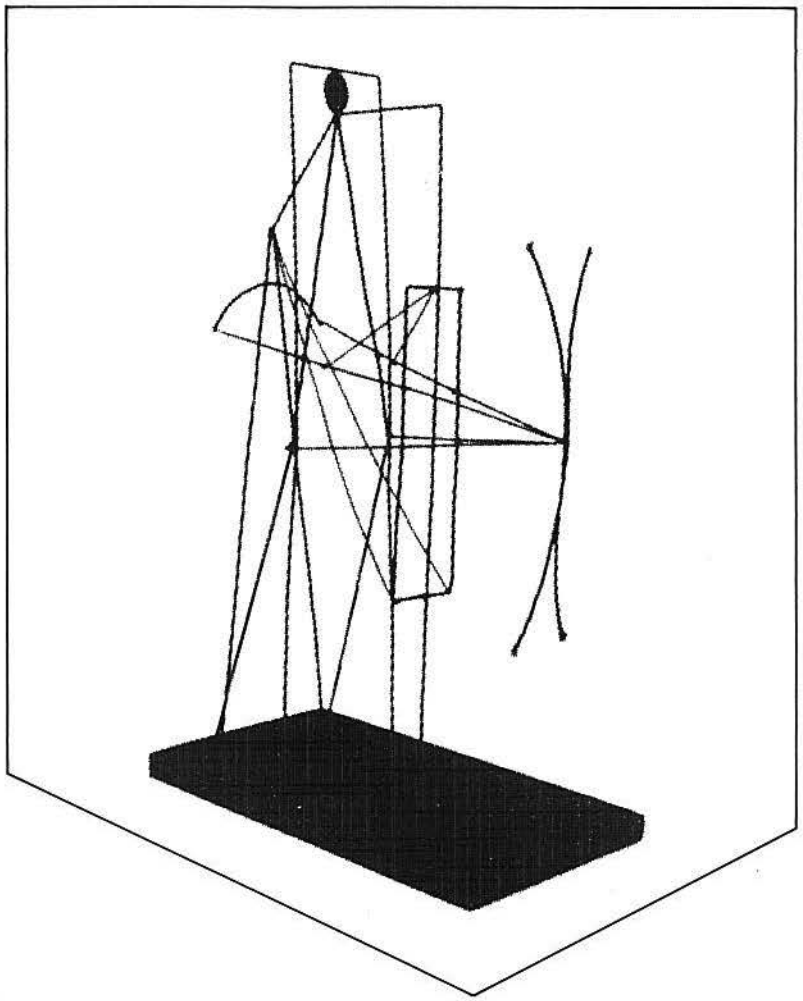
H

Regression Analysis and Empirical Processes



Sara van de Geer

REGRESSION ANALYSIS
AND
EMPIRICAL PROCESSES

PROEFSCHRIFT

ter verkrijging van de graad van Doctor in de Wiskunde en Natuurwetenschappen aan de Rijksuniversiteit te Leiden op gezag van de Rector Magnificus Dr. J.J.M. Beenakker, hoogleraar in de faculteit der Wiskunde en Natuurwetenschappen, volgens besluit van het college van dekanen

te verdedigen op
woensdag 30 september 1987 te klokke 14.15 uur

door

SARA ANNA VAN DE GEER

geboren te Leiden in 1958

SAMENSTELLING PROMOTIECOMMISSIE

Promotoren	prof.dr. W.R. van Zwet
	prof.dr. R.D. Gill
Referent	prof.dr. J. Fabius
Overige leden	prof.dr. J. Oosterhoff
	prof.dr. F.H. Ruymgaart
	prof.dr. G. van Dijk

ACKNOWLEDGEMENTS

I am grateful to the staff of the Centre for Mathematics and Computer Science. They provided me a very stimulating working environment.

I thank Wilbert Kallenberg for his suggestions that led to the proof of Theorem 7.2.6, Michel Voors for carrying out the computations that are reported in Chapter 8 and Jossi Kustina for her excellent typing.

То Валери

CONTENTS

1. Introduction

1.1 Goal and itinerary of this study

The problem we investigate is least squares estimation of a regression function. We have n observations $(\mathbf{x}_k, \mathbf{y}_k)$, $k = 1, \cdots, n$, which are assumed to satisfy

$$\mathbf{y}_k = g(\mathbf{x}_k) + \boldsymbol{\epsilon}_k \ , \ k = 1, \cdots, n,$$

where the disturbances $\boldsymbol{\epsilon}_k$ are independent and all have expectation zero and finite variance, and where the $\mathbf{x}_k$ are vectors in some Euclidean space. The function $g(\cdot)$ is in part unknown. The least squares method for estimating g is: find a $\hat{\mathbf{g}}_n$ such that

$$\frac{1}{n} \sum_{k=1}^{n} (\mathbf{y}_k - g(\mathbf{x}_k))^2$$

is minimized, where the minimization is over the class $\mathcal{G}$ of the regression functions that one considers feasible. The properties of the least squares estimator $\hat{\mathbf{g}}_n$ depend on the extent to which extent g is unknown, i.e. on $\mathcal{G}$. If it is known that the regression is linear, then $\mathcal{G} = \{g(x) = x\theta : \theta \in \Theta\}$ is the class of linear functions and we are in a classical situation. Linear regression has been studied extensively. More recent work in this field focuses e.g. on necessary conditions for consistency (Lai, Robbins and Wei (1978)).

Linear regression is a special case of the situation where g is known up to a finite-dimensional parameter. This more general case is called nonlinear regression. The class $\mathcal{G}$ is $\mathcal{G} = \{g = g(\cdot, \theta) : \theta \in \Theta\}$, with $\Theta \subset \mathbb{R}^r$. Because of the possible nonlinearity, the approach to the study of the least squares estimator is mostly asymptotic. Hartley and Baker (1965) prove asymptotic normality under the assumption of normally distributed errors. As in Jennrich (1969), we shall not specify the distribution of the $\boldsymbol{\epsilon}_k$. Jennrich obtains consistency and asymptotic normality under regularity conditions on the $g(\cdot, \theta)$. Later, these conditions have been refined (Wu (1981)). However, there still remain nonlinear models that have only been investigated on an ad hoc basis. As an example we present a two-phase regression model in its simplest form.

Example 1.1.

$$\mathbf{y}_k = \begin{cases} \alpha^{(1)} + \boldsymbol{\epsilon}_k \ , \text{ if } \mathbf{x}_k \leqslant \gamma \\ \alpha^{(2)} + \boldsymbol{\epsilon}_k \ , \text{ if } \mathbf{x}_k > \gamma \end{cases} .$$

Both the $\alpha^{(i)}, i = 1,2$, and γ are unknown parameters. The class $\mathcal{G}$ is

$$\mathcal{G} = \{g = \alpha^{(1)} 1_{(-\infty, \gamma]} + \alpha^{(2)} 1_{(\gamma, \infty)} : \alpha^{(1)}, \alpha^{(2)}, \gamma \in \mathbb{R}\}.$$

Nonlinear regression , in turn, is a special case of a even more general class of models which includes non- and semiparametric regression. In the latter cases the regression functions can no longer be indexed by a finite-dimensional

parameter.

EXAMPLE 1.2.

$$\mathbf{y}_k = g(\mathbf{x}_k)+\epsilon_k,$$
$$g\in\mathcal{G} = \{g:\mathbb{R}\to\mathbb{R},\ g \text{ has } m \text{ derivatives},\ \int |g^{(m)}|^2\leqslant K\},$$

with K a known constant.

Another example of nonparametric regression is e.g. the situation where only monotonicity of the regression function is assumed.

We shall take a unified approach in investigating the asymptotic properties of the least squares estimator. We regard the function g itself as unknown parameter and we shall study how well g can be estimated by the least squares method, given that g is a member of a class $\mathcal{G}$ of regression functions. It is to be expected that the asymptotic behaviour of $\hat{\mathbf{g}}_n$ is primarily determined by the properties of $\mathcal{G}$, the parameter space. In particular, the larger or richer $\mathcal{G}$ is, the harder it will be to estimate g. Using concepts of *empirical process theory*, we shall give a precise description of the link between the 'size' of $\mathcal{G}$ and the behaviour of $\hat{\mathbf{g}}_n$. Empirical process theory is the theory of uniform laws of large numbers and uniform central limit theorems. Its topics are limit theorems for processes indexed by sets or functions. For instance, let $\mathbf{H}_n$ be the empirical distribution based on n independent observations $\mathbf{x}_k$ from H. $\mathbf{H}_n$ puts mass $1/n$ on each of the $\mathbf{x}_k$, $k=1,\cdots,n$. The theory supplies us with sufficient and - modulo measurability - also necessary conditions such that for a class $\mathcal{G}$ of H-square integrable functions g

$$\sup_{g\in\mathcal{G}} \left|\int |g|^2 d(\mathbf{H}_n - H)\right| \to 0 \text{ almost surely,} \tag{1.1}$$

as n tends to infinity (VAPNIK and CHERVONENKIS (1971,1981), POLLARD (1984), DUDLEY (1984)). A result like (1.1) is very helpful for proving consistency of $\hat{\mathbf{g}}_n$.

We shall now present one more two-phase regression model. This model drew our attention to empirical processes indexed by sets because it has sets as unknown parameters.

EXAMPLE 1.3.

$$\mathbf{y}_k = \min(\alpha^{(1)}+\mathbf{x}_{k,1}\beta_1^{(1)}+\mathbf{x}_{k,2}\beta_2^{(1)},\ \alpha^{(2)}+\mathbf{x}_{k,1}\beta_1^{(2)}+\mathbf{x}_{k,2}\beta_2^{(2)})+\epsilon_k. \tag{1.2}$$

Here, the measurements $\mathbf{y}_k$, $k=1,\cdots,n$ are the log-lifetimes of plastic pipes for the transportation of fluids. The $\mathbf{x}_k=(\mathbf{x}_{k,1},\mathbf{x}_{k,2})$ are (stress)/(absolute temperature) and (absolute temperature)$^{-1}$. The idea is that at high stress and temperature the pipes become brittle and break due to a mechanism different from the one at low stress and temperature.

Related to (1.2) is the model

$$\mathbf{y}_k = \begin{cases} \alpha^{(1)}+\mathbf{x}_{k,1}\beta_1^{(1)}+\mathbf{x}_{k,2}\beta_2^{(1)}+\epsilon_k & \text{if } \mathbf{x}_k\in A \\ \alpha^{(2)}+\mathbf{x}_{k,1}\beta_1^{(2)}+\mathbf{x}_{k,2}\beta_2^{(2)}+\epsilon_k & \text{if } \mathbf{x}_k\notin A \end{cases},$$

where $A=\{\mathbf{x}_k\colon \mathbf{x}_{k,1}\gamma_1+\mathbf{x}_{k,2}\gamma_2\leqslant 1\}$. The class of regression functions is now

$$\begin{aligned}\mathcal{G} = \{&g(x_1,x_2)=(\alpha^{(1)}+x_1\beta_1^{(1)}+x_2\beta_2^{(1)})1\!\!1_A(x_1,x_2)\\ &+(\alpha^{(2)}+x_1\beta_1^{(2)}+x_2\beta_2^{(2)})1\!\!1_{A^c}(x_1,x_2)\colon\\ &(\alpha^{(i)},\beta_1^{(i)},\beta_2^{(i)})^T\in\mathbb{R}^3,\ i=1,2,\ A\in\mathcal{A}\},\end{aligned} \tag{1.3}$$

with $\mathcal{A}$ the collection of halfspaces in $\mathbb{R}^2$. The only difference between this model and (1.2) is that in the latter one imposes the restriction $\gamma_t=(\beta_t^{(1)}-\beta_t^{(2)})/(\alpha^{(2)}-\alpha^{(1)})$, $t=1,2$. In both models, the halfspace A is an unknown parameter. In (1.2) the the set A is a function of the other unknown parameters $\alpha^{(i)},\beta^{(i)}$ and in (1.3) it is a function of the Euclidean parameter γ. However, in the general two-phase regression model, the class $\mathcal{A}$ in (1.3) need not be indexed by a finite-dimensional parameter. An example is the case where $\mathcal{A}$ is the collection of all monotone sets, i.e. the class of sets A such that if $(x_1,x_2)\in A$ also $(\tilde{x}_1,\tilde{x}_2)\in A$ for all $(\tilde{x}_1,\tilde{x}_2)$ with $\tilde{x}_1\leqslant x_1$ and $\tilde{x}_2\leqslant x_2$.

We shall take two-phase regression models of the form presented in Example 1.3 as the major illustration of the theory we develop for general regression models. In this way, we hope to provide some insight into the significance of our results. Examples concerning other (nonparametric) models occur throughout the manuscript and are sometimes not explored in full detail.

The presentation is organized as follows. Chapter 2 sets the background for proving consistency. We give an overview of the history that led to the uniform law of large numbers (1.1), which goes from sets via bounded functions to integrable functions. We extend the uniform law of large numbers to the case of non-identically distributed variables and allow virtually everything to depend on the number of observations (i.e. on the n-th experiment). With these tools, we prove in Chapter 3 a general consistency theorem, followed by some applications to nonlinear and nonparametric regression. We must stress however that the general theorem should be regarded rather as expressing a general viewpoint on regression than as a recipe for checking consistency. One of its conditions often does not hold for the original $\mathcal{G}$, but only for a subclass of $\mathcal{G}$, c.f. the assumption in parametric maximum likelihood that the parameter space is compact. In specific situations one faces the problem of proving that eventually $\hat{\mathbf{g}}_n$ lies in this subset, which can be just as difficult as showing consistency directly. We elaborate on this in Section 3.4, where we apply the general theorem to the models of Example 1.3.

Chapter 4 summarizes some results from the literature on uniform central limit theorems. We use these in Chapter 5 to prove asymptotic normality of the least squares estimator of the $\alpha^{(i)}$ and $\beta^{(i)}$ of Example 1.3. In Chapter 6 we return to the more general case. We exploit the techniques for proving uniform central limit theorems to obtain rates of convergence for $\hat{\mathbf{g}}_n$. Here, we make the distinction between finite-dimensional models and infinite-dimensional models more explicit. We show to what extent the speed of estimation, i.e. the rate at which the estimation error goes to zero, can be deduced

from the *entropy* of $\mathcal{G}$. In Section 6.4 the theory is applied to two-phase regression and the results are compared with those of Chapter 5.

Because two-phase regression is closely related to change-point models, we devote a separate chapter to the latter: Chapter 7 concentrates on tests for a change-point. Finally, in Chapter 8 we compute the least squares estimators for the model of Example 1.3, using simulated and real data.

Throughout, we make extensive use of Chapters II and VII from POLLARD (1984). In fact, this present study is very much in the spirit of this book.

We now mention some of our notational conventions:

- $\mathbb{P}$ is the probability measure underlying either the whole sequence of random variables, or the random variables involved in the n-th experiment,
- boldface symbols will always represent random quantities but not vise versa: some random quantities are not boldface because of the limited possibilities of a word processor,
- $\boldsymbol{\epsilon}$ (in boldface) is always the disturbance term. Unfortunately, this typographic distinction is hard to see (c.f. ϵ),
- for small numbers we mostly use the greek letter η,
- θ is a finite-dimensional parameter that possibly indexes g,
- δ is usually employed for defining δ-entropy, but it can also be a small number such as η, or the point mass $\delta(.)$,
- $\mathbf{x}$ or x is always a row-vector,
- L^2 is a Hilbert space of real functions on some Euclidean space, but with functions not identified with equivalence classes,
- $\|.\|$ is the norm of a Euclidean vector or of a function in L^2 (in that case it is a pseudo-norm),

Theorems, lemmas and corollaries will be numbered according to the section they are part of whereas examples and equations are numbered throughout the chapter they are in.

Although many other models also fit into the theory, we mainly consider two-phase regression as an application. For this reason, we shall present a brief overview of the literature on this subject in the next section.

1.2 Multi-phase regression and change-point models

QUANDT (1958) is one of the earlier workers on two-phase regression. He considers the model

$$\mathbf{y}_k = \begin{cases} \alpha^{(1)} + \mathbf{x}_k \beta^{(1)} & + \boldsymbol{\epsilon}_k \text{ if } \mathbf{x}_k \leqslant \gamma \\ \alpha^{(1)} + \gamma \beta^{(1)} + (\mathbf{x}_k - \gamma)\beta^{(2)} & + \boldsymbol{\epsilon}_k \text{ if } \mathbf{x}_k \geqslant \gamma \end{cases}, \tag{1.4}$$

with $\alpha^{(1)}, \beta^{(1)}, \beta^{(2)}$ and the *change-point* γ unknown parameters. The model arises in many fields. A famous example (BACON and WATTS (1971)) is the relation between stagnant surface layer height and flow rate in an inclined channel. The model also describes the influence of warfarin concentration on blood factor VII, of nitrogen concentration on the intake of protein, of after-

tax income on the expenditure on luxury goods, etc.. Recently, IPPEL and BEEM (1986) fitted the model to reaction times as function of some measure of discrepancy between stimuli.

Methods for finding the exact solution for the least squares minimization problem are discussed in HUDSON (1966) and WILLIAMS (1970) extended these techniques to the case of linear three-phase regression. Smooth approximations to the non-differentiable model are given by BACON and WATTS (1971) and TISHLER and ZANG (1981). HINKLEY (1969,1971) studies the asymptotic properties of parameter estimators and procedures for obtaining approximate confidence intervals. FEDER (1975) establishes asymptotic theory for a continuous model of the form

$$\mathbf{y}_k = \begin{cases} g^{(1)}(\mathbf{x}_k,\theta^{(1)})+\epsilon_k \text{ if } \mathbf{x}_k \leqslant \gamma \\ g^{(2)}(\mathbf{x}_k,\theta^{(2)})+\epsilon_k \text{ if } \mathbf{x}_k \geqslant \gamma \end{cases}.$$

He provides conditions for consistency, and - for the situation with $g^{(i)}(x,\theta^{(i)})$ linear in $\theta^{(i)}$, $i=1,2$ - asymptotic normality, assuming that the model is identified at the underlying true state of nature.

A more general model does not impose continuity in the parameters, e.g.

$$\mathbf{y}_k = \begin{cases} \alpha^{(1)}+\mathbf{x}_k\beta^{(1)}+\epsilon_k \text{ if } \mathbf{x}_k \leqslant \gamma \\ \alpha^{(2)}+\mathbf{x}_k\beta^{(2)}+\epsilon_k \text{ if } \mathbf{x}_k > \gamma \end{cases}.$$

An example is the model for eruptions of the Old Faithful Geyser in Yellowstone National Park (COOK and WEISBER (1982)). I am not aware of any asymptotic theory for this model.

An identification problem comes up if for instance in (1.4) $\beta^{(1)}=\beta^{(2)}$. For testing the constancy of the regression relationship, BROWN, DURBIN and EVANS (1975) propose a *cusum* and *cusum of squares* test. They assume normality of the errors, so that their tests can be compared with the likelihood ratio test. Asymptotic comparison in the *large deviations* sense is carried out by DESHAYES and PICARD (1982). Many other tests have been developed (e.g. FERREIRA (1975) and MOEN and e "Broemeling" (1984) propose Bayesian test procedures). In Chapter 7 we shall give our contribution to this discussion.

Example 1.3 of the previous section deals with another extension of (1.4). Here, the regressors are in higher-dimensional Euclidean space $\mathbb{R}^d$, and one can no longer speak of a change-point. The general linear two-phase regression model - with obvious extension to p-phase regression - assumes functions of the form

$$g(x) = \begin{cases} g^{(1)}(x,\theta^{(1)}) \text{ if } x \in A \\ g^{(2)}(x,\theta^{(2)}) \text{ if } x \notin A \end{cases},$$

where $g^{(i)}:\mathbb{R}^d \times \Theta \to \mathbb{R}$ is linear in the parameter, $i=1,2$, $\Theta \subset \mathbb{R}^r$ and where A varies in a class $\mathcal{A}$ of subsets of $\mathbb{R}^d$. In Section 3.4 we shall formulate conditions on $\mathcal{A}$ that can lead to consistency of the least squares estimators of $\theta^{(i)}$, $i=1,2$ and A and Chapter 5 presents conditions for asymptotic normality of the estimators of the Euclidean parameters. In cluster analysis (see e.g.

POLLARD (1981) and two-lines least squares (LENSTRA et al. (1982)) $\mathcal{C}$ is the collection of all subsets of $\mathbb{R}^d$. In that case the least squares estimator of g will generally be inconsistent. However, the aim in cluster analysis and two-lines least squares is not to estimate the regression but some other quantity of interest.

Let us return for a moment to the model in Example (1.1). It is widely used, e.g. in ROYSTON and ABRAMS (1980) it describes the shift in basal body temperature of a woman. It can be written in the more conventional form

$$\mathbf{y}_k = \begin{cases} \alpha^{(1)} + \epsilon_k \ , k = 1, \cdots \tau \\ \alpha^{(2)} + \epsilon_k \ , k = \tau + 1, \cdots, n \end{cases} . \tag{1.4}$$

In a general change-point model , one has observations $\mathbf{y}_1, \cdots, \mathbf{y}_\tau$ from distibution $F^{(1)}$ and $\mathbf{y}_{\tau+1}, \cdots, \mathbf{y}_n$ from $F^{(2)}$, where τ as well as $F^{(1)}$ and $F^{(2)}$ are in whole or in part unknown. In HINKLEY (1970) and HINKLEY and HINKLEY (1970), this model is considered for the normal and the binomial distribution respectively. WORSLEY (1985) studies the model for a one-parameter exponential family. Of special interest is testing $F^{(1)} = F^{(2)}$. Worsley considers the exact distribution of the likelihood ratio test and confidence intervals for the change-point τ. The asymptotic null-distibution is given in HACCOU et al. (1985) in the case of exponential distributions, and in Chapter 7 in the case of normal errors. In Chapter 7 also Bahadur efficiency in the situation of a one-parameter exponential family is obtained and contrasted with efficiency at local alternatives.

WOLFE and SCHECHTMAN (1984) establish nonparametric confidence intervals for τ. PETTITT (1979) investigates a nonparametric procedure for testing $F^{(1)} = F^{(2)}$. His statistic is an extension of the Mann-Whitney test for the two-sample problem. PICARD and DESHAYES (1983) propose a Kolmogorov-Smirnov type of test. In PRAAGMAN (1986), the asymptotic efficiencies of a broad class of linear rank statistics are compared.

Change-points can occur anywhere, for instance in hazard rates (see e.g. NGUYEN, ROGERS and WALKER (1984)) and in time-series (PICARD (1983)). We shall only investigate changes in parameters in a sequence of independent random variables, i.e. two-phase regression type of models. We also point out that in the literature mentioned above, the sample size is nonrandom. The problem is to be distinguished from what one could call 'alarm detection', where a process is followed in time and the aim is to react as quickly as possible when it is likely enough that a change has occurred (see e.g. SHIRYAYEV (1963)).

2. Empirical process theory I

2.1. Vapnik and Chervonenkis theory

Let us reconsider the multi-dimensional two-phase regression model of Example 1.3:

$$\mathbf{y}_k = \begin{cases} \alpha^{(1)} + \mathbf{x}_k \beta^{(1)} + \epsilon_k & \text{if } \mathbf{x}_k \gamma \leqslant 1 \\ \alpha^{(2)} + \mathbf{x}_k \beta^{(2)} + \epsilon_k & \text{if } \mathbf{x}_k \gamma > 1 \end{cases},$$

with $\mathbf{x}_1, \cdots, \mathbf{x}_n$ i.i.d. (row-)vectors in $\mathbb{R}^d$ with distribution H, and $\theta^{(i)} = (\alpha^{(i)}, \beta^{(i)T})^T$ and γ unknown (column-)vectors. Example 1.3 is about the case $d = 2$. In the more simple situation with $d = 1$, the subsets $A = \{x: x\gamma \leqslant 1\}$ are half-lines, and the model can be written as

$$\mathbf{y}_{(k)} = \begin{cases} \alpha^{(1)} + \mathbf{x}_{(k)} \beta^{(1)} + \epsilon_{(k)} & \text{if } k \leqslant \tau \\ \alpha^{(2)} + \mathbf{x}_{(k)} \beta^{(2)} + \epsilon_{(k)} & \text{if } k > \tau \end{cases}$$

with $\mathbf{x}_{(1)} \leqslant \cdots \leqslant \mathbf{x}_{(n)}$ the order statistics, and $\mathbf{y}_{(k)}$ and $\epsilon_{(k)}$ the regressor and disturbance term corresponding to $\mathbf{x}_{(k)}$, respectively. The least squares estimators are obtained in the following way. For each l, compute (if possible) ordinary least squares estimators $\theta_l^{(i)}$, $i = 1,2$ of $\theta^{(i)}$, $i = 1,2$, and the residual sum of squares $(\mathbf{S}_l^{(i)})^2$, $i = 1,2$, given that the change-point is at l. Let $\hat{\tau}$ be the value of l where $(\mathbf{S}_l^{(1)})^2 + (\mathbf{S}_l^{(2)})^2$ has its minimum. Then $\hat{\theta}_i = \theta_{\hat{\tau}}^{(i)}$, $i = 1,2$ is the least squares estimator in the two-phase regression model (without the continuity restriction $\alpha^{(1)} + \gamma\beta^{(1)} = \alpha^{(2)} + \gamma\beta^{(2)}$). The subsets of the form $\{x\gamma \leqslant 1\}$ of the data are

$$\{\mathbf{x}_{(1)}\}, \{\mathbf{x}_{(1)}, \mathbf{x}_{(2)}\}, \cdots, \{\mathbf{x}_{(1)}, \cdots, \mathbf{x}_{(n)}\}$$

and complements. Hence, the number of times one has to do ordinary least squares is at most $2(n-3)+1$, since it suffices to consider only those partitions where both $\theta^{(1)}$ and $\theta^{(2)}$ are identified. If all $\mathbf{x}_k$'s are different, l can take the values $\{2,3, \cdots, n-2\}$ and n.

In the case $d > 1$, the $\mathbf{x}_k$ can no longer be ordered. Still, it is not difficult to generate all different subsets of the form $\{x: x\gamma \leqslant 1\}$ of the data (see also Steiner (1826), Schläfli (1901), Cover (1965), Harding (1967) and Watson (1969) for combinatorial results). Let $\mathbf{x}_{l_1}, \cdots, \mathbf{x}_{l_d}$ be a d-tuple from $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$. Write $\mathbf{X}_{l_1, \cdots, l_d} = (\mathbf{x}_{l_1}^T, \cdots, \mathbf{x}_{l_d}^T)^T$ and let e be the d-dimensional vector $(1, \cdots, 1)$. For $\mathbf{X}_{l_1, \cdots, l_d}$ non-singular, we can take as the partition corresponding to $\mathbf{X}_{l_1, \cdots, l_d}$: $\{\mathbf{A}_{l_1, \cdots, l_d} = \{x: x\gamma_{l_1, \cdots, l_d} \leqslant 1\}, \mathbf{A}_{l_1, \cdots, l_d}^c\}$, with $\gamma_{l_1, \cdots, l_d} = \mathbf{X}_{l_1, \cdots, l_d}^{-1} e$. Since these are at most $\binom{n}{d}$ d-tuples for which $\mathbf{X}_{l_1, \cdots, l_d}$ is non-singular, the number of times one has to do ordinary least squares is $\mathcal{O}(n^d)$. The computation of the least squares estimator can be done in polynomial time.

As we shall see, the fact that the number of different partitions is polynomial in n can also be used to derive some asymptotic properties of the least

squares estimator. So-called *empirical process theory* provides the theoretical background.

Let $\mathcal{A}$ be a class of measurable subsets of $\mathbb{R}^d$, and let $\Delta^{\mathcal{A}}(\mathbf{x}_1, \cdots, \mathbf{x}_n)$ be the number of different partitions of $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ of the form $A \cap \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$, $A^c \cap \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$, $A \in \mathcal{A}$. Then $\Delta^{\mathcal{A}}(\mathbf{x}_1, \cdots, \mathbf{x}_n)$ is always at most 2^n. We have seen that for

$$\mathcal{A} = \{\{x: x\gamma \leqslant 1\}: \gamma \in \mathbb{R}^d\}$$

$\Delta^{\mathcal{A}}(\mathbf{x}_1, \cdots, \mathbf{x}_n) = \mathcal{O}(n^d)$. Let $\mathbf{H}_n = 1/n \sum_{k=1}^n \delta_{\mathbf{x}_k}$ be the empirical measure based on $\mathbf{x}_1, \cdots, \mathbf{x}_n$. The Glivenko-Cantelli Theorem states that if $\mathcal{A}$ is the collection of lower-orthants $\{(-\infty, x]: x \in \mathbb{R}^d\}$, then

$$\lim_{n \to \infty} \sup_{A \in \mathcal{A}} |\mathbf{H}_n(A) - H(A)| = 0 \quad \text{almost surely.} \tag{2.1}$$

VAPNIK and CHERVONENKIS (1971) extended this to more general classes of subsets $\mathcal{A}$ than lower orthants. They show that if $\Delta^{\mathcal{A}}(\mathbf{x}_1, \cdots, \mathbf{x}_n)$ does not grow exponentially fast, then (2.1) holds for $\mathcal{A}$ - provided some conditions on measurability are fulfilled.

We have to impose measurability conditions, because the supremum of an uncountable set of measurable functions need not be measurable. We shall assume that $\mathcal{A}$ is *permissible* in the sense of POLLARD (1984). The definition of permissibility is given in Section 2.4. At this stage, it is only necessary to know that for a permissible class $\mathcal{A}$, $\sup_{A \in \mathcal{A}} |\mathbf{H}_n(A) - H(A)|$ is measurable.

Also, quantities like $\Delta^{\mathcal{A}}(\mathbf{x}_1, \cdots, \mathbf{x}_n)$ need not be measurable, even if $\mathcal{A}$ is permissible. However, it turns out that if probability statements about $\Delta^{\mathcal{A}}(\mathbf{x}_1, \cdots, \mathbf{x}_n)$ are replaced by statements in terms of outer-probabilities and upper-expectations, the theory goes through. For definiteness, let $(\Omega, \mathcal{E}, \mathbb{P})$ be the underlying probability space, and write $\mathbb{E}(\cdot)$ for taking expectations under $\mathbb{P}$. Define for $A \subset \Omega$,

$$\mathbb{P}^*(A) = \inf\{\mathbb{P}(B): B \supset A,\ B \in \mathcal{E}\}$$

and for a real function f on Ω and the Borel σ-algebra $\mathcal{B}$ on $\mathbb{R}$,

$$\mathbb{E}^*(f) = \inf\{\mathbb{E}(g): g \geqslant f,\ g\ \mathcal{E}/\mathcal{B}-\text{measurable}\}.$$

THEOREM 2.1.1. *For a permissible class $\mathcal{A}$ the following statements are equivalent*

(i) $\mathbb{E}^*\left(\frac{1}{n} \log \Delta^{\mathcal{A}}(\mathbf{x}_1, \cdots, \mathbf{x}_n)\right) \to 0$,

(ii) $\sup_{A \in \mathcal{A}} |\mathbf{H}_n(A) - H(A)| \xrightarrow{\mathbb{P}} 0$,

(iii) $\frac{1}{n} \log \Delta^{\mathcal{A}}(\mathbf{x}_1, \cdots, \mathbf{x}_n) \xrightarrow{\mathbb{P}^*} 0$,

(iv) $\sup_{A \in \mathcal{A}} |\mathbf{H}_n(A) - H(A)| \to 0$ *almost surely.*

PROOF. See VAPNIK and CHERVONENKIS (1971), and for measurability issues and (iv) STEELE (1978) and POLLARD (1981). □

Results of this type can be used in two-phase regression to obtain strong consistency. But there are also results available that are even more directly applicable.

Let $\mathcal{G}$ be a class of measurable real functions on $\mathbb{R}^d$. Suppose that the functions in $\mathcal{G}$ are uniformly bounded, i.e.

$$\sup_{g \in \mathcal{G}} |g| \leqslant M,$$

for some constant M. Endow $\mathcal{G}$ with $L^\infty(\mathbb{R}^d, \mathbf{H}_n)$ semi-norm $\|\cdot\|_{\infty,n}$:

$$\|g\|_{\infty,n} = \max_{1 \leqslant k \leqslant n} |g(\mathbf{x}_k)|.$$

For each $\delta > 0$, let $N_\infty(\delta, \mathbf{H}_n, \mathcal{G})$ be the minimum value of $\mathbf{m}$, such that there exist functions $\mathbf{g}_1, \cdots, \mathbf{g}_\mathbf{m}$, in $\mathcal{G}$, such that for each $g \in \mathcal{G}$

$$\min_{j=1,\cdots,\mathbf{m}} \|g - \mathbf{g}_j\|_{\infty,n} < \delta.$$

For example, if $\mathcal{G}$ is a class $\mathcal{A}$ of indicator functions, then (identify sets with their indicators) $N_\infty(\delta, \mathbf{H}_n, \mathcal{A}) = \Delta^{\mathcal{A}}(\mathbf{x}_1, \cdots, \mathbf{x}_n)$, $\delta < 1$.

We call $N_\infty(\delta, \mathbf{H}_n, \mathcal{G})$ the (δ-)covering number of $\mathcal{G}$ with respect to the $L^\infty(\mathbb{R}^d, \mathbf{H}_n)$-norm. This terminology is also used by POLLARD (1984), but he does not require that the covering set $\{\mathbf{g}_j, j = 1, \cdots, \mathbf{m}\}$ is a subset of $\mathcal{G}$. Note that if $\mathbf{g}_1, \cdots, \mathbf{g}_\mathbf{m}$ form a δ-covering set, not necessarily in $\mathcal{G}$, one can always construct a 2δ-covering set $\tilde{\mathbf{g}}_1, \cdots, \tilde{\mathbf{g}}_\mathbf{m}$ with $\tilde{\mathbf{g}}_j \in \mathcal{G}$.

In the following theorem, we assume *permissibility* of $\mathcal{G}$. In fact this concept is defined for classes of functions, with a collection of sets as special case. Permissibility of $\mathcal{G}$ implies measurability of

$$\sup_{g \in \mathcal{G}} |\int g d(\mathbf{H}_n - H)|.$$

Again, permissibility need not result in measurable covering numbers $N_\infty(\delta, \mathbf{H}_n, \mathcal{G})$ (see Section 2.4).

THEOREM 2.1.2. *For a permissible class $\mathcal{G}$ of uniformly bounded functions, the following statements are equivalent*

(i) $\mathbb{E}^*(\frac{1}{n} \log N_\infty(\delta, \mathbf{H}_n, \mathcal{G})) \to 0$ *for all* $\delta > 0$,

(ii) $\sup_{g \in \mathcal{G}} |\int g d(\mathbf{H}_n - H)| \xrightarrow{\mathbf{P}} 0$,

(iii) $\frac{1}{n} \log N_\infty(\delta, \mathbf{H}_n, \mathcal{G}) \xrightarrow{\mathbf{P}^*} 0$ *for all* $\delta > 0$,

(iv) $\sup_{g \in \mathcal{G}} |\int g d(\mathbf{H}_n - H)| \to 0$ *almost surely.*

PROOF. VAPNIK and CHERVONENKIS (1981) obtained the uniform weak law of large numbers, and STEELE (1978) shows that convergence in probability implies almost sure convergence, by noting that

$$\sup_{g \in \mathcal{G}} |\int g d(\mathbf{H}_n - H)|$$

is a *subadditive* process. Statement (iv) of Theorem 2.1.1 is a special case of this. □

In two-phase regression, least squares estimators can be obtained in polynomial time, if the covering number of the class of feasible partitions does not grow exponentially fast. This property also leads to a uniform law of large numbers, as Theorems 2.1.1 and 2.1.2 assert. We shall briefly indicate why.

For bounded random variables (such as $1_A(\mathbf{x})$ or $g(\mathbf{x})$, g bounded), one has exponential probability inequalities (see e.g. BERNSTEIN (1924, 1927), HOEFFDING (1963)). For instance, for $|g| \leqslant M$, Berstein's inequality says that

$$\mathbb{P}(|\int g d(\mathbf{H}_n - H)| > t) \leqslant 2 \exp\left[\frac{-nt^2}{2\sigma^2 + \frac{2}{3}Mt}\right],$$

where $\sigma^2 = \mathbb{E}(g(\mathbf{x}) - \mathbb{E}g(\mathbf{x}))^2$. Now if the covering number of $\mathcal{G}$ does not grow exponentially fast, there are only $m = \exp(o(n))$ essentially different g's in $\mathcal{G}$. Moreover, if card $(\mathcal{G}) = m$

$$\mathbb{P}(\sup_{g \in \mathcal{G}} |\int g d(\mathbf{H}_n - H)| > t) \leqslant m \max_{g \in \mathcal{G}} \mathbb{P}(|\int g d(\mathbf{H}_n - H)| > t).$$

These observations, and a randomization device (which is necessary because $N_\infty(\delta, \mathbf{H}_n, \mathcal{G})$ is random) are the major ingredients of the proof of the sufficiency part of Theorem 2.1.2 (2.1.1).

2.2. Pollard's law of large numbers

For $1 \leqslant s < \infty$ and for $\mathbf{Q}$ some probability measure on $\mathbb{R}^d$, we denote by $L^s(\mathbb{R}^d, \mathbf{Q})$ the space of measurable real functions g on $\mathbb{R}^d$ with $(\int |g|^s d\mathbf{Q})^{1/s} < \infty$. In most of what follows, $\mathbf{Q}$ will be the empirical measure $\mathbf{H}_n$ or the (theoretical) measure H. We denote the $L^s(\mathbb{R}^d, \mathbf{H}_n)$-(pseudo)norm by

$$\|\cdot\|_{s,n} = (\int |\cdot| d\mathbf{H}_n)^{1/s}$$

and we sometimes call this the *empirical* norm. The theoretical counterpart is

$$\|\cdot\|_s = (\int |\cdot|^s dH)^{1/s}.$$

For $\mathcal{G}$ a class of functions, the *envelope* G of $\mathcal{G}$ is defined as

$$G = \sup_{g \in \mathcal{G}} |g|.$$

Moreover, for $\mathcal{G} \subset L^s(\mathbb{R}^d, \mathbf{Q})$, we define the covering number $N_s(\delta, \mathbf{Q}, \mathcal{G})$ as the smallest value of $\mathbf{m}$ such that there exist $\mathbf{g}_1, \cdots, \mathbf{g}_\mathbf{m}$ in $\mathcal{G}$ such that for all $g \in \mathcal{G}$

$$\min_{j=1,\cdots,\mathbf{m}} (\int |g - \mathbf{g}_j|^s d\mathbf{Q})^{1/s} < \delta.$$

The logarithm of $N_s(\delta, \mathbf{Q}, \mathcal{G})$ is called the *δ-entropy* of $\mathcal{G}$ for the metric

$(\int |\cdot|^s d\mathbf{Q})^{1/s}$.

In the previous subsection, we considered a class of uniformly bounded functions, i.e. $G \in L^\infty(\mathbb{R}^d, \mathbf{Q})$ for all $\mathbf{Q}$. In that case $L^\infty(\mathbb{R}^d, \mathbf{H}_n)$-covering numbers are useful. For a class of possibly unbounded functions, with $G \in L^s(\mathbb{R}^d, H)$, $1 \leqslant s < \infty$, it is more appropriate to work with the $N_s(\delta, \mathbf{H}_n, \mathcal{G})$-covering number of $\mathcal{G}$ equipped with $\|\cdot\|_{s,n}$-norm. We shall first treat the case $s = 1$ and afterwards extend this to arbitrary $s \geqslant 1$.

THEOREM 2.2.1. *Suppose $\mathcal{G}$ is a permissible class with envelope G. Then*

$$\sup_{g \in \mathcal{G}} |\int g d(\mathbf{H}_n - H)| \to 0 \tag{2.2}$$

almost surely if and only if both $G \in L^1(\mathbb{R}^d, H)$ and

$$\frac{1}{n} \log N_1(\delta, \mathbf{H}_n, \mathcal{G}) \overset{\mathbf{P}^*}{\to} 0 \tag{2.3}$$

for all $\delta > 0$.

PROOF. POLLARD (1981) shows that if $G \in L^1(\mathbb{R}^d, H)$, (2.3) implies (2.2), and GINÉ and ZINN (1984) prove necessity of (2.3) and of the envelope condition $G \in L^1(\mathbb{R}^d, H)$. □

Remember that for bounded random variables, exponential probability inequalities are available, whereas this need not be the case for unbounded random variables. Therefore, one might have expected that in the unbounded case a more stringent condition than (2.3) on the covering numbers is needed, in order to arrive at the uniform law of large numbers (2.2). The following theorem shows that if $N_1(\delta, \mathbf{H}_n, \mathcal{G})$ does not grow exponentially fast, it does not grow at all. This result is due to VAPNIK and CHERVONENKIS (1981) and GINÉ and ZINN (1984). Because the result is somewhat hidden in literature, we give a full proof.

THEOREM 2.2.2. *Suppose $\mathcal{G}$ is a permissible class with envelope $G \in L^1(\mathbb{R}^d, H)$. Then*

$$\frac{1}{n} \log N_1(\delta, \mathbf{H}_n, \mathcal{G}) \overset{\mathbf{P}^*}{\to} 0 \tag{2.4}$$

for all $\delta > 0$ implies that the theoretical covering number $N_1(\delta, H, \mathcal{G})$ is finite, i.e.

$$T_1(\delta) = N_1(\delta, H, \mathcal{G})$$

is a finite function of $\delta > 0$. Furthermore

$$\mathbb{P}^*(\limsup_{n \to \infty} N_1(\delta, \mathbf{H}_n, \mathcal{G}) > T_1(\delta - \eta)) = 0, \quad 0 < \eta < \delta, \quad \delta > 0. \tag{2.5}$$

PROOF. Consider the class $\mathcal{G}' = \{|g - \tilde{g}| : g, \tilde{g} \in \mathcal{G}\}$. This class has envelope $2G \in L^1(\mathbb{R}^d, H)$, and moreover (2.4) implies that also

$$\frac{1}{n}\log N_1(\delta, \mathbf{H}_n, \mathcal{G}') \xrightarrow{\mathbb{P}^*} 0 \quad \text{for all } \delta>0.$$

Hence, we can apply Theorem 2.2.1 to $\mathcal{G}'$, provided it is permissible. Indeed, this follows easily from the permissibility of $\mathcal{G}$, as we show in Section 2.4. It follows that

$$\sup_{g,\tilde{g}\in\mathcal{G}} \left| \int |g-\tilde{g}|\, d(\mathbf{H}_n - H) \right| \tag{2.6}$$

is measurable, and that

$$\sup_{g,\tilde{g}\in\mathcal{G}} \left| \int |g-\tilde{g}|\, d(\mathbf{H}_n - H) \right| \to 0 \quad \text{almost surely.}$$

Or, using the notation in $L^1(\mathbb{R}^d, \cdot)$-norms

$$\sup_{g,\tilde{g}\in\mathcal{G}} \left| \|g-\tilde{g}\|_{1,n} - \|g-\tilde{g}\|_1 \right| \to 0 \quad \text{almost surely.} \tag{2.7}$$

Let

$$A_n = \{\omega\in\Omega: \sup_{g,\tilde{g}\in\mathcal{G}} \left| \|g-\tilde{g}\|_{1,n} - \|g-\tilde{g}\|_1 \right|(\omega) \leqslant \frac{\delta}{2}\}.$$

Note that $A_n\in\mathcal{E}$, i.e. A_n is measurable. Moreover, the almost sure convergence (2.7) implies convergence in probability. So, for $n\geqslant n_0'(=n_0'(\delta))$, n_0' sufficiently large

$$\mathbb{P}(A_n) > 1-\delta.$$

Let $\{\mathbf{g}_1, \cdots, g_{\mathbf{m}}\}$ be a $\delta/2$-covering set of $\mathcal{G}$ endowed with $\|\cdot\|_{1,n}$-norm. On the set A_n, we have

$$\min_{j=1,\cdots,\mathbf{m}} \|g-\mathbf{g}_j\|_1 \leqslant \min_{j=1,\cdots,\mathbf{m}} \|g-g_j\|_{1,n} + \frac{\delta}{2} < \delta.$$

Hence, for $\omega\in A_n$, $N_1(\delta, H, \mathcal{G}) \leqslant N_1(\delta/2, H_n, \mathcal{G})(\omega)$.

Condition (2.4) means by definition that there exists a $B_n\in\mathcal{E}$ such that $\mathbb{P}(B_n)>1-\delta$, and $1/n \log N_1(\delta/2, H_n, \mathcal{G})(\omega)\leqslant\delta$ for $\omega\in B_n$ and for all $n\geqslant n_0''(=n_0''(\delta))$. It follows that for $n_0=\max(n_0', n_0'')$

$$\mathbb{P}(A_{n_0}\cap B_{n_0}) > 1-2\delta.$$

But for $\omega\in A_{n_0}\cap B_{n_0}$

$$N_1(\delta, H, \mathcal{G}) \leqslant \exp(n_0\delta). \tag{2.8}$$

Since (2.8) does not depend on $\omega\in\Omega$, this proves that $N_1(\delta, H, \mathcal{G})$ is finite for all $\delta>0$.

The almost sure convergence (2.7) means that for some $A\in\mathcal{E}$ with $\mathbb{P}(A)=1$, and all $0<\eta<\delta$

$$\sup_{g,g} \left| \|g-\tilde{g}\|_{1,n} - \|g-\tilde{g}\|_1 \right|(\omega) \leqslant \eta$$

for all $n\geqslant n_0(\omega)(=n_0(\delta, \eta, \omega))$ and all $\omega\in A$. Thus

$$N_1(\delta, H_n, \mathcal{G})(\omega) \leqslant N_1(\delta-\eta, H, \mathcal{G}) = T_1(\delta-\eta).$$

for all $n \geqslant n_0(\omega)$, $\omega \in A$. This shows that

$$\mathbb{P}^*(\limsup_{n\to\infty} N_1(\delta, \mathbf{H}_n, \mathcal{G}) > T_1(\delta-\eta)) = 0. \quad \square$$

VAPNIK and CHERVONENKIS (1981) proved that for a uniformly bounded class $\mathcal{G}$,

$$\frac{1}{n} \log N_\infty(\delta, \mathbf{H}_n, \mathcal{G}) \overset{\mathbb{P}^*}{\to} 0 \quad \text{for all} \quad \delta > 0$$

implies that $N_1(\delta, H, \mathcal{G})$ is finite, for all $\delta > 0$, and that this in turn implies that $N_1(\delta, \mathbf{H}_n, \mathcal{G})$ remains finite in probability, for all $\delta > 0$. They do not concern themselves with measurability problems.

The situation with unbounded functions is treated in GINÉ and ZINN (1984). Their approach to measurability issues differs somewhat from ours. Modulo measurability, their Remark 8.9 asserts that for a class $\mathcal{G}$ with $G \in L^1(\mathbb{R}^d, H)$ and for $\mathcal{G}_C = \{g 1_{G \leqslant C} : g \in \mathcal{G}\}$, $C > 0$,

$$\frac{1}{n} \log N_1(\delta, \mathbf{H}_n, \mathcal{G}_C) \overset{\mathbb{P}^*}{\to} 0 \quad \text{for all} \quad \delta > 0, \quad C > 0 \tag{2.9}$$

implies that there exists a finite function $T(\delta)$ such that

$$\lim_{n\to\infty} \mathbb{P}^*(N_1(\delta, \mathbf{H}_n, \mathcal{G}) > T(\delta)) = 0, \quad \text{for all} \quad \delta > 0.$$

It is easy to see that if $G \in L^1(\mathbb{R}^d, H)$, than (2.9) and (2.4) are equivalent.

We call a class $\mathcal{G}$ equipped with some metric *totally bounded* if for all $\delta > 0$, the number of elements of a minimal δ-covering set is finite. Since (2.4) is a necessary condition for the uniform law of large numbers over a permissible $\mathcal{G}$, a reformulation of one of the results of Theorem 2.2.2 says that a necessary condition for the uniform law of large numbers, is that $\mathcal{G}$ is totally bounded for $\|\cdot\|_1$. In other words, the closure of $\mathcal{G}$ should be compact.

We shall now investigate the relation between $L^\infty(\mathbb{R}^d, \mathbf{H}_n)$-, $L^1(\mathbb{R}^d, \mathbf{H}_n)$- and other $L^s(\mathbb{R}^d, \mathbf{H}_n)$-covering numbers, and what consequences conditions like (2.3) on these covering numbers have if $G \in L^s(\mathbb{R}^d, H)$. Note first of all, that combination of Theorems 2.1.2 and 2.2.1 yields that for a permissible class $\mathcal{G}$ of uniformly bounded functions

$$\frac{1}{n} \log N_1(\delta, \mathbf{H}_n, \mathcal{G}) \overset{\mathbb{P}^*}{\to} 0$$

iff

$$\frac{1}{n} \log N_\infty(\delta, \mathbf{H}_n, \mathcal{G}) \overset{\mathbb{P}^*}{\to} 0.$$

For classes of unbounded functions, it is often easier to employ a truncation device. GINÉ and ZINN (1984) use truncation at $\{G > C\}$ and work with $\mathcal{G}_C = \{g 1_{G \leqslant C} : g \in \mathcal{G}\}$, $C > 0$. For reasons that will become clear in Section 2.3,

we introduce an other way of truncation. Define for all $C>0$

$$(g)_C = \begin{cases} C & \text{if} \quad g>C \\ g & \text{if} \quad -C\leqslant g\leqslant C \\ -C & \text{if} \quad g<-C \end{cases}.$$

Let $(\mathcal{G})_C=\{(g)_C\colon g\in\mathcal{G}\}$.

LEMMA 2.2.3. *If $G\in L^s(\mathbb{R}^d, H)$, $1\leqslant s<\infty$, then for all $\delta>0$ there exists a $C>0$ such that*

$$N_s(\delta, H, \mathcal{G}) \leqslant N_s(\frac{\delta}{2}, H, (\mathcal{G})_C), \tag{2.10}$$

and with probability 1 for n sufficiently large

$$N_s(\delta, \mathbf{H}_n, \mathcal{G}) \leqslant N_s(\frac{\delta}{3}, \mathbf{H}_n, (\mathcal{G})_C). \tag{2.11}$$

Moreover, for $1\leqslant s<\infty$ and arbitrary probability measure $\mathbf{Q}$, $\delta>0$, $C>0$

$$N_s(\delta, \mathbf{Q}, (\mathcal{G})_C) \leqslant N_s(\delta, \mathbf{Q}, \mathcal{G}) \tag{2.12}$$

$$N_1(\delta, \mathbf{Q}, (\mathcal{G})_C) \leqslant N_s(\delta, \mathbf{Q}, (\mathcal{G})_C)\leqslant N_1(\frac{\delta^s}{(2C)^{s-1}}, \mathbf{Q}, (\mathcal{G})_C) \tag{2.13}$$

and, if we denote by $\mathcal{G}^s$

$$\mathcal{G}^s = \{|g|^s\colon g\in\mathcal{G}\}$$

$$N_1(\delta, \mathbf{Q}, (\mathcal{G}^s)_C) \leqslant N_1(\frac{\delta}{(2C)^{\frac{s-1}{s}}}, \mathbf{Q}, (\mathcal{G})_{C^{1/s}}). \tag{2.14}$$

PROOF. Let $g, \tilde{g}\in\mathcal{G}$ be arbitrary. If $G\in L^s(\mathbb{R}^d, H)$, then

$$\lim_{C\to\infty} \|G-(G)_C\|_s = 0$$

as well as

$$\lim_{C\to\infty} \limsup_{n\to\infty} \|G-(G)_C\|_{s,n} = 0 \quad \text{almost surely.}$$

Since for arbitrary $\mathbf{Q}$

$$(\int |g-\tilde{g}|^s d\mathbf{Q})^{1/s} \leqslant (\int |(g)_C-(\tilde{g})_C|^s d\mathbf{Q})^{1/s}+2(\int |G-(G)_C|^s d\mathbf{Q})^{1/s},$$

this implies (2.10) and (2.11).

Of course, $|(g)_C-(\tilde{g})_C|\leqslant|g-\tilde{g}|$, so (2.12) follows easily. Furthermore, for arbitrary $\mathbf{Q}$,

$$\begin{aligned}\int |(g)_C-(\tilde{g})_C|\, d\mathbf{Q} &\leqslant (\int |(g)_C-(\tilde{g})_C|^s d\mathbf{Q})^{1/s} \\ &\leqslant ((2C)^{s-1}\int |(g)_C-(\tilde{g})_C|\, d\mathbf{Q})^{1/s},\end{aligned}$$

which yields (2.13).

Finally, (2.14) follows from

$$\int |(|g|^s)_C - (|\tilde{g}|^s)_C| \, d\mathbf{Q} \leqslant (2C)^{\frac{s-1}{s}} \int |(|g|)_{C^{1/s}} - (|\tilde{g}|)_{C^{1/s}}| \, d\mathbf{Q}$$
$$\leqslant (2C)^{\frac{s-1}{s}} \int |(g)_{C^{1/s}} - (\tilde{g})_{C^{1/s}}| \, d\mathbf{Q}. \quad \square$$

The following theorem is the analogue of Theorem 2.2.1, albeit that we do not present necessary conditions.

THEOREM 2.2.4. *Suppose $\mathcal{G}$ is a permissible class with envelope $G \in L^s(\mathbb{R}^d, H)$, $1 \leqslant s < \infty$. Then*

$$\frac{1}{n} \log N_s(\delta, \mathbf{H}_n, \mathcal{G}) \xrightarrow{\mathbf{P}^*} 0 \quad \textit{for all} \quad \delta > 0 \tag{2.15}$$

implies

$$\sup_{g \in \mathcal{G}} | \, \|g\|_{s,n} - \|g\|_s \, | \to 0 \quad \textit{almost surely.}$$

PROOF. We show in Section 2.4 that also $\mathcal{G}^s$ is permissible. Thus, the theorem is proved if (2.15) implies

$$\frac{1}{n} \log N_1(\delta, \mathbf{H}_n, \mathcal{G}^s) \xrightarrow{\mathbf{P}^*} 0 \quad \text{for all} \quad \delta > 0, \tag{2.16}$$

because then, we can apply Theorem 2.2.1 to $\mathcal{G}^s$. But application of (2.11) and (2.12) with $s = 1$ to $\mathcal{G}^s$, shows that it suffices to prove that (2.16) holds for the truncated class, i.e. that

$$\frac{1}{n} \log N_1(\delta, \mathbf{H}_n, (\mathcal{G}^s)_C) \xrightarrow{\mathbf{P}^*} 0 \quad \text{for all} \quad \delta > 0, \; C > 0.$$

And this follows immediatly from (2.13) and (2.14):

$$N_1(\delta, \mathbf{H}_n, (\mathcal{G}^s)_C) \leqslant N_1\left(\frac{\delta}{(2C)^{\frac{s-1}{s}}}, \mathbf{H}_n, (\mathcal{G})_{C^{1/s}}\right)$$
$$\leqslant N_s\left(\frac{\delta}{(2C)^{\frac{s-1}{s}}}, \mathbf{H}_n, (\mathcal{G})_{C^{1/s}}\right). \quad \square$$

Of course, it also follows from Lemma 2.2.3 that it doesn't really matter which covering numbers are used. This is made explicit in Lemma 2.2.5 below, where we show the analogue of Theorem 2.2.2.

LEMMA 2.2.5. *Suppose that $\mathcal{G}$ is a permissible class with envelope $G \in L^s(\mathbb{R}^d, H)$, $1 \leqslant s < \infty$. Then*

$$\frac{1}{n} \log N_1(\delta, \mathbf{H}_n, \mathcal{G}) \xrightarrow{\mathbf{P}^*} 0 \quad \textit{for all} \quad \delta > 0 \tag{2.17}$$

implies that $\mathcal{G}$ is totally bounded for $\|\cdot\|_s$, i.e.

$$T_s(\delta) = N_s(\delta, H, \mathcal{G})$$

is a finite function of $\delta>0$. Furthermore

$$\mathbb{P}^*(\limsup_{n\to\infty} N_s(\delta, \mathbf{H}_n, \mathcal{G})>T_s(\delta-\eta)) = 0, \quad 0<\eta<\delta, \quad \delta>0. \tag{2.18}$$

PROOF. We have seen in Theorem 2.2.2 that (2.17) implies that $T_1(\delta)=N_1(\delta, H, \mathcal{G})$ is a finite function of δ. In view of (2.12) and (2.13), for all $C>0$

$$N_s(\delta, H, (\mathcal{G})_C) \leqslant N_1(\frac{\delta^s}{(2C)^{s-1}}, H, (\mathcal{G})_C) \leqslant T(\frac{\delta^s}{(2C)^{s-1}})$$

and moreover, by (2.10)

$$N_s(\delta, H, \mathcal{G}) \leqslant N_s(\frac{\delta}{2}, H, (\mathcal{G})_C)$$

for C sufficiently large. This gives that $T_s(\delta)=N_s(\delta, H, \mathcal{G})$ is a finite function of δ.

Using again (2.12), (2.13), we see that (2.17) also implies that for all $C>0$, $\delta>0$

$$\frac{1}{n}\log N_s(\delta, \mathbf{H}_n, (\mathcal{G})_C) \xrightarrow{\mathbf{P}^*} 0$$

and from (2.11), for all $\delta>0$

$$\frac{1}{n}\log N_s(\delta, \mathbf{H}_n, \mathcal{G}) \xrightarrow{\mathbf{P}^*} 0.$$

Hence, in view of Theorem 2.2.4

$$\sup_{g\in\mathcal{G}} |\,\|g\|_{s,n} - \|g\|_s\,| \to 0 \quad \text{almost surely.}$$

But this means that for δ arbitrary, $0<\eta<\delta$, a $(\delta-\eta)$-covering set of $\mathcal{G}$ for $\|\cdot\|_s$ is for n sufficiently large a δ-covering set of $\mathcal{G}$ for $\|\cdot\|_{s,n}$, almost surely. Thus, by the same argument as in the proof of Theorem 2.2.2

$$\mathbb{P}^*(\limsup_{n\to\infty} N_s(\delta, \mathbf{H}_n, \mathcal{G})>T_s(\delta-\eta)) = 0. \quad \square$$

We conclude that if (2.17) holds and $G\in L^s(\mathbb{R}^d, H)$, $1\leqslant s<\infty$ then $\mathcal{G}$ is totally bounded for $\|\cdot\|_s$. If $s=\infty$, (2.17) is equivalent to

$$\frac{1}{n}\log N_\infty(\delta, \mathbf{H}_n, \mathcal{G}) \xrightarrow{\mathbf{P}^*} 0, \quad \text{for all} \quad \delta>0,$$

in particular, if $G\in L^s(\mathbb{R}^d, H)$, (2.17) is equivalent to

$$\frac{1}{n}\log N_\infty(\delta, \mathbf{H}_n, (\mathcal{G})_C) \xrightarrow{\mathbf{P}^*} 0, \quad \text{for all} \quad \delta>0, \quad C>0.$$

This observation is useful because $L^\infty(\mathbb{R}^d, \mathbf{H}_n)$-covering numbers are often easier to compute.

We shall illustrate the results of this subsection with an example. In a substantial number of applications the conditions on the covering numbers can be checked without imposing distributional assumptions, apart from a moment condition on the envelope G. An important special case occurs when a collection $\mathcal{C}$ of sets satisfies

$$\sup_{\{x_1, \cdots, x_n\}} \Delta^{\mathcal{C}}(x_1, \cdots, x_n) \leqslant n^r, \tag{2.19}$$

for some r and all n, $\Delta^{\mathcal{C}}(x_1, \cdots, x_n)$ being defined in Section 2.1. Recall for instance that if $\mathcal{C}=\{\{x: x\gamma\leqslant 1\}, \gamma\in\mathbb{R}^d\}$

$$\sup_{\{x_1, \cdots, x_n\}} \Delta^{\mathcal{C}}(x_1, \cdots, x_n) \leqslant \binom{n}{d} \leqslant n^d.$$

An $\mathcal{C}$ satisfying (2.19) is called a *VC-class* (VAPNIK and CHERVONENKIS (1971)).

For classes of functions, POLLARD (1984) introduces the related concept of *VC-graph classes.* Let $g: \mathbb{R}^d \to \mathbb{R}$ be some function and define the *graph* of g as the subset

$$\{(x,t): 0\leqslant t\leqslant g(x) \quad \text{or} \quad g(x)\leqslant t\leqslant 0\}$$

of $\mathbb{R}^{d+1}$. A class $\mathcal{G}$ is a VC-graph class if the collection of graphs of functions in $\mathcal{G}$ from a VC-class.

THEOREM 2.2.6. *Let* $\mathbf{Q}$ *be some probability measure on* $\mathbb{R}^d$, *and let* $\mathcal{G}$ *be a VC-graph class with envelope* $\int G d\mathbf{Q} = C_Q$ *say. Then*

$$N_1(\delta, \mathbf{Q}, \mathcal{G}) \leqslant A_1 C_Q^{r'} \delta^{-r'} \quad \textit{for all} \quad \delta>0,$$

where A_1 *and* r' *are constants independent of* $\mathbf{Q}$.

PROOF. See POLLARD (1984). □

It is easy to see that if $\mathcal{G}$ is a *VC*-graph class, then so is $(\mathcal{G})_C$. Thus, then

$$N_1(\delta, \mathbf{Q}, (\mathcal{G})_C) \leqslant A_1 C^{r'} \delta^{-r'} \quad \text{for all} \quad \delta>0, \quad C>0$$

and from Lemma 2.2.3, $1\leqslant s<\infty$

$$N_s(\delta, \mathbf{Q}, (\mathcal{G})_C) \leqslant A_1 C^{r'} \left(\frac{\delta^s}{(2C)^{s-1}}\right)^{-r'} = A_s C^{sr'} \delta^{-sr'}, \ \delta>0, \ C>0.$$

Note that if $\mathcal{C}$ is a *VC*-class, then $\{1_A: A\in\mathcal{C}\}$ is a *VC*-graph class. Since the envelope of a collection of indicator functions is bounded by 1, this gives for $\mathcal{C}$ a *VC*-class

$$N_s(\delta, \mathbf{Q}, \mathcal{C}) \leqslant A_s \delta^{-r's} \quad \text{for all} \quad \delta>0, \quad 1\leqslant s<\infty$$

for some A_s and r', and by (2.19)

$$N_\infty(\delta, \mathbf{Q}, \mathcal{G}) \leqslant n^r$$

for some r.

EXAMPLE 1.3 continued. In this two-phase regression model, $\mathcal{G}$ is a class of regression functions of the form

$$\begin{aligned}\mathcal{G} = \{g(x) = &(\alpha^{(1)} + x\beta^{(1)})1_{\{x\gamma \leqslant 1\}}(x) \\ &+ (\alpha^{(2)} + x\beta^{(2)})1_{\{x\gamma > 1\}}(x): \\ &\alpha^{(i)} \in \mathbb{R},\ \beta^{(i)} \in \mathbb{R}^d,\ i = 1,2,\ \gamma \in \mathbb{R}^d\}.\end{aligned}$$

The graph of a $g \in \mathcal{G}$ is the union of two intersections of three halfspaces. Now, the class of halfspaces forms a *VC*-class. And it is easy to see that the *VC*-property is preserved under taking finite unions and intersections. Hence, $\mathcal{G}$ is a *VC*-graph class.

2.3. Extensions

In many regression models, the class of feasible regression functions is allowed to vary with the number of observations. Also, the independent variables and disturbances are often not identically distributed, and their distributions might vary with n too. To handle these situations, we generalize some of the results of the previous sections.

Let for each $n = 1,2, \cdots$, $\mathbf{x}_{n,1}, \cdots, \mathbf{x}_{n,n}$ be independent random vectors in $\mathbb{R}^d$, $\mathbf{x}_{n,k}$ having distribution $H_{n,k}$. Furthermore, let for each $n \in \mathbb{N}$, $\mathcal{G}_n$ be a class of functions on $\mathbb{R}^d$ with envelope $G_n = \sup_{g \in \mathcal{G}_n} |g|$. Define

$$H^{(n)} = 1/n \sum_{k=1}^{n} H_{n,k}$$

and let $\mathbf{H}_n$ be the empirical measure generated by $\mathbf{x}_{n,1}, \cdots, \mathbf{x}_{n,n}$.

To establish a uniform law of large numbers, we make use of Hoeffding's inequality.

LEMMA 2.3.1 (Hoeffding's inequality). *Let* $\mathbf{y}_1, \cdots, \mathbf{y}_n$ *be independent random variables with zero means and bounded ranges:* $a_k \leqslant \mathbf{y}_k \leqslant b_k$. *Then for each* $\eta > 0$

$$\mathbb{P}\left(\frac{1}{n}\sum_{k=1}^{n} \mathbf{y}_k \geqslant \eta\right) \leqslant \exp\left[-2n\eta^2 \Big/ \frac{1}{n}\sum_{k=1}^{n}(b_k - a_k)^2\right].$$

PROOF. HOEFFDING (1963). □

We have seen that in the i.i.d. case with $\mathcal{G}_n = \mathcal{G}$ (Section 2.2), necessary conditions for the uniform law of large numbers are that the covering numbers $N_1(\delta, \mathbf{H}_n, \mathcal{G})$ remain bounded in probability, and that the envelope of $\mathcal{G}$ is in $L^1(\mathbb{R}^d, H)$. In general however, the covering numbers are allowed to grow with n. Furthermore, the $L^1(\mathbb{R}^d, H^{(n)})$-norm of the envelope of $\mathcal{G}_n$ is allowed to grow with n too, but the faster $N_1(\delta, \mathbf{H}_n, \mathcal{G}_n)$ grows, the more stringent the envelope conditions become. This result is stated in Theorem 2.3.2 below. We

shall also show that for the case of i.i.d. random variables and $\mathcal{G}_n$ not depending on n, the conditions of Theorem 2.3.2 reduce to those of Theorem 2.2.1.

In the general set up, with triangular arrays, it is not possible to obtain a strong uniform law of large numbers: all results only concern convergence in probability. The assumption of permissibility is needed again to guard against measurability difficulties (see Section 2.4). We shall prove the uniform law of large numbers exactly according to the recipe Pollard supplies for the i.i.d. case (POLLARD (1984), Ch. II). This illustrates the power of the techniques Pollard proposes.

THEOREM 2.3.2. *Let $\{\mathcal{G}_n\}$ be a sequence of permissible classes with envelopes $G_n = \sup_{g \in \mathcal{G}_n} |g|$. Suppose that for some sequence $c_n \geq 1$, $c_n = \mathcal{O}(n)$*

$$\limsup_{n \to \infty} \int_{G_n > c_n} G_n dH^{(n)} = 0, \tag{2.20}$$

$$\limsup_{n \to \infty} \frac{1}{c_n} \int_{G_n \leq c_n} G_n^2 dH^{(n)} = 0, \tag{2.21}$$

and that $(c_n/n) \log N_1(\delta, \mathbf{H}_n, \mathcal{G}_n)$ remains bounded in probability, i.e.

$$\lim_{T \to \infty} \limsup_{n \to \infty} \mathbb{P}^*\left(\frac{c_n}{n} \log N_1(\delta, \mathbf{H}_n, \mathcal{G}_n) > T\right) = 0 \tag{2.22}$$

for all $\delta > 0$. Then

$$\sup_{g \in \mathcal{G}_n} \left| \int g d(\mathbf{H}_n - H^{(n)}) \right| \xrightarrow{\mathbf{P}} 0. \tag{2.23}$$

PROOF. First, we shall show that it suffices to prove a uniform law of large numbers for the truncated class $\{g 1_{G_n \leq c_n} : g \in \mathcal{G}_n\}$. Let $0 < \delta \leq 1$ be arbitrary. In view of (2.20)

$$\int_{G_n > c_n} G_n dH^{(n)} < \frac{\delta^2}{4}$$

for all n sufficiently large. Apply Chebyshev's inequality to see that

$$\mathbb{P}\left(\int_{G_n > c_n} G_n d\mathbf{H}_n > \frac{\delta}{4} \right) \leq \frac{\int_{G_n > c_n} G_n dH^{(n)}}{\delta/4} < \delta.$$

Hence

$$\begin{aligned} &\mathbb{P}\left(\sup_{g \in \mathcal{G}_n} \left| \int g d(\mathbf{H}_n - H^{(n)}) \right| > \delta\right) \\ &\leq \mathbb{P}\left(\sup_{g \in \mathcal{G}_n} \left| \int_{G_n \leq c_n} g d(\mathbf{H}_n - H^{(n)}) \right| > \frac{\delta}{2}\right) + \mathbb{P}\left(\int_{G_n > c_n} G_n d\mathbf{H}_n > \frac{\delta}{4} \right) \\ &\leq \mathbb{P}\left(\sup_{g \in \mathcal{G}_n} \left| \int_{G_n \leq c_n} g d(\mathbf{H}_n - H^{(n)}) \right| > \frac{\delta}{2}\right) + \delta. \end{aligned} \tag{2.24}$$

Next, we *symmetrize* the process. For this purpose, we use that for arbitrary $\eta>0$, and for all n sufficiently large

$$\frac{1}{c_n}\int_{G_n\leqslant c_n} G_n^2 dH^{(n)}<\eta$$

by assumption (2.21). Application of Chebyshev's inequality gives that for each $g\in\mathcal{G}_n$

$$\mathbb{P}(|\int_{G_n\leqslant c_n} g d(\mathbf{H}_n-H^{(n)})|>\frac{\delta}{4}) \leqslant \frac{\frac{1}{n}\int_{G_n\leqslant c_n} g^2 dH^{(n)}}{(\delta/4)^2} \tag{2.25}$$

$$\leqslant \frac{\frac{1}{n}\int_{G_n\leqslant c_n} G_n^2 dH^{(n)}}{(\delta/4)^2} \leqslant \frac{\frac{c_n}{n}\eta}{(\delta/4)^2} \leqslant \frac{1}{2},$$

for η sufficiently small, and all n sufficiently large. For the symmetrization, we introduce an independent copy $\{\mathbf{x}'_{n,1},\cdots,\mathbf{x}'_{n,n}\}$ of $\{\mathbf{x}_{n,1},\cdots,\mathbf{x}_{n,n}\}$, i.e. $\mathbf{x}_{n,1},\cdots,\mathbf{x}_{n,n},\mathbf{x}'_{n,1},\cdots,\mathbf{x}'_{n,n}$ are independent and $\mathbf{x}'_{n,k}$ has distribution $H_{n,k}$. Let $\mathbf{H}'_n$ be the empirical distribution, based on $\mathbf{x}'_{n,1},\cdots,\mathbf{x}'_{n,n}$. Since (2.25) holds, we have for all $g\in\mathcal{G}_n$

$$\mathbb{P}(|\int_{G_n\leqslant c_n} g d(\mathbf{H}_n-H^{(n)})|\leqslant\frac{\delta}{4}) \geqslant \frac{1}{2}. \tag{2.26}$$

The assumption of permissibility of $\mathcal{G}_n$ ensures that for some random $\mathbf{g}^*\in\mathcal{G}_n$, independent of $\mathbf{H}'_n$

$$|\int_{G_n\leqslant c_n} g^* d(H_n-H^{(n)})|>\frac{\delta}{2}$$

on the set

$$\{\sup_{g\in\mathcal{G}_n}|\int_{G_n\leqslant c_n} g d(H_n-H^{(n)}|>\frac{\delta}{2}\}$$

(see Section 2.4). Because (2.26) holds for $\mathbf{g}^*$ too,

$$\frac{1}{2}\mathbb{P}(\sup_{g\in\mathcal{G}_n}|\int_{G_n\leqslant c_n} g d(\mathbf{H}_n-H^{(n)})|>\frac{\delta}{2}) \tag{2.27}$$

$$\leqslant \mathbb{P}(|\int_{G_n\leqslant c_n} \mathbf{g}^* d(\mathbf{H}_n-H^{(n)})|>\frac{\delta}{2}\ ,\ |\int_{G_n\leqslant c_n} \mathbf{g}^* d(\mathbf{H}_n'-H^{(n)})|\leqslant\frac{\delta}{4})$$

$$\leqslant \mathbb{P}(|\int_{G_n\leqslant c_n} \mathbf{g}^* d(\mathbf{H}_n-\mathbf{H}_n')|>\frac{\delta}{4})\leqslant\mathbb{P}(\sup_{g\in\mathcal{G}_n}|\int_{G_n\leqslant c_n} g d(\mathbf{H}_n-\mathbf{H}_n')|>\frac{\delta}{4}).$$

We shall now describe the *randomization* device. Let $\boldsymbol{\sigma}_1,\cdots,\boldsymbol{\sigma}_n$ be independent random variables, independent of $\{\mathbf{x}_{n,1},\cdots,\mathbf{x}_{n,n},\mathbf{x}'_{n,1},\cdots,\mathbf{x}'_{n,n}\}$, with

$$\mathbb{P}(\boldsymbol{\sigma}_k=1)=\mathbb{P}(\boldsymbol{\sigma}_k=-1)=\frac{1}{2}.$$

Write $\mathbf{H}_n^0$ for the signed measure that puts mass $1/n\,\boldsymbol{\sigma}_k$ at $\mathbf{x}_{n,k}$, e.g.

$$\int_{G_n \leqslant c_n} g d\mathbf{H}_n^0 = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{\sigma}_k g(\mathbf{x}_{n,k}) 1_{\{G_n \leqslant c_n\}}(\mathbf{x}_{n,k}).$$

Then

$$\begin{aligned}
&\mathbb{P}(\sup_{g\in\mathcal{G}_n} \left| \int_{G_n \leqslant c_n} g d(\mathbf{H}_n - \mathbf{H}_n') \right| > \frac{\delta}{4}) \\
&= \mathbb{P}(\sup_{g\in\mathcal{G}_n} \left| \frac{1}{n}\sum_{k=1}^{n} (g(\mathbf{x}_{n,k}) 1_{\{G_n \leqslant c_n\}}(\mathbf{x}_{n,k}) - g(\mathbf{x}'_{n,k}) 1_{\{G_n \leqslant c_n\}}(\mathbf{x}'_{n,k})) \right| > \frac{\delta}{4}) \\
&= \mathbb{P}(\sup_{g\in\mathcal{G}_n} \left| \frac{1}{n}\boldsymbol{\sigma}_k (\frac{1}{n}\sum (g(\mathbf{x}_{n,k}) 1_{\{G_n \leqslant c_n\}}(\mathbf{x}_{n,k}) - g(\mathbf{x}'_{n,k}) 1_{\{G_n \leqslant c_n\}}(\mathbf{x}'_{n,k})) \right| > \frac{\delta}{4}) \\
&\leqslant \mathbb{P}(\sup_{g\in\mathcal{G}_n} \left| \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{\sigma}_k g(\mathbf{x}_{n,k}) 1_{\{G_n \leqslant c_n\}}(\mathbf{x}_{n,k}) \right| > \frac{\delta}{8}) \\
&+ \mathbb{P}(\sup_{g\in\mathcal{G}_n} \left| \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{\sigma}_k g(\mathbf{x}'_{n,k}) 1_{\{G_n \leqslant c_n\}}(\mathbf{x}'_{n,k}) \right| > \frac{\delta}{8}) \\
&= 2\mathbb{P}(\sup_{g\in\mathcal{G}_n} \left| \int_{G_n \leqslant c_n} g d\mathbf{H}_n^0 \right| > \frac{\delta}{8}) . \qquad (2.28)
\end{aligned}$$

Let $\mathbf{g}_1, \cdots, \mathbf{g}_{\mathbf{m}}$, $\mathbf{m} = N_1(\frac{\delta}{16}, \mathbf{H}_n, \mathcal{G}_n)$ be a minimal $\delta/16$-covering set of $\mathcal{G}_n$. Observe that if

$$\int |g - \mathbf{g}_j| d\mathbf{H}_n < \frac{\delta}{16},$$

also

$$\left| \int_{G_n \leqslant c_n} |g - \mathbf{g}_j| d\mathbf{H}_n^0 \right| < \frac{\delta}{16}.$$

Now, given $(\mathbf{x}_{n,1}, \cdots, \mathbf{x}_{n,n}) = (x_{n,1}, \cdots, x_{n,n})$, with $x_{n,1}, \cdots, x_{n,n}$ satisfying

$$\frac{1}{nc_n}\sum_{k=1}^{n} G_n^2(x_{n,k}) 1_{\{G_n \leqslant c_n\}}(x_{n,k}) \leqslant \frac{\delta^2}{T},$$

we have by Hoeffding's inequality

$$\begin{aligned}
&\mathbb{P}(\sup_{g\in\mathcal{G}_n} \left| \int_{G_n \leqslant c_n} g d\mathbf{H}_n^0 \right| > \frac{\delta}{8} \mid x_{n,1}, \cdots, x_{n,n}) \\
&\leqslant N_1(\frac{\delta}{16} H_n, \mathcal{G}_n) 2 \exp\left[-\frac{n(\frac{\delta}{16})^2}{2c_n\delta^2/T} \right] \\
&= 2N_1(\frac{\delta}{16}, H_n, \mathcal{G}_n) \exp\left[-\frac{nT}{512\, c_n} \right].
\end{aligned}$$

Therefore, by Fubini's theorem

$$\mathbb{P}(\sup_{g\in\mathcal{G}_n}|\int_{G_n\leq c_n} g d\mathbf{H}_n^0|>\frac{\delta}{8}) \leq 2\exp[-\frac{nT}{1024\,c_n}]+\mathbb{P}(A_n)+\mathbb{P}^*(B_n) \tag{2.29}$$

with

$$A_n = \{\frac{1}{c_n}\int_{G_n\leq c_n} G_n^2 dH_n > \frac{\delta^2}{T}\}$$

and

$$B_n = \{\frac{c_n}{n}\log N_1(\frac{\delta}{16}, H_n, \mathcal{G}_n) > T/1024\}.$$

We shall now show that $\mathbb{P}(A_n)$ and $\mathbb{P}(B_n)$ can be made arbitrarily small. Using (2.21), we see that

$$\frac{1}{c_n}\int_{G_n\leq c_n} G_n^2 dH^{(n)} < \frac{\delta^3}{T}$$

for all n sufficiently large. Again by Chebyshev's inequality, this implies

$$\mathbb{P}(A_n) < \frac{\delta^3/T}{\delta^2/T} = \delta.$$

Moreover

$$\mathbb{P}^*(B_n) < \delta$$

for T large enough and all n large enough, because of assumption (2.22).

Returning to (2.29), we see that

$$\mathbb{P}(\sup_{g\in\mathcal{G}_n}|\int_{G_n\leq c_n} g d\mathbf{H}_n^0|>\frac{\delta}{8}) < 2\exp[-\frac{nT}{1024\,c_n}]+2\delta \leq 3\delta$$

for T sufficiently large and all n sufficiently large. In view of the truncation, symmetrization and randomization inequalities ((2.24), (2.27) and (2.28) respectively), this completes the proof. □

We present a weaker version of Theorem 2.3.2 for two reasons. First, this clarifies that Theorem 2.3.2 is a generalization of the sufficiency part of Theorem 2.2.1 and secondly, the weaker version will be used in Chapter 3 to prove consistency of the least squares estimators.

LEMMA 2.3.3. *Suppose $\{\mathcal{G}_n\}$ is a sequence of permissible classes with envelopes G_n. Assume that for some sequence $b_n \geq 1$, $b_n = o(n^{1/2})$*

$$\limsup_{n\to\infty}\int_{G_n>b_n} G_n dH^{(n)} = 0 \tag{2.30}$$

and

$$\frac{b_n^2}{n}\log N_1(\delta, \mathbf{H}_n, \mathcal{G}_n) \xrightarrow{\mathbb{P}^*} 0 \quad \text{for all} \quad \delta>0. \tag{2.31}$$

Then

$$\sup_{g \in \mathcal{G}_n} | \int g d(\mathbf{H}_n - H^{(n)}) | \xrightarrow{\mathbf{P}} 0.$$

PROOF. Since

$$\mathbf{Z}_n(\delta) = \frac{b_n^2}{n} \log N_1(\delta, \mathbf{H}_n, \mathcal{G}_n)$$

is nondecreasing in δ, (2.31) ensures the existence of sequences $\eta_n \downarrow 0$ and $\delta_n \downarrow 0$ such that

$$\mathbb{P}^*(\sup_{\delta \geqslant \delta_n} \mathbf{Z}_n(\delta) > \eta_n) = \mathbb{P}^*(\mathbf{Z}_n(\delta_n) > \eta_n) \to 0 .$$

This implies that there exists a sequence $\tilde{b}_n \geqslant 1$ with $\tilde{b}_n / b_n \to \infty$, $\tilde{b}_n = o(n^{1/2})$, such that

$$\frac{\tilde{b}_n^2}{n} \log N_1(\delta, \mathbf{H}_n, \mathcal{G}_n) \xrightarrow{\mathbf{P}^*} 0 \quad \text{for all} \quad \delta > 0 . \tag{2.32}$$

By (2.30) we have

$$\int_{G_n > \tilde{b}_n^2} G_n dH^{(n)} \leqslant \int_{G_n > b_n} G_n dH^{(n)} \to 0. \tag{2.33}$$

Moreover, also

$$\begin{aligned} \frac{1}{\tilde{b}_n^2} \int_{G_n \leqslant \tilde{b}_n^2} G_n^2 dH^{(n)} &= \frac{1}{\tilde{b}_n^2} \int_{G_n \leqslant b_n} G_n^2 dH^{(n)} + \frac{1}{\tilde{b}_n^2} \int_{b_n < G_n \leqslant \tilde{b}_n^2} g_n^2 dH^{(n)} \\ &\leqslant \frac{b_n^2}{\tilde{b}_n^2} + \int_{G_n > b_n} G_n dH^{(n)} \to 0. \end{aligned} \tag{2.34}$$

Together, (2.32), (2.33) and (2.34) ensure that the conditions of Theorem 2.3.2 are fulfilled with $c_n = \tilde{b}_n^2$. □

Recall that in the i.i.d. case with $\mathcal{G}_n = \mathcal{G}$, a necessary condition for the uniform law of large numbers is that the envelope G is integrable. This corresponds to imposing (2.30) with $\{b_n\}$ any sequence tending to infinity. Letting b_n grow slowly enough, we see that (2.31) reduces to condition (2.3) of Theorem 2.2.1:

$$\frac{1}{n} \log N_1(\delta, \mathbf{H}_n, \mathcal{G}) \xrightarrow{\mathbf{P}^*} 0.$$

Moreover, we showed in Theorem 2.2.2 that under the conditions of Theorem 2.2.1 the covering numbers in fact remain bounded. Obviously, if $\mathcal{G}_n$ varies with n the uniform law of large numbers no longer implies that $N_1(\delta, \mathbf{H}_n, \mathcal{G}_n)$ does not grow with n.

EXAMPLE 2.1. Let $\mathcal{G}$ be a permissible *VC*-graph class with envelope not

necessarily in $L^1(\mathbb{R}^d, H^{(n)})$. As in Section 2.2 we define $(\mathcal{G})_C$ as the class of functions truncated at C:

$$(\mathcal{G})_C = \{\text{sign}(g)[|g| \wedge C]: g \in \mathcal{G}\}.$$

The class $(\mathcal{G})_C$ is still a VC-graph class (with envelope the constant function C). Application of Theorem 2.2.6 yields that for all $\delta > 0$

$$N_1(\delta, \mathbf{H}_n, (\mathcal{G})_C) \leqslant AC^r\delta^{-r}$$

for some constants A and r. Also, if $\mathcal{G}$ is permissible, then so is $(\mathcal{G})_C$ for all $C > 0$.

Let $\eta > 0$ be arbitrary and take $c_n = n(\log n)^{-1}$, then (2.20) and (2.21) hold for $(\mathcal{G})_{n^{1/2}(\log n)^{-1/2-\eta}}$

$$\sup_{g \in \mathcal{G}} |(g)_{n^{1/2}(\log n)^{-1/2-\eta}}| \leqslant n^{1/2}(\log n)^{-1/2-\eta} \leqslant c_n \quad \text{for } n \text{ sufficiently large}$$

$$\sup_{g \in \mathcal{G}} |(g)_{n^{1/2}(\log n)^{-1/2-\eta}}|^2 \leqslant n(\log n)^{-1-2\eta} = o(c_n).$$

Also, (2.23) is met for $(\mathcal{G})_{n^{1/2}(\log n)^{-1/2-\eta}}$:

$$\frac{c_n}{n} \log N_1(\delta, \mathbf{H}_n, (\mathcal{G})_{n^{1/2}(\log n)^{-1/2-\eta}}) = \frac{1}{\log n} O(\log n) = O(1).$$

Hence, for a permissible VC-graph class

$$\sup_{g \in \mathcal{G}} \left| \int (g)_{n^{1/2}(\log n)^{-1/2-\eta}} d(\mathbf{H}_n - H^{(n)}) \right| \xrightarrow{\mathbb{P}} 0.$$

The remainder of this section is devoted to the situation where higher order moments of the envelopes exist:

$$G_n \in L^s(\mathbb{R}^d, H^{(n)}), \quad 1 \leqslant s < \infty.$$

As before, we write

$$\|g\|_{s,n} = \left(\int |g|^s d\mathbf{H}_n\right)^{1/s}$$

for the empirical norm of g. The theoretical norm now also depends on n, and is denoted by

$$\|g\|_{s,(n)} = \left(\int |g|^s dH^{(n)}\right)^{1/s}.$$

Define

$$\mathcal{G}_n^s = \{|g|^s: g \in \mathcal{G}_n\}.$$

Because in general the $L^s(\mathbb{R}^d, H^{(n)})$ norm will be allowed to grow with n, it is no longer possible to replace conditions on $L^1(\mathbb{R}^d, \mathbf{H}_n)$-covering numbers by conditions on $L^s(\mathbb{R}^d, \mathbf{H}_n)$-covering numbers. We present a lemma to clarify this.

LEMMA 2.3.4. *For $1 \leqslant s < \infty$ and all $\delta > 0$*

$$N_1(\delta, \mathbf{H}_n, \mathcal{G}_n^s) \leqslant N_s(\delta/(s(2\sup_{g\in\mathcal{G}_n}\|g\|_{s,n})^{s-1}), \mathbf{H}_n, \mathcal{G}_n). \tag{2.35}$$

PROOF. For $a \geqslant b \geqslant 0$ $a^s - b^s \leqslant s(a-b)a^{s-1}$ for all $1 \leqslant s < \infty$. Using this and Hölder's inequality, we obtain that for all $g, \tilde{g} \in \mathcal{G}_n$

$$\begin{aligned}\int \left| |g|^s - |\tilde{g}|^s \right| d\mathbf{H}_n &\leqslant s \int \left| |g| - |\tilde{g}| \right| [\max(|g|, |\tilde{g}|)]^{s-1} d\mathbf{H}_n \\ &\leqslant s \int |g - \tilde{g}| [|g| + |\tilde{g}|]^{s-1} d\mathbf{H}_n \leqslant s \|g - \tilde{g}\|_{s,n} \| |g| + |\tilde{g}| \|_{s,n}^{s-1} \\ &\leqslant s \|g - \tilde{g}\|_{s,n} (2\sup_{g\in\mathcal{G}_n}\|g\|_{s,n})^{s-1}. \quad \square\end{aligned}$$

Hence, if $\sup_{g\in\mathcal{G}_n}\|g\|_{s,n}$ remains bounded, say

$$\sup_{g\in\mathcal{G}_n}\|g\|_{s,n} \leqslant K \tag{2.36}$$

with arbitrary large probability for all n sufficiently large, then $N_1(\delta, \mathbf{H}_n, \mathcal{G}_n^s)$ and $N_s(\delta, \mathbf{H}_n, \mathcal{G}_n)$ are of the same order of magnitude.

THEOREM 2.3.5. *Let $\{\mathcal{G}_n\}$ be a sequence of permissible classes with envelopes G_n satisfying*

$$\limsup_{n\to\infty} \|G_n\|_{s,(n)} < \infty, \quad 1 \leqslant s < \infty \tag{2.37}$$

Suppose that for some sequence $c_n \geqslant 1$, $c_n = o(n^{\frac{1}{s}})$

$$\limsup_{n\to\infty} \int_{G_n > c_n} G_n^s dH^{(n)} = 0 \tag{2.38}$$

$$\limsup_{n\to\infty} \frac{1}{c_n^s} \int_{G_n \leqslant c_n} G_n^{2s} dH^{(n)} = 0 \tag{2.39}$$

and

$$\frac{c_n^s}{n} \log N_s(\delta, \mathbf{H}_n, \mathcal{G}_n) \xrightarrow{\mathbf{P}^*} 0, \quad \textit{for all} \quad \delta > 0 \tag{2.40}$$

Then

$$\sup_{g\in\mathcal{G}_n} | \|g\|_{s,n} - \|g\|_{s,(n)} | \xrightarrow{\mathbf{P}} 0. \tag{2.41}$$

PROOF. Conditions (2.38) and (2.39) imply that

$$| \|G_n\|_{s,n} - \|G_n\|_{s,(n)} | \xrightarrow{\mathbf{P}} 0.$$

It now follows from (2.37) that for some $K < \infty$

$$\sup_{g\in\mathcal{G}_n}\|g\|_{s,n} \leqslant \|G_n\|_{s,n} \leqslant K$$

with arbitrary large probability for all n sufficiently large. Apply Lemma 2.3.4 to see that (2.40) implies

$$\frac{c_n^s}{n} \log N_1(\delta, \mathbf{H}_n, \mathcal{G}_n^s) \xrightarrow{\mathbf{P}^*} 0 \quad \text{for all} \quad \delta>0\,. \tag{2.42}$$

The conclusion of the theorem now follows easily from Theorem 2.3.2. □

If (2.37) is not fulfilled, one can check uniform convergence of $\|g\|_{s,n}$ to $\|g\|_{s,(n)}$ by verifying (2.42) directly.

2.4 Measurability I

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots$ be independent, identically distributed random variables, with distribution H on $\mathbb{R}^d$. As underlying probability space, we take the product space

$$(\Omega, \mathcal{E}, \mathbf{P}) = ((\mathbb{R}^d)^\infty, \mathcal{B}^\infty, H^\infty) \otimes (M, \mathcal{M}, Q))$$

where $(M, \mathcal{M}, Q)$ is some probability space on which some auxiliary random variables live (we need some additional space for randomization). Without loss of generality, $(\Omega, \mathcal{E}, \mathbf{P})$ is assumed to be complete. We observed that

$$\omega \mapsto \sup_{g \in \mathcal{G}} \int g d(\mathbf{H}_n - H)(\omega)$$

need not be measurable. Of course if $\mathcal{G}$ is a countable class of measurable functions, there are no problems. Suppose now that there exists a countable ${}_0\mathcal{G}$ such that

$$\mathbf{P}\left[\sup_{g \in \mathcal{G}} |\int g d(\mathbf{H}_n - H)| \neq \sup_{g \in {}_0\mathcal{G}} |\int g d(\mathbf{H}_n - H)|\right] = 0, \quad n \geqslant 1. \tag{2.43}$$

Then application of Theorem 2.2.1 to ${}_0\mathcal{G}$ yields

$$\sup_{g \in \mathcal{G}} |\int g d(\mathbf{H}_n - H)| \to 0 \quad \text{almost surely} \tag{2.44}$$

iff both

$$\sup_{g \in {}_0\mathcal{G}} |g| \in L^1(\mathbb{R}^d, H) \quad \text{and}$$

$$\frac{1}{n} \log N_1(\delta, \mathbf{H}_n, {}_0\mathcal{G}) \xrightarrow{\mathbf{P}} 0 \quad \text{for all} \quad \delta>0.$$

Now, suppose $\mathcal{G}$ is separable. The process $g \mapsto \int g d(\mathbf{H}_n - H)$ is called *stochastically separable* if there exists a countable ${}_0\mathcal{G} \subset \mathcal{G}$ such that for all closed $\tilde{\mathcal{G}} \subset \mathcal{G}$ and open $B \subset \mathbb{R}$

$$\int g d(\mathbf{H}_n - H) \in B \quad \text{for all} \quad g \in \tilde{\mathcal{G}} \cap {}_0\mathcal{G}$$

implies

$$\int g d(\mathbf{H}_n - H) \in B \quad \text{for all} \quad g \in \tilde{\mathcal{G}}$$

with probability one (GIHMAN and SKOROHOD (1974)). If $g \mapsto \int g d(\mathbf{H}_n - H)$ is stochastically separable, (2.43) holds.

Stochastic separability suffices for most practical purposes (DUDLEY (1984), Section 11.3). Note that it implies measurability of

$$\sup_{g \in \mathcal{G}} |\int g d(\mathbf{H}_n - H)|. \tag{2.45}$$

However, the proof of a uniform law of large numbers needs measurability of other quantities too. If one assumes that $\mathcal{G}$ is *nearly linearly supremum measurable* (ALEXANDER (1984), GINÉ and ZINN (1984)), measurability difficulties are overcome without the assumption of stochastic separability.

POLLARD (1984) introduces the concept of *permissibility*. A permissible class $\mathcal{G}$ is also nearly linearly supremum measurable, but need not result in stochastic separability of the process. We shall now copy the definition of permissibility - of a class of functions on $\mathbb{R}^d$ - from Pollard's book (POLLARD (1984), Appendix C). We say that $\mathcal{G}$ can be indexed by T if $\mathcal{G} = \{g(\cdot, t): t \in T\}$.

DEFINITION: $\mathcal{G}$ is permissible if $\mathcal{G}$ can be indexed by a separable metric space T such that

(i) $g(\cdot, \cdot)$ is $\mathcal{B} \otimes \mathcal{B}(T)$ - measurable on $\mathbb{R}^d \otimes T \to \mathbb{R}$ ($\mathcal{B}$ is the Borel σ-algebra on $\mathbb{R}^d$, $\mathcal{B}(T)$ the Borel σ-algebra on T),

(ii) T is an analytic subset of a compact metric space $\overline{T}$ (from which it inherits its metric and Borel σ-field).

POLLARD (1984) eleborates on the merits of assuming permissibility. He shows that (among other things) permissibility of $\mathcal{G}$ implies measurability of (2.45).

Note that if $\mathcal{G}$ is permissible, then so is $\{|g - \tilde{g}|: g, \tilde{g} \in \mathcal{G}\}$ (see (2.6)) and

$$\mathcal{G}^s = \{|g|^s: g \in \mathcal{G}\}, \quad 1 \leqslant s < \infty,$$

and also the class of truncated functions

$$(\mathcal{G})_C = \{\text{sign}(g)(|g| \wedge C): g \in \mathcal{G}\}, \quad C > 0.$$

The quantities $N_s(\delta, \mathbf{H}_n, \mathcal{G})$ still need not be measurable even if $\mathcal{G}$ is permissible. However, the use of outer-probabilities for statements about the possibly non-measurable covering numbers does not interfere with proving laws of large numbers.

Suppose now that $\mathbf{x}_{n,1}, \cdots, \mathbf{x}_{n,n}$ are independent random variables, $\mathbf{x}_{n,k}$ having distribution $H_{n,k}$, $k = 1, \cdots, n$, $n \geqslant 1$. For each n, we denote the underlying probability space by $(\Omega_n, \mathcal{E}_n, \mathbf{P}_n)$, and we shall assume that it is complete. Let $\{\mathcal{G}_n\}$ be a sequence of classes of measurable functions on $\mathbb{R}^d$. In order to handle measurability for the non i.i.d. case and triangular arrays, it suffices to assume permissibility of each $\mathcal{G}_n$. To see this, recall the proof of Theorem 2.3.2. Note that all probability statements are for fixed (sufficiently large, but nonrandom) n. For each n,

$$\sup_{g \in \mathcal{G}_n} |\int g d(\mathbf{H}_n - H^{(n)})|$$

is measurable, provided $\sup_{g \in \mathcal{G}_n} \|g\|_{1,(n)} < \infty$. POLLARD (1984) shows that for fixed n, the symmetrization device

$$\int g d(\mathbf{H}_n - H^{(n)}) \mapsto \int g d(\mathbf{H}_n - \mathbf{H}'_n)$$

is valid. Of course, if $\mathcal{G}_n$ is permissible, then $\{g(x)\sigma: g \in \mathcal{G}_n\}$ is a permissible class of functions on $\mathbb{R}^{d+1}$. This makes it possible to randomize the process. The use of Fubini's Theorem in (2.30) is thus legitimate.

3. Consistent least squares estimation

3.1. L^2-consistency

Consider the regression model

$$\mathbf{y} = g(\mathbf{x}) + \boldsymbol{\epsilon}$$

where $\mathbf{x}$ is a $\mathbb{R}^d$-valued random vector with distribution H, $\boldsymbol{\epsilon}$ is independent of $\mathbf{x}$ and has expectation zero and finite variance, and g is a member of a class $\mathcal{G}$ of regression functions on $\mathbb{R}^d$. For an estimator of the unknown g to be statistically meaningful, it should at least be consistent in some sense. In the least squares context the most natural requirement is *L^2-consistency*. In this chapter we show that entropy conditions on a (rescaled and truncated version of) $\mathcal{G}$ imply this type of consistency. The results from Chapter 2 are used to prove this.

Let $L^2(\mathbb{R}^d, H)$ be the Hilbert space of H-square integrable functions on $\mathbb{R}^d$. Writing K for the distribution of $\boldsymbol{\epsilon}$, let $L^2(\mathbb{R}^d \times \mathbb{R}, P)$ be the Hilbert space of measurable $P = H \times K$-square integrable functions on $\mathbb{R}^d \times \mathbb{R}$ with norm $\|\cdot\|_2$. For convenience, we omit the subscript 2, i.e. we write $\|\cdot\|$. Confusion is not likely, because from now on L^s-norms with $s \neq 2$ will only appear sporadically and then we shall use our old notation.

Denote by x and ϵ the first and second coordinate projections into $\mathbb{R}^d$ and $\mathbb{R}$ respectively, and write $g = g(x)$, $g_0 = g_0(x)$, $y = g_0 + \epsilon$, where we assume that g_0, the true state of nature, is in $L^2(\mathbb{R}^d, H)$. We have for $g \in L^2(\mathbb{R}^d, H)$

$$\|y - g\|^2 = \mathbb{E}(\mathbf{y} - g(\mathbf{x}))^2 = \|\epsilon\|^2 + \|g - g_0\|^2,$$

since $\mathbf{x}$ and $\boldsymbol{\epsilon}$ independent.

Let $(\mathbf{x}_1, \boldsymbol{\epsilon}_1)$, $(\mathbf{x}_2, \boldsymbol{\epsilon}_2)$, $\cdots$ be independent copies of $(\mathbf{x}, \boldsymbol{\epsilon})$ with $\mathbf{y}_k = g_0(\mathbf{x}_k) + \boldsymbol{\epsilon}_k$. Write $\mathbf{P}_n$ for the empirical distribution based on $(\mathbf{x}_1, \boldsymbol{\epsilon}_1), \cdots, (\mathbf{x}_n, \boldsymbol{\epsilon}_n)$ and $\mathbf{H}_n$ for the marginal empirical distribution generated by $\mathbf{x}_1, \cdots, \mathbf{x}_n$. Suppressing the subscript 2, we write $\|\cdot\|_n$ for the corresponding $L^2(\mathbb{R}^d \times \mathbb{R}, \mathbf{P}_n)$-norm:

$$\|g\|_n^2 = \frac{1}{n} \sum_{k=1}^{n} g(\mathbf{x}_k)^2,$$

$$\|y - g\|_n^2 = \frac{1}{n} \sum_{k=1}^{n} (\mathbf{y}_k - g(\mathbf{x}_k))^2 = \|\epsilon - (g - g_0)\|_n^2.$$

The least squares estimator $\hat{\mathbf{g}}_n$ is - not necessarily uniquely - defined by

$$\|y - \hat{\mathbf{g}}_n\|_n^2 = \inf_{g \in \mathcal{G}} \|y - g\|_n^2.$$

The estimator $\hat{\mathbf{g}}_n$ is *strongly $L^2(\mathbb{R}^d, H)$-consistent* if

$$\|\hat{\mathbf{g}}_n - g_0\| \to 0 \quad \text{almost surely.} \tag{3.1}$$

Strong $L^2(\mathbb{R}^d, \mathbf{H}_n)$-consistency is defined in a similar manner. We concentrate on convergence with respect to these metrics because the information on the

regression function is determined by the distribution of the data. The additional knowledge that $\hat{\mathbf{g}}_n$ is in a class of regression functions $\mathcal{G}$ can sometimes be used to prove consistency in, for instance, the sup-norm.

Observe that g_0 is the essentially unique minimizer of $\|y-g\|$, whereas $\hat{\mathbf{g}}_n$ minimizes the empirical counter part $\|y-g\|_n$. By the strong law, $\|y-g\|_n$ converges for each fixed $g \in L^2(\mathbb{R}^d, H)$ to $\|y-g\|$ almost surely, and if this convergence is uniform, consistency in both $\|\cdot\|$- and $\|\cdot\|_n$-norm follows almost immediately. The almost sure convergence, uniformly over a class of functions $\mathcal{G}$, was studied in the previous chapter. Recall Theorem 2.2.4. For the case $s=2$, it states that, for $\mathcal{G}$ a permissible class with envelope G,

$$\sup_{g \in \mathcal{G}} | \|g\|_n - \|g\| | \to 0 \quad \text{almost surely} \tag{3.2}$$

if the envelope condition

$$\int G^2 dH < \infty \tag{3.3}$$

and the entropy condition

$$\frac{1}{n} \log N_2(\delta, \mathbf{H}_n, \mathcal{G}) \xrightarrow{\mathbf{P}^*} 0 \quad \text{for all} \quad \delta > 0 \tag{3.4}$$

are fulfilled. Remember that $\log N_2(\delta, \mathbf{H}_n, \mathcal{G})$ is called the entropy of $\mathcal{G}$ for $\|\cdot\|_n$.

PROPOSITION 3.1.1. *Suppose that $\mathcal{G}$ is a permissible class with $g_0 \in \mathcal{G}$ and that (3.3) and (3.4) are fulfilled, then*

$$\|\hat{\mathbf{g}}_n - g_n\| \to 0 \quad \textit{almost surely},$$

as well as

$$\|\hat{\mathbf{g}}_n - g_n\|_n \to 0 \quad \textit{almost surely}.$$

PROOF. Obviously, conditions (3.3) and (3.4) ensure that we can apply Theorem 2.2.4 to the class $\{y-g: g \in \mathcal{G}\}$, so

$$\sup_{g \in \mathcal{G}} | \|y-g\|_n - \|y-g\| | \to 0 \quad \text{almost surely.}$$

Now, $\|y-g\|^2 = \|\epsilon\|^2 + \|g-g_0\|^2$, and since $g_0 \in \mathcal{G}$, $\|y-\hat{\mathbf{g}}_n\|_n^2 \leqslant \|\epsilon\|_n^2$. Hence, for arbitrary $\eta > 0$, and for all n sufficiently large

$$\|\epsilon\|^2 + \|\hat{\mathbf{g}}_n - g_0\|^2 \leqslant \|y-\hat{\mathbf{g}}_n\|_n^2 + \eta \leqslant \|\epsilon\|_n^2 + \eta \leqslant \|\epsilon\|^2 + 2\eta$$

almost surely. Or

$$\|\hat{\mathbf{g}}_n - g_0\|^2 \leqslant 2\eta \quad \text{almost surely.}$$

Thus $\|\hat{\mathbf{g}}_n - g_0\| \to 0$ almost surely, and since $\|g-g_0\|_n \to \|g-g_0\|$ almost surely, uniformly in $g \in \mathcal{G}$, this implies that also $\|\hat{\mathbf{g}}_n - g_0\|_n \to 0$ almost surely. □

The uniform convergence (3.2) is certainly not necessary for consistency and it is clear that condition (3.3) and (3.4) from empirical process theory will hardly ever be satisfied for a class of regression functions $\mathcal{G}$. For example, for $\mathcal{G}=\{g(x,\theta)=x\theta=\theta_1 x_1+\cdots+\theta_d x_d\colon \theta\in\mathbb{R}^d\}$ (3.3) and (3.4) do not hold. This partly due to the fact that $\mathcal{G}$ is a cone (i.e. if $g\in\mathcal{G}$ also $\alpha g\in\mathcal{G}$ for all $\alpha>0$). Therefore, we consider a class scaled functions

$$\mathcal{F} = \{f=\frac{g}{1+\|g\|}: g\in\mathcal{G}\}.$$

Then $\|f\|\leqslant 1$ for all $f\in\mathcal{F}$, and $\mathcal{F}$ is often essentially smaller that $\mathcal{G}$, e.g. if $\mathcal{G}$ is a cone. In smooth enough models, (3.3) and (3.4) will hold for $\mathcal{F}$. This is for instance the case in linear regression. However, the envelope condition on $\mathcal{F}$ still seems to rule out many interesting models. Therefore, we propose to weaken (3.3) to uniform square integrability of $\mathcal{F}$ and to impose the entropy condition on a class of truncated functions.

A class $\mathcal{F}$ is uniformly square integrable if

$$\lim_{C\to\infty}\sup_{f\in\mathcal{F}}\int_{|f|>C} f^2 dH = 0. \tag{3.5}$$

The class of truncated versions of functions in $\mathcal{F}$ is defined as before: i.e. let C be a positive number and denote

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

and $(\mathcal{F})_C=\{(f)_C\colon f\in\mathcal{F}\}$.

THEOREM 3.1.2. *Suppose that $\mathcal{G}$ is a permissible class with $g_0\in\mathcal{G}$, that $\mathcal{F}$ is uniformly square integrable and that for each $C>0$*

$$\frac{1}{n}\log N_2(\delta, \mathbf{H}_n, (\mathcal{F})_C) \xrightarrow{\mathbf{P}^*} 0 \quad \textit{for all} \quad \delta>0. \tag{3.6}$$

Then $\hat{\mathbf{g}}_n$ is strongly $L^2(\mathbb{R}^d, H)$-consistent.

PROOF. We shall first construct a covering set of the class

$$\mathcal{H}_C = \left\{ \left(\frac{\epsilon+g_0}{1+\|g\|}\right)_C - \left(\frac{g}{1+\|g\|}\right)_C : g\in\mathcal{G} \right\}.$$

Let $\mathbf{f}_j$, $j=1,2,\cdots,N_2(\delta, \mathbf{H}_n, (\mathcal{F})_C)$ be a covering set of $(\mathcal{F})_C$, i.e. for each $f=g/(1+\|g\|)\in\mathcal{F}$ there exists an $\mathbf{f}_j$ such that

$$\|(f)_C-\mathbf{f}_j\|_n<\delta. \tag{3.7}$$

For all $j=1,\cdots,N_2(\delta, \mathbf{H}_n(\mathcal{F})_C)$, define

$$\mathbf{h}_{j,k} = (k\delta(\epsilon+g_0))_C - \mathbf{f}_j, \quad k=0,1,\cdots,[1/\delta].$$

Then for all n sufficiently large, $\{\mathbf{h}_{j,k}: j=1,\cdots,N_2(\delta,\mathbf{H}_n,(\mathcal{F})_C), k=0,1,\cdots,[1/\delta]\}$ is a covering set of $\mathcal{K}_C$. To see this, choose $f=g/(1+\|g\|)$, $\mathbf{f}_j$ as in (3.7) and $k=[1/(\delta(1+\|g\|))]$. Then

$$\begin{aligned}
&\left\| \left[\frac{\epsilon+g_0}{1+\|g\|}\right]_C - \left[\frac{g}{1+\|g\|}\right]_C - \mathbf{h}_{j,k} \right\|_n \\
&\leqslant \left\| \left(\frac{1}{1+\|g\|} - k\delta\right)(\epsilon+g_0) \right\|_n + \left\| \left[\frac{g}{1+\|g\|}\right]_C - \mathbf{f}_j \right\|_n \\
&< \delta\|\epsilon+g_0\|_n + \delta \leqslant \delta\|\epsilon-g_0\| + 2\delta
\end{aligned}$$

almost surely, for n sufficiently large. Thus, we can apply Theorem 2.2.4 to $\mathcal{K}_C$, which yields that

$$\sup_{g\in\mathcal{G}} \left| \left\| \left[\frac{\epsilon+g_0}{1+\|g\|}\right]_C - \left[\frac{g}{1+\|g\|}\right]_C \right\| - \left\| \left[\frac{\epsilon+g_0}{1+\|g\|}\right]_C - \left[\frac{g}{1+\|g\|}\right]_C \right\|_n \right| \to 0 \tag{3.8}$$

almost surely, for all $C>0$.

Let $\eta>0$ be arbitrary. Then from (3.8) we have that for all $g\in\mathcal{G}$, $C>0$ and n sufficiently large

$$\begin{aligned}
&\left\| \left[\frac{\epsilon+g_0}{1+\|g\|}\right]_C - \left[\frac{g}{1+\|g\|}\right]_C \right\|^2 \\
&\leqslant \left\| \left[\frac{\epsilon+g_0}{1+\|g\|}\right]_C - \left[\frac{g}{1+\|g\|}\right]_C \right\|_n^2 + \eta \quad \text{almost surely.}
\end{aligned} \tag{3.9}$$

To get rid of the truncation in (3.9), we argue as follows. Obviously,

$$\left\| \left[\frac{\epsilon+g_0}{1+\|g\|}\right]_C - \left[\frac{g}{1+\|g\|}\right]_C \right\|_n^2 \leqslant \left\| \frac{\epsilon+g_0-g}{1+\|g\|} \right\|_n^2.$$

For the lefthand side of (3.9), we have

$$\begin{aligned}
&\left\| \left[\frac{\epsilon+g_0}{1+\|g\|}\right]_C - \left[\frac{g}{1+\|g\|}\right]_C \right\| \\
&\geqslant \left\| \frac{\epsilon+g_0-g}{1+\|g\|} \right\| - \left\| \left[\frac{g}{1+\|g\|}\right]_C - \frac{g}{1+\|g\|} \right\| - \left\| \left[\frac{\epsilon+g_0}{1+\|g\|}\right]_C - \frac{\epsilon+g_0}{1+\|g\|} \right\|.
\end{aligned} \tag{3.10}$$

Because of the assumed uniform square integrability, $\|(g/(1+\|g\|))_C - g/(1+\|g\|)\|$ can be made arbitrary small by taking C

sufficiently large. Moreover, $\|\epsilon+g_0\|$ is finite, so $\{(\epsilon+g_0)/(1+\|g\|: g\in\mathcal{G}\}$ is also uniformly square integrable. Hence, for C large enough

$$\| \left[\frac{\epsilon+g_0}{1+\|g\|}\right]_C - \left[\frac{g}{1+\|g\|}\right]_C \|^2 \geqslant \|\frac{\epsilon+g_0-g}{1+\|g\|}\|^2 - \eta.$$

Thus, (3.9) implies that for n sufficiently large

$$\|\frac{\epsilon+g_0-g}{1+\|g\|}\|^2 \leqslant \|\frac{\epsilon+g_0-g}{1+\|g\|}\|_n^2 + 2\eta \quad \text{almost surely.}$$

Since ϵ and $\mathbf{x}$ are independent, this can be written as

$$\begin{aligned} &\|\epsilon\|^2 + \|g-g_0\|^2 \qquad (3.11)\\ &\leqslant \|\epsilon+g_0-g\|_n^2 + 2\eta(1+\|g\|)^2 \quad \text{almost surely,} \end{aligned}$$

for all $g\in\mathcal{G}$.

For $\hat{\mathbf{g}}_n$, we have

$$\|\epsilon+g_0-\hat{\mathbf{g}}_n\|_n^2 \leqslant \|\epsilon\|_n^2,$$

because $g_0\in\mathcal{G}$. Hence (3.11) implies that for all n sufficiently large

$$\begin{aligned} \|\epsilon\|^2 + \|\hat{\mathbf{g}}_n - g_0\|^2 &\leqslant \|\epsilon\|_n^2 + 2\eta(1+\|\hat{\mathbf{g}}_n\|)^2 \\ &\leqslant \|\epsilon\|^2 + 3\eta(1+\|\hat{\mathbf{g}}_n\|)^2 \quad \text{almost surely,} \end{aligned}$$

or

$$\|\frac{\hat{\mathbf{g}}_n - g_0}{1+\|\hat{\mathbf{g}}_n\|}\|^2 \leqslant 3\eta \quad \text{almost surely.}$$

Since η was arbitrary we can take $3\eta<1$. But then $((\|g_0-\hat{\mathbf{g}}_n\|)/(1+\|\hat{\mathbf{g}}_n\|))^2<1$ for all n sufficiently large implies that for some constant $K<\infty$

$$\|\hat{\mathbf{g}}_n\| \leqslant K$$

for all n sufficiently large.
This yields

$$\|g_0-\hat{\mathbf{g}}_n\|^2 \leqslant 3\eta(1+K)^2 \quad \text{almost surely,}$$

which completes the proof. □

It is easy to see that the conditions of Theorem 3.1.2 are implied by those of Proposition 3.1.1, but that in general they do not imply $L^2(\mathbb{R}^d, \mathbf{H}_n)$-consistency. Consistency properties of regression estimators for more specific models have been studied by other authors. In nonlinear regression, $\mathcal{G}$ is a class of functions of the form $\{g(x, \theta): \theta\in\Theta\}$ with Θ some metric space and $g(x, \theta)$ continuous is θ for H-almost all x. It is shown in Section 3.2 that condition (3.6) is fulfilled for this $\mathcal{G}$ if Θ is compact. JENNRICH (1969) proves consistency

under the assumption that Θ is compact and that the envelope condition on $\mathcal{G}$ holds:

$$\int \sup_{\theta \in \Theta} |g(x, \theta)|^2 dH(x) < \infty.$$

HUBER (1967) imposes an envelope condition on a rescaled version of $\mathcal{G}$. He allows for more general scale transformations, but there appears to be not much loss of generality if we restrict ourselves to the choice of $\mathcal{F}$. If the envelope F of $\mathcal{F}$ belongs to $L^2(\mathbb{R}^d, H)$, then it can be shown that if (3.6) holds, $\hat{\mathbf{g}}_n$ is also strongly $L^2(\mathbb{R}^d, \mathbf{H}_n)$-consistent. Moreover, the truncation device becomes redundant.

In nonparametric regression, there is usually no parametrization such that the regression functions are continuous in the parameter for H-almost all x. In Theorem 3.2, this continuity assumption is not required. The relation with the assumption of compactness of parameter space is made clear in the following lemma. Remember that a class $\mathcal{F}$ is called totally bounded for $\|\cdot\|$ if for all $\delta > 0$ the δ-entropy $\log N_2(\delta, H, \mathcal{F})$ with respect to the $L^2(\mathbb{R}^d, H)$-norm, is finite. The closure of a totally bounded $\mathcal{F}$ is compact.

LEMMA 3.1.3 *The conditions of Theorem 3.1.2 imply that $\mathcal{F}$ is totally bounded for $\|\cdot\|$. Moreover, if $\mathcal{F}$ is totally bounded for $\|\cdot\|$, then $\mathcal{F}$ is uniformly square integrable.*

PROOF. In view of condition (3.6), application of Lemma 2.2.5 to $(\mathcal{F})_C$ yields that $(\mathcal{F})_C$ is totally bounded for $\|\cdot\|$. The uniform square integrability now gives that $\mathcal{F}$ is also totally bounded. This proves the first assertion.

Suppose now that $\mathcal{F}$ is totally bounded for $\|\cdot\|$. Let δ be arbitrary and let $f_1, \ldots, f_m$, $m = 1, \ldots, N_2(\delta, H, \mathcal{F})$, be a δ-covering set of $\mathcal{F}$. Then for C sufficiently large

$$\max_{j=1,\ldots,m} \|(f_j)_C - f_j\| \leq \delta.$$

Furthermore, for $f \in \mathcal{F}$, $\|f - f_j\| \leq \delta$

$$\|(f)_C - f\| \leq \|(f)_C - (f_j)_C\| + \|(f_j)_C - f_j\|$$
$$+ \|f_j - f\| \leq 2\|f - f_j\| + \|(f_j)_C - f_j\| \leq 3\delta.$$

It follows that

$$\lim_{C \to \infty} \sup_{f \in \mathcal{F}} \|(f)_C - f\| = 0.$$

This is equivalent to uniform square integrability. □

So far we did not consider classes of regression functions depending on n, $\mathcal{G}_n$ say. Such a situation arises for instance in spline regression, nearest neighbour regression and some other nonparametric regression models. The situation with $\mathcal{G}_n$ depending on n will be treated in detail in Section 3.3. Here, we maintain the assumption of i.i.d. random variables, but because of the practical

importance we consider a simple application of Lemma 2.3.3. Suppose $\{\mathcal{G}_n\}$ is a permissible sequence, then Lemma 2.3.3 asserts that

$$\frac{1}{n} \log N_2(\delta, \mathbf{H}_n, (\mathcal{G}_n)_C) \xrightarrow{\mathbf{P}^*} 0 \quad \text{for all} \quad \delta>0 \tag{3.12}$$

implies

$$\sup_{g \in \mathcal{G}_n} |\|(g)_C\|_n - \|(g)_C\|| \xrightarrow{\mathbf{P}} 0.$$

Note that the convergence is now in probability (almost sure results can only be obtained if the entropy remains small). It is now not difficult to adjust Theorem 3.1.2 to this situation, assuming uniform square integrability of $\cup \mathcal{F}_n$, $\mathcal{F}_n = \{g/(1+\|g\|): g \in \mathcal{G}_n\}$, together with (3.12) for $(\mathcal{F}_n)_C$, $C>0$.

3.2 Applications

In this section we shall concentrate on conditions for the entropy condition (3.6) on $(\mathcal{F})_C$ to hold. The technique to prove the lemmas is construction of a covering set and some combinatorics to count the number of elements. The uniform square integrability of $\mathcal{F}$ imposes requirements on the (unknown) H. Often, it has to be shown by separate means that $\hat{\mathbf{g}}_n/(1+\|\hat{\mathbf{g}}_n\|)$ is eventually in a totally bounded subset of $\mathcal{F}$ (see e.g. HUBER (1967)). To avoid digressions, we shall not elaborate on the uniform square integrability condition for specific situations, but only highlight that (3.6) is a common feature of regression models.

An important special class of functions, that appears in several applications, is the collection of indicator functions of VC-classes of sets. A minor modification of Theorem 2.2.6 says that for a VC-class of sets, and more generally, for a VC-graph class $\mathcal{F}$ of functions

$$N_2(\delta, Q, (\mathcal{F})_C) \leqslant AC^r\delta^{-r} \quad \text{for all} \quad \delta>0$$

where A and r are constants not depending on Q. Examples of VC-graph classes will be given below.

3.2.1. Nonlinear regression. If the functions in $\mathcal{G}$ form a (subset of a)finite-dimensional vector space, then both $\mathcal{G}$ and $\mathcal{F}$ are VC-graph classes (see POLLARD (1984, Ch. II, Lemma 28), DUDLEY (1984)). This is a consequence of the fact that the collection of half-spaces is a VC-class. Here is one more example where the regression functions form a VC-graph class.

EXAMPLE. A model considered in BARD (1974) is

$$\mathbf{y} = \exp(-\boldsymbol{\theta}_1 \mathbf{x}_1 e^{-\boldsymbol{\theta}_2 \mathbf{x}_2}) + \epsilon, \quad \theta_i \geqslant 0, \quad \mathbf{x}_i \geqslant 0, \quad i=1,2.$$

The graphs are of the form

$$\{(x_1,x_2,t): 0 \leqslant t \leqslant \exp(\theta_1 x_1 e^{-\theta_2 x_2}),\ \theta_i \geqslant 0,\ x_i \geqslant 0,\ i=1,2\}$$

$$= \{(x_1,x_2,t): \log\log \frac{1}{t} \geqslant \log\theta_1 + \log\theta_1 - \theta_2 x_2,\ \theta_i \geqslant 0,\ x_i \geqslant 0,\ i=1,2\}.$$

Thus (use Theorem 9.2.2 of DUDLEY (1984)) $\mathcal{G}$ is a *VC*-graph class and since $\mathcal{G}$ is uniformly bounded, this implies that $\mathcal{F}$ satisfies (3.6).

EXAMPLE. The p-compartment model

$$\mathbf{y} = \sum_{i=1}^{p} \alpha_i e^{\lambda_i \mathbf{x}} + \epsilon, \ \alpha_i \geqslant 0, \ \lambda_i \geqslant 0, \ i = 1, \cdots, p, \ \mathbf{x} \geqslant 0.$$

If $p = 1$, the class of regression functions $\mathcal{G}$ forms a *VC*-graph class, so then we have for some A and r

$$N_2(\delta, \mathbf{H}_n, (\mathcal{G})_C) \leqslant AC^r \delta^{-r}, \quad 0 < \delta < 1.$$

This yields for the case $p \neq 1$ (apply the triangle inequality)

$$N_2(\delta, \mathbf{H}_n, (\mathcal{G})_C) \leqslant \left[AC^r \left(\frac{\delta}{p}\right)^{-r} \right]^p,$$

and since $\mathcal{G}$ is a cone, the same holds for the $(\mathcal{F})_C$.

In general, let $\mathcal{G} = \{g(\cdot, \theta): \theta \in \Theta\}$, with $(\Theta, \|\cdot\|)$ some metric space. If $\mathcal{F}$ is not a *VC*-graph class, one can handle the entropy condition by assuming compactness of the parameter space.

LEMMA 3.2.1. *Suppose that $g(x, \theta)$ is continuous in θ for H-almost all x, and that $(\Theta, \|\cdot\|)$ is compact. Then for all $C > 0$, $\delta > 0$*

$$\frac{1}{n} \log N_2(\delta, \mathbf{H}_n, (\mathcal{G})_C \xrightarrow{\mathbf{P}^*} 0$$

as well as

$$\frac{1}{n} \log N_2(\delta, \mathbf{H}_n, (\mathcal{F})_C) \xrightarrow{\mathbf{P}^*} 0.$$

PROOF. The proof shows that for all $\delta > 0$ there exists a finite δ-bracketing -set, i.e. a set of functions $\{g_j^{(L)}, g_j^{(R)}\}$ such that for each $g \in \mathcal{G}$ there exists a pair $[g_j^{(L)}, g_j^{(R)}]$ with $g_j^{(L)} \leqslant (g)_C \leqslant g_j^{(R)}$ and $\|g_j^{(L)} - g_j^{(R)}\| < \delta$ (see DEHARDT (1971)).

Define for all $x \in \mathbb{R}^d$, $\theta \in \Theta$

$$w(x, \theta, \rho) = \sup_{(\tilde{\theta}: \|\theta - \tilde{\theta}\| \leqslant \rho)} |(g(x, \tilde{\theta}))_C - (g(x, \theta))_C|.$$

Then

$$\lim_{\rho \to 0} w(x, \theta, \rho) = 0$$

for every θ and H-almost all x. Since $(g(x, \theta))_C \leqslant C$ for all x, dominated convergence implies that also

$$\lim_{\rho \to 0} \|w(\cdot, \theta, \rho)\|^2 = 0.$$

Hence for arbitrary $\delta > 0$ there exists a finite covering set of Θ by balls with

radius ρ_i and centres θ_i, such that

$$\|w(\cdot,\theta_i,\rho_i)\|^2 < \tfrac{1}{2}\delta^2.$$

For all n sufficiently large, also

$$\|w(\cdot,\theta_i,\rho_i)\|_n^2 < \delta^2.$$

But then $\{(g(\cdot,\theta_i))_C\}$ is a finite covering set of $(\mathcal{G})_C$ with $L^2(\mathbf{R}^d, \mathbf{H}_n)$-norm:

$$\|(g(\cdot,\theta))_C - (g(\cdot,\theta_i))_C\|_n \leqslant \|w(\cdot,\theta_i,\rho_i)\|_n < \delta,$$

for all $\|\theta-\theta_i\|<\rho_i$.

In the same way, one can construct a finite covering ser of $\mathcal{F}$, since the class $\{\alpha g\colon \alpha\in[0,1],\ g\in\mathcal{G}\}$ also satisfies the assumptions of Lemma 3.2.1. □

If the regression functions are not continuous in θ, one can often split them up into continuous parts. An example is *multi-phase* regression, which is treated in detail in Section 3.4.

In the next three applications $\mathcal{G}$ is always a cone. Thus, to check the entropy condition for the $(\mathcal{F})_C$ it certainly suffices to verify the entropy condition for the $(\mathcal{G})_C$. In the proofs, the order symbol $\mathcal{O}(\cdot)$ holds for $n\to\infty$.

3.2.2. Monotone functions (isotonic regression).

LEMMA 3.2.2. *Let* $\mathcal{G}=\{g\colon \mathbf{R}\to\mathbf{R},\ g \text{ is increasing}\}$, *then for all* $\delta>0$, $C>0$

$$\frac{1}{n}\log N_2(\delta, \mathbf{H}_n, (\mathcal{G})_C) \xrightarrow{\mathbf{P}^*} 0.$$

PROOF. For $g\in\mathcal{G}$, define $k=[C/\delta]$ and $A^{(i)}=\{x\colon i\delta\leqslant(g(x))_C<(i+1)\delta\}$, for $i=-(k+1), -k, \cdots, k$. Take $g^{(i)}=i\delta$ and approximate $(g)_C$ by $\Sigma_i g^{(i)}1_{A^{(i)}}$. The $\{A^{(i)}\}$ form a partition of $\mathbf{R}$ with $T=2(k+1)$ elements. As g varies, the $A^{(i)}$ are in a class $\mathcal{A}^{(i)}$ of intervals, for which

$$\Delta^{\mathcal{A}^{(i)}}(\mathbf{x}_1, \cdots, \mathbf{x}_n) = \mathcal{O}(n^2).$$

Thus, we have $\mathcal{O}(n^{2T})$ functions of the type $\Sigma_i g^{(i)}1_{A^{(i)}}$. Also,

$$\sup_x |(g(x))_C - \sum_i g^{(i)}(x)1_{A^{(i)}}| < \delta.$$

Thus,

$$N_\infty(\delta, \mathbf{H}_n(\mathcal{G})_C) = \mathcal{O}(n^{2T}). \quad \square$$

The result can be extended to functions of bounded variation and unimodal functions. If $d>1$, further conditions are in general necessary to make sure that the entropy condition is fulfilled, e.g. assumptions on H or the condition that $\mathcal{G}$ is a class of distribution functions of bounded Stieltjes-Lebesgue measures.

3.2.3. Smooth functions. Let $\mathcal{G}_n$, $n \geqslant 1$, be a sequence of classes such that the elements of $\bigcup \mathcal{G}_n$ have all partial derivatives of order $s \leqslant m$, $m \geqslant 0$.

LEMMA 3.2.3. *For $x \in \mathbb{R}^d$, let $\|x\|$ denote the Euclidean norm of x. Suppose there exists an $\alpha \leqslant 1$ and*

$$L_n = o(n^{\frac{m+\alpha}{d}})$$

such that

$$|g^{(m)}(x) - g^{(m)}(\tilde{x})| \leqslant L_n \|x - \tilde{x}\|^{\alpha}$$

for all x, $\tilde{x}$, $g \in \mathcal{G}_n$. There for all $\delta > 0$, $C > 0$

$$\frac{1}{n} \log N_2(\delta, \mathbf{H}_n, (\mathcal{G})_C) \xrightarrow{\mathbf{P}^*} 0.$$

PROOF. Without loss of generality we can assume that H has compact support K. If this is not the case, take a K with $H(K) > 1 - \delta^2/C^2$. Then for any g

$$\|(g 1_K)_C - (g)_C\|_n \leqslant C(1 - \mathbf{H}_n(K))^{1/2} \to C(1 - H(K))^{1/2} < \delta,$$

almost surely. Let $\{B^{(i)}\}$ be a covering of K by balls with centres $x^{(i)}$ and radius $m!(\delta/L_n)^{1/m+\alpha}$. The number of balls needed is $\mathcal{O}(L_n/\delta)^{d/m+\alpha}$.
Construct from the $\{B^{(i)}\}$ a partition $\{A^{(i)}\}$ of K, e.g. take $A^{(i)} = \{x \in B^{(i)},\ x \notin B^{(j)},\ j < i\}$.

Let $g \in \mathcal{G}_n$ be arbitrary, and expand $g(x)$ for $x \in A^{(i)}$ in a Taylor series around $x^{(i)}$,

$$g(x) = g^{(i)}(x) + R^{(i)}(x), \quad x \in A^{(i)},$$

where $g^{(i)}(x)$ is the m-th order Taylor expansion. The Lipschitz condition tells us that

$$|R^{(i)}(x)| \leqslant L_n/m! \|x - x^{(i)}\|^{m+\alpha} < \delta.$$

Thus we have that

$$\sup_x |(g(x))_C - (\sum_i (g^{(i)}(x))_C 1_{A^{(i)}}(x))| < \delta.$$

As g varies in $\mathcal{G}_n$, the $g^{(i)}$ form a class of polynomials of fixed degree, $\mathcal{G}$ say. This class is a finite-dimensional vector space, so there exist constants A and r such the for arbitrary measure Q

$$N_2(\delta, Q, (\mathcal{G})_C) \leqslant AC^r \delta^{-r}.$$

For each i with $\mathbf{H}_n(A^{(i)}) \neq 0$ we make the following choice for Q

$$\mathbf{Q} = \mathbf{Q}_n^{(i)} = \frac{\mathbf{H}_n}{\mathbf{H}_n(A^{(i)})}, \quad \text{on } A^{(i)}.$$

This shows that there is a covering set $\{g_j^{(i)}\}$ of $(\mathcal{G})_C$ with at most $AC^r\delta^{-r}$

elements, such that for arbitrary $g^{(i)} \in \mathcal{G}$ there is a $\mathbf{g}_{j_i}^{(i)}$ with

$$\|(g^{(i)})_C 1_{A^{(i)}} - \mathbf{g}_{j_i}^{(i)} 1_{A^{(i)}}\|_n^2 = \int_{A^{(i)}} |(g^{(i)})_C - \mathbf{g}_{j_i}^{(i)}|^2 d\mathbf{H}_n$$

$$= \mathbf{H}_n(A^{(i)}) \int |(g^{(i)})_C - g_{j_i}^{(i)}|^2 d\mathbf{Q}_n^{(i)} < \mathbf{H}_n(A^{(i)}) \delta^2, \quad \mathbf{H}_n(A^{(i)}) \neq 0.$$

But then

$$\|\sum_i (g^{(i)})_C 1_{A^{(i)}} - \sum_i \mathbf{g}_{j_i}^{(i)} 1_{A^{(i)}}\|_n^2 = \sum_{i:\, \mathbf{H}_n(A^{(i)}) \neq 0} \mathbf{H}_n(A^{(i)}) \int |(g^{(i)})_C - \mathbf{g}_{j_i}^{(i)}|^2 d\mathbf{Q}_n^{(i)} < \delta^2$$

and

$$\|(g)_C - \sum_i \mathbf{g}_{j_i}^{(i)} 1_{A^{(i)}}\|_n < 2\delta.$$

Hence, the functions $\{\sum_i \mathbf{g}_{j_i}^{(i)} 1_{A^{(i)}}\}$ form a 2δ-covering set of $(\mathcal{G}_n)_C$. The number of different functions in this covering set is

$$\mathcal{O}\left((AC^r \delta^{-r})^{\mathcal{O}(\frac{L_n}{\delta})^{\frac{d}{m+\alpha}}} \right)$$

i.e.

$$\frac{1}{n} \log N_2(\delta, \mathbf{H}_n, (\mathcal{G}_n)_C) = \mathcal{O}(\frac{1}{n} L_n^{\frac{d}{m+\alpha}}) = o(1). \quad \square$$

If the functions in $\mathcal{G}_n$ are uniformly bounded and H has compact support, then $\mathcal{G}_n$ is totally bounded with respect to the sup-norm (see KOLMOGOROV and TIKHOMIROV (1959)). In our situation, $\mathcal{G}_n$ need not be uniformly bounded. The functions in $(\mathcal{G}_n)_C$ no longer have m derivatives, except in the case $m = 0$.

The result of Lemma 3.2.3 can be applied in penalized least squares. Let $d = 1$ and let the penalized least squares estimator $\tilde{\mathbf{g}}_n$ be obtained by minimizing

$$\|y - g\|_n^2 + \lambda_n^2 J(g),$$

where $J(g)$ is the penalty

$$J(g) = \int (g^{(m+1)}(x))^2 dx, \quad m \geq 0$$

(see e.g. WAHBA (1984)). We use Lemma 3.2.3 with $d = 1$ and $\alpha = 1$ to establish the following.

LEMMA 3.2.4. *Suppose $J(g_0) < \infty$ and $n^{m+1}\lambda_n \to \infty$, then there exists a sequence $\mathcal{G}_n$ such that $\tilde{\mathbf{g}}_n \in \mathcal{G}_n$ almost surely for all n sufficiently large, and such that for all $\delta > 0$, $C > 0$*

$$\frac{1}{n} \log N_2(\delta, \mathbf{H}_n, (\mathcal{G}_n)_C) \xrightarrow{\mathbf{P}^*} 0.$$

PROOF. The penalized least squares estimator $\tilde{\mathbf{g}}_n$ has $2m$ continuous derivatives (see WAHBA (1984)). We have

$$|\tilde{\mathbf{g}}^{(m)}(x)-\tilde{\mathbf{g}}_n^{(m)}(\tilde{x})| \leqslant J^{1/2}(\tilde{\mathbf{g}}_n)\|x-\tilde{x}\|$$

(see IBRAGIMOV and HAS'MINSKII (1981, page 81)). Also

$$\|y-\tilde{\mathbf{g}}_n\|_n^2+\lambda_n^2 J(\tilde{\mathbf{g}}_n) \leqslant \|\epsilon\|_n^2+\lambda_n^2 J(g_0),$$

which implies that for all n sufficiently large,

$$J^{1/2}(\tilde{\mathbf{g}}_n) \leqslant 2\frac{\|\epsilon\|}{\lambda_n}+J^{1/2}(g_0)$$

almost surely. Take

$$\mathcal{G}_n = \{g: \sup_{x,\tilde{x}}\|g^{(m)}(x)-g^{(m)}(\tilde{x})\| \leqslant L_n\|x-\tilde{x}\|\}$$

with $L_n=2\|\epsilon\|/\lambda_n+J^{1/2}(g_0)=o(n^{m+1})$ and apply Lemma 3.2.3 with $\alpha=1$ and $d=1$. □

3.2.4. Nearest neighbour regression. We consider the nearest neighbour regression estimator of the form

$$\hat{\mathbf{g}}_n = \sum_{i=1}^{p_n} \mathbf{g}_n^{(i)} 1_{\mathbf{A}_n^{(i)}}$$

where the $\mathbf{g}_n^{(i)}$ are polynomials of fixed degree and $\mathbf{A}_n^{(i)}$, $i=1, \cdots, p_n$ forms a random partition of $\mathbb{R}^d$. For instance, one may take the $\mathbf{A}_n^{(i)}$ as the set containing the $N=[n/p_n]$ nearest neighbours of some $\mathbf{x}_k$. In general, let

$$\mathcal{G}_n = \{\sum_{i=1}^{p_n} g^{(i)}\mathbf{1}_{A_n^{(i)}}: g^{(i)} \in \mathcal{G},\ A_n^{(i)} \in \mathcal{A},$$

$$A_n^{(i)} \cap A_n^{(j)} = \varnothing,\ i \neq j,\ \bigcup_{i=1}^{p_n} A_n^{(i)} = \mathbb{R}^d\}. \tag{3.13}$$

In a sense, this is an extension of a p-phase regression model to p_n-phase regression.

LEMMA 3.2.5. *Suppose that in (3.13) $\mathcal{G}$ is a VC-graph class and $\mathcal{A}$ a VC-class, and that $p_n=o(n/\log n)$, then for all $\delta>0$, $C>0$*

$$\frac{1}{n}\log N_2(\delta, \mathbf{H}_n, (\mathcal{G}_n)_C) \xrightarrow{\mathbf{P}^*} 0.$$

PROOF. Since $\mathcal{G}$ is a *VC*-graph class, we have

$$N_2\left(\frac{\delta}{p_n}, \mathbf{H}_n, (\mathcal{G})_C\right) \leqslant AC^r\left[\frac{\delta}{p_n}\right]^{-r}$$

for some constants A and r.

Let $\{\mathbf{g}_j\}$ be a (δ/p_n)-covering class of $(\mathcal{G})_C$, such that for arbitrary $g^{(i)} \in \mathcal{G}$ there is a $\mathbf{g}_j \in \{\mathbf{g}_j\}$ such that

$$\|(g^{(i)})_C - \mathbf{g}_{j_i}\|_n < \frac{\delta}{p_n}.$$

Then

$$\|\sum_{i=1}^{p_n} (g^{(i)})_C 1_{A^{(i)}} - \sum_{i=1}^{p_n} \mathbf{g}_{j_i} 1_{A^{(i)}}\|_n \leqslant \sum_{i=1}^{p_n} \|g^{(i)})_C - \mathbf{g}_{j_i}\|_n < \delta.$$

For a fixed partition $A^{(1)}, \cdots, A^{(p_n)}$, there are at most $(AC^r(\delta/p_n)^{-r})^{p_n}$ different functions of the type $\Sigma_{i=1}^{p_n} \mathbf{g}_{j_i} 1_{A^{(i)}}$. Since $\mathcal{C}$ is a VC-class,

$$\Delta^{\mathcal{C}}(\mathbf{x}_1, \cdots, \mathbf{x}_n) = \mathcal{O}(n^s)$$

for some $s \geqslant 0$. Thus the number of $L^\infty(\mathbb{R}^d, \mathbf{H}_n)$-different partitions is $\mathcal{O}(n^{sp_n})$. The total number of $L^\infty(\mathbb{R}^d, \mathbf{H}_n)$-different functions $\Sigma_{i=1}^{p_n} \mathbf{g}_{j_i} 1_{A^{(i)}}$ is thus

$$\left[AC^r\left(\frac{\delta}{p_n}\right)^{-r}\right]^{p_n} \mathcal{O}(n^{sp_n}).$$

And $1/n \log N_2(\delta, \mathbf{H}_n, (\mathcal{G}_n)_C) = \mathcal{O}(1/n\, p_n \log(np_n)) = o(1)$. □

3.3. The non-i.i.d. case and triangular arrays

In this section, we assume that for each n, $\mathbf{x}_{n,1}, \cdots, \mathbf{x}_{n,n}$ are independent random vectors in $\mathbb{R}^d$, $\mathbf{x}_{n,k}$ having distribution $H_{n,k}$. Furthermore, $\epsilon_{n,1}, \cdots, \epsilon_{n,n}$ are independent random variables with distribution $K_{n,k}$, $\mathbb{E}\epsilon_{n,k} = 0$, $k = 1, \cdots, n$, and $\{\epsilon_{n,1}, \cdots, \epsilon_{n,n}\}$ is independent of $\{\mathbf{x}_{n,1}, \cdots, \mathbf{x}_{n,n}\}$. We observe $(\mathbf{x}_{n,k}, \mathbf{y}_{n,k})$, $k = 1, \cdots, n$, where

$$\mathbf{y}_{n,k} = g_{0,n}(\mathbf{x}_{n,k}) + \epsilon_{n,k}, \quad k = 1, \cdots, n$$

and where $g_{0,n}$ is a member of a class $\mathcal{G}_n$ of regression functions. The least squares estimator $\hat{\mathbf{g}}_n$ is defined as a solution of the minimization problem

$$\inf_{g \in \mathcal{G}_n} \frac{1}{n} \sum_{k=1}^{n} (\mathbf{y}_{n,k} - g(\mathbf{x}_{n,k}))^2.$$

As in Section 3.1, $\mathbf{P}_n$ denotes the empirical measure based on $(\mathbf{x}_{n,1}, \epsilon_{n,1}), \cdots, (\mathbf{x}_{n,n}, \epsilon_{n,n})$ and $\mathbf{H}_n$ is the empirical measure generated by $\mathbf{x}_{n,1}, \cdots, \mathbf{x}_{n,n}$. Moreover, we write

$$H^{(n)} = \frac{1}{n} \sum_{k=1}^{n} H_{n,k},$$

$$P^{(n)} = \frac{1}{n} \sum_{k=1}^{n} P_{n,k} = \frac{1}{n} \sum_{k=1}^{n} H_{n,k} \times K_{n,k}$$

$$K^{(n)} = \frac{1}{n} \sum_{k=1}^{n} K_{n,k}.$$

The theoretical norm on $L^2(\mathbb{R}^d \times \mathbb{R}, P^{(n)})$ is denoted by $\|\cdot\|_{(n)}$, i.e. for $g \in L^2(\mathbb{R}^d, H^{(n)})$

$$\|g\|_{(n)}^2 = \int |g|^2 dH^{(n)},$$

and for g, $g_{0,n} \in L^2(\mathbb{R}^d, H^{(n)})$ and $\int |\epsilon|^2 dK^{(n)}(\epsilon) < \infty$

$$\begin{aligned} \|y - g\|_{(n)}^2 &= \int |\epsilon + g_{0,n}(x) - g(x)|^2 dP^{(n)}(x, \epsilon) \\ &= \|\epsilon\|_{(n)}^2 + \|g - g_{0,n}\|_{(n)}^2. \end{aligned}$$

The empirical norm on $L^2(\mathbb{R}^d \times \mathbb{R}, \mathbf{P}_n)$ is denoted again by $\|\cdot\|_n$, e.g.

$$\|g\|_n^2 = \int |g|^2 d\mathbf{H}_n,$$

$$\|y - g\|_n^2 = \int |\epsilon + g_{0,n}(x) - g(x)|^2 d\mathbf{P}_n(x, \epsilon) = \|\epsilon + g_{0,n} - g\|_n^2.$$

Finally, the class $\mathcal{F}_n$ of rescaled functions is defined as

$$\mathcal{F}_n = \{g/(1 + \|g\|_{(n)}) : g \in \mathcal{G}_n\}.$$

Throughout this section, we assume that $\|\epsilon\|_{(n)}$ as well as $\|g_{0,n}\|_{(n)}$ remain bounded. Moreover, we shall impose conditions that ensure that $|\|\epsilon\|_{(n)} - \|\epsilon\|_n|$ and $|\|g_{0,n}\|_{(n)} - \|g_{0,n}\|_n|$ converge to zero in probability. We impose an entropy condition on the class of truncated functions $(\mathcal{F}_n)_{C_n}$, $C_n = \sqrt{b_n}$, $b_n \geqslant 1$, $b_n = o(n^{1/2})$, endowed with $L^2(\mathbb{R}^d, \mathbf{H}_n)$-norm, as well as on $(\mathcal{F}_n)_{C_n}^2 = \{f^2 \wedge C_n^2 : f \in \mathcal{F}_n\}$ endowed with $L^1(\mathbb{R}^d, \mathbf{H}_n)$-norm. Recall Lemma 2.3.4, where a relation between these covering numbers is presented.

THEOREM 3.3.1. *Suppose that $\{\mathcal{G}_n\}$ is a sequence of permissible classes with $g_{0,n} \in \mathcal{G}_n$, $n \geqslant 1$. Assume that for some sequence $\{b_n\}$, $b_n \geqslant 1$, $b_n = o(n^{1/2})$*

$$\frac{b_n^2}{n} \log N_2(\delta, \mathbf{H}_n, (\mathcal{F}_n)_{b_n^{1/2}}) \xrightarrow{\mathbf{P}^*} 0 \quad \textit{for all} \quad \delta > 0, \tag{3.15}$$

$$\frac{b_n^2}{n} \log N_1(\delta, \mathbf{H}_n, (\mathcal{F}_n)_{b_n^{1/2}}^2) \xrightarrow{\mathbf{P}^*} 0 \quad \textit{for all} \quad \delta > 0, \tag{3.16}$$

and

$$\limsup_{n \to \infty} \sup_{f \in \mathcal{F}_n} \int_{|f|^2 > b_n} |f|^2 dH^{(n)} = 0. \tag{3.17}$$

Moreover, assume that

$$\limsup_{n \to \infty} \int_{|\epsilon|^2 > b_n} |\epsilon|^2 dK^{(n)}(\epsilon) = 0 \tag{3.18}$$

and

$$\limsup_{n \to \infty} \|\epsilon\|_{(n)} < \infty, \quad \limsup_{n \to \infty} \|g_{0,n}\|_{(n)} < \infty. \tag{3.19}$$

Then $\hat{\mathbf{g}}_n$ is $L^2(\mathbb{R}^d, H^{(n)})$-consistent, i.e.

$$\|\hat{\mathbf{g}}_n - g_{0,n}\|_{(n)} \xrightarrow{\mathbf{P}} 0.$$

PROOF. The proof is very similar to the proof for the i.i.d. case. Define $C_n = \sqrt{b_n}$. We construct a covering set of the class

$$\mathcal{H}_{C_n} = \left\{ \left[\frac{\epsilon + g_{0,n}}{1 + \|g\|_{(n)}} \right]_{C_n} - \left[\frac{g}{1 + \|g\|_{(n)}} \right]_{C_n} : g \in \mathcal{G}_n \right\}$$

as before: let $\mathbf{f}_j$, $j = 1, \cdots, N_2(\delta, \mathbf{H}_n, (\mathcal{F}_n)_{C_n})$ be a covering set of $(\mathcal{F}_n)_{C_n}$, take for $f = (g/(1 + \|g\|_{(n)})) \in \mathcal{F}_n$, $\mathbf{f}_j$ the corresponding neighbour of $(f)_{C_n}$ (as in (3.7)) and take

$$\mathbf{h}_{j,k} = (k\delta(\epsilon + g_{0,n}))_{C_n} - \mathbf{f}_j, \quad k = [1/\delta(1 + \|g\|_{(n)})].$$

Then

$$\left\| \left[\frac{\epsilon + g_{0,n}}{1 + \|g\|_{(n)}} \right]_{C_n} - \left[\frac{g}{1 + \|g\|_{(n)}} \right]_{C_n} - \mathbf{h}_{j,k} \right\|_n < \delta \|\epsilon + g_{0,n}\|_n + \delta.$$

Use Lemma 2.3.3 to see that conditions (3.17), (3.18) and (3.19) imply that $\|\epsilon + g_{0,n}\|_n = \mathcal{O}_{\mathbf{P}}(1)$. Thus from (3.15)

$$\frac{b_n^2}{n} \log N_2(\delta, H_n, \mathcal{H}_{C_n}) \xrightarrow{\mathbf{P}^*} 0 \quad \text{for all} \quad \delta > 0. \tag{3.20}$$

If we apply Lemma 2.3.3 to $(\mathcal{F}_n)_{C_n}^2$, we obtain that

$$\sup_{f \in \mathcal{F}_n} |\|(f)_{C_n}\|_n^2 - \|(f)_{C_n}\|_{(n)}^2| \xrightarrow{\mathbf{P}} 0.$$

Therefore

$$\sup_{h \in \mathcal{H}_{C_n}} \|h\|_n \leqslant \|\epsilon + g_{0,n}\|_n + \sup_{f \in \mathcal{F}_n} \|(f)_{C_n}\|_n = \mathcal{O}_{\mathbf{P}}(1).$$

Application of Lemma 2.3.4 now gives that (3.20) implies

$$\frac{b_n^2}{n} \log N_1(\delta, \mathbf{H}_n, \mathcal{H}_{C_n}^2) \xrightarrow{\mathbf{P}^*} 0 \quad \text{for all} \quad \delta > 0.$$

Use Lemma 2.3.3 now for $\mathcal{H}_{C_n}^2$ to get

$$\sup_{h \in \mathcal{H}_{C_n}} |\|h\|_n^2 - \|h\|_{(n)}^2| \xrightarrow{\mathbf{P}} 0.$$

In other words, for arbitrary $\eta > 0$

$$\mathbb{P}\left[\begin{array}{l} \sup_{g \in \mathcal{G}_n} \left[\left\| \left[\frac{\epsilon + g_{0,n}}{1 + \|g\|_{(n)}} \right]_{C_n} - \left[\frac{g}{1 + \|g\|_{(n)}} \right]_{C_n} \right\|_{(n)}^2 - \right. \\ \left. \left\| \left[\frac{\epsilon + g_{0,n}}{1 + \|g\|_{(n)}} \right]_{C_n} - \left[\frac{g}{1 + \|g\|_{(n)}} \right]_{C_n} \right\|_n^2 \right] \leqslant \eta \end{array} \right] > 1 - \eta \tag{3.21}$$

for all n sufficiently large.

Using inequality (3.10) and assumptions (3.17) and (3.18), we get that for all n sufficiently large

$$\| \left[\frac{\epsilon + g_{0,n}}{1+\|g\|_{(n)}} \right]_{C_n} - \left[\frac{g}{1+\|g\|_{(n)}} \right]_{C_n} \|_n^2 \tag{3.22}$$

$$\geqslant \frac{\|\epsilon\|_{(n)}^2 + \|g_{0,n} - g\|_{(n)}^2}{(1+\|g\|_{(n)})^2} - \eta,$$

for all $g \in \mathcal{G}_n$. The fact that $g_{0,n} \in \mathcal{G}_n$ for all n gives that

$$\|\epsilon + g_{0,n} - \hat{\mathbf{g}}_n\|_n^2 \leqslant \|\epsilon\|_n^2.$$

Combine (3.21), (3.22) and (3.23) and use (3.18) and (3.19) for $\|\epsilon\|_n^2$, to obtain that for all n sufficiently large

$$\mathbb{P}\left[\| \frac{\hat{\mathbf{g}}_n - g_{0,n}}{1+\|g\|_{(n)}} \|_{(n)}^2 \leqslant 3\eta \right] > 1 - 2\eta.$$

Since $\|g_{0,n}\|_{(n)}$ is assumed to remain bounded, we can complete the proof as before. □

We can now establish consistency in the empirical metric $\|\cdot\|_n$ using two approaches which depart from apparently different sets of assumptions. The first approach resembles the one for the i.i.d. case: assume that the envelope F_n of $\mathcal{F}_n$ is square integrable. The second approach is to work conditionally on $x_{n,1}, \cdots, x_{n,n}$. We summarize the result in two lemmas.

LEMMA 3.3.2. *Suppose that $\{\mathcal{G}_n\}$ is a sequence of permissible classes, that $g_{0,n} \in \mathcal{G}_n$ for all n and that for some $b_n \geqslant 1$, $b_n = o(n^{1/2})$*

$$\frac{b_n^2}{n} \log N_2(\delta, \mathbf{H}_n, \mathcal{F}_n) \xrightarrow{\mathbf{P}^*} 0, \quad \textit{for all} \quad \delta > 0. \tag{3.24}$$

$$\frac{b_n^2}{n} \log N_1(\delta, \mathbf{H}_n, \mathcal{F}_n^2) \xrightarrow{\mathbf{P}^*} 0, \quad \textit{for all} \quad \delta > 0. \tag{3.25}$$

and

$$\limsup_{n\to\infty} \int_{F_n^2 > b_n} F_n^2 dH^{(n)} = 0. \tag{3.26}$$

Moreover, suppose that (3.18) and (3.19) hold for this $\{b_n\}$. Then $\|\hat{\mathbf{g}}_n - g_{0,n}\|_{(n)}$ as well as $\|\hat{\mathbf{g}}_n - g_{0,n}\|_n$ converge to zero in probability.

PROOF. Of course (3.26) implies (3.17). It is also obvious under (3.26), (3.25) and (3.16) are equivalent, and that (3.24) and (3.15) are equivalent too. So

$$\|\hat{\mathbf{g}}_n - g_{0,n}\|_{(n)} \xrightarrow{\mathbf{P}} 0$$

In other words, for all $\eta>0$, $\|\hat{g}_n - g_{0,n}\|_{(n)}<\eta$ with large probability for all n sufficiently large. It now suffices to show that

$$\sup_{\substack{\|g-g_{0,n}\|_{(n)}<\eta \\ g\in\mathcal{G}_n}} |\|g-g_{0,n}\|_n - \|g-g_{0,n}\|_{(n)}| \xrightarrow{\mathbf{P}} 0. \tag{3.27}$$

Now, application of Lemma 2.3.3 to $\mathcal{F}_n^2$ yields

$$\sup_{f\in\mathcal{F}_n} |\|f\|_n - \|f\|_{(n)}| \xrightarrow{\mathbf{P}} 0,$$

which easily leads to (3.27). □

Recall that under (3.24)

$$\limsup_{n\to\infty} \|F_n\|_{(n)} < \infty$$

implies (3.25).

We now discuss the alternative approach. Conditioning on $\mathbf{x}_{n,k}=x_{n,k}$ $k=1,\ldots,n$, $n=1,2,\cdots$ can be seen as assuming nonstochastic regressors. Therefore, we take $H_{n,k}=\delta_{x_{n,k}}$ in the following theorem.

LEMMA 3.3.3. *Suppose* $\{\mathcal{G}_n\}$ *is a sequence of permissible classes,* $g_{0,n}\in\mathcal{G}_n$, $n\geqslant 1$. *Suppose* $H_{n,k}=\delta_{x_{n,k}}$, $k=1,\ldots,n$, $n=1,2,\cdots$. *If for some* $b_n\geqslant 1$, $b_n=o(n^{1/2})$

$$\frac{b_n^2}{n}\log N_2(\delta, H_n, \mathcal{F}_n)\to 0 \quad \textit{for all} \quad \delta>0 \tag{3.28}$$

$$\limsup_{n\to\infty}\sup_{f\in\mathcal{F}_n}\int_{|f|^2>b_n} |f|^2 dH_n = 0 \tag{3.29}$$

and (3.18) and (3.19) are met, then

$$\|\hat{\mathbf{g}}_n - g_{0,n}\|_n \xrightarrow{\mathbf{P}} 0.$$

PROOF. Conditions (3.28) and (3.29) correspond to (3.15) and (3.17) respectively, with $H_{n,k}=\delta_{x_{n,k}}$, $k=1,\ldots,n$ (under (3.29), truncation becomes redundant). Also (3.16) holds, since Lemma 2.3.4 can be applied:

$$\sup_{f\in\mathcal{F}_n}\|f\|_n = \sup_{g\in\mathcal{G}_n}\frac{\|g\|_n}{1+\|g\|_n}\leqslant 1. \quad \square$$

If the $x_{n,k}$ are actually stochastic, condition (3.29) is to be replaced by

$$\limsup_{n\to\infty}\mathbb{P}(\sup_{f\in\mathcal{F}}\int_{|f|^2>b_n}|f|^2 d\mathbf{H}_n>\eta)=0 \quad \text{for all} \quad \eta>0.$$

Then, provided (3.28) holds in $\mathbb{P}^*$-probability, consistency in empirical norm follows. Lemma 3.3.3 does not give any clue on consistency in theoretical

norm and (3.28) and (3.29) seem to be substantially weaker than the conditions of Lemma 3.3.2. We shall consider the particular case of i.i.d. $\mathbf{x}_k$ and $\mathcal{G}_n = \mathcal{G}$, where nevertheless the assumptions of Lemma 3.3.3 imply those of Lemma 3.3.2.

LEMMA 3.3.4 *Suppose that* $\mathbf{x}_1, \mathbf{x}_2, \cdots$ *are i.i.d. with distribution* H *and that* $\mathcal{F}$ *is a permissible class with envelope* F. *If*

$$\frac{1}{n} \log N_2(\delta, \mathbf{H}_n, \mathcal{F}) \xrightarrow{\mathbf{P}^*} 0 \quad \textit{for all} \quad \delta > 0$$

and if for all $\eta > 0$

$$\lim_{n \to \infty} \sup \mathbb{P}(\sup_{f \in \mathcal{F}} \int_{|f|^2 > b_n} |f|^2 d\mathbf{H}_n > \eta) = 0, \tag{3.30}$$

for all b_n *tending to infinity arbitrarily slowly then*

$$F \in L^2(\mathbb{R}^d, H) .$$

PROOF. As in the proof of Lemma 2.3.3, we can choose sequences $\epsilon_n \downarrow 0$, $\delta_n \downarrow 0$ such that

$$\mathbb{P}^*(\frac{1}{n} \log N_2(\delta_n, \mathbf{H}_n, \mathcal{F}) > \epsilon_n) \to 0 .$$

It now follows from application of Lemma 2.3.4 that for some sequences $b_n \to \infty$, $b_n = o(n^{1/2})$,

$$\frac{b_n^2}{n} \log N_1(\delta, \mathbf{H}_n, (\mathcal{F})_{b_n}^{2,1/2}) \xrightarrow{\mathbf{P}^*} 0 \quad \text{for all} \quad \delta > 0 . \tag{3.31}$$

Let $\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2, \cdots$ be independent random variables with $\mathbb{P}(\boldsymbol{\sigma}_k = 1) = \mathbb{P}(\boldsymbol{\sigma}_k = -1) = 1/2$. It follows from (3.31) that

$$\sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{k=1}^{n} \boldsymbol{\sigma}_k (f(\mathbf{x}_k))_{b_n}^{2,1/2}| \xrightarrow{\mathbf{P}} 0.$$

Hence by (3.30)

$$\sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{k=1}^{n} \boldsymbol{\sigma}_k f^2(\mathbf{x}_k)| \leqslant$$

$$\sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{k=1}^{n} \boldsymbol{\sigma}_k (f(\mathbf{x}_k))_{b_n}^{2,1/2}| + 2 \sup_{f \in \mathcal{F}} \int_{|f|^2 > b_n} |f|^2 d\mathbf{H}_n \xrightarrow{\mathbf{P}} 0.$$

Since

$$\sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{k=1}^{n} \boldsymbol{\sigma}_k f^2(\mathbf{x}_k)|$$

is a reversed submartingale (see e.g. POLLARD (1984)) this implies

$$\sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{k=1}^{n} \boldsymbol{\sigma}_k f^2(\mathbf{x}_k)| \to 0 \quad \text{almost surely.}$$

But by the Borel Cantelli Lemma, the strong uniform law of large numbers implies that the envelope is integrable

$$\int \sup_{f \in \mathcal{G}} f^2 dH < \infty$$

(see GINÉ and ZINN (1984)). □

It turns out that in case of stochastic $\mathbf{x}_{n,k}$ (i.e. $H_{n,k}$ does not degenerate at $\mathbf{x}_{n,k} = x_{n,k}$) it is often difficult to verify whether the entropy condition (3.28) holds in $\mathbb{P}^*$-probability, unless the envelope condition (3.26) holds. For obtaining consistency in both $\|\cdot\|_{(n)}$ - and $\|\cdot\|_n$-norm, our approach indeed needs the envelope condition (3.26).

EXAMPLE 3.1. Suppose (for simplicity) that $(\mathbf{x}_1, \boldsymbol{\epsilon}_1), (\mathbf{x}_2, \boldsymbol{\epsilon}_2), \cdots$ are i.i.d. and that g_0 is fixed. Suppose that $\mathcal{G} \subset L^2(\mathbb{R}^d, H)$ is a permissible VC-graph class with $g_0 \in \mathcal{G}$. Let $b_n \to \infty$, $b_n = o(n^{1/2}(\log n)^{-1/2})$ and define

$$\mathcal{G}_n = \{(g)_{b_n^{1/2}} : g \in \mathcal{G}\}.$$

Let $\hat{\mathbf{g}}_n$ be the function in $\mathcal{G}_n$ which minimizes $\|y - g\|_n$. Then one can prove that

$$\|\hat{\mathbf{g}}_n - g_0\| \xrightarrow{\mathbb{P}} 0$$

as well as

$$\|\hat{\mathbf{g}}_n - g_0\|_n \xrightarrow{\mathbb{P}} 0.$$

To see this, recall Theorem 2.2.6, which says that for all $C > 0$, $\delta > 0$, $n \geq 1$ and for some constants A and r

$$N_1(\delta, \mathbf{H}_n, (\mathcal{G})_C) \leq AC^r \delta^{-r}.$$

Let $\mathcal{F}_n = \{g/(1 + \|g\|) : g \in \mathcal{G}_n\}$. By straightforward computation

$$N_2(\delta, \mathbf{H}_n, \mathcal{F}_n) \leq A' b_n^{r+1/2}, \delta^{-2r-1}, \quad \delta > 0$$

for some A', and

$$N_1(\delta, \mathbf{H}_n, \mathcal{F}_n^2) \leq 4A' b_n^{2r+1} \delta^{-2r-1}, \quad \delta > 0.$$

Thus, the conditions of Lemma 3.3.2 are met, except that g_0 need not be $\mathcal{G}_n$ for all n, i.e. (3.23) need not hold. However, we can replace (3.23) by

$$\|y - \hat{\mathbf{g}}_n\|_n \leq \|\boldsymbol{\epsilon}\|_n + \|(g_0 - (g_0)_{b_n^{1/2}}) 1_{|g_0|^2 > b_n}\|_n \leq \|\boldsymbol{\epsilon}\|_n + \eta$$

almost surely, since $\|(g_0 - (g_0)_{b_n^{1/2}}) 1_{|g_0|^2 > b_n}\|_n \to 0$ almost surely.

We end this section with the following observation. Since everything may depend on n, one can define a new class $\mathcal{G}_n^* = \{a_n g : g \in \mathcal{G}_n\}$, with $\{a_n\}$ some sequence converging to infinity, and use the uniform laws of large numbers of the previous chapter to prove that $\|a_n(\hat{\mathbf{g}}_n - g_{0,n})\|_n$ converges to zero. In other words, in this way one obtains a rate of convergence. However, the resulting

rate will not always be the best possible. Note that so far, we only assumed existence of second order moments of the $\epsilon_{n,k}$. We shall show in Chapter 6 how the existence of higher order moments of disturbances can lead to optimal rates and laws of large deviations.

Nevertheless, consistency of $\|a_n(\hat{\mathbf{g}}_n - g_{0,n})\|_n$ can be concern in certain parametric models, where

$$\mathcal{G} = \{g_\theta : \theta \in \Theta\} \tag{3.32}$$

with $\Theta \subset \mathbb{R}^r$.

EXAMPLE 3.2. In linear regression

$$g_\theta(x) = x\theta$$

with x a row-vector in $\mathbb{R}^d$ and θ a column vector. Let $H_{n,x} = \delta_{x_{n,k}}$ and let

$$X_n = \begin{bmatrix} x_{n,1} \\ \cdot \\ \cdot \\ \cdot \\ x_{n,n} \end{bmatrix}$$

be the design matrix. Denote by $\lambda_{1,n}$ and $\lambda_{2,n}$ the smallest and largest eigenvalue of $X_n^T X_n$ respectively. It is easy to see that if

$$\lambda_{2,n}/\lambda_{1,n} = \mathcal{O}(n^{1/2(1-c)}) \text{ for some } 0 < c \leqslant 1,$$

then conditions (3.29) and (3.30) of Lemma 3.3.3. are fulfilled with $b_n = n^{1/2(1-c)}$. It follows that

$$\|\hat{\mathbf{g}}_n - g_{0,n}\|_n \xrightarrow{\mathbf{P}} 0,$$

provided that the regularity conditions (3.18) and (3.19) are met. If

$$\liminf_{n\to\infty} \frac{1}{n}\lambda_{1,n} > 0,$$

this in turn implies

$$\|\hat{\boldsymbol{\theta}}_n - \theta_{0,n}\| \xrightarrow{\mathbf{P}} 0,$$

$\hat{\mathbf{g}}_n = g_{\hat{\boldsymbol{\theta}}_n}, g_{0,n} = g_{\theta_{0,n}}$. However, if $X_n' X_n$ is ill-conditioned, i.e. if $n^{-1}\lambda_{1,n}$ goes to zero, consistency of $\hat{\mathbf{g}}_n$ in $\|\cdot\|$-norm no longer implies consistency of $\hat{\boldsymbol{\theta}}_n$.

The following lemma presents a direct proof of consistency of the least squares estimator of a finite-dimensional parameter. It is a straightforward application of Theorem 2.3.2. To arrive at the same result as in WU (1981), we assume compactness of parameter space. By a simple argument, this assumption can be dropped at the cost of strengthening (3.33) (see also Section 6.2). Moreover, we assume nonstochastic $x_{n,k}$.

LEMMA 3.3.5. *Let* $\mathcal{G} = \{g_\theta : \theta \in \Theta\}$, *with* Θ *a compact subset of* $\mathbb{R}^d$, $g_0 = g_{\theta_0}$,

$\theta_0 \in \Theta$, and let $H_{n,k} = \delta_{x_{n,k}}$, $k = 1, \ldots, n$, $n = 1,2, \cdots$. Suppose

$$\|g_\theta - g_{\theta_0}\|_n \geqslant K_{1,n}\|\theta - \theta_0\|$$

for all $\theta \in \Theta$, where $K_{1,n} > 0$,

$$|g_\theta(x) - g_{\theta'}(x)| \leqslant \Lambda_{2,n}(x)\|\theta - \theta'\|$$

for all $\theta, \theta' \in \Theta$, and where $\|\Lambda_{2,n}\|_n = K_{2,n} = \mathcal{O}(1)$ and

$$\frac{K_{2,n}^{1+c}}{K_{1,n}^2} = \mathcal{O}(n^{1/2(1-c)}) \tag{3.33}$$

for some $0 < c \leqslant 1$. Moreover, impose the regularity conditions

$$\limsup_{n\to\infty} \|\epsilon\|_n < \infty, \quad |\|\epsilon\|_n - \|\epsilon\|_{(n)}| \xrightarrow{\mathbf{P}} 0.$$

Then

$$\|\hat{\boldsymbol{\theta}}_n - \theta_0\| \xrightarrow{\mathbf{P}} 0.$$

PROOF. Since

$$\|y - \hat{\mathbf{g}}_n\|_n \leqslant \|\epsilon\|_n, \quad \text{or}$$

$$\frac{2}{n}\sum_{k=1}^{n} \epsilon_{n,k}(\hat{\mathbf{g}}_n(x_{n,k}) - g_0(x_{n,k})) \geqslant \|\hat{\mathbf{g}}_n - g_0\|_2^2,$$

it suffices to show that for all $\eta > 0$

$$\sup_{\substack{\|\theta-\theta_0\|>\eta \\ \theta\in\Theta}} \frac{\frac{1}{n}\sum_{k=1}^{n} \epsilon_{n,k}(g_\theta(x_{n,k}) - g_{\theta_0}(x_{n,k}))}{\|g_\theta - g_{\theta_0}\|_n^2} \xrightarrow{\mathbf{P}} 0.$$

Define

$$\mathcal{H}_n = \{h(\epsilon, x) = \frac{\epsilon(g_\theta(x) - g_{\theta_0}(x))}{\|g_\theta - g_{\theta_0}\|_n^2} : \|\theta - \theta_0\| > \eta, \ \theta \in \Theta\}$$

and $c_n = n^{1-c/2}$. It is now easy to see that Theorem 2.3.2 can be applied to $\mathcal{H}_n$. Thus

$$\sup_{h\in\mathcal{H}_n} |\frac{1}{n}\sum_{k=1}^{n} h(\epsilon_{n,k}, \mathbf{x}_{n,k})| \xrightarrow{\mathbf{P}} 0,$$

and the proof is complete. □

3.4. Two-phase regression in detail: identified case

We noted already in Example 1.3 that the class $\mathcal{G}$ of functions of the form

$$g(x) = \begin{cases} \alpha^{(1)} + x\beta^{(1)} & \text{if } x\gamma \leqslant 1 \\ \alpha^{(2)} + x\beta^{(2)} & \text{if } x\gamma > 1, \end{cases} \quad \theta^{(i)} = \begin{bmatrix} \alpha^{(i)} \\ \beta^{(i)} \end{bmatrix} \in \mathbb{R}^{d+1}, \quad i = 1,2, \quad \gamma \in \mathbb{R}^d \tag{3.34}$$

is a *VC*-graph class. Thus there exist constants A and r such that for arbitrary probability measure Q

$$N_1(\delta, Q, (\mathcal{G})_C) \leqslant A\delta^{-r}C^r, \quad \text{for all} \quad C > 0, \quad \delta > 0. \tag{3.35}$$

Since, $\mathcal{G}$ is a cone, the same holds for any rescaled version of $\mathcal{G}$, e.g. $\mathcal{F}_n = \{g/(1 + \|g\|_{(n)}) : g \in \mathcal{G}\}$. In other words, no distributional assumptions are needed to verify the entropy conditions (3.6) (or (3.15) and (3.16)) of the previous sections. To investigate consistency, we now have to check some uniform square integrability condition. Here, we do need to specify the distributional assumptions.

By making use of the results of Section 3.3, one can study the general setup with possibly non-i.i.d. random variables. However, to simplify the exposition we mainly restrict ourselves to the i.i.d. case and only briefly address the non-i.i.d. case at the end of this section. We assume that $\mathbf{x}_1, \mathbf{x}_2, \cdots$ are i.i.d. with distribution H, and $\epsilon_1, \epsilon_2, \cdots$ are i.i.d. with expectation zero and finite variance and independent of the $\mathbf{x}_k$, $k = 1,2, \cdots$. Also g_0 is assumed to be fixed. We consider the class of regression functions

$$\mathcal{G} = \{g(x) = \sum_{i=1,2} (\alpha^{(i)} + x\beta^{(i)}) 1_{A^{(i)}}(x) : \theta^{(i)} = \begin{bmatrix} \alpha^{(i)} \\ \beta^{(i)} \end{bmatrix} \in \mathbb{R}^{d+1}, \quad A^{(i)} \subset \mathbb{R}^d, \quad i = 1,2,$$

$$A^{(1)} \cup A^{(2)} = \mathbb{R}^d, \; A^{(1)} \cap A^{(2)} = \varnothing, \; A = A^{(1)} \in \mathcal{A}\} \tag{3.36}$$

where $\mathcal{A}$ is a permissible class of subsets of $\mathbb{R}^d$. For convenience, we often write $\mathcal{A}^{(1)} = \mathcal{A}$ and $\mathcal{A}^{(2)} = \{A^c : A \in \mathcal{A}\}$. We do not restrict $\mathcal{A}$ to be the class of halfspaces $\{\{x : x\gamma \leqslant 1\}, \gamma \in \mathbb{R}^d\}$. Moreover, the regression functions are allowed to be discontinuous. The least squares estimator is defined by

$$\|y - \hat{\mathbf{g}}_n\|_n = \inf_{g \in \mathcal{G}} \|y - g\|_n,$$

where $\mathcal{G}$ is given in (3.36).

Theorem 3.1.2 asserts that $\hat{\mathbf{g}}_n$ is $L^2(\mathbb{R}^d, H)$-consistent if both

$$\frac{1}{n} \log N_2(\delta, \mathbf{H}_n, \mathcal{A}) \xrightarrow{\mathbf{P}^*} 0$$

and $\mathcal{F} = \{g/(1 + \|g\|) : g \in \mathcal{G}\}$ uniformly (H-)square integrable. However, it turns out that even if the regression functions are of the form (3.34), $\mathcal{F}$ is in general not uniformly square integrable. Here are three examples.

EXAMPLE 3.3. Take $d=1$ and consider the class $\mathcal{G}_s$ defined by

$$\mathcal{G}_s = \{\alpha 1_{(-\infty,\gamma]}: \alpha\in\mathbb{R},\ \gamma\in\mathbb{R}\}.$$

Note that $\mathcal{G}_s$ is a subclass of $\mathcal{G}$ in case $\mathcal{C}$ is the collection of halfspaces. Define $H(\gamma)=H(-\infty,\gamma]$. Suppose there exists a sequence $\{\gamma_m\}_{m=1}^{\infty}$, with $H(\gamma_m)>0$, $m=1,2,\cdots$ and

$$\lim_{m\to\infty} H(\gamma_m) = 0.$$

Let $g_m=\alpha_m 1_{(-\infty,\gamma_m]}$, $\alpha_m=H(\gamma_m)^{-1/2}$, $m=1,2,\cdots$. Then $\|g_m\|=1$ and

$$\int_{|g_m|/(1+\|g_m\|)>C}\left[\frac{g_m}{1+\|g_m\|}\right]^2 dH = \frac{1}{4}\int_{|g_m|>2C}(g_m)^2 dH\to 1/4,\ m\to\infty,$$

since $|g_m|>2C$ for m sufficiently large.

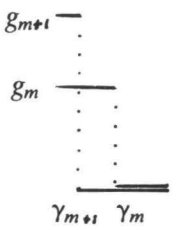


FIGURE 3.1. H is the uniform distribution on (0,1)

EXAMPLE 3.4. Let $d=1$ and

$$\mathcal{G}_s = \{g_\beta(x)=\min(\beta^3+\beta x,0):\ \beta>0\}.$$

Let $H(x)=-\dfrac{1}{x^3}$, $-\infty<x\leqslant 1$. Then $\|g_\beta\|=1$, $g_\beta\in\mathcal{G}_S$ and

$$\lim_{\beta\to\infty}\int_{|g_\beta|/(1+\|g_\beta\|)>C}\left[\frac{g_\beta}{1+\|g_\beta\|}\right]^2 dH = \frac{1}{4}.$$

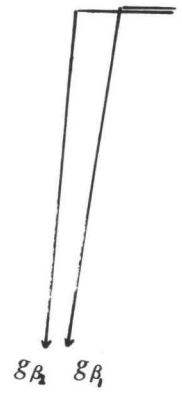


FIGURE 3.2. $\beta_1<\beta_2$

EXAMPLE 3.5. Let $d=1$ and

$$\mathcal{G}_s = \{g_\gamma(x)= \sqrt{\frac{6}{\gamma^3}}\, 1_{(-\infty,\gamma]}(x)\colon \gamma>0\}.$$

Let $H(x)=\frac{1}{2}x+\frac{1}{2}$, $0\leqslant x<1$, $H(\{0\})=\frac{1}{2}$.
Then $\|g_\gamma\|=1$ for all $g_\gamma\in\mathcal{G}_s$ and

$$\lim_{\gamma\to\infty} \int_{|g_\gamma|/(1+\|g_\gamma\|)>C} \left(\frac{g_\gamma}{1+\|g_\gamma\|}\right)^2 = \frac{1}{4}.$$

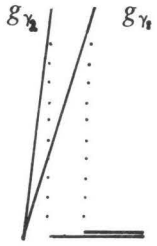


FIGURE 3.3. $\gamma_1<\gamma_2$

Our conclusion is that Theorem 3.1.2 cannot be applied under fairly general conditions on H. We shall now take the following approach. We first show that for a subclass $\mathcal{G}_R$ of $\mathcal{G}$, $\mathcal{F}_R=\{g/(1+\|g\|)\colon g\in\mathcal{G}_R\}$ is uniformly square integrable, provided of course that

$$\mathbb{E}\mathbf{x}^T\mathbf{x} < \infty. \tag{3.37}$$

In the sequel, we assume throughout that (3.37) is fulfilled. Next, we show that under certain conditions on g_0 and H, $\hat{\mathbf{g}}_n$ automatically belongs to this subclass $\mathcal{G}_R$ for all n sufficiently large (see Lemma 3.4.2).

As before, write $\|\theta^{(i)}\|$ for the norm of the Euclidean vector $\theta^{(i)}$. Define

$$g_{\theta^{(i)}}(x) = \alpha^{(i)}+x\beta^{(i)}=(1,x)\theta^{(i)},\ \theta^{(i)}=\begin{bmatrix}\alpha^{(i)}\\ \beta^{(i)}\end{bmatrix},\quad i=1,2. \tag{3.38}$$

Define for $A\subset\mathbb{R}^d$

$$\Sigma(A) = \int_A \begin{bmatrix}1 & x\\ x^T & x^Tx\end{bmatrix} dH(x).$$

If $H(A)\neq 0$ we denote by λ_A the smallest non-zero eigenvalue of $\Sigma(A)$, and otherwise we take $\lambda_A=1$. Note that in all three examples 3.3, 3.4 and 3.5, we constructed a sequence of functions $g=g_{\theta^{(i)}}1_A$ with $\lambda_A\to 0$. The following lemma asserts that if one prevents λ_A from becoming arbitrarily small this results in uniform square integrability.

LEMMA 3.4.1. *For $\eta>0$, consider the restricted class of regression functions*

$$\mathcal{G}_R = \{g = \sum_{i=1,2} g_{\theta^{(i)}} 1_{A^{(i)}} : g \in \mathcal{G},\ \lambda_{A^{(i)}} > \eta,\ i = 1,2\}. \tag{3.39}$$

The class

$$\mathcal{F}_R = \{g/(1+\|g\|) : g \in \mathcal{G}_R\}$$

is uniformly square integrable.

PROOF. Take $g = \sum_{i=1,2} g_{\theta^{(i)}} 1_{A^{(i)}} \in \mathcal{G}_R$ and $C > 0$. We have

$$\int_{|g/(1+\|g\|)|>C} \left[\frac{g}{1+\|g\|}\right]^2 dH \leqslant \sum_{i=1,2} \int_{\frac{|g_{\theta^{(i)}} 1_{A^{(i)}}|}{1+\|g_{\theta^{(i)}} 1_{A^{(i)}}\|} > C} \left[\frac{g_{\theta^{(i)}}}{1+\|g_{\theta^{(i)}} 1_{A^{(i)}}\|}\right]^2 dH.$$

For $A \subset \mathbb{R}^d$, $H(A) \neq 0$, let $\tilde{\Lambda}_A$ be the diagonal matrix of eigenvalues of $\Sigma(A)$, and $\tilde{P}_A$ the matrix of eigenvectors:

$$\Sigma(A) = \tilde{P}_A \tilde{\Lambda}_A \tilde{P}_A^T, \quad \tilde{P}_A \tilde{P}_A^T = \tilde{P}_A^T \tilde{P}_A = I.$$

The diagonal matrix of non-zero eigenvalues is denoted by Λ_A, and the corresponding matrix of eigenvectors by P_A:

$$\Sigma(A) = P_A \Lambda_A P_A^T.$$

So $\tilde{P}_A = (P_A, P_{0,A})$, with $P_{0,A}$ the eigenvectors corresponding to the eigenvalues equal to zero.

We have

$$\int_A (1,x) P_{0,A} P_{0,A}^T (1,x)^T dH(x) = 0.$$

Hence

$$\int_{\frac{|g_{\theta^{(i)}} 1_A|}{1+\|g_{\theta^{(i)}} 1_A\|} > C} \left[\frac{g_{\theta^{(i)}}}{1+\|g_{\theta^{(i)}} 1_A\|}\right]^2$$

$$= \int_{\frac{|(1,x)P_A P_A^T \theta^{(i)}| 1_A|}{1+\|g_{\theta^{(i)}} 1_A\|} > C} \left[\frac{(1,x) P_A P_A^T \theta^{(i)}}{1+\|g_{\theta^{(i)}} 1_A\|}\right]^2 dH(x)$$

$$\leqslant \int_{\frac{|(1,x)P_A P_A^T \theta^{(i)}|}{1+\|g_{\theta^{(i)}} 1_A\|} > C} \left[\frac{(1,x) P_A P_A^T \theta^{(i)}}{1+\|g_{\theta^{(i)}} 1_A\|}\right]^2 dH(x).$$

If $\lambda_A > \eta$, then

$$\|g_{\theta^{(i)}} 1_A\| = (\theta^{(i)T} \Sigma(A) \theta^{(i)})^{1/2} > \eta^{1/2} \|P_A P_A^T \theta^{(i)}\|.$$

Therefore

$$\sup_{\{A:\lambda_A>\eta\}} \sup_{\theta^{(i)}} \int_{\frac{|(1,x)P_A P_A^T \theta^{(i)} 1_A|}{1+\|g_{\theta^{(i)}} 1_A\|}>C} \left[\frac{(1,x)P_A P_A^T \theta^{(i)}}{1+\|g_{\theta^{(i)}} 1_A\|}\right]^2 dH(x)$$

$$< \sup_{\{A:\lambda_A>\eta\}} \sup_{\mu^{(i)}=P_A P_A^T \theta^{(i)}} \int_{\frac{|(1,x)\mu^{(i)}|}{1+\eta^{1/2}\|\mu^{(i)}\|}>C} \left[\frac{(1,x)\mu^{(i)}}{1+\eta^{1/2}\|\mu^{(i)}\|}\right]^2 dH(x)$$

$$\leqslant \sup_{\mu^{(i)}} \int_{\frac{|(1,x)\mu^{(i)}|}{1+\eta^{1/2}\|\mu^{(i)}\|}>C} \left[\frac{(1,x)\mu^{(i)}}{1+\eta^{1/2}\|\mu^{(i)}\|}\right]^2 dH(x).$$

But the class

$$\{g_{\mu^{(i)}}/(1+\eta^{1/2}\|\mu^{(i)}\|): \mu^{(i)} \in \mathbb{R}^{d+1}\}$$

is uniformly square integrable. □

Write

$$g_0 = \sum_{i=1,2} g_{\theta_0^{(i)}} 1_{A_0^{(i)}}$$

and

$$\hat{\mathbf{g}}_n = \sum_{i=1,2} g_{\hat{\theta}_n^{(i)}} 1_{\hat{A}_n^{(i)}}.$$

Moreover, let for $A \subset \mathbb{R}^d$,

$$A \setminus \tilde{A} = A \cap \tilde{A}^c, \quad A \Delta \tilde{A} = (A \setminus \tilde{A}) \cup (\tilde{A} \setminus A).$$

To show that eventually $\hat{\mathbf{g}}_n \in \mathcal{G}_R$, with $\mathcal{G}_R$ the restricted class defined in (3.39), we first of all need an entropy condition on $\mathcal{A}$. Secondly, we require that g_0 is actually a two-phase regression function, not a one-phase regression function. This can be seen as an identifiability condition, since if g_0 consists of only one phase one cannot identify the $A_0^{(i)}$ or equivalently, one of the $\theta_0^{(i)}$. However, this type of identifiability is not a necessary condition for $\|\cdot\|$-consistency, as we shall see in Chapter 6.

Thirdly, we impose a regularity condition on H. For this purpose, we introduce the class $\mathcal{C}$ of all hyperplanes in $\mathbb{R}^d$, i.e.

$$\mathcal{C} = \{C = \{x: (1,x)^T = PP^T(1,x)^T\}: P \in \mathcal{P}\}$$

were $\mathcal{P}$ is the class of $(d+1)\times s$ matrices P, $1 \leqslant s \leqslant d+1$, $P^T P = I$. As in the proof of the previous lemma, let P_A denote the eigenvectors corresponding to non-zero eigenvalues of $\Sigma(A)$, $A \subset \mathbb{R}^d$, $H(A) \neq 0$. Then $C_A = \{x: x = P_A P_A^T x\}$ is an element of $\mathcal{C}$ with positive mass. Such hyperplanes will play an

important role in Lemma 3.4.2 below. We shall assume that the probability in Hausdorff-neighbourhoods of each $C \in \mathcal{C}$ - neighbourhoods not including C itself - is uniformly small (see (3.42)). This assumption is e.g. fulfilled if H has a uniformly bounded density with respect to Lebesgue measure at all x in such neighbourhoods. The Hausdorff-distance is denoted by

$$d(x,C) = \inf_{\tilde{x} \in C} \|x - \tilde{x}\|, \quad C \subset \mathbb{R}^d,$$

where $\|x - \tilde{x}\|$ is the Euclidean distance between x and $\tilde{x}$.

LEMMA 3.4.2. *Suppose that the entropy condition:*

$$\frac{1}{n} \log N_2(\delta, \mathbf{H}_n, \mathcal{G}) \xrightarrow{\mathbf{P}^*} 0 \quad \textit{for all} \quad \delta > 0, \tag{3.40}$$

the identifiability condition:

$$\|g_0 - g_k\| \to 0 \quad \textit{for some sequence} \quad g_k = \sum_{i=1,2} g_{\theta_k^{(i)}} 1_{A_k^{(i)}} \in \mathcal{G} \tag{3.41}$$

$$\textit{implies} \quad \lambda_{A_k^{(i)}} \to 0, \ i = 1,2,$$

and the regularity condition:

$$\lim_{\eta \downarrow 0} \sup_{C \in \mathcal{C}} H(\{x: 0 < d(x, C) \leqslant \eta\}) = 0 \tag{3.42}$$

are fulfilled. Then there exists an $\eta > 0$ such that eventually $\lambda_{\hat{\mathbf{A}}_n^{(i)}} > \eta$ almost surely.

PROOF. Define

$$\mathbf{\Sigma}_n(A) = \int_A \begin{bmatrix} 1 & x \\ x^T & x^T x \end{bmatrix} d\mathbf{H}_n(x), \quad A \subset \mathbb{R}^d.$$

We have

$$\|y - \hat{\mathbf{g}}_n\|_n^2 \leqslant \|\epsilon\|_n^2,$$

which implies

$$\|\hat{\mathbf{g}}_n\|_n \leqslant 2\|\epsilon\|_n + \|g_0\|_n.$$

Hence for some constant $K < \infty$

$$\|\hat{\mathbf{g}}_n\|_n^2 \leqslant K$$

for all sufficiently large n. Write this as

$$\sum_{i=1,2} \hat{\boldsymbol{\theta}}_n^{(i)T} \mathbf{\Sigma}_n(\hat{\mathbf{A}}_n^{(i)}) \hat{\boldsymbol{\theta}}_n^{(i)} \leqslant K. \tag{3.43}$$

Now, let $\lambda_{\hat{\mathbf{A}}_{n,1}^{(i)}} \geqslant \cdots \geqslant \lambda_{\hat{\mathbf{A}}_{n,d+1}^{(i)}}$ be the eigenvalues of $\Sigma(\hat{\mathbf{A}}_n^{(i)})$ in decreasing order, and define $\lambda_{\hat{\mathbf{A}}_{n,0}^{(i)}} = 1$ and $\lambda_{\hat{\mathbf{A}}_{n,d+2}^{(i)}} = 0$, $i = 1,2$. In other words, $\lambda_{\hat{\mathbf{A}}_n^{(i)}} = \lambda_{\hat{\mathbf{A}}_{n,s_i}^{(i)}}$ for some $0 \leqslant s_i \leqslant d+1$ and $\lambda_{\hat{\mathbf{A}}_{n,s}^{(i)}} = 0$ for $s > s_i$. For each infinite subsequence $\{n'\} \subset \{n\}$ one can construct a further infinite subsequence $\{n^*\} \subset \{n'\}$ such that for some

$0<s_i\leqslant d+2$ and some $\eta_i>0$, $\lambda^{(\omega)}_{\hat{\mathbf{A}}^{\cdot}_{n^{\cdot},s_i}}\to 0$ and $\lambda^{(\omega)}_{\hat{\mathbf{A}}^{\cdot}_{n^{\cdot},s_i-1}}>\eta_i$, $i=1,2$. In view of assumption (3.40), $\boldsymbol{\Sigma}_n(A^{(i)})\to\Sigma(A^{(i)})$ almost surely, uniformly in $A^{(i)}\in\mathcal{A}^{(i)}$, $i=1,2$.

If we denote by $\lambda^{(i)}_{n,1}\geqslant\cdots\geqslant\lambda^{(i)}_{n,d+1}$ the eigenvalues of $\boldsymbol{\Sigma}_n(\hat{\mathbf{A}}^{(i)}_n)$ and define $\lambda^{(i)}_{n,1}=1$ and $\lambda^{(i)}_{n,d+2}=0$, it follows that $\boldsymbol{\lambda}^{(i)}_{n^{\cdot},s_i}\to 0$ and for $n^{\cdot}$ sufficiently large, $\boldsymbol{\lambda}_{n^{\cdot},s_i-1}>\frac{1}{2}\eta_i$, $i=1,2$.

Let $\mathbf{P}^{(i)}_n$ be the matrix of eigenvectors corresponding to the eigenvalues of $\boldsymbol{\Sigma}_n(\hat{\mathbf{A}}^{(i)}_n)$ that are larger than $\frac{1}{2}\eta_i$, with the convention $\mathbf{P}^{(i)}_n=0$ if all eigenvalues are smaller than $\frac{1}{2}\eta_i$. Then (3.43) implies

$$\|\boldsymbol{\mu}^{(i)}_n\|\leqslant\frac{2}{\eta_i}K,$$

where $\boldsymbol{\mu}^{(i)}_n=\mathbf{P}^{(i)}_n\mathbf{P}^{(i)T}_n\hat{\boldsymbol{\theta}}^{(i)}_n$, $i=1,2$. Define

$$\mathbf{C}^{(i)}_n=\{x\colon (1,x)^T=\mathbf{P}^{(i)}_n\mathbf{P}^{(i)T}_n(1,x)^T\},\quad i=1,2.$$

Because for the subsequence, $\boldsymbol{\lambda}^{(i)}_{n^{\cdot},s_i}\to 0$, we have that for each $\eta>0$

$$H(\hat{\mathbf{A}}^{(i)}_{n^{\cdot}}\cap\{x\colon d(x_1,\mathbf{C}^{(i)}_{n^{\cdot}})>\eta\})\to 0.$$

But then from (3.43)

$$H(\hat{\mathbf{A}}^{(i)}_{n^{\cdot}}\setminus\mathbf{C}^{(i)}_{n^{\cdot}})\leqslant H(\hat{\mathbf{A}}^{(i)}_{n^{\cdot}}\cap\{x,\ d(x_1,\mathbf{C}^{(i)}_{n^{\cdot}})>\eta\})$$
$$+\ H(\{x\colon 0<d(x,\mathbf{C}^{(i)}_{n^{\cdot}})\leqslant\eta\})\to 0,\quad\text{as}\quad\eta\to 0,$$

or equivalently

$$H(\hat{\mathbf{A}}^{(i)}_{n^{\cdot}}\setminus\mathbf{B}^{(i)}_{n^{\cdot}})\to 0, \tag{3.44}$$

where $\mathbf{B}^{(i)}_{n^{\cdot}}=\hat{\mathbf{A}}^{(i)}_{n^{\cdot}}\cap\mathbf{C}^{(i)}_{n^{\cdot}}$, $i=1,2$.

The class

$$\mathcal{G}^{(i)}=\{g_{\mu^{(\omega)}}1_{B^{(\omega)}}\colon\|\mu^{(i)}\|\leqslant\frac{2}{\eta_i}K,\ B^{(i)}=A^{(i)}\cap C,\ A^{(i)}\in\mathcal{A}^{(i)},\ C\in\mathcal{C}\}$$

satisfies

$$\frac{1}{n}\log N_2(\delta,\mathbf{H}_n,\mathcal{G}^{(i)})\xrightarrow{\mathbf{P}^{\cdot}}0\quad\text{for all}\quad\delta>0,$$

because (3.40) holds. Moreover, the envelope of $\mathcal{G}^{(i)}$ is in $L^2(\mathbb{R}^d,H)$. Thus

$$\Big|\sum_{i=1,2}\Big[\|(\epsilon+g_0-g_{\boldsymbol{\mu}^{(\omega)}_{n^{\cdot}}})1_{\mathbf{B}^{(\omega)}_{n^{\cdot}}}\|^2_{n^{\cdot}}-\|(\epsilon+g_0-g_{\boldsymbol{\mu}^{(i)}_{n^{\cdot}}})1_{\mathbf{B}^{(\omega)}_{n^{\cdot}}}\|^2\Big]\Big|\to 0\quad\text{almost surely.}$$

Furthermore

$$\sum_{i=1,2}\|(\epsilon+g_0-g_{\boldsymbol{\mu}^{(\omega)}_{n^{\cdot}}})1_{\mathbf{B}^{(\omega)}_{n^{\cdot}}}\|^2_{n^{\cdot}}=\sum_{i=1,2}\|(\epsilon+g_0-g_{\hat{\boldsymbol{\theta}}^{(\omega)}_{n^{\cdot}}})1_{\mathbf{B}^{(\omega)}_{n^{\cdot}}}\|^2_{n^{\cdot}}$$
$$\leqslant\sum_{i=1,2}\|(\epsilon+g_0-g_{\hat{\boldsymbol{\theta}}^{(\omega)}_{n^{\cdot}}})1_{\mathbf{A}^{(\omega)}_{n^{\cdot}}}\|^2_{n^{\cdot}}=\|\epsilon+g_0-\hat{\mathbf{g}}_{n^{\cdot}}\|^2_{n^{\cdot}}\leqslant\|\epsilon\|^2_{n^{\cdot}}.$$

This yields that

$$\limsup_{n^{\cdot}\to\infty}\left[\sum_{i=1,2}\|(g_0-g_{\mu_{n^{\cdot}}^{(\omega)}})1_{\mathbf{B}_{n^{\cdot}}^{(\omega)}}\|^2-\sum_{i=1,2}\|\epsilon 1_{\hat{\mathbf{A}}_{n^{\cdot}}^{(\omega)}\setminus\mathbf{B}_{n^{\cdot}}^{(\omega)}}\|^2\right]\leqslant 0 \quad \text{almost surely.}$$

From (3.44) it now follows that

$$\sum_{i=1,2}\|(g_0-g_{\mu_{n^{\cdot}}^{(\omega)}})1_{\hat{\mathbf{A}}_{n^{\cdot}}^{(\omega)}}\|^2\to 0 \quad \text{almost surely,}$$

or

$$\|g_0-\sum_{i=1,2}g_{\mu_{n^{\cdot}}^{(\omega)}}1_{\hat{\mathbf{A}}_{n^{\cdot}}^{(\omega)}}\|^2\to 0 \quad \text{almost surely.}$$

But then by (3.41), $\lambda_{\hat{\mathbf{A}}_{n^{\cdot}}^{(\omega)}}\to 0$, $i=1,2$.

Summarizing, we have that for each infinite subsequence $\{n'\}\subset\{n\}$ there exists a further infinite subsequence $\{n^*\}\subset\{n'\}$ such that $\lambda_{\hat{\mathbf{A}}_{n^{\cdot}}^{(\omega)}}$ does not converge to 0, $i=1,2$. This shows that there exists an $\eta>0$ such that $\lambda_{\hat{\mathbf{A}}_n^{(\omega)}}>\eta$ for all n sufficiently large. □

It requires virtually no additional effort to conclude from the proof of Lemma 3.4.2 that under assumptions (3.40), (3.41) and (3.42), $\hat{g}_n$ is $\|\cdot\|$-consistent. However, we alternatively use Theorem 3.12 to show this.

PROPOSITION 3.4.3. *Suppose (3.40), (3.41) and (3.42) are met, then* $\|\hat{\mathbf{g}}_n-g_0\|\to 0$ *almost surely.*

PROOF. By Lemma 3.4.2 there exists an $\eta>0$ such that $\lambda_{\hat{\mathbf{A}}_n^{(\omega)}}>\eta$, $i=1,2$, almost surely for all n sufficiently large. Thus it suffices to show that the conditions of Theorem 3.1.2 are fulfilled for the restricted class

$$\mathcal{G}_R=\{g=\sum_{i=1,2}g_{\theta^{(\omega)}}1_{A^{(\omega)}},\ \theta^{(i)}\in\mathbb{R}^{d+1},\ A^{(i)}\in\mathcal{A}^{(i)},\ \lambda_{A^{(\omega)}}>\eta,\ i=1,2\}.$$

The entropy condition follows from (3.40):

$$\frac{1}{n}\log N_2(\delta,\mathbf{H}_n,(\mathcal{F}_R)_C)\xrightarrow{\mathbf{P}^{\cdot}}0 \quad \text{for all} \quad C>0,\ \delta>0.$$

Furthermore, we have shown in Lemma 3.4.1 that $\mathcal{F}_R$ is uniformly square integrable. □

For consistency of the estimators of the parameters $\theta_0^{(i)}$ and $A_0^{(i)}$, $i=1,2$, we of course need a further identifiability condition. If A_0 is known, $\theta_0^{(i)}$ is identified if $\Sigma(A_0^{(i)})$ is of full rank. In the situation with A_0 unknown, this is no longer true, even when $\|\theta_0^{(i)}-\theta_0^{(j)}\|\neq 0$.

EXAMPLE 3.6. Let $d=2$ and suppose H puts all its mass on 8 points $x_1^{(1)},\cdots,x_4^{(1)}$, $x_1^{(2)},\cdots,x_4^{(2)}$, $H(x_t^{(i)})>0$, $t=1,\cdots,4$, $i=1,2$. Let $A_0^{(1)}=\{x_1^{(1)},x_2^{(1)},x_3^{(1)},x_4^{(1)}\}$ and $A^{(1)}=\{x_1^{(1)},x_2^{(1)},x_1^{(2)},x_2^{(2)}\}$ (see Figure 3.4).

Then there exists a θ_0 such that $\|\theta_0^{(1)} - \theta_0^{(2)}\| \neq 0$, and a θ such that $\|\theta - \theta_0\| \neq 0$, with

$$\| \sum_{i=1,2} g_{\theta^{(i)}} 1_{A^{(i)}} - g_0 \| = 0.$$

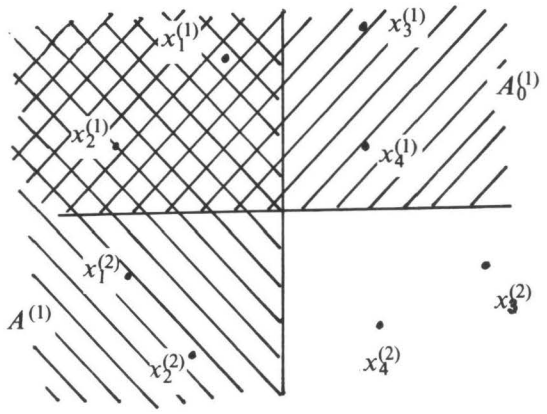


FIGURE 3.4.

Obviously, the roles of $(\theta^{(1)}, A^{(1)})$ and $(\theta^{(2)}, A^{(2)})$ can often be interchanged. Identifiability should be understood in the wide sense, i.e. modulo a possible re-indexing of the $\{(\theta^{(i)}, A^{(i)}): i = 1,2\}$. A sufficient condition for identifiability that can easily be verified, is given in Lemma 3.4.4. Let

$$T = \left\{ \tilde{x}: H(\{x: \|x - \tilde{x}\| < \eta\}) > 0 \quad \text{for all} \quad \eta > 0 \right\}$$

be the support of H. We assume below that there are sufficiently many points in $T \cap A_0^{(i)}$, $i = 1,2$, in order to identify θ.

LEMMA 3.4.4. *Suppose that* $\|\theta_0^{(1)} - \theta_0^{(2)}\| \neq 0$ *and that there are* $2(d+1)-1$ *points* $\{x_t^{(i)}: t = 1, \cdots, 2(d+1)-1\} \subset A_0^{(i)} \cap T$, *with no* $d+1$ $x_t^{(i)}$ *on a* $(d-1)$*-dimensional hyperplane,* $i = 1,2$. *Suppose furthermore that (3.40) and (3.42) are met. Then* $\|\hat{\boldsymbol{\theta}}_n - \theta_0\| \to 0$ *almost surely. If moreover*

$$H(\{x: g_{\theta_0^{(1)}}(x) = g_{\theta_0^{(2)}}(x)\}) = 0, \tag{3.45}$$

then also $H(\hat{\mathbf{A}}_n \Delta A_0) \to 0$ *almost surely. (These convergence results should be understood modulo replacement of* $(\hat{\boldsymbol{\theta}}_n^{(i)}, \mathbf{A}_n^{(i)})$ *by* $(\hat{\boldsymbol{\theta}}_n^{(j)}, \hat{\mathbf{A}}_n^{(j)})$, $i \neq j$*).*

PROOF. We shall first show that the identifiability condition (3.41) is fulfilled. Let $g_k = \Sigma_{i=1,2}\, g_{\theta_k^{(i)}} 1_{A_k^{(i)}} \in \mathcal{G}$ be some sequence with $\|g_k - g_0\| \to 0$. Either $A_k^{(1)}$ or $A_k^{(2)}$ contains at least $2(d+1)-1$ points from $\{x_t^{(i)}: t = 1, \cdots, 2(d+1)-1,\ i = 1,2\}$. Therefore, at least $(d+1)$ of the $x_t^{(i)}$ in one of the $A_k^{(i)}$, say $A_k^{(1)}$, must all be $x_t^{(1)}$'s or $x_t^{(2)}$'s. Without loss of generality,

we can therefore assume that at least $(d+1)$ $x_i^{(1)}$'s are in $A_k^{(1)}$ for all k. This implies that $\lambda_{A_k^{(1)}}$ stays away from zero. Moreover, it implies that $\theta_k^{(1)} \to \theta_0^{(1)}$. This in turn yields that $A_k^{(1)}$ cannot contain more than d $x_i^{(2)}$'s, since $\|\theta_0^{(1)} - \theta_0^{(2)}\| \neq 0$. So $A_k^{(2)}$ must contain more than $(d+1)$ $x_i^{(2)}$'s and thus $\lambda_{A_k^{(2)}}$ stays away from zero and $\theta_k^{(2)} \to \theta_0^{(2)}$. In other words (3.41) holds.

Conditions (3.40) and (3.42) yield consistency of $\hat{\mathbf{g}}_n$, and obviously this now results in consistency of $\hat{\boldsymbol{\theta}}_n$.

Since

$$\|\hat{\mathbf{g}}_n - g_0\|^2 = \sum_{i \neq j \in \{1,2\}} \Big\{ (\hat{\boldsymbol{\theta}}_n^{(i)} - \theta_0^{(i)})^T \Sigma(\hat{\mathbf{A}}_n^{(i)} \cap A_0^{(i)})(\hat{\boldsymbol{\theta}}_n^{(i)} - \theta_0^{(i)}) + (\hat{\boldsymbol{\theta}}_n^{(i)} - \theta_0^{(j)})^T \Sigma(\hat{\mathbf{A}}_n^{(i)} \setminus A_0^{(j)})(\hat{\boldsymbol{\theta}}_n^{(i)} - \theta_0^{(j)}) \Big\}$$

the consistency of $\hat{\mathbf{g}}_n$ and $\hat{\boldsymbol{\theta}}_n$ implies that

$$(\theta_0^{(1)} - \theta_0^{(2)})^T \Sigma(\hat{\mathbf{A}}_n \Delta A_0)(\theta_0^{(1)} - \theta_0^{(2)}) \to 0. \tag{3.46}$$

Now, let $\lambda_{\hat{\mathbf{A}}_n \Delta A_0, 1} \geqslant \cdots \geqslant \lambda_{\hat{\mathbf{A}}_n \Delta A_0, d+1}$ be the eigenvalues of $\Sigma(\hat{\mathbf{A}}_n \Delta A_0)$, and take $\lambda_{\hat{\mathbf{A}}_n \Delta A_0, 0} = 1$ and $\lambda_{\hat{\mathbf{A}}_n \Delta A_0, d+2} = 0$. Construct an infinite subsequence $\{n^*\} \subset \{n\}$ such that for some $0 < s \leqslant d+2$ and some $\eta_0 > 0$, $\lambda_{\hat{\mathbf{A}}_{n^*} \Delta A_0, s} \to 0$ and $\lambda_{\hat{\mathbf{A}}_{n^*} \Delta A_0, s-1} > \eta_0$. Let $P_{\hat{\mathbf{A}}_n \Delta A_0}$ be the matrix of eigenvectors corresponding to the eigenvalues larger than η_0, with $P_{\hat{\mathbf{A}}_n \Delta A_0} = 0$ if all eigenvalues are smaller than η_0. Define

$$B_{\hat{\mathbf{A}}_n \Delta A_0} = \{x: (1,x)^T = P_{\hat{\mathbf{A}}_n \Delta A_0} P_{\hat{\mathbf{A}}_n \Delta A_0}^T (1,x)^T\}$$

and

$$C_0 = \{x: g_{\theta_0^{(1)}}(x) = g_{\theta_0^{(2)}}(x)\}.$$

It follows from (3.46) that

$$(\theta_0^{(1)} - \theta_0^{(2)})^T P_{\hat{\mathbf{A}}_{n^*} \Delta A_0} P_{\hat{\mathbf{A}}_{n^*} \Delta A_0}^T (\theta_0^{(1)} - \theta_0^{(2)}) \to 0.$$

Therefore for each $\eta > 0$

$$H(B_{\hat{\mathbf{A}}_{n^*} \Delta A_0} \cap \{x: d(x, C_0) > \eta\}) \to 0.$$

Assumptions (3.42) and (3.45) now yield

$$H(B_{\hat{\mathbf{A}}_{n^*} \Delta A_0}) \leqslant H(\{x: d(x, C_0) = 0\}) + H(\{x: 0 < d(x, C_0) \leqslant \eta\}) + H(B_{\hat{\mathbf{A}}_{n^*} \Delta A_0} \cap \{x: d(x, C_0) > \eta\}) \to 0.$$

Again by (3.42) this implies that also $H(\hat{\mathbf{A}}_{n^*} \Delta A_0) \to 0$. □

Here is an example where (3.45) is not fulfilled.

EXAMPLE 3.7. Take $d = 1$ and $g_0(x) = \min(\alpha_0 + x\beta_0, 0)$. Suppose that $\beta_0 \neq 0$

and that there is positive mass m concentrated at the change point $-\alpha_0/\beta_0$. Let $A_0^{(i)}=(-\infty, -\alpha_0/\beta_0]$ and $A_n^{(i)}=(-\infty, -\alpha_n/\beta_n)$ with α_n/β_n some sequence converging from above to α_0/β_0. Then

$$\Sigma(A_n \Delta A_0) = \int_{\frac{-\alpha_n}{\beta_n} < x \leqslant \frac{-\alpha_0}{\beta_0}} \begin{bmatrix} 1 & x \\ x & x^2 \end{bmatrix} dH(x) \to m \begin{bmatrix} 1 & \frac{-\alpha_0}{\beta_0} \\ \frac{-\alpha_0}{\beta_0} & \left[\frac{\alpha_0}{\beta_0}\right]^2 \end{bmatrix}.$$

Thus the limiting matrix is singular and has $\begin{bmatrix} \alpha_0 \\ \beta_0 \end{bmatrix}$ in its null-space.

In the non-i.i.d. case, consistency of $\hat{\mathbf{g}}_n$ and of the parameter estimates can be proved using e.g. Theorem 3.3.1. We shall not do this, but only investigate one particular case for later reference (Examples 6.6 and 6.7). We take $\epsilon_1, \cdots, \epsilon_n$ i.i.d. and $x_{n,1}, \cdots, x_{n,n}$ fixed points on a uniform lattice in the d-dimensional unit cube. Furthermore, we let $\mathcal{A}=\{\{x: x\gamma \leqslant 1\}, \gamma \in \mathbb{R}^d\}$ be the class of halfspaces. Finally, we take $g_0=\Sigma_{i=1,2} g_{\theta_0^{(i)}} 1_{A_0^{(i)}}$ fixed. Define $H_n=1/n\Sigma_{k=1}^n \delta_{x_{n,k}}$ and

$$\Sigma_n(A) = \int_A \begin{bmatrix} 1 & x \\ x^T & x^T x \end{bmatrix} dH_n(x).$$

The conditions of Proposition 3.4.3 and of the lemma following it all have their counterpart for the non-i.i.d. case. In the particular situation we have now, we have introduced so much regularity that the only additional assumption we need is some kind identifiability.

PROPOSITION 3.4.5. *Suppose that the $A_0^{(i)}$, $i=1,2$, have positive Lebesgue measure and that $\|\theta_0^{(1)}-\theta_0^{(2)}\| \neq 0$, then*

$$\|\hat{\boldsymbol{\theta}}_n - \theta_0\| \xrightarrow{\mathbf{P}} 0 \quad \textit{and} \quad H_n(\hat{\mathbf{A}}_n \Delta A_0) \xrightarrow{\mathbf{P}} 0.$$

PROOF. We have as before that for some constant K

$$\sum_{i=1,2} \hat{\boldsymbol{\theta}}_n^{(i)^T} \Sigma_n(\hat{\mathbf{A}}_n^{(i)}) \hat{\boldsymbol{\theta}}_n^{(i)} \leqslant K.$$

We can without loss of generality assume that for all n the eigenvalues of one of the $\Sigma_n(\hat{\mathbf{A}}_n^{(i)})$, say of $\Sigma_n(\hat{\mathbf{A}}_n^{(2)})$, are all bounded away from zero. This implies that $\hat{\boldsymbol{\theta}}_n^{(2)}$ remains bounded. But then by Theorem 2.3.5

$$|\|(\epsilon + g_0 - \hat{\mathbf{g}}_n) 1_{\hat{\mathbf{A}}_n^{(2)}}\|_n^2 - \|(\epsilon + g_0 - \hat{\mathbf{g}}_n) 1_{\hat{\mathbf{A}}_n^{(2)}}\|_{(n)}^2| \xrightarrow{\mathbf{P}} 0,$$

which yields

$$\limsup_{n\to\infty}(\|(g_0-\hat{\mathbf{g}}_n)1_{\hat{\mathbf{A}}_n^{(2)}}\|_{(n)}^2-\|\epsilon 1_{\hat{\mathbf{A}}_n^{(1)}}\|_{(n)}^2) \overset{\mathbf{P}}{\leqslant} 0.$$

The identifiability at g_0 now implies that $H(\hat{\mathbf{A}}_n^{(1)})$ is bounded away from zero. But then also $\hat{\boldsymbol{\theta}}_n^{(1)}$ remains bounded. Application of Theorem 2.3.5 gives

$$\|g_0-\hat{\mathbf{g}}_n\|_{(n)} \overset{\mathbf{P}}{\to} 0.$$

The consistency of $\hat{\boldsymbol{\theta}}_n^{(i)}$, $i=1,2$, and $\hat{\mathbf{A}}_n$ now follows easily. □

4. Empirical process theory II

4.1. Introduction

Just as a uniform law of large numbers can be a tool to prove consistency of the least squares estimator, a uniform central limit theorem can be applied to obtain rates of convergence and asymptotic distributions. First, we briefly discuss the idea which led to Proposition 3.1.1 of the previous chapter. Let $\mathcal{G}$ be a class of measurable functions in $L^2(\mathbb{R}^d, H)$, let $\mathbf{y}=g(\mathbf{x})+\epsilon$, $g\in\mathcal{G}$ be a regression model, where it is assumed that $\mathbb{E}\epsilon=0$, $\mathbb{E}|\epsilon|^2<\infty$ and that $\mathbf{x}$ and ϵ are independent, and let $(\mathbf{x}_1,\epsilon_1)$, $(\mathbf{x}_2,\epsilon_2)$, $\cdots$ be independent copies of $(\mathbf{x},\epsilon)$.

If we define $\prec\epsilon,g\succ_n$ as

$$\prec\epsilon,g\succ_n \;=\; \int\epsilon g d\mathbf{P}_n$$

then we can write

$$\|y-g\|_n^2 \;=\; \|\epsilon\|_n^2-2\prec\epsilon,g-g_0\succ_n+\|g-g_0\|_n^2.$$

Since $\|y-\hat{\mathbf{g}}_n\|_n^2 \leqslant \|\epsilon\|_n^2$,

$$\|\hat{\mathbf{g}}_n-g_0\|_n^2 \leqslant 2\prec\epsilon,\hat{\mathbf{g}}_n-g_0\succ_n. \tag{4.1}$$

The uniform law of large numbers says that if $\mathcal{G}$ is a permissible class satisfying some entropy condition and with envelope $G\in L^2(\mathbb{R}^d, H)$, then

$$\sup_{g\in\mathcal{G}}|\prec\epsilon,g-g_0\succ_n| \to 0 \quad \text{almost surely.}$$

This implies by (4.1) that $\|\hat{\mathbf{g}}_n-g_0\|_n\to 0$ almost surely (see Proposition 3.1.1).

By the central limit theorem, we have that for each $g\in\mathcal{G}$ with $\|g-g_0\|\neq 0$

$$\frac{\sqrt{n}\prec\epsilon,g-g_0\succ_n}{\|g-g_0\|_n} \xrightarrow{\mathcal{L}} \mathcal{N}(0,\|\epsilon\|^2). \tag{4.2}$$

Thus

$$\frac{\prec\epsilon,g-g_0\succ_n}{\|g-g_0\|_n} \;=\; \mathcal{O}_{\mathbf{P}}(n^{-1/2}). \tag{4.3}$$

Suppose now that (4.3) also holds for $\hat{\mathbf{g}}_n$, then (4.1) shows that $\|\hat{\mathbf{g}}_n-g_0\|_n=\mathcal{O}_{\mathbf{P}}(n^{-1/2})$. Indeed under entropy conditions on $\mathcal{G}$, (4.3) holds for $\hat{\mathbf{g}}_n$. We shall establish this in Chapters 5 and 6. The present chapter provides the theoretical background. Uniform central limit theorems will be used directly in Chapter 5. The techniques for proving uniform central limit theorems are adjusted for proving rates of convergence in Chapter 6.

Let us return to (4.1) for a moment. Obviously, for each $g\in\mathcal{G}$

$$\sqrt{n}\prec\epsilon,g-g_0\succ_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,\|\epsilon\|^2\|g-g_0\|^2). \tag{4.4}$$

Now, think of $\sqrt{n}\prec\epsilon,g-g_0\succ_n$ as a process indexed by functions $g\in\mathcal{G}$, and suppose this process converges to some limiting process with uniformly

continuous sample paths for $\|\cdot\|$ (we shall make this more precise in the next section). In view of the $\|\cdot\|$ consistency of $\hat{\mathbf{g}}_n$, this would imply

$$\prec \epsilon, \hat{\mathbf{g}}_n - g_0 \succ_n = o_{\mathbf{P}}(n^{-1/2})$$

and by (4.1), one obtains

$$\|\hat{\mathbf{g}}_n - g_0\|_n = o_{\mathbf{P}}(n^{-1/4}).$$

In Chapter 6, we shall also encounter this rate, and in fact all rates ranging from $\mathcal{O}_{\mathbf{P}}(n^{-1/2})$ to $o_{\mathbf{P}}(n^{-1/4})$.

4.2. Uniform central limit theorems

Define $\mathcal{H} = \{\epsilon(g(x) - g_0(x)) : g \in \mathcal{G}\}$ and

$$\nu_n(h) = \sqrt{n} \int h d\mathbf{P}_n, \quad h \in \mathcal{H}.$$

The process $\nu_n(\cdot)$ is an element of some space $\mathcal{X}$ of real valued functions on $\mathcal{H}$, $\mathcal{X}$ being equipped with supremum norm. A function $y \in \mathcal{X}$ is continuous if $\|h - \tilde{h}\| \to 0$ implies $|y(h) - y(\tilde{h})| \to 0$. We introduce a Gaussian process $\mathbf{G}_p$ on $\mathcal{H}$ with mean zero and covariance structure

$$\operatorname{cov}(\mathbf{G}_p(h), \mathbf{G}_p(\tilde{h})) = \int h\tilde{h} dP,$$

where we assume that $\mathcal{H}$ is G_pBUC (or *P-pregaussian*), i.e. $\mathcal{H}$ is such that $\mathbf{G}_p$ admits a version with *b*ounded and *u*niformly *c*ontinuous sample paths. A sufficient condition for $\mathcal{H}$ to be G_pBUC is the entropy-integrability condition

$$\int_0^1 (\log N_2(x, P, \mathcal{H}))^{1/2} dx < \infty \tag{4.5}$$

(see DUDLEY (1967)).

We first present the definition of a *functional Donsker class.* The word *functional* refers to the fact that convergence in law is strengthened to convergence in probability. This makes it possible to postpone some measurability considerations.

DEFINITION. $\mathcal{H}$ is called a functional Donsker class if

(i) $\mathcal{H}$ is G_pBUC,

(ii) there exist independent copies $Y_k(h, \omega)$ of $\mathbf{G}_p$ such that $h \mapsto Y_k(h, \omega)$ is bounded and uniformly continuous on $\mathcal{H}$ for all k, and such that for all $\eta > 0$

$$\mathbf{P}^*(n^{-1/2} \max_{m \leq n} \sup_{h \in \mathcal{H}} |\sum_{k=1}^{m} (h(\epsilon_k, \mathbf{x}_k) - \mathbf{Y}_k(h))| > \eta) \to 0.$$

Here is a characterization of a functional Donsker class.

THEOREM 4.2.1. *$\mathcal{H}$ is a functional Donsker class iff $\mathcal{H}$ is totally bounded for $\|\cdot\|$ and for all $\eta > 0$ there exists a $\delta > 0$ such that*

$$\mathbb{P}^* \left[\sup_{\substack{h,\tilde{h}\in\mathcal{H} \\ \|h-\tilde{h}\|<\delta}} |\nu_n(h)-\nu_n(\tilde{h})|>\eta \right] <\eta \tag{4.6}$$

for all n sufficiently large.

PROOF. See DUDLEY (1984). □

Condition (4.6) is called the *asymptotic equicontinuity criterion.*

In the literature on empirical processes, there are several results available which make it feasible to check whether a particular $\mathcal{H}$ is a Donsker class. We present one of these results. Let S be a finite collection of points in $\mathbb{R}^{d+1}$ and denote by P_S the empirical distribution based on S. Write $\|\cdot\|_S^2 = \int(\cdot)^2 dP_S$. Define for $H = \sup_{h\in\mathcal{H}}|h|$

$$D_2(\delta, \mathcal{H}) = \sup_S N_2(\delta\|H\|_S, P_S, \mathcal{H}). \tag{4.7}$$

THEOREM 4.2.2. *Suppose that $H\in L^2(\mathbb{R}^{d+1}, P)$, that $\mathcal{H}$ is permissible, and that the entropy integrability condition*

$$\int_0^1 (\log D_2(x, \mathcal{H}))^{1/2} dx < \infty \tag{4.8}$$

holds. Then $\mathcal{H}$ is a functional Donsker class.

PROOF. POLLARD (1982). □

Recall that $\mathcal{H}=\{\epsilon(g(x)-g_0(x)): g\in\mathcal{G}\}$. We use Theorem 4.2.2 to show that under entropy conditions on $\mathcal{G}$, $\mathcal{H}$ is a functional Donsker class, provided a higher order moment of ϵ exists. Observe that (4.8) is met if

$$\log D_2(\delta, \mathcal{H}) \leqslant M\delta^{-\nu} \tag{4.9}$$

for some constants M and $0<\nu<2$ and for all δ. In the following theorem, we impose (4.9) on $\mathcal{G}$.

THEOREM 4.2.3. *Suppose that $\mathcal{G}$ is a permissible class with envelope $G\in L^2(\mathbb{R}^d, H)$, and with*

$$\log D_2(\delta, \mathcal{G}) \leqslant M\delta^{-\nu} \tag{4.10}$$

for some constants M and $0<\nu<2$ and for all $\delta>0$. Moreover, suppose that $\mathbb{E}|\epsilon|^{2p}<\infty$ for some $p>2/(2-\nu)$. Then $\mathcal{H}=\{\epsilon(g(x)-g_0(x)): g\in\mathcal{G}\}$ is a functional Donsker class.

PROOF. If $D_2(\delta, \mathcal{G})<\infty$ for all $\delta>0$, then $\mathcal{G}$ is totally bounded for $\|\cdot\|$ (see DUDLEY (1984)). Since ϵ and $\mathbf{x}$ are independent, this yields that $\mathcal{H}$ is totally bounded for $\|\cdot\|$ too. Thus, the theorem is proved if we show that the

asymptotic equicontinuity criterion (4.6) holds. In fact, the envelope and entropy condition on $\mathcal{G}$ imply that $\|\cdot\|_n \to \|\cdot\|$ almost surely uniformly on $\mathcal{G}$, so it is also sufficient to show that for all $\eta>0$ there exists a $\delta>0$ such that

$$\mathbb{P}\Big(\sup_{\substack{\|g-\tilde{g}\|_n<\delta \\ g,\tilde{g}\in\mathcal{G}}} |n^{1/2}\prec\epsilon, g-\tilde{g}\succ_n|>3\eta\Big)<5\eta. \tag{4.11}$$

Without loss of generality, we may assume that ϵ is symmetric about zero (see the symmetrization device in Section 2.3). Let $\sigma_1, \cdots, \sigma_n$ be independent random variables, independent of $(\mathbf{x}_1, \epsilon_1), \cdots, (\mathbf{x}_n, \epsilon_n)$, with $\mathbb{P}(\sigma_k=1)=\mathbb{P}(\sigma_k=-1)=1/2$. Write

$$\prec\epsilon, g-\tilde{g}\succ_n^0 = \frac{1}{n}\sum_{k=1}^n \sigma_k\epsilon_k(g(\mathbf{x}_k)-\tilde{g}(\mathbf{x}_k)).$$

For each $j\in\mathbb{N}$, let $\mathcal{G}^{(j)}$ be a minimal $(2^{-\frac{jp}{2}}\delta)$-covering set of $\mathcal{G}$ endowed with $L^2(\mathbb{R}^d, \mathbf{H}_n)$-norm. Then

$$\mathrm{Card}(\mathcal{G}^{(j)}) \leqslant \exp(M\delta^{-\nu}2^{jp\nu/2}\|G\|_n^2)$$

for all $j\in\mathbb{N}$ and $\delta>0$. Define $g_{(j)}=g^{(j)}1_{|\epsilon|\leqslant 2^{j/2}\delta^{-1/p}}$, $g^{(j)}\in\mathcal{G}^{(j)}$. We have

$$\mathbb{P}\Big(\sup_{\substack{\|g-\tilde{g}\|<\delta \\ g,g\in\mathcal{G}}} |n^{1/2}\prec\epsilon,g-\tilde{g}\succ_n^0|>3\eta\Big)\leqslant 2\mathbb{P}(\sup_{g\in\mathcal{G}}|n^{1/2}\prec\epsilon,g-g_{(0)}\succ_n^0|>\eta)$$
$$+\mathbb{P}\Big(\sup_{\|g_{(0)}-\tilde{g}_{(0)}\|_n<3\delta}|n^{1/2}\prec\epsilon,\tilde{g}_{(0)}-g_{(0)}\succ_n^0|>\eta\Big)=2\mathbb{P}^{(1)}+\mathbb{P}^{(2)},$$

where $g_{(0)}=g^{(0)}1_{|\epsilon|\leqslant\delta^{-1/p}}$, $g^{(0)}=g^{(0)}(g)\in\mathcal{G}^{(0)}$, $\|g^{(0)}-g\|_n<\delta$.

Let $r=r(n)$ be the smallest integer such that $2^{rp}\geqslant n$ and write

$$\prec\epsilon,g-g_{(0)}\succ_n^0 = \prec\epsilon,g-g_{(r)}\succ_n^0+\sum_{j=0}^{r-1}\prec\epsilon,g_{(j+1)}-g_{(j)}\succ_n^0, \tag{4.12}$$

$g_{(j)}=g^{(j)}1_{|\epsilon|\leqslant 2^{j/2}\delta^{-1/p}}$, $g^{(j)}=g^{(j)}(g)\in\mathcal{G}^{(j)}$, $j=0,1,\cdots,r$. Now,

$$\|g_{(j+1)}-g_{(j)}\|_n \leqslant \|(g^{(j+1)}-g^{(j)})1_{|\epsilon|\leqslant 2^{j/2}\delta^{-1/p}}\|_n+\|G1_{|\epsilon|>2^{j/2}\delta^{-1/p}}\|_n,$$

and

$$\mathbb{P}(\|G1_{|\epsilon|>2^{j/2}\delta^{-1/p}}\|_n^2 \geqslant (j+1)^2 2^{-jp}\delta^2\log\frac{1}{\delta}\|G\|^2\mathbb{E}|\epsilon|^{2p})$$
$$\leqslant\frac{\|G\|^2\mathbb{P}(|\epsilon|>2^{j/2}\delta^{-1/p})}{(j+1)^2 2^{-jp}\delta^2\log\frac{1}{\delta}\|G\|^2\mathbb{E}|\epsilon|^{2p}} \leqslant \frac{1}{(j+1)^2\log\frac{1}{\delta}}.$$

Thus

$$\mathbb{P}\left(\|G1_{|\epsilon|>2^{j/2}\delta^{-1/p}}\|_n^2 \geqslant (j+1)^2 2^{-jp}\delta^2\log\frac{1}{\delta}\|G\|^2\mathbb{E}|\epsilon|^{2p}\text{ for some }\ j\in\mathbb{N}\right)$$
$$\leqslant\sum_{j=0}^{\infty}\frac{1}{(j+1)^2\log\frac{1}{\delta}}<\frac{1}{2}\eta,$$

for δ sufficiently small. Hence, with probability $>1-\frac{1}{2}\eta$

$$\|g_{(j+1)}-g_{(j)}\|_n \leqslant 2.2^{-jp/2}\delta+(j+1)2^{-jp/2}\delta(\log\frac{1}{\delta})^{1/2}\|G\|(\mathbb{E}|\epsilon|^{2p})^{1/2} \quad (4.13)$$

$$\leqslant C_0(j+1)2^{-jp/2}\delta(\log\frac{1}{\delta})^{1/2},$$

for some constant C_0.

Take $E=\sum_{j=0}^{r-1}\frac{1}{(j+1)^2}$ and $\eta_j=\frac{\eta}{2E(j+1)^2}$. Then $\sum_{j=0}^{r-1}\eta_j=\frac{\eta}{2}$ and by (4.13)

$$\mathbb{P}\left(\sup_{\substack{g\in\mathcal{G}\\ g_{(j)}=g_{(j)}(g)\\ g_{(j+1)}=g_{(j+1)}(g)}}|\sum_{j=0}^{r-1}n^{1/2}\prec\epsilon,g_{(j+1)}-g_{(j)}\succ_n^0|>\tfrac{1}{2}\eta\right) \quad (4.14)$$

$$\leqslant\sum_{j=0}^{r-1}\mathbb{P}\left(\sup_{\substack{g\in\mathcal{G}\\ g_{(j)}=g_{(j)}(g)\\ g_{(j+1)}=g_{(j+1)}(g)\\ \|g_{(j+1)}-g_{(j)}\|_n\leqslant C_0(j+1)2^{-jp/2}\delta(\log\frac{1}{\delta})^{1/2}}}|n^{1/2}\prec\epsilon,g_{(j+1)}-g_{(j)}\succ_n^0|>\eta_j\right)+\tfrac{1}{2}\eta.$$

By Hoeffding's inequality (Lemma 2.3.1), for

$$\|g_{(j+1)}-g_{(j)}\|_n\leqslant C_0(j+1)2^{-jp/2}\delta(\log\frac{1}{\delta})^{1/2},$$

$$\mathbb{P}(|n^{1/2}\prec\epsilon,g_{(j+1)}-g_{(j)}\succ_n^0|>\eta_j|(\mathbf{x}_1,\epsilon_1),\cdots,(x_n,\epsilon_n))$$

$$\leqslant 2\exp\left[-\frac{(\frac{1}{2}\eta/\{E(j+1)^2\})^2}{2(2^{j+1}\delta^{-\frac{2}{p}})(C_0^2(j+1)^2 2^{-jp}\delta^2\log\frac{1}{\delta})}\right]\leqslant 2\exp(-C_1\eta^2 2^{j\beta/2}\delta^{-\beta/p}),$$

for some constant C_1 and with $p\nu<\beta<2p-2$. Thus, on the set with $\|G\|_n\leqslant 2\|G\|$,

$$\mathbb{P}\left(\sup_{\|g_{(j+1)}-g_{(j)}\|_n\leqslant C_0(j+1)2^{-jp/2}\delta(\log\frac{1}{\delta})^{1/2}}|n^{1/2}\prec\epsilon,g_{(j+1)}-g_{(j)}\succ_n^0|>\eta_j|(x_1,\epsilon_1),\cdots,(x_n,\epsilon_n)\right)$$

$$\leqslant N_2^2(2^{-(j+1)p/2}\delta,H_n,\mathcal{G})2\exp(-C_1\eta^2 2^{j\beta/2}\delta^{-\beta/p})$$

$$\leqslant\exp(4M\delta^{-\nu}2^{jp\nu/2}\|G\|^{\nu})2\exp(-C_1\eta^2 2^{j\beta/2}\delta^{-\beta/p})\leqslant 2\exp(-C_3\eta^2 2^{jp\nu/2}\delta^{-\nu}),$$

for some C_2, C_3 and all $\delta>0$ sufficiently small. Insert this in (4.14) to see that for n sufficiently large

$$\mathbb{P}\left[\sup_{\substack{g_{(j)}=g_{(j)}(g)\\ g_{(j+1)}=g_{(j+1)}(g)\\ g\in\mathcal{G}}} |\sum_{j=0}^{r-1} n^{1/2}\prec\epsilon,g_{(j+1)}-g_{(j)}\succ_n^0|>\frac{1}{2}\eta\right]$$

$$\leqslant \sum_{j=0}^{r-1} 2\exp(-C_3\eta^2 2^{jp\nu/2}\delta^{-\nu})+\frac{1}{2}\eta<\eta,$$

for δ sufficiently small.

Representation (4.12) now shows that

$$\mathbb{P}^{(1)} \leqslant \mathbb{P}(\sup_g |n^{1/2}\prec\epsilon,g-g_{(r)}(g)\succ_n^0|>\frac{1}{2}\eta)+\eta. \tag{4.15}$$

But

$$|n^{1/2}\prec\epsilon,g-g_{(r)}(g)\succ_n^0| \leqslant n^{1/2}\|\epsilon\|_n\|g-g^{(r)}(g)\|_n \tag{4.16}$$
$$+n^{1/2}\prec|\epsilon|1_{|\epsilon|>2^{r/2}\delta^{-1/p}},G\succ_n^0.$$

Since $2^{rp}\geqslant n$ and $\|g-g^{(r)}(g)\|_n\leqslant 2^{-rp/2}\delta$,

$$n^{1/2}\|\epsilon\|_n\|g-g^{(r)}(g)\|_n\leqslant 2^{rp/2}2^{-rp/2}\delta\|\epsilon\|_n\leqslant\frac{1}{4}\eta, \tag{4.17}$$

for $\delta<\frac{1}{8}\eta$ and all n sufficiently large. Also,

$$n^{1/2}\mathbb{E}|\epsilon|1_{|\epsilon|>2^{r/2}\delta^{-1/p}}\mathbb{E}(G(\mathbf{x}))\leqslant n^{1/2}2^{r/2}\delta^{-1/p}2^{-rp}\delta^2\mathbb{E}|\epsilon|^{2p}\mathbb{E}(G(\mathbf{x}))\leqslant\frac{1}{8}\eta,$$

for δ small. Thus for the second term on the right hand side of (4.16) we have

$$\mathbb{P}\ (n^{1/2}|\prec|\epsilon|1_{|\epsilon|>2^{r/2}\delta^{-1/p}},G\succ_n^0|>\frac{1}{4}\eta) \tag{4.18}$$

$$\leqslant(4/\eta)^2 2^{r(1-p)}\delta^{2-2/p}\mathbb{E}|\epsilon|^{2p}\|G\|^2<\eta,$$

for δ small enough. Combination of (4.16), (4.17) and (4.18) gives

$$\mathbb{P}(\sup_g |n^{1/2}\prec\epsilon,g-g^{(r)}(g)\succ_n^0|>\frac{1}{2}\eta)<\eta,$$

and it follows from (4.15) that $\mathbb{P}^{(1)}<2\eta$.

It remains to show that $\mathbb{P}^{(2)}\leqslant\eta$, where

$$\mathbb{P}^{(2)} = \mathbb{P}(\sup_{\|g_{(0)}-\tilde{g}_{(0)}\|_n<3\delta} |n^{1/2}\prec\epsilon,g_{(0)}-\tilde{g}_{(0)}\succ_n^0|>\eta).$$

Again by Hoeffding's inequality , we have that on the set $\|G\|_n\leqslant 2\|G\|$,

$$\mathbb{P}(\sup_{\|g_{(0)}-\tilde{g}_{(0)}\|_n<3\delta} |n^{1/2}\prec\epsilon,g_{(0)}-\tilde{g}_{(0)}\succ_n^0|>\eta|(x_1,\epsilon_1),\cdots,(x_n,\epsilon_n))$$

$$\leqslant N_2^2(\delta,H_n,\mathcal{G})2\exp(\frac{-\eta^2}{2\delta^{-2/p}9\delta^2})$$

$$\leqslant \exp(4M\delta^{-\nu}\|G\|^{\nu})2\exp\left(\frac{-\eta^2}{18\delta^{2-2/p}}\right) \leqslant 2\exp(-C_4\delta^{-\nu}) \leqslant \eta,$$

with C_4 some constant and with δ sufficiently small. □

4.3. Measurability II

We have specified $\nu_n(\cdot)$ as an element of some space $\mathcal{X}$ of functions on $\mathcal{H}$. The problem is that the supremum metric generally makes $\mathcal{X}$ into a nonseparable space. As a consequence , $\nu_n(\cdot)$ is not Borel measurable. Now, denote by $\mathcal{B}^p$ the σ-algebra on $\mathcal{X}$ that makes all finite-dimensional projections measurable and that contains all closed balls with centres $y \in \mathcal{X}$ that are uniformly continuous. E.g. in D[0,1], the space of functions on [0,1] that are right-continuous and have left hand limits, the σ-algebra generated by closed balls coincides with the smallest σ -algebra that makes all coordinate projections measurable. Denote by $(\Omega, \mathcal{E}, \mathbb{P})$ the underlying probability space. If $\mathcal{H}$ is permissible and separable for $\|\cdot\|$, then $\nu(\cdot)$ is $\mathcal{E}/\mathcal{B}^p$ - measurable (POLLARD (1984)). Then by definition the random process ν_n converges in law to some limiting process $\nu(\cdot)$ if

$$\mathbb{P}(g(\nu_n)) \to \mathbb{P}(g(\nu))$$

for all real continuous measurable functions g on $\mathcal{X}$.

If the limiting process $\nu(\cdot)$ concentrates on a separable set, then some important theorems for the Euclidean case (the *Continuous mapping theorem* and the *Almost sure representation theorem*) go through for the situation with more general $\mathcal{X}$. A separable set in $\mathcal{X}$ is for instance the set of bounded continuous functions. Now, the limiting distribution of $\nu_n(\cdot)$, if it exists, must be some Gaussian process on $\mathcal{H}$. If $\mathcal{H}$ is $G_p BUC$, then the limiting distribution of $\nu_n(\cdot)$ concentrates on a separable set.

DEFINITION. A permissible class $\mathcal{H}$ is a *Donsker* class if

(i) $\mathcal{H}$ is $G_p BUC$

(ii) $\nu_n(\cdot) \to \mathbf{G}_p(\cdot)$.

In Theorem 4.2.2 we have presented sufficient conditions for $\mathcal{H}$ to be a functional Donsker class. POLLARD (1982) assumed stochastic separability of the process $\nu_n(\cdot)$ (see Section 2.4) and only proved the Donsker property (not the functional Donsker property). Using the results of POLLARD (1984) and DUDLEY and PHILIPP (1983), one sees that stochastic separability of $\nu_n(\cdot)$ can be replaced by permissibility of $\mathcal{H}$, and that the word functional can be added.

5. Asymptotic theory in two-phase regression: identified case

5.1. Introduction

In this chapter, we study the model

$$\mathbf{y} = g(\mathbf{x}) + \boldsymbol{\epsilon} \tag{5.1}$$

with

$$g = \sum_{i=1,2} g^{(i)} 1_{A^{(i)}}$$

where $g^{(i)}$ is in the class of linear functions, i.e. $g^{(i)}(x) = g_{\theta^{(i)}}(x) = \alpha^{(i)} + x\beta^{(i)}$, $\theta^{(i)} = \binom{\alpha^{(i)}}{\beta^{(i)}} \in \mathbb{R}^{d+1}$, $i = 1,2$, and where $\{A^{(i)}\}_{i=1,2}$ forms a partition of $\mathbb{R}^d$. We write $\theta = \binom{\theta^{(1)}}{\theta^{(2)}}$ and $A = A^{(1)}$. The set A is an unknown parameter and it is assumed to be an element of a class $\mathcal{A}$ of subsets of $\mathbb{R}^d$. As in Section 3.4, we sometimes write $\mathcal{A}^{(1)} = \mathcal{A}$ and $\mathcal{A}^{(2)} = \{A^c : A \in \mathcal{A}\}$.

We are interested in conditions for asymptotic normality of the least squares estimator of θ, based on n copies of $(\mathbf{x},\mathbf{y})$. First, the continuous model is investigated. In this model,

$$\mathcal{A} = \{A(\gamma) = \{x : x\gamma \leqslant 1\} : \gamma \in \mathbb{R}^d\}$$

and the class of regression functions is

$$\mathcal{G} = \{g_{\theta,c}(x) = \min(\alpha^{(1)} + x\beta^{(1)}, \alpha^{(2)} + x\beta^{(2)}) : \theta^{(i)} = \begin{bmatrix} \alpha^{(i)} \\ \beta^{(i)} \end{bmatrix} \in \mathbb{R}^{d+1}, \ i = 1,2\}. \tag{5.2}$$

Thus, in this model the least squares estimator of $A(\gamma)$ is a function of the estimator of θ. Also, a discontinuous model is considered, where

$$\mathcal{G} = \{g_{\theta,A}(x) = \sum_{i=1,2} (\alpha^{(i)} + x\beta^{(i)}) 1_{A^{(i)}}(x), \ \theta^{(i)} = \begin{bmatrix} \alpha^{(i)} \\ \beta^{(i)} \end{bmatrix} \in \mathbb{R}^{d+1}, \tag{5.3}$$

$$A^{(i)} \in \mathcal{A}^{(i)}, \ i = 1,2\}$$

and where $\mathcal{A}$ is e.g. $\{A(\gamma) = \{x : x\gamma \leqslant 1\} : \gamma \in \mathbb{R}^d\}$, but also more general classes are allowed. Note that in this model, the sets A are actually unknown parameters. We also derive the asymptotic distribution of the estimator of A.

Let

$$g_0 = \sum_{i=1,2} g_{\theta_0^{(i)}} 1_{A_0^{(i)}}$$

be the true underlying regression function. Throughout, the identifiability condition $\|\theta_0^{(1)} - \theta_0^{(2)}\| \neq 0$ is imposed. Moreover, for model (5.3) it is assumed that g_0 is discontinuous in some sense. Chapter 6 treats the situation where $d = 1$, $\|\theta_0^{(1)} - \theta_0^{(2)}\| = 0$ and also the case where $d = 1$ and g_0 continuous, but the continuity is not taken into account in the estimation procedure.

5.2. The continuous model

The regression function is assumed to be of the form

$$g_{\theta,c}(x) = \min(\alpha^{(1)}+x\beta^{(1)},\ \alpha^{(2)}+x\beta^{(2)}), \quad \theta=\begin{bmatrix}\theta^{(1)}\\ \theta^{(2)}\end{bmatrix},\ \theta^{(i)}=\begin{bmatrix}\alpha^{(i)}\\ \beta^{(i)}\end{bmatrix}.$$

Define

$$A_\theta = A_\theta^{(1)} = \{x\colon \alpha^{(1)}+x\beta^{(1)} \leqslant \alpha^{(2)}+x\beta^{(2)}\}$$

and $A_\theta^{(2)} = A_\theta^c$. Write $A_0^{(i)} = A_{\theta_0}^{(i)}$ and $\hat{\mathbf{A}}_n^{(i)} = A_{\hat{\theta}_n}^{(i)}$, $i=1,2$.

Theorem 5.2.1 below asserts that the asymptotic distribution of $\hat{\boldsymbol{\theta}}_n$ does not differ from the asymptotic distribution of the least squares estimator for the case $A_0^{(i)}$ known but without continuity restriction. In the latter situation the of regression functions are of the form

$$\sum_{i=1,2} (\alpha^{(i)}+x\beta^{(i)})1_{A_0^{(i)}}(x) \tag{5.4}$$

with $\binom{\alpha^{(i)}}{\beta^{(i)}}$, $i=1,2$ unknown parameters, i.e. the regression functions are not necessarily continuous. The conditions of Theorem 5.2.1 are those of Lemma 3.4.4 plus the assumption that an arbitrary higher order moment of $|\epsilon|^2$ exists.

THEOREM 5.2.1. *Suppose that the conditions of Lemma 3.4.4 are met:*

$$(i) \qquad \lim_{\eta \downarrow 0} \sup_{C\in\mathcal{C}} H(\{x\colon 0<d(x,C)\leqslant\eta\}) = 0 \tag{5.5}$$

where $\mathcal{C}$ is the class of hyperplanes in $\mathbb{R}^d$ and $d(\cdot,\cdot)$ is the Hausdorff distance,

(ii) there exist

$$\{x_t^{(i)}\colon t=1,\ \cdots,2(d+1)-1\} \subset A_0^{(i)} \cap T, \tag{5.6}$$

where T is the support of H, such that no $d+1$ $x_t^{(i)}$ lie on a $(d-1)$-dimensional hyperplane, $i=1,2$,

$$(iii) \qquad H(\{x\colon g_{\theta_0^{(1)}}(x)=g_{\theta_0^{(2)}}\})=0 \tag{5.7}$$

and

$$(iv) \qquad \|\theta_0^{(1)}-\theta_0^{(2)}\|\neq 0. \tag{5.8}$$

Assume that $\mathbb{E}|\epsilon|^{2p}<\infty$ for some $p>1$. Then $\hat{\boldsymbol{\theta}}_n^{(1)}$ and $\hat{\boldsymbol{\theta}}_n^{(2)}$ are asymptotically independent, with limiting distribution

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n^{(i)}-\theta_0^{(i)}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \|\epsilon\|^2\Sigma^{-1}(A_0^{(i)})), \quad i=1,2. \tag{5.9}$$

PROOF. The functions $g_{\theta,c}(x)=\min(g_{\theta^{(1)}}(x), g_{\theta^{(2)}}(x))$ are Lipschitz continuous in θ for every x, i.e. at θ_0

$$|g_{\theta,c}(x)-g_{\theta_0,c}(x)| \leqslant J(x)\|\theta-\theta_0\|,$$

where $J(x)=1+|z_1|+\cdots+|z_d|$, with $z_1,\cdots,z_d$ the coordinates of $x\in\mathbb{R}^d$. Consider the functions

$$j_\theta(x) = \begin{cases} \dfrac{g_{\theta,c}(x)-g_{\theta_0,c}(x)}{\|\theta-\theta_0\|} & \text{if} \quad \|\theta-\theta_0\|\neq 0 \\ 1 & \text{otherwise.} \end{cases}$$

These functions form a VC-graph class $\mathcal{J}=\{j_\theta\colon \theta\in\mathbb{R}^{2(d+1)}\}$ with envelope $J\in L^2(\mathbb{R}^d, H)$. Thus (Theorem 2.2.4)

$$\sup_\theta |\|j_\theta\|_n - \|j_\theta\|| \to 0 \quad \text{almost surely.} \tag{5.10}$$

But

$$\|j_{\hat{\boldsymbol{\theta}}_n}\|^2 = \frac{\|g_{\hat{\boldsymbol{\theta}}_n,c}-g_{\boldsymbol{\theta}_0,c}\|^2}{\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\|^2} \geqslant \sum_{i=1,2} \frac{(\hat{\boldsymbol{\theta}}_n^{(i)}-\boldsymbol{\theta}_0^{(i)})^T \Sigma(\hat{\mathbf{A}}_n^{(i)} \cap A_0^{(i)})(\hat{\boldsymbol{\theta}}_n^{(i)}-\boldsymbol{\theta}_0^{(i)})}{\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\|^2},$$

and since $\Sigma(\hat{\mathbf{A}}_n^{(i)} \cap A_0^{(i)}) \to \Sigma(A_0^{(i)})$ and $\Sigma(A_0^{(i)})$ is of full rank, this implies that

$$\|j_{\hat{\boldsymbol{\theta}}_n}\|_n \geqslant K_1 \tag{5.11}$$

for some constant $K_1>0$.

Write

$$\begin{aligned} 0 \geqslant \|y-\hat{\mathbf{g}}_n\|_n^2 - \|\epsilon\|_n^2 &= -2\prec \epsilon,\ \hat{\mathbf{g}}_n - g_0 \succ_n + \|\hat{\mathbf{g}}_n - g_0\|_n^2 \qquad (5.12) \\ &= -2\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\| \prec \epsilon,\ j_{\hat{\boldsymbol{\theta}}_n} \succ_n + \|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\|^2 \|j_{\hat{\boldsymbol{\theta}}_n}\|_n^2. \end{aligned}$$

Because $\mathcal{J}$ is a VC-graph class

$$D_2(\delta,\mathcal{J}) = \sup_S N_2(\delta\|J\|_s, H_s, \mathcal{J}) \leqslant \exp(M\delta^{-\nu})$$

for all $\nu>0$ and some M (see Theorem 2.2.6). Take $\nu<2-2/p$ and conclude from Theorem 4.2.3 (take $\mathcal{G}=\mathcal{J}$ and $g_0\equiv 0$ in this theorem) that $\{\epsilon j_\theta(x)\colon \theta\in\mathbb{R}^{2(d+1)}\}$ is a (functional) Donsker class, which implies

$$\prec \epsilon, j_{\hat{\boldsymbol{\theta}}_n} \succ_n = \mathcal{O}_{\mathbf{P}}(n^{-1/2}). \tag{5.13}$$

Insert (5.13) in (5.12) to obtain that

$$-2\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\| \mathcal{O}_{\mathbf{P}}(n^{-1/2}) + \|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\|^2 \|j_{\hat{\boldsymbol{\theta}}_n}\|_n^2 \leqslant 0,$$

or

$$\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\| \|j_{\hat{\boldsymbol{\theta}}_n}\|_n^2 = \mathcal{O}_{\mathbf{P}}(n^{-1/2}).$$

Hence by (5.11)

$$\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\| = \mathcal{O}_{\mathbf{P}}(n^{-1/2}).$$

Let $\tilde{\boldsymbol{\theta}}_n$ be a $\sqrt{n}$-consistent estimator and define $\tilde{\mathbf{A}}_n^{(i)} = \mathbf{A}_{\tilde{\boldsymbol{\theta}}_n}^{(i)}$, $i=1,2$. By (5.7),

we have $H(\tilde{\mathbf{A}}_n \Delta A_0) \xrightarrow{\mathbf{P}} 0$. Thus, again because of the VC-graph property of all classes of functions involved

$$\begin{aligned}
&\|(g_{\tilde{\boldsymbol{\theta}}_n,c} - g_0) 1_{\tilde{\mathbf{A}}_n^{(i)}}\|_n^2 \\
&= (\tilde{\boldsymbol{\theta}}_n^{(i)} - \boldsymbol{\theta}_0^{(i)})^T \Sigma_n(\tilde{\mathbf{A}}_n^{(i)} \cap A_0^{(i)})(\tilde{\boldsymbol{\theta}}_n^{(i)} - \boldsymbol{\theta}_0^{(i)}) + \|\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|^2 \|j_{\tilde{\boldsymbol{\theta}}_n} 1_{\tilde{\mathbf{A}}_n^{(i)} \setminus A_0^{(i)}}\|_n^2 \\
&= (\tilde{\boldsymbol{\theta}}_n^{(i)} - \boldsymbol{\theta}_0^{(i)})^T \Sigma(A_0^{(i)})(\tilde{\boldsymbol{\theta}}_n^{(i)} - \boldsymbol{\theta}_0^{(i)}) + o_{\mathbf{P}}(\frac{1}{n}), \quad i = 1,2.
\end{aligned}$$

Similarly, since the Donsker property implies asymptotic equicontinuity

$$\begin{aligned}
&\prec \epsilon, (g_{\tilde{\boldsymbol{\theta}}_n,c} - g_0) 1_{\tilde{\mathbf{A}}_n^{(i)}} \succ_n \\
&= \left[\frac{1}{n} \sum_{\mathbf{x}_k \in \tilde{\mathbf{A}}_n^{(i)} \cap A_0^{(i)}} \epsilon_k(1, \mathbf{x}_k)\right](\tilde{\boldsymbol{\theta}}_n^{(i)} - \boldsymbol{\theta}_0^{(i)}) + \|\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \left[\frac{1}{n} \sum_{\mathbf{x}_k \in \tilde{\mathbf{A}}_n^{(i)} \setminus A_0^{(i)}} \epsilon_k j_{\tilde{\boldsymbol{\theta}}_n}(\mathbf{x}_k)\right] \\
&= \left[\frac{1}{n} \sum_{\mathbf{x}_k \in A_0^{(i)}} \epsilon_k(1, \mathbf{x}_k)\right](\tilde{\boldsymbol{\theta}}_n^{(i)} - \boldsymbol{\theta}_0^{(i)}) + o_{\mathbf{P}}(\frac{1}{n}), \quad i = 1,2.
\end{aligned}$$

Thus, if we write $\|\boldsymbol{\Sigma}^{1/2}(A_0^{(i)})a\|^2 = a^T \boldsymbol{\Sigma}(A_0^{(i)})a$, $a \in \mathbb{R}^{d+1}$, and

$$\left[\frac{1}{n} \sum_{\mathbf{x}_k \in A_0^{(i)}} \epsilon_k(1, \mathbf{x}_k)^T\right] = \mathbf{N}_n^{(i)} \in \mathbb{R}^{d+1}, \quad i = 1,2,$$

$$\begin{aligned}
&\|y - g_{\tilde{\boldsymbol{\theta}}_n,c}\|_n^2 - \|\epsilon\|_n^2 \\
&= \sum_{i=1,2} \{-2\mathbf{N}_n^{(i)T}(\tilde{\boldsymbol{\theta}}_n^{(i)} - \boldsymbol{\theta}_0^{(i)}) + \|\Sigma^{1/2}(A_0^{(i)})(\tilde{\boldsymbol{\theta}}_n^{(i)} - \boldsymbol{\theta}_0^{(i)})\|\} + o_{\mathbf{P}}(\frac{1}{n}).
\end{aligned} \tag{5.14}$$

In particular (5.14) holds for $\tilde{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n$, so that

$$\begin{aligned}
0 \geqslant &\|y - \hat{\mathbf{g}}_n\|_n^2 - \|\epsilon\|_n^2 \\
&= \sum_{i=1,2} \{\|\Sigma^{1/2}(A_0^{(i)})(\hat{\boldsymbol{\theta}}_n^{(i)} - \boldsymbol{\theta}_0^{(i)}) - \Sigma^{-1/2}(A_0^{(i)})\mathbf{N}_n^{(i)}\|^2 \\
&\quad - \|\Sigma^{-1/2}(A_0^{(i)})\mathbf{N}_n^{(i)}\|^2\} + o_{\mathbf{P}}(\frac{1}{n}).
\end{aligned} \tag{5.15}$$

If we take $\tilde{\boldsymbol{\theta}}_n^{(i)} = \boldsymbol{\theta}_0^{(i)} + \Sigma^{-1}(A_0^{(i)})\mathbf{N}_n^{(i)}$, $i = 1,2$, we get from (5.14)

$$\|y - g_{\tilde{\boldsymbol{\theta}}_n,c}\|_n - \|\epsilon\|_n^2 = -\sum_{i=1,2} \|\Sigma^{-1/2}\mathbf{N}_n^{(i)}\|^2 + o_{\mathbf{P}}(\frac{1}{n}). \tag{5.16}$$

Since $\|y - \hat{\mathbf{g}}_n\|_n \leqslant \|y - g_{\tilde{\boldsymbol{\theta}}_n,c}\|_n$ combination of (5.15) and (5.16) yields

$$\sum_{i=1,2} \{\|\Sigma^{1/2}(A_0^{(i)})(\hat{\boldsymbol{\theta}}_n^{(i)} - \boldsymbol{\theta}_0^{(i)}) - \Sigma^{-1/2}(A_0^{(i)})\mathbf{N}_n^{(i)}\|^2\} = o_{\mathbf{P}}(\frac{1}{n}),$$

or

$$\Sigma^{1/2}(A_0^{(i)})(\hat{\boldsymbol{\theta}}_n^{(i)} - \boldsymbol{\theta}_0^{(i)}) = \Sigma^{-1/2}(A_0^{(i)})\mathbf{N}_n^{(i)} + o_{\mathbf{P}}(n^{-1/2}), \quad i = 1,2.$$

Because $\sqrt{n}\,\mathbf{N}_n^{(i)} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \|\epsilon\|^2 \Sigma(A_0^{(i)}))$, $i = 1,2$, this proves the required result. □

Condition (5.7) ensures that $\theta \to \theta_0$ implies $H(A_\theta \Delta A_0) \to 0$. If it is not fulfilled, then a large fraction of observations is concentrated at $\{x: g_{\theta_0^{(1)}}(x) = g_{\theta_0^{(2)}}(x)\}$, and the $\hat{\boldsymbol{\theta}}_n^{(i)}$ will no longer be asymptotically independent. This situation can be compared with the case A_0 known, where because of the continuity restriction, the least squares estimators of $\theta^{(1)}$ and $\theta^{(2)}$ are also not independent.

The object of study in FEDER (1975) is a continuous model with one-dimensional change-point, i.e. $d = 1$ and

$$\mathbf{y} = g_{\theta^{(1)}}(\mathbf{x}) 1_{(-\infty,\gamma]}(\mathbf{x}) + g_{\theta^{(2)}}(\mathbf{x}) 1_{[\gamma,\infty)}(\mathbf{x}) + \boldsymbol{\epsilon},$$

where the $g_{\theta^{(i)}}$ are linear in $\theta^{(i)}$, $i = 1,2$, and satisfy $g_{\theta^{(1)}}(\gamma) = g_{\theta^{(2)}}(\gamma)$. Feder obtains asymptotic normality of the least squares estimators, under identifiability conditions. His method of proof makes extensive use of the special structure of the class $\mathcal{A} = \{(-\infty, \gamma]: \gamma \in \mathbb{R}\}$ of subsets of $\mathbb{R}$. Extension of Feder's method to two-phase regression models with sets in higher-dimensional Euclidean space as unknown parameters appears to be cumbersome.

5.3. The discontinuous model

In this section, we deal with two-phase regression functions of the form

$$g_{\theta,A}(x) = \begin{cases} g_{\theta^{(1)}}(x) & \text{if } x \in A \\ g_{\theta^{(2)}}(x) & \text{if } x \notin A \end{cases} \tag{5.17}$$

with $g_{\theta^{(i)}}(x) = \alpha^{(i)} + x\beta^{(i)}$, $\theta^{(i)} = \begin{pmatrix} \alpha^{(i)} \\ \beta^{(i)} \end{pmatrix}$ and with $A \in \mathcal{A}$. The class $\mathcal{A}$ may be the class $\{A(\gamma) = \{x: x\gamma \leqslant 1\}: \gamma \in \mathbb{R}^d\}$, but we shall not require this because it turns out that also other classes can be handled without too much increase of complexity.

We assume again that the conditions for consistency of $\hat{\boldsymbol{\theta}}_n$ and $\hat{\mathbf{A}}_n$ are fulfilled, i.e. $\mathcal{A}$ is a permissible class satisfying the entropy condition

$$\frac{1}{n} \log N_2(\delta, \mathbf{H}_n, \mathcal{A}) \xrightarrow{\mathbf{P}^*} 0 \quad \text{for all} \quad \delta > 0, \tag{5.18}$$

and moreover

$$\lim_{\eta \downarrow 0} \sup_{C \in \mathcal{C}} H(\{x: 0 < d(x,C) \leqslant \eta\}) = 0 \tag{5.19}$$

and there exist

$$\{x_t^{(i)}: t = 1, \cdots, 2(d+1)-1\} \subset A_0^{(i)} \cap T, \quad i = 1,2, \tag{5.20}$$

with no $d+1$ $x_t^{(i)}$ on a $(d-1)$-dimensional hyperplane,

$$\|\boldsymbol{\theta}_0^{(1)} - \boldsymbol{\theta}_0^{(2)}\| \neq 0, \quad H(\{x: g_{\theta_0^{(1)}}(x) = g_{\theta_0^{(2)}}(x)\}) = 0. \tag{5.21}$$

Here is a description of discontinuity at the true parameter value.

DISCONTINUITY ASSUMPTION. There exists an $\eta>0$ such that

$$\inf_{A\in\mathcal{A}:\ 0<H(A\Delta A_0)<\eta} \mathbb{E}(g_{\theta_0^{(1)}}(\mathbf{x})-g_{\theta_0^{(2)}}(\mathbf{x})|\mathbf{x}\in A\Delta A_0)>0. \tag{5.22}$$

EXAMPLE 5.1. Take $d=1$ and $\mathcal{A}=\{(-\infty,\gamma]:\ \gamma\in\mathbb{R}\}$. If $\alpha_0^{(1)}+\gamma_0\beta_0^{(1)}\neq\alpha_0^{(2)}+\gamma_0\beta_0^{(2)}$, and if H puts positive mass on some interval around γ_0, then the discontinuity assumption is satisfied.

In the discontinuous model, with discontinuity assumption, the least squares estimators $\hat{\boldsymbol{\theta}}_n^{(1)}$ and $\hat{\boldsymbol{\theta}}_n^{(2)}$ are asymptotically independent and asymptotically equivalent to the least squares estimators of the $\boldsymbol{\theta}^{(i)}, i=1,2$ in the case A_0 known. This is asserted in Theorem 5.3.2, and the result is called *adaptation*: the fact that A_0 is unknown has asymptotically no influence on the estimators of the $\boldsymbol{\theta}^{(i)}, i=1,2$.

THEOREM 5.3.2. *Suppose that the conditions (5.19), . . , (5.21) and the discontinuity assumption are fulfilled, and that (5.18) can be strengthened to*

$$D_2(\delta,\mathcal{A}) = \sup_S N_2(\delta,H_s,\mathcal{A}) \leqslant \exp(M\delta^{-\nu}) \tag{5.23}$$

for some constants M and $0<\nu<2$. Assume that $\mathbf{x}1_{A\Delta A_0}(\mathbf{x})$ is bounded uniformly in $A\in\mathcal{A}$, A in a neighbourhood of A_0, i.e. there exists a constant $K_0<\infty$ such that for some $\eta_0>0$

$$H(\{x:\ \sup_{A\in\mathcal{A}:\ H(A\Delta A_0)<\eta_0} |x1_{A\Delta A_0}(x)|>K_0\}) = 0. \tag{5.24}$$

Finally, assume that $\mathbb{E}|\epsilon|^{2p}<\infty$ for some $p>2/(2-\nu)$. Then $\hat{\boldsymbol{\theta}}_n^{(1)}$ and $\hat{\boldsymbol{\theta}}_n^{(2)}$ are asymptotically independent with limiting distribution

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n^{(i)}-\boldsymbol{\theta}_0^{(i)}) \xrightarrow{\mathcal{L}} \mathcal{N}(0,\|\epsilon\|^2\Sigma^{-1}(A_0^{(i)})),\quad i=1,2. \tag{5.25}$$

PROOF. We shall first show that $H(\hat{\mathbf{A}}_n\Delta A_0)=\mathcal{O}_{\mathbf{P}}(n^{-1/2})$. Of course, for $A=A_0$ fixed the class

$$\{\epsilon(g_{\theta^{(i)}}-g_{\theta_0^{(i)}})1_{A_0^{(i)}}:\ \boldsymbol{\theta}^{(i)} \text{ in a neighbourhood of } \boldsymbol{\theta}_0^{(i)}\}$$

is a Donsker class, $i=1,2$. Since $\hat{\boldsymbol{\theta}}_n^{(i)}\to\boldsymbol{\theta}_0^{(i)}$ this implies that

$$|\prec\epsilon,(g_{\hat{\boldsymbol{\theta}}_n^{(i)}}-g_{\boldsymbol{\theta}_0^{(i)}})1_{A_0^{(i)}}\succ_n| = o_{\mathbf{P}}(n^{-1/2}),\quad i=1,2.$$

By (5.23) and (5.24) and using the assumption that $\mathbb{E}|\epsilon|^{2p}<\infty$, we see that also the class

$$\{\epsilon(g_{\theta^{(i)}}-g_{\theta_0^{(j)}})1_{A\Delta A_0}:\ \boldsymbol{\theta}^{(i)} \text{ in a neighbourhood of } \boldsymbol{\theta}_0^{(j)},\ A\in\mathcal{A},$$
$$H(A\Delta A_0)<\eta_0\}$$

is a Donsker class, $i,\ j\in\{1,2\}$. Hence, since $H(\hat{\mathbf{A}}_n\Delta A_0)\to 0$,

$$|\prec\epsilon,(g_{\hat{\boldsymbol{\theta}}_n^{(i)}}-g_{\boldsymbol{\theta}_0^{(i)}})1_{A_0^{(i)}\setminus\hat{\mathbf{A}}_n^{(i)}}\succ_n| \;=\; o_{\mathbf{P}}(n^{-1/2}),\quad i=1,2,$$

as well as

$$|\prec\epsilon,(g_{\hat{\boldsymbol{\theta}}_n^{(i)}}-g_{\boldsymbol{\theta}_0^{(j)}})1_{\hat{\mathbf{A}}_n^{(i)}\setminus A_0^{(j)}}\succ_n| \;=\; o_{\mathbf{P}}(n^{-1/2}),\quad i\neq j\in\{1,2\}.$$

But then also

$$\begin{aligned}|\prec\epsilon,\hat{\mathbf{g}}_n-g_0\succ_n| &\leqslant \sum_{i=1,2}|\prec\epsilon,(g_{\hat{\boldsymbol{\theta}}_n^{(i)}}-g_{\boldsymbol{\theta}_0^{(i)}})1_{A_0^{(i)}}\succ_n|\\ &\quad+\sum_{i=1,2}|\prec\epsilon,(g_{\hat{\boldsymbol{\theta}}_n^{(i)}}-g_{\boldsymbol{\theta}_0^{(i)}})1_{A_0^{(i)}\setminus\hat{\mathbf{A}}_n^{(i)}}\succ_n|\\ &\quad+\sum_{i\neq j\in\{1,2\}}|\prec\epsilon,(g_{\hat{\boldsymbol{\theta}}_n^{(i)}}-g_{\boldsymbol{\theta}_0^{(j)}})1_{\hat{\mathbf{A}}_n^{(i)}\setminus A_0^{(j)}}\succ_n| \;=\; o_{\mathbf{P}}(n^{-1/2}).\end{aligned}$$

This shows that

$$\|(\hat{\mathbf{g}}_n-g_0)1_{\hat{\mathbf{A}}_n\Delta A_0}\|_n^2\leqslant\|\hat{\mathbf{g}}_n-g_0\|_n^2\leqslant 2|\prec\epsilon,\hat{\mathbf{g}}_n-g_0\succ_n| \;=\; o_{\mathbf{P}}(n^{-1/2}). \tag{5.26}$$

Assumptions (5.23) and (5.24) also imply that the class

$$\begin{aligned}&\{(g_{\theta^{(i)}}-g_{\boldsymbol{\theta}_0^{(j)}})^2 1_{A\Delta A_0}\colon \theta \text{ in a neighbourhood of } \boldsymbol{\theta}_0,\ A\in\mathcal{A},\\ &H(A\Delta A_0)<\eta_0\},\quad i,j\in\{1,2\}\end{aligned}$$

is Donsker, so from (5.26)

$$\|(\hat{\mathbf{g}}_n-g_0)1_{\hat{\mathbf{A}}_n\Delta A_0}\|^2 \;=\; \|(\hat{\mathbf{g}}_n-g_0)1_{\hat{\mathbf{A}}_n\Delta A_0}\|_n^2+\mathcal{O}_{\mathbf{P}}(n^{-1/2})=\mathcal{O}_{\mathbf{P}}(n^{-1/2}). \tag{5.27}$$

We shall now utilize the discontinuity assumption. In view of (5.24), for all n sufficiently large,

$$|g_{\hat{\boldsymbol{\theta}}_n^{(i)}}(x)-g_{\boldsymbol{\theta}_0^{(i)}}(x)|1_{\hat{\mathbf{A}}_n\Delta A_0}(x)\leqslant K_0^2\|\hat{\boldsymbol{\theta}}_n^{(i)}-\boldsymbol{\theta}_0^{(i)}\|^2.$$

Thus, for arbitrary $\eta_1>0$

$$\|(\hat{\mathbf{g}}_n-g_0)1_{\hat{\mathbf{A}}_n\Delta A_0}\|^2 \;=\; \|(g_{\boldsymbol{\theta}_0^{(1)}}-g_{\boldsymbol{\theta}_0^{(2)}})1_{\hat{\mathbf{A}}_n\Delta A_0}\|^2-\eta_1 H(\hat{\mathbf{A}}_n\Delta A_0),$$

for all n sufficiently large. Combine this with (5.22) to obtain that for some constant $K_1>0$

$$\|(\hat{\mathbf{g}}_n-g_0)1_{\hat{\mathbf{A}}_n\Delta A_0}\|^2\geqslant K_1 H(\hat{\mathbf{A}}_n\Delta A_0)$$

for all n sufficiently large. In view of (5.27), we thus obtain

$$H(\hat{\mathbf{A}}_n\Delta A_0) \;=\; \mathcal{O}_{\mathbf{P}}(n^{-1/2}).$$

Since $\boldsymbol{\Sigma}_n(\hat{\mathbf{A}}_n^{(i)})\to\boldsymbol{\Sigma}(A_0^{(i)})$ almost surely, we see by explicitly writing down the expression for the least squares estimator

$$(\hat{\boldsymbol{\theta}}_n^{(i)}-\boldsymbol{\theta}_0^{(i)}) \;=\; (\Sigma^{-1}(A_0^{(i)})+o(1))\left\{\frac{1}{n}\sum_{\mathbf{x}_k\in\hat{\mathbf{A}}_n^{(i)}}\epsilon_k(1,\mathbf{x}_k)^T\right.$$

$$-\frac{1}{n}\sum_{\mathbf{x}_k\in\hat{\mathbf{A}}_n^{(i)}\setminus A_0^{(i)}}(\hat{\mathbf{g}}_n-g_0)(1,\mathbf{x}_k)^T\Bigg\},\quad i=1,2. \tag{5.28}$$

By (5.24) and (5.26)

$$\frac{1}{n}\sum_{\mathbf{x}_k\in\hat{\mathbf{A}}_n^{(i)}\setminus A_0^{(i)}}(\hat{\mathbf{g}}_n-g_0)(1,\mathbf{x}_k)^T \leqslant \|\hat{\mathbf{g}}_n-g_0\|_n K_0\mathbf{H}_n^{1/2}(\hat{\mathbf{A}}_n\Delta A_0)$$

$$= o_{\mathbf{P}}(n^{-1/4})\mathbf{H}_n^{1/2}(\hat{\mathbf{A}}_n\Delta A_0).$$

But the Donsker-property for $\{A\Delta A_0\colon A\in\mathcal{A}\}$ and $H(\hat{\mathbf{A}}_n\Delta A_0)=\mathcal{O}_{\mathbf{P}}(n^{-1/2})$ imply $\mathbf{H}_n(\hat{\mathbf{A}}_n\Delta A_0)=\mathcal{O}_{\mathbf{P}}(n^{-1/2})$. Therefore, we can write (5.28) as

$$(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0^{(i)}) = (\Sigma^{-1}(A_0^{(i)})+o(1))\left\{\frac{1}{n}\sum_{\mathbf{x}_k\in\hat{\mathbf{A}}_n^{(i)}}\boldsymbol{\epsilon}_k(1,\mathbf{x}_k)^T+o_{\mathbf{P}}(n^{-1/2})\right\}$$

$$= \Sigma^{-1}(A_0^{(i)})\frac{1}{n}\sum_{\mathbf{x}_k\in\hat{\mathbf{A}}_n^{(i)}}\boldsymbol{\epsilon}_k(1,\mathbf{x}_k)^T+o_{\mathbf{P}}(n^{-1/2}),$$

because $\{\epsilon(1,x)^T 1_{A\Delta A_0}(x)\colon A\in\mathcal{A},\ H(A\Delta A_0)<\eta_0\}$ is also a Donsker class. □

We shall derive an expression for the limiting distribution of $\hat{\mathbf{A}}_n$. This limiting distribution does not depend on $\hat{\boldsymbol{\theta}}_n^{(i)}$, $i=1,2$.

LEMMA 5.3.3. *Under the conditions of Theorem 5.3.2*

$$\hat{\mathbf{A}}_n\Delta A_0 = \arg\sup_{A\in\mathcal{A}}\mathbf{R}_n(A)+o_{\mathbf{P}}(1),$$

with

$$\mathbf{R}_n(A) \tag{5.29}$$

$$= \sum_{\substack{i=1,2\\ j\neq i}}\left[2\sum_{A^{(i)}\setminus A_0^{(i)}}\boldsymbol{\epsilon}_k(g_{\theta_0^{(i)}}(\mathbf{x}_k)-g_{\theta_0^{(j)}}(\mathbf{x}_k))-\sum_{A^{(i)}\setminus A_0^{(i)}}(g_{\theta_0^{(i)}}(\mathbf{x}_k)-g_{\theta_0^{(j)}}(\mathbf{x}_k))^2\right].$$

PROOF. It is easy to see that for $\{A_n\}\subset\mathcal{A}$ some sequence of subsets in $\mathbb{R}^d$ and $\tau=\begin{pmatrix}\tau^{(1)}\\ \tau^{(2)}\end{pmatrix}$

$$n(\|\epsilon 1_{A_n^{(i)}\cap A_0^{(i)}}\|_n^2-\|(y-g_{(\theta_0+n^{-1/2}\tau),A_n})1_{A_n^{(i)}\cap A_0^{(i)}}\|_n^2)$$

$$= 2\frac{1}{\sqrt{n}}\sum_{\mathbf{x}_k\in A_n^{(i)}\cap A_0^{(i)}}\boldsymbol{\epsilon}_k(1,\mathbf{x}_k)\tau^{(i)}-\tau^{(i)T}\Sigma(A_n^{(i)}\cap A_0^{(i)})\tau^{(i)},\quad i=1,2.$$

Thus, using the Donsker-property

$$n(\|\epsilon 1_{A_n^{(i)}\cap A_0^{(i)}}\|_n^2-\|(y-g_{(\theta_0+n^{-1/2}\tau),A_n})1_{A_n^{(i)}\cap A_0^{(i)}}\|_n^2)$$

$$= 2\frac{1}{\sqrt{n}}\sum_{\mathbf{x}_k\in A_n^{(i)}}\boldsymbol{\epsilon}_k(1,\mathbf{x}_k)\tau^{(i)}-\tau^{(i)T}\Sigma(A_n^{(i)})\tau^{(i)}+o_{\mathbf{P}}(1),\quad i=1,2,$$

uniformly in $\|\tau\| \leqslant L$ and $\{A_n\} \subset \mathcal{A}$, $H(A_n \Delta A_0) \leqslant \eta_n \to 0$. Furthermore

$$n(\|\epsilon 1_{A_n^{(i)} \setminus A_0^{(i)}}\|_n^2 - \|(y - g_{(\theta_0 + n^{-1/2}\tau), A_n}) 1_{A_n^{(i)} \setminus A_0^{(i)}}\|_n^2)$$

$$= 2 \sum_{\mathbf{x}_k \in A_n^{(i)} \setminus A_0^{(i)}} \epsilon_k (g_{(\theta_0^{(i)} + \tau^{(i)}/\sqrt{n})}(\mathbf{x}_k) - g_{\theta_0^{(i)}}(\mathbf{x}_k))$$

$$- \sum_{\mathbf{x}_k \in A_n^{(i)} \setminus A_0^{(i)}} (g_{(\theta_0^{(i)} + \tau^{(i)}/\sqrt{n})}(\mathbf{x}_k) - g_{\theta_0^{(i)}}(\mathbf{x}_k))^2 = 2 \sum_{\mathbf{x}_k \in A_n^{(i)} \setminus A_0^{(i)}} \epsilon_k (g_{\theta_0^{(i)}}(\mathbf{x}_k) - g_{\theta_0^{(i)}}(\mathbf{x}_k))$$

$$- \sum_{\mathbf{x}_k \in A_n^{(i)} \setminus A_0^{(i)}} (g_{\theta_0^{(i)}}(\mathbf{x}_k) - g_{\theta_0^{(i)}}(\mathbf{x}_k))^2 + o_{\mathbf{P}}(1), \quad i = 1,2,$$

$i \neq j \in \{1,2\}$, uniformly in $\|\tau\| \leqslant L$, $\{A_n\} \subset \mathcal{A}$, $H(A_n \Delta A) \leqslant \eta_n \to 0$.

We have seen that with arbitrary large probability, $\sqrt{n}\|\hat{\boldsymbol{\theta}}_n - \theta_0\| \leqslant L$ for L and n sufficiently large. Moreover , $H(\hat{\mathbf{A}}_n \Delta A_0) \to 0$ implies that there exists a sequence $\eta_n \downarrow 0$ such that with arbitrary large probability $H(\hat{\mathbf{A}}_n \Delta A_0) \leqslant \eta_n$ for all n sufficiently large. Thus with arbitrary large probability

$$n(\|\epsilon\|_n^2 - \|y - \hat{\mathbf{g}}_n\|_n^2) = \sup_{\substack{\|\tau\| \leqslant L \\ A \in \mathcal{A}}} n(\|\epsilon\|_n^2 - \|y - g_{(\theta_0 + n^{-1/2}\tau), A}\|_n^2)$$

$$= \sup_{\substack{\|\tau\| \leqslant L \\ A \in \mathcal{A}:\ H(A \Delta A_0) \leqslant \eta_n}} \left\{ \sum_{i=1,2} \{2 \frac{1}{\sqrt{n}} \sum_{A_0^{(i)}} \epsilon_k (1, \mathbf{x}_k)^T \tau^{(i)} - \tau^{(i)T} \Sigma(A_0^{(i)}) \tau^{(i)}\} + \mathbf{R}_n(A) + o_{\mathbf{P}}(1) \right\}$$

$$= \sup_{\|\tau\| \leqslant L} \left\{ \sum_{i=1,2} \{2 \frac{1}{\sqrt{n}} \sum_{A_0^{(i)}} \epsilon_k (1, \mathbf{x}_k)^T \tau^{(i)}\} \right\} + \sup_{A \in \mathcal{A}} \mathbf{R}_n(A) + o_{\mathbf{P}}(1). \quad \square$$

EXAMPLE 5.2. Take $d = 1$, $\mathcal{A} = \{(-\infty, \gamma]: \gamma \in \mathbb{R}\}$ and $g_{\theta^{(i)}} = \alpha^{(i)}$. Then for $\gamma > \gamma_0$

$$\mathbf{R}_n((-\infty, \gamma]) = 2 \sum_{\gamma_0 < \mathbf{x}_k \leqslant \gamma} \epsilon_k (\alpha_0^{(1)} - \alpha_0^{(2)}) - (\alpha_0^{(1)} - \alpha_0^{(2)})^2 n \mathbf{H}_n(\gamma_0, \gamma].$$

Apply the law of the iterated logarithm for partial sums to see that conditionally on $\mathbf{x}_1, \mathbf{x}_2, \cdots = x_1, x_2 \cdots$

$$\sum_{\gamma_0 < \mathbf{x}_k \leqslant \gamma} \epsilon_k (\alpha_0^{(1)} - \alpha_0^{(2)}) = \mathcal{O}((nH_n(\gamma_0, \gamma])^{1/2} \log\log(nH_n(\gamma_0, \gamma])^{1/2}),$$

uniformly in $nH_n((\gamma_0, \gamma]) \to 0$. Hence

$$\sup_{\gamma} \mathbf{R}_n((-\infty, \gamma]) = \mathcal{O}_{\mathbf{P}}(1),$$

and

$$H_n((-\infty, \hat{\gamma}_n] \Delta (-\infty, \gamma_0]) = \mathcal{O}_{\mathbf{P}}\left(\frac{1}{n}\right).$$

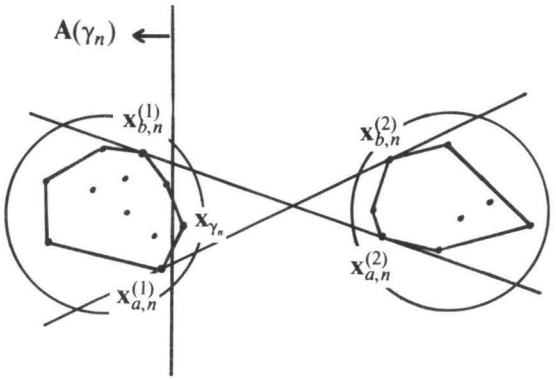
This result is comparable with HINKLEY (1970), who assumes normality of the ϵ_k, $k = 1,2, \cdots$.

In Example 5.2, we showed that $\sup_{A \in \mathcal{A}} \mathbf{R}_n(A) = \mathcal{O}_{\mathbf{P}}(1)$, and this in turn implies that $n(\|\epsilon\|_n^2 - \|y - \hat{\mathbf{g}}_n\|_n^2) = \mathcal{O}_{\mathbf{P}}(1)$ and $\|\hat{\mathbf{g}}_n - g_0\|_n = \mathcal{O}_{\mathbf{P}}(n^{-1/2})$. Here is an

example where $\sup_{a\in\mathcal{A}}\mathbf{R}_n(A)$ does not remain bounded in probability.

EXAMPLE 5.3. Take $d=2$, $\mathcal{A}=\{A(\gamma)=\{x: x\gamma\leqslant 1\},\ \gamma\in\mathbb{R}^2\}$ and $g_{\theta^{(i)}}=\alpha^{(i)}$, $i=1,2$. Let $A_0^{(1)}=\{x=(z_1,z_2): (z_1+2)^2+z_2^2\leqslant 1\}$ and let H be the uniform distribution on $A_0^{(1)}\cup A_0^{(2)}$, where $A_0^{(2)}=\{x=(z_1,z_2): (z_1-2)^2+z_2^2\leqslant 1\}$. Observe that if $\alpha_0^{(1)}\neq\alpha_0^{(2)}$, the discontinuity assumption is fulfilled. Also all further conditions of Theorem 5.3.2 hold, provided $\mathbb{E}|\epsilon|^{2p}<\infty$ for some $p>1$.

Now, consider the convex hull of the data $\{\mathbf{x}_1,\cdots,\mathbf{x}_n\}$ in $A_0^{(1)}$ and $A_0^{(2)}$ respectively. To every point $\mathbf{x}_{\gamma_n}$ on the convex hull, which lies between $\mathbf{x}_{a,n}^{(1)}$ and $\mathbf{x}_{b,n}^{(1)}$, there corresponds an $\mathbf{A}(\gamma_n)=A_0^{(1)}\setminus\{\mathbf{x}_{\gamma_n}\}$.



Obviously

$$\sup_{A\in\mathcal{A}}\mathbf{R}_n(A)\geqslant\sup_{\gamma_n}\mathbf{R}_n(\mathbf{A}(\gamma_n))=\max_{\gamma_n}2\epsilon_{\gamma_n}(\alpha_0^{(1)}-\alpha_0^{(2)})-(\alpha_0^{(1)}-\alpha_0^{(2)})^2. \tag{5.30}$$

As n tends to infinity, the number of points $\mathbf{x}_{\gamma_n}$ also tends to infinity, so the maximum in (5.30) will be taken over an increasing number of independent copies of ϵ. This maximum will not remain bounded.

6. Rates of convergence

6.1. Introduction

This chapter is inspired by LeCam (1973) and Birgé (1983). We shall first sketch some of their results.

Let $\mathcal{G}$ be an index set and $\{P_g: g \in \mathcal{G}\}$ a collection of probability measures on a Euclidean space. One can equip $\mathcal{G}$ with the Hellinger-metric, defined as

$$h(g,\tilde{g}) = \{\tfrac{1}{2}\int |(dP_g)^{1/2} - (dP_{\tilde{g}})^{1/2}|^2\}^{1/2}.$$

Let $\mathbf{g}_n^{ML}$ be the maximum likelihood estimator of g based on n independent observations from P_{g_0}. LeCam (1973) shows that if $\mathcal{G}$ satisfies certain dimensionality restrictions

$$h(\mathbf{g}_n^{ML}, g_0) = \mathcal{O}_{\mathbf{P}}(n^{-1/2}).$$

These dimensionality restrictions are entropy conditions on $\mathcal{G}$ endowed with the Hellinger-metric.

Birgé (1983) investigates the minimax risk for estimation. For example, let P_g be the probability measure on $\mathbb{R}$ with density g with respect to Lebesgue measure and let $\mathcal{G}$ be a class of densities on $\mathbb{R}$. Define $d(g,\tilde{g}) = \int |g(x) - \tilde{g}(x)|dx$. The minimax risk is

$$R_n(d) = \inf_{\mathbf{T}_n} \sup_{g_0 \in \mathcal{G}} \mathbb{E}_{g_0}(d(\mathbf{T}_n, g_0)),$$

where $\mathbf{T}_n$ is any estimator of g_0. Denote by $\log N_d(\delta, \mathcal{G})$ the δ-entropy of $\mathcal{G}$ for d. Birgé shows that

$$\log N_d(\delta, \mathcal{G}) \leqslant M\delta^{-\nu} \quad \text{for all} \quad \delta > 0$$

implies

$$R_n(d) \leqslant M' n^{-\frac{1}{2+\nu}}.$$

In regression theory, least squares estimators coincide with maximum likelihood estimators if the disturbances are i.i.d. and normally distributed. Thus, in that case LeCam's theory can be applied to obtain conditions under which $\hat{\mathbf{g}}_n$ converges with rate $\mathcal{O}_{\mathbf{P}}(n^{-1/2})$ in the Hellinger-metric. We shall prove that if the disturbances are not necessarily normally distributed, but satisfy some moment conditions, and if certain dimensionality restrictions on $\mathcal{G}$ endowed with $\|\cdot\|_n$-norm are met, then $\hat{\mathbf{g}}_n$ converges with rate $\mathcal{O}_{\mathbf{P}}(n^{-1/2})$ in $\|\cdot\|_n$-norm. This result is established in Theorem 6.2.2 while Corollary 6.2.6 contains the result as a special case of a partly more general situation, where the parameter space may be infinite-dimensional, but stronger moment conditions are imposed.

The relation with Birgé's work becomes clear from Corollary 6.2.7. Here, it is shown that

$$\log N_2(\delta, \mathbf{H}_n, \mathcal{G}) \leqslant M\delta^{-\nu} \quad \text{for all} \quad \delta>0, \quad n\geqslant 1$$

and $0<\nu<2$, implies

$$\|\hat{\mathbf{g}}_n - g_0\|_n = \mathcal{O}_{\mathbb{P}}(n^{-\frac{1}{2+\nu}}).$$

Because the minimax theorem of Birgé in the situation of density estimation has its obvious counterpart in regression analysis, this means that the least squares estimator is minimax in the sense of rates of convergence in $\|\cdot\|_n$-norm.

Theorem 6.2.5 gives the most general result, albeit under fairly strong moment conditions on the disturbances. We allow for classes of regression functions $\mathcal{G}_n$ depending on n and the true underlying $g_{0,n}\in\mathcal{G}_n$ may vary with n too. In some situations, the rate of convergence can actually depend on $g_{0,n}$, which generally means that a rate faster than minimax is obtained. We denote by

$$\mathbf{B}_n(\rho, \mathcal{G}_n, g_{0,n}) = \{g\in\mathcal{G}_n: \|g-g_{0,n}\|_n\leqslant\rho\}, \quad \rho>0 \tag{6.1}$$

a ball with radius ρ around $g_{0,n}$, intersected with $\mathcal{G}_n$. The covering number of this neighbourhood of $g_{0,n}$ is

$$\mathbf{N}_n(\delta,\rho, \mathcal{G}_n, g_{0,n}) = N_2(\delta, \mathbf{H}_n, \mathbf{B}_n(\rho, \mathcal{G}_n, g_{0,n})), \quad \rho\geqslant\delta>0. \tag{6.2}$$

In the following section, we prove that the behaviour of $\mathbf{N}_n(\delta, \rho, \mathcal{G}_n, g_{0,n})$ as function of δ, ρ and n determines the speed of estimation. We call a model finite-dimensional if $\mathbf{N}_n(\delta, \rho, \mathcal{G}_n, g_{0,n})$ remains in some sense small (see (6.3)). In Subsection 6.2.1 we obtain rates under moment conditions depending on the dimension. Subsection 6.2.2 deals with infinite-dimensional models. Here, we impose an entropy-integrability condition on $\mathbf{N}_n(\delta, \rho, \mathcal{G}_n, g_{0,n})$, which is similar to condition (4.8) of Theorem 4.2.4 (see (6.21)).

Now, in general $\mathbf{N}_n(\delta, \rho, \mathcal{G}_n, g_{0,n})$ is random. However, to simplify the exposition, we assume throughout Section 6.2 that $H_{n,k}=\delta_{x_{n,k}}$, $k=1, \cdots, n$, $n\geqslant 1$. If the $\mathbf{x}_{n,k}$ are actually stochastic, this is equivalent to working conditionally on $(\mathbf{x}_{n,1}, \cdots, \mathbf{x}_{n,n})=(x_{n,1}, \cdots, x_{n,n})$. It is not difficult to adjust the results of the next section for the case of stochastic $\mathbf{x}_{n,k}$: one simply imposes the condition that for each n $\mathcal{G}_n$ is permissible (in order that Fubini's theorem can be applied) and assumes that the appropriate entropy-conditions hold in $\mathbb{P}^*$-probability. We elaborate on this in Section 6.3, Corollary 6.3.1, in the situation of i.i.d. $\mathbf{x}_{n,k}$. Theorem 6.3.2 presents sufficient conditions such that the rates of convergence in $\|\cdot\|_n$- and $\|\cdot\|$-norm are the same.

In Section 6.4 the results are applied to two-phase regression and compared with those of Chapter 5.

6.2. The rate of convergence of the least squares estimator
Let

$$\mathbf{y}_{n,k} = g(x_{n,k}) + \boldsymbol{\epsilon}_{n,k}, \quad k = 1, \cdots, n, \; g \in \mathcal{G}_n, \quad n = 1,2, \cdots,$$

where $x_{n,1}, \cdots, x_{n,n}$ are vectors in $\mathbb{R}^d$ and $\boldsymbol{\epsilon}_{n,1}, \cdots, \boldsymbol{\epsilon}_{n,n}$ are independent random variables with expectation zero and finite variance. The finite-dimensional case and the (possibly) infinite-dimensional case are treated separately, because in the latter we need more stringent moment conditions on the $\boldsymbol{\epsilon}_{n,k}$.

6.2.1. The finite-dimensional case. Call the sequence $\{\mathcal{G}_n, \|\cdot\|_n\}$ of *finite metric dimension r* at $\{g_{0,n}\}$ if there exist constants n_0, j_0, δ_0 such that

$$\sup_{n \geqslant n_0} \sup_{j \geqslant j_0} \sup_{0 < \delta \leqslant \delta_0} \frac{N_n(\delta, 2^j\delta, \mathcal{G}_n, g_{0,n})}{2^{jr}} \leqslant A < \infty. \tag{6.3}$$

For instance, suppose $\mathcal{G}_n$ can be indexed by an $\mathbb{R}^r$-valued parameter:

$$\mathcal{G}_n = \{g_\theta : \theta \in \Theta_n\}, \quad \Theta_n \subset \mathbb{R}^r.$$

Then $\{\mathcal{G}_n, \|\cdot\|_n\}$ is of finite metric dimension r at $\{g_{0,n}\}$ if for some $0 < K_{1,n} \leqslant K_{2,n} < \infty$ with

$$\limsup_{n \to \infty} \frac{K_{2,n}}{K_{1,n}} < \infty \tag{6.5}$$

the following holds:

$$\|g_\theta - g_{\theta_{0,n}}\|_n \geqslant K_{1,n} \|\theta - \theta_{0,n}\| \quad \text{for all} \quad \theta \in \Theta_n, \tag{6.6}$$

where $g_{\theta_{0,n}} = g_{0,n}$, and

$$\|g_\theta - g_{\tilde{\theta}}\|_n \leqslant K_{2,n} \|\theta - \tilde{\theta}\| \quad \text{for all} \quad \theta, \tilde{\theta} \in \Theta_n. \tag{6.7}$$

Observe that if $g_\theta(x)$ is differentiable with respect to θ for all x, this can be exploited to compute $K_{1,n}$ and $K_{2,n}$. We also remark that it is of course sufficient to consider neighbourhoods of $\theta_{0,n}$ once consistency is already established. We shall see examples of this in Section 6.4.

To establish a rate of convergence for $\hat{\mathbf{g}}_n$, we need a probability inequality for the random variables

$$\prec \boldsymbol{\epsilon}, g - \tilde{g} \succ_n = \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{\epsilon}_{n,k} (g(x_{n,k}) - \tilde{g}(x_{n,k})).$$

LEMMA 6.2.1. *If for some* $p \geqslant 1$

$$\sup_n \max_{1 \leqslant k \leqslant n} \mathbb{E} |\boldsymbol{\epsilon}_{n,k}|^{2p} = \gamma < \infty, \tag{6.8}$$

then for some C depending only on p and γ

$$\mathbb{P}(|\prec \boldsymbol{\epsilon}, g - \tilde{g} \succ_n| \geqslant a) \leqslant C \frac{\|g - \tilde{g}\|_n^{2p}}{n^p a^{2p}},$$

for all $a>0$, all $g, \tilde{g}$ and all $n \geqslant 1$.

PROOF. WHITTLE (1960) shows that for some C_p depending only on p

$$\mathbb{E}|\prec \epsilon, g-\tilde{g} \succ_n|^{2p} \leqslant \frac{C_p}{n^{2p}} \left[\sum_{k=1}^{n} (g(x_{n,k})-\tilde{g}(x_{n,k}))^2 (\mathbb{E}|\epsilon_{n,k}|^{2p})^{1/p} \right]^p.$$

Application of Chebyshev's inequality now gives the required result. □

THEOREM 6.2.2. *If $\{\mathcal{G}_n, \|\cdot\|_n\}$ is of finite metric dimension r at $\{g_{0,n}\}$ and (6.8) holds for some $p>r$, then there exist constants A', L' and n' such that for all $L \geqslant L'$ and $n \geqslant n'$*

$$\mathbb{P}(\|\hat{\mathbf{g}}_n - g_{0,n}\|_n > n^{-\frac{1}{2}} L) \leqslant A' L^{-(2p-r)}. \tag{6.9}$$

PROOF. Define $\delta_n = n^{-\frac{1}{2}}$. Remember that $g_{0,n} \in \mathcal{G}_n$ implies

$$2 \prec \epsilon, \hat{\mathbf{g}}_n - g_{0,n} \succ_n - \|\hat{\mathbf{g}}_n - g_{0,n}\|_n^2 \geqslant 0.$$

Therefore, replacing L by 2^L in (6.9), the theorem is proved if we show that for all L sufficiently large and n sufficiently large

$$\mathbb{P}\left(\sup_{\substack{g \in \mathcal{G}_n \\ \|g-g_{0,n}\|_n > 2^L \delta_n}} 2 \prec \epsilon, g-g_{0,n} \succ_n - \|g-g_{0,n}\|_n^2 \geqslant 0 \right) \leqslant A' 2^{-L(2p-r)}.$$

In particular, we shall take $L \geqslant j_0$, where j_0 is defined in (6.3).

Clearly,

$$\mathbb{P}\left(\sup_{\substack{g \in \mathcal{G}_n \\ \|g-g_{0,n}\|_n > 2^L \delta_n}} 2 \prec \epsilon, g-g_{0,n} \succ_n - \|g-g_{0,n}\|_n^2 \geqslant 0 \right) \tag{6.10}$$

$$\leqslant \sum_{j \geqslant L} \mathbb{P}\left(\sup_{\substack{g \in \mathcal{G}_n \\ 2^j \delta_n < \|g-g_{0,n}\|_n \leqslant 2^{j+1} \delta_n}} 2 \prec \epsilon, g-g_{0,n} \succ_n - \|g-g_{0,n}\|_n^2 \geqslant 0 \right)$$

$$\leqslant \sum_{j \geqslant L} \mathbb{P}\left(\sup_{g \in B_n(2^{j+1}\delta_n, \mathcal{G}_n, g_{0,n})} 2 \prec \epsilon, g-g_{0,n} \succ_n \geqslant 2^{2j} \delta_n^2 \right) = \sum_{j \geqslant L} \mathbb{P}_j, \quad \text{say.}$$

Let $\{g^{(0)}\}$ be a minimal δ_n-covering set of $B_n(2^{j+1}\delta_n, \mathcal{G}_n, g_{0,n})$, i.e. for each $g \in B_n(2^{j+1}\delta_n, \mathcal{G}_n, g_{0,n})$ there exists a $g^{(0)}(g) \in \{g^{(0)}\}$ such that

$$\|g - g^{(0)}(g)\|_n < \delta_n.$$

Since $\{\mathcal{G}_n, \|\cdot\|_n\}$ is of finite metric dimension r at $\{g_{0,n}\}$,

$$\text{card}\,(\{g^{(0)}\}) \leqslant A 2^{(j+1)r} \tag{6.11}$$

for all n and j sufficiently large. We get

$$\mathbb{P}_j = \mathbb{P}\left(\sup_{g\in B_n(2^{j+1}\delta_n,\,\mathcal{G}_n,\,g_{0,n})} 2\prec\epsilon, g-g_{0,n}\succ_n \geqslant 2^{2j}\delta_n^2\right)$$

$$\leqslant \mathbb{P}\left(\sup_{\{g^{(0)}\}}|\prec\epsilon, g^{(0)}-g_{0,n}\succ_n|\geqslant 2^{2(j-1)}\delta_n^2\right)$$

$$+\ \mathbb{P}\left(\sup_{g\in B_n(2^{j+1}\delta_n,\,\mathcal{G}_n,\,g_{0,n})}|\prec\epsilon, g-g^{(0)}(g)\succ_n|\geqslant 2^{2(j-1)}\delta_n^2\right) = \mathbb{P}_j^{(1)}+\mathbb{P}_j^{(2)}.$$

Since $\|g^{(0)}-g_{0,n}\|_n \leqslant 2^{j+2}\delta_n$, application of Lemma 6.2.1 yields

$$\mathbb{P}_j^{(1)} = \mathbb{P}\left(\sup_{\{g^{(0)}\}}|\prec\epsilon, g^{(0)}-g_{0,n}\succ_n|\geqslant 2^{2(j-1)}\delta_n^2\right)$$

$$\leqslant \operatorname{card}(\{g^{(0)}\})C\frac{(2^{j+2}\delta_n)^{2p}}{n^p(2^{2(j-1)}\delta_n^2)^{2p}} \leqslant A2^{(j+1)r}C\frac{(2^{j+2}\delta_n)^{2p}}{n^p(2^{2(j-1)}\delta_n^2)^{2p}}$$

for all n sufficiently large. This can tidied up to

$$\mathbb{P}_j^{(1)} \leqslant AC2^{r+8p}2^{-j(2p-r)}. \tag{6.12}$$

Next, we shall use the *chaining method* to show that the $\mathbb{P}_j^{(2)}$ are also small (see e.g. POLLARD (1984), Ch. VII). Let for $k\in\mathbb{N}$, $\{g^{(k)}\}$ be a minimal $2^{-k}\delta_n$-covering set of $B_n(2^{j+1}\delta_n, \mathcal{G}_n, g_{0,n})$. Then for $g\in B_n(2^{j+1}\delta_n, \mathcal{G}_n, g_{0,n})$, $\|g-g^{(k)}(g)\|_n < 2^{-k}\delta_n$, $k\in\mathbb{N}$

$$g-g^{(0)}(g) = \sum_{k=1}^{\infty} g^{(k)}(g)-g^{(k-1)}(g)$$

pointwise on $x_{n,1}, \cdots, x_{n,n}$. Define

$$s = 1-(r/p) \tag{6.13}$$

and $E=\Sigma_{k=1}^{\infty} k2^{-ks}$, $\eta_k = k2^{-ks}/E$. Then

$$\mathbb{P}\left(\sup_{g\in B_n(2^{j+1}\delta_n,\,\mathcal{G}_n,\,g_{0,n})}|\prec\epsilon, g-g^{(0)}(g)\succ_n|\geqslant 2^{2(j-1)}\delta_n^2\right)$$

$$\leqslant \sum_{k=1}^{\infty}\mathbb{P}\left(\sup_{g\in B_n(2^{j+1}\delta_n,\,\mathcal{G}_n,\,g_{0,n})}|\prec\epsilon, g^{(k)}(g)-g^{(k-1)}(g)\succ_n|\geqslant \eta_k 2^{2(j-1)}\delta_n^2\right).$$

The number of pairs $\{g^{(k)}(g), g^{(k-1)}(g)\}$ is at most

$$N_n(2^{-k}\delta_n, 2^{j+1}\delta_n, \mathcal{G}_n, g_{0,n})N_n(2^{-(k-1)}\delta_n, 2^{j+1}\delta_n, \mathcal{G}_n, g_{0,n}) \tag{6.14}$$
$$\leqslant N_n(2^{-k}\delta_n, 2^{j+1}\delta_n, \mathcal{G}_n, g_{0,n})^2 \leqslant (A2^{(j+k+1)r})^2$$

for all n sufficiently large.

Hence, application of Lemma 6.2.1 gives

$$\mathbb{P}_j^{(2)} \leqslant \sum_{k=1}^{\infty} (A2^{(j+k+1)r})^2 C \frac{(2^{-(k-2)}\delta_n)^{2p}}{n^p(\eta_k 2^{2(j-1)}\delta_n^2)^{2p}} \tag{6.15}$$

$$= A^2 C 2^{2r+8p} E^{2p} 2^{-2j(2p-r)} \sum_{k=1}^{\infty} k^{-2p}.$$

Returning to (6.10), we see that

$$\mathbb{P}\left[\sup_{\substack{g\in\mathcal{G}_n \\ \|g-g_{0,n}\|_n > 2^L\delta_n}} 2\langle \epsilon, g-g_{0,n}\rangle_n - \|g-g_{0,n}\|_n^2 \geqslant 0\right] \leqslant \sum_{j\geqslant L} (\mathbb{P}_j^{(1)} + \mathbb{P}_j^{(2)})$$

$$\leqslant \sum_{j\geqslant L} (AC2^{r+8p} + A^2 C 2^{2r+8p} E^{2p} \sum_{k=1}^{\infty} k^{-2p}) 2^{-j(2p-r)} \leqslant A' 2^{-(2p-r)L}$$

for L and n sufficiently large. Thus the proof is complete. □

In (6.3), where we defined finite-dimensionality, we assumed that the constant A does not depend on n. A weaker version of (6.3) would be

$$\sup_{n\geqslant n_0} \sup_{j\geqslant j_0} \sup_{0<\delta\leqslant\delta_0} \left[\frac{N_n(\delta, 2^j\delta, \mathcal{G}_n, g_{0,n})}{A_n 2^{jr}}\right] < \infty, \tag{6.16}$$

where $\{A_n\}$ is some possibly unbounded sequence. One can easily adjust the proof of Theorem 6.2.2 to show that under (6.16) the rate becomes $\mathcal{O}_{\mathbb{P}}(n^{-1/2}A_n^{1/p})$ (replace $\delta_n = n^{-1/2}$ by $\delta_n = n^{-1/2}A_n^{1/p}$).

Now, let us reconsider the case

$$\mathcal{G}_n = \{g_\theta : \theta\in\Theta_n\},\ \Theta_n \subset \mathbb{R}^r, \tag{6.17}$$

with $\{\mathcal{G}_n, \|\cdot\|_n\}$ satisfying (6.6) and (6.7), but not necessarily (6.5). Obviously, then (6.16) is met with $A_n = (K_{2,n}/K_{1,n})^r$, and the rate is thus $\mathcal{O}_{\mathbb{P}}(n^{-1/2}A_n^{1/p})$. However, careful inspection of the metric structure of Euclidean space reveals that this is not the most refined result: it turns out that it suffices to assume $p > \frac{1}{2}r$ in (6.8) and that the rate is $\mathcal{O}_{\mathbb{P}}(n^{-1/2}A_n^{1/(2p)})$. This is shown below.

LEMMA 6.2.3. *Suppose that $\mathcal{G}_n$ is of the form (6.17) and that for some $0 < K_{1,n} \leqslant K_{2,n} < \infty$*

$$\|g_\theta - g_{\theta_{0,n}}\|_n \geqslant K_{1,n}\|\theta - \theta_{0,n}\| \quad \textit{for all} \quad \theta\in\Theta_n,$$

$$\|g_\theta - g_{\tilde{\theta}}\|_n \leqslant K_{2,n}\|\theta - \tilde{\theta}\| \quad \textit{for all} \quad \theta, \tilde{\theta}\in\Theta_n.$$

If (6.8) is met for some $p > \frac{1}{2}r$, $p \geqslant 1$, and if $K_{2,n}/K_{1,n} = o(n^{p/r})$, then there exist constants A', L' and n' such that for all $L \geqslant L'$ and all $n \geqslant n'$

$$\mathbb{P}(\|\hat{\boldsymbol{\theta}}_n - \theta_{0,n}\| \geqslant n^{-1/2} \frac{K_{2,n}^{r/(2p)}}{K_{1,n}^{r/(2p)+1}} L) \leqslant A' L^{-(2p-r)}$$

($\hat{\boldsymbol{\theta}}_n$ *being defined by* $\hat{\mathbf{g}}_n = g_{\hat{\boldsymbol{\theta}}_n}$).

PROOF. Take

$$\delta_n = n^{-1/2} A_n^{1/2p}, \quad A_n = \left[\frac{K_{2,n}}{K_{1,n}}\right]^r, \tag{6.18}$$

and consider the set

$$b_{n,j+1} = \{\theta=(\theta_1, \cdots, \theta_r): \max_{1\leqslant s\leqslant r} |\theta_s - \theta_{0,n,s}| \leqslant 2^{j+1}\frac{\delta_n}{K_{1,n}}\}.$$

Since for $\theta \in B_n(2^{j+1}\delta_n, \mathcal{G}_n, g_{0,n})$

$$\|\theta-\theta_{0,n}\| \leqslant \frac{1}{K_{1,n}}\|g_\theta - g_{\theta_{0,n}}\|_n \leqslant 2^{j+1}\frac{\delta_n}{K_{1,n}},$$

we have

$$B_n(2^{j+1}\delta_n, \mathcal{G}_n, g_{0,n}) \subset \{g_\theta: \theta \in b_{n,j+1}\}.$$

The r-dimensional cube $b_{n,j+1}$ can be covered by

$$\left(\left[\frac{2^{j+k+1}2\sqrt{r}K_{2,n}}{K_{1,n}}\right]+1\right)^r$$

small cubes with side of length $2^{-k}(\delta_n/(\sqrt{r}K_{2,n}))$. We have

$$\left(\left[\frac{2^{j+k+1}2\sqrt{r}K_{2,n}}{K_{1,n}}\right]+1\right)^r \leqslant C_r A_n 2^{(j+k+1)r}$$

for some constant C_r depending only on r. Write $N_n^c(2^{-k}\delta_n, 2^{j+1}\delta_n, \mathcal{G}_n, g_{0,n}) = C_r A_n 2^{(j+k+1)r}$. Let $\{c^{(k)}\}$ be the collection of corners with the smallest co-ordinates of the cubes covering $b_{n,j+1}$. Then $\text{card}(\{c^{(k)}\}) \leqslant N_n^c(2^{-k}\delta_n, 2^{j+1}\delta_n, \mathcal{G}_n, g_{0,n})$. For $\theta=(\theta_1, \cdots, \theta_r) \in b_{n,j+1}$, write

$$g_c^{(k)}(g_\theta) = g_{c^{(k)}} \quad \text{if} \quad \max_{1\leqslant s\leqslant r} |\theta_s - c_s^{(k)}| < 2^{-k}(\delta_n/(\sqrt{r}K_{2,n})).$$

Then

$$\|g_\theta - g_c^{(k)}(g_\theta)\|_n \leqslant K_{2,n}\|\theta - c^{(k)}\| < 2^{-k}\delta_n.$$

So $\{g_{c^{(k)}}\}$ forms a $2^{-k}\delta_n$-covering set of $B_n(2^{j+1}\delta_n, \mathcal{G}_n, g_{0,n})$ with

$$\text{card}(\{g_{c^{(k)}}\}) \leqslant N_n^c(2^{-k}\delta_n, 2^{j+1}\delta_n, \mathcal{G}_n, g_{0,n}).$$

The covering sets $\{g_{c^{(k)}}\}$ have as special feature that the number of pairs $\{g_{c^{(k)}}(g_\theta), g_{c^{(k-1)}}(g_\theta)\}$, with $c^{(k)} \neq c^{(k-1)}$, is at most

$$(2^r - 1)N_n^c(2^{-k}\delta_n, 2^{j+1}\delta_n, \mathcal{G}_n, g_{0,n}).$$

Now, in the proof Theorem 6.2.2 one can make the following adjustments. Take δ_n as in (6.18), replace $\{g^{(k)}\}$ by $\{g_{c^{(k)}}\}$, $k=0,1,2, \cdots$ and

$N_n(2^{-k}\delta_n, 2^{j+1}\delta_n, \mathcal{G}_n, g_{0,n})$ by $N_n^c(2^{-k}\delta_n, 2^{j+1}\delta_n, \mathcal{G}_n, g_{0,n})$, $k=1,2,\cdots$, $j\in\mathbb{N}$. Define in (6.13), $s=1-(r/2p)$ and replace the bound in (6.14) by $(2^r-1)N_n^c(2^{-k}\delta_n, 2^{j+1}\delta_n, \mathcal{G}_n, g_{0,n})$. The rate $\mathcal{O}_{\mathbb{P}}(\delta_n)$ for $\hat{\mathbf{g}}_n$ now follows easily and this rate implies the $\mathcal{O}_{\mathbb{P}}(K_{1,n}^{-1}\delta_n)$-rate for $\hat{\boldsymbol{\theta}}_n$. □

EXAMPLE 6.1. In Example 3.2 of Chapter 3, we studied the linear model

$$g_\theta(x) = x\theta, \quad \theta\in\Theta_n,$$

with $\Theta_n=\Theta$, $\theta_{0,n}=\theta_0$. The smallest and largest eigenvalue of $X_n^TX_n$, $X_n=(x_{n,1}^T,\ldots,x_{n,n}^T)^T$, are denoted by $\lambda_{1,n}$ and $\lambda_{2,n}$ respectively. We showed in Lemma 3.3.5 that under regularity conditions on the second moments of the $\boldsymbol{\epsilon}_{n,k}$

$$\|\hat{\boldsymbol{\theta}}_n-\theta_0\| \xrightarrow{\mathbb{P}} 0$$

provided that for some $c>0$

$$\frac{\lambda_{2,n}^{\frac{1}{2}(1+c)}}{\lambda_{1,n}} = \mathcal{O}(1)$$

and provided Θ is compact.

If (6.8) holds for some $p>1$, then the regularity conditions on the second moments of the $\boldsymbol{\epsilon}_{n,k}$ are met. Now, obviously (6.6) and (6.7) are fulfilled, with $K_{i,n}=(\frac{1}{n}\lambda_{i,n})^{\frac{1}{2}}$, $i=1,2$. So, if $p>\frac{1}{2}r(=\frac{1}{2}(d+1))$, then it follows from Lemma 6.2.3 that

$$\|\hat{\boldsymbol{\theta}}_n-\theta_0\| \xrightarrow{\mathbb{P}} 0$$

provided

$$\frac{\lambda_{2,n}^{\frac{1}{2}(1-\frac{2p-r}{2p+r})}}{\lambda_{1,n}} = o(1).$$

Compactness of Θ is not needed.

6.2.2. The infinite-dimensional case. The condition on the $|\boldsymbol{\epsilon}_{n,k}|^2$ we need in finite-dimensional models is the existence of an absolute moment of order larger than the dimension of parameter space. In possibly infinite-dimensional models, we assume existence of the moment generating function of $|\boldsymbol{\epsilon}_{n,k}|^2$. Of course, this assumption also establishes an improvement of the bound in (6.9) (see Corollary 6.2.6). We start off by formulating a pendant of the Chebyshev-type inequality which we presented in Lemma 6.2.1.

LEMMA 6.2.4. *If for some* $\beta>0$

$$\sup_n \max_{1\leqslant k\leqslant n} \mathbb{E}(\exp(\beta|\boldsymbol{\epsilon}_{n,k}|^2)) \leqslant \Gamma<\infty, \tag{6.19}$$

then there exists an $\alpha>0$ *depending only on* β *and* Γ *such that*

$$\mathbb{P}(|\prec \epsilon, g-\tilde{g}\succ_n| \geqslant a) \leqslant \exp\left[-\frac{\alpha n a^2}{\|g-\tilde{g}\|_n^2}\right],$$

for all $a>0$, *all* $g,\tilde{g}$ *and all* $n \geqslant 1$.

PROOF. For all $h>0$

$$\mathbb{P}(|\prec \epsilon, g-\tilde{g}\succ_n| \geqslant a) \leqslant \exp(-hna)\mathbb{E}[\exp(hn \prec \epsilon, g-\tilde{g}\succ_n)]$$

$$\leqslant \exp(-hna)\prod_{k=1}^{n}\mathbb{E}[\exp(h|\epsilon_{n,k}||g(x_{n,k})-\tilde{g}(x_{n,k})|)].$$

KUELBS (1978) shows that under (6.19) for some Λ depending only on β and Γ

$$\mathbb{E}[\exp(h|\epsilon_{n,k}||g(x_{n,k})-\tilde{g}(x_{n,k})|)] \leqslant \exp[h^2(g(x_{n,k})-\tilde{g}(x_{n,k}))^2\Lambda^2].$$

Thus

$$\mathbb{P}(|\prec \epsilon, g-\tilde{g}\succ_n| \geqslant a) \leqslant \exp(-hna)\exp(h^2 n\|g-\tilde{g}\|_n^2\Lambda^2).$$

Take $h=(2\alpha a)/\|g-\tilde{g}\|_n^2$, with $\alpha=(4\Lambda^2)^{-1}$. Then

$$\mathbb{P}(|\prec \epsilon, g-\tilde{g}\succ_n| \geqslant a) \leqslant \exp\left[-\frac{na^2}{2\Lambda^2\|g-\tilde{g}\|_n^2}\right]\exp\left[\frac{a^2\Lambda^2 n\|g-\tilde{g}\|_n^2}{4\Lambda^4\|g-\tilde{g}\|_n^4}\right]$$

$$=\exp\left[-\frac{\alpha n a^2}{\|g-\tilde{g}\|_n^2}\right]. \quad \square$$

In Theorem 6.2.5 below, the entropy conditions (6.20) and (6.21) are perhaps at first sight rather unappealing. However, after proving the theorem we shall give several clarifying examples.

THEOREM 6.2.5. *Let* $\delta_n \to 0$ *be some sequence with* $\liminf_{n\to\infty} n^{1/2}\delta_n > 0$ *and suppose that*

$$\lim_{j\to\infty}\sup_{n\geqslant n_0}\frac{\sqrt{\log N_n(\delta_n, 2^j\delta_n, \mathcal{G}_n, g_{0,n})}}{n^{1/2}\delta_n 2^j} = 0 \tag{6.20}$$

$$\lim_{j\to\infty}\sup_{n\geqslant n_0}\int_0^1 \frac{\sqrt{\log N_n(u\delta_n, 2^j\delta_n, \mathcal{G}_n, g_{0,n})}}{n^{1/2}\delta_n 2^j}\,du \leqslant M<\infty. \tag{6.21}$$

If moreover (6.19) holds for some $\beta>0$, *then* $\hat{\mathbf{g}}_n$ *converges with rate* $\mathcal{O}_{\mathbb{P}}(\delta_n)$. *In fact, there exist constants* M', L' *and* n' *such that for all* $L \geqslant L'$ *and all* $n \geqslant n'$

$$\mathbb{P}(\|\hat{\mathbf{g}}_n - g_{0,n}\|_n > \delta_n L) \leqslant \exp(-M'L^2 n\delta_n^2).$$

PROOF. As in the proof of Theorem 6.2.2 we replace L by 2^L and write

$$\mathbb{P}\left\{\sup_{\substack{g\in\mathcal{G}_n\\ \|g-g_{0,n}\|_n>2^L\delta_n}} 2\prec\epsilon, g-g_{0,n}\succ_n-\|g-g_{0,n}\|_n^2\geqslant 0\right\}$$

$$\leqslant \sum_{j\geqslant L}\mathbb{P}\left\{\sup_{\substack{g\in\mathcal{G}_n\\ g\in B_n(2^{j+1}\delta_n,\mathcal{G}_n,g_{0,n})}} 2\prec\epsilon, g-g_{0,n}\succ_n\geqslant 2^{2j}\delta_n^2\right\}\leqslant \sum_{j\geqslant L}\mathbb{P}_j.$$

Let, for each $k\in\{0,1,2,\cdots\}$, $\{g^{(k)}\}$ be a minimal $2^{-k}\delta_n$-covering set of $B_n(2^{j+1}\delta_n,\mathcal{G}_n,g_{0,n})$ and let $g^{(k)}(g)$ be defined by

$$\|g-g^{(k)}(g)\|_n=\min_{\{g^{(k)}\}}\|g-g^{(k)}\|_n.$$

As before

$$\mathbb{P}_j=\mathbb{P}\left\{\sup_{g\in B_n(2^{j+1}\delta_n,\mathcal{G}_n,g_{0,n})} 2\prec\epsilon, g-g_{0,n}\succ_n\geqslant 2^{2j}\delta_n^2\right\}$$
$$\leqslant \mathbb{P}\left\{\sup_{\{g^{(0)}\}}|\prec\epsilon,g^{(0)}-g_{0,n}\succ_n|\geqslant 2^{2(j-1)}\delta_n^2\right\}$$
$$+\mathbb{P}\left\{\sup_{g\in B_n(2^{j+1}\delta_n,\mathcal{G}_n,g_{0,n})}|\prec\epsilon,g-g^{(0)}(g)\succ_n|\geqslant 2^{2(j-1)}\delta_n^2\right\}=\mathbb{P}_j^{(1)}+\mathbb{P}_j^{(2)}.$$

Application of Lemma 6.2.4 gives

$$\mathbb{P}_j^{(1)}=\mathbb{P}\left\{\sup_{\{g^{(0)}\}}|\prec\epsilon,g^{(0)}-g_{0,n}\succ_n|\geqslant 2^{2(j-1)}\delta_n^2\right\}$$
$$\leqslant \exp(\log N_n(\delta_n,2^{j+1}\delta_n,\mathcal{G}_n,g_{0,n})-\alpha 2^{-6}2^{2j}n\delta_n^2).$$

By (6.20), $\log N_n(\delta_n,2^{j+1}\delta_n,\mathcal{G}_n,g_{0,n})\leqslant\frac{1}{2}(\alpha 2^{-6}2^{2j}n\delta_n^2)$ for all j and n sufficiently large, so

$$\mathbb{P}_j^{(1)}\leqslant\exp(-\alpha 2^{-7}2^{2j}n\delta_n^2).$$

We use the chaining again to bound $\mathbb{P}_j^{(2)}$:

$$\mathbb{P}_j^{(2)}=\mathbb{P}\left\{\sup_{g\in B_n(2^{j+1}\delta_n,\mathcal{G}_n,g_{0,n})}|\prec\epsilon,g-g^{(0)}(g)\succ_n|\geqslant 2^{2(j-1)}\delta_n^2\right\}$$
$$\leqslant\sum_{k=1}^{\infty}\mathbb{P}\left\{\sup_{g\in B_n(2^{j+1}\delta_n,\mathcal{G}_n,g_{0,n})}|\prec\epsilon,g^{(k)}(g)-g^{(k-1)}(g)\succ_n|\geqslant 2^{2(j-1)}\delta_n^2\eta_{j,k}\right\},$$

where $\{\eta_{j,k}\}_{k=1}^{\infty}$ is a sequence satisfying $\Sigma_{k=1}^{\infty}\eta_{j,k}\leqslant 1$. Define $E=\Sigma_{k=1}^{\infty}2^{-k}k^{1/2}$ and take

$$\eta_{j,k}=\max\left[\left(\frac{\sqrt{\log N_n(2^{-k}\delta_n,2^{j+1}\delta_n,\mathcal{G}_n,g_{0,n})}}{2Mn^{1/2}\delta_n 2^{j+k+1}(\log 2)^{-1}}\right),\left(\frac{2^{-k}k^{1/2}}{2E}\right)\right].$$

Then in view of (6.21)

$$\sum_{k=1}^{\infty}\eta_{j,k}\leqslant\sum_{k=1}^{\infty}\left[\left(\frac{\sqrt{\log N_n(2^{-k}\delta_n,2^{j+1}\delta_n,\mathcal{G}_n,g_{0,n})}}{2Mn^{1/2}\delta_n 2^{j+k+1}(\log 2)^{-1}}\right)+\left(\frac{2^{-k}k^{1/2}}{2E}\right)\right]=1$$

for all $j \geqslant L_0$ and all n sufficiently large. Use Lemma 6.2.4 again to establish

$$\begin{aligned}
\mathbb{P}_j^{(2)} &\leqslant \sum_{k=1}^{\infty} \exp\left[2 \log N_n(2^{-k}\delta_n, 2^{j+1}\delta_n, \mathcal{G}_n, g_{0,n}) - \alpha 2^{-6} 2^{4j} 2^{2k} \eta_{j,k}^2 n\delta_n^2\right] \\
&\leqslant \sum_{k=1}^{\infty} \exp\left[2(2Mn^{1/2}\delta_n 2^{j+k+1})^2 \eta_{j,k}^2 - \alpha 2^{-6} 2^{4j} 2^{2k} n\delta_n^2 \eta_{j,k}^2\right] \\
&\leqslant \sum_{k=1}^{\infty} \exp\left[-\alpha 2^{-7} 2^{4j} 2^{2k} n\delta_n^2 \eta_{j,k}^2\right] \\
&\leqslant \sum_{k=1}^{\infty} \exp\left[-\alpha 2^{-7} 2^{4j} 2^{2k} n\delta_n^2 \left(\frac{2^{-k} k^{1/2}}{2E}\right)^2\right] \\
&= \sum_{k=1}^{\infty} \exp\left[-\alpha 2^{-7} 2^{4j} n\delta_n^2 \frac{k}{(2E)^2}\right].
\end{aligned}$$

Hence for L sufficiently large, n sufficiently large

$$\begin{aligned}
&\sum_{j \geqslant L} (\mathbb{P}_j^{(1)} + \mathbb{P}_j^{(2)}) \\
&\leqslant \sum_{j \geqslant L} \left[\exp(-\alpha 2^{-7} 2^{2j} n\delta_n^2) + \sum_{k=1}^{\infty} \exp(-\alpha 2^{-7} 2^{4j} n\delta_n^2 \frac{k}{(2E)^2})\right] \\
&\leqslant \exp(-M' 2^{2L} n\delta_n^2). \quad \square
\end{aligned}$$

The *entropy-integrability condition* (6.21) makes the chaining method work. POLLARD (1982) uses this method to establish the uniform central limit theorem that was reproduced here as Theorem 4.2.2. We have adopted his technique in the proofs of Theorems 6.2.2 and 6.2.5. We also mention Pollard's chaining lemma (POLLARD (1984) Ch. VII), which presents the relation between entropy-integrability and asymptotic equicontinuity in a more general context.

A first corollary of Theorem 6.2.5 concerns the finite-dimensional case.

COROLLARY 6.2.6. *Suppose that (6.16) holds, i.e.*

$$\sup_{n \geqslant n_0} \sup_{j \geqslant j_0} \sup_{0 < \delta \leqslant \delta_0} \frac{N_n(\delta, 2^j\delta, \mathcal{G}_n, g_{0,n})}{A_n 2^{jr}} < \infty$$

for some r and some sequence $\{A_n\}$, $\liminf A_n > 0$. Without loss of generality we assume $A_n \geqslant 2$ for all n, so that $\log A_n > 0$. Then for $\delta_n = n^{-1/2}(\log A_n)^{1/2}$, (6.20) and (6.21) are fulfilled:

$$\frac{\sqrt{\log N_n(\delta_n, 2^j\delta_n, \mathcal{G}_n, g_{0,n})}}{n^{1/2}\delta_n 2^j} \leqslant C_r \frac{\sqrt{\log 2^j}}{2^j}$$

and

$$\int_0^1 \frac{\sqrt{\log N_n(u\delta_n, 2^j\delta_n, \mathcal{G}_n, g_{0,n})}}{n^{1/2}\delta_n 2^j}\, du \leqslant C'_r \int_0^1 (\log \frac{1}{u})^{1/2} du,$$

for all $n \geqslant n_0$, $j \geqslant j_0$, $\delta_n \leqslant \delta_0$, *where* C_r *and* C'_r *only depend on* r. *Thus provided (6.19) holds for some* $\beta > 0$,

$$\|\hat{\mathbf{g}}_n - g_{0,n}\|_n = \mathcal{O}_{\mathbb{P}}(n^{-1/2}(\log A_n)^{1/2}).$$

In particular, if $\{\mathcal{G}_n, \|\cdot\|_n\}$ *is of finite metric dimension i.e.* $\limsup A_n < \infty$

$$\frac{1}{n}\log \mathbb{P}(\|\hat{\mathbf{g}}_n - g_{0,n}\|_n > a) \leqslant -M'a^2$$

for all $n \geqslant n'$ *and* $a > L'n'^{-1/2}$. *This is called a law of large deviations for* $\hat{\mathbf{g}}_n$.

EXAMPLE 6.1 CONTINUED. In the linear model, application of Corollary 6.2.6 yields that if (6.19) holds for some $\beta > 0$, then

$$\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n\| = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\log\left(\sqrt{2} \vee \frac{\lambda_{2,n}}{\lambda_{1,n}}\right)/\lambda_{1,n}}\right).$$

The remainder of this section deals with application of Theorem 6.2.5 to (truly) infinite-dimensional models. The first example we give, however, shares a common feature with finite-dimensional models. Consider the *global* entropy $\log N_2(\delta_n, H_n, \mathcal{G}_n)$ of the space $\mathcal{G}_n$. Provided $\|\cdot\|_n$ remains bounded on $\mathcal{G}_n$, we have that if $\{\mathcal{G}_n, \|\cdot\|_n\}$ is of finite metric dimension r, then

$$\sup_{n \geqslant n_0} \sup_{\delta \leqslant \delta_0} \delta^r N_2(\delta, H_n, \mathcal{G}_n) \leqslant A.$$

This is also true for $\mathcal{G}_n$ in Example 6.2.

EXAMPLE 6.2. Let $\mathcal{G}_n = \mathcal{G}$ be a *VC*-graph class with envelope G, and let $\{\|\cdot\|_n\}$ be such that

$$\limsup_{n \to \infty} \|G\|_n < \infty. \tag{6.22}$$

Then by Theorem 2.2.6

$$\sup_{n \geqslant n_0} \sup_{\delta > 0} \delta^r N_2(\delta, H_n, \mathcal{G}) \leqslant A$$

for some constants r and A, where A only depends on $\{\|\cdot\|_n\}$ via the left-hand side of (6.22). It is straightforward to see that (6.20) and (6.21) hold with $\delta_n = n^{-1/2}(\log n)^{1/2}$, using the bound $N_n(\delta, 2^j\delta, \mathcal{G}_n, g_{0,n}) \leqslant N_2(\delta, H_n, \mathcal{G})$. Hence under (6.19), $\|\hat{\mathbf{g}}_n - g_{0,n}\|_n = \mathcal{O}_{\mathbb{P}}(n^{-1/2}(\log n)^{1/2})$ for all sequences $\{g_{0,n}\} \subset \mathcal{G}$.

Corollary 6.2.7 below clarifies the relation with Birgé's results (BIRGÉ (1983)).

COROLLARY 6.2.7. *Suppose that for some constants* $\nu > 0$ *and* M

$$\sup_{n \geqslant n_0} \sup_{0 < \delta \leqslant \delta_0} \delta^\nu \log N_2(\delta, H_n, \mathcal{G}_n) \leqslant M.$$

Take $\delta_n = n^{-\frac{1}{2+\nu}}$. *Then (6.20) holds, and if* $\nu < 2$, *(6.21) holds too. It follows*

that under (6.19)

$$\|\hat{\mathbf{g}}_n - g_{0,n}\|_n = \mathcal{O}_{\mathbf{P}}(n^{-\frac{1}{2+\nu}})$$

for $\nu<2$ *and for all* $\{g_{0,n}\}$, $g_{0,n}\in\mathcal{G}_n$, $n\geqslant n_0$.

Here is an application of the previous corollary.

EXAMPLE 6.3. Let

$$\mathcal{G} = \{g: K\to\mathbb{R},\ g \text{ has } m \text{ derivatives}, \tag{6.23}$$

$$\sup_{x,\tilde{x}\in K} \frac{|g^{(m)}(x)-g^{(m)}(\tilde{x})|}{\|x-\tilde{x}\|^{\alpha}} \leqslant L,\ |g|\leqslant C\},$$

with $\alpha>0$, K is a compact subset of $\mathbb{R}^d$ and where $\|x-\tilde{x}\|$ is the Euclidean distance between x and $\tilde{x}$. KOLMOGOROV and TIHOMIROV (1959) show that

$$\sup_{\delta>0} \delta^{\frac{d}{m+\alpha}} \log N_{\infty}(\delta, H_n, \mathcal{G}) \leqslant M. \tag{6.24}$$

Thus if $d/(m+\alpha)<2$ and (6.19) holds

$$\|\hat{\mathbf{g}}_n - g_{0,n}\|_n = \mathcal{O}_{\mathbf{P}}(n^{-\frac{m+\alpha}{2(m+\alpha)+d}})$$

for all $\{g_{0,n}\}\subset\mathcal{G}$ Similarly, let

$$\mathcal{G} = \{g: K\to\mathbb{R},\ g \text{ has } m \text{ derivatives}, \tag{6.25}$$

$$\int |g^{(m)}(x)|^2 dx \leqslant L,\ |g|\leqslant C\}$$

where K is a compact subset of $\mathbb{R}$. Given the result (6.24) for $\mathcal{G}$ defined in (6.23), it is easy to see that the $\mathcal{G}$ of (6.25) satisfies

$$\sup_{\delta>0} \delta^m \log N_2(\delta, H_n, \mathcal{G}) \leqslant M$$

so under (6.19) , $\|\hat{\mathbf{g}}_n - g_{0,n}\|_n = \mathcal{O}_{\mathbf{P}}(n^{-m/(2m+1)})$ for all $\{g_{0,n}\}\subset\mathcal{G}$. STONE (1982) proves that these rates are optimal.

EXAMPLE 6.4. Let

$$\mathcal{G} = \{g: \mathbb{R}\to\mathbb{R},\ g \text{ increasing}, |g|\leqslant C\}. \tag{6.26}$$

BIRGÉ (1980) shows that the L^1-entropy of $\mathcal{G}$ is of order δ^{-1}. It is not clear whether the L^2-entropy is also of this order. Lemma 6.2.8 below presents a bound for the L^2-covering number that by application of Theorem 6.2.5 leads to the rate $\|\hat{\mathbf{g}}_n - g_{0,n}\|_n = \mathcal{O}_{\mathbf{P}}(n^{-1/3}(\log n)^{1/2})$ for all $\{g_{0,n}\}\subset\mathcal{G}$.

LEMMA 6.2.8. *For* $\mathcal{G}$ *defined in (6.26)*

$$\log N_2(\delta, Q, \mathcal{G}) \leqslant M\delta^{-1}\log(\delta^{-1}) \quad \textit{for all} \quad \delta>0, \tag{6.27}$$

where Q is any probability measure on $\mathbb{R}$ and where M only depends on C.

PROOF. Without loss of generality we assume that $0 \leqslant g \leqslant 1$ for all $g \in \mathcal{G}$. Define $T = [1/\delta^2] + 1$ and let $-\infty = a_0 < a_1 < \cdots < a_{T-1} < a_T = \infty$ be such that $Q(a_{i-1}, a_i] \leqslant \delta^2$ for $i = 1, \cdots, T$. Define for each $g \in \mathcal{G}$

$$\bar{g}_i(g) = \int_{(a_{i-1}, a_i]} g dQ / Q(a_{i-1}, a_i]$$

and

$$k_i(g) = \left[\frac{\bar{g}_i(g)}{\delta} \right], \quad i = 1, \cdots, T. \tag{6.28}$$

Then

$$\int_{(a_{i-1}, a_i]} |g - \delta k_i(g)|^2 dQ \leqslant Q(a_{i-1}, a_i] \{ \mathrm{var}_Q(g(\mathbf{x}) | \mathbf{x} \in (a_{i-1}, a_i]) + \delta^2 \}$$
$$\leqslant Q(a_{i-1}, a_i] \{ g(a_i)^2 - g(a_{i-1})^2 \} + Q(a_{i-1}, a_i] \delta^2, \quad i = 1, \cdots, T.$$

Hence

$$\int |g - \delta \sum_{k=1}^{T} k_i(g) 1_{(a_{i-1}, a_i]}|^2 dQ \leqslant \delta^2 (g(a_n)^2 - g(a_0)^2) + \delta^2 \leqslant 2\delta^2. \tag{6.29}$$

We have that $0 \leqslant k_1(g) \leqslant \cdots \leqslant k_T(g) \leqslant [1/\delta]$ and $k_i(g) \in \mathbb{Z}$, $i = 1, \cdots, T$. The number of functions of the form

$$\sum_{i=1}^{T} k_i 1_{(a_{i-1}, a_i]}, \ 0 \leqslant k_1 \leqslant \cdots \leqslant k_T \leqslant [1/\delta], \ k_i \in \mathbb{Z}, \ i = 1, \cdots, T, \tag{6.30}$$

is equal to

$$\begin{pmatrix} (T+1) + [1/\delta] - 1 \\ [1/\delta] \end{pmatrix} = \begin{pmatrix} [1/\delta^2] + [1/\delta] + 1 \\ [1/\delta] \end{pmatrix}. \tag{6.31}$$

Thus

$$\log N_2(\sqrt{2}\delta, Q, \mathcal{G}) \leqslant \log \begin{pmatrix} [1/\delta^2] + [1/\delta] + 1 \\ [1/\delta] \end{pmatrix} \leqslant M \frac{1}{\delta} \log(\frac{1}{\delta}). \quad \square$$

We end this section with some remarks. First, Theorem 6.2.5 presents a fairly general result, but since the calculation of entropies is often quite difficult, the merit of the theorem is primarily that it shows that the statistical problem can be replaced by a combinatorial one.

It should secondly be noted that if the rate δ_n is slower than $n^{-1/2}$, then the probability inequality of Theorem 6.2.5 implies that for some constant L_0

$$\mathbb{P}(\|\hat{\mathbf{g}}_n - g_{0,n}\|_n \leqslant L_0 \delta_n) \to 1. \tag{6.32}$$

Moreover, if the rate is slow enough - e.g. $\delta_n = n^{-\frac{1}{2+\nu}}$, $\nu > 0$- then by Borel-Cantelli's theorem $\|\hat{\mathbf{g}}_n - g_{0,n}\|_n \leqslant L_0 \delta_n$ almost surely, provided of course that

the sequence of disturbances all live on the same probability space.

Finally, due to the entropy integrability condition (6.21) Theorem 6.2.5 cannot handle optimal rates slower than $o_{\mathbb{P}}(n^{-1/4})$. Such slow rates are the consequence of large entropies, meaning that $\mathcal{G}_n$ has so little metric structure that the process $\sqrt{n}\prec\epsilon, g-g_0\succ_n$ might not be asymptotically equicontinuous (see also Chapter 4).

6.3. Stochastic design

Let $\mathbf{x}_1, \mathbf{x}_2, \cdots$ be independent random vectors with distribution H, and let $\mathbf{N}_n(\delta, \rho, \mathcal{G}_n, g_{0,n})$ be defined as in (6.2). The randomness of this covering number prohibits direct application of Theorem 6.2.5, but of course by conditioning one can easily adjust this theorem to the case of stochastic design. Before doing this, we make some simplifying assumptions to facilitate the exposition. We assume that also $\epsilon_1, \epsilon_2, \cdots$ are i.i.d. (of course with expectation zero, finite variance and independent of the $\mathbf{x}_k$) and that $\mathcal{G}_n = \mathcal{G}$ and $g_{0,n} = g_0 (\in \mathcal{G})$. This brings us back to the situation of Section 3.1. Finally, we restrict ourselves to $\mathcal{O}_{\mathbb{P}}(n^{-1/(2+\nu)})$-rates, $0 \leqslant \nu < 2$. Then the stochastic counterpart of Theorem 6.2.5 becomes:

COROLLARY 6.3.1. *Suppose $\mathcal{G}$ is a permissible class, satisfying*

$$\limsup_{n\to\infty} \mathbb{P}^* \left(\sup_{j \geqslant j_0} \sup_{0<\delta\leqslant\delta_0} \frac{\delta^\nu \log \mathbf{N}_n(\delta, 2^j\delta, \mathcal{G}_n, g_{0,n})}{\log 2^j} > M \right) = 0, \quad (6.30)$$

for some $L>0$, $M>0$ and $0\leqslant\nu<2$. If

$$\mathbb{E}\exp(\beta|\epsilon_1|^2) < \infty \quad \text{for some} \quad \beta>0, \quad (6.31)$$

then

$$\|\hat{\mathbf{g}}_n - g_{0,n}\|_n = \mathcal{O}_{\mathbb{P}}(n^{-\frac{1}{2+\nu}}). \quad (6.32)$$

It appears to be difficult to check (6.30). However, we have seen examples (e.g. Examples 6.3 and 6.4) where covering numbers can be computed even when one has virtually no knowledge about the metric used (i.e. $\|\cdot\|_n$). Nevertheless, in general one faces the problem of drawing conclusions about the random $L^2(\mathbb{R}^d, \mathbf{H}_n)$-covering numbers from the theoretical $L^2(\mathbb{R}^d, H)$-covering numbers. In other words, one is asking for the order of magnitude of the ratio $\|\cdot\|_n / \|\cdot\|$. We address this problem in Lemma 6.3.4.

The main aim of this section is to present sufficient conditions such that a rate of convergence in $\|\cdot\|_n$-norm implies the same rate in $\|\cdot\|$-norm (see Theorem 6.3.2). A natural question is whether it is possible to prove rates in $\|\cdot\|$-norm directly. Recall that the conditions we needed in Section 3.1 for consistency in $\|\cdot\|_n$-norm are stronger than those for consistency in $\|\cdot\|$-norm: in the latter case an envelope condition could be replaced by a uniform square integrability condition. Indeed, an envelope condition is implicit in (6.30). This is illustrated by Lemma 3.3.4 and also by for instance Examples 6.3 and

6.4. It is not clear whether anything can be gained on the assumptions if one is only interested in rates in $\|\cdot\|$-norm.

A situation where $\|\cdot\|$- and $\|\cdot\|_n$- norms can in a certain sense be interchanged frecly, arises when there exist covering sets with *bracketing*. A δ-bracketing with respect to $\|\cdot\|$ of a function $g \in L^2(\mathbb{R}^d, H)$ is a pair $[g_1, g_2]$ such that $g_1 \leqslant g \leqslant g_2$ and $\|g_1 - g_2\| < \delta$. The minimum number of brackets necessary to cover $\mathcal{G}$ is denoted by $N_2^{[]}(\delta, H, \mathcal{G})$. Lemma 6.3.4 will show that under appropriate conditions on $N_2^{[]}(\delta, H, \mathcal{G})$ the metrics $\|\cdot\|_n$ and $\|\cdot\|$ are asymptotically equivalent.

We already encountered covering sets with bracketing in Application 3.2.1. Here,

$$\mathcal{G} = \{g_\theta : \theta \in \Theta\}$$

with $g_\theta(x)$ continuous in θ for all x, Θ compact and

$$\sup_{\theta \in \Theta} |g_\theta| \in L^2(\mathbb{R}^d, H).$$

We asserted in Application 3.2.1 that $N_2(\delta, \mathbf{H}_n, \mathcal{G})$ remains bounded almost surely for all $\delta > 0$. To prove this, we showed that $N_2^{[]}(\delta, H, \mathcal{G})$ is finite.

Another illustration is given by Example 6.4. It is not difficult to see that in this example $N_2(\delta, Q, \mathcal{G})$ and $N_2^{[]}(\delta, Q, \mathcal{G})$ are of the same order of magnitude (in δ) for all probability measures Q.

Now, let $B(\rho, \mathcal{G}, g_0)$, $\rho > 0$, be a ball with radius ρ for $\|\cdot\|$ around g_0 intersected with $\mathcal{G}$ and let

$$N^{[]}(\delta, \rho, \mathcal{G}, g_0) = N_2^{[]}(\delta, H, B(\rho, \mathcal{G}, g_0)), \quad 0 < \delta \leqslant \rho.$$

THEOREM 6.3.2. *Suppose $\mathcal{G}$ is a uniformly bounded permissible class with*

$$\sup_{L \geqslant L_0} \sup_{0 < \delta \leqslant \delta_0} \frac{\delta^\nu \log N^{[]}(\delta, L\delta, \mathcal{G}, g_0)}{\log L} \leqslant M, \quad \nu \geqslant 0, \tag{6.33}$$

then $\|\hat{\mathbf{g}}_n - g_0\|_n = \mathcal{O}_{\mathbb{P}}(n^{-\frac{1}{2+\nu}})$ *implies* $\|\hat{\mathbf{g}}_n - g_0\| = \mathcal{O}_{\mathbb{P}}(n^{-\frac{1}{2+\nu}})$.

PROOF. This follows from Lemmas 6.3.3 and 6.3.4 below. □

We first present the probability inequality we use and then prove that the ratio $\|g - g_0\|_n / \|g - g_0\|$ cannot differ too much from 1 if $\|g - g_0\|$ is large enough. Theorem 6.3.2 then follows immediately.

LEMMA 6.3.3. *If* $|g| \leqslant 1$, $|\tilde{g}| \leqslant 1$, *then*

$$\mathbb{P}(|\|g - \tilde{g}\|_n^2 - \|g - \tilde{g}\|^2| \geqslant a) \leqslant 2 \exp\left[-\frac{na^2}{8\|g - \tilde{g}\|^2 + \frac{8}{3}a}\right], \quad a > 0.$$

PROOF. If $\mathbf{z}_1, \cdots, \mathbf{z}_n$ are independent random variables with expectation zero, variance $\mathbb{E}\mathbf{z}_k^2 = \sigma_k^2$ and with $|\mathbf{z}_k| \leqslant M$, $k = 1, \cdots, n$, then Bernstein's inequality (BERNSTEIN (1924, 1927), BENNETT (1962)) says that

$$\mathbb{P}(|\sum_{k=1}^{n} \mathbf{z}_k| \geqslant a) \leqslant 2\exp\left[-\frac{a^2}{2(\sum_{k=1}^{n}\sigma_k^2) + \frac{2}{3}Ma}\right].$$

Apply this with $\mathbf{z}_k = (g(\mathbf{x}_k) - \tilde{g}(\mathbf{x}_k))^2 - \|g - \tilde{g}\|^2$, $|\mathbf{z}_k| \leqslant 4$ and $\mathbb{E}|\mathbf{z}_k|^2 \leqslant 4\|g - g_0\|^2$, $k = 1, \cdots, n$. □

LEMMA 6.3.4. *Suppose that $\mathcal{G}$ is a uniformly bounded class satisfying (6.33) for some $\nu \geqslant 0$ and M. Then for all $\eta > 0$ there exists an $L_\eta > 0$ and $\alpha_\eta > 0$ such that for all $n \geqslant n_0 (= \delta_0^{-(2+\nu)}$, with δ_0 defined in (6.33))*

$$\mathbb{P}^*\left(\sup_{\substack{g \in \mathcal{G} \\ \|g - g_0\| \geqslant L_\eta n^{-\frac{1}{2+\nu}}}} \left|\frac{\|g - g_0\|_n}{\|g - g_0\|} - 1\right| > \eta\right) \leqslant \frac{8}{\alpha_\eta} \exp\left[-\alpha_\eta L_\eta^2 n^{\frac{\nu}{2+\nu}}\right].$$

PROOF. Define $\delta_n = n^{-\frac{1}{2+\nu}}$. Assume without loss of generality that $|g| \leqslant 1$ for all $g \in \mathcal{G}$. Let $\{[g_1, g_2]\}$ be a minimal δ_n-bracketing set of $B(L\delta_n, \mathcal{G}, g_0)$, where $L \geqslant L_\eta$, L_η to be specified later. We shall first show that for all $L \geqslant L_\eta$, $n \geqslant n_0$

$$\begin{aligned} \mathbb{P}_L &= \mathbb{P}\left(\sup_{g \in B(L\delta_n, \mathcal{G}, g_0)} \|g - g_0\|_n > (1 + \tfrac{1}{2}\eta)L\delta_n\right) \qquad (6.35) \\ &\leqslant 4\exp\left[-\alpha_\eta L^2 n^{\nu/(2+\nu)}\right]. \end{aligned}$$

Let $\{g_1\}$ be the set of left brackets from $\{[g_1, g_2]\}$. We have

$$\begin{aligned} \mathbb{P}_L &\leqslant \mathbb{P}^*\left(\sup_{\substack{\|g_1 - g_0\| \leqslant (L+1)\delta_n \\ g_1 \in \{g_1\}}} \|g_1 - g_0\|_n > (1 + \tfrac{1}{4}\eta)L\delta_n\right) \\ &+ \mathbb{P}^*\left(\sup_{[g_1, g_2] \in \{[g_1, g_2]\}} \|g_1 - g_2\|_n > \tfrac{1}{4}\eta L\delta_n\right) = \mathbb{P}_L^{(1)} + \mathbb{P}_L^{(2)}. \end{aligned}$$

If we take L_η sufficiently large, such that $((1 + \frac{1}{4}\eta)L)^2 - (L+1)^2 > \frac{1}{8}\eta L^2$ for all $L \geqslant L_\eta$, then

$$\begin{aligned} \mathbb{P}_L^{(1)} &= \mathbb{P}^*\left(\sup_{\substack{\|g_1 - g_0\| \leqslant (L+1)\delta_n \\ g_1 \in \{g_1\}}} \|g_1 - g_0\|_n > (1 + \tfrac{1}{4}\eta)L\delta_n\right) \\ &\leqslant \mathbb{P}^*\left(\sup_{\substack{\|g_1 - g_0\| \leqslant (L+1)\delta_n \\ g_1 \in \{g_1\}}} \|g_1 - g_0\|_n^2 - \|g_1 - g_0\|^2 > ((1 + \tfrac{1}{4}\eta)L\delta_n)^2 - ((L+1)\delta_n)^2\right) \end{aligned}$$

$$\leqslant \mathbb{P}^* \left[\sup_{\substack{\|g_1 - g_0\| \leqslant (L+1)\delta_n \\ g_1 \in \{g_1\}}} \|g_1 - g_0\|_n^2 - \|g_1 - g_0\|^2 > \tfrac{1}{8}\eta L^2 \delta_n^2 \right].$$

From Lemma 6.3.3 we see that for $\|g_1 - g_0\| \leqslant (L+1)\delta_n$

$$\mathbb{P}\left[\|g_1 - g_0\|_n^2 - \|g_1 - g_0\|^2 > \tfrac{1}{8}\eta L^2 \delta_n^2 \right] \leqslant 2\exp\left[-n\alpha_\eta^{(1)} L^2 \delta_n^2 \right],$$

for all $L \geqslant L_\eta$, L_η sufficiently large and for some constant $\alpha_\eta^{(1)}$ depending only on η. Moreover, for $L \geqslant L_0$ and $n \geqslant n_0 = \delta_0^{-(2+\nu)}$

$$\log N^{[]}(\delta_n, L\delta_n, \mathcal{G}, g_{0,n}) \leqslant M(\log L)\delta_n^{-\nu},$$

and $M(\log)\delta_n^{-\nu} \leqslant \frac{1}{2} n\alpha_\eta^{(1)} L^2 \delta_n^2$ for all $L \geqslant L_\eta$, L_η sufficiently large. Hence

$$\begin{aligned} \mathbb{P}_L^{(1)} &\leqslant N^{[]}(\delta_n, L\delta_n, \mathcal{G}, g_0) 2\exp\left[-n\alpha_\eta^{(1)} L^2 \delta_n^2 \right] \\ &\leqslant 2\exp\left[-\tfrac{1}{2}\alpha_\eta^{(1)} L^2 n^{\nu/(2+\nu)} \right], \quad L \geqslant L_\eta, \quad n \geqslant n_0. \end{aligned}$$

As for $\mathbb{P}_L^{(2)}$, we have that for $L \geqslant L_\eta$, L_η sufficiently large, $n \geqslant n_0$

$$\begin{aligned} \mathbb{P}_L^{(2)} &= \mathbb{P}^* \left[\sup_{[g_1, g_2] \in \{[g_1, g_2]\}} \|g_1 - g_2\|_n > \tfrac{1}{4}\eta L \delta_n \right] \\ &\leqslant \mathbb{P}^* \left[\sup_{[g_1, g_2] \in \{[g_1, g_2]\}} \|g_1 - g_2\|_n^2 - \|g_1 - g_2\|^2 > (\tfrac{1}{4}\eta L\delta_n)^2 - \delta_n^2 \right] \\ &\leqslant \mathbb{P}^* \left[\sup_{[g_1, g_2] \in \{[g_1, g_2]\}} \|g_1 - g_2\|_n^2 - \|g_1 - g_2\|^2 > \frac{1}{32}\eta^2 L^2 \delta_n^2 \right] \\ &\leqslant 2\exp\left[-\tfrac{1}{2}\alpha_\eta^{(2)} L^2 n^{\nu/(2+\nu)} \right], \end{aligned}$$

for some constant $\alpha_\eta^{(2)}$.

Thus for $\tilde{\alpha}_\eta \leqslant \frac{1}{2}\min(\alpha_\eta^{(1)}, \alpha_\eta^{(2)})$

$$\mathbb{P}_L \leqslant \mathbb{P}_L^{(1)} + \mathbb{P}_L^{(2)} \leqslant 4\exp\left[-\tilde{\alpha}_\eta L^2 n^{\nu/(2+\nu)} \right],$$

for all $L \geqslant L_\eta$, $n \geqslant n_0$. This proves (6.35). Assertion (6.35) in turn implies that if we take $L \geqslant L_\eta$, L_η sufficiently large, $n \geqslant n_0$

$$\mathbb{P}^* \left[\sup_{\substack{(L-1)\delta_n \leqslant \|g - g_0\| \leqslant L\delta_n \\ g \in \mathcal{G}}} \frac{\|g - g_0\|_n}{\|g - g_0\|} > (1+\eta) \right] \tag{6.36}$$

$$\leqslant \mathbb{P}^*\left[\sup_{g\in B(L\delta_n,\,\hat{\mathcal{G}},\,g_0)} \|g-g_0\|_n > (1+\eta)(L-1)\delta_n\right]$$

$$\leqslant \mathbb{P}^*\left[\sup_{g\in B(L\delta_n,\,\hat{\mathcal{G}},\,g_0)} \|g-g_0\|_n > (1+\tfrac{1}{2}\eta)L\delta_n\right] = \mathbb{P}_L$$

$$\leqslant 4\exp(-\tilde{\alpha}_\eta L^2 n^{\nu/(2+\nu)}).$$

Similarly,

$$\mathbb{P}^*\left[\inf_{\substack{(L-1)\delta_n\leqslant\|g-g_0\|\leqslant L\delta_n \\ g\in\hat{\mathcal{G}}}} \frac{\|g-g_0\|_n}{\|g-g_0\|} < 1-\eta\right]$$

$$\leqslant \mathbb{P}^*\left[\inf_{(L-1)\delta_n\leqslant\|g-g_0\|\leqslant L\delta_n} \|g-g_0\|_n < (1-\eta)L\delta_n\right]$$

$$\leqslant \mathbb{P}^*\left[\inf_{\substack{(L-2)\delta_n\leqslant\|g_1-g_0\|\leqslant((L+1)\delta_n \\ g_1\in\{g_1\}}} \|g_1-g_0\|_n < (1-\tfrac{1}{2}\eta)L\delta_n\right]$$

$$+\ \mathbb{P}^*\left[\sup_{[g_1,\,g_2]\in\{[g_1,\,g_2]\}} \|g_1-g_2\|_n \geqslant \tfrac{1}{2}\eta L\delta_n\right]$$

$$\leqslant \mathbb{P}^*\left[\sup_{\substack{\|g_1-g_0\|\leqslant(L+1)\delta_n \\ g_1\in\{g_1\}}} \|g_1-g_0\|^2 - \|g_1-g_0\|_n^2 > (L-2)^2\delta_n^2 - (1-\tfrac{1}{2}\eta)^2 L^2\delta_n^2\right]$$

$$+\ \mathbb{P}^*\left[\sup_{[g_1,\,g_2]\in\{[g_1,\,g_2]\}} \|g_1-g_2\|_n \geqslant \tfrac{1}{4}\eta L\delta_n\right] = \mathbb{P}_L^{(3)} + \mathbb{P}_L^{(2)}.$$

We already showed that for $L\geqslant L_\eta$, $n\geqslant n_0$

$$\mathbb{P}_L^{(2)} \leqslant 2\exp\left[-\tfrac{1}{2}\alpha_\eta^{(1)} L^2 n^{\nu(2+\nu)}\right].$$

If we take L_η sufficiently large then for $L\geqslant L_\eta$

$$\mathbb{P}_L^{(3)} = \mathbb{P}^*\left[\sup_{\substack{\|g_1-g_0\|\leqslant(L+1)\delta_n \\ g_1\in\{g_1\}}} \|g_1-g_0\|^2 - \|g_1-g_0\|_n^2 > (L-2)^2\delta_n^2 - (1-\tfrac{1}{2}\eta)^2 L^2\delta_n^2\right]$$

$$\leqslant \mathbb{P}^*\left[\sup_{\substack{\|g_1-g_0\|\leqslant(L+1)\delta_n \\ g_1\in\{g_1\}}} \|g_1-g_0\|^2 - \|g_1-g_0\|_n^2 > \tfrac{1}{4}\eta(1-\tfrac{1}{2}\eta)L^2\delta_n^2\right]$$

$$\leqslant 2\exp\left[-\tfrac{1}{2}\alpha_\eta^{(3)} L^2 n^{\nu/(2+\nu)}\right]$$

for some $\alpha_\eta^{(3)}$, $n \geqslant n_0$. Hence for $\alpha_\eta \leqslant \frac{1}{2}\min(\alpha_\eta^{(2)}, \alpha_\eta^{(3)})$

$$\mathbb{P}^* \left[\inf_{\substack{(L-1)\delta_n \leqslant \|g-g_0\| \leqslant L\delta_n \\ g \in \mathcal{G}}} \frac{\|g-g_0\|_n}{\|g-g_0\|} > 1-\eta \right] \tag{6.37}$$
$$\leqslant 4\exp\left[-\alpha_\eta L^2 n^{\nu/(2+\nu)}\right].$$

Finally, combine (6.36) and (6.37) to obtain that

$$\mathbb{P}^* \left[\inf_{\substack{\|g-g_0\| \geqslant L_\eta \delta_n \\ g \in \mathcal{G}}} \frac{\|g-g_0\|_n}{\|g-g_0\|} > 1-\eta \right] \leqslant \sum_{L \geqslant L_\eta} 8\exp\left[-\alpha_\eta L^2 n^{\nu/(2+\nu)}\right]$$

$$\leqslant \frac{8}{\alpha_\eta} \exp\left[-\alpha_\eta L_\eta^2 n^{\nu/(2+\nu)}\right]. \quad \square$$

6.4. Application to two-phase regression

We consider the models of Chapter 5 and compare the various sets of assumptions and outcomes with those of Section 6.2. To avoid digressions, we assume throughout that the disturbances $\epsilon_1, \epsilon_2, \cdots$ form an i.i.d. sequence (ϵ_1 having expection zero and finite variance) and that $g_{0,n} = g_0$ is fixed. We start with the continuous model:

$$\mathcal{G} = \{g_{\theta,c}(x) = \min(\alpha^{(1)} + x\beta^{(1)}, \alpha^{(2)} + x\beta^{(2)}): \tag{6.38}$$
$$\theta = \begin{pmatrix} \theta^{(1)} \\ \theta^{(2)} \end{pmatrix}, \ \theta^{(i)} = \begin{pmatrix} \alpha^{(i)} \\ \beta^{(i)} \end{pmatrix}, \ i = 1,2\}$$

LEMMA 6.4.1. *Let $\mathbf{x}_{n,k} = \mathbf{x}_k$, $k = 1, \cdots, n$, with $\mathbf{x}_1, \mathbf{x}_2, \cdots$ a sequence of i.i.d. random vectors with distribution H. Let $\mathcal{G}$ be given by (6.38). Then there exists a constant $K_2 < \infty$ such that for all n sufficiently large and for all $g_{\theta,c}, g_{\tilde{\theta},c} \in \mathcal{G}$*

$$\|g_{\theta,c} - g_{\tilde{\theta},c}\|_n \leqslant K_2\|\theta - \tilde{\theta}\| \quad \textit{almost surely.} \tag{6.39}$$

Define for all $\eta > 0$ the restricted class

$$\mathcal{G}_R(\eta) = \{g_{\theta,c} \in \mathcal{G}: \|\theta - \theta_0\| < \eta\}.$$

Suppose that there exists a set of points

$$\{x_t^{(i)}: t = 1, \cdots, 2(d+1)-1\} \subset A_0^{(i)} \cap T, \tag{6.40}$$

where T is the support of H, and no $d+1$ $x_t^{(i)}$ lie on a $(d-1)$-dimensional hyperplane, $i = 1,2$, and that

$$\|\theta_0^{(1)} - \theta_0^{(2)}\| \neq 0. \tag{6.41}$$

Then there exists an $\eta > 0$ and a constant $K_1 > 0$ such that

$$\|g_{\theta,c}-g_{\theta_0,c}\|_n \geqslant K_1\|\theta-\theta_0\| \quad \textit{almost surely}, \tag{6.42}$$

for all n sufficiently large and all $g_{\theta,c}\in\mathcal{G}_R(\eta)$.

PROOF. Result (6.39) follows from the fact that the functions $g_{\theta,c}(x)$ are Lipschitz continuous in θ for every x:

$$|g_{\theta,c}(x)-g_{\tilde{\theta},c}(x)| \leqslant J(x)\|\theta-\tilde{\theta}\|,$$

where $J(x)=1+|z_1|+\cdots+|z_d|$, $x=(z_1,\cdots,z_d)$. Since $\|J\|_n\to\|J\|$ almost surely,

$$\|g_{\theta,c}-g_{\tilde{\theta},c}\|_n \leqslant 2\|J\|\|\theta-\tilde{\theta}\| \quad \text{almost surely,}$$

for all n sufficiently large.

Inequality (6.42) is of course closely related to (5.11) (see the proof of Theorem 5.2.1) which asserts that (6.42) holds for $\theta=\hat{\theta}_n$. Condition (6.40) implies that if the $\theta^{(i)}$, $i=1,2$, in θ are appropriately indexed, then there are at least $(d+1)$ $x_t^{(1)}$'s in $A_\theta^{(1)}$. This implies that, from a possible re-indexing, the smallest eigenvalue of $\mathbf{\Sigma}_n(A_\theta^{(1)}\cap A_0^{(i)})$ is bounded away from zero for all θ and all n sufficiently large. Hence

$$(\theta^{(1)}-\theta_0^{(1)})^T\mathbf{\Sigma}_n(A_\theta^{(1)}\cap A_0^{(1)})(\theta^{(1)}-\theta_0^{(1)}) \geqslant K_{1,1}^2\|\theta^{(1)}-\theta_0^{(1)}\|^2 \tag{6.43}$$

for some constant $K_{1,1}>0$, all properly indexed θ and all n sufficiently large. Moreover, by taking η sufficiently small we see that $\|\theta-\theta_0\|<\eta$ and (6.39) imply that $A_\theta^{(1)}$ cannot contain more than d $x_t^{(2)}$'s, because of assumption (6.41). Thus, for η sufficiently small the eigenvalues of $\mathbf{\Sigma}_n(A_\theta^{(2)}\cap A_0^{(2)})$, $\|\theta-\theta_0\|<\eta$, are eventually also bounded away from zero, and so

$$(\theta^{(2)}-\theta_0^{(2)})^T\mathbf{\Sigma}_n(A_\theta^{(2)}\cap A_0^{(2)})(\theta^{(2)}-\theta_0^{(2)}) \geqslant K_{2,2}^2\|\theta^{(2)}-\theta_0^{(2)}\|^2$$

for some constant $K_{2,2}>0$, all $\|\theta-\theta_0\|<\eta$ and all n sufficiently large. (In fact, if η is sufficiently small re-indexing of θ, $\|\theta-\theta_0\|<\eta$, in (6.43) is not needed). Thus

$$\|g_{\theta,c}-g_{\theta_0,c}\|_n^2 \geqslant \sum_{i=1,2}(\theta^{(i)}-\theta_0^{(i)})^T\mathbf{\Sigma}_n(A_\theta^{(i)}\cap A_0^{(i)})(\theta^{(i)}-\theta_0^{(i)})$$

$$\geqslant \left\{\min_{i=1,2} K_{i,i}^2\right\}\|\theta-\theta_0\|^2 \quad \text{almost surely}$$

for all $\|\theta-\theta_0\|<\eta$, η sufficiently small, and all n sufficiently large. □

In other words, under (6.40) and (6.41) the sequence $\{\mathcal{G}_R(\eta), \|\cdot\|_n\}$, with $\mathcal{G}_R(\eta)$ defined in Lemma 6.4.1, is for η small enough of finite metric dimension $2(d+1)$. We can now apply the results of Section 6.2, because in Lemma 3.4.4 the strong consistency of $\tilde{\theta}_n$ is established, i.e. for every $\eta>0$ $\tilde{\theta}_n\in\mathcal{G}_R(\eta)$ for all n sufficiently large. The conditions of Lemma 3.4.4 include (6.40) and (6.41). Recall furthermore that in Theorem 5.2.1 we also needed the conditions of Lemma 3.4.4. The following proposition collects previous results and those

obtained by application of the theory in Section 6.2.

PROPOSITION 6.4.2. *Define $\mathcal{G}$ as in (6.8) and let $\mathbf{x}_{n,k}=\mathbf{x}_k$, with $\mathbf{x}_1, \mathbf{x}_2, \cdots$ i.i.d. with distribution H. Suppose that the conditions of Lemma 3.4.4 are fulfilled. We have*

(i) $\|\hat{\boldsymbol{\theta}}_n-\theta_0\|\to 0$ *and* $H(\hat{\mathbf{A}}_n \Delta A_0)\to 0$, *almost surely.*

(ii) *if $\mathbb{E}|\epsilon_1|^{2p}<\infty$ for some $p>1$, then $\hat{\boldsymbol{\theta}}_n^{(1)}$ and $\hat{\boldsymbol{\theta}}_n^{(2)}$ are asymptotically independent and*

$$n^{1/2}(\hat{\boldsymbol{\theta}}_n^{(i)}-\theta_0^{(i)}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \|\epsilon\|^2 \Sigma^{-1}(A_0^{(i)})),\ i=1,2,$$

(iii) *if $\mathbb{E}|\epsilon_1|^{2p}<\infty$ for some $p>\frac{1}{2}r(=d+1)$, then for all $L\geqslant L'$, $n\geqslant n'$*

$$\mathbb{P}(\|\hat{\boldsymbol{\theta}}_n-\theta_0\| > n^{-1/2}L) \leqslant A' L^{-(2p-r)},$$

(iv) *if $\mathbb{E}\exp(\beta|\epsilon_1|^2)<\infty$ for some $\beta>0$, then for all $L\geqslant L'$, $n\geqslant n'$*

$$\mathbb{P}(\|\hat{\boldsymbol{\theta}}_n-\theta_0\|>n^{-1/2}L) \leqslant \exp(-M'L^2).$$

PROOF.

(i) This is Lemma 3.4.4.

(ii) This is Theorem 5.2.1.

(iii) Combine Lemma 3.4.4, Lemma 6.2.3 and Lemma 6.4.1.

(iv) Combine Lemma 3.4.4, Theorem 6.2.5 and Lemma 6.4.1. □

Note that the $\mathcal{O}_{\mathbb{P}}(n^{-1/2})$-rate for $\hat{\boldsymbol{\theta}}_n$ in (iii) and (iv) of Proposition 6.4.2 follows from the $\mathcal{O}_{\mathbb{P}}(n^{-1/2})$-rate for $\hat{\mathbf{g}}_n$. The situation is somewhat different in the discontinuous model. Here, the class of regression functions is

$$\mathcal{G} = \{g_{\theta,A}(x)=\sum_{i=1,2}(\alpha^{(i)}+x\beta^{(i)})1_{A^{(i)}}(x):\ \theta=\begin{bmatrix}\theta^{(1)}\\ \theta^{(2)}\end{bmatrix}, \tag{6.44}$$

$$\theta^{(i)}=\begin{bmatrix}\alpha^{(i)}\\ \beta^{(i)}\end{bmatrix},\ A^{(i)}\in\mathcal{A}^{(i)},\ i=1,2\}.$$

We shall first consider a special case with $d=1$. This will clarify the difficulties in higher dimensions.

LEMMA 6.4.3. *Let $d=1$, $\mathbf{x}_{n,k}=\mathbf{x}_k$, $k=1,\cdots,n$, with $\mathbf{x}_1, \mathbf{x}_2, \cdots$ i.i.d. with distribution function $H: \mathbb{R}\to\mathbb{R}$. Let $\mathcal{G}$ be defined in (6.44), with $\mathcal{A}=\{A_\gamma=(-\infty,\gamma]:\ \gamma\in\mathbb{R}\}$. Define for all $\eta>0$*

$$\mathcal{G}_R(\eta) = \{g_{\theta,A_\gamma}\in\mathcal{G}:\ \|\theta-\theta_0\|<\eta,\ H(A_\gamma\Delta A_0)<\eta\}.$$

Suppose there exist $\{x_t^{(i)}:\ t=1,2,3,\ x_{t_1}^{(i)}\neq x_{t_2}^{(i)},\ t_1\neq t_2\}\subset A_0^{(i)}\cap T$, $i=1,2$. Furthermore, suppose that the discontinuity assumption (5.22) can be strengthened to: for some $\eta_1>0$, $K>0$ we have

$$H(\gamma_0-\eta_1,\ \gamma_0+\eta_1] > 0$$

and

$$|g_{\theta_0^{(1)}}(x)-g_{\theta_0^{(2)}}(x)| > K,$$

for all $x\in(\gamma_0-\eta_1, \gamma_0+\eta_1]$. *Then for* η *sufficiently small*

$$\mathbf{N}_n(\delta, 2^j\delta, \mathcal{G}_R(\eta), g_0) \leqslant A\ 2^{rj}, \quad \textit{almost surely},$$

for all n sufficiently large, where $r=2(d+1)+2=6$.

PROOF. Let $g_{\theta,A_\gamma}\in\mathbf{B}_n(2^j\delta, \mathcal{G}_R(\eta), g_0)$:

$$\|g_{\theta,A_\gamma}-g_0\|_n \leqslant 2^j\delta.$$

Since for η sufficiently small $H(A_\gamma\Delta A_0)<\eta$ implies that $\gamma\in(\gamma_0-\eta_1, \gamma_0+\eta_1]$, we have

$$(2^j\delta)^2 \geqslant \|(g_{\theta,A_\gamma}-g_0)1_{A_\gamma\Delta A_0}\|_n^2 \geqslant K\mathbf{H}_n(A_\gamma\Delta A_0).$$

Also

$$(2^j\delta) \geqslant \|(g_{\theta,A_\gamma}-g_0)1_{A_\gamma^{(i)}\cap A_0^{(i)}}\|_n \geqslant K_i\|\theta^{(i)}-\theta_0^{(i)}\|, \quad \text{almost surely}, \quad i=1,2$$

for some $K_i>0$, η sufficiently small and n sufficiently large. Hence

$$\mathbf{B}_n(2^j\delta, \mathcal{G}_R(\eta), g_0)\subset\{g_{\theta,A_\gamma}\in\mathcal{G}_R(\eta): \|\theta^{(i)}-\theta_0^{(i)}\|\leqslant 2^j\delta/K_i,\ i=1,2,$$
$$H(A_\gamma\Delta A_0)\leqslant(2^j\delta)^2/K\}.$$

Since

$$N_2(\delta, \mathbf{H}_n, \{A_\gamma: H(A_\gamma\Delta A_0)\leqslant(2^j\delta)^2/K\}) \leqslant \tilde{A}2^{2j} \tag{6.45}$$

for all $0<\delta<1$, this implies that for η sufficiently small

$$\mathbf{N}_n(\delta, 2^j\delta, \mathcal{G}_R(\eta), g_0) \leqslant A2^{rj}, \quad r=2(d+1)+2. \quad \square$$

Equality (6.45) in the proof of Lemma 6.4.3 is a special feature of the class of intervals $\{(-\infty, \gamma]: \gamma\in\mathbb{R}\}$. If $d=2$ and $\mathcal{A}=\{x: x\gamma\leqslant 1\}: \gamma\in\mathbb{R}^2\}$, then in general the number of $\mathbf{x}_k$ in the set

$$\bigcup\{A\in\mathcal{A}: \mathbf{H}_n(A\Delta A_0) \leqslant (2^j n^{-1/2})^2\} \tag{6.46}$$

need not remain bounded (see Example 6.5). It is not clear how to calculate the entropy of neighbourhoods like (6.46) for general $\mathcal{A}$. An upper bound is of course the global entropy of $\mathcal{A}$. We use this upper bound in (iv) of Proposition 6.4.4.

PROPOSITION 6.4.4. *Let* $\mathcal{G}$ *be defined in (6.44) and let* $\mathbf{x}_{n,k}=\mathbf{x}_k$, $\mathbf{x}_1, \mathbf{x}_2, \cdots$ *i.i.d. with distribution* $\mathcal{G}$. *Suppose that the conditions of Lemma 3.4.4 are fulfilled. We have*

(i) $\|\hat{\boldsymbol{\theta}}_n-\theta_0\|\to 0$, $H(\hat{\mathbf{A}}_n\Delta A_0)\to 0$ *almost surely.*

Suppose in addition that H satisfies (5.24).

(ii) *If the discontinuity assumption (5.22) holds and if moreover* $D_2(\delta, \mathcal{G}) \leq \exp(M\delta^{-\nu})$ *and* $\mathbb{E}|\epsilon_1|^{2p} < \infty$, $p > 2/(2-\nu)$, $0 < \nu < 2$, *then* $\hat{\boldsymbol{\theta}}_n^{(1)}$ *and* $\hat{\boldsymbol{\theta}}_n^{(2)}$ *are asymptotically independent, and*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n^{(i)} - \boldsymbol{\theta}_0^{(i)}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \|\epsilon\|^2 \Sigma^{-1}(A_0^{(i)})), \quad i = 1,2.$$

(iii) *If* $d = 1$ *and the conditions of Lemma 6.4.3 hold (i.e. if (5.22) is replaced by the stronger assumption), then* $\mathbb{E}|\epsilon_n|^{2p} < \infty$ *for some* $p > 6$ *implies that*

$$\|\hat{\mathbf{g}}_n - g_0\|_n = \mathcal{O}_{\mathbb{P}}(n^{-1/2}).$$

(iv) *If* $D_2(\delta, \mathcal{G}) \leq \tilde{A}\delta^{-\tilde{r}}$, $\tilde{r} > 0$, *and* $\mathbb{E}\exp(\beta|\epsilon_1|^2) < \infty$ *for some* $\beta > 0$, *then*

$$\|\hat{\mathbf{g}}_n - g_0\|_n = \mathcal{O}_{\mathbb{P}}(n^{-1/2}(\log n)^{1/2}).$$

If $D_2(\delta, \mathcal{G}) \leq \exp(M\delta^{-\nu})$, $0 < \nu < 2$ *and* $\mathbb{E}\exp(\beta|\epsilon_1|^2) < \infty$ *for some* $\beta > 0$, *then*

$$\|\hat{\mathbf{g}}_n - g_0\|_n = \mathcal{O}_{\mathbb{P}}(n^{-\frac{1}{2+\nu}}).$$

PROOF.

(i) This is Lemma 3.4.4.

(ii) This is Lemma 5.3.1.

(iii) Combine Lemma 3.4.4, Theorem 6.2.2 and Lemma 6.4.3.

(iv) For η_0 defined in (5.24), the class

$$\mathcal{G}_R(\eta_0) = \{g_{\theta,A} \in \mathcal{G}: \|\theta - \theta_0\| \leq \eta_0, H(A \Delta A_0) < \eta_0\}$$

satisfies

$$N_2(\delta, \mathbf{H}_n, \mathcal{G}_R(\eta_0)) \leq A\delta^{-2(d+1)} D_2(\delta, \mathcal{G}).$$

Insert this in conditions (6.20) and (6.21) of Theorem 6.2.5, with $\delta_n = n^{-1/2}(\log n)^{1/2}$ and $\delta_n = n^{-1/(2+\nu)}$ respectively. □

EXAMPLE 6.5. Let $d = 2$, $\mathcal{A} = \{A(\gamma) = \{x: x\gamma \leq 1\}, \gamma \in \mathbb{R}^2\}$, $g_{\theta^{(i)}} = \alpha^{(i)}$, $i = 1,2$, $\alpha_0^{(1)} \neq \alpha_0^{(2)}$ (i.e. we assume for simplicity that $\beta_0^{(i)} = 0$, $i = 1,2$ is known), and let H be the uniform distribution on $A_0^{(1)} \cup A_0^{(2)}$, where $A_0^{(1)}$ and $A_0^{(2)}$ are the two disjoint discs defined in Example 5.3. Since $\mathcal{A}$ is a VC-class, it follows from Proposition 6.4.3 (iii) that if $\mathbb{E}\exp(\beta|\epsilon_1|^2) < \infty$, then $\|\hat{\mathbf{g}}_n - g_0\|_n = \mathcal{O}_{\mathbb{P}}(n^{-1/2}(\log n)^{1/2})$. This implies the rate $\mathcal{O}_{\mathbb{P}}(n^{-1/2}(\log n)^{1/2})$ for the estimator of $\alpha_0^{(i)}$, $i = 1,2$, but from (ii) of Proposition 6.4.3 we know that in fact $|\hat{\boldsymbol{\alpha}}_n^{(i)} - \alpha_0^{(i)}| = \mathcal{O}_{\mathbb{P}}(n^{-1/2})$, $i = 1,2$. The rate for $\hat{\mathbf{g}}_n$ also implies that $\mathbf{H}_n(\hat{\mathbf{A}}_n \Delta A_0) = \mathcal{O}_{\mathbb{P}}(n^{-1}\log n)$, i.e. $\mathcal{O}_{\mathbb{P}}(\log n)$ observations are assigned to the wrong sample. This rate cannot be improved, in the sense that if e.g. $\epsilon_1, \epsilon_2, \cdots$ are normally distributed, then one can show that for some $a > 0$

$$\liminf_{n \to \infty} \mathbb{P}(\mathbf{H}_n(\hat{\mathbf{A}}_n \Delta A_0) > a\frac{\log n}{n}) > 0.$$

In the following three examples, we again restrict ourselves to the case $d=1$ and $\mathcal{A}=\{(-\infty,\gamma]\colon \gamma\in\mathbb{R}\}$. We take nonrandom $x_{n,k}$, with the particular choice $x_{n,k}=k/n$. Speeds of estimation are investigated in the discontinuous model, with the assumption of discontinuity of the underlying true regression function (Example 6.6), the assumption of continuity and identifiability of the underlying regression (Example 6.7) or without identifiability at g_0 (Example 6.8). The first example treats virtually the same situation as the one in Lemma 6.4.3. We present it to facilitate the comparison with Example 6.7.

EXAMPLE 6.6. Let $d=1$, $x_{n,k}=\dfrac{k}{n}$, $k=-[(n-1)/2],\ \cdots,\ [n/2]$,

$$g_{\theta,A_\gamma}(x) = \sum_{i=1,2}(\alpha^{(i)}+x\beta^{(i)})1_{A^{(i)}}(x),\ \theta\in\mathbb{R}^4,\ A^{(1)}=(-\infty,\gamma],$$

and

$$g_0(x) = \sum_{i=1,2}(\alpha_0^{(i)}+x\beta_0^{(i)})1_{A_0^{(i)}}(x),\ \alpha_0^{(1)}\neq\alpha_0^{(2)}, A_0^{(1)}=(-\infty,\gamma_0],\ \gamma_0=0.$$

Application of Proposition 3.4.5 yields that $\|\hat{\boldsymbol{\theta}}_n-\theta_0\| \xrightarrow{\mathbb{P}} 0$ and $|\hat{\gamma}_n-\gamma_0| \xrightarrow{\mathbb{P}} 0$. Moreover, for η sufficiently small the class

$$\mathcal{G}_R(\eta) = \{g_{\theta,A}\colon \|\theta-\theta_0\|\leqslant\eta,\ |\gamma-\gamma_0|<\eta\}$$

satisfies for some constant A

$$\sup_{j\geqslant j_0}\sup_{n\geqslant n_0}\sup_{\delta\leqslant\delta_0}\frac{N_n(\delta_1 2^j\delta,\mathcal{G}_R(\eta),g_0)}{2^{6j}}\leqslant A$$

Hence if $\mathbb{E}|\epsilon_1|^{2p}<\infty$ for some $p>6$

$$\|\hat{\mathbf{g}}_n-g_0\|_n = \mathcal{O}_{\mathbb{P}}(n^{-1/2})$$

which implies

$$\|\hat{\boldsymbol{\theta}}_n-\theta_0\| = \mathcal{O}_{\mathbb{P}}(n^{-1/2}),\quad |\hat{\gamma}_n-\gamma_0| = \mathcal{O}_{\mathbb{P}}(\frac{1}{n}).$$

It is now not difficult to prove that $\hat{\boldsymbol{\theta}}_n^{(1)}$, $\hat{\boldsymbol{\theta}}_n^{(2)}$ and $\hat{\gamma}_n$ are asymptotically independent, with limiting distributions

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n^{(1)}-\theta_0^{(1)}) \xrightarrow{\mathfrak{L}} \mathcal{N}(0,\|\epsilon\|^2(\int_{-1/2}^{0}\begin{bmatrix}1 & x\\ x & x^2\end{bmatrix}dx)^{-1}),$$

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n^{(2)}-\theta_0^{(2)}) \xrightarrow{\mathfrak{L}} \mathcal{N}(0,\|\epsilon\|^2(\int_{0}^{1/2}\begin{bmatrix}1 & x\\ x & x^2\end{bmatrix}dx)^{-1}),$$

$$n(\hat{\gamma}_n-\gamma_0) \xrightarrow{\mathfrak{L}} \arg\sup_{l\geqslant 0}2(\alpha_0^{(1)}-\alpha_0^{(2)})\sum_{k=0}^{l}\epsilon_k-(\alpha_0^{(1)}-\alpha_0^{(2)})^2 l$$

(compare with Example 5.2).

EXAMPLE 6.7. Let $d=1$, $x_{n,k}=\frac{k}{n}$, $k=-[(n-1)/2], \cdots, [n/2]$,

$$g_{\theta,A_\gamma}(x) = \sum_{i=1,2} (\alpha^{(i)}+x\beta^{(i)})1_{A^{(i)}}(x),\ \theta\in\mathbb{R}^4,\ A^{(1)}=(-\infty,\gamma],$$

and

$$g_0(x) = \min(0, x\beta_0),\ \beta_0>0\ (A_0=(-\infty,\gamma_0],\ \gamma_0=0).$$

From Proposition 3.4.5 we obtain that $\|\hat{\boldsymbol{\theta}}_n-\theta_0\| \xrightarrow{\mathbb{P}} 0$ and $|\hat{\gamma}_n-\gamma_0| \xrightarrow{\mathbb{P}} 0$. For η sufficiently small, the class

$$\mathcal{G}_R(\eta) = \{g_{\theta,A}: \|\theta-\theta_0\|\leqslant\eta,\ |\eta-\eta_0|\leqslant\eta\}$$

satisfies for some constant A

$$\sup_{j\geqslant j_0}\sup_{n\geqslant n_0}\sup_{\delta\leqslant\delta_0}\frac{N_n(\delta,2^j\delta,\mathcal{G}_R(\eta),g_0)}{2^{(4+2/3)j}} \leqslant A$$

Hence if $\mathbb{E}|\epsilon_1|^{2p}<\infty$ for some $p>4+\frac{2}{3}$

$$\|\hat{\mathbf{g}}_n-g_0\|_n = \mathcal{O}_{\mathbb{P}}(n^{-1/2})$$

which implies

$$\|\hat{\boldsymbol{\theta}}_n-\theta_0\| = \mathcal{O}_{\mathbb{P}}(n^{-1/2}),\quad |\hat{\gamma}_n-\gamma_0| = \mathcal{O}_{\mathbb{P}}(n^{-1/3}).$$

It can be shown that $\hat{\boldsymbol{\theta}}_n^{(2)}$ and $\hat{\gamma}_n$ are asymptotically independent, with limiting distributions

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n^{(1)}-\theta_0^{(1)}) \xrightarrow{\mathcal{L}} \mathcal{N}(0,\|\epsilon\|^2(\int_{-1/2}^{0}\begin{bmatrix}1 & x\\ x & x^2\end{bmatrix}dx)^{-1}),$$

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n^{(2)}-\theta_0^{(2)}) \xrightarrow{\mathcal{L}} \mathcal{N}(0,\|\epsilon\|^2(\int_{0}^{1/2}\begin{bmatrix}1 & x\\ x & x^2\end{bmatrix}dx)^{-1}),$$

$$n^{1/3}(\hat{\gamma}_n-\gamma_0) \xrightarrow{\mathcal{L}} \arg\sup_{l\geqslant 0}(2\beta_0\|\epsilon\|\int_0^{\mu} x d\mathbf{W}(x)-\beta_0^2\mu^3/3)$$

where $\mathbf{W}(\cdot)$ is standard Brownian motion. So the difference with Example 6.6 lies in the slower rate for $\hat{\gamma}_n$.

EXAMPLE 6.8. Let $d=1$, $x_{n,k}=k/n$, $k=1,\cdots,n$,

$$g_{\alpha,A_\gamma} = \alpha 1_{(-\infty,\gamma]},\ \alpha>0,\ A_\gamma=(-\infty,\gamma],\ \gamma\geqslant 0$$

and

$$g_0\equiv 0.$$

Then for $\mathcal{G}=\{g_{\alpha,A_\gamma}: \alpha\in\mathbb{R}, \gamma\geqslant 0\}$

$$\sup_{j\geqslant j_0}\sup_{n\geqslant n_0}\sup_{\delta\leqslant\delta_0}\frac{N_n(\delta,2^j\delta,\mathcal{G}_R(\eta),g_0)}{2^{3j}\log n} \leqslant A<\infty. \tag{6.47}$$

To see this, let $g_{\alpha,A_\gamma} \in \mathcal{G}$ with $|\alpha| \leqslant 2^j \delta / H_n^{1/2}(\gamma)$. Then $\|g_{\alpha,\gamma} - g_0\|_n = |\alpha|^2 H_n(\gamma) \leqslant 2^j \delta$. Define $g_i = \alpha_i 1_{(-\infty,\gamma_i]}$, where

$$\gamma_i = n^{-1}(1-2^{-2j})^{-i}, \quad i = \left[\frac{\log(nH_n(\gamma))}{\log(1-2^{-2j})^{-1}}\right]$$

$$\alpha_i = \frac{\delta}{\sqrt{\gamma_i}} k_i$$

and

$$k_i = \left[\frac{\alpha\sqrt{\gamma_i}}{\delta}\right].$$

Then $(1-2^{-2j}) \leqslant \gamma_i / H_n(\gamma) \leqslant 1$. Furthermore

$$0 \geqslant \alpha_i - \alpha \geqslant \left[\frac{\alpha\sqrt{\gamma_i}}{\delta} - 1\right]\frac{\delta}{\sqrt{\gamma_i}} - \alpha = -\frac{\delta}{\sqrt{\gamma_i}}.$$

It follows that

$$\begin{aligned}\|g_{\alpha,A_\gamma} - g_i\|_n &= (\alpha-\alpha_i)^2\gamma_i + \alpha^2(H_n(\gamma)-\gamma_i) \\ &\leqslant \delta^2 + 2^{2j}\delta^2(1-\gamma_i/H_n(\gamma)) \leqslant 2\delta^2.\end{aligned}$$

We have

$$0 \leqslant k_i = \left[\frac{\alpha\sqrt{\gamma_i}}{\delta}\right] \leqslant \left[\frac{2^j\delta}{\delta}\sqrt{\frac{\gamma_i}{H_n(\gamma)}}\right] \leqslant 2^j.$$

The number of functions of the form

$$k\frac{\delta}{\sqrt{\gamma_i}} 1_{(-\infty,\gamma_i]}$$

with $k \in \mathbb{Z}$, $0 \leqslant k \leqslant 2^j$, and with

$$\gamma_i = n^{-1}(1-2^{-2j})^{-i}, \quad i \in \mathbb{Z}, \quad 0 \leqslant i \leqslant \left[\frac{\log n}{\log(1-2^{-2j})}\right]$$

is equal to

$$(2^j+1)\left[\frac{\log n}{\log(1-2^{-2j})}\right] \leqslant a\,(\log n)2^{3j}.$$

It follows that if $\mathbb{E}(\exp(\beta|\epsilon_1|^2)) < \infty$ for some $\beta > 0$, then

$$\|\hat{\mathbf{g}}_n - g_0\|_n = \mathcal{O}_{\mathbf{P}}(n^{-1/2}(\log\log n)^{1/2})$$

(see Corollary 6.2.6). In fact, DARLING and ERDÖS (1956) prove that if $\mathbb{E}|\epsilon_1|^3 < \infty$, then

$$\limsup_{n\to\infty} \frac{\|\hat{\mathbf{g}}_n - g_0\|_n}{n^{-1/2}(\log\log n)^{1/2}} \leqslant \sqrt{2} \quad \text{almost surely}$$

and

$$\lim_{n\to\infty} \mathbb{P}\left[\|\hat{\mathbf{g}}_n - g_0\|_n \leqslant \frac{a + 2\log\log n + \frac{1}{2}\log\log\log n - \frac{1}{2}\log\pi}{n^{1/2}(2\log\log n)^{1/2}}\right]$$

$$= \exp(-2e^{-a}), \quad -\infty < a < \infty$$

(see also Chapter 7.4 for related results).

REMARK. In the continuous model without identification at g_0, the same rates as in Example 6.8 can occur, i.e. the continuity restriction cannot prevent $n^{1/2}\|\hat{\mathbf{g}}_n - g_0\|_n$ from exploding.

We conclude that the application of the theory of Section 6.2 to two-phase regression problems can lead to some extent to more refined results than the ones obtained by the direct methods of Chapter 5. It shows that the continuous model - with identification at g_0 - is of finite metric dimension, whereas for $d>1$ the discontinuous model can be infinite-dimensional. Example 6.5 illustrates this. However, Proposition 6.4.4 reveals a major shortcoming: the rate for $\hat{\mathbf{g}}_n$ does not always determine the rate for the $\hat{\boldsymbol{\theta}}_n^{(i)}$. Since Section 6.2 concentrates on rates for $\hat{\mathbf{g}}_n$, the techniques there cannot produce possible faster rates for the $\hat{\boldsymbol{\theta}}_n^{(i)}$.

In Examples 6.6 and 6.7, where $d=1$, the models are again finite-dimensional. These examples only differ as regards the assumptions on g_0. In Example 6.6 the rate $\mathcal{O}_{\mathbb{P}}(n^{-1/2})$ for $\hat{\mathbf{g}}_n$ implies that $|\hat{\gamma}_n - \gamma_0| = \mathcal{O}_{\mathbb{P}}(n^{-1})$, whereas in Example 6.7 we have that $\|\hat{\mathbf{g}}_n - g_0\|_n = \mathcal{O}_{\mathbb{P}}(n^{-1/2})$ leads to $|\hat{\gamma}_n - \gamma_0| = \mathcal{O}_{\mathbb{P}}(n^{-1/4})$. If in Example 6.7 the continuity of g_0 were known and a continuity restriction were super imposed on the estimated model, then the rate for $\hat{\gamma}_n$ would of course have been $\mathcal{O}_{\mathbb{P}}(n^{-1/2})$. It is important to note that in Chapter 5 we could not handle the model of Examples 6.6 and 6.7 without restricting g_0 to satisfy the discontinuity assumption (5.22). Example 6.7 now treats a situation where (5.22) (or rather its counterpart for the non- i.i.d. case) is violated.

Given the rate of convergence, the asymptotic distributions in Examples 6.6 and 6.7 are relatively easy to find. We remark that in e.g. LECAM (1970), the rate $\mathcal{O}_{\mathbb{P}}(n^{-1/2})$ for the Euclidean parameters indexing a parametric model is taken as a starting point. Then asymptotic normality can be proved without assuming the existence of first and second derivatives almost everywhere: essentially only differentiability in quadratic mean is required. The continuous model of Section 5.2 can be viewed in this light, since there the estimator of θ indexing $g_{\theta,c}$ converges with $\mathcal{O}_{\mathbb{P}}(n^{-1/2})$-rate and it can be shown that $g_{\theta,c}$ is differentiable in quadratic mean $\|\cdot\|$ at θ_0. Also for other non-linear regression models, it may be convenient to prove the $\mathcal{O}_{\mathbb{P}}(n^{-1/2})$-rate for the Euclidean parameters first, using the results of Section 6.2 (more specifically, Lemma 6.2.3), and then establishing asymptotic normality given this rate.

If the rate for $\hat{\mathbf{g}}_n$ is $\mathcal{O}_{\mathbb{P}}(n^{-1/2})$ but the rate for some of the Euclidean parameters indexing g differs from $\mathcal{O}_{\mathbb{P}}(n^{-1/2})$, then ad hoc methods are necessary in

order to obtain asymptotic distributions. Yet, Examples 6.6 and 6.7 suggest that they can again be found more easily, once the rates have already been established. Observe that the limiting distributions of the $\hat{\boldsymbol{\theta}}_n^{(i)}$ that we have encountered so far were always of the same kind, i.e. $\hat{\boldsymbol{\theta}}_n^{(1)}$ and $\hat{\boldsymbol{\theta}}_n^{(2)}$ asymptotically independent and $\sqrt{n}(\hat{\boldsymbol{\theta}}_n^{(i)} - \boldsymbol{\theta}_0^{(i)})$ converges to a normal law with covariance matrix $\|\epsilon\|^2 \Sigma^{-1}(A_0^{(i)})$, $i = 1,2$.

In Example 6.8, the model is again as in Examples 6.6 and 6.7, but g_0 is now assumed to be a one-phase function. The example shows that the rate for $\hat{\mathbf{g}}_n$ can depend on g_0. It illustrates the merit of concentrating on $\hat{\mathbf{g}}_n$ instead of $\boldsymbol{\theta}_n$ and $\hat{\mathbf{A}}_n$: the latter are not identifiable at g_0. We already elaborated on this in Section 3.4. However, even though we did not assume identifiability of all $\boldsymbol{\theta}^{(i)}$, we did need condition (3.41), which can be seen as an identifiability condition on A. Example 6.8 now suggests that if (3.41) is not imposed, then techniques that g_0 beyond uniform laws of large numbers are needed to prove consistency of $\hat{\mathbf{g}}_n$. To find the limiting distribution of $\hat{\mathbf{g}}_n$ in this example, we used the fact that the expression for $\|\hat{\mathbf{g}}_n - g_0\|_n$ coincides with the maximum of the absolute value of weighted partial sums. The question arises whether in general the knowledge of the rate of convergence - possibly slower than $\mathcal{O}_{\mathbb{P}}(n^{-1/2})$- for $\hat{\mathbf{g}}_n$ can substantially facilitate the investigation of its asymptotic distributional behaviour.

7. Tests for a change-point

7.1 Introduction

Example (1.1) deals with the change-point model

$$\mathbf{y}_k = \begin{cases} \lambda^{(1)} + \epsilon_k \ , \ k = 1, \cdots, \tau \\ \lambda^{(2)} + \epsilon_k \ , \ k = \tau + 1, \cdots, n \end{cases} \ , \ \lambda^{(i)} \in \mathbb{R}, \ i = 1,2.$$

In Section 6.4, Examples 6.6 and 6.8, entropy considerations led to the conclusion that if there is no a priori knowledge about $\lambda^{(i)}$, $i = 1,2$ or τ then

$$\|\hat{g}_n - g_0\|_n = \mathcal{O}_{\mathbb{P}}(n^{-1/2}), \quad \text{if} \quad \lambda_0^{(1)} \neq \lambda_0^{(2)}$$

whereas

$$\|\hat{g}_n - g_0\|_n = \mathcal{O}_{\mathbb{P}}(n^{-1/2}(\text{loglog}n)^{1/2}), \quad \text{if} \quad \lambda_0^{(1)} = \lambda_0^{(2)},$$

provided that the proper moment conditions on ϵ_k hold.

In this Chapter, we shall study the model where $\mathbf{y}_1, \cdots, \mathbf{y}_n$ are independent random variables, $\mathbf{y}_1, \cdots, \mathbf{y}_\tau$ having distribution $F_{\lambda^{(1)}}$ and $\mathbf{y}_{\tau+1}, \cdots, \mathbf{y}_n$ having distribution $F_{\lambda^{(2)}}$. $\{F_\lambda : \lambda \in \Lambda\}$ is a set of probability measures, with probability densities f_λ with respect to some σ-finite measure μ. We are interested in the testing problem $H_0 : \lambda^{(1)} = \lambda^{(2)}$ against $H_1 : \lambda^{(1)} \neq \lambda^{(2)}$.

The (log)likelihood ratio test statistic is

$$\mathbf{T}_n = \max_{1 \leq \tau \leq n-1} \bar{\mathbf{T}}_n(\frac{\tau}{n}),$$

where

$$\bar{\mathbf{T}}_n(\frac{\tau}{n}) = \inf_{\lambda \in \Lambda} \left[\sup_{\lambda^{(1)} \in \Lambda} 2\log \left(\frac{\prod_{k=1}^{\tau} f_{\lambda^{(1)}}(\mathbf{y}_k)}{\prod_{k=1}^{\tau} f_\lambda(\mathbf{y}_k)} \right) + \sup_{\lambda^{(2)} \in \Lambda} 2\log \left(\frac{\prod_{k=\tau+1}^{n} f_{\lambda^{(2)}}(\mathbf{y}_k)}{\prod_{k=\tau+1}^{n} f_\lambda(\mathbf{y}_k)} \right) \right].$$

The rate $\mathcal{O}_{\mathbb{P}}(n^{-1/2}(\text{loglog}n)^{1/2})$ that we encountered in Example 6.8 suggests that under H_0, $\mathbf{T}_n = \mathcal{O}_{\mathbb{P}}(\text{loglog}n)$. In fact, if F_λ is the normal distribution with variance 1, then this is a straightforward consequence of Example 6.8. In other words, $\mathbf{T}_n$ behaves in a non-standard way.

We shall consider two approaches for investigating the asymptotic efficiency of $\mathbf{T}_n$: efficiency in the sense of Bahadur and efficiency at local alternatives. We show in Section 7.2 that if $\{F_\lambda : \lambda \in \Lambda\}$ is e.g. a one-parameter exponential family, then $\mathbf{T}_n$ is optimal in the sense of Bahadur. Section 7.3 compares the Bahadur slope of $\mathbf{T}_n$ with the slopes of some alternative tests. We shall however also give evidence that $\mathbf{T}_n$'s optimality in Bahadur's sense is for practical purposes not very relevant. In Section 7.4 we show that if F_λ is the normal distribution or the exponential distribution, then at local alternatives $\mathbf{T}_n$ has asymptotic power equal to its asymptotic significance level. Local alternatives will be those alternatives with $|\lambda^{(1)} - \lambda^{(2)}| = \mathcal{O}(n^{-1/2})$.

Section 7.5 deals with the testing problem for a regression model with change-point. There is an obvious analogue of $\mathbf{T}_n$ in a regression model with possibly unknown error distribution. However, the theory developed in Sections 7.3 and 7.4 indicates that this analogue has too many unfavourable properties. Therefore, we shall propose several alternative test statistics, also bearing in mind that a more user-friendly test is desirable.

7.2 Bahadur efficiency of likelihood ratio tests

For a description of the concepts of Bahadur slope and efficiency, we refer to BAHADUR (1967,1971) and GROENEBOOM and OOSTERHOFF (1977). Bahadur looks at probabilities of large deviations, i.e. probabilities which are exponentially small as $n \to \infty$. We shall first review some general results.

Let $\{P_\theta : \theta \in \Theta_0 \cup \Theta_1\}$ be a set of probability measures dominated by a σ-finite measure μ. Let $p_\theta = dP_\theta / d\mu$ and let $\{\mathbf{T}_n\}$ be a sequence of test statistics, based on n i.i.d. observations from P_θ, for testing H_0: $\theta \in \Theta_0$ against H_1: $\theta \in \Theta_1$. Define for all $t > 0$

$$G_n(t) = \mathbb{P}_{H_0}(\mathbf{T}_n \geqslant t) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\mathbf{T}_n \geqslant t).$$

The sequence $\{\mathbf{T}_n\}$ has *(exact) Bahadur slope* $c(\theta)$ at $\theta \in \Theta_1$ if

$$\frac{1}{n} \log G_n(\mathbf{T}_n) \xrightarrow{\mathbb{P}_\theta} -\tfrac{1}{2} c(\theta).$$

The word 'exact' refers to the fact that one uses the exact null-distribution of $\mathbf{T}_n$, as opposed to its asymptotic null-distribution.

For the evaluation of the Bahadur slope, the following theorem is useful.

THEOREM 7.2.1. *Suppose that*

$$\frac{1}{n} \mathbf{T}_n \xrightarrow{\mathbb{P}_\theta} c(\theta), \ \theta \in \Theta_1$$

and that for all $a > 0$ in a neighbourhood of $c(\theta)$

$$\frac{1}{n} \log \mathbb{P}_{H_0}(\mathbf{T}_n \geqslant na) = -l(a),$$

where $l(a)$ is a nonnegative function, continuous at $c(\theta)$, then the Bahadur slope of $\{\mathbf{T}_n\}$ is equal to $2l(c(\theta))$.

PROOF. See BAHADUR (1967,1971). □

An upper bound for the Bahadur slope is twice the Kullback-Leibler information $J(\theta)$, defined as

$$J(\theta) = \inf_{\theta \in \Theta_0} K(\theta, \tilde{\theta}),$$

with

$$K(\theta,\tilde{\theta}) = \begin{cases} \int p_\theta \log(p_\theta / p_{\tilde{\theta}}) d\mu \text{ if } P_\theta \ll P_{\tilde{\theta}} \\ \infty \qquad\qquad\qquad \text{otherwise} \end{cases}.$$

THEOREM 7.2.2. *For each* θ

$$\mathbb{P}_\theta \left[\frac{1}{n} \log G_n(\mathbf{T}_n) \leqslant -J(\theta) - \eta \right] \to 0 \text{ for all } \eta > 0.$$

PROOF. See BAHADUR (1971). □

The following lemma is a minor modification of Corollary 5 in BAHADUR and RAGHAVACHARI (1972).

LEMMA 7.2.3. *Suppose that*

$$\lim_{n\to\infty} \mathbb{P}_\theta(\frac{1}{n}\mathbf{T}_n \leqslant 2J(\theta) - \eta) = 0 \text{ for all } \eta > 0, \tag{7.1}$$

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}_{\mathrm{H}_0}(\mathbf{T}_n \geqslant na) \leqslant -\tfrac{1}{2}a \text{ for all } a > 0, \tag{7.2}$$

then $\{\mathbf{T}_n\}$ *is optimal in the sense of Bahadur, i.e. its Bahadur slope is equal to* $2J(\theta)$.

PROOF. Let $\eta > 0$ be arbitrary. Then

$$\begin{aligned} &\limsup_{n\to\infty} \mathbb{P}_\theta(\frac{1}{n}\log G_n(\mathbf{T}_n) \geqslant -J(\theta) + \eta) \\ &\leqslant \limsup_{n\to\infty} \mathbb{P}_\theta(\frac{1}{n}\log G_n(\mathbf{T}_n) \geqslant -J(\theta) + \eta,\ \frac{1}{n}\mathbf{T}_n > 2J(\theta) - \eta) \\ &+ \limsup_{n\to\infty} \mathbb{P}_\theta(\frac{1}{n}\mathbf{T}_n \leqslant 2J(\theta) - \eta) \\ &= \limsup_{n\to\infty} \mathbb{P}_\theta(\frac{1}{n}\log G_n(\mathbf{T}_n) \geqslant -J(\theta) + \eta,\ \frac{1}{n}\mathbf{T}_n > 2J(\theta) - \eta). \end{aligned}$$

If $n^{-1}T_n > 2J(\theta) - \eta$, then

$$\frac{1}{n}\log G_n(T_n) \leqslant \frac{1}{n}\log G_n(n(2J(\theta) - \eta)),$$

and application of (7.2) with $a = 2J(\theta) - \eta$ gives that for all n sufficiently large

$$\frac{1}{n}\log G_n(n(2J(\theta) - \eta)) \leqslant -J(\theta) + \frac{3}{4}\eta.$$

Thus

$$\begin{aligned} &\limsup_{n\to\infty} \mathbb{P}_\theta(\frac{1}{n}\log G_n(\mathbf{T}_n) \geqslant -J(\theta) - \eta) \\ &\leqslant \limsup_{n\to\infty} \mathbb{P}_\theta(\frac{1}{n}\log G_n(\mathbf{T}_n) \geqslant -J(\theta) + \eta, \end{aligned}$$

$$\frac{1}{n}\log G_n(\mathbf{T}_n) \leqslant -J(\theta)+\frac{3}{4}\eta)$$
$$= 0.$$

Since according to Theorem 7.2.2 we have

$$\lim_{n\to\infty} \mathbb{P}_\theta(\frac{1}{n}\log G_n(\mathbf{T}_n) \leqslant -J(\theta)-\eta) = 0,$$

this completes the proof. □

Lemma 7.2.3 is the basic tool for proving optimality in Bahadur's sense of the statistic

$$\mathbf{T}_n = \max_{1\leqslant\tau\leqslant n-1} \bar{\mathbf{T}}_n(\frac{\tau}{n}).$$

We shall first describe the change-point model in an i.i.d. setting to enable us to use the previous results. Let $\theta=(\lambda^{(1)},\lambda^{(2)},\gamma)$, and let $(\mathbf{x}_k,\tilde{\mathbf{y}}_k)$, $i=1,\cdots,n$, be independent observations from the probability distribution

$$P_\theta(\mathbf{x}_1 \leqslant x,\ \tilde{\mathbf{y}}_1 \leqslant y) = \begin{cases} xF_{\lambda^{(1)}}(y) & \text{if } x \leqslant \gamma \\ \gamma F_{\lambda^{(1)}}(y)+(x-\gamma)F_{\lambda^{(2)}}(y) & \text{if } x > \gamma \end{cases}.$$

In the sequel, we shall assume that $\mathbf{y}_{\mathbf{r}_k}=\tilde{\mathbf{y}}_k$, where $\mathbf{r}_k$ is the rank of $\mathbf{x}_k$ in the ordered sequence $\mathbf{x}_{(1)} \leqslant \cdots \leqslant \mathbf{x}_{(n)}$. Then given $(\mathbf{x}_1,\cdots,\mathbf{x}_n)=(x_1,\cdots,x_n)$ we have that $\mathbf{y}_1,\cdots,\mathbf{y}_{\tau_n}$ are i.i.d. with distribution function $F_{\lambda^{(1)}}$ and $\mathbf{y}_{\tau_n+1},\cdots,\mathbf{y}_n$ are i.i.d. with distribution function $F_{\lambda^{(2)}}$, where $\tau_n=\tau_n(\gamma)=$ {number of $x_k \leqslant \gamma$, $1\leqslant k\leqslant n$}. We shall regard $\mathbf{T}_n$ as the unconditional likelihood.

The parameter space is

$$\Theta = \{\theta=(\lambda^{(1)},\lambda^{(2)},\gamma):\ \lambda^{(i)}\in\Lambda,\ i=1,2,\ \gamma\in(0,1)\}.$$

For $J(\theta)$, $\theta=(\lambda^{(1)},\lambda^{(2)},\gamma)$, we find the following expression:

$$J(\theta) = \inf_{\lambda\in\Lambda} \left[\gamma\int f_{\lambda^{(1)}}\log(f_{\lambda^{(1)}}/f_\lambda)d\mu+(1-\gamma)\int f_{\lambda^{(2)}}\log(f_{\lambda^{(2)}}/f_\lambda)d\mu\right].$$

Lemmas 7.2.4 and 7.2.5 below present sufficient conditions such that the assumptions (7.1) and (7.2) of Lemma 7.2.3 hold for $\{\mathbf{T}_n\}$.

LEMMA 7.2.4. *Suppose that for* $\theta=(\lambda^{(1)},\lambda^{(2)},\gamma)$,

$$\lim_{n\to\infty} \mathbb{P}_\theta(\frac{1}{n}\bar{\mathbf{T}}_n(\frac{\tau_n(\gamma)}{n}) \leqslant 2J(\theta)-\eta) = 0 \text{ for all } \eta>0. \tag{7.3}$$

Then also

$$\lim_{n\to\infty} \mathbb{P}_\theta(\mathbf{T}_n \leqslant 2J(\theta)-\eta) = 0 \text{ for all } \eta>0.$$

PROOF. This follows immediately from the fact that

$$\mathbf{T}_n = \max_{1 \leqslant k \leqslant n-1} \bar{\mathbf{T}}_n\left(\frac{k}{n}\right) \geqslant \bar{\mathbf{T}}_n\left(\frac{\tau_n(\gamma)}{n}\right). \quad \square$$

If we define

$$\mathbf{l}_n^{(1)}(\lambda,\tau) = \sup_{\lambda^{(1)} \in \Lambda} 2\log\left[\frac{\prod_{k=1}^{\tau} f_{\lambda^{(1)}}(\mathbf{y}_k)}{\prod_{k=1}^{\tau} f_{\lambda}(\mathbf{y}_k)}\right]$$

and

$$\mathbf{l}_n^{(2)}(\lambda,\tau) = \sup_{\lambda^{(2)} \in \Lambda} 2\log\left[\frac{\prod_{k=\tau+1}^{n} f_{\lambda^{(2)}}(\mathbf{y}_k)}{\prod_{k=\tau+1}^{n} f_{\lambda}(\mathbf{y}_k)}\right],$$

then

$$\mathbf{T}_n = \max_{1 \leqslant k \leqslant n-1} \inf_{\lambda \in \Lambda} [\mathbf{l}_n^{(1)}(\lambda,k) + \mathbf{l}_n^{(2)}(\lambda,k)].$$

LEMMA 7.2.5. *Suppose that for every sequence* $\{k_n\}$, $1 \leqslant k_n \leqslant n-1$, $n = 1,2,\cdots$

$$\limsup_{n\to\infty} \frac{1}{n}\log\left[\sup_{\lambda \in \Lambda} \mathbb{P}_\lambda(\mathbf{l}_n^{(i)}(\lambda,k_n) \geqslant na)\right] \leqslant -\tfrac{1}{2}a,\ a > 0,\ i = 1,2. \quad (7.4)$$

Then also

$$\limsup_{n\to\infty} \frac{1}{n}\log \mathbb{P}_{\mathrm{H}_0}(\mathbf{T}_n \geqslant na) \leqslant -\tfrac{1}{2}a.$$

PROOF. For each $\lambda_0 \in \Lambda$

$$\begin{aligned}\bar{\mathbf{T}}_n\left(\frac{k_n}{n}\right) &= \inf_{\lambda \in \Lambda} [\mathbf{l}_n^{(1)}(\lambda,k_n) + \mathbf{l}_n^{(2)}(\lambda,k_n)] \\ &\leqslant \mathbf{l}_n^{(1)}(\lambda_0,k_n) + \mathbf{l}_n^{(2)}(\lambda_0,k_n).\end{aligned}$$

Hence

$$\mathbb{P}_{\mathrm{H}_0}\left(\bar{\mathbf{T}}_n\left(\frac{k_n}{n}\right) \geqslant na\right) \leqslant \sup_{\lambda_0 \in \Lambda} \mathbb{P}_{\lambda_0}(\mathbf{l}_n^{(1)}(\lambda_0,k_n) + \mathbf{l}_n^{(2)}(\lambda_0,k_n) \geqslant na). \quad (7.5)$$

Let $\eta > 0$ be arbitrary. Then for all $\lambda_0 \in \Lambda$

$$\begin{aligned}&\mathbb{P}_{\lambda_0}(\mathbf{l}_n^{(1)}(\lambda_0,k_n) + \mathbf{l}_n^{(2)}(\lambda_0,k_n) \geqslant na) \\ &\leqslant \sum_{i=0}^{[a/\eta]} \mathbb{P}_{\lambda_0}(\mathbf{l}_n^{(1)}(\lambda_0,k_n) \in [ni\eta, n(i+1)\eta),\ \mathbf{l}_n^{(2)}(\lambda_0,k_n) \geqslant na - n(i+1)\eta)\end{aligned}$$

$$+ \mathbb{P}_{\lambda_0}(\mathbf{l}_n^{(2)}(\lambda_0,k_n)\geqslant na)$$

$$\leqslant \sum_{i=0}^{[a/\eta]} \mathbb{P}_{\lambda_0}(\mathbf{l}_n^{(1)}(\lambda_0,k_n)\geqslant ni\eta)\mathbb{P}_{\lambda_0}(\mathbf{l}_n^{(2)}(\lambda_0,k_n)\geqslant na-n(i+1)\eta)$$

$$+ \mathbb{P}_{\lambda_0}(\mathbf{l}_n^{(2)}(\lambda_0,k_n)\geqslant na).$$

From (7.4) we know that for arbitrary $\delta>0$ and for n sufficiently large

$$\sup_{\lambda_0\in\Lambda} \mathbb{P}_{\lambda_0}(\mathbf{l}_n^{(1)}(\lambda_0,k_n)\geqslant ni\eta) \leqslant \exp(-n(\frac{i\eta}{2}-\delta))$$

and

$$\sup_{\lambda_0\in\Lambda} \mathbb{P}_{\lambda_0}(\mathbf{l}_n^{(2)}(\lambda_0,k_n)\geqslant na-n(i+1)\eta) \leqslant \exp(-n(\frac{a}{2}-\frac{(i+1)\eta}{2}-\delta)),$$

which implies

$$\sup_{\lambda_0\in\Lambda} \mathbb{P}_{\lambda_0}(\mathbf{l}_n^{(1)}(\lambda_0,k_n)+\mathbf{l}_n^{(2)}(\lambda_0,k_n)\geqslant na)$$

$$\leqslant \sum_{i=0}^{[a/\eta]} \exp(-n(\frac{a}{2}-\frac{\eta}{2}-2\delta))+\exp(-n(\frac{a}{2}-\delta))$$

$$\leqslant \left[([\frac{a}{\eta}]+1)e^{n\eta/2}+1\right]\exp(-n(\frac{a}{2}-2\delta)).$$

Since η and δ are arbitrary, this implies

$$\limsup_{n\to\infty} \sup_{\lambda_0\in\Lambda} \frac{1}{n}\log\mathbb{P}_{\lambda_0}(\mathbf{l}_n^{(1)}(\lambda_0,k_n)+\mathbf{l}_n^{(2)}(\lambda_0,k_n)\geqslant na) \leqslant -\tfrac{1}{2}a.$$

From (7.5) it follows that also

$$\limsup_{n\to\infty} \log\mathbb{P}_{H_0}(\bar{\mathbf{T}}_n(\frac{k_n}{n})\geqslant na) \leqslant -\tfrac{1}{2}a.$$

And since this is true for all sequences $\{k_n\}$, also

$$\limsup_{n\to\infty}\frac{1}{n}\mathbb{P}_{H_0}(\mathbf{T}_n\geqslant na)$$

$$\leqslant \limsup_{n\to\infty} \frac{1}{n}\log\left\{n \max_{1\leqslant k\leqslant n} \mathbb{P}_{H_0}(\bar{\mathbf{T}}_n(\frac{k}{n})\geqslant na)\right\}$$

$$\leqslant \limsup_{n\to\infty} \frac{1}{n}\log n - \tfrac{1}{2}a = -\tfrac{1}{2}a. \quad \square$$

Now, KALLENBERG (1978) shows that (7.3) holds for $\{F_\lambda: \lambda\in\Lambda\}$ an exponential family in standard representation and $\lambda^{(i)}$, $i=1,2$, in the interior of parameter space. Moreover, he proves that (7.4) also holds if $\{F_\lambda: \lambda\in\Lambda\}$ is a one-parameter exponential family. Thus, we arrive at the following theorem.

THEOREM 7.2.6. *For $\{F_\lambda: \lambda\in\Lambda\}$ a one-parameter exponential family in standard*

representation, $\{\mathbf{T}_n\}$ *is optimal in the sense of Bahadur at all alternatives* $\theta=(\lambda^{(1)},\lambda^{(2)},\gamma)$, $\lambda^{(i)}$, $i=1,2$, *in the interior of* Λ, $\gamma\in(0,1)$. □

Note that for k-parameter exponential families ($k>1$), Bahadur-optimality of $\{\mathbf{T}_n\}$ follows if (7.4) holds.

Related results have been obtained by DESHAYES and PICARD (1982). They consider the normal distribution and derive large deviations results both at H_0 and H_1.

7.3. Bahadur efficiency in the normal and exponential case

Examples of one-parameter exponential families are the normal distribution with known variance and the exponential distribution. We shall treat these in some more detail. In Subsection 7.3.1 we compute the slopes for $\mathbf{T}_n$ and some alternative tests that are easier to use in practice. Furthermore, the fact that these alternative tests are $\mathcal{O}_{\mathbb{P}}(1)$ under H_0 might also be considered as a theoretical advantage. To explain why, we actually need the results of Section 7.4, which imply that the alternative tests always behave better than $\mathbf{T}_n$ at local alternatives.

Subsection 7.3.2 presents a test statistic which is asymptotically equivalent to $\mathbf{T}_n$ under H_0, but which has Bahadur slope zero.

7.3.1. The normal case

For $F_\lambda=\Phi(\cdot-\lambda)$, $\lambda\in\mathbb{R}$, Φ the standard normal distribution, we have

$$\mathbf{T}_n = \max_{1\leq k\leq n-1} \bar{\mathbf{T}}_n\left(\frac{k}{n}\right),$$

with $\bar{\mathbf{T}}_n(k/n)=\bar{\mathbf{t}}_n^2(k/n)$,

$$\bar{\mathbf{t}}_n\left(\frac{k}{n}\right) = \sqrt{\frac{k(n-k)}{n}}\left[\frac{1}{k}\sum_{i=1}^{k}\mathbf{y}_i-\frac{1}{n-k}\sum_{i=k+1}^{n}\mathbf{y}_i\right]. \tag{7.6}$$

The exact null-distribution of $\mathbf{T}_n$ is quite cumbersome and it turns out that the limiting null-distibution of the appropriately normalized $\mathbf{T}_n$ is not a good approximation for finite sample sizes.

We propose statistics of the form

$$\mathbf{T}_{n,\psi} = \max_{1\leq k\leq n-1} \psi\left(\frac{k}{n}\right)\bar{\mathbf{T}}_n\left(\frac{k}{n}\right),$$

where $\psi(.)$ is a function that diminishes the weights in the tails. For practical purposes it is convenient to take $\psi(s)=s(1-s)$, because then the approximate significance level can be found in standard tables: under H_0

$$\mathbf{T}_{n,\psi} \xrightarrow{\mathcal{L}} \sup_{0<s<1} \mathbf{B}^2(s), \text{ for } \psi(s)=s(1-s),$$

where $\mathbf{B}(.)$ is a standard Brownian bridge.

Other relatively easy to use tests statistics are

$$|\mathbf{t}_n^S| = |\sum_{k=1}^{n-1} \mathbf{t}_n(\frac{k}{n})|$$

and more generally

$$|\mathbf{t}_{n,\psi}^S| = |\sum_{k=1}^{n-1} \psi^{1/2}(\frac{k}{n})\bar{\mathbf{t}}_n(\frac{k}{n})|.$$

The superscript 'S' refers to *sum-statistic*, as in PRAAGMAN (1986). Under H_0

$$|\mathbf{t}_n^S| \xrightarrow{\mathcal{L}} |\mathcal{N}(0,(\frac{1}{4}\pi^2-2))|$$

and

$$|\mathbf{t}_{n,\psi}^S| \xrightarrow{\mathcal{L}} |\mathcal{N}(0,\frac{1}{12})|, \text{ for } \psi(s)=s(1-s).$$

Let $c(\mathbf{T}_\psi,\theta)$ and $c(|\mathbf{t}_\psi^S|,\theta)$ denote the Bahadur slope at $\theta=(\lambda^{(1)},\lambda^{(2)},\gamma)$ of $\{\mathbf{T}_{n,\psi}\}$ and $\{|\mathbf{t}_{n,\psi}^S|\}$ respectively.

LEMMA 7.3.1. *If* $F_\lambda=\Phi(\cdot-\lambda)$,

$$c(\mathbf{T},\theta) = \gamma(1-\gamma)(\lambda^{(1)}-\lambda^{(2)})^2,$$

$$c(|\mathbf{t}^S|,\theta) = \frac{(\frac{1}{2}\pi-\sqrt{\gamma(1-\gamma)}-(1-\gamma)\arcsin\sqrt{1-\gamma}-\gamma\arcsin\sqrt{\gamma})^2}{\frac{1}{4}\pi^2-2}(\lambda^{(1)}-\lambda^{(2)})^2$$

and for $\psi(s)=s(1-s)$

$$c(\mathbf{T}_\psi,\theta) = 4\gamma^2(1-\gamma)^2(\lambda^{(1)}-\lambda^{(2)})^2,$$

$$c(|\mathbf{t}_\psi^S|,\theta) = 3\gamma^2(1-\gamma)^2(\lambda^{(1)}-\lambda^{(2)})^2.$$

PROOF. The Kullback-Leibler information number is

$$J(\theta) = \tfrac{1}{2}\gamma(1-\gamma)(\lambda^{(1)}-\lambda^{(2)})^2.$$

Hence $c(\mathbf{T},\theta)=\gamma(1-\gamma)(\lambda^{(1)}-\lambda^{(2)})^2$.

We apply Theorem 7.2.1 to calculate the slopes of the other statistics. It is easy to see that for a sequence of normally distributed random variables $\mathbf{N}_n$, with expectation zero and variance $\sigma_n^2\to\sigma^2$

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}(|\mathbf{N}_n|\geq n^{1/2}a) = -\tfrac{1}{2}\frac{a^2}{\sigma^2}.$$

Straightforward calculation now gives

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}_{H_0}(|\mathbf{t}_n^S|\geq n^{1/2}a) = -\tfrac{1}{2}\frac{a^2}{\frac{1}{4}\pi^2-2},$$

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}_{H_0}(\mathbf{T}_{n,\psi}\geqslant(n^{1/2}a)^2) = -\tfrac{1}{2}\inf_{0<s<1}\frac{a^2}{\psi(s)} = -2a^2,$$

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}_{H_0}(|\mathbf{t}^S_{n,\psi}|\geqslant n^{1/2}a) = -\tfrac{1}{2}\frac{a^2}{\frac{1}{12}} = -6a^2.$$

Moreover

$$n^{-1/2}\bar{\mathbf{t}}_n(\frac{[ns]}{n}) \xrightarrow{\mathbb{P}_\theta} \begin{cases}(s/(1-s))^{1/2}(1-\gamma)|\lambda^{(1)}-\lambda^{(2)}| & \text{if } s\leqslant\gamma\\ ((1-s)/s)^{1/2}\gamma|\lambda^{(1)}-\lambda^{(2)}| & \text{if } s\geqslant\gamma\end{cases}$$

uniformly in $s\in(0,1)$. Thus

$$n^{-1/2}|\mathbf{t}^S_n| \xrightarrow{\mathbb{P}_\theta} \left[(1-\gamma)\int_0^\gamma\sqrt{\frac{s}{1-s}}ds+\gamma\int_\gamma^1\sqrt{\frac{1-s}{s}}ds\right]|\lambda^{(1)}-\lambda^{(2)}|$$

$$= (\tfrac{1}{2}\pi-\sqrt{\gamma(1-\gamma)}-(1-\gamma)\arcsin\sqrt{1-\gamma}-\gamma\arcsin\sqrt{\gamma})|\lambda^{(1)}-\lambda^{(2)}|,$$

$$n^{-1}\mathbf{T}_{n,\psi} \xrightarrow{\mathbb{P}_\theta} \psi(\gamma)\gamma(1-\gamma)(\lambda^{(1)}-\lambda^{(2)})^2 = \gamma^2(1-\gamma)^2(\lambda^{(1)}-\lambda^{(2)})^2$$

and

$$n^{-1/2}|\mathbf{t}^S_{n,\psi}| \xrightarrow{\mathbb{P}_\theta} \left[(1-\gamma)\int_0^\gamma sds+\gamma\int_\gamma^1(1-s)ds\right]|\lambda^{(1)}-\lambda^{(2)}|$$

$$= \tfrac{1}{2}\gamma(1-\gamma)|\lambda^{(1)}-\lambda^{(2)}|. \quad \square$$

As is to be expected, the loss of Bahadur efficiency for the alternative tests is always the most substantial for values of γ near 0 or 1.

7.3.2. The exponential case

Suppose $F_\lambda(y)=1-\exp(-\lambda y)$, $\lambda>0$, $y\geqslant 0$. Then

$$\mathbf{T}_n = \max_{1\leqslant k\leqslant n}\bar{\mathbf{T}}_n(\frac{k}{n}),$$

with

$$\bar{\mathbf{T}}_n(\frac{k}{n}) = -2k\log\left[\frac{\beta_n(\mathbf{Y}_n,k)}{k/n}\right]-2(n-k)\log\left[\frac{1-\beta_n(\mathbf{Y}_n,k)}{1-k/n}\right], \quad (7.7)$$

$$\beta_n(\mathbf{Y}_n,k) = \frac{\sum_{i=1}^k\mathbf{y}_i}{\sum_{i=1}^n\mathbf{y}_i},\ k=1,\cdots,n.$$

At $\theta=(\lambda^{(1)},\lambda^{(2)},\gamma)$ we have

$$J(\theta) = \log(\frac{\gamma}{\lambda^{(1)}}+\frac{1-\gamma}{\lambda^{(2)}})-\gamma\log\frac{1}{\lambda^{(1)}}-(1-\gamma)\log\frac{1}{\lambda^{(2)}}$$

and the Bahadur slope of $\{\mathbf{T}_n\}$ is $2J(\theta)$.

The second order Taylor expansion of the right-hand side of (7.7) at $\beta_n(\mathbf{Y}_n,k)=k/n$ is equal to

$$\bar{\mathbf{T}}_n^*(\frac{k}{n}) = n\frac{(\beta_n(\mathbf{Y}_n,k)-\frac{k}{n})^2}{(k/n)(1-k/n)}.$$

Define

$$\mathbf{T}_n^* = \max_{1\leqslant k\leqslant n-1} \bar{\mathbf{T}}_n^*(\frac{k}{n}).$$

It is shown in HACCOU et al (1985) that after the appropriate normalization, $\mathbf{T}_n$ and $\mathbf{T}_n^*$ have the same limiting null-distribution. Moreover,

$$\mathbf{T}_n^* = \frac{\mathbf{T}_n^\Phi}{(\frac{1}{n}\sum_{i=1}^n \mathbf{y}_i)^2},$$

where

$$\mathbf{T}_n^\Phi = \max_{1\leqslant k\leqslant n} \frac{k(n-k)}{n}\left[\frac{1}{k}\sum_{i=1}^k \mathbf{y}_i - \frac{1}{n-k}\sum_{i=k+1}^n \mathbf{y}_i\right]^2$$

is the likelihood ratio test for the case of normally distributed random variables (see equation (7.6)).

LEMMA 7.3.2. *If* $F_\lambda(y)=1-\exp(-\lambda y)$ *then* $\mathbf{T}_n^*$ *has Bahadur slope zero.*

PROOF. It is easy to prove that $\mathbf{T}_n^*$ converges in $\mathbb{P}_\theta$-probability for each θ. Thus, it remains to show that for all $a>0$

$$\lim_{n\to\infty} \frac{1}{n}\log\mathbb{P}_{H_0}(\mathbf{T}_n^*\geqslant na) = 0.$$

Now, under H_0, $\beta_n(\mathbf{Y}_n,k)$ has the same distribution as the k-th order statistic $\mathbf{U}_n(k)$ from a sample of size $n-1$ from the uniform distribution. Hence, if we take n sufficiently large

$$\begin{aligned}
&\mathbb{P}_{H_0}(\mathbf{T}_n^*\geqslant na) \geqslant \mathbb{P}_{H_0}(\bar{\mathbf{T}}_n^*(\frac{1}{n})\geqslant na)\\
&\geqslant \mathbb{P}\left[\frac{\mathbf{U}_n(1)-\frac{1}{n}}{((1/n)(1-1/n))^{1/2}}\geqslant a^{1/2}\right]\\
&\geqslant \mathbb{P}(\mathbf{U}_n(1)\geqslant(\frac{a}{n})^{1/2}+\frac{1}{n}) \geqslant \mathbb{P}(\mathbf{U}_n(1)\geqslant 2(\frac{a}{n})^{1/2})\\
&= \left[1-2(\frac{a}{n})^{1/2}\right]^{n-1}.
\end{aligned}$$

Thus,

$$\frac{1}{n}\log \mathbb{P}_{H_0}(\mathbf{T}_n^* \geqslant na) \geqslant \frac{n-1}{n}\log\left[1-2\left(\frac{a}{n}\right)^{1/2}\right] \to 0. \qquad \square$$

In the same way it can be shown that $\mathbf{T}_n^{\Phi}$ also has Bahadur slope zero. In Section 7.5, we shall introduce the residuals-test, based on the least squares estimators of the parameters in a two-phase regression model. The residuals-test has the same appearance as the likelihood ratio test for normally distributed random variables. The result of this subsection therefore indicates that from the point of view of Bahadur efficiency it is not sensible to use the residuals test. Furthermore, the following section implies that also its Pitman efficiency is zero.

7.4. Efficiency of the likelihood ratio test at local alternatives

We study the behaviour of $\mathbf{T}_n$ at alternatives $\theta_n = (\lambda_n^{(1)}, \lambda_n^{(2)}, \tau_n / n)$ for which the following holds: for some $\{\lambda_n\}$,

$$\begin{aligned} |\lambda_n^{(1)} - \lambda_n| &= \mathcal{O}(\tau_n^{-1/2}), \\ |\lambda_n^{(2)} - \lambda_n| &= \mathcal{O}((n-\tau_n)^{-1/2}). \end{aligned} \tag{7.8}$$

Again, we shall only consider the normal case with known variance and the exponential case. Then, condition (7.8) defines exactly the alternatives which are contiguous to the null-hypothesis and it is equivalent to the condition that the Hellinger distance between $(F_{\lambda_n^{(1)}})^{\tau_n}(F_{\lambda_n^{(2)}})^{n-\tau_n}$ and $(F_{\lambda_n})^n$ remains bounded (see e.g. OOSTERHOFF and VAN ZWET (1975)). We shall only study the situation where $\eta < (\tau_n / n) < 1-\eta$ for some $\eta > 0$ and for all n sufficiently large. Then we can assume without loss of generality that $\tau_n / n \to \gamma \in (0,1)$ and (7.8) reduces to

$$|\lambda_n^{(1)} - \lambda_n^{(2)}| = \mathcal{O}(n^{-1/2}).$$

7.4.1. The normal case. Let $F_\lambda = \Phi(\cdot - \lambda)$. The limiting null-distribution of $\mathbf{T}_n$ is given in Lemma 7.4.1.1 below. Since $\mathbf{T}_n = \mathcal{O}_{\mathbb{P}}(\log\log n)$ under H_0 we need to renormalize it. Define for $0 < \eta_n < 1-\delta_n < 1$,

$$\rho(\eta_n, \delta_n) = \tfrac{1}{2}\log\left[\frac{(1-\eta_n)(1-\delta_n)}{\eta_n \delta_n}\right].$$

Furthermore, write

$$\begin{aligned} b(x) &= 2\log x + \tfrac{1}{2}\log\log x - \tfrac{1}{2}\log\pi, \\ a(x) &= 2(\log x)^{1/2}, \\ b_n &= b(\log n), \\ a_n &= a(\log n). \end{aligned} \tag{7.9}$$

Let $\bar{\mathbf{t}}_n(k/n)$, $k=1,\cdots,n-1$ be defined as in (7.6).

LEMMA 7.4.1.1.

$$\lim_{n\to\infty} \mathbb{P}_{H_0}\left[\max_{\eta_n<k/n<1-\delta_n} |\bar{\mathbf{t}}_n(\frac{k}{n})| \leqslant \frac{s+b(\rho(\eta_n,\delta_n))}{a(\rho(\eta_n,\delta_n))}\right] = \exp(-2e^{-s}),$$

$$-\infty<s<\infty.$$

PROOF. A minor extension of Corollary 1.9.1, page 57 in CSÖRGÖ and RÉVÉSZ (1981) says that for $\mathbf{B}(x)$ a standard Brownian bridge

$$\lim_{n\to\infty} \mathbb{P}\left[\sup_{\eta_n<x<\delta_n} \frac{|\mathbf{B}(x)|}{\sqrt{x(1-x)}} \leqslant \frac{s+b(\rho(\eta_n,\delta_n))}{a(\rho(\eta_n\delta_n))}\right] = \exp(-2e^{-s}),$$

$$-\infty<s<\infty.$$

Now, $\bar{\mathbf{t}}_n(k/n)$, $k=1,\cdots,n-1$, is under H_0 in distribution equal to

$$\frac{\mathbf{B}(\frac{k}{n})}{\sqrt{\frac{k}{n}(1-\frac{k}{n})}},\ k=1,\cdots,n-1.$$

The increments of $\mathbf{B}$ satisfy

$$\lim_{n\to\infty} \sup_{0\leqslant u\leqslant 1-1/n} \sup_{0<x<1/n} \frac{|\mathbf{B}(u+x)-\mathbf{B}(u)|}{\sqrt{2(\log n)/n}} = 1 \tag{7.10}$$

almost surely (CSÖRGÖ and RÉVÉSZ (1981), Theorem 1.4.1, page 42). For simplicity, we only consider the interval $(0,\frac{1}{2}]$. Take $\tilde{\eta}_n = a(\rho(\eta_n,\delta_n))(\log n)^2/n$, then

$$a(\rho(\eta_n,\delta_n))\left[\max_{\tilde{\eta}_n\leqslant k/n\leqslant \frac{1}{2}} \sup_{k/n<x\leqslant(k+1)/n} \left|\frac{\mathbf{B}(x)}{\sqrt{x(1-x)}} - \frac{\mathbf{B}(k/n)}{\sqrt{(k/n)(1-k/n)}}\right|\right] \to 0,$$

almost surely, in view of (7.10). On the remaining subinterval $(\eta_n,\tilde{\eta}_n)$ we have

$$\mathbb{P}\left[\max_{\eta_n<k/n<\tilde{\eta}_n} \frac{|\mathbf{B}(k/n)|}{\sqrt{(k/n)(1-k/n)}} \geqslant \frac{s+b(\rho(\eta_n,\delta_n))}{a(\rho(\eta_n,\delta_n))}\right]$$

$$\leqslant \mathbb{P}\left[\sup_{\eta_n<x<\tilde{\eta}_n} \frac{|\mathbf{B}(x)|}{\sqrt{x(1-x)}} \geqslant \frac{s+b(\rho(\eta_n,\delta_n))}{a(\rho(\eta_n\delta_n))}\right] \to 0,$$

since $\rho(\eta_n,1-\tilde{\eta}_n)=o(a(\rho(\eta_n,\delta_n))$. □

It follows that

$$\mathbb{P}_{H_0}\left[\mathbf{T}_n \leqslant (\frac{s+b_n}{a_n})^2\right] \to \exp(-2e^{-s}),\ -\infty<s<\infty.$$

We can also use Lemma 7.4.1.1 to draw conclusions about the behaviour of the maximum likelihood estimator of the change-point. Let $\hat{\tau}_n$ be defined by

$$\hat{\tau}_n = \arg \max_{1 \leqslant k \leqslant n-1} \bar{\mathbf{T}}_n(\frac{k}{n}).$$

It follows from Lemma 7.4.1.2 below that $\hat{\tau}_n / n \to 0$ or 1 in $\mathbb{P}_{H_0}$-probability, so under H_0 $\hat{\tau}_n / n$ is in a sense a consistent estimator of the change-point. However, at contiguous alternatives θ_n also $\hat{\tau}_n / n \to 0$ or 1 in $\mathbb{P}_{\theta_n}$-probability, so in general $\hat{\tau}_n$ is inconsistent.

LEMMA 7.4.1.2.

$$\lim_{n\to\infty} \mathbb{P}_{H_0}(\hat{\tau}_n \leqslant \frac{n}{\log n} \text{ or } \hat{\tau}_n \geqslant n - \frac{n}{\log n})$$
$$= \lim_{n\to\infty} \mathbb{P}_{\theta_n}(\hat{\tau}_n \leqslant \frac{n}{\log n} \text{ or } \hat{\tau}_n \geqslant n - \frac{n}{\log n}) = 1,$$

for all contiguous alternatives $\theta_n = (\lambda_n^{(1)}, \lambda_n^{(2)}, \tau_n / n)$.

PROOF. We have that

$$\mathbb{P}_{H_0}\left(\max_{\frac{n}{\log n} < k < n - \frac{n}{\log n}} |\bar{\mathbf{t}}_n(\frac{k}{n})| > \frac{s + b_n}{a_n} \right)$$
$$= \mathbb{P}_{H_0}\left(\max_{\frac{n}{\log n} < k < n - \frac{n}{\log n}} |\bar{\mathbf{t}}_n(\frac{k}{n})| > \frac{s(\rho_n) + b(\rho_n)}{a(\rho_n)} \right),$$

where $s(\rho_n) = (a(\rho_n) / a_n)(s + b_n) - b(\rho_n)$ and $\rho_n = \log[\log n(1 - 1 / \log n)]$. Since $s(\rho_n) \to \infty$ as $n \to \infty$, application of Lemma 7.4.1.1 now implies that under H_0, $\hat{\tau}_n \leqslant n / \log n$ or $\hat{\tau}_n \geqslant n - n / \log n$ with probability tending to 1.

Because $(F_{\lambda_n^{(1)}})^{\tau_n}(F_{\lambda_n^{(2)}})^{n-\tau_n}$ is assumed to be contiguous to $(F_{\lambda_n})^n$, the same is true in $\mathbb{P}_{\theta_n}$-probability, $\theta_n = (\lambda_n^{(1)}, \lambda_n^{(2)}, \tau_n / n)$. □

Define $\mathbf{y}_{n,k}^{(0)} = \mathbf{y}_{n,k} - \mathbb{E}_{\theta_n} \mathbf{y}_{n,k}$, $k = 1, \cdots, n$, and let

$$\mathbf{T}_n^{(0)} = \max_{1 \leqslant k \leqslant n-1} |\bar{\mathbf{t}}_n^{(0)}(\frac{k}{n})|^2$$

be the likelihood ratio evaluated at $\mathbf{y}_{n,1}^{(0)}, \cdots, \mathbf{y}_{n,n}^{(0)}$. Then under $\mathbb{P}_{\theta_n}$

$$|\bar{\mathbf{t}}_n(\frac{k}{n})|^2 = \begin{cases} \left[\bar{\mathbf{t}}_n^{(0)}(\frac{k}{n}) + (\frac{k}{n-k} \frac{n-\tau_n}{\tau_n})^{1/2} C_n \right]^2 & \text{if } k \leqslant \tau_n \\ \left[\bar{\mathbf{t}}_n^{(0)}(\frac{k}{n}) + (\frac{n-k}{k} \frac{\tau_n}{n-\tau_n})^{1/2} C_n \right]^2 & \text{if } k \geqslant \tau_n \end{cases}, \tag{7.11}$$

where

$$C_n = (n-\tau_n)^{1/2}\lambda_n^{(1)} - \tau_n^{1/2}\lambda_n^{(2)}.$$

In view of (7.11), we have at contiguous alternatives with $(\tau_n / n) \to \gamma$

$$|\bar{\mathbf{t}}_n(\frac{k}{n})| \leqslant |\bar{\mathbf{t}}_n^{(0)}(\frac{k}{n})| + \mathcal{O}(\frac{1}{\log n})^{1/2}$$

uniformly in $k \leqslant n / \log n$ or $k \geqslant n - n / \log n$. Thus the extra term added to $|\bar{\mathbf{t}}_n^{(0)}(k / n)|$ is small. The consequence is that $\mathbf{T}_n$ has asymptotic power equal to its significance level at alternatives $\theta_n = (\lambda_n^{(1)}, \lambda_n^{(2)}, \gamma)$, $|\lambda_n^{(1)} - \lambda_n^{(2)}| = \mathcal{O}(n^{-1/2})$.

THEOREM 7.4.1.3.

$$\left| \mathbb{P}_{\theta_n}\left[\mathbf{T}_n > (\frac{s+b_n}{a_n})^2\right] - \mathbb{P}_{\mathrm{H}_0}\left[\mathbf{T}_n > (\frac{s+b_n}{a_n})^2\right] \right| \to 0,$$

$$-\infty < s < \infty,$$

for all contiguous alternatives $\theta_n = (\lambda_n^{(1)}, \lambda_n^{(2)}, \tau_n / n)$ *with* $(\tau_n / n) \to \gamma \in (0,1)$.

PROOF. For n sufficiently large,

$$\mathbb{P}_{\theta_n}\left[\max_{1 \leqslant k \leqslant \frac{n}{\log n} \text{ or } n - \frac{n}{\log n} \leqslant k \leqslant n-1} |\bar{\mathbf{t}}_n^{(0)}(\frac{k}{n})| > \frac{s+q_n+b_n}{a_n}\right]$$

$$\leqslant \mathbb{P}_{\theta_n}\left[\max_{1 \leqslant k \leqslant \frac{n}{\log n} \text{ or } n - \frac{n}{\log n} \leqslant k \leqslant n-1} |\bar{\mathbf{t}}_n(\frac{k}{n})| > \frac{s+b_n}{a_n}\right]$$

$$\leqslant \mathbb{P}_{\theta_n}\left[\max_{1 \leqslant k \leqslant \frac{n}{\log n} \text{ or } n - \frac{n}{\log n} \leqslant k \leqslant n-1} |\bar{\mathbf{t}}_n^{(0)}(\frac{k}{n})| > \frac{s-q_n+b_n}{a_n}\right],$$

where

$$q_n = a_n(\log n - 1)^{-1/2}\left[(\frac{n-\tau_n}{\tau_n})^{1/2} + (\frac{\tau_n}{n-\tau_n})^{1/2}\right] |C_n| \to 0.$$

The theorem now follows from:

$$\lim_{n\to\infty} \mathbb{P}_{\theta_n}\left[\max_{1 \leqslant k \leqslant \frac{n}{\log n} \text{ or } n - \frac{n}{\log n} \leqslant k \leqslant n-1} |\bar{\mathbf{t}}_n(\frac{k}{n})| > \frac{s+b_n}{a_n}\right]$$

$$= \lim_{n\to\infty} \mathbb{P}_{\theta_n}\left[\mathbf{T}_n > (\frac{s+b_n}{a_n})^2\right]$$

and

$$\lim_{n\to\infty} \mathbb{P}_{\theta_n}\left[\max_{1 \leqslant k \leqslant \frac{n}{\log n} \text{ or } n - \frac{n}{\log n} \leqslant k \leqslant n-1} |\bar{\mathbf{t}}_n^{(0)}(\frac{k}{n})| > \frac{s+\tilde{q}_n+b_n}{a_n}\right]$$

$$= \limsup_{n\to\infty} \mathbb{P}_{H_0}\left[\max_{1\leq k\leq \frac{n}{\log n} \text{ or } n-\frac{n}{\log n}\leq k\leq n-1} |\bar{\mathbf{t}}_n(\frac{k}{n})| > \frac{s+\tilde{q}_n+b_n}{a_n}\right]$$

$$= 1-\exp(-2e^{-s}),$$

for all $\tilde{q}_n\to 0$. □

In Theorem 7.4.1.3 we excluded the cases $(\tau_n/n)\to 0$ or 1. One can however also show that if (τ_n/n) converges to zero very fast (e.g. $\tau_n = o(\log n/\operatorname{loglog} n)$), then $\mathbf{T}_n$ has again asymptotic power equal to its asymptotic significance level at contiguous alternatives $(\lambda_n^{(1)},\lambda_n^{(2)},\tau_n/n)$. On the other hand, at contiguous alternatives with e.g. $\tau_n = \mathcal{O}((\operatorname{loglog} n)\log n)$, lim inf $\tau_n/\log n>0$, $\mathbf{T}_n$ does have some nontrivial power.

7.4.2. The exponential case

Let $F_\lambda(y)=1-\exp(-\lambda y)$, $y\geq 0,\lambda>0$. Most of the results of the previous subsection also hold for the case of exponentially distributed random variables. We shall again only consider contiguous alternatives $\theta_n=(\lambda_n^{(1)},\lambda_n^{(2)},\tau_n/n)$ with $(\tau_n/n)\to\gamma\in(0,1)$, so that

$$|\lambda_n^{(1)}-\lambda_n^{(2)}| = \mathcal{O}(n^{-\frac{1}{2}}). \tag{7.12}$$

Let a_n and b_n be defined as in (7.9).

THEOREM 7.4.2.1.

$$\lim_{n\to\infty} \mathbb{P}_{H_0}\left[\mathbf{T}_n > (\frac{s+b_n}{a_n})^2\right] = \exp(-2e^{-s}),\ -\infty<s<\infty.$$

PROOF. See HACCOU et al. (1985). □

Define $\mathbf{y}_{n,k}^{(0)}=\mathbf{y}_{n,k}/\mathbb{E}_{\theta_n}(\mathbf{y}_{n,k})$, $k=1,\cdots,n$. Let

$$\mathbf{T}_n^{(0)} = \max_{1\leq k\leq n-1} \bar{\mathbf{T}}_n^{(0)}(\frac{k}{n})$$

be the likelihood ratio evaluated at $(\mathbf{y}_{n,1}^{(0)},\cdots,\mathbf{y}_{n,n}^{(0)})$. We compare $\bar{\mathbf{T}}_n(k/n)$ and $\bar{\mathbf{T}}_n^{(0)}(k/n)$ at contiguous alternatives of the type (7.12). For simplicity, we only consider the case $k\geq\tau_n$.

LEMMA 7.4.2.2. *At contiguous alternatives* $\theta_n=(\lambda_n^{(1)},\lambda_n^{(2)},\tau_n/n)$ *with* $(\tau_n/n)\to\gamma\in(0,1)$

$$\bar{\mathbf{T}}_n(\frac{k}{n})-\bar{\mathbf{T}}_n^{(0)}(\frac{k}{n}) = \frac{2(\frac{1}{\lambda_n^{(1)}}-\frac{1}{\lambda_n^{(2)}})}{\frac{1}{\lambda_n^{(2)}}}\left[\frac{\sum_{i=1}^{\tau_n}\mathbf{y}_{n,i}^{(0)}}{\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_{n,i}^{(0)}}-\frac{\sum_{i=1}^{\tau_n}\mathbf{y}_{n,i}^{(0)}}{\frac{1}{k}\sum_{i=1}^{k}\mathbf{y}_{n,i}^{(0)}}\right]$$

$$+ \frac{(\frac{1}{\lambda_n^{(1)}} - \frac{1}{\lambda_n^{(2)}})^2}{(\frac{1}{\lambda_n^{(2)}})^2} \frac{\tau_n^2}{n} \frac{n-k}{n} + \mathcal{O}_{\mathbb{P}_{\theta_n}}(n^{-1/2}),$$

uniformly in $k \geqslant \tau_n$.

PROOF. We have

$$\bar{\mathbf{T}}_n(\frac{k}{n}) = 2n\log\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_{n,i}\right]$$
$$- 2k\log\left[\frac{1}{k}\sum_{i=1}^{k}\mathbf{y}_{n,i}\right] - 2(n-k)\log\left[\frac{1}{n-k}\sum_{i=k+1}^{n}\mathbf{y}_{n,i}\right],$$

so for $k \geqslant \tau_n$,

$$\bar{\mathbf{T}}_n(\frac{k}{n}) - \bar{\mathbf{T}}_n^{(0)}(\frac{k}{n}) = -2k\log\left[1 + \frac{(\frac{1}{\lambda_n^{(1)}} - \frac{1}{\lambda_n^{(2)}})\sum_{i=1}^{\tau_n}\mathbf{y}_{n,i}^{(0)}}{\frac{1}{\lambda_n^{(2)}}\sum_{i=1}^{k}\mathbf{y}_{n,i}}\right] \tag{7.13}$$

$$+ 2n\log\left[1 + \frac{(\frac{1}{\lambda_n^{(1)}} - \frac{1}{\lambda_n^{(2)}})\sum_{i=1}^{\tau_n}\mathbf{y}_{n,i}^{(0)}}{\frac{1}{\lambda_n^{(2)}}\sum_{i=1}^{n}\mathbf{y}_{n,i}}\right]$$

$$= -2k\log(1+\mathbf{x}_k) + 2n\log(1+\mathbf{x}_n) \text{ say.}$$

Note that $\mathbf{x}_k = \mathcal{O}_{\mathbb{P}_{\theta_n}}(n^{-1/2})$ uniformly in $k \geqslant \tau_n$. Expand the two terms on the right-hand side of (7.13) in a second order Taylor series around $\mathbf{x}_k$ and $\mathbf{x}_n$ respectively, to obtain that uniformly in $k \geqslant \tau_n$

$$\bar{\mathbf{T}}_n(\frac{k}{n}) - \bar{\mathbf{T}}_n^{(0)}(\frac{k}{n}) = \frac{2(\frac{1}{\lambda_n^{(1)}} - \frac{1}{\lambda_n^{(2)}})}{\frac{1}{\lambda_n^{(2)}}}\left[\frac{\sum_{i=1}^{\tau_n}\mathbf{y}_{n,i}^{(0)}}{\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_{n,i}^{(0)}} - \frac{\sum_{i=1}^{\tau_n}\mathbf{y}_{n,i}^{(0)}}{\frac{1}{k}\sum_{i=1}^{k}\mathbf{y}_{n,i}^{(0)}}\right]$$

$$+ \frac{(\frac{1}{\lambda_n^{(1)}} - \frac{1}{\lambda_n^{(2)}})^2}{(\frac{1}{\lambda_n^{(2)}})^2}\left[k(\frac{\sum_{i=1}^{\tau_n}\mathbf{y}_{n,i}^{(0)}}{\sum_{i=1}^{k}\mathbf{y}_{n,i}^{(0)}})^2 - n(\frac{\sum_{i=1}^{\tau_n}\mathbf{y}_{n,i}^{(0)}}{\sum_{i=1}^{n}\mathbf{y}_{n,i}^{(0)}})^2\right] + \mathcal{O}_{\mathbb{P}_{\theta_n}}(n^{-1/2})$$

$$= \frac{2(\frac{1}{\lambda_n^{(1)}} - \frac{1}{\lambda_n^{(2)}})}{\frac{1}{\lambda_n^{(2)}}} \left[\frac{\sum_{i=1}^{\tau_n} \mathbf{y}_{n,i}^{(0)}}{\frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_{n,i}^{(0)}} - \frac{\sum_{i=1}^{\tau_n} \mathbf{y}_{n,i}^{(0)}}{\frac{1}{k}\sum_{i=1}^{k} \mathbf{y}_{n,i}^{(0)}} \right]$$

$$+ \frac{(\frac{1}{\lambda_n^{(1)}} - \frac{1}{\lambda_n^{(2)}})^2}{(\frac{1}{\lambda_n^{(2)}})^2} \frac{\tau_n^2}{n} \frac{n-k}{k} + \mathcal{O}_{\mathbb{P}_{\theta_n}}(n^{-1/2}). \qquad \square$$

We proceed by showing that the maximum likelihood estimator $\hat{\tau}_n / n$ is inconsistent under contiguous alternatives, i.e. the pendant of Lemma 7.4.1.2 for exponentially distributed random variables.

LEMMA 7.4.2.3.

$$\mathbb{P}_{H_0}(\frac{n}{\log n} < \hat{\tau}_n < n - \frac{n}{\log n}) \to 0$$

as well as

$$\mathbb{P}_{\theta_n}(\frac{n}{\log n} < \hat{\tau}_n < n - \frac{n}{\log n}) \to 0$$

for all θ_n of the type (7.12).

PROOF. Let $\overline{\mathbf{T}}_n(x) = \overline{\mathbf{T}}_n(k / n)$, $x \in (\frac{k-1}{n-1}, \frac{k}{n-1}]$, and

$$\mathbf{U}_n(x) = \frac{\sum_{i=1}^{k} \mathbf{y}_{n,i}}{\sum_{i=1}^{n} \mathbf{y}_{n,i}}, \quad x \in (\frac{k-1}{n-1}, \frac{k}{n-1}].$$

From HACCOU et al. (1985), we have that under H_0

$$\sup_{\frac{(\log\log n)^4}{n} < x < 1 - \frac{(\log\log n)^4}{n}} \left| \overline{\mathbf{T}}_n(x) - \frac{n(\mathbf{U}_n(x) - x)^2}{x(1-x)} \right| \tag{7.14}$$

$$= o(\log\log n), \text{ almost surely.}$$

On a rich enough probability space, one can define a sequence of Brownian bridges $\{\mathbf{B}_n(x): 0 \leqslant x \leqslant 1\}$ such that

$$\sup_{\frac{\log\log n}{n} < x < 1 - \frac{\log\log n}{n}} |n^{1/2}(\mathbf{U}_n(x) - x) - \mathbf{B}_n(x)|$$

$$= \mathcal{O}(n^{-1/2} \log n), \text{ almost surely}$$

under H_0 (see CSÖRGÖ and RÉVÉSZ (1981)). Thus under H_0

$$\sup_{\frac{1}{\log n}<x<1-\frac{1}{\log n}} \left| \frac{n^{1/2}(\mathbf{U}_n(x)-x)}{(x(1-x))^{1/2}} - \frac{\mathbf{B}_n(x)}{(x(1-x))^{1/2}} \right| \tag{7.15}$$

$$= \mathcal{O}\left(\frac{(\log n)^{3/2}}{n^{1/2}}\right) \text{ almost surely.}$$

Combination of (7.14) and (7.15) yields that under H_0

$$\sup_{\frac{1}{\log n}<x<1-\frac{1}{\log n}} [a_n \bar{\mathbf{T}}_n^{1/2}(x) - b_n] \tag{7.16}$$

$$= \sup_{\frac{1}{\log n}<x<1-\frac{1}{\log n}} \left[a_n \frac{|\mathbf{B}_n(x)|}{(x(1-x))^{1/2}} - b_n\right] + o(1), \text{ almost surely.}$$

Define $a(\rho_n)$ and $b(\rho_n)$ as in (7.9). From CSÖRGÖ and RÉVÉSZ (1981)

$$\mathbb{P}\left\{ \sup_{\frac{1}{\log n}<x<1-\frac{1}{\log n}} \left[a(\rho_n)\frac{|\mathbf{B}_n(x)|}{(x(1-x))^{1/2}} - b(\rho_n)\right] \leqslant s \right\}$$

$$\to \exp(-2e^{-s}), \quad -\infty < s < \infty.$$

In view of (7.16), this gives that under H_0

$$\sup_{\frac{1}{\log n}<x<1-\frac{1}{\log n}} |\bar{\mathbf{T}}_n(x)|^{1/2} = \mathcal{O}(\log\log\log n)^{1/2}, \text{ almost surely.}$$

Theorem 7.4.2.1 now implies that under H_0, $\hat{\tau}_n / n \leqslant 1/\log n$ or $\hat{\tau}_n / n \geqslant 1 - 1/\log n$ with probability tending to one.

This is also true under $\mathbb{P}_{\theta_n}$ because θ_n is contiguous to H_0. □

Finally, we show that $\mathbf{T}_n$ has asymptotic power equal to its asymptotic significance level at contiguous alternatives of the type (7.12).

THEOREM 7.4.2.4. *For all contiguous alternatives* $\theta_n = (\lambda_n^{(1)}, \lambda_n^{(2)}, \tau_n / n)$ *with* $(\tau_n / n) \to \gamma \in (0,1)$

$$\left| \mathbb{P}_{\theta_n}(a_n |\mathbf{T}_n|^{1/2} - b_n > s) - \mathbb{P}_{H_0}(a_n |\mathbf{T}_n|^{1/2} - b_n > s) \right| \to 0, \quad -\infty < s < \infty.$$

PROOF. Application of Lemma 7.4.2.2 gives that

$$\sup_{1 \leqslant k \leqslant \frac{n}{\log n} \text{ or } n - \frac{n}{\log n} \leqslant k \leqslant n-1} \left| \bar{\mathbf{T}}_n\left(\frac{k}{n}\right) - \bar{\mathbf{T}}_n^{(0)}\left(\frac{k}{n}\right) \right| = \mathcal{O}_{\mathbb{P}_{\theta_n}}\left(\frac{1}{\log n}\right)^{1/2}.$$

The same line of reasoning as in the proof of Theorem 7.4.1.3 now leads to the required result. □

Of course, the results of this section can be extended to other families of distributions. Now, consider the statistic

$$\tilde{\mathbf{T}}_n = \max_{\eta_n < \frac{k}{n} < 1-\eta_n} \bar{\mathbf{T}}_n(\frac{k}{n}),$$

where $0<\eta_n<\frac{1}{2}$. If no a priori knowledge about τ_n is available, it is desirable to let η_n tend to zero. But then $\tilde{\mathbf{T}}_n$ still cannot detect local alternatives $\theta_n=(\lambda_n^{(1)},\lambda_n^{(2)},\tau_n/n)$ with $(\tau_n/n)\to\gamma\in(0,1)$. The order of magnitude of $\tilde{\mathbf{T}}_n$ for the case of normally distributed random variables is given in Lemma 7.4.1.1.

7.5. Hypothesis testing in a regression model with a change-point

The two-phase regression model we study in this section is

$$\mathbf{y}_k = \begin{cases} g(\mathbf{x}_k)\theta^{(1)}+\epsilon_k & \text{if } \mathbf{x}_k \leqslant \gamma \\ g(\mathbf{x}_k)\theta^{(2)}+\epsilon_k & \text{if } \mathbf{x}_k > \gamma \end{cases}$$

where $\epsilon_1,\epsilon_2,\cdots$ are i.i.d. random variables with variance σ^2, $\mathbf{x}_1,\mathbf{x}_2,\cdots$ are i.i.d. random variables, independent of $\epsilon_1,\epsilon_2,\cdots$, with distribution $H:\mathbb{R}\to\mathbb{R}$, and where $g:\mathbb{R}\to\mathbb{R}^r$ is a known function, with

$$G = \int_{-\infty}^{\infty} g(x)^T g(x)dH(x) < \infty.$$

The $\theta^{(i)}$, $i=1,2$, are unknown elements of $\mathbb{R}^r$ and γ is the unknown change-point. The continuous version of this model, where it is assumed that $g(\gamma)\theta^{(1)}=g(\gamma)\theta^{(2)}$, is studied in FEDER (1975) (see also Section 5.2) and the discontinuous model is a special case of the one considered in Section 5.3. Also Section 6.4 treats models of this form.

We showed that under regularity conditions, the least squares estimators of $\theta^{(i)}$, $i=1,2$, are asymptotically normal, as long as the true underlying regression function actually obeys two different regimes. Example 6.8 clarifies what goes wrong if there is only one phase instead of two, and Section 7.3.1 and 7.4.1 give some more precise results for the case with ϵ_1 normally distributed and $g\equiv 1$ (i.e. $r=1$). We shall now provide some heuristics for the testing problem H_0: $\theta^{(1)}=\theta^{(2)}$ against H_1: $\theta^{(1)}\neq\theta^{(2)}$.

Let $\hat{\theta}_n^{(1)}$, $\hat{\theta}_n^{(2)}$ and $\hat{\gamma}_n$ be the least squares estimators without continuity restriction and let $\hat{\theta}_{n,H_0}$ be the least squares estimator given that H_0 is true. The residuals test statistic is

$$\mathbf{T}_n = \sum_{k=1}^{n}(\mathbf{y}_k - g(\mathbf{x}_k)\hat{\theta}_{n,H_0})^2 - \sum_{\mathbf{x}_k\leqslant\hat{\gamma}_n,\,1\leqslant k\leqslant n}(\mathbf{y}_k - g(\mathbf{x}_k)\hat{\theta}_n^{(1)})^2 - \sum_{\mathbf{x}_k>\hat{\gamma}_n,\,1\leqslant k\leqslant n}(\mathbf{y}_k - g(\mathbf{x}_k)\hat{\theta}_n^{(2)})^2.$$

Example 6.8 shows that $\mathbf{T}_n$ generally explodes at rate $\mathcal{O}_{\mathbb{P}}(\log\log n)$. Section 7.4.1 establishes its local inefficiency. Therefore, we shall consider other test statistics, which are the counterparts of the tests $\mathbf{T}_{n,\psi}$ and $|\mathbf{t}_{n,\psi}^S|$ introduced in

Section 7.3.1, and for the situation with a priori knowledge about the change-point, we present the analogue of $\tilde{\mathbf{T}}_n$ which was mentioned at the end of Section 7.4.

We shall first write $\mathbf{T}_n$ in a convenient form. Let $\bar{\mathbf{T}}_n(\gamma)$ be the residuals test statistic given that the change-point is at γ:

$$\bar{\mathbf{T}}_n(\gamma) = \sum_{k=1}^{n} (\mathbf{y}_k - g(\mathbf{x}_k)\hat{\boldsymbol{\theta}}_{n,\mathrm{H}_0})^2$$
$$- \sum_{\mathbf{x}_k \leqslant \gamma,\ 1 \leqslant k \leqslant n} (\mathbf{y}_k - g(\mathbf{x}_k)\hat{\boldsymbol{\theta}}^{(1)}_{n,\gamma})^2 - \sum_{\mathbf{x}_k > \gamma,\ 1 \leqslant k \leqslant n} (\mathbf{y}_k - g(\mathbf{x}_k)\hat{\boldsymbol{\theta}}^{(2)}_{n,\gamma})^2,$$

with $\hat{\boldsymbol{\theta}}^{(i)}_{n,\gamma}$, $i = 1,2$, the least squares estimators given γ. Of course $\bar{\mathbf{T}}_n(\gamma) \geqslant 0$. We shall write $\bar{\mathbf{T}}_n(\gamma)$ in the form

$$\bar{\mathbf{T}}_n(\gamma) = \bar{\mathbf{t}}^T_n(\gamma)\bar{\mathbf{t}}_n(\gamma),$$

$\bar{\mathbf{t}}_n(\gamma)$ defined below, and we shall consider test statistics that are functions of $\bar{\mathbf{t}}_n(\gamma)$.

Let $\mathbf{H}_n$ be the empirical distribution function based on $\mathbf{x}_1, \cdots, \mathbf{x}_n$ and define $\boldsymbol{\tau}_n(\gamma) = n\mathbf{H}_n(\gamma)$. Let $\mathbf{x}_{(1)} \leqslant \cdots \leqslant \mathbf{x}_{(n)}$ be the order statistics and write

$$\mathbf{X}_{n,\gamma} = \begin{pmatrix} g(\mathbf{x}_{(1)}) \\ \cdot \\ \cdot \\ \cdot \\ g(\mathbf{x}_{(\tau_n(\gamma))}) \end{pmatrix}, \quad \mathbf{Y}_{n,\gamma} = \begin{pmatrix} \mathbf{y}_{(1)} \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{y}_{(\tau_n(\gamma))} \end{pmatrix},$$

where $\mathbf{y}_{(k)}$ corresponds to the k-th order statistic $\mathbf{x}_{(k)}$, $k = 1, \cdots n$. Write

$$\mathbf{G}_{n,\gamma} = \mathbf{X}^T_{n,\gamma}\mathbf{X}_{n,\gamma}, \ \mathbf{G}_n = \mathbf{G}_{n,\infty}, \ \mathbf{X}_n = \mathbf{X}_{n,\infty}, \ \mathbf{Y}_n = \mathbf{Y}_{n,\infty}.$$

Then $\bar{\mathbf{t}}_n(\gamma)$ is defined for $\mathbf{G}_{n,\gamma}$ and $\mathbf{G}_n - \mathbf{G}_{n,\gamma}$ non-singular:

$$\bar{\mathbf{t}}_n(\gamma) = \mathbf{Q}^{-\frac{1}{2}}_{n,\gamma}\mathbf{A}_{n,\gamma},$$

with

$$\mathbf{Q}_{n,\gamma} = \frac{1}{n}\mathbf{G}_{n,\gamma}\mathbf{G}_n^{-1}(\mathbf{G}_n - \mathbf{G}_{n,\gamma}),$$

$$\mathbf{A}_{n,\gamma} = \frac{1}{n^{\frac{1}{2}}}(\mathbf{X}^T_{n,\gamma}\mathbf{Y}_{n,\gamma} - \mathbf{G}_{n,\gamma}\mathbf{G}_n^{-1}\mathbf{X}^T_n\mathbf{Y}_n).$$

Given $(\mathbf{x}_1, \cdots, \mathbf{x}_n) = (x_1, \cdots, x_n)$, $\bar{\mathbf{t}}_n(\gamma)$ has expectation $n^{\frac{1}{2}}\mu_n(\gamma,\gamma_0)(\theta_0^{(1)} - \theta_0^{(2)})$, where

$$\mu_n(\gamma,\gamma_0) = Q^{-\frac{1}{2}}_{n,\gamma}Q_n(\gamma,\gamma_0),$$

$$Q_n(\gamma,\gamma_0) = \begin{cases} R_n(\gamma,\gamma_0) \text{ if } \gamma \geqslant \gamma_0 \\ R_n(\gamma_0,\gamma) \text{ if } \gamma \leqslant \gamma_0 \end{cases},$$

and

$$R_n(\gamma,\gamma_0) = \frac{1}{n^{1/2}} G_{n,\gamma_0} G_n^{-1}(G_n - G_{n,\gamma}).$$

Now, define

$$G_\gamma = \int_{-\infty}^{\gamma} g^T(x)g(x)dH(x),\ Q_\gamma = G_\gamma G^{-1}(G - G_\gamma),$$

and let $\mathbf{W}$ be a standard Brownian motion. Then

$$\bar{\mathbf{t}}_n(\gamma) - n^{1/2}\boldsymbol{\mu}_n(\gamma,\gamma_0)(\theta_0^{(1)} - \theta_0^{(2)}) \xrightarrow{\mathcal{L}} Q_\gamma^{-1/2}\mathbf{B}(\gamma)$$

as process in $\gamma \in \{\gamma\colon G_\gamma$ and $G - G_\gamma$ have all eigenvalues $> \eta\}$, $\eta > 0$. Here

$$\mathbf{B}(\gamma) = \sigma \int_{-\infty}^{\gamma} g(x)d\mathbf{W}(H(x)) - G_\gamma G^{-1}\sigma \int_{-\infty}^{\infty} g(x)d\mathbf{W}(H(x)).$$

We also have that for $0 < \nu \leqslant 1/2$

$$\mathbf{Q}_{n,\gamma}^{\nu}\bar{\mathbf{t}}_n(\gamma) - n^{1/2}\mathbf{Q}_{n,\gamma}^{\nu}\boldsymbol{\mu}_n(\gamma,\gamma_0)(\theta_0^{(1)} - \theta_0^{(2)}) \xrightarrow{\mathcal{L}} Q_\gamma^{-1/2+\nu}\mathbf{B}(\gamma)$$

as process in $\gamma \in \mathbb{R}$.

This suggests test statistics of the form

$$\tilde{\mathbf{T}}_n = \sup_{\{\gamma\colon G_\gamma \text{ and } G - G_\gamma \text{ have all eigenvalues } > \eta\}} \bar{\mathbf{t}}_n^T(\gamma)\bar{\mathbf{t}}_n(\gamma)$$

which has limiting null-distribution

$$\sup_{\{\gamma\colon G_\gamma \text{ and } G - G_\gamma \text{ have all eigenvalues } > \eta\}} \mathbf{B}^T(\gamma)Q_\gamma^{-1}\mathbf{B}(\gamma)$$

and

$$\mathbf{T}_{n,\psi} = \sup_{-\infty < \gamma < \infty} \bar{\mathbf{t}}_n^T(\gamma)\mathbf{Q}_{n,\gamma}^{2\nu}\bar{\mathbf{t}}_n(\gamma)$$

with limiting null-distribution

$$\sup_{-\infty < \gamma < \infty} \mathbf{B}^T(\gamma)Q_\gamma^{-1+2\nu}\mathbf{B}(\gamma).$$

Moreover, one can construct uniform asymptotic confidence intervals for $\boldsymbol{\mu}_n(\gamma,\gamma_0)(\theta_0^{(1)} - \theta_0^{(2)})$, $\gamma,\ \gamma_0 \in \{\gamma\colon G_\gamma$ and $G - G_\gamma$ have all eigenvalues $> \eta\}$ and for $\mathbf{Q}_{n,\gamma}^{\nu}\boldsymbol{\mu}_n(\gamma,\gamma_0)(\theta_0^{(1)} - \theta_0^{(2)})$, $\nu > 0$, $\gamma \in \mathbb{R}$.

It will be clear however, that the asymptotic distributions are hardly of any practical use. One could alternatively approximate the level of the tests proposed so far, by simulating from the null-distribution. However, in general the distribution of ϵ_1 will be unknown. One could start up a simulation procedure with the disturbances normally distributed with variance $\hat{\boldsymbol{\sigma}}_n^2$, where $\hat{\boldsymbol{\sigma}}_n^2$ is some consistent estimator of σ^2, and with $(\mathbf{x}_1, \cdots, \mathbf{x}_n) = (x_1, \cdots, x_n)$. Two drawbacks are of course the computer time needed and the assumption of normality.

A more simple test statistic is

$$\mathbf{T}_{n,\psi}^{S} = \left[\int_{-\infty}^{\infty} \mathbf{Q}_{n,\gamma}^{\nu}\bar{\mathbf{t}}_n(\gamma)d\gamma\right]^T \mathbf{C}_n \left[\int_{-\infty}^{\infty} \mathbf{Q}_{n,\gamma}^{\nu}\bar{\mathbf{t}}_n(\gamma)d\gamma\right], \quad 0 \leqslant \nu \leqslant \tfrac{1}{2},$$

where $\mathbf{C}_n$ is some positive (semi-)definite matrix depending on $(\mathbf{x}_1, \cdots, \mathbf{x}_n)$ and where the integral is taken over those values of γ where $\bar{\mathbf{t}}_n(\gamma)$ is well-defined. Note that under H_0

$$\int_{-\infty}^{\infty} \mathbf{Q}_{n,\gamma}\bar{\mathbf{t}}_n(\gamma)d\gamma \xrightarrow{\mathcal{L}} \int_{-\infty}^{\infty} Q_{\gamma}^{-\frac{1}{2}+\nu}\mathbf{B}(\gamma)d\gamma, \quad 0 \leqslant \nu \leqslant \tfrac{1}{2},$$

i.e. the limiting null-distribution is multi-dimensional normal, with covariance matrix V say. One can estimate V consistently by $\hat{\mathbf{V}}_n$ say, using $(\mathbf{x}_n, \cdots, \mathbf{x}_n)$ and a $\hat{\boldsymbol{\sigma}}_n^2$. If one chooses for $\mathbf{C}_n$

$$\mathbf{C}_n = \hat{\mathbf{V}}_n^{-1}$$

then the limiting null-distribution of $\mathbf{T}_{n,\psi}^{S}$ is chi-squared with r degrees of freedom.

The Pitman efficiency of $\mathbf{T}_{n,\psi}^{S}$ at some alternative $(\theta_0^{(1)}, \theta_0^{(2)}, \gamma_0)$, with $|\theta_0^{(1)} - \theta_0^{(2)}| = n^{-\frac{1}{2}}\Delta$, can be approximated by

$$\frac{\left\{\Delta^T\left[\int_{-\infty}^{\infty} \mathbf{Q}_{n,\gamma}\boldsymbol{\mu}_n(\gamma,\gamma_0)d\gamma\right]^T \mathbf{C}_n \left[\int_{-\infty}^{\infty} \mathbf{Q}_{n,\gamma}\boldsymbol{\mu}_n(\gamma,\gamma_0)d\gamma\right]\Delta\right\}}{(\text{sum eigenvalues } \hat{\mathbf{V}}_n^{-\frac{1}{2}}\mathbf{C}_n\hat{\mathbf{V}}_n^{-\frac{1}{2}})}.$$

In the case of normally distributed errors with known variance, this is also an approximation of the Bahadur slope (see Section 7.3.1, where the Bahadur slopes for a special case are computed).

Test statistics of the type $\mathbf{T}_{n,\psi}^{S}$ could be called *sum-type* statistics and the tests based on the supremum over γ *max-type* statistics. PRAAGMAN (1986) shows for a related problem (i.e. linear rank tests for a change-point) that for every sum-type statistic there exists a max-type statistic that is at least as efficient in Bahadur's sense. This indicates that our sum-type statistics $\mathbf{T}_{n,\psi}^{S}$ are not efficient in the sense of Bahadur. However, the practical significance of this may be exponentially small.

The sum-type statistics we mentioned above are easier to use in practice than the max-type statistics $\tilde{\mathbf{T}}_n$ and $\mathbf{T}_{n,\psi}$. BROWN, DURBIN and WATSON (1975) propose the *CUSUM* test statistic, a max-type statistic that is also easy to use in practice. This test statistic is

$$CUSUM = \sup_{\gamma} \mathbf{t}_n^*(\gamma)$$

where

$$\mathbf{t}_n^*(\gamma) = \frac{1}{n^{-\frac{1}{2}}} \int_{-\infty}^{\gamma} \frac{\mathbf{y}_{(\tau_n(s)+1)} - g(\mathbf{x}_{(\tau_n(s)+1)})\mathbf{G}_{n,s}^{-1}\mathbf{X}_s^T\mathbf{Y}_s}{(1+g(\mathbf{x}_{(\tau_n(s)+1)})\mathbf{G}_{n,s}^{-1}g(\mathbf{x}_{(\tau_n(s)+1)})^T)^{\frac{1}{2}}} ds.$$

The limiting null-distribution of $\mathbf{t}_n^*(\gamma)$ is

$$\mathbf{t}_n^*(\gamma) \xrightarrow{\mathcal{L}} \sigma\mathbf{W}^*(H(\gamma)),$$

as process in γ, with $\mathbf{W}^*$ a standard Brownian motion.

8. Computation of least squares estimates in a multi-dimensional two-phase regression model

8.1 Description of the algorithm

We calculate estimates for the two-dimensional version of the two-phase regression model of Section 5.2:

$$\mathbf{y} = \min(\alpha^{(1)} + \mathbf{x}\beta^{(1)}, \alpha^{(2)} + \mathbf{x}\beta^{(2)}) + \boldsymbol{\epsilon},$$

with $\mathbf{x} = (\mathbf{z}_1, \mathbf{z}_2) \in \mathbb{R}^2$. This model is used to describe the lifetimes of plastic pipes for transportation of fluids as function of temperature and stress. The class $\mathcal{C}$ is of the form

$$\mathcal{C} = \{\{x : x\gamma \leqslant 1\} : \gamma \in \mathbb{R}^2\}.$$

Estimates are obtained from realizations $\{(x_k, y_k),\ k = 1, \cdots, n\}$ by the method of least squares. We mentioned already in Section 2.1 that the computation can be done in polynomial time. At each partition it takes $\mathcal{O}(n)$ time to find the least squares estimates given this partition. Since there are $\mathcal{O}(n^2)$ different partitions of the data $\{x_1, \cdots, x_n\}$, the total computation takes $\mathcal{O}(n^3)$ time. We shall present an algorithm that reduces this to $\mathcal{O}(n^2)$. The algorithm needs constant time to find the estimates at a given partition. Our experience however is that although asymptotically this is an improvement, the constant time needed at each partition is still substantial, i.e. of the same order of magnitude as n for moderate sample sizes ($n \simeq 70$). Some numerical results are given in the next section (Tables 3 and 4).

The main idea of the algorithm is to exploit the fact that estimates corresponding to one partition can be easily calculated from those at another partition, provided these partitions differ with respect to a limited number of points. The complexity of the calculations increases as a function of the number of points at which two partitions differ. Therefore, we aim at a sequence of partitions such that successive partitions differ in only one point.

Denote the partitions of $\{x_1, \cdots, x_n\}$ by $P_j = \{J_j^{(1)}, J_j^{(2)}\}$, with $J_j^{(1)} = A \cap \{x_1, \cdots, x_n\}$ and $J_j^{(2)} = A^c \cap \{x_1, \cdots, x_n\}$ for some $A \in \mathcal{C}$. For simplicity, we assume that no three points of $\{x_1, \cdots, x_n\}$ are on a line. In Section 8.2 we shall elaborate on the case with some points in $\{x_1, \cdots, x_n\}$ coinciding. Other violations of the assumption that there are no three points on a line necessitate only minor adjustments in the algorithm. There are now exactly $M = \binom{n}{2}$ different partitions P_j. Here, we do not include the partition $\{\{x_1, \cdots, x_n\}, \varnothing\}$ because the least squares estimate will not consider this partition as feasible.

The M partitions P_j are represented as vertices in a graph $\tilde{G} = (\mathcal{P}, \tilde{\Gamma})$, where $\mathcal{P} = \{P_1, \cdots, P_M\}$ (we identify vertices with the partitions they represent) and where $\tilde{\Gamma}$ denotes the collection of edges. Two partitions are connected by an edge in $\tilde{\Gamma}$ iff they differ in only one point. We shall now describe a method to recognize some (not all) of the adjacent vertices with little effort. The method

defines a subgraph $G=(\mathcal{P},\Gamma)$ of $\tilde{G}$ with $\Gamma\subset\tilde{\Gamma}$.

Let $x_k=(z_{k,1},z_{k,2})$, $k=1,\cdots,n$. We assume that the first co-ordinates $z_{1,1}\leqslant\cdots\leqslant z_{n,1}$ are in increasing order and that if $z_{k,1}=z_{l,1}$ for some $k\neq l$, then $z_{k,2}<z_{l,2}$. Consider the line $L_{k,l}$ through x_k and x_l. Denote by $X_{k,l}$ the 2×2-matrix

$$X_{k,l} = \begin{bmatrix} x_l \\ x_k \end{bmatrix} = \begin{bmatrix} z_{l,1} & z_{l,2} \\ z_{k,1} & z_{k,2} \end{bmatrix}.$$

Write

$$X_{k,l}^{-} = \begin{bmatrix} z_{k,2} & -z_{l,2} \\ -z_{k,1} & z_{l,1} \end{bmatrix}$$

and

$$c_{k,l} = X_{k,l}^{-}\begin{bmatrix}1\\1\end{bmatrix},$$

$$d_{k,l} = \det(X_{k,l}).$$

Then $L_{k,l}=\{x: xc_{k,l}=d_{k,l}\}$. We define $P_{k,l}=\{J_{k,l}^{(1)},J_{k,l}^{(2)}\}$ as the partition with $x_l\in J_{k,l}^{(1)}$, $x_k\in J_{k,l}^{(2)}$ and for $m\neq k,l$, $x_m\in J_{k,l}^{(1)}$ iff $x_m c_{k,l}<d_{k,l}$ (see Figure 8.1).

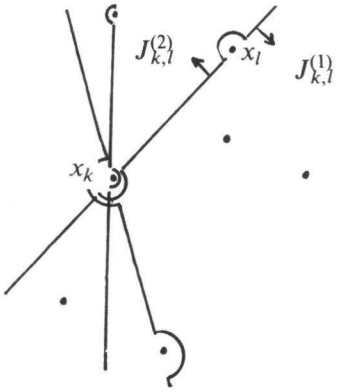


FIGURE 8.1. $P_{k,l}$ and some other partitions $P_{k,m}$

In this way, we have defined a one-to-one correspondence between all pairs $\{(x_k,x_l): k<l\in\{1,\cdots,n\}\}$ and all partitions $\{P_j: j=1,\cdots,M\}$.

The slope of $L_{k,l}$ is

$$s_{k,l} = \frac{z_{k,2}-z_{l,2}}{z_{k,1}-z_{l,1}} = -\frac{c_{k,l,1}}{c_{k,l,2}},\quad k<l,$$

with $s_{k,l}=\infty$ if $z_{k,1}=z_{l,1}$. We put the slopes in increasing order: $s_1\leqslant\cdots\leqslant s_M$ (equal slopes are ordered arbitrarily in this sequence). Let P_j be the partition corresponding to the j-th slope in the ordered sequence. Define a graph $G=(\mathcal{P},\Gamma)$, with two partitions P_{j_1} and P_{j_2}, $j_1<j_2$, adjacent iff one of the following conditions holds:

(i) $P_{j_1}=P_{k,l}$, $P_{j_2}=P_{k,m}$ and for $j_1<j<j_2$, $P_j=P_{q,r}$ where $q\neq k$, $r\neq k$,
(ii) $P_{j_1}=P_{k,l}$, $P_{j_2}=P_{k,m}$ and for $j_1<j<j_2$, $P_j=P_{q,r}$ where $q\neq l$, $r\neq l$.

EXAMPLE 8.1. Let n be equal to 5 (see Figure 8.2).

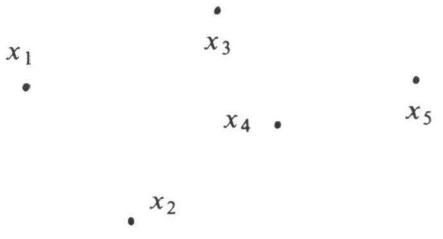


FIGURE 8.2. $n=5$

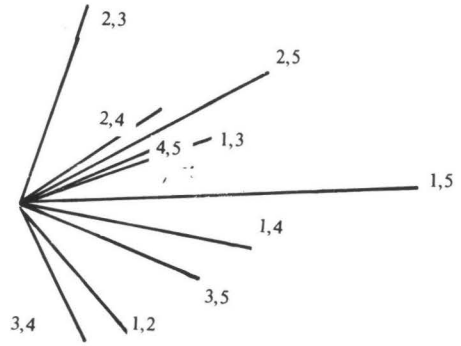


FIGURE 8.3. Ordered slopes

The ordered slopes and the corresponding partitions are

slope	partition
3,4	1 2 4|3 5
1,2	2|1 4 3 5
3,5	2 1 4 5|3
1,4	2 4|1 5 3
1,5	2 4 5|1 3
1,3	2 4 5 3|1
4,5	2 5|4 3 1
2,5	5|2 4 3 1
2,4	5 4|2 3 1
2,3	5 4 3|2 1

The graph $G=(\mathcal{P},\Gamma)$ is given in Figure 8.4.

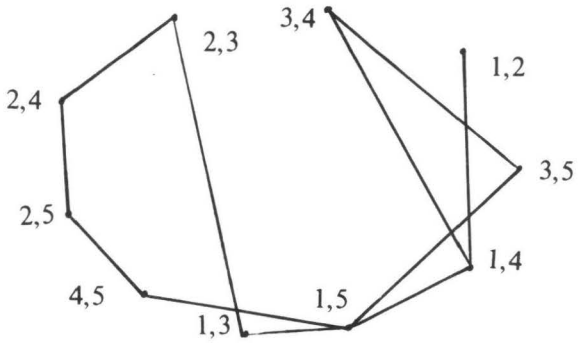


FIGURE 8.4. $(\mathcal{P},\Gamma)$ corresponding to the data of Figure 8.2

Lemma 8.1.1 asserts that two adjacent partitions in $G=(\mathcal{P},\Gamma)$ differ with respect to only one point.

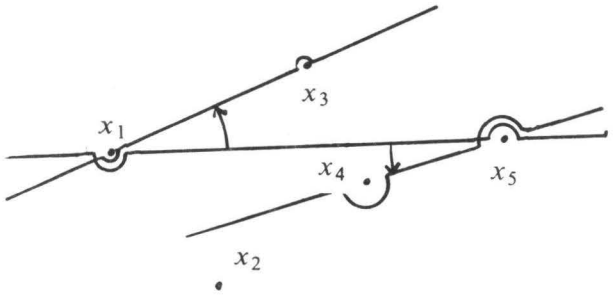


FIGURE 8.5. $P_{1,5}=\{\{2,4,5\},\{1,3\}\}$ is connected with e.g. $P_{1,3}=\{\{2,4,3,5\},\{1\}\}$

LEMMA 8.1.1.

$$\Gamma\subset\tilde{\Gamma}$$

PROOF. Let P_{j_1} and P_{j_2}, $j_1<j_2$, be adjacent in Γ, with $P_{j_1}=P_{k,l}$ and $P_{j_2}=P_{k,m}$. Then there are no data-points x_0 such that the slope between x_0 and x_k is larger than $s_{k,l}$ and smaller than $s_{k,m}$. Hence, the only point at which P_{j_1} and P_{j_2} differ is x_m. Similarly, if $P_{j_1}=P_{k,l}$ and $P_{j_2}=P_{m,l}$ are adjacent, they

can only differ in x_m. □

We shall show that $G=(\mathcal{P},\Gamma)$ is a connected graph, i.e. there is a path from each vertex to any other vertex. This is a desirable property because given estimates at one partition, one can follow the path to obtain estimates at any other partition. Let

$$\Gamma(P_j) = \{ \text{ all vertices adjacent in } G \text{ to } P_j, \text{ including } P_j \text{ itself}\}.$$

LEMMA 8.1.2. *$G=(\mathcal{P},\Gamma)$ is a connected graph.*

PROOF. This can be shown by induction. Let $G_n=(\mathcal{P}_n,\Gamma_n)$ be the graph representing the partitions and edges for a data set of size n. Obviously, the lemma holds for $n=2$.

Now, let G_{n-1} be the graph corresponding to $\{x_1,\cdots,x_{n-1}\}$ and suppose that G_{n-1} is connected. All vertices in $\mathcal{P}_n \setminus \mathcal{P}_{n-1}$ are of the form $P_j=P_{a,n}$: $a\in\{1,\cdots,n-1\}$. Let P_{j_1} and P_{j_2}, $j_1<j_2$, be two vertices in $\mathcal{P}_{n-1}$ which were adjacent in G_{n-1}, i.e. $P_{j_1}\in\Gamma_{n-1}(P_{j_2})$. Define

$$B = \begin{cases} \{a\colon P_{a,n}=P_j \text{ for some } j_1<j<j_2\}=\{a_1,\cdots,a_T\} \text{ say} \\ \varnothing \text{ if no such } a \text{ exists} \end{cases}.$$

We consider four cases:

(i) If $P_{j_1}=P_{k,l}$ and $P_{j_2}=P_{k,m}$, $k\notin B$, then $P_{j_1}\in\Gamma_n(P_{j_2})$, i.e. the edge between P_{j_1} and P_{j_2} remains in G_n.

(ii) Similarly, if $P_{j_1}=P_{k,l}$ and $P_{j_2}=P_{m,l}$, $l\notin B$, then $P_{j_1}\in\Gamma_n(P_{j_2})$.

(iii) If $P_{j_1}=P_{k,l}$ and $P_{j_2}=P_{k,m}$, $k\in B$, then there is a path $P_{k,l}\to P_{k,n}\to P_{k,m}$.

(iv) If $P_{j_1}=P_{k,l}$ and $P_{j_2}=P_{m,l}$, $l\in B$, then the situation is as in Figure 8.6. Assume without loss of generality that $x_l=(0,0)$.

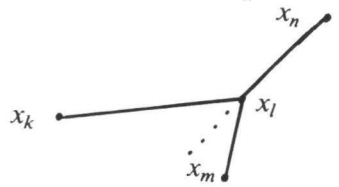


FIGURE 8.6.

Then $s_{k,l}<s_{l,n}$ is equivalent to

$$\frac{z_{k,2}}{z_{k,1}}<\frac{z_{n,2}}{z_{n,1}}.$$

This implies

$$\frac{z_{k,2}(z_{n,1}-z_{k,1})}{z_{k,1}z_{n,1}}<\frac{z_{k,1}(z_{n,2}-z_{k,2})}{z_{k,1}z_{n,1}}$$

or, since $z_{n,1}>0$ and $z_{n,2}-z_{k,2}>0$,

$$s_{k,l}=\frac{z_{k,2}}{z_{k,1}}<\frac{z_{n,2}-z_{k,2}}{z_{n,1}-z_{k,1}}=s_{k,n}.$$

In the same way, one can show that $s_{k,n}<s_{l,n}$, $s_{l,n}<s_{m,n}$ and $s_{m,n}<s_{m,l}$. Thus, one obtains a path

$$P_{k,l}\to P_{k,n}\to P_{l,n}\to P_{m,n}\to P_{m,l}.$$

In all four cases, we found that the edge between P_{j_1} and P_{j_2} remained in G_n or was replaced by a path. Clearly, all of the $(n-1)$ vertices added to $\mathfrak{P}_{n-1}$ are adjacent to at least one vertex of $\mathfrak{P}_{n-1}$. Since by induction G_{n-1} is connected, the lemma follows. □

The connected graph $G=(\mathfrak{P},\Gamma)$ has a connected subgraph $G_T=(\mathfrak{P},\Gamma_T)$, $\Gamma_T\subset\Gamma$ with the minimum number $(M-1)$ of edges. Such a subgraph is called a *generating tree.*

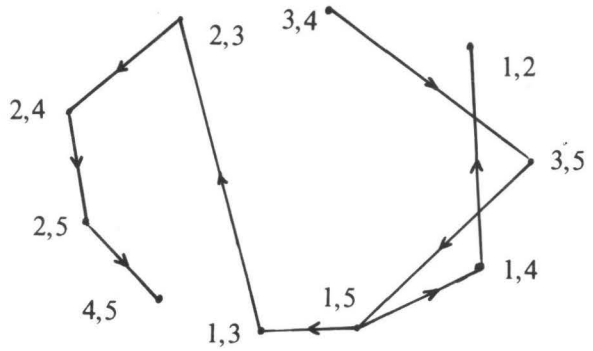


FIGURE 8.7. Generating tree for Figure 8.4

One can supply each branch in Γ_T with an orientation such that for some vertex - called the *root* of the directed graph - there is a directed path from this vertex to all other vertices.

The tree G_T endowed with orientations will define a path through the partitions. Starting in the root, one follows the directed branches until one reaches a vertex where there is no way out. Then one follows back the same path against the stream, until a vertex is entered from whence one can take a directed edge to a not previously visited vertex. The formal description of this walk is given below. We create for the original graph G a generating tree G_T including root and orientations.

(I) ALGORITHM FOR FINDING A GENERATING TREE

(1) Start in an arbitrary $P^{(1)} \in \mathcal{P}$.

(2) Given the vertices $P^{(1)}, \cdots, P^{(s)}$:

(a) find $r = \max\{t: 1 \leqslant t \leqslant s, \Gamma(P^{(t)}) \text{ not a subset of } \{P^{(1)}, \cdots, P^{(s)}\}\}$,

(b) choose a $P^{(s+1)} \in \Gamma(P^{(r)}) \setminus \{P^{(1)}, \cdots, P^{(s)}\}$,

(c) take the orientation from $P^{(r)}$ to $P^{(s+1)}$.

(3) Stop if all vertices have been visited.

While creating the generating tree, we simultaneously compute estimates. Thus, estimates corresponding to partitions are found according to the ordering $P^{(1)}, \cdots, P^{(M)}$ of the tree. We postpone the exact formulas for the estimates to the next section. Here, we only present a more or less verbal description.

The least squares estimates without continuity restriction at partition P_j are denoted by θ_j and the residual sum of squares at θ_j is denoted by S_j^2. For convenience, and to stress the fact that these estimates need not respect the continuity restriction, we sometimes write $\theta_j = \theta_{j,0}$ and $S_j^2 = S_{j,0}^2$. The issue of continuity follows now.

Define for each θ_j

$$\gamma_j = \beta_j^{(1)} - \beta_j^{(2)}, \ \delta_j = \alpha_j^{(2)} - \alpha_j^{(1)}.$$

Since θ_j does not take the continuity restriction into account, partitions of the form

$$\{\{x_k: x_k \gamma_j \leqslant \delta_j\}, \{x_k: x_k \gamma_j \geqslant \delta_j\}\}$$

need not coincide with P_j. Therefore, we consider at P_j three types of restricted estimates. Suppose $P_j = P_{k,l}$. We let $\theta_{j,1}$ be the least squares estimate at P_j under the restriction

$$\alpha_{j,1}^{(1)} + x_k \beta_{j,1}^{(1)} = \alpha_{j,1}^{(2)} + x_k \beta_{j,1}^{(2)}, \tag{8.1}$$

where $(\alpha_{j,1}^{(1)}, \beta_{j,1}^{(1)T}, \alpha_{j,1}^{(2)}, \beta_{j,1}^{(2)T}) = \theta_{j,1}^T$. Similarly, $\theta_{j,2}$ is the least squares estimate at P_j under the restriction

$$\alpha_{j,2}^{(1)} + x_l \beta_{j,2}^{(1)} = \alpha_{j,2}^{(2)} + x_l \beta_{j,2}^{(2)}. \tag{8.2}$$

Furthermore, $\theta_{j,3}$ will be the estimate at P_j under both restrictions (8.1) and (8.2). Obviously, the continuity restriction is always fulfilled at $\theta_{j,3}$. Denote by $S_{j,q}^2$ the residual sum of squares at $\theta_{j,q}$, $q = 1,2,3$.

Now, let $\theta_{j,opt}$ be the optimal solution at P_j under the continuity restriction that some partition of the form

$$\{\{x_k: x_k \gamma_{j,opt} \leqslant \delta_{j,opt}\}, \{x_k: x_k \gamma_{j,opt} \geqslant \delta_{j,opt}\}\}$$

is the same as P_j. Here, $\gamma_{j,opt}$ and $\delta_{j,opt}$ are defined by

$$\gamma_{j,opt} = \beta_{j,opt}^{(1)} - \beta_{j,opt}^{(2)}, \ \delta_{j,opt} = \alpha_{j,opt}^{(2)} - \alpha_{j,opt}^{(1)}.$$

Note that $\theta_{j,opt}$ need not be one of the $\theta_{j,q}$, $q = 0,1,2,3$. However, the algorithm is such that nevertheless the overall optimal solution $\hat{\theta}$ will be found (see

Lemma 8.1.3).

(II) ALGORITHM FOR FINDING THE LEAST SQUARES ESTIMATE $\hat{\theta}$

(1) At the root $P^{(1)}$ of the tree, the least squares estimates without continuity restriction are calculated, using a standard least squares program. These estimates - and some auxiliary variables - are stored.

(2) Given estimates and auxiliary variables at $P^{(1)}, \cdots, P^{(s)}$, we choose an r as in step (2) of algorithm (I). The least squares estimates without continuity restriction at $P^{(s+1)}$ are computed from those at $P^{(r)}$ according to the formulas given in Section 8.2.

(3) Let $j_0 = \arg\min\{S_j^2 : j \in \{1, \cdots, M\}\}$. If at θ_{j_0} the continuity restriction is fulfilled, $\hat{\theta} = \theta_{j_0}$, $\hat{S}^2 = S_{j_0}^2$ and the algorithm stops.

(4) If at θ_{j_0} the continuity restriction is not fulfilled, this necessitates the calculation of $\theta_{j_0,q}$ and $S_{j_0,q}^2$, $q = 1,2$. This can be done using the formulas of Section 8.2. The algorithm replaces $S_{j_0}^2$ by $\min\{S_{j_0,q}^2, q = 1,2\}$ and searches anew for $j_1 = \arg\min\{S_j^2 : j \in \{1, \cdots, M\}\}$. Continuing this procedure, one ends up with a sequence of indices $j_0, j_1, \cdots, j_r$ say.

(5) If $S_{j_r}^2$ has already been replaced by an $S_{j_r,q}^2$, $q = 1,2$, the algorithm calculates $\theta_{j_r,3}$ and $S_{j_r,3}^2$ and replaces $S_{j_r,q}^2$, $q = 1,2$, by $S_{j_r,3}^2$.

Algorithm (II) results in an estimate $\theta_{j_{opt},q_{opt}}$ corresponding to $S_{j_{opt},q_{opt}}^2 = \min\{S_{j,q}^2 : j \in \{1, \cdots, M\}, q \in \{0,1,2,3\}\}$. Note that algorithm (II) does not compute all $S_{j,q}^2$, $q = 1,2,3$. That $\theta_{j_{opt},q_{opt}}$ is actually the overall optimal solution $\hat{\theta}$ is shown in the following lemma.

LEMMA 8.1.3.

$$\theta_{j_{opt},q_{opt}} = \hat{\theta}.$$

PROOF. Clearly, if at each partition $\theta_{j,opt}$ were calculated, then $\hat{\theta} = \theta_{j_{opt},opt}$, where $\theta_{j_{opt},opt}$ is the estimate corresponding to

$$S_{j_{opt},opt}^2 = \min\{S_{j,opt}^2 : j \in \{1, \cdots, M\}\}.$$

Thus, we only need to show that the $\theta_{j,opt}$ that are not considered are not the overall optimal solution $\hat{\theta}$.

Let $P_j = P_{k,l}$. If $\theta_{j,opt} \notin \{\theta_{j,q} : q = 0,1,2,3\}$ then there is an x_a, $a \neq k,l$ on the line $\{x : x\gamma_{j,opt} = \delta_{j,opt}\}$. Suppose there is exactly one x_a, $a \neq k,l$ on this line. Consider the partition P_h generated by the line through x_a and some other point x_b say. If θ_h satisfies the continuity restriction, then $S_h^2 < S_{h,q}^2$, $q = 1,2$, so then $\theta_{j,opt}$ is not the overall optimum. If alternatively θ_h does not satisfy the continuity restriction, then $\theta_{j,opt} = \theta_{h,q}$ for some $q \in \{1,2\}$, so then $\theta_{j,opt}$ is considered.

Suppose there are two points x_a and x_b, $a \neq k,l$, $b \neq k,l$ on the line $\{x : x\gamma_{j,opt} = \delta_{j,opt}\}$. Let P_h now be the partition generated by this x_a and x_b.

The same line of reasoning shows that either $\theta_{j,opt}$ is not the overall optimal solution, and/or $\theta_{j,opt}=\theta_{h,3}$. □

8.2 Numerical results

We shall first present the formulas for the $\theta_{j,q}$ and $S^2_{j,q}$. Let $P_{j_1}=P^{(r)}$ and $P_{j_2}=P^{(s+1)}$ be two successive partitions in the tree that differ in one single point x_m say. Suppose $P_{j_1}=\{J^{(1)}_{j_1},J^{(2)}_{j_1}\}$, with $x_m\in J^{(1)}_{j_1}$. Then $P^{(2)}_{j_2}=\{J^{(1)}_{j_1}\setminus\{x_m\},J^{(2)}_{j_1}\cup\{x_m\}\}$. Let $Z^{(i)}_{j_1}$ be the matrix of design-points $z_k=(1,x_k)$ with $x_k\in J^{(i)}_{j_1}$, $i=1,2$. When the algorithm arrives at P_{j_2} the following quantities are in store at P_{j_1}:

1) $B^{(i)}_{j_1}=\left[Z^{(i)T}_{j_1}Z^{(i)}_{j_1}\right]^{-1}$, $i=1,2$,

2) the parameter estimates $\theta^{(i)}_{j_1}$, $i=1,2$,

3) the residual sum of squares $S^2_{j_1}$.

From these the $B^{(i)}_{j_2}$, $\theta^{(i)}_{j_2}$ and $S^2_{j_2}$ can be calculated:

1)
$$B^{(1)}_{j_2}=B^{(1)}_{j_1}-\frac{B^{(1)}_{j_1}z_mz^T_mB^{(1)}_{j_1}}{1+z_mB^{(1)}_{j_1}z^T_m}, \tag{8.3}$$
$$B^{(2)}_{j_2}=B^{(2)}_{j_1}+\frac{B^{(2)}_{j_1}z_mz^T_mB^{(2)}_{j_1}}{1-z_mB^{(2)}_{j_1}z^T_m},$$

2)
$$\theta^{(1)}_{j_2}=\theta^{(1)}_{j_1}+B^{(1)}_{j_2}z^T_m(y_m-z_m\theta^{(1)}_{j_1}), \tag{8.4}$$
$$\theta^{(2)}_{j_2}=\theta^{(2)}_{j_1}-B^{(2)}_{j_2}z^T_m(y_m-z_m\theta^{(2)}_{j_1}),$$

3)
$$S^2_{j_2}=S^2_{j_1}+\frac{(y_m-z_m\theta^{(1)}_{j_1})^2}{1+z_mB^{(1)}_{j_1}z^T_m}-\frac{(y_m-z_m\theta^{(2)}_{j_1})^2}{1-z_mB^{(2)}_{j_1}z^T_m}. \tag{8.5}$$

Given the unrestricted estimates θ_j at some partition $P_j=P_{k,l}$ say, one can also calculate the restricted estimates $\theta_{j,q}$, $q=1,2,3$. Let

$$C_j=\begin{bmatrix}B^{(1)}_j & 0\\ 0 & B^{(2)}_j\end{bmatrix}$$

and

$$r_{j,1}=(z_k,-z_k),\ z_k=(1,x_k),$$
$$r_{j,2}=(z_l,-z_l),\ z_l=(1,x_l).$$

Calculate for $q=1,2$

$$1)\quad C_{j,q} = C_j - \frac{C_j r_{j,q}^T r_{j,q} C_j}{r_{j,q} C_j r_{j,q}^T},$$

$$2)\quad \theta_{j,q} = \theta_j - \frac{C_j r_{j,q}^T r_{j,q} \theta_j}{r_{j,q} C_j r_{j,q}^T},$$

$$3)\quad S_{j,q}^2 = S_j^2 + \frac{\theta_j^T r_{j,q}^T r_{j,q} \theta_j}{r_{j,q} C_j r_{j,q}^T}.$$

Given $C_{j,q}$, $\theta_{j,q}$ and $S_{j,q}^2$ for some $q \in \{1,2\}$, say for $q=1$, we have

1) no need for further matrices

$$2)\quad \theta_{j,3} = \theta_{j,1} - \frac{C_{j,1} r_{j,2} r_{j,2}^T \theta_{j,1}}{r_{j,2} C_{j,1} r_{j,2}^T},$$

$$3)\quad S_{j,3}^2 = S_{j,1}^2 + \frac{\theta_{j,1}^T r_{j,2}^T r_{j,2} \theta_{j,1}}{r_{j,2} C_{j,1} r_{j,2}^T}.$$

We now describe how equal points in $\{x_1, \cdots, x_n\}$ are handled. Slopes $s_{k,l}$ are computed for the subset $\{x_{k_1}, \cdots, x_{k_{n'}}\}$ of different points. At the root of G_T the initial estimates at $P^{(1)}$ are calculated using the complete data set $\{(x_k, y_k),\ k=1, \cdots, n\}$. Estimates at $P_{j_2} = P^{(s+1)}$ are found from those at $P_{j_1} = P^{(r)}$ using the following transformation. Let $x_m \in \{x_{k_1}, \cdots, x_{k_{n'}}\}$ be the point at which P_{j_1} and P_{j_2} differ and suppose that there are p observations $y_m^{(1)}, \cdots, y_m^{(p)}$ at x_m, i.e. there is a group of the form $\{(x_m, y_m^{(t)}),\ t=1, \cdots, p\}$ in $\{(x_k, y_k)\colon k=1, \cdots, n\}$. In the expressions (8.3), (8.4) and (8.5) we replace $z_m = (1, x_m)$ by $\tilde{z}_m = p^{1/2} z_m$ and y_m by $\tilde{y}_m = p^{-1/2} \sum_{t=1}^{p} y_m^{(t)}$.

For the algorithm of Section 8.1 the computer program *NEWP* was written in Pascal by M. Voors. A full description of NEWP can be found in Van de Geer and Voors (1986). We first present the result of a simulation, with $n=20$ and low noise level (Table 1).

	$\alpha^{(1)}$	$\beta_1^{(1)}$	$\beta_2^{(1)}$	$\alpha^{(2)}$	$\beta_1^{(2)}$	$\beta_2^{(2)}$
θ_0	1	3	5	4	1	2
$\hat{\theta}$	1.03	3.00	5.00	4.14	0.98	2.00

Table 1. Simulation results, $n=20$, $\hat{S}^2 = 1.13$

Real data were supplied by a firm for the production of plastic pipes:

$$y = \log(\text{life-time of a pipe})$$

$$z_1 = \frac{\text{stress}}{\text{absolute temperature}}$$

$$z_2 = \frac{1}{\text{absolute temperature}}$$

There are $n = 295$ observations. We first used the program *NONLINWOOD* (see DANIEL and WOOD (1980)). This is a program for computation of least squares estimates in a general nonlinear regression model. To obtain starting values for NONLINWOOD, F. Burger wrote a special program for life-times-of-pipes-data, which calculates estimates at a hopefully representative subset of all possible partitions. The program NONLINWOOD was run several times with varying starting values and step sizes. From the outcomes we took the one with the smallest residual sum of squares. The result is given in the first row of Table 2.

The program NEWP is too costly to handle the complete data set on the interactive system to which we had access, even though after grouping equal x_k there remained only $n' = 71$ observations (see Table 3 and 4). Therefore, we simply threw away 11 observations. The data turned out to be more or less ordered with respect to temperature: in the second row of Table 2 high temperatures are disregarded whereas in the fourth row low temperatures are omitted. Note that throwing away observations from the reduced data set means not using more than four times as many observations from the original data set.

	n or n'	$\alpha^{(1)}$	β_1^1	$\beta_2^{(1)}$	$\alpha^{(2)}$	$\beta_1^{(2)}$	$\beta_2^{(2)}$	S^2
(1)	295	-45.13	-50.19	21.82	-26.79	-22.55	12.21	47.20
(2)	60	-56.20	-62.63	27.19	-41.93	-25.95	12.21	10.12
(3)	60	-51.45	-69.99	26.74	-44.08	-27.09	18.47	10.45
(4)	60	-41.20	-51.27	20.78	-39.11	-26.13	16.72	10.88

TABLE 2. (1)NONLINWOOD, (2)NEWP 1 -60, (3)NEWP 6-65, (4)NEWP 12-71

Table 3 and Table 4 present an overview of the relative cost of NEWP as function of n'.

n'	SIMP	NEWP	TREE
20	3.69	4.31	4.27
30	6.32	8.01	9.04
40	13.27	17.89	21.05
50	29.86	73.33	82.24
60	65.80	199.93	218.31
71	***	***	***

TABLE 3. NP-costs, *** insufficient field length for load

n'	SIMP	NEWP	TREE
20	41735	43157	42401
30	62724	64204	63376
40	112567	114105	113277
50	151306	152662	152054
60	216733	220313	217505
71	***	***	***

TABLE 4. CM-costs, *** insufficient field length for load

The program SIMP uses straightforward calculations, i.e. no generating tree is created and at each partition the estimates are computed directly without making use of previously obtained estimates at other partitions. TREE does create the generating tree but it does not use it: estimates are computed as in SIMP. As to be expected, NEWP is cheaper than TREE as regards NP-costs (normal priority costs) but less economical with CM-costs (central memory costs). Roughly speaking, the difference between TREE and SIMP represents the time needed for creating a generating tree. This turns out to be very costly. In order to decide which partition will be next in the generating tree, the program has to make about 20 comparisons at each partition. This may be substantial, but we did not expect it to outweigh the $\mathcal{O}(n)$ effort needed for recalculating estimates at each partition.

9. References

ALEXANDER, K.S. (1984). Probability inequalities for empirical processes and a law of the iterated logarithm. *Ann. probability 12,* 1041-1067.

BACON, D.W. and D.G. WATTS (1971). Estimating the transition between two intersecting straight lines. *Biometrika 58,* 525-534.

BAHADUR, R.R. (1967). An optimal property of the likelihood ratio statistic. *Proc. Fifth Berkeley Symp. Math. Stat. Prob. 1,* 13-26.

BAHADUR, R.R. (1971). Some limit theorems in statistics. *SIAM,* Philadelphia.

BAHADUR, R.R. and M. RAGHAVACHARI (1972). Some asymptotic properties of likelihood ratios on general sample spaces. *Proc. Sixth Berkeley Symp. Math. Stat. Prob. 1,* 129-152.

BARD, Y. (1974). *Nonlinear parameter estimation,* Academic Press, New York.

BENNETT, G. (1962). Probability inequalities for the sum of independent random variables. *J. Amer. Statist. Assoc. 57,* 33-45.

BERNSTEIN, S. (1924). Sur un modification de l'inéqualité de Tchebichef. *Annals Science Institute Sav. Ukraine, Sect. Math. I,* (Russian, Franch Summary).

BERNSTEIN, S. (1927) *Theory of Probability,* Moscow.

BIRGÉ L. (1980). Thèse. *Université Paris VII.*

BIRGÉ, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrscheinlichkeitstheorie verw. Gebiete 65,* 181-237.

BROWN, R.L., J. DURBIN and J.M. EVANS (1975). Techniques for testing the constancy of regression relationships over time (with discussion). *J. Roy. Statist. Soc. B, 37,* 149-192.

COOK, R.D. and S. WEISBERG (1982). *Residuals and influence in regression,* London: Chapman and Hall.

COVER, T.M. (1965). Geometric and statistical properties of systems of linear inequalities with applications to pattern recognition. *IEEE Trans. Elec. Comp. EC-14,* 326-334.

CSÖRGÖ, M. and P. RÉVÉSZ (1981). *Strong approximations in probability and statistics,* Academic Press.

DANIEL, C. and F.S. WOOD (1980). *Fitting equations to data,* Wiley and Sons.

DARLING, D.A. and P. ERDÖS (1956). A limit theorem for the maximum of normalized sums of independent random variables. *Duke Math. J. 23,* 143-155.

DEHARDT, J. (1971). Generalizations of the Glivenko-Cantelli theorem. *Ann. Math. Statist. 42,* 2050-2055.

DESHAYES, J. and D. PICARD (1982). Test of disorder of regression: asymptotic comparison. *Theory of Prob. and Appl. 27,* 100-115.

DESHAYES, J. and D. PICARD (1983). *Rupture de modelès en statistique,* Thèses d'Etat, Orsay.

DUDLEY, R.M. (1967). The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *J. Functional analysis 1,* 290-330.

DUDLEY, R.M. and W. PHILIPP (1983). Invariance principles for sums of Banach space valued random elements and empirical processes. *Z. Wahrscheinlichkeitstheorie verw. Gebiete 62,* 509-552.

DUDLEY, R.M. (1984). A course on empirical processes. *Springer Lecture Notes in Math. (Lectures given at Ecole d'Eté de Probabilités de St. Flour, 1982),* 1-142.

FEDER, P.I. (1975). On asymptotic distribution theory in segmented regression problems - identified case. *Ann. Stat. 3,* 49-83.

FERREIRA, P.E. (1975). Bayes switching regressions. *J. Amer. Statist. Ass. 350,* 370-374.

GIHMAN, I.I. and A.V. SKOROHOD (1974). *The theory of stochastic processes, I,* Springer Verlag, New York.

GINÉ, E. and J. ZINN (1984). On the central limit theorem for empirical processes. *Ann. Prob. 12,* 929-989.

GROENEBOOM, P. and J. OOSTERHOFF (1977). Bahadur efficiency and probabilities of large deviations. *Statist. Neerl. 31,* 1-24.

HACCOU, P., E. MEELIS and S.A. VAN DE GEER (1985). On the likelihood ratio test for a change point in a sequence of independent exponentially distributed random variables. *Report MS R8507,* Centre for mathematics and Computer Science, Amsterdam.

HARDING, E.F. (1967). The number of partitions of a set of N points in k dimensions induced by hyperplanes. *Proc. Edinburgh Math. Soc. (Ser. II) 15,* 285-289.

HARTLEY, H.O. and A. BOOKER (1965). Nonlinear least squares estimation. *Ann. Math. Statist. 36,* 638-650.

HINKLEY, D.V. (1969). Inference about the intersection in two-phase regression. *Biometrika 56,* 495-504.

HINKLEY, D.V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika 57,* 1-17.

HINKLEY, D.V. and E.A. HINKLEY (1970). Inference about the change-point in a sequence of binomial variables. *Biometrika 57,* 477-488.

HINKLEY, D.V. (1971). Inference in two-phase regression. *J. Amer. Statist. Assoc. 66,* 736-743.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc. 58,* 13-30.

HUBER, P.J. (1967). The behaviour of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Symp. Math. Stat. Prob.,* 221-233.

HUDSON, D.J. (1966). Fitting segmented curves whose join points have to be estimated. *J. Amer. Statist. Assoc. 61,* 1097-1129.

IBRAGIMOV, I.A. and R.Z. HAS'MINSKII (1981). *Statistical estimation : asymptotic theory,* Springer-Verlag, New York.

IPPEL, M.J. and L.A. BEEM (1986). A theory of antagonistic strategies. In: *Learning and Instruction,* New York, Pergamon Press., Vol. I, 111-121.

JENNRICH, R.I. (1969). Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist. 40,* 633-643.

KALLENBERG, W.C.M. (1978). *Asymptotic optimality of likelihood ratio tests in exponential families,* Mathematical Centre tracts 77.

KOLMOGOROV, A.N. and V.M. TIHOMIROV (1959). ϵ-entropy and ϵ-capacity of sets in function spaces. *Uspehi Mat. Nauk. 14, 3-86;* English transl., *Amer. Math. Soc. Transl. (2, 1961), 17,* 277-364.

KUELBS, J. (1978). Some exponential moments of sums of independent random variables. *Trans. Amer. Math. Soc. 240,* 145-162.

LAI, T.L. H. ROBBINS and C.Z. WEI (1978). Strong consistency of least squares estimators in multiple regression. *Proc. Nat. Acad. Sci. U.S.A. 75,* 3034-3036.

LECAM, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *Ann. Statist. 41,* 802-828.

LECAM, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist. 1,* 38-53.

LENSTRA, A.K., J.K. LENSTRA, A.H.G. RINNOOY KAN and T.J. WANSBEEK (1982). Two lines least squares. *Ann. Discrete Math. 16,* 201-211.

MOEN, D.H. and L.P. BROEMELING (1984). Testing for a change in the regression matrix of a multivariate linear model. *Commun. Statist. - Theor. Math. 13,* 1521-1531.

NGUYEN, H.T., G.S. ROGERS and E.A. WALKER (1984). Estimating in change-point hazard rate models. *Biometrika 71,* 299-304.

OOSTERHOFF, J. and W.R. VAN ZWET (1975). A note on contiguity and Hellinger distance. *SW 36/75,* Stichting Mathematisch Centrum.

PETTITT, A.N. (1979). A non-parametric approach to the change-point problem. *Appl. Statist. 28,* 126-135.

PICARD, D. (1983). Testing and estimating change-points in time series. *Technical report,* Orsay.

POLLARD, D. (1981). Strong consistency of k-means clustering. *Ann. Statist. 9,* 135-140.

POLLARD, D. (1981). Limit theorems for empirical processes. *Z. Wahrscheinlichkeitstheorie verw. Gebiete 57,* 181-195.

POLLARD, D. (1982). A central limit theorem for empirical processes. *J. Austr. Math. Soc. (Series A) 33,* 235-248.

POLLARD, D. (1984). *Convergence of stochastic processes,* Springer Series in Statistics, Springer-Verlag, New York.

PRAAGMAN, J. (1986). *Efficiency of change-point tests,* Ph. D. Thesis, T.H. Eindhoven.

QUANDT, R.E. (1958). The estimation of a linear regression obeying two separate regimes. *J. Amer. Statist. Assoc. 51,* 873-886.

ROYSTON, J.P. and R.M. ABRAMS (1980). An objective method for detecting the shift in basal body temperature in women. *Biometrics 36,* 217-224.

SCHLÄFLI, L. (1901, Posth.). *Theorie der vielfachen Kontinuität.* In: Gesammelte Math. Abhandlungen I, (Basel, Birkhäuser, 1950).

SHIRYAYEV, A.N. (1963). On optimum methods in quickest detection problems. *Th. Prob. and Appl. 8,* 22-46.

STEELE, J.M. (1978). Empirical discrepancies and subadditive processes. *Ann.*

Probability 6, 118-127.

STEINER, J. (1826). Einige Gesetze über die Theilung der Ebene und des Raumes. *J. Reine Angew. Math. 1,* 349-364.

STONE, C.J. (1982). Optimal rates of convergence for nonparametric regression. *Ann. Statist. 10,* 1040-1053.

TISHLER, A. and I. ZANG (1981). A new maximum likelihood algorithm for piecewise regression. *J. Amer. Statist. Assoc. 76,* 980-987.

VAN DE GEER, S.A. and M. VOORS (1986). A computer program for the two-phase broken-hyperplane regression model, (in Dutch). *Report MS-N8602,* Centre for Mathematics and Computer Science.

VAPNIK, V.N. and Y.A. CHERVONENKIS (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Prob. and Appl. 16,* 264-280.

VAPNIK, V.N. and Y.A. CHERVONENKIS (1981). Necessary and sufficient conditions for the uniform convergence of means to their expectations. *Theory of Prob. and Appl. 26,* 532-553.

WAHBA, G. (1984). Partial spline models for the semi-parametric estimation of functions of several variables. *Statistical analysis of time series. Tokyo: Institute of Statistical Mathematics,* 319-329.

WATSON, P. (1969). On partitions of *n* points. *Proc. Edinburgh Math. Soc. 16,* 263-264.

WHITTLE, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory of Prob. and Appl. 5,* 302-305.

WILLIAMS, D.A. (1970). Discrimination between regression models to determine the pattern of enzyme synthesis in synchronous cell cultures. *Biometrics 26,* 23-32.

WOLFE, A. and E. SCHECHTMAN (1984). Nonparametric statistical procedures for the change-point problem. *J. Statist. Planning and Inference 9,* 389-396.

WORSELEY, K.J. (1985). Confidence intervals and tests for a change-point in a sequence of exponential family random variables. *Biometrics 72,* 1-14.

WU, C.F. (1981). Asymptotic theory of nonlinear least squares estimation. *Ann. Statist. 9,* 501-513.

Samenvatting

Regressieanalyse en empirische processen

De klasse van regressiemodellen die in dit proefschrift wordt bestudeerd is

$$\mathbf{y}_k = g(\mathbf{x}_k)+\boldsymbol{\epsilon}_k,\ k=1,\cdots,n,$$

met $\boldsymbol{\epsilon}_1,\cdots,\boldsymbol{\epsilon}_n$ onderling onafhankelijke stochastische grootheden met verwachting nul en eindige variantie, en $\mathbf{x}_1,\cdots,\mathbf{x}_n$ vectoren in $\mathbb{R}^d$. De funktie g wordt verondersteld een element te zijn van een collectie $\mathcal{G}$ van regressiefunkties. Voorbeelden zijn niet-lineaire regressie, waarbij $\mathcal{G}$ een klasse is van funkties geïndexeerd door een Euclidische parameter, en niet-parametrische regressie met bijvoorbeeld $\mathcal{G}$ een klasse van gladde funkties.

We onderzoeken de relatie tussen de 'grootte' van $\mathcal{G}$ en het asymptotisch gedrag van de kleinste-kwadratenschatter $\hat{\mathbf{g}}_n$. Zij $g_0 \in \mathcal{G}$ de ware onderliggende regressie. Des te minder men van g_0 bekend veronderstelt, des te groter is $\mathcal{G}$ en des te moeilijker zal het zijn g_0 te schatten. We preciseren dit door de *entropie* van $\mathcal{G}$ te beschouwen en maken daarbij gebruik van de theorie over *empirische processen*. Ter illustratie gaan we in op het twee-fasen regressiemodel.

In (lineaire) twee-fasen regressie, de klasse $\mathcal{G}$ is de verzameling van funkties van de vorm

$$g = g^{(1)}1_A + g^{(2)}1_{A^c},$$

met $g^{(1)}$ en $g^{(2)}$ lineair en de verzameling A variërend in een klasse $\mathcal{A}$ van deelverzamelingen van $\mathbb{R}^d$. Hoofdstuk 1 geeft een aantal voorbeelden van klassieke twee-fasen regressie, waar $d=1$ en waar de funkties g een knik of sprong hebben. In klassieke twee-fasen regressie is $\mathcal{A}$ de collectie van halfrechten; in het algemeen kan men ook andere klassen $\mathcal{A}$ beschouwen.

Empirische proces-theorie betreft met name de uitbreiding van de Glivenko-Cantelli-stelling naar algemene uniforme wetten van grote aantallen en uniforme centrale-limietstellingen. Hoofdstuk 2 geeft een overzicht van de literatuur over uniforme wetten van grote aantallen en generaliseert de theorie naar het geval van niet-identiek verdeelde stochastische grootheden. In Hoofdstuk 3 worden deze resultaten toegepast op regressie. Er wordt beschreven in hoeverre entropievoorwaarden op $\mathcal{G}$ leiden tot consistentie van de kleinste-kwadratenschatter $\hat{\mathbf{g}}_n$.

Hoofdstuk 4 behandelt de uniforme centrale-limietstellingen die in de navolgende hoofdstukken als referentiekader zullen dienen. In Hoofdstuk 5 wordt ingegaan op het twee-fasen regressiemodel. Het blijkt relatief eenvoudig om - gegeven consistentie en de theorie van Hoofdstuk 4 - asymptotische

normaliteit van de kleinste-kwadratenschatters van de Euclidische parameters af te leiden.

In Hoofdstuk 6 keren we terug naar het algemene regressiemodel. Hier wordt op een wat subtielere manier gebruik gemaakt van de entropie van $\mathcal{G}$, waardoor het mogelijk wordt de convergentiesnelheid voor $\hat{\mathbf{g}}_n$ te bepalen. De entropievoorwaarden in dit hoofdstuk gelden echter vaak alleen lokaal, d.w.z. in een omgeving van g_0. Met behulp van de resultaten in Hoofdstuk 3 kan men nagaan of $\hat{\mathbf{g}}_n$ op den duur in zo'n omgeving belandt. Naast niet-parametrische regressie dient het twee-fasen model weer ter illustratie.

Twee-fasen regressie is sterk verwant met de situatie waarbij men een abrupte verandering modelleert in de verdelingsfunkties van een rij van onafhankelijke stochastische grootheden. In Hoofdstuk 7 besteden we aandacht aan dit laatste geval. We onderzoeken de asymptotische efficiëntie van de likelihood-ratio toets voor de aanwezigheid van een verandering.

Tenslotte presenteert Hoofdstuk 8 een algoritme voor het berekenen van de kleinste-kwadratenschatter van een twee-fasen regressiemodel.

Curriculum vitae

Sara van de Geer werd op 7 mei 1958 geboren te Leiden. Na het V.W.O. aan de Rembrandt Scholengemeenschap in Leiden, begon zij in september 1976 de studie in de wiskunde bij de Rijksuniversiteit Leiden. In januari 1982 behaalde zij het doctoraalexamen met bijvak Psychologie.

Van mei 1982 tot oktober 1983 werkte zij in Tilburg aan het Z.W.O.-projekt "Latente variabelen en sociale invloeden op consumptiepatronen en individuele voorkeuren". Daarna was zij tot oktober 1987 verbonden aan het Centrum voor Wiskunde en Informatica te Amsterdam.

Naar het zich laat aanzien zal zij gedurende het academisch jaar 1987/1988 werkzaam zijn bij de School of Mathematics van de Universiteit van Bristol, UK.

Stellingen

bij het proefschrift

Regression Analysis and Empirical Processes

Sara van de Geer

- 1 -

Het bewijs van Stelling 2.3.2 in dit proefschrift kan eenvoudig getransformeerd worden om de volgende uniforme wet van de grote aantallen, uniform over een klasse $\mathcal{K}$ van verdelingen op $\mathbf{R}^d$, af te leiden: Zij $\mathbf{H}_n$ de empirische maat gebaseerd op n o.o. waarnemingen met verdeling $H \in \mathcal{K}$ en zij $\mathcal{F}$ een klasse van reële funkties op $\mathbf{R}^d$ met $F = \sup_{f \in \mathcal{F}} |f|$. Stel dat $\lim_{C \to \infty} \sup_{H \in \mathcal{K}} \int_{F>C} F dH = 0$ en dat voor alle $\delta > 0$, $(1/n) \log N_1(\delta, \mathbf{H}_n, \mathcal{F}) \to^{\mathbf{P}_H} 0$, uniform in $H \in \mathcal{K}$, waarbij $\log N_1(\delta, \mathbf{H}_n, \mathcal{F})$ de δ-entropie van $\mathcal{F}$ is voor de (pseudo-)metriek $\int |\cdot| d\mathbf{H}_n$. Dan geldt (onder meetbaarheidsvoorwaarden)

$$\sup_{f \in \mathcal{F}} \left| \int f d(\mathbf{H}_n - H) \right| \to^{\mathbf{P}_H} 0,$$

uniform in $H \in \mathcal{K}$.

- 2 -

Zij $\mathcal{F}$ een klasse van reële uniform begrensde funkties op $\mathbf{R}^d$ en laat voor alle $\delta > 0$, $\mathcal{F}_\delta$ een overdekking in supremum-norm van $\mathcal{F}$ zijn. Zij $\mathcal{Q}_\delta$ de collectie van *graphs* van funkties in $\mathcal{F}_\delta$ en zij $\Delta^{\mathcal{Q}_\delta}(x_1, \cdots, x_n)$ het aantal verschillende verzamelingen van de vorm $A_\delta \cap \{x_1, \cdots, x_n\}$, $A_\delta \in \mathcal{Q}_\delta$. Als $\mathcal{F}_\delta$ zó gekozen kan worden dat

$$\sup_{x_1, \cdots, x_n} \log \Delta^{\mathcal{Q}_\delta}(x_1, \cdots, x_n) \leqslant W \delta^{-\nu} \log n, \quad \delta > 0,$$

voor zekere constanten W en $\nu \geqslant 0$ en alle $n \geqslant 1$, dan geldt voor de δ-entropie $\log N_2(\delta, Q, \mathcal{F})$ behorende bij de (pseudo-)metriek $(\int |\cdot|^2 dQ)^{1/2}$:

$$\log N_2(\delta, Q, \mathcal{F}) \leqslant M \delta^{-\nu} \log\left(\frac{1}{\delta}\right), \quad \delta > 0,$$

voor zekere constante M die alleen van de maat Q afhangt. Als $\mathcal{F}$ een *VC-graph* klasse is, kan men $\mathcal{F}_\delta = \mathcal{F}$ en $\nu = 0$ kiezen en komt dit resultaat overeen met het *Approximation Lemma* in [1].

[1]POLLARD, D. (1984). *Convergence of Stochastic Processes.* Springer Series in Statistics, Springer Verlag, New York.

- 3 -

Laat $\{P_\theta : \theta \in \mathbf{R}^r\}$ een collectie kansmaten zijn met dichtheid $p_\theta = dP_\theta / d\mu$ ten opzichte van een σ-finiete maat μ. Zij $\hat{\boldsymbol{\theta}}_n$ de meest aannemelijke schatter gebaseerd op n o.o. waarnemingen $\mathbf{x}_k$, $k = 1, \cdots, n$, met verdeling P_θ. Stel dat er een oneindige verzameling $\Theta \subset \mathbf{R}^r$ is met dimensie kleiner dan r, zodanig dat $\forall\, \theta \in \Theta\ \exists\, \tilde{\theta} \neq \theta$ met $P_\theta = P_{\tilde{\theta}}$, en zodanig dat $\forall\, \theta \notin \Theta$, $P_\theta = P_{\tilde{\theta}}$ d.e.s.d. als $\theta = \tilde{\theta}$. Dan is in het algemeen voor $n \to \infty$ de logaritme van het aannemelijkheidsquotiënt

$$\sum_{k=1}^{n} \log p_{\hat{\boldsymbol{\theta}}_n}(\mathbf{x}_k) - \sum_{k=1}^{n} \log p_\theta(\mathbf{x}_k)$$

niet begrensd in $\mathbb{P}_\theta$-kans. Voorbeelden zijn het twee-fasen regressiemodel en het twee-compartimentenmodel.

- 4 -

De beperking tot $n^{-\frac{1}{2}}$-omgevinkjes van de oneindig-dimensionale component van de onbekende parameter, zoals in [2] gebeurt, verdient te worden gerechtvaardigd.

[2]BEGUN, J.M., W.J. HALL, W.M. HUANG en J. WELLNER (1983). Information and asymptotic efficiency in parameteric–nonparametric models. *Ann. Statist. 11,* 435-452

- 5 -

Grote-afwijkingen en *lokale asymptotiek* zijn twee wiskundige technieken ter benadering van een experiment $\mathcal{E}^n = \{P_\theta^n : \theta \in \Theta\}$ voor grote waarden van n. Aangezien deze benaderingen tot tegengestelde conclusies kunnen leiden, is ten minste één ervan alleen statistisch zinvol onder extra regulariteitsvoorwaarden.

- 6 -

Bij een model met abrupte verandering in de parameters van orde $n^{-\frac{1}{2}}$ voor $n \to \infty$ kan de lokatie de verandering niet geschat worden maar het bestaan ervan kan wel worden getoetst.

- 7 -

Een inkomenspolitiek die rekening houdt met een subjectief oordeel van het individu over de subjectieve waarde van het inkomen, kan in abstracto bestudeerd worden (zie [3]), maar is in praktijk onuitvoerbaar.

[3]KAPTEYN, A., S. VAN DE GEER en H. VAN DE STADT (1985). The impact of changes in income and family composition on subjective measures of well-being. In: *Horizontal Equity, Uncertainty, and Economic Well-Being.* Studies in Income and Wealth, Vol. 50, 35-67, The University of Chicago Press

- 8 -

Aangezien 'commerciële kunst' een contradictio in terminis is, betekent de afschaffing van de BKR dat hedendaagse beeldende kunst als overbodig wordt gezien.

- 9 -

In [4] wordt het volgende knapzakprobleem onderzocht:

$$\max\left\{\sum_{j=1}^{n} \mathbf{c}_j x_j : \sum_{j=1}^{n} \mathbf{a}_{ij} x_j \leqslant n b_i,\ i = 1, \cdots, m,\ x_j \in \{0,1\},\ j = 1, \cdots, n\right\},$$

met $\mathbf{c}_1, \mathbf{c}_2, \cdots$ en $\mathbf{a}_{i1}, \mathbf{a}_{i2}, \cdots$, $i = 1, \cdots, m$, onafhankelijke identiek verdeelde stochastische grootheden met waarden in $[0,1]$. De Lagrange-relaxatie van het continue probleem is

$$\mathbf{L}_n(\lambda) = \max\left\{\sum^{m} \lambda_i b_i + \frac{1}{n}\sum^{n}(\mathbf{c}_j - \sum^{m} \lambda_i \mathbf{a}_{ij}) x_j : 0 \leqslant x_j \leqslant 1,\ j = 1, \cdots, n\right\}.$$

Laat $\lambda_n^* \geq 0$ een oplossing van de Lagrange-relaxatie zijn. Zij $L(\lambda)=\mathbf{E}\mathbf{L}_n(\lambda)$ en stel dat $L(\lambda)$ een uniek minimum λ^* heeft. Dan geldt voor $n \to \infty$

$$\left(\frac{n}{\text{loglog}n}\right)^{1/2} \left|\mathbf{L}_n(\lambda_n^*) - L(\lambda^*)\right| = \mathcal{O}(1)$$

met kans 1.

[4]VAN DE GEER, S. en L. STOUGIE (1987). A note on the rate of convergence of the multi-knapsack value function. *To appear*

- 10 -

Beschouwt men de statistische consultaties waar ik mee te maken heb gehad als representatieve steekproef uit het universum van statistische consultaties, dan leidt dit tot de conclusie dat proeven met muizen een onaanvaardbaar groot bestanddeel van het wetenschappelijk onderzoek vormen.

- 11 -

Het sex-gedrag van de Chlamydomonas engametos kan worden beschreven door middel van een statistisch model door de overgangswaarschijnlijkheden van vrije, ongebonden cel naar geëxciteerde of gebonden cel, te relateren aan het aantal ongebonden cellen van het andere geslacht. In de limiet levert dit model een aantal differentiaalvergelijkingen op die analoog zijn aan de Boltzmann-vergelijking.

[5]DEMETS, R., A.M. TOMSON, S VAN DE GEER en A. TIP (1987). A statistical description of sexual cell interaction in Chlamydomonas engametos. *Pre-print,* FOM-institute for Atomic and Molecular Physics

- 12 -

Een vertolking van *Das Wohltemperierte Klavier* op piano kan deze muziek een dynamische dimensie geven, maar doet tekort aan het karakter van de verschillende toonsoorten.