALGOL 60.* 1967.

17 R.P. van de Riet. *Formula manipulation in ALGOL 60, part 1.* 1968.

18 R.P. van de Riet. *Formula manipulation in ALGOL 60, part 2.* 1968.

19 J. van der Slot. *Some properties related to compactness.* 1968.

20 P.J. van der Houwen. *Finite difference methods for solving partial differential equations.* 1968.

21 E. Wattel. *The compactness operator in set theory and topology.* 1968.

22 T.J. Dekker. *ALGOL 60 procedures in numerical algebra, part 1.* 1968.

23 T.J. Dekker, W. Hoffmann. *ALGOL 60 procedures in numerical algebra, part 2.* 1968.

24 J.W. de Bakker. *Recursive procedures.* 1971.

25 E.R. Paërl. *Representations of the Lorentz group and projective geometry.* 1969.

26 European Meeting 1968. *Selected statistical papers, part I.* 1968.

27 European Meeting 1968. *Selected statistical papers, part II.* 1968.

28 J. Oosterhoff. *Combination of one-sided statistical tests.* 1969.

29 J. Verhoeff. *Error detecting decimal codes.* 1969.

30 H. Brandt Corstius. *Exercises in computational linguistics.* 1970.

31 W. Molenaar. *Approximations to the Poisson, binomial and hypergeometric distribution functions.* 1970.

32 L. de Haan. *On regular variation and its application to the weak convergence of sample extremes.* 1970.

33 F.W. Steutel. *Preservations of infinite divisibility under mixing and related topics.* 1970.

34 I. Juhász, A. Verbeek, N.S. Kroonenberg. *Cardinal functions in topology.* 1971.

35 M.H. van Emden. *An analysis of complexity.* 1971.

36 J. Grasman. *On the birth of boundary layers.* 1971.

37 J.W. de Bakker, G.A. Blaauw, A.J.W. Duijvestijn, E.W. Dijkstra, P.J. van der Houwen, G.A.M. Kamsteeg-Kemper, F.E.J. Kruseman Aretz, W.L. van der Poel, J.P. Schaap-Kruseman, M.V. Wilkes, G. Zoutendijk. *MC-25 Informatica Symposium.* 1971.

38 W.A. Verloren van Themaat. *Automatic analysis of Dutch compound words.* 1972.

39 H. Bavinck. *Jacobi series and approximation.* 1972.

40 H.C. Tijms. *Analysis of (s,S) inventory models.* 1972.

41 A. Verbeek. *Superextensions of topological spaces.* 1972.

42 W. Vervaat. *Success epochs in Bernoulli trials (with applications in number theory).* 1972.

43 F.H. Ruymgaart. *Asymptotic theory of rank tests for independence.* 1973.

44 H. Bart. *Meromorphic operator valued functions.* 1973.

45 A.A. Balkema. *Monotone transformations and limit laws.* 1973.

46 R.P. van de Riet. *ABC ALGOL, a portable language for formula manipulation systems, part 1: the language.* 1973.

47 R.P. van de Riet. *ABC ALGOL, a portable language for formula manipulation systems, part 2: the compiler.* 1973.

48 F.E.J. Kruseman Aretz, P.J.W. ten Hagen, H.L. Oudshoorn. *An ALGOL 60 compiler in ALGOL 60, text of the MC-compiler for the EL-X8.* 1973.

49 H. Kok. *Connected orderable spaces.* 1974.

50 A. van Wijngaarden, B.J. Mailloux, J.E.L. Peck, C.H.A. Koster, M. Sintzoff, C.H. Lindsey, L.G.L.T. Meertens, R.G. Fisker (eds.). *Revised report on the algorithmic language ALGOL 68.* 1976.

51 A. Hordijk. *Dynamic programming and Markov potential theory.* 1974.

52 P.C. Baayen (ed.). *Topological structures.* 1974.

53 M.J. Faber. *Metrizability in generalized ordered spaces.* 1974.

54 H.A. Lauwerier. *Asymptotic analysis, part 1.* 1974.

55 M. Hall, Jr., J.H. van Lint (eds.). *Combinatorics, part 1: theory of designs, finite geometry and coding theory.* 1974.

56 M. Hall, Jr., J.H. van Lint (eds.). *Combinatorics, part 2: graph theory, foundations, partitions and combinatorial geometry.* 1974.

57 M. Hall, Jr., J.H. van Lint (eds.). *Combinatorics, part 3: combinatorial group theory.* 1974.

58 W. Albers. *Asymptotic expansions and the deficiency concept in statistics.* 1975.

59 J.L. Mijnheer. *Sample path properties of stable processes.* 1975.

60 F. Göbel. *Queueing models involving buffers.* 1975.

63 J.W. de Bakker (ed.). *Foundations of computer science.* 1975.

64 W.J. de Schipper. *Symmetric closed categories.* 1975.

65 J. de Vries. *Topological transformation groups, 1: a categorical approach.* 1975.

66 H.G.J. Pijls. *Logically convex algebras in spectral theory and eigenfunction expansions.* 1976.

68 P.P.N. de Groen. *Singularly perturbed differential operators of second order.* 1976.

69 J.K. Lenstra. *Sequencing by enumerative methods.* 1977.

70 W.P. de Roever, Jr. *Recursive program schemes: semantics and proof theory.* 1976.

71 J.A.E.E. van Nunen. *Contracting Markov decision processes.* 1976.

72 J.K.M. Jansen. *Simple periodic and non-periodic Lamé functions and their applications in the theory of conical waveguides.* 1977.

73 D.M.R. Leivant. *Absoluteness of intuitionistic logic.* 1979.

74 H.J.J. te Riele. *A theoretical and computational study of generalized aliquot sequences.* 1976.

75 A.E. Brouwer. *Treelike spaces and related connected topological spaces.* 1977.

76 M. Rem. *Associons and the closure statements.* 1976.

77 W.C.M. Kallenberg. *Asymptotic optimality of likelihood ratio tests in exponential families.* 1978.

78 E. de Jonge, A.C.M. van Rooij. *Introduction to Riesz spaces.* 1977.

79 M.C.A. van Zuijlen. *Empirical distributions and rank statistics.* 1977.

80 P.W. Hemker. *A numerical study of stiff two-point boundary problems.* 1977.

81 K.R. Apt, J.W. de Bakker (eds.). *Foundations of computer science II, part 1.* 1976.

82 K.R. Apt, J.W. de Bakker (eds.). *Foundations of computer science II, part 2.* 1976.

83 L.S. van Benthem Jutting. *Checking Landau's "Grundlagen" in the AUTOMATH system.* 1979.

84 H.L.L. Busard. *The translation of the elements of Euclid from the Arabic into Latin by Hermann of Carinthia (?), books vii-xii.* 1977.

85 J. van Mill. *Supercompactness and Wallmann spaces.* 1977.

86 S.G. van der Meulen, M. Veldhorst. *Torrix I, a programming system for operations on vectors and matrices over arbitrary fields and of variable size.* 1978.

88 A. Schrijver. *Matroids and linking systems.* 1977.

89 J.W. de Roever. *Complex Fourier transformation and analytic functionals with unbounded carriers.* 1978.

90 L.P.J. Groenewegen. *Characterization of optimal strategies in dynamic games.* 1981.

91 J.M. Geysel. *Transcendence in fields of positive characteristic.* 1979.

92 P.J. Weeda. *Finite generalized Markov programming.* 1979.

93 H.C. Tijms, J. Wessels (eds.). *Markov decision theory.* 1977.

94 A. Bijlsma. *Simultaneous approximations in transcendental number theory.* 1978.

95 K.M. van Hee. *Bayesian control of Markov chains.* 1978.

96 P.M.B. Vitányi. *Lindenmayer systems: structure, languages, and growth functions.* 1980.

97 A. Federgruen. *Markovian control problems; functional equations and algorithms.* 1984.

98 R. Geel. *Singular perturbations of hyperbolic type.* 1978.

99 J.K. Lenstra, A.H.G. Rinnooy Kan, P. van Emde Boas (eds.). *Interfaces between computer science and operations research.* 1978.

100 P.C. Baayen, D. van Dulst, J. Oosterhoff (eds.). *Proceedings bicentennial congress of the Wiskundig Genootschap, part 1.* 1979.

101 P.C. Baayen, D. van Dulst, J. Oosterhoff (eds.). *Proceedings bicentennial congress of the Wiskundig Genootschap, part 2.* 1979.

102 D. van Dulst. *Reflexive and superreflexive Banach spaces.* 1978.

103 K. van Harn. *Classifying infinitely divisible distributions by functional equations.* 1978.

104 J.M. van Wouwe. *GO-spaces and generalizations of metrizability.* 1979.

105 R. Helmers. *Edgeworth expansions for linear combinations of order statistics.* 1982.

106 A. Schrijver (ed.). *Packing and covering in combinatorics.* 1979.

107 C. den Heijer. *The numerical solution of nonlinear operator equations by imbedding methods.* 1979.

108 J.W. de Bakker, J. van Leeuwen (eds.). *Foundations of computer science III, part 1.* 1979.

109 J.W. de Bakker, J. van Leeuwen (eds.). *Foundations of computer science III, part 2.* 1979.

110 J.C. van Vliet. *ALGOL 68 transput, part I: historical review and discussion of the implementation model.* 1979.

111 J.C. van Vliet. *ALGOL 68 transput, part II: an implementation model.* 1979.

112 H.C.P. Berbee. *Random walks with stationary increments and renewai theory.* 1979.

113 T.A.B. Snijders. *Asymptotic optimality theory for testing problems with restricted alternatives.* 1979.

114 A.J.E.M. Janssen. *Application of the Wigner distribution to harmonic analysis of generalized stochastic processes.* 1979.

115 P.C. Baayen, J. van Mill (eds.). *Topological structures II, part 1.* 1979.

116 P.C. Baayen, J. van Mill (eds.). *Topological structures II, part 2.* 1979.

117 P.J.M. Kallenberg. *Branching processes with continuous state space.* 1979.

118 P. Groeneboom. *Large deviations and asymptotic efficiencies.* 1980.

119 F.J. Peters. *Sparse matrices and substructures, with a novel implementation of finite element algorithms.* 1980.

120 W.P.M. de Ruyter. *On the asymptotic analysis of large-scale ocean circulation.* 1980.

121 W.H. Haemers. *Eigenvalue techniques in design and graph theory.* 1980.

122 J.C.P. Bus. *Numerical solution of systems of nonlinear equations.* 1980.

123 I. Yuhász. *Cardinal functions in topology - ten years later.* 1980.

124 R.D. Gill. *Censoring and stochastic integrals.* 1980.

125 R. Eising. *2-D systems, an algebraic approach.* 1980.

126 G. van der Hoek. *Reduction methods in nonlinear programming.* 1980.

127 J.W. Klop. *Combinatory reduction systems.* 1980.

128 A.J.J. Talman. *Variable dimension fixed point algorithms and triangulations.* 1980.

129 G. van der Laan. *Simplicial fixed point algorithms.* 1980.

130 P.J.W. ten Hagen, T. Hagen, P. Klint, H. Noot, H.J. Sint, A.H. Veen. *ILP: intermediate language for pictures.* 1980.

131 R.J.R. Back. *Correctness preserving program refinements: proof theory and applications.* 1980.

132 H.M. Mulder. *The interval function of a graph.* 1980.

133 C.A.J. Klaassen. *Statistical performance of location estimators.* 1981.

134 J.C. van Vliet, H. Wupper (eds.). *Proceedings international conference on ALGOL 68.* 1981.

135 J.A.G. Groenendijk, T.M.V. Janssen, M.J.B. Stokhof (eds.). *Formal methods in the study of language, part I.* 1981.

136 J.A.G. Groenendijk, T.M.V. Janssen, M.J.B. Stokhof (eds.). *Formal methods in the study of language, part II.* 1981.

137 J. Telgen. *Redundancy and linear programs.* 1981.

138 H.A. Lauwerier. *Mathematical models of epidemics.* 1981.

139 J. van der Wal. *Stochastic dynamic programming, successive approximations and nearly optimal strategies for Markov decision processes and Markov games.* 1981.

140 J.H. van Geldrop. *A mathematical theory of pure exchange economies without the no-critical-point hypothesis.* 1981.

141 G.E. Welters. *Abel-Jacobi isogenies for certain types of Fano threefolds.* 1981.

142 H.R. Bennett, D.J. Lutzer (eds.). *Topology and order structures, part 1.* 1981.

143 J.M. Schumacher. *Dynamic feedback in finite- and infinite-dimensional linear systems.* 1981.

144 P. Eijgenraam. *The solution of initial value problems using interval arithmetic; formulation and analysis of an algorithm.* 1981.

145 A.J. Brentjes. *Multi-dimensional continued fraction algorithms.* 1981.

146 C.V.M. van der Mee. *Semigroup and factorization methods in transport theory.* 1981.

147 H.H. Tigelaar. *Identification and informative sample size.* 1982.

148 L.C.M. Kallenberg. *Linear programming and finite Markovian control problems.* 1983.

149 C.B. Huijsmans, M.A. Kaashoek, W.A.J. Luxemburg, W.K. Vietsch (eds.). *From A to Z, proceedings of a symposium in honour of A.C. Zaanen.* 1982.

150 M. Veldhorst. *An analysis of sparse matrix storage schemes.* 1982.

151 R.J.M.M. Does. *Higher order asymptotics for simple linear rank statistics.* 1982.

152 G.F. van der Hoeven. *Projections of lawless sequencies.* 1982.

153 J.P.C. Blanc. *Application of the theory of boundary value problems in the analysis of a queueing model with paired services.* 1982.

154 H.W. Lenstra, Jr., R. Tijdeman (eds.). *Computational methods in number theory, part I.* 1982.

155 H.W. Lenstra, Jr., R. Tijdeman (eds.). *Computational methods in number theory, part II.* 1982.

156 P.M.G. Apers. *Query processing and data allocation in distributed database systems.* 1983.

157 H.A.W.M. Kneppers. *The covariant classification of two-dimensional smooth commutative formal groups over an algebraically closed field of positive characteristic.* 1983.

158 J.W. de Bakker, J. van Leeuwen (eds.). *Foundations of computer science IV, distributed systems, part 1.* 1983.

159 J.W. de Bakker, J. van Leeuwen (eds.). *Foundations of computer science IV, distributed systems, part 2.* 1983.

160 A. Rezus. *Abstract AUTOMATH.* 1983.

161 G.F. Helminck. *Eisenstein series on the metaplectic group, an algebraic approach.* 1983.

162 J.J. Dik. *Tests for preference.* 1983.

163 H. Schippers. *Multiple grid methods for equations of the second kind with applications in fluid mechanics.* 1983.

164 F.A. van der Duyn Schouten. *Markov decision processes with continuous time parameter.* 1983.

165 P.C.T. van der Hoeven. *On point processes.* 1983.

166 H.B.M. Jonkers. *Abstraction, specification and implementation techniques, with an application to garbage collection.* 1983.

167 W.H.M. Zijm. *Nonnegative matrices in dynamic programming.* 1983.

168 J.H. Evertse. *Upper bounds for the numbers of solutions of diophantine equations.* 1983.

169 H.R. Bennett, D.J. Lutzer (eds.). *Topology and order structures, part 2.* 1983.

CWI TRACTS

1 D.H.J. Epema. *Surfaces with canonical hyperplane sections.* 1984.
2 J.J. Dijkstra. *Fake topological Hilbert spaces and characterizations of dimension in terms of negligibility.* 1984.
3 A.J. van der Schaft. *System theoretic descriptions of physical systems.* 1984.
4 J. Koene. *Minimal cost flow in processing networks, a primal approach.* 1984.
5 B. Hoogenboom. *Intertwining functions on compact Lie groups.* 1984.
6 A.P.W. Böhm. *Dataflow computation.* 1984.
7 A. Blokhuis. *Few-distance sets.* 1984.
8 M.H. van Hoorn. *Algorithms and approximations for queueing systems.* 1984.
9 C.P.J. Koymans. *Models of the lambda calculus.* 1984.
10 C.G. van der Laan, N.M. Temme. *Calculation of special functions: the gamma function, the exponential integrals and error-like functions.* 1984.
11 N.M. van Dijk. *Controlled Markov processes; time-discretization.* 1984.
12 W.H. Hundsdorfer. *The numerical solution of nonlinear stiff initial value problems: an analysis of one step methods.* 1985.
13 D. Grune. *On the design of ALEPH.* 1985.
14 J.G.F. Thiemann. *Analytic spaces and dynamic programming: a measure theoretic approach.* 1985.
15 F.J. van der Linden. *Euclidean rings with two infinite primes.* 1985.
16 R.J.P. Groothuizen. *Mixed elliptic-hyperbolic partial differential operators: a case-study in Fourier integral operators.* 1985.
17 H.M.M. ten Eikelder. *Symmetries for dynamical and Hamiltonian systems.* 1985.
18 A.D.M. Kester. *Some large deviation results in statistics.* 1985.
19 T.M.V. Janssen. *Foundations and applications of Montague grammar, part 1: Philosophy, framework, computer science.* 1986.
20 B.F. Schriever. *Order dependence.* 1986.
21 D.P. van der Vecht. *Inequalities for stopped Brownian motion.* 1986.
22 J.C.S.P. van der Woude. *Topological dynamix.* 1986.
23 A.F. Monna. *Methods, concepts and ideas in mathematics: aspects of an evolution.* 1986.
24 J.C.M. Baeten. *Filters and ultrafilters over definable subsets of admissible ordinals.* 1986.
25 A.W.J. Kolen. *Tree network and planar rectilinear location theory.* 1986.
26 A.H. Veen. *The misconstrued semicolon: Reconciling imperative languages and dataflow machines.* 1986.
27 A.J.M. van Engelen. *Homogeneous zero-dimensional absolute Borel sets.* 1986.
28 T.M.V. Janssen. *Foundations and applications of Montague grammar, part 2: Applications to natural language.* 1986.
29 H.L. Trentelman. *Almost invariant subspaces and high gain feedback.* 1986.
30 A.G. de Kok. *Production-inventory control models: approximations and algorithms.* 1987.
31 E.E.M. van Berkum. *Optimal paired comparison designs for factorial experiments.* 1987.
32 J.H.J. Einmahl. *Multivariate empirical processes.* 1987.
33 O.J. Vrieze. *Stochastic games with finite state and action spaces.* 1987.
34 P.H M. Kersten. *Infinitesimal symmetries: a computational approach.* 1987.
35 M.L. Eaton. *Lectures on topics in probability inequalities.* 1987.
36 A.H.P. van der Burgh, R.M.M. Mattheij (eds.). *Proceedings of the first international conference on industrial and applied mathematics (ICIAM 87).* 1987.
37 L. Stougie. *Design and analysis of algorithms for stochastic integer programming.* 1987.
38 J.B.G. Frenk. *On Banach algebras, renewal measures and regenerative processes.* 1987.
39 H.J.M. Peters, O.J. Vrieze (eds.). *Surveys in game theory and related topics.* 1987.
40 J.L. Geluk, L. de Haan. *Regular variation, extensions and Tauberian theorems.* 1987.
41 Sape J. Mullender (ed.). *The Amoeba distributed operating system: Selected papers 1984-1987.* 1987.
42 P.R.J. Asveld, A. Nijholt (eds.). *Essays on concepts, formalisms, and tools.* 1987.
43 H.L. Bodlaender. *Distributed computing: structure and complexity.* 1987.
44 A.W. van der Vaart. *Statistical estimation in large parameter spaces.* 1988.
45 S.A. van de Geer. *Regression analysis and empirical processes* 1988.
46 S.P. Spekreijse. *Multigrid solution of the steady Euler equations.* 1988.
47 J.B. Dijkstra. *Analysis of means in some non-standard situations.* 1988.
48 F.C. Drost. *Asymptotics for generalized chi-square goodness-of-fit tests.* 1988.
49 F.W. Wubs. *Numerical solution of the shallow-water equations.* 1988.
50 F. de Kerf. *Asymptotic analysis of a class of perturbed Korteweg-de Vries initial value problems.* 1988.
51 P.J.M. van Laarhoven. *Theoretical and computational aspects of simulated annealing.* 1988.
52 P.M. van Loon. *Continuous decoupling transformations for linear boundary value problems.* 1988.
53 K.C.P. Machielsen. *Numerical solution of optimal control problems with state constraints by sequential quadratic programming in function space.* 1988.
54 L.C.R.J. Willenborg. *Computational aspects of survey data processing.* 1988.
55 G.J. van der Steen. *A program generator for recognition, parsing and transduction with syntactic patterns.* 1988.
56 J.C. Ebergen. *Translating programs into delay-insensitive circuits.* 1989.
57 S.M. Verduyn Lunel. *Exponential type calculus for linear delay equations.* 1989.
58 M.C.M. de Gunst. *A random model for plant cell population growth.* 1989.
59 D. van Dulst. *Characterizations of Banach spaces not containing l^1.* 1989.
60 H.E. de Swart. *Vacillation and predictability properties of low-order atmospheric spectral models.* 1989.

61 P. de Jong. *Central limit theorems for generalized multilinear forms.* 1989.
62 V.J. de Jong. *A specification system for statistical software.* 1989.
63 B. Hanzon. *Identifiability, recursive identification and spaces of linear dynamical systems, part I.* 1989.
64 B. Hanzon. *Identifiability, recursive identification and spaces of linear dynamical systems, part II.* 1989.
65 B.M.M. de Weger. *Algorithms for diophantine equations.* 1989.
66 A. Jung. *Cartesian closed categories of domains.* 1989.
67 J.W. Polderman. *Adaptive control & identification: Conflict or conflux?.* 1989.
68 H.J. Woerdeman. *Matrix and operator extensions.* 1989.
69 B.G. Hansen. *Monotonicity properties of infinitely divisible distributions.* 1989.
70 J.K. Lenstra, H.C. Tijms, A. Volgenant (eds.). *Twenty-five years of operations research in the Netherlands: Papers dedicated to Gijs de Leve.* 1990.
71 P.J.C. Spreij. *Counting process systems. Identification and stochastic realization.* 1990.
72 J.F. Kaashoek. *Modeling one dimensional pattern formation by anti-diffusion.* 1990.
73 A.M.H. Gerards. *Graphs and polyhedra. Binary spaces and cutting planes.* 1990.
74 B. Koren. *Multigrid and defect correction for the steady Navier-Stokes equations. Application to aerodynamics.* 1991.
75 M.W.P. Savelsbergh. *Computer aided routing.* 1992.
76 O.E. Flippo. *Stability, duality and decomposition in general mathematical programming.* 1991.
77 A.J. van Es. *Aspects of nonparametric density estimation.* 1991.
78 G.A.P. Kindervater. *Exercises in parallel combinatorial computing.* 1992.
79 J.J. Lodder. *Towards a symmetrical theory of generalized functions.* 1991.
80 S.A. Smulders. *Control of freeway traffic flow.* 1996.
81 P.H.M. America, J.J.M.M. Rutten. *A parallel object-oriented language: design and semantic foundations.* 1992.
82 F. Thuijsman. *Optimality and equilibria in stochastic games.* 1992.
83 R.J. Kooman. *Convergence properties of recurrence sequences.* 1992.
84 A.M. Cohen (ed.). *Computational aspects of Lie group representations and related topics. Proceedings of the 1990 Computational Algebra Seminar at CWI, Amsterdam.* 1991.
85 V. de Valk. *One-dependent processes.* 1994.
86 J.A. Baars, J.A.M. de Groot. *On topological and linear equivalence of certain function spaces.* 1992.
87 A.F. Monna. *The way of mathematics and mathematicians.* 1992.
88 E.D. de Goede. *Numerical methods for the three-dimensional shallow water equations.* 1993.
89 M. Zwaan. *Moment problems in Hilbert space with applications to magnetic resonance imaging.* 1993.
90 C. Vuik. *The solution of a one-dimensional Stefan problem.* 1993.
91 E.R. Verheul. *Multimedians in metric and normed spaces.* 1993.
92 J.L.M. Maubach. *Iterative methods for non-linear partial differential equations.* 1994.
93 A.W. Ambergen. *Statistical uncertainties in posterior probabilities.* 1993.
94 P.A. Zegeling. *Moving-grid methods for time-dependent partial differential equations.* 1993.
95 M.J.C. van Pul. *Statistical analysis of software reliability models.* 1993.
96 J.K. Scholma. *A Lie algebraic study of some integrable systems associated with root systems.* 1993.
97 J.L. van den Berg. *Sojourn times in feedback and processor sharing queues.* 1993.
98 A.J. Koning. *Stochastic integrals and goodness-of-fit tests.* 1993.
99 B.P. Sommeijer. *Parallelism in the numerical integration of initial value problems.* 1993.
100 J. Molenaar. *Multigrid methods for semiconductor device simulation.* 1993.
101 H.J.C. Huijberts. *Dynamic feedback in nonlinear synthesis problems.* 1994.
102 J.A.M. van der Weide. *Stochastic processes and point processes of excursions.* 1994.
103 P.W. Hemker, P. Wesseling (eds.). *Contributions to multigrid.* 1994.
104 I.J.B.F. Adan. *A compensation approach for queueing problems.* 1994.
105 O.J. Boxma, G.M. Koole (eds.). *Performance evaluation of parallel and distributed systems - solution methods. Part 1.* 1994.
106 O.J. Boxma, G.M. Koole (eds.). *Performance evaluation of parallel and distributed systems - solution methods. Part 2.* 1994.
107 R.A. Trompert. *Local uniform grid refinement for time-dependent partial differential equations.* 1995.
108 M.N.M. van Lieshout. *Stochastic geometry models in image analysis and spatial statistics.* 1995.
109 R.J. van Glabbeek. *Comparative concurrency semantics and refinement of actions.* 1996.
110 W. Vervaat, H. Holwerda (ed.). *Probability and lattices.* 1997.
111 I. Helsloot. *Covariant formal group theory and some applications.* 1995.
112 R.N. Bol. *Loop checking in logic programming.* 1995.
113 G.J.M. Koole. *Stochastic scheduling and dynamic programming.* 1995.
114 M.J. van der Laan. *Efficient and inefficient estimation in semiparametric models.* 1995.
115 S.C. Borst. *Polling models.* 1996.
116 G.D. Otten. *Statistical test limits in quality control.* 1996.
117 K.G. Langendoen. *Graph reduction on shared-memory multiprocessors.* 1996.
118 W.C.A. Maas. *Nonlinear $\mathcal{H}_\infty$ control: the singular case.* 1996.
119 A. Di Bucchianico. *Probabilistic and analytical aspects of the umbral calculus.* 1997.
120 M. van Loon. *Numerical methods in smog prediction.* 1997.
121 B.J. Wijers. *Nonparametric estimation for a windowed line-segment process.* 1997.
122 W.K. Klein Haneveld, O.J. Vrieze, L.C.M. Kallenberg (editors). *Ten years LNMB – Ph.D. research and graduate courses of the Dutch Network of Operations Research.* 1997.
123 R.W. van den Hofstad. *One-dimensional random polymers.* 1998.
124 W.J.H. Stortelder. *Parameter estimation in nonlinear dynamical systems.* 1998.
125 M.H. Wegkamp. *Entropy methods in statistical estimation.* 1998.