Estimation of a Regression Function by Maxima of Minima of Linear Functions

Dissertation
zur Erlangung des Grades
des Doktors der Naturwissenschaften
der Naturwissenschaftlich-Technischen Fakultäten
der Universität des Saarlandes

vorgelegt von

Conny Clausen

Saarbrücken 2008

Tag des Kolloquiums:	13.06.2008
Dekan:	Prof. Dr. Joachim Weickert
Prüfungsausschuss:	Vorsitzender
	Prof. Dr. Jörg Eschmeier
	Berichterstatter
	Prof. Dr. Michael Kohler
	Prof. Dr. Alfred K. Louis
	Akademischer Mitarbeiter
	Dr. Christoph Barbian

To
my parents

Abstract

The estimation of a multivariate regression function from independent and identically distributed random variables is considered. First we propose and analyse estimates which are defined by minimisation of the empirical L_2 risk over a class of functions consisting of maxima of minima of linear functions. It is shown that the estimates are strongly universally consistent. Moreover results concerning the rate of convergence of the estimates with data-dependent parameter choice using 'splitting the sample' are derived in the case of an unbounded response variable. In particular it is shown that, for smooth regression functions satisfying the assumptions of single index models, the estimate is able to achieve (up to some logarithmic factor) the corresponding optimal one–dimesional rate of convergence. In this context it is remarkable that this newly proposed estimate can be computed in applications (see the appendix).

Furthermore an L_2 boosting algorithm for estimation of a regression function is presented. This method repeatedly fits a function from a fixed function space to the residuals of the data and the number of iteration steps is chosen data–dependently by 'splitting the sample'. A general result concerning the rate of convergence of the algorithm is derived in the case of an unbounded response variable. Finally this method is used to fit a sum of maxima of minima of linear functions to a given set of data. The derived rate of convergence of the corresponding estimate does not depend on the dimension of the observation variable.

Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit der Schätzung multivariater Regressionsfunktionen anhand von unabhängig und identisch verteilten Zufallsvariablen. Zunächst wird ein neues Schätzverfahren vorgestellt, welches auf der Minimierung des empirischen L_2–Risikos bezüglich einer Funktionenklasse, die aus Maxima von Minima von linearen Funtionen besteht, basiert. Für dieses Schätzverfahren wird zunächst die starke universelle Konsistenz nachgewiesen. Weiterhin werden sowohl für diesen Schätzer als auch für das entsprechende Schätzverfahren mit datenabhängiger Parameterwahl (mittels „Splitting the Sample“) die entsprechenden Konvergenzraten hergeleitet. Diese Konvergenzraten gelten insbesondere auch dann, wenn die abhängige Variable unbeschränkt ist. Insbesondere wird gezeigt, dass unter den Voraussetzungen des „Single Index Models“ die (bis auf einen logarithmischen Faktor) zugehörige optimale eindimensionale Konvergenzrate erreicht wird.

Weiterhin wird in dieser Arbeit ein L_2–Boosting–Algorithmus zur Schätzung multivariater Regressionsfunktionen vorgestellt. Bei diesem Verfahren werden schrittweise Funktionen eines festgewählten Funktionenraumes an die Residuen der Daten angepasst. Auch hierbei erfolgt die Wahl der Anzahl der Iterationsschritte wieder datenabhängig. Es wird für diesen L_2–Boosting–Algorithmus zunächst ein allgemeines Resultat bezüglich der Konvergenzrate hergeleitet, welches auch in dem Fall einer unbeschränkten abhängigen Variablen gilt. Abschließend wird dieses Verfahren verwendet, um einen Schätzer zu konstruieren, der als Summe von Maxima von Minima von linearen Funktionen dargestellt werden kann. Die für diesen Schätzer hergeleitete Konvergenzrate hängt nicht mehr von der Dimension der unabhängigen Variablen ab.

Contents

List of symbols 3

Introduction 5

Chapter 1. Preliminaries 11
1.1. Regression Analysis 11
1.2. Least Squares Method 14
1.3. Consistency and Rate of Convergence 16
1.4. Vapnik-Chervonenkis Theory 19
1.5. Auxiliary Results 23

Chapter 2. Maxima of Minima of Linear Functions 27
2.1. Definition of the Estimate 27
2.2. Characterisation of $\mathcal{F}_n$ 29
2.3. Covering Numbers of $\mathcal{F}_n$ 38

Chapter 3. Analysis of Asymptotic Behaviour 43
3.1. Universal Consistency 43
3.2. Rate of Convergence 50
3.3. Splitting the Sample 58

Chapter 4. Dimension Reduction 67
4.1. Single Index Models 67
4.2. Projection Pursuit 70

Chapter 5. L_2 Boosting 77
5.1. A general L_2 Boosting Result 77
5.2. L_2 Boosting with Maxmin Functions 89

Appendix 95
A.1. The Algorithm 95
A.2. Application to Simulated Data 96

Bibliography 107

List of symbols

$\mathcal{F}$ ..... A class of functions $f : \mathbb{R}^d \to \mathbb{R}$.
$\mathcal{F}_n$ 27
$\mathcal{F}_{m,n}$ 28
$\mathcal{F}_n^1$ 68
$\mathcal{H}_N^{\mathcal{F}}$ 79
$\mathcal{F}_C$ 89
$\bigoplus_{i=1}^K \mathcal{F}$ 70

$\mathcal{D}_n$ 12
$\mathcal{Q}_n$ 59
S_r 90

$\arg\min_{f \in \mathcal{F}_n}$ 14
log ..... natural logarithm
m ..... Regression function
$sign(z)$ ..... algebraic sign of $z \in \mathbb{R}$.
T_β 28

$\mathcal{M}_p(\varepsilon, \mathcal{G}, z_1^n)$ 21
$\mathcal{N}_p(\varepsilon, \mathcal{G}, z_1^n)$ 20
$V_{\mathcal{G}}$ 21

$||x||$ ..... Euclidean norm of $x \in \mathbb{R}^d$.
$||x||_1$ ..... $||x||_1 = \sum_{i=1}^d |x^{(i)}|$, for $x \in \mathbb{R}^d$.
$\lceil x \rceil$ ..... upper integer part of $x \in \mathbb{R}$.
$x \cdot y$ ..... Scalar product between $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^d$.
$||f||_\infty$ ..... Supremum norm of $f : \mathbb{R}^d \to \mathbb{R}$.

$a.s.$ ..... almost surely
A^C ..... Complement of an event A.
$\mathbf{E}$ ..... Expected value
$\mathbf{Var}$ ..... Variance
μ ..... Distribution of X

$L_2(\mu)$ ..... Set of all square integrable functions with respect to μ.
$C_0(\mathbb{R}^d)$ ..... Set of all continuous functions $f : \mathbb{R}^d \to \mathbb{R}$ with bounded support.
$C_0^\infty(\mathbb{R}^d)$ ..... Set of all infinitely often continuously differentiable functions.
..... $f : \mathbb{R}^d \to \mathbb{R}$ with bounded support.

$|A|$ ..... Cardinality of a set A.
e ..... Euler's number
$\mathbf{1}_A$ ..... Indicator function of a set A.
(p, C)–smooth 18
$\mathcal{O}$ ..... Landau notation

We assume that all random variables in a joint context are defined on the same probability space $(\Omega, \mathcal{A}, \mathbf{P})$.

Whenever we say a function $f : \mathbb{R}^d \to \mathbb{R}$ is measurable, this denotes the measurability with respect to the Borel algebras on $\mathbb{R}^d$ and $\mathbb{R}$, respectively.

Introduction

The regression estimation problem is one of the most important subjects in statistics. Regression estimation is a technique for the modeling and analysis of observed data consisting of values of a dependent variable (response variable) Y and one or more independent variables (observation variables) X. In its earliest form this field of activity goes back to A.M. Legendre and C.F. Gauß at the beginning of the 19th century. The problem of regression estimation is of increasing importance today, not least because of the enormous growth of information technology. Whereas in the early days the underlying questions came from industrial experiments or agricultural issues, and therefore the statistical problems were relatively simple, the appearance of computers has entailed a massive growth in both the number and complexity of statistical problems.

This dissertation considers the problem of estimating a multivariate regression function given a sample of the underlying distribution. That is, we try to estimate a regression function which describes the relationship between the dependent real–valued random variable Y and the $\mathbb{R}^d$–valued random vector X. The most famous method for this purpose is the principle of least squares. It was first used in linear regression, where, roughly speaking, the aim is to fit a line through a cloud of points. Since a linear relationsship between X and Y is a very simple model, this method has been applied to other parametric as well as nonparametric settings.

In applications no a priori information about the regression function is usually known and it is therefore necessary to apply nonparametric methods to this estimation problem. There are several established methods for nonparametric regression, including regression trees such as CART, which were proposed by Breiman et al. (1984), adaptive spline fitting such as MARS, as introduced by Friedman (1991), or least squares neural network estimates (cf. Chapter 11 in Hastie, Tibshirani and Friedmann (2001)). All these methods also minimize a kind of least squares risk of the regression estimate. For neural networks this is done heuristically over a fixed and very complex function space, whereas regression trees and spline fitting use this principle over a stepwise defined data dependent space of piecewise constant functions or piecewise polynomials, respectively.

In this dissertation we consider a rather complex function space consisting of maxima of minima of linear functions, and we also minimize a least squares risk over this class of functions in order to define our estimate. To be more precise, we deal with functions $f : \mathbb{R}^d \to \mathbb{R}$ of the form

$$f(x) = \max_{k=1,...,K_n} \min_{l=1,...,L_{k,n}} (a_{k,l} \cdot x + b_{k,l}) \quad (x \in \mathbb{R}^d),$$

for some $a_{k,l} \in \mathbb{R}^d$ and $b_{k,l} \in \mathbb{R}$. K_n and $L_{1,n}, ..., L_{K_n,n}$ are parameters of the class of functions or, in other words, parameters of the corresponding regression estimate. Since each maximum of minima of linear functions is in fact continuous and piecewise linear function (cf. Example 2.1), we actually fit a linear spline function with free knots to the data. However, in contrast to MARS, we do not need heuristics to choose these free knots, but use instead advanced methods from optimization theory of nonlinear and nonconvex functions to compute our estimate approximately in applications.

In general, there is a gap between theory and practice in multivariate nonparametric regression function estimation. The established estimates as CART, MARS or least squares neural networks need a some heuristics for their computation, and this makes it practically impossible to analyse their rate of convergence theoretically. On the other hand, a definition of these estimates without any heuristics allows a theoretical analysis of their rates of convergence, but in this form the estimates cannot be computed in an application. Results of this kind concerning the rate of convergence can be found in Barron (1993, 1994) for neural networks, and for CART in Kohler (1999).

A similar phenomenon also occurs for our estimate, since we need heuristics to compute it approximately in an application. However, in contrast to the above-mentioned estimates, we use heuristics from advanced optimization theory. In particular we use methods from nonlinear and nonconvex optimization theory (see Bagirov (1999, 2002) and Bagirov and Ugon (2006)) instead of complicated heuristics from statistics for stepwise computation as for CART or MARS, or a simple gradient descent as for least squares neural networks.

We now give an outline of the dissertation and summarise our results. The first chapter provides an introduction to regression function estimation and briefly describes the central ideas in the analysis of nonparametric regression estimates. Furthermore, it deals with all results necessary for the analysis of our estimate, particularly results from the Vapnik-Chervonenkis theory.

In Chapter 2 we introduce a class of functions consisting of maxima of minima of linear functions and discuss some of its properties. In particular, we discuss how it

is related to the class of linear spline functions. Based on this function space, we define a regression function estimate by minimising a corresponding least squares risk. A subsequent truncation of this estimate yields the maxmin estimate in which we are interested. Moreover, Chapter 2 discusses bounds on the covering numbers of classes of maxima of minima of linear functions, in order to obtain bounds on the estimation error of the estimate.

These bounds are used in Chapter 3, to prove results concerning consistency and rate of convergence. Firstly, we see there that the maxmin estimate presented in Chapter 2 is strongly universally consistent (cf. Definition 1.3) for all distributions of (X, Y) with $X \in [0, 1]^d$. Secondly, Section 3.2 provides a bound on the expectation of the estimation error of the maxmin estimate and therefore on its L_2 error as well. For this purpose we use a theorem of Lee, Bartlett and Williamson (1996).

The approach of Lee, Bartlett and Williamson is described in detail in Section 11.3 in Györfi at al. (2002). We extend this approach to unbounded data which satisfy a modified Sub-Gaussian condition (cf. Inequality 1.12) by introducing new truncation arguments. In this way we are able to derive a rate of convergence under similar general assumptions on the distribution of Y as in alternative methods from empirical process theory (see van de Geer (2000), or Kohler (2000, 2006)). From the bound on the L_2 error and an approximation result from Schumaker (cf. Lemma 1.16) we infer that our estimate has the rate of convergence

$$C^{2d/(2p+d)} \cdot \left(\frac{\log(n)^3}{n} \right)^{2p/(2p+d)}$$

if the underlying regression function is (p, C)–smooth (cf. Definition 1.4). This rate also holds for unbounded Y which satisfies the Sub-Gaussian condition. Moreover, it follows from Stone (1982) that this rate of convergence is optimal (in some minimax sense) up to a logarithmic factor.

Since these results hold only for a certain choice of parameters (depending on the smoothness of the regression function) we complete Chapter 3 with the definition of an estimate with data-dependent parameter choice using 'splitting the sample'. Such an adaptive parameter choice is very important because in applications we have usually no information about the smoothness of the underlying regression function. We obtain the same rate of convergence for the so–defined estimate under similar assumptions.

The above rate of convergence is obviously not completely satisfactory in the high-dimensional case, that is, for large dimension d of the observation variable X.

Therefore, Chapter 4 describes two methods of dimension reduction in which additional assumptions are made in order to derive better rates of convergence. The idea of imposing additional restrictions on the structure of the regression function (such as additivity or the assumption in the single index model) and so to derive better rates of convergence is due to Stone (1985, 1994). We shall prove that, even for large dimension of X, the L_2 error of our estimate quickly converges to zero if the regression function satisfies the assumption of single index models (see Theorem 3.6). Similar results are shown in Section 22.2 of Györfi et al. (2002), but in contrast to the estimate defined there our newly proposed estimate can be computed in applications.

In Section 4.2 we consider so–called projection pursuit, which is a generalisation of additive models. We derive the one-dimensional rate of convergence in this setting as well (cf. Theorem 4.2). However, the estimate used in projection pursuit is different from the maxmin estimate presented in Chapter 2. Namely, we consider an estimate which is defined by minimizing the least squares risk over a class of functions consisting of sums of maxima of minima of linear functions. Therefore, this estimate unfortunately exhibits the same computability problems as the estimates in Section 22.2 of Györfi et al. (2002).

In order to overcome these difficulties, Chapter 5 provides an L_2 boosting estimate, which can be computed in applications, and in addition is a sum of maxima of minima of linear functions. Boosting is a very well-known method proposed by Freund and Schapire (1996). It is based on the idea of repeatedly fitting a function from a fixed function space to the residuals of the data. In Section 5.1 we present a general L_2 boosting result by using ideas from Barron et al. (2006), and extend them to unbounded data and parameter choice via splitting the sample in place of complexity regularisation. In Section 5.2 this result is applied to a class of maxima of minima of linear functions. From this and an approximation result for neural networks (cf. Lemma 16.8 in Györfi et al. (2002)), we can infer a parametric rate of convergence for the L_2 boosting estimate based on maxmin functions.

The appendix completes this dissertation by examining the behaviour of the maxmin estimate from Chapter 2 in a small simulation study. Since the development of the algorithm used for the computation of the maxmin estimate was mainly made by Bagirov, and is therefore not part of this thesis, we refer for detailed information to Bagirov, Clausen and Kohler (2007). Here we only give a brief description of the algorithm in Appendix A.1. In part A.2 of the appendix we provide an application of our estimate to simulated data for different regression functions of varying dimension.

In the case $d = 1$ we compare the maxmin estimate to kernel estimates (with Gaussian kernel), local linear kernel estimates, smoothing splines, neural networks and regression trees, whereas for $d > 1$, our estimate is only compared to the last two estimates used in the one–dimensional simulations. In summary, we can state that our estimate certainly performs well in comparison with the established estimates. Even in the univariate case the maxmin estimate can actually outperfrom the other estimates for large sample sizes. In the multivariate case our estimate is generally better than regression trees and moreover it often outperforms neural networks even for small sample sizes.

There are a number of people whom I would like to thank. First, I am very grateful to my supervisor Prof. Dr. Michael Kohler for suggesting the subject of this dissertation and for always listening patiently to all my questions. I would like to thank him for the valuable advice he gave me throughout the development of this thesis and in addition, for making possible my visit to Melbourne. For this great experience and the generous support and hospitality during my stay I am also very grateful to Dr. Gleb Beliakov. Furthermore, I would like to thank Prof. Dr. Mark Groves and Prof. Dr. Alfred Louis for their advice and encouragement. I am also very grateful to Dr. Christoph Barbian for always helping with words and deeds. Moreover, I want and need to thank all my dear friends for looking after me. Finally, I would like to thank my whole family (and especially my mom) for their loving support and confidence throughout the last years.

CHAPTER 1

Preliminaries

This first chapter represents a general introduction in nonparametric regression estimation and analysis. The first two sections describe the problem of estimating a multivariate regression function given a sample of the underlying distribution, and point out the advantages of the nonparametric regression estimation and particularly the convenience of least squares estimates. In Sections 1.3 and 1.4, we overview the main ideas in the analysis of regression estimates, and summarise some results from Vapnik-Chervonenkis Theory which permit the analysis of nonparametric estimates.

1.1. Regression Analysis

In regression analysis one considers an $\mathbb{R}^d \times \mathbb{R}$-valued random vector (X, Y) with $E(Y^2) < \infty$, and one is interested in the dependency of the *response variable* Y on the value of the *observation variable* X. Roughly speaking this means that we have a set of points in the $(d+1)$–dimensional space, where the x–coordinate is d–dimensional and the y–coordinate is one-dimensional and it is our aim to describe the path in average of the y–coordinate dependent on the x–coordinates.

Thus we want to find a function $f : \mathbb{R}^d \to \mathbb{R}$, such that $f(X)$ is close to Y in some sense or in other words $f(X)$ should be 'a good approximation of Y'. This problem can be resolved by the introduction of the so-called L_2 *risk* or *mean squared error of* f,

$$\mathbf{E}\left(|f(X) - Y|^2\right), \tag{1.1}$$

and the requirement that it is as small as possible. It is not immediately obvious why the minimisation of the L_2 risk is reasonable. However, if one restates '$f(X)$ is close to Y' into '$|f(X) - Y|$ is small' and reminds that (X, Y) is a random vector and therefore $|f(X) - Y|$ is random as well, the use of the expectation in (1.1) is reasonable instantly.
It is well-known that the so-called *regression function*

$$m : \mathbb{R}^d \to \mathbb{R},\ m(x) = \mathbf{E}\left(Y|X = x\right) \tag{1.2}$$

minimizes the L_2 risk under all measurable functions. This fact results directly from the following equation for arbitrary measurable functions $f : \mathbb{R}^d \to \mathbb{R}$. We have

$$\begin{aligned}\mathbf{E}\left(|f(X)-Y|^2\right) &= \mathbf{E}\left(\big((f(X)-m(X))+(m(X)-Y)\big)^2\right)\\ &= \mathbf{E}\left(|f(X)-m(X)|^2\right)+\mathbf{E}\left(|m(X)-Y|^2\right)\\ &= \mathbf{E}\left(|m(X)-Y|^2\right)+\int |f(x)-m(x)|^2\mu(dx), \qquad (1.3)\end{aligned}$$

where μ denotes the distribution of X, and the second equation follows from

$$\begin{aligned}\mathbf{E}\big(\,(f(X)-m(X))\,(m(X)-Y)\,\big) &= \mathbf{E}\big(\,(f(X)-m(X))\cdot\mathbf{E}\,((m(X)-Y)\,|X)\,\big)\\ &= \mathbf{E}\,((f(X)-m(X))\cdot(m(X)-m(X)))\\ &= 0.\end{aligned}$$

Due to the fact that the so-called L_2 *error*

$$\int |f(x)-m(x)|^2\mu(dx) \qquad (1.4)$$

is always nonnegative, it is clear that

$$\mathbf{E}\left(|m(X)-Y|^2\right)=\min_{f:\mathbb{R}^d\to\mathbb{R},f \text{ measurable}}\mathbf{E}\left(|f(X)-Y|^2\right)$$

holds for the regression function m. Hence the optimal approximation of Y with respect to the L_2 risk by a function of X is given by the regression function.

So far we did not take into account that in applications the distribution of (X,Y) is usually unknown. Hence the regression function is unknown as well and therefore cannot be used as predictor of Y. However, in many applications it is possible to observe a sample of the underlying distribution and to estimate the regression function from this known sample.

Let us suppose that $(X,Y),(X_1,Y_1),(X_2,Y_2),\dots$ are independent and identically distributed random variables with $\mathbf{E}(Y^2)<\infty$, and that we have given a set of data

$$\mathcal{D}_n=\{(X_1,Y_1),\dots,(X_n,Y_n)\}.$$

Our main aim is to construct an estimate of the regression function, which clearly should depend on this sample. To be more precise, we want to construct an estimate

$$m_n(\cdot)=m_n(\cdot,\mathcal{D}_n):\mathbb{R}^d\to\mathbb{R}, \qquad (1.5)$$

such that the L_2 error

$$\int |m_n(x)-m(x)|^2\mu(dx)$$

is small. Since equation (1.3) shows that the L_2 risk $\mathbf{E}(|m_n(X) - Y|^2|\mathcal{D}_n)$ of a measurable estimate m_n is close to the optimal value if and only if the L_2 error

$$\int_{\mathbb{R}^d} |m_n(x) - m(x)|^2 \mu(dx)$$

is small the L_2 error is a plausible error criterion in the context of regression analysis. Therefore we are using the L_2 error in order to measure the quality of an estimate. Here it should be mentioned that one can find different error criteria for regression analysis in the literature such as the pointwise error or the supremum norm error for example, and of course every criterion has its assets and drawbacks. However, we are using the L_2 error as measure of the performance of regression function estimates.

The traditional approach to estimate regression functions assumes that the regression function is included in a known class of functions, which can be described by finitely many parameters. This approach corresponds to the so-called *parametric regression* estimation. In the parametric case one uses the given data to estimate the unknown values of the parameters. The most popular parametric regression estimate is the linear regression estimate where one assumes that the regression function is linear, that is,

$$m(x^{(1)}, \dots, x^{(d)}) = a_0 + \sum_{i=1}^{d} a_i x^{(i)} \qquad ((x^{(1)}, \dots, x^{(d)})^T \in \mathbb{R}^d),$$

for unknown real numbers $a_o, a_1, \dots, a_d$. Thus one just has to estimate $d + 1$ parameters and this estimation is usually quite easy and moreover suitable even if the sample size is small.

In spite of these advantages the parametric regression estimation has one serious drawback. It is very unflexible in terms of the shape of the regression function, that is, the method is only promising if the underlying regression function is contained in the assumed class of functions. Otherwise the resulting estimate cannot approximate the regression function better than the best function with the assumed structure, and hence the resulting estimate produces a large error even for large sample sizes.

To avoid this disadvantage we consider *nonparametric regression* estimates. Nonparametric methods do not make the assumption that the regression function has a certain shape, which can be described by several parameters, and hence allow statements for more general distributions of (X, Y). Therefore we do not need informations about the shape of the regression function to calculate nonparametric estimates. Especially in the multivariate case this can be a huge advantage, because

in most high–dimensional cases it is just impossible to make assumptions concerning the constitution of the distribution, since for example graphic tools cannot be considered for $d > 2$.

1.2. Least Squares Method

A very famous principle to construct regression estimates (both in parametric and nonparametric regression analysis) is the *principle of least squares*. This classical method was independently proposed by A. Legendre in 1805 and C.F. Gauß in 1809, and results from the following equality

$$\mathbf{E}\left(|m(X)-Y|^2\right) = \min_{f:\mathbb{R}^d\to\mathbb{R}, f \text{ measurable}} \mathbf{E}\left(|f(X)-Y|^2\right),$$

which we have already seen earlier. The central idea is to estimate the L_2 risk, $\mathbf{E}\left(|f(X)-Y|^2\right)$, of a function f by the so-called empirical L_2 risk

$$\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-Y_i|^2, \tag{1.6}$$

for a given set of data. Afterwards one chooses a function which minimizes the empirical L_2 risk over some given class of functions as estimate for the regression function. In the parametric case this class of functions again is determined by finitely many parameters but in the nonparametric case there are no such restrictions. However it is self-evident that not every class of functions is reasonable even in the nonparametric approach. Thus, one has to choose a suitable class of functions $\mathcal{F}_n$, which may (and in many cases actually does) depend on the sample size n, and the resulting estimate m_n is defined by

$$m_n(\cdot) = \arg\min_{f\in\mathcal{F}_n} \frac{1}{n}\sum_{1=1}^{n}|f(X_i)-Y_i|^2, \tag{1.7}$$

which on the other hand is definded by

$$m_n \in \mathcal{F}_n \text{ and } \frac{1}{n}\sum_{1=1}^{n}|m_n(X_i)-Y_i|^2 = \min_{f\in\mathcal{F}_n}\frac{1}{n}\sum_{1=1}^{n}|f(X_i)-Y_i|^2.$$

Here we assume that the minimum exists, but we do not require its uniqueness.

Usually the set of functions $\mathcal{F}_n$ grows as the sample size grows. This idea goes back to Grenander (1981) and is known as 'method of thieves'. Moreover, some approaches even use function spaces $\mathcal{F}_n$, which does depend not only on the sample size but also depend on the sample.

However as already mentioned, the choice of $\mathcal{F}_n$ is very important but it is not very easy. On the one hand a large underlying class of functions has the advantage that

it is more likely that it will contain functions, which can approximate the unknown regression function very well. Basically this is owing to the requirement that the estimate is contained in $\mathcal{F}_n$ and hence cannot approximate the regression function better than the best function in $\mathcal{F}_n$.

On the other hand, if for example $X_1, ..., X_n$ are all distinct, which, in the case that X has a density, is almost sure, and $\mathcal{F}_n$ is too massive then this method leads to an estimate that just interpolates the data points $(X_1, Y_1), \ldots, (X_n, Y_n)$. Obviously such an estimate is not a reasonable estimate for $m(x) = \mathbf{E}(Y|X = x)$. Therefore it is really important to choose a sufficiently large (but not too large) class of functions $\mathcal{F}_n$. The following lemma restates this difficulty exactly.

LEMMA 1.1. *Let $\mathcal{F}_n$ be a class of measurable functions $f : \mathbb{R}^d \to \mathbb{R}$, that maybe depends on the data $\mathcal{D}_n$. Then for every estimate $m_n : \mathbb{R}^d \to \mathbb{R}$ satisfying (1.7) the inequality*

$$\begin{aligned}\int |m_n(x) - m(x)|^2 \mu(dx) \quad &\leq \quad 2 \sup_{f \in \mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2 - \mathbf{E}\left(|f(X) - Y|^2\right) \right| \\ &\quad + \inf_{f \in \mathcal{F}_n} \int |f(x) - m(x)| \mu(dx)\end{aligned}$$

holds.

PROOF. This lemma is well known and a proof can be found in Lugosi and Zeger (1995). □

In fact, this lemma provides a decomposition of the L_2 error of the estimate into (up to a factor two) the so-called *estimation error*,

$$\sup_{f \in \mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2 - \mathbf{E}\left(|f(X) - Y|^2\right) \right|, \tag{1.8}$$

and the so-called *approximation error*,

$$\inf_{f \in \mathcal{F}_n} \int |f(x) - m(x)| \mu(dx). \tag{1.9}$$

The estimation error (1.8) can be seen as the maximal difference between the L_2 risk of the estimate and the L_2 risk of the functions contained in $\mathcal{F}_n$, whereas the approximation error (1.9) measures how well the regression function can be approximated by functions of $\mathcal{F}_n$.

In this dissertation we will use the principle of least squares in order to construct suitable nonparametric regression estimates. However, before we start with the

introduction of the underlying function space and the exact definition of our regression function estimates, firstly we consider properties concerning the performance of regression estimates and particularly least squares estimates.

1.3. Consistency and Rate of Convergence

In Section 1.1 we have already motivated the use of the L_2 error

$$\int |m_n(x) - m(x)|^2 \mu(dx)$$

to measure the error of regression estimates. As a matter of course, the L_2 error of a good regression estimate should be very small and therefore, the weakest property a regression estimate should have, is the convergence of its L_2 error to zero, for a sample size tending to infinity. This attribute is called consistency and is defined next.

DEFINITION 1.2. *A sequence of regression function estimates $(m_n)_{n\in\mathbb{N}}$ is called* ***weakly consistent*** *for a certain distribution of (X,Y), if*

$$\lim_{n\to\infty} \mathbf{E}\left(\int |m_n(x) - m(x)|^2 \mu(dx)\right) = 0,$$

and it is called ***strongly consistent*** *for a certain distribution of (X,Y), if*

$$\lim_{n\to\infty} \int |m_n(x) - m(x)|^2 \mu(dx) = 0 \qquad a.s.$$

However, consistency for a certain distribution is just the weakest requirement a reasonable estimate should fulfil. Even if it is consistent for a certain class of distributions of (X,Y) we do not know its performance for anything but these. Since in most applications the distribution of (X,Y) is exactly what is unknown it would be of high interest to exhibit an estimate which is consistent for all distributions or at least for a large class of distributions of (X,Y). This desirable distribution-free consistency goes back to Stone (1977), and is defined as follows:

DEFINITION 1.3. *A sequence of regression function estimates $(m_n)_{n\in\mathbb{N}}$ is called* ***weakly universally consistent****, if it is weakly consistent for all distributions of (X,Y) with $\mathbf{E}(Y^2) < \infty$. Analogously the sequence $(m_n)_{n\in\mathbb{N}}$ is called* ***strongly universally consistent****, if it is strongly consistent for all distributions of (X,Y) with $\mathbf{E}(Y^2) < \infty$.*

For the first time, the existence of weakly universally consistent estimates was proved in Stone (1977). More precisely Stone has shown that nearest-neighbour-estimates have this attribute. About twenty years later it was shown in Devroye,

Györfi, Krzyżak and Lugosi (1994) that nearest-neighbour-estimates are actually strongly universally consistent. In the meantime universal consistency (both weak and strong) was shown for a number of estimates. Detailed descriptions and proofs, in particular for *partitioning estimates*, *kernel estimates*, *smoothing spline estimates*, and *least squares estimates*, can be found in Györfi et al. (2002).

In order to prove universal consistency for least squares esimates it is common to study the approximation error (1.9) and the estimation error (1.8) seperately. In this manner, an upper bound on the L_2 error is obtained for all distributions of (X, Y) with $\mathbf{E}(Y^2) < \infty$, and its convergence ensures that the regression estimate is universally consistent.

The analysis of the approximation error mostly is the simpler one. Since $C_0(\mathbb{R}^d)$ is dense in $L_2(\mu)$ for every distribution μ of X (cf. Lemma 1.15) we can require that $m \in C_0(\mathbb{R}^d)$. Due to the inequality

$$\inf_{f \in \mathcal{F}_n} \int |f(x) - m(x)|^2 \mu(dx) \leq \inf_{f \in \mathcal{F}_n} ||f - m||_\infty^2$$

the class of functions $\mathcal{F}_n$ just has to be chosen such that functions in $C_0(\mathbb{R}^d)$ can be approximated arbitrarily close with respect to the $||\cdot||_\infty$ by functions of $\mathcal{F}_n$. From this we can infer directly that the approximation error tends to zero.

On the other hand, bounding the estimation error often is more difficult. Here it is necessary to require the uniform boundedness of $|f(X) - Y|$ over $\mathcal{F}_n$ in order to apply the so-called *Vapnik-Chervonenkis theory* and it is usually a quite difficult task to prove this boundedness over a class of functions $\mathcal{F}_n$. However, as one can see in Section 10.2 of Györfi et al. (2002), it is sufficient to prove the convergence of the estimation error to zero only for bounded Y. Moreover, in the case of bounded Y it suffices to choose the class $\mathcal{F}_n$ such that its functions are uniformly bounded by some constant depending on the sample size n, in order to obtain the uniform boundedness of $|f(X) - Y|$ (cf. Lugosi and Zeger (1995) and Haussler (1992)).

Even though universal consistency is a quite strong property, it is not the only thing we need to know in practical applications. The consistency guarantees the convergence of the L_2 error to zero for a growing sample size n, but especially in applications the sample size often is prescribed, and hence it is desirable that the L_2 error of an estimate tends to zero as fast as possible. Thus together with the consistency, the rate of convergence is of high interest during the analyis of regression estimates. To provide a rate of convergence of an estimate m_n we will analyse

$$\mathbf{E}\left(\int |m_n(x) - m(x)|^2 \mu(dx) \right), \tag{1.10}$$

for fixed $n \in \mathbb{N}$.

Unfortunately there exists no estimate which converges to zero at some fixed nontrivial rate for all distributions of (X, Y) with $\mathbf{E}(Y^2) < \infty$, as Devroye has proved in 1982. Thus one has to make some restrictions on the distribution of (X, Y) to get nontrivial rates of convergence of (1.10), which for example can be found in Györfi et al. (2002) and Devroye and Wagner (1980).

A widely accepted restriction is to impose smoothness assumptions on the regression function. To be more precise it was shown in Stone (1982), that for distributions of (X, Y), which satisfy that $X \in [0,1]^d$ *a.s.*, $Y = m(X) + N$, where N is standard normal distributed and independent of X with an (p, C)–smooth regression function m, the lower minimax rate of convergence is $n^{-2p/(2p+d)}$, that is in particular,

$$\liminf_{n\to\infty} \inf_{m_n} \sup_{(X,Y)} C^{-2d/(2p+d)} \cdot n^{2p/(2p+d)} \, \mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) \geq C_1, \quad (1.11)$$

where the minimum is taken with respect to all possible regression estimates, and C_1 is some positive constant independent of C. Roughly speaking, the required (p, C)–smoothness means that all derivates of order p exist, but a detailed definition is given next.

Definition 1.4. *Let $p = k + \beta$ for some $k \in \mathbb{N}_0$ and $0 < \beta \leq 1$ and let $C > 0$. A function $f : [a,b]^d \to \mathbb{R}$ is called $(\mathbf{p}, \mathbf{C})$–**smooth** if for every $\alpha = (\alpha_1, ..., \alpha_d), \alpha_i \in \mathbb{N}_0, \sum_{j=1}^d \alpha_j = k$ the partial derivative*

$$\frac{\partial^k f}{\partial x_1^{\alpha_1} ... \partial x_d^{\alpha_d}}$$

exists and satisfies

$$\left| \frac{\partial^k f}{\partial x_1^{\alpha_1} ... \partial x_d^{\alpha_d}}(x) - \frac{\partial^k f}{\partial x_1^{\alpha_1} ... \partial x_d^{\alpha_d}}(z) \right| \leq C \cdot ||x - z||^{\beta},$$

for all $x, z \in [a,b]^d$.

For further results on the general minimax theory of statistical estimates we refer to Ibragimov and Khasminskii (1980, 1981, 1982) and Birgé (1983).

In order to obtain our rate of convergence results under these assumptions we will use a theorem of Lee, Bartlett and Williamson (1996) (cf. Theorem 1.17). However in order to use this theorem, one has to suppose the boundedness of Y which for example does not hold in the common case that $\mathbf{P}_{Y|X=x}$ is the normal distribution $\mathcal{N}_{(m(x),\sigma)}$. Therefore we will extend the approach of Lee, Bartlett and Williamson

(cf. Section 11.3 of Györfi et al. (2002)) to unbounded data by introducing some new truncation arguments.

This extension enables us to prove rate of convergence results without assuming the boundedness of Y, but suppose instead that the distribution of (X, Y) satisfies a modified Sub-Gaussian condition or, to be more precise, that

$$\mathbf{E}\left(e^{c\cdot|Y|^2}\right) < \infty \tag{1.12}$$

holds for some constant $c > 0$.

Since the analysis of nonparametric regression function estimates, with regard to both consistency and rate of convergence, requires a basic knowledge of the Vapnik-Chervonenkis theory, the next section overviews the accordant results which are necessary for the analysis of our estimates.

1.4. Vapnik-Chervonenkis Theory

The idea of minimizing the empirical risk in the context of decision rules was developed to a great extent by Vapnik and Chervonenkis (1971). They started publishing a series of papers which revolutionised the field of pattern recognition, and therefore affected nonparametric regression estimation strongly, too.

As already mentioned, the theory based on these papers enables us to bound the estimation error or to be more precise to show, for $n \to \infty$,

$$\sup_{f\in\mathcal{F}_n}\left|\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-Y_i|^2 - \mathbf{E}\left(|f(X)-Y|^2\right)\right| \to 0 \quad a.s. \tag{1.13}$$

We can rephrase our goal in a different notation to make this section easier to handle. Let $Z, Z_1, Z_2, \ldots$ be independent and identically distributed random variables with values in $\mathbb{R}^{d+1}$, and let $\mathcal{G}$ denote a class of functions $g : \mathbb{R}^{d+1} \to \mathbb{R}^+$.

Thus we want to derive conditions for

$$\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n}g(Z_i) - \mathbf{E}\left(g(Z)\right)\right| \to 0 \quad (n\to\infty) \quad a.s.$$

Hoeffdings inequality, which is specified in the next lemma, will be an important device in this context.

Lemma 1.5 (Hoeffding (1963)). *Let $X_1, \ldots, X_n$ be independent real-valued random variables, let $a_1, b_1, \ldots, a_n, b_n \in \mathbb{R}$, and assume that $X_i \in [a_i, b_i]$ almost surely for*

all $i = 1, ..., n$. Then, for all $\varepsilon > 0$,

$$\mathbf{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbf{E}(X_i))\right| > \varepsilon\right\} \leq 2 \cdot \exp\left(-\frac{2n\varepsilon^2}{\frac{1}{n}\sum_{i=1}^{n}|b_i - a_i|^2}\right).$$

PROOF. The proof can found in Hoeffding (1963) and moreover, it is given in Devroye, Györfi and Lugosi (1996), Theorem 8.1. □

Obviuosly this inequality from Hoeffding implies for a fixed function $g \in \mathcal{G}$ which is bounded by $B \in \mathbb{R}^+$, that

$$\mathbf{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}(g(Z_i) - \mathbf{E}(g(Z))\right| > \varepsilon\right\} \leq 2 \cdot \exp\left(-\frac{2n\varepsilon^2}{B^2}\right). \tag{1.14}$$

Furthermore, this conclusion can be extended to

$$\mathbf{P}\left\{\sup_{g\in\mathcal{G}_n}\left|\frac{1}{n}\sum_{i=1}^{n}(g(Z_i) - \mathbf{E}(g(Z))\right| > \varepsilon\right\}$$

if one introduces a measure of the complexity of the function space $\mathcal{G}_n$. The massiveness of a class of functions $\mathcal{F}$ can be measured in many ways, but in our context it is reasonable to take so–called L_p–covering numbers, which were suggested in the paper of Kolmogorov and Tikhomirov (1961).

DEFINITION 1.6. *Let $z_1, ..., z_n \in \mathbb{R}^d$ and set $z_1^n = (z_1, ..., z_n)$. Let $\mathcal{G}$ be a set of functions $g : \mathbb{R}^d \to \mathbb{R}$. An L_p-ε-**cover of $\mathcal{G}$ on** z_1^n is a finite set of functions $g_1, ..., g_k : \mathbb{R}^d \to \mathbb{R}$ with the property*

$$\min_{1\leq j\leq k}\left(\frac{1}{n}\sum_{i=1}^{n}|g(z_i) - g_j(z_i)|^p\right)^{1/p} < \varepsilon \qquad \text{for all } g \in \mathcal{G}. \tag{1.15}$$

*The L_p-ε-**covering number** $\mathcal{N}_p(\varepsilon, \mathcal{G}, z_1^n)$ of $\mathcal{G}$ on z_1^n is the minimal size of a L_p-ε-cover of $\mathcal{G}$ on z_1^n. In case that there exists no finite L_p-ε-cover of $\mathcal{G}$ the L_p-ε-covering number of $\mathcal{G}$ on z_1^n is defined by $\mathcal{N}_p(\varepsilon, \mathcal{G}, z_1^n) = \infty$.*

The desirable extension from (1.14) is now given by a general version of Pollards Lemma.

LEMMA 1.7 (Pollards Lemma (1984)). *Let G be a set of functions $g : \mathbb{R}^d \to [0, B]$. For any n, and any $\varepsilon > 0$,*

$$\mathbf{P}\left\{\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n}(g(Z_i) - \mathbf{E}(g(Z))\right| > \varepsilon\right\} \leq 8 \cdot \mathbf{E}\left(\mathcal{N}_1(\varepsilon/8, \mathcal{G}, Z_1^n)\right) \cdot \exp\left(-\frac{n\varepsilon^2}{128B^2}\right).$$

For the sake of completness it should be mentioned that in this lemma $Z, Z_1, \ldots Z_n$ are random variables and hence $\mathcal{N}_1(\varepsilon, \mathcal{G}, Z_1^n)$ is a random variable as well.

PROOF. The proof can be found in Devroye, Györfi and Lugosi (1996), Theorem 29.1. □

Consequently the problem of bounding the estimation error for least squares regression estimates is reduced to find bounds on the covering number of the underlying class of functions, (at least if, for the moment, we overlook that Z is random in the above lemma). For this purpose the concept of L_p packing numbers is very helpful.

DEFINITION 1.8. *Let $z_1, ..., z_n \in \mathbb{R}^d$ and set $z_1^n = (z_1, ..., z_n)$. Let $\mathcal{G}$ be a set of functions $g : \mathbb{R}^d \to \mathbb{R}$. An L_p-ε-**packing of $\mathcal{G}$ on** z_1^n is a finite set of functions $g_1, ..., g_N : \mathbb{R}^d \to \mathbb{R}$, with the property*

$$\left(\frac{1}{n}\sum_{i=1}^{n}|g_k(z_i) - g_j(z_i)|^p\right)^{1/p} \geq \varepsilon, \qquad \text{for all } 1 \leq j < k \leq N. \tag{1.16}$$

*The L_p-ε-**packing number** $\mathcal{M}_p(\varepsilon, \mathcal{G}, z_1^n)$ of $\mathcal{G}$ on z_1^n is the maximal $N \in \mathbb{N}$ such that there exist functions $g_1, ..., g_N \in \mathcal{G}$ with*

$$\left(\frac{1}{n}\sum_{i=1}^{n}|g_k(z_i) - g_j(z_i)|^p\right)^{1/p} \geq \varepsilon$$

for all $1 \leq j < k \leq N$. Take $\mathcal{M}_p(\varepsilon, \mathcal{G}, z_1^n) = \infty$, if there exists a L_p-ε-packing of $\mathcal{G}$ on z_1^n of size N for every $N \in \mathbb{N}$.

L_p–covering numbers and L_p–packing numbers are closely related to each other, as the next Lemma shows.

LEMMA 1.9. *Let $z_1, ..., z_n \in \mathbb{R}^d$ and set $z_1^n = (z_1, ..., z_n)$. Let $\mathcal{G}$ be a set of functions $g : \mathbb{R}^d \to \mathbb{R}$, $p \geq 1$ and let $\varepsilon > 0$. Then*

$$\mathcal{M}_p(2\varepsilon, \mathcal{G}, z_1^n) \leq \mathcal{N}_p(\varepsilon, \mathcal{G}, z_1^n) \leq \mathcal{M}_p(\varepsilon, \mathcal{G}, z_1^n).$$

PROOF. Even though the proof is quite simple, we want to refer to the proof of Lemma 9.2. in Györfi et al. (2002). □

Hence, with this result it makes no difference if one has bounds on packing or covering numbers for a class of functions, owing to the close relationship of both. However, it is usually easier to bound the packing numbers and therefore, the following definition is needed.

DEFINITION 1.10. *Let $\mathcal{G}$ be a class of subsets of $\mathbb{R}^d$ with $\mathcal{G} \neq \emptyset$, and let $n \in \mathbb{N}$. For a set $A \subset \mathbb{R}^d$ with $|A| = n$, one says $\mathcal{G}$ **shatters** A if each subset of A can be represented in the form $G \cap A$, for some $G \in \mathcal{G}$. The **Vapnik-Chervonenkis dimension** (VC dimension) $V_{\mathcal{G}}$ of $\mathcal{G}$ is the largest interger n such that there exists a set of n points in $\mathbb{R}^d$ which can be shattered by $\mathcal{G}$.*

The next theorem will give an upper bound on the packing number of relatively general classes of functions. But before we can state this result we need one more notation. For a class of functions $\mathcal{G}$ with elements $g : \mathbb{R}^d \to \mathbb{R}$, we define

$$\mathcal{G}^+ := \Big\{\{(z,t) \in \mathbb{R}^d \times \mathbb{R}; t \leq g(z)\}; g \in \mathcal{G}\Big\}$$

the set of all subgraphs of functions of $\mathcal{G}$.

THEOREM 1.11. *Let $\mathcal{G}$ be a class of functions $g : \mathbb{R}^d \to [0, B]$ with $V_{\mathcal{G}^+} \geq 2$, let $p \geq 1$ and $0 < \varepsilon < B/4$. Then*

$$\mathcal{M}_p(\varepsilon, \mathcal{G}, z_1^n) \leq 3\left(\frac{2eB^p}{\varepsilon^p}\log\left(\frac{3eB^p}{\varepsilon^p}\right)\right)^{V_{\mathcal{G}^+}},$$

for all $z_1^n = (z_1, ..., z_n)$ with $z_1, ..., z_n \in \mathbb{R}^d$.

PROOF. This theorem goes back to Haussler (1992), who proved this inequality for $p = 1$. A general proof can be found again in Györfi et al. (2002), Theorem 9.4. □

Now in order to get a bound on the covering number of a certain class of bounded functions, it suffices to find an upper bound on the VC dimension $V_{\mathcal{G}^+}$. The following theorem, which goes back to Steele (1975) and Dudley (1978), will give exactly such a bound in the case that $\mathcal{G}$ is a linear vector space with finite dimension.

THEOREM 1.12. *Let $\mathcal{G}$ be an r-dimensional vector space of functions $g : \mathbb{R}^d \to \mathbb{R}$ and set*

$$\mathcal{A} = \Big\{\{z : g(z) \geq 0\} : g \in \mathcal{G}\Big\}.$$

Then $V_{\mathcal{A}} \leq r$.

PROOF. Amongst others a proof can be found in Devroye, Györfi and Lugosi (1996), Theorem 13.9. □

In Chapters 4 and 5 we shall consider estimates, which are defined as a sum of certain functions, and hence we also need bounds on the covering numbers of such function spaces. For this purpose, we define the class of functions,

$$\mathcal{F} \oplus \mathcal{G} = \Big\{h : \mathbb{R}^d \to \mathbb{R},\ h(x) = f(x) + g(x),\ \text{ for some } f \in \mathcal{F} \text{ and } g \in \mathcal{G}\Big\},$$

for classes $\mathcal{F}$ and $\mathcal{G}$ of functions $f : \mathbb{R}^d \to \mathbb{R}$.

As presumably expected there exists a connection between the covering number of $\mathcal{F} \oplus \mathcal{G}$ and the covering numbers of $\mathcal{F}$ and $\mathcal{G}$, and this connection is given in the following lemma. Similar results can be found in Nobel (1992), Nolan and Pollard (1987) and Pollard (1990).

LEMMA 1.13. *Let $\mathcal{F}$ and $\mathcal{G}$ be two families of functions $f : \mathbb{R}^d \to \mathbb{R}$. Then, for $\varepsilon, \delta > 0$, we have*

$$\mathcal{N}_1(\varepsilon + \delta, \mathcal{F} \oplus \mathcal{G}, z_1^n) \leq \mathcal{N}_1(\varepsilon, \mathcal{F}, z_1^n) \cdot \mathcal{N}_1(\delta, \mathcal{G}, z_1^n).$$

PROOF. For the proof we refer to the proof of Theorem 29.6 in Devroye, Györfi and Lugosi (1996). □

Now, we have collected all tools from VC-theory needed in this dissertation. Since we try to make this work self-contained the next section provides a couple of different results we will need during the analyse of our estimates and which are not linked up closely.

1.5. Auxiliary Results

Firstly we want to refer to the inequation from Bernstein, which is closely related to Hoeffding's inequality and typically can outperform it if the underlying random variables have a small variance.

LEMMA 1.14 (Bernstein (1946)). *Let $X_1, ..., X_n$ be independent real–valued random variables, let $a, b \in \mathbb{R}$ with $a < b$, and assume that $X_i \in [a, b]$ a.s. $(i = 1, ..., n)$. Let*

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} \mathbf{Var}(X_i) > 0.$$

Then, for all $\varepsilon > 0$,

$$\mathbf{P}\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} (X_i - \mathbf{E}(X_i)) \right| > \varepsilon \right\} \leq 2 \cdot \exp\left(- \frac{n\varepsilon^2}{2\sigma^2 + 2\varepsilon(b-a)/3} \right)$$

Therefore, Bernstein's inequality kicks in when ε is larger than about

$$\max\left\{ \sigma/\sqrt{n}, (b-a)/\sqrt{n} \right\},$$

and it is typically stronger than Hoeffding's inequality if $\sigma \ll b - a$.

In order to prove universal consistency of nonparametric regression estimates, one usually proves consistency for continuous regression functions (or for infinitely often continuously differentiable regression functions) first, and then extends this result to arbitrary functions. For this purpose one needs the following general denseness result.

LEMMA 1.15. *For any $p \geq 1$ and any probability measure μ, the set of continuous functions of bounded support is dense in $L_p(\mu)$, that is, for any $f \in L_p(\mu)$ and $\varepsilon > 0$ there exists a continuous function g with compact support such that*

$$\int |f(x) - g(x)|^p \mu(dx) \leq \varepsilon.$$

PROOF. A proof can be found in Elstrodt (1996) (cf. Theorem 2.31). □

Note that Lemma 1.15 involves directly the denseness of $C_0^\infty(\mathbb{R}^d)$ (set of all infinitely often continuously differentiable functions with bounded support) in $L_2(\mu)$, due to the well-known fact that $C_0^\infty(\mathbb{R}^d)$ is dense in $C_0(\mathbb{R}^d)$ (the set of continuous functions with bounded support). Therefore it is completely sufficient to prove consistency only for regression functions which are contained in $C_0^\infty(\mathbb{R}^d)$ (or supersets) in order to obtain universal consistency.

Since the class of functions we consider in this dissertation is closely related to the class of linear spline functions (cf. Lemma 2.2), we can use an approximation result from Schumaker (1981) in order to get bounds on the approximation error of our estimate and therefore to derive the desired rate of convergence.

LEMMA 1.16. *Let $a, b \in \mathbb{R}, a < b$ and $h : [a,b]^d \to \mathbb{R}$ be a (p,C)–smooth function, for some $0 < p \leq 2, C > 1$. Furthermore, let $\mathcal{G}^d$ denote the set of all continuous piecewise linear functions $g : [a,b]^d \to \mathbb{R}$, with respect to a partition of $[a,b]^d$ in n equivolume cubes. Then,*

$$\inf_{g \in \mathcal{G}^d} \left(\max_{x \in [a,b]^d} |g(x) - h(x)|^2 \right) \leq c_1 \cdot C^2 \cdot n^{-2p/d},$$

holds, for a sufficiently large constant $c_1 > 0$.

PROOF. This result is a consequence of Theorem 12.8, Example 13.27 and inequality (13.62) in Schumaker (1981). □

Furthermore we have already pointed out the impact of the approach of Lee, Bartlett and Williamson. Hence, a very important instrument, in order to derive our rate of convergence results, is the corresponding theorem from Lee, Bartlett and Williamson.

THEOREM 1.17. *Assume $|Y| \leq B$ a.s. and $B \geq 1$. Let $\mathcal{F}$ be a class of functions $f : \mathbb{R}^d \to \mathbb{R}$ and let $|f(x)| \leq B$. Then for each $n \geq 1$,*

$$\begin{aligned}\mathbf{P}\Bigg\{\exists f \in \mathcal{F} : \mathbf{E}\left(|f(X) - Y|^2\right) - \mathbf{E}\left(|m(X) - Y|^2\right) \\ - \frac{1}{n}\sum_{i=1}^{n}\left(|f(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2\right) \\ \geq \varepsilon \cdot \left(\alpha + \beta + \mathbf{E}\left(|f(X) - Y|^2\right) - \mathbf{E}\left(|m(X) - Y|^2\right)\right)\Bigg\} \\ \leq 14 \sup_{x_1^n} \mathcal{N}_1\left(\frac{\beta\varepsilon}{20B}, \mathcal{F}, x_1^n\right) \exp\left(-\frac{\varepsilon^2(1-\varepsilon)\alpha n}{214(1+\varepsilon)B^4}\right),\end{aligned}$$

where $\alpha, \beta > 0$ and $0 < \varepsilon \leq 1/2$.

PROOF. This theorem was proved in Lee, Bartlett and Williamson (1996) and another proof can be found in Györfi et al. (2002). □

We want to complete this section by stating the well-known Borel-Cantelli lemma.

LEMMA 1.18 (Borel-Cantelli lemma). *Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of events in some probability space $(\Omega, \mathcal{A}, \mathbf{P})$. If the sum of the probabilities of the A_n is finite, that is*

$$\sum_{i=1}^{\infty} \mathbf{P}(A_n) < \infty,$$

then

$$\mathbf{P}(\limsup_{n\to\infty} A_n) = 0.$$

PROOF. A proof can be found for example in Billingsley (1995), Theorem 4.3. □

In this chapter we have motivated the use of least squares estimates in regression estimation problems. Furthermore, we gave some basic ideas how one can analyse such estimates in terms of universal consistency and their corrsponding rate of convergence, and we provided all important tools we shall need during this analyse.

CHAPTER 2

Maxima of Minima of Linear Functions

In this chapter we will introduce the class of functions standing in the centre of our attention troughout this thesis. This class consists of maxima of minima of linear functions, where linear means actually affine linear. The first section contains the definition of this class of functions as well as the definition of the estimate we will analyse in Chapter 3. In Section 2.2, we examine how these functions are generated and discuss some of their properties. Furthermore we formulate some helpful relations to linear spline functions, but also point out some inconveniences in this context. After that, Section 2.3 discusses bounds on the covering numbers of a truncated version of the function space consisting of maxima of minima of linear functions and therefore provides implicit bounds on the estimation error of least squares estimates over these functions.

2.1. Definition of the Estimate

In the sequel we will use the principle of least squares to fit maxima of minima of linear functions to the data. More precisely, let $K_n \in \mathbb{N}$ and $L_{1,n}, \dots, L_{K_n,n} \in \mathbb{N}$ be parameters of the estimate and set

$$\mathcal{F}_n = \Bigg\{ f : \mathbb{R}^d \to \mathbb{R} \,:\, f(x) = \max_{k=1,\dots,K_n} \min_{l=1,\dots,L_{k,n}} (a_{k,l} \cdot x + b_{k,l}) \quad (x \in \mathbb{R}^d), \\ \text{for some } a_{k,l} \in \mathbb{R}^d, b_{k,l} \in \mathbb{R} \Bigg\}, \tag{2.1}$$

where

$$a_{k,l} \cdot x = a_{k,l}^{(1)} \cdot x^{(1)} + \cdots + a_{k,l}^{(d)} \cdot x^{(d)}$$

denotes the scalar product between the vectors $a_{k,l} = (a_{k,l}^{(1)}, \dots, a_{k,l}^{(d)})^T$ and $x = (x^{(1)}, \dots, x^{(d)})^T$. From now on $\mathcal{F}_n$ denotes the class of functions defined by (2.1), where $K_n, L_{1,1}, ..., L_{K_n,n}$ are parameters depending on n. Note that the class $\mathcal{F}_n$ therefore depends on n as well and that we have inclusions of the form

$$\mathcal{F}_m \subset \mathcal{F}_n, \quad \text{if } K_m \leq K_n \text{ and } L_{i,m} \leq L_{i,n} \text{ for all } 1 \leq i \leq K_m.$$

Furthermore, for fixed $K, L \in \mathbb{N}$, we define

$$\mathcal{F}_{K,L} = \Bigg\{ f : \mathbb{R}^d \to \mathbb{R} \,:\, f(x) = \max_{k=1,\dots,K} \min_{l=1,\dots,L} (a_{k,l} \cdot x + b_{k,l}) \quad (x \in \mathbb{R}^d), \\ \text{for some } a_{k,l} \in \mathbb{R}^d, b_{k,l} \in \mathbb{R} \Bigg\},$$

corresponding to $\mathcal{F}_n$, but for $L_{1,n} = ... = L_{K,n}$.

Now for a given set of data $\mathcal{D}_n = \{(X_1, Y_1), ..., (X_n, Y_n)\}$ we define an estimate $\tilde{m}_n$ by

$$\tilde{m}_n(\cdot) \quad = \quad \arg\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2. \tag{2.2}$$

Here we assume again that the minimum exists, however we do not require that it is unique. Since the elements of $\mathcal{F}_n$ (which are referred to as *maxmin functions* in the following) are unbounded in general, also the estimate $\tilde{m}_n$ may happen to be unbounded. Hence we consider the truncated version of this least squares estimate, that is,

$$m_n = T_{\beta_n} \circ \tilde{m}_n, \text{ where } T_{\beta_n}(z) = \begin{cases} \beta_n & z > \beta_n, \\ z & -\beta_n \le z \le \beta_n, \\ -\beta_n & z < -\beta_n \end{cases} \tag{2.3}$$

for some $\beta_n \in \mathbb{R}_+$. This truncation provides a bounded estimate and therefore allows us to use results from VC theory in order to obtain bounds on the estimation error, although we actually use the principle of least squares with respect to unbounded functions. As can be seen in (2.2), we choose $\mathcal{F}_n$ dependent on the sample size or, in other words, we have to choose the parameters somehow. Later we shall see, how one can do this choice data-dependent. For now we just mention, that the size of the parameters will grow with growing sample size.

Before we take a more detailed look at the underlying class of functions $\mathcal{F}_n$, we want to give an overview of earlier appearances of the function class $\mathcal{F}_n$ in the literature.

To our knowledge, the use of maxima of minima of linear functions in order to represent continuous piecewise linear functions goes back to Bartels, Kuntz and Sholtes (1995). In 2003 Bagirov, Rubinov, Soukhoroukova and Yearwood have used maxima of minima of linear functions in connection with pattern recognition. In this context the class $\mathcal{F}_n$ stands out as a rather complex and highly flexible class of functions. Furthermore maxima of minima of linear functions have been used in regression estimation already by Beliakov and Kohler (2005). There, least squares estimates are derived by minimizing the empirical L_2 risk over classes of

functions consisting of Lipschitz smooth functions where a bound on the Lipschitz constant is given. It is shown that the resulting estimate in fact is a maxmin function, where the number of minima occurring in the maxima is equal to the sample size. Additional restrictions (for example on the linear functions in the minima) ensure that no overfitting can happen. In contrast, the number of linear functions we consider in this dissertation is much smaller, and restrictions on these linear functions are therefore not necessary. This seems to be promising, because we do not fit too many parameters to the data.

2.2. Characterisation of $\mathcal{F}_n$

Maxima of minima of linear functions are in fact continuous piecewise linear functions with respect to partitions of finite size, and this size only depends on the number of linear functions, which induce the maxmin function. Hence using the principle of least squares with respect to the class of functions consisting of maxmin functions corresponds with fitting linear spline functions with free knots to the data.

As seen in the above definition $\mathcal{F}_n$ depends on the parameters

$$K_n, L_{1,n}, \ldots, L_{K_n,n} \in \mathbb{N}.$$

$L_{i,n}$ declares the number of linear functions under the i-th minimum, and K_n declares the number of minima functions under the maximum. To make this a bit more perspicuous, we want to give an example in the univariate case $d = 1$.

EXAMPLE 2.1. *As parameters we choose $K = 3$ and $L_1, L_2, L_3 = 2$ and thus consider the class $\mathcal{F}_{3,2}$. Figure 1 shows two linear functions with their minimum function (bold).*

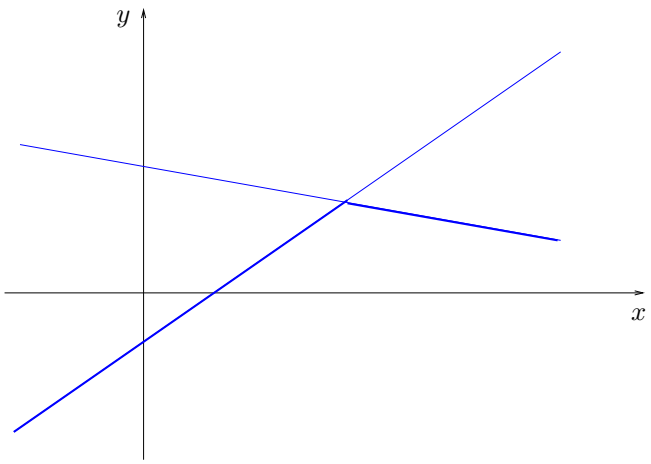

FIGURE 1. Two linear functions with their minimum function.

Due to the choice $K = 3$ we need two more minimum functions under the maximum to get a function that fulfils the definition, and under each of them we need two linear functions. Figure 2 shows four linear functions more together with their two minimum functions.

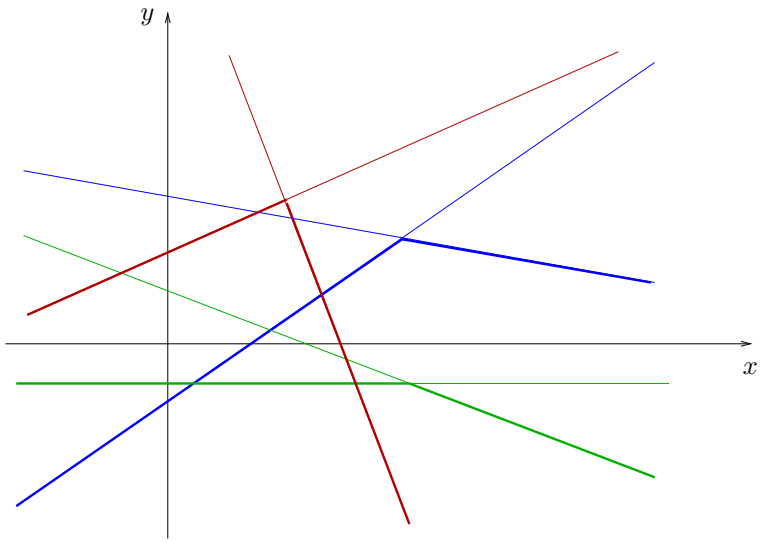

FIGURE 2. Three minima functions with their generating linear functions.

In Figure 3 we can see the resulting maxmin function (which belongs to $\mathcal{F}_{3,2}$) with its three generating minimum functions, but without the six underlying linear functions. Apparently the same linear functions could induce a different maxmin function, because the generated function obviously depends on how one chooses the pairs of linear functions, which belong to the same minimum function.

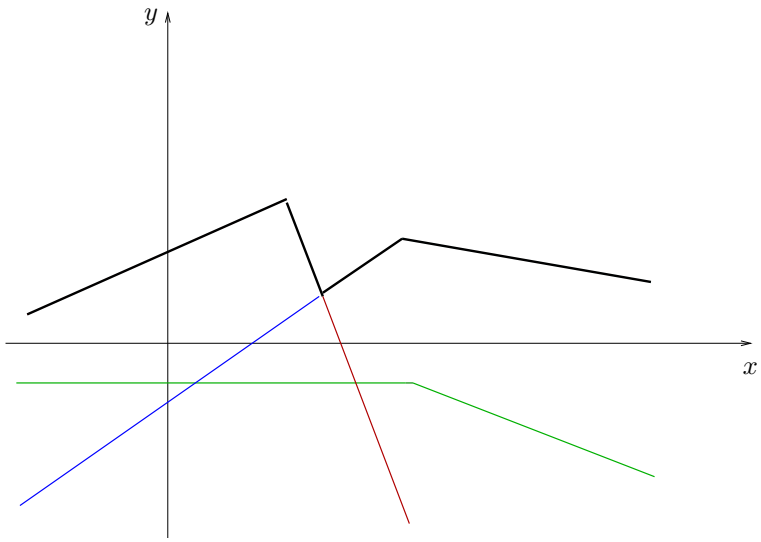

FIGURE 3. The resulting maximum function (black).

At least for the univariate case this short example should have suggested that maxmin functions (functions that can be constructed as maxima of minima of linear functions) are continuous piecewise linear. In the multivariate case the continuity also results directly from the fact that both the maximum and the minimum function are continuous, which shows that maxmin functions are just compositions of continuous functions. The piecewise linearity is self-evident, because maxmin functions are induced by linear functions.

The next lemma shows a connection in the opposite direction. It demonstrates how linear spline functions can be interpolated by maxmin functions with parameters depending on the size of the partition belonging to the spline and on the dimension of X.

LEMMA 2.2. *Let $K_n \in \mathbb{N}$ and let Π be a partition of $[a,b]^d$ consisting of K_n rectangulars. Assume that $f^{lin} : [a,b]^d \to \mathbb{R}$ is a piecewise linear function with respect to Π and assume that f^{lin} is continuous. Furthermore let $x_1, ..., x_n \in \mathbb{R}^d$ be n fixed points in $[a,b]^d$. Then there exist linear functions*

$$f_{1,0}, ..., f_{1,2d}, ..., f_{K_n,0}, ..., f_{K_n,2d} : \mathbb{R}^d \to \mathbb{R},$$

such that

$$f^{lin}(z) = \max_{i=1,...,K_n} \min_{k=0,..,2d} f_{i,k}(z), \qquad \text{for all } z \in \{x_1, ..., x_n\}.$$

PROOF. Since f^{lin} is a continuous piecewise linear function, it is of the shape

$$f^{lin}(z) = \sum_{i=1}^{K_n} f_i^{lin}(z) \cdot \mathbf{1}_{A_i} = \sum_{i=1}^{K_n} \left(\sum_{j=1}^{d} \alpha_{i,j} \cdot z^{(j)} + \alpha_{i,0} \right) \cdot \mathbf{1}_{A_i},$$

for suitable constants $\alpha_{i,j} \in \mathbb{R}$ $(i = 1, ..., K_n, j = 0, ..., d)$, and moreover $\Pi = \{A_1, ..., A_{K_n}\}$ is a partition of $[a,b]^d$ with

$$A_i = I_i^{(1)} \times \cdots \times I_i^{(d)},$$

for some univariate intervals $I_i^{(j)}$ $(i = 1, \ldots, K_n)$. We denote the left and the right endpoint of $I_i^{(j)}$ by $a_{i,j}$ and $b_{i,j}$, respectively, that is,

$$I_i^{(j)} = [a_{i,j}, b_{i,j}) \quad \text{or} \quad I_i^{(j)} = [a_{i,j}, b_{i,j}].$$

This choice is without restriction of any kind because f^{lin} is continuous by assumption. Now we choose, for every $i \in \{1, ..., K_n\}$,

$$f_{i,0}(x) = f_i^{lin}(x) = \sum_{j=1}^{d} \alpha_{i,j} \cdot x^{(j)} + \alpha_{i,0}.$$

This implies that $f_{i,0}$ and the given piecewise polynomial f^{lin} coincide on A_i for every $i = 1, ..., K_n$. Furthermore, for $i = 1, ..., K_n$ and $j = 1, ..., d$, we define

$$f_{i,2j-1}(x) = f_i^{lin}(x) + (x^{(j)} - a_{i,j}) \cdot \beta_{i,j},$$

where $\beta_{i,j} \geq 0$ is such that

$$f_{i,2j-1}(z) \leq f^{lin}(z),$$

for all $z = (z^{(1)}, ..., z^{(d)}) \in \{x_1, ..., x_n\}$ satisfying $z^{(j)} < a_{i,j}$ and

$$f_{i,2j-1}(z) \geq f^{lin}(z),$$

for all $z = (z^{(1)}, ..., z^{(d)}) \in \{x_1, ..., x_n\}$ satisfying $z^{(j)} > a_{i,j}$.

The above conditions are satisfied in particular, if

$$\beta_{i,j} \geq \max_{k=1,...,n; x_k^{(j)} \neq a_{i,j}} \frac{f^{lin}(x_k) - f_i^{lin}(x_k)}{x_k^{(j)} - a_{i,j}},$$

and obviously, for $z^{(j)} = a_{i,j}$, we have$f_{i,2j-1}(z) = f_i^{lin}(z)$.

Analogously we choose

$$f_{i,2j}(x) = f_i^{lin}(x) - (x^{(j)} - b_{i,j}) \cdot \gamma_{i,j},$$

where $\gamma_{i,j} \geq 0$ is such that

$$f_{i,2j}(z) \geq f^{lin}(z),$$

for all $z = (z^{(1)}, ..., z^{(d)}) \in \{x_1, ..., x_n\}$ satisfying $z^{(j)} < b_{i,j}$ and

$$f_{i,2j}(z) \leq f^{lin}(z),$$

for all $z = (z^{(1)}, ..., z^{(d)}) \in \{x_1, ..., x_n\}$ satisfying $z^{(j)} > b_{i,j}$.

In this case the conditions from above are satisfied, if

$$\gamma_{i,j} \quad \geq \quad \max_{k=1,...,n; x_k^{(j)} \neq a_{i,j}} \frac{f_i^{lin}(x_k) - f^{lin}(x_k)}{x_k^{(j)} - b_{i,j}}.$$

From this choice of the functions $f_{i,j}$ $(i = 1, ..., K_n)$, $(j = 0, ..., 2d)$ results directly, that

$$\min_{k=0,..,2d} f_{i,k}(z) \begin{cases} = & f_i^{lin}(z) = f^{lin}(z) \qquad \text{for } z \in A_i \cap \{x_1, ..., x_n\} \\ \leq & f^{lin}(z) \qquad \text{for } z \in \{x_1, ..., x_n\} \end{cases}$$

holds for all $i = 1, ..., K_n$, which implies the assertion. □

In the course of this dissertation we shall see that this connection between maxmin functions and continuous piecewise linear functions ease the bounding of the approximation error of our estimate, because we can use well-known approximation results from spline theory.

Although the above lemma is the only result concerning the connection of maxmin functions and linear spline functions needed in this thesis, we want to make a view remarks on the correlation of these two function spaces, and we hope these remarks will support the comprehension and figuring of maxmin functions.

First, it should be mentioned that in general maxmin functions are not necessarily piecewise linear with respect to a partition that consists of rectangulars. Usually, the underlying partitions are not of this form, because the intersection of two graphs of linear functions $f, g : \mathbb{R}^d \to \mathbb{R}$ is a hyperplane of at most dimension d that lies arbitrarily in the $d+1$–dimensional space. Since exactly such intersections induce the partition (owing to the continuity of maxmin functions), it is clear that maxmin functions are piecewise linear with respect to arbitrary partitions of the underlying space and not necessarily to partitions consisting of rectangulars.

Secondly, it is remarkable that there exists no simple connection between the number of knots of a linear spline and the number of parameters needed to express splines with this certain number of knots. Although it is clear that there exists a class of functions $\mathcal{F}_n$ which contains all linear spline functions with a certain number of knots, the number of the parameters needed will be comparatively large. Actually it will be so large that it also contains spline functions with a mutiple of the number of given knots. The next example is supposed to sample this challenge in view of Lemma 2.2 and in the case $d = 1$.

EXAMPLE 2.3. *Firstly we consider the function*

$$f(x) \quad = \quad \begin{cases} 3/4 - 2x =: f_1(x); & x \in \big[0, 1/4\big) \\ 1/2 - x =: f_2(x); & x \in \big[1/4, 1/2\big) \\ -1/2 + x =: f_3(x); & x \in \big[1/2, 3/4\big) \\ -5/4 + 2x =: f_4(x); & x \in \big[3/4, 1\big], \end{cases}$$

which is obviously a piecewise linear function with respect to the partition

$$\Pi = \left\{ \left[0 \, , \, \frac{1}{4} \right), \left[\frac{1}{4} \, , \, \frac{1}{2} \right), \left[\frac{1}{2} \, , \, \frac{3}{4} \right), \left[\frac{3}{4} \, , \, 1 \right] \right\} \tag{2.4}$$

of the interval $[0, 1]$. *It is easy to see that* f *is also continuous, since we have that*

$$f_i(i/4) = f_{i+1}(i/4) \qquad (i = 1, 2, 3).$$

Moreover we can deduce from Figure 4 that f *is a convex function and we shall see that usually a large number of minimum functions under the maximum is necessary to induce a convex spline function by maxmin functions. This property is caused by the concave structure of minimum functions.*

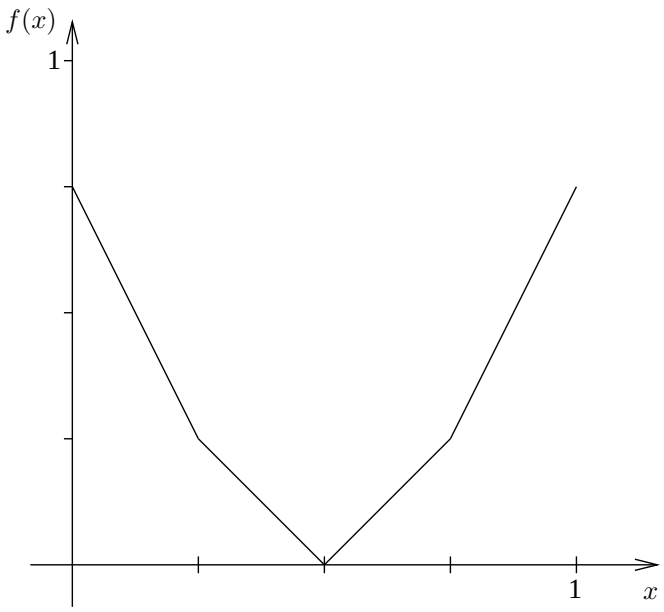

FIGURE 4. The convex spline f

Apparently we can rewrite f as a maxmin function, namely

$$f(x) = \max\big\{\min\{f_1(x)\}, \min\{f_2(x)\}, \min\{f_3(x)\}, \min\{f_4(x)\}\big\},$$

and therefore f is obviously contained in $\mathcal{F}_{4,1}$. This representation shows that f is actually not a maxmin function but a maximum function (cf. Figure 5). Hence, in

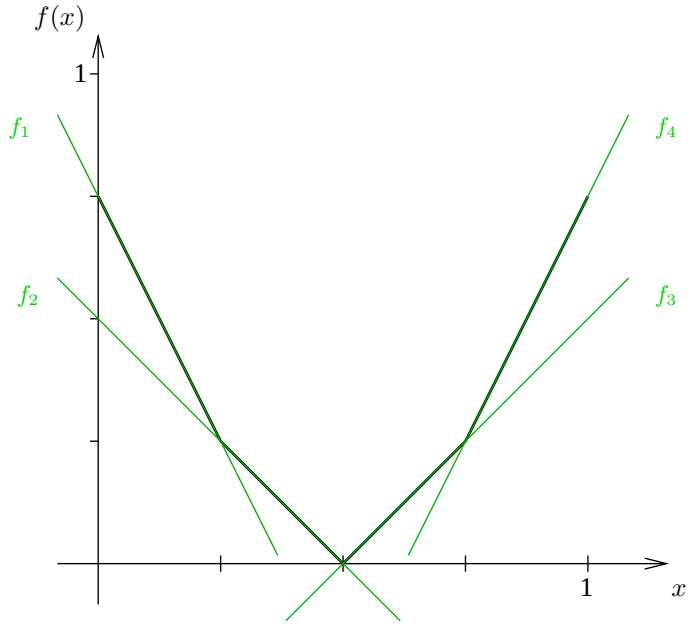

FIGURE 5. f as maxmin function

view of Lemma 2.2, the class of functions $\mathcal{F}_{4,1}$ instead of $\mathcal{F}_{4,3}$ would be sufficient

in order to generate the convex spline function f. Note that we can induce an arbitrary convex spline with respect to a partition of $[0,1]$ into 4 intervals in a similar way. However, on the other hand we actually need 4 minimum functions under the maximum to represent f as a maxmin function. Indeed, this results from the fact that the points $i/4$ $(i = 1,2,3)$ need to be intersection points of the functions under the maximum, since otherwise we would not be able to generate the convex shape in the neighbourhood of these points.

Therefore, we can infer that a class of maxmin functions that contains all continuous piecewise linear functions with respect to a partition of $[0,1]$ into at least 4 subintervals need to be at least $\mathcal{F}_{4,1}$.

Let us now consider the concave function

$$g(x) \quad = \quad \begin{cases} 1/4 + 2x =: g_1(x) & x \in \big[0, 1/4\big) \\ 1/2 + x =: g_2(x) & x \in \big[1/4, 1/2\big) \\ 3/2 - x =: g_3(x) & x \in \big[1/2, 3/4\big) \\ 9/4 - 2x =: g_4(x) & x \in \big[3/4, 1\big], \end{cases}$$

which is sketched in Figure 6. Obviously g is piecewise linear with respect to the

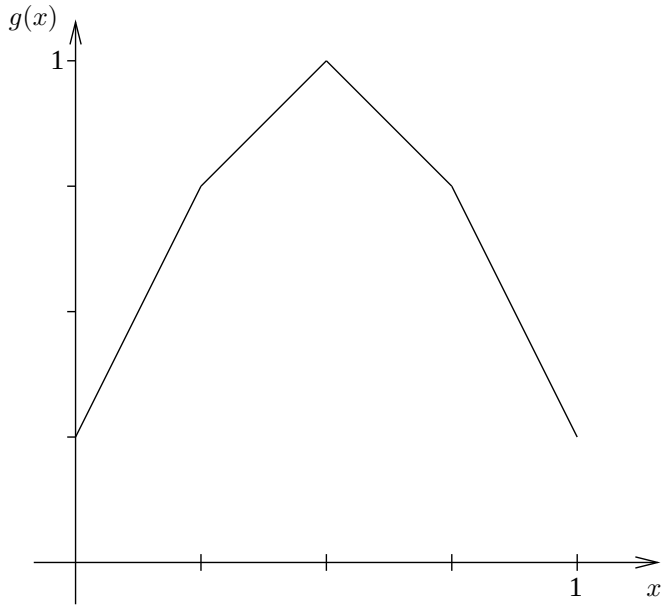

FIGURE 6. The concave spline g

partition Π defined in (2.4) and its continuity can be deduced again by verifying

$$g_i(i/4) = g_{i+1}(i/4) \qquad (i = 1,2,3).$$

Moreover it is easy to see that we can rewrite g as maxmin function, that is

$$g(x) = \max\big\{\min\{g_1(x), g_2(x), g_5(x)\}, \min\{g_3(x), g_4(x), g_6(x)\}\big\},$$

with $g_5 : [0,1] \to \mathbb{R}, x \mapsto 2 - 2x$, *and* $g_6 : [0,1] \to \mathbb{R}, x \mapsto 2x$ *(cf. Figure 7). Note that we could also choose different functions for* g_5 *and* g_6 *in order to induce the same spline function* g. *But we have to choose* g_5 *such that it goes through the point* $(1/2, 1)$ *and such that its slope is smaller than or equal to the slopes of* g_3 *and* g_4. *Because otherwise either we would obtain an additional knot in the intersection point of* g_5 *and* g_4, *or the resulting maxmin function would just be exactly the first minimum function, that is* $\min\{g_1(x), g_2(x), g_5(x)\}$. *Moreover* g_6 *must be chosen*

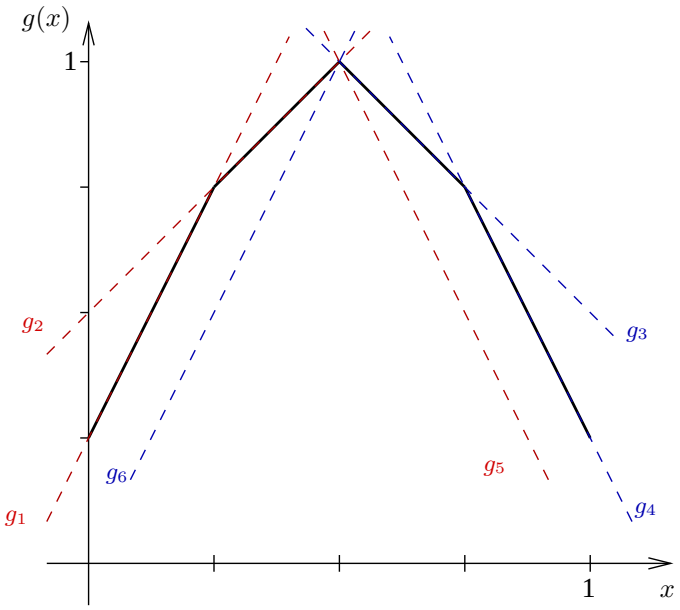

FIGURE 7. g as maxmin function

such that it also goes through the point $(1/2, 1)$ *but in such a way that its slope is greater than or equal to the slopes of* g_1 *and* g_2, *because otherwise we would also obtain a different spline function (which would not necessarily be concave anymore).*

Now, we have seen that g *is contained in* $\mathcal{F}_{2,3}$ *and moreover, one can actually justify that it is necessary to have at least* 3 *linear functions under the minimum. If one tries to induce* g *by using only functions of the form* $\min\{ax+b, cx+d\}$, *that is functions from* $\mathcal{F}_{n,2}$ *for some* $n \in \mathbb{N}$, *one will see directly that it is impossible, because the functions under the minimum need to have at least two knots in order to induce a concave spline function.*

In summary, we need on the one hand a superset of $\mathcal{F}_{4,1}$ *to be able to represent convex linear spline functions with respect to the partition* Π, *and on the other hand we need a superset of* $\mathcal{F}_{2,3}$ *in order to provide all concave linear spline functions with respect to* Π. *From this we can infer that the class of maxmin functions* $\mathcal{F}_{4,3}$ *is*

the 'smallest' class of the form $\mathcal{F}_{m,n}$ which could contain all continuous piecewise linear functions with respect to an arbitrary partition of $[0,1]$ into 4 subintervals.

Furthermore in the case $d = 1$, one can show that $\mathcal{F}_{4,3}$ actually contains all such linear spline functions. This can be deduced from the construction in the proof of Lemma 2.2. It is possible to choose the real numbers $\beta_{i,j}$ and $\gamma_{i,j}$ in that proof in such a way that the corresponding construction provides the desired maxmin function. Since moreover the above arguments concerning the concave and convex splines can be easily extended to partitions of $[0,1]$ into N subintervals, we can infer that $\mathcal{F}_{N,3}$ is the smallest class of maxmin functions which contains all continuous piecewise linear functions with respect to a partition of $[0,1]$ into N subintervals.

Let us complete this example by demonstrating that the class $\mathcal{F}_{4,3}$ also contains spline functions with more than three knots. Figure 8 sketches a maxmin function

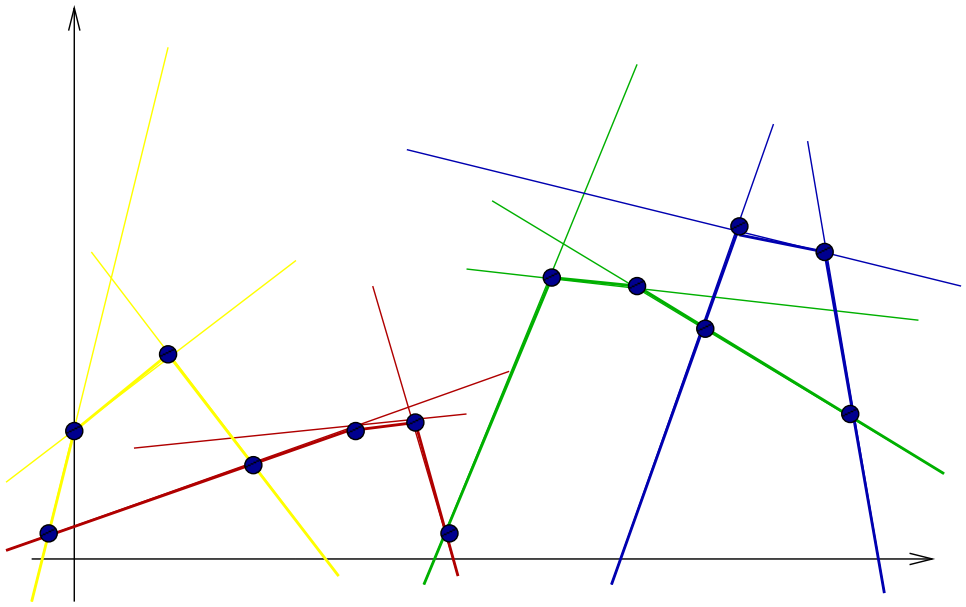

FIGURE 8. A member of $\mathcal{F}_{4,3}$ with 13 knots

which has 13 knots and which is contained in $\mathcal{F}_{4,3}$. We have marked those linear functions with the same colour, which belong to the same minimum function. Hence we have four different colours, each with three belonging linear functions. The resulting minimum functions are shown in the same colour but bold.

From Sections 1.3 and 1.4 we can appraise the importance of covering numbers of $\mathcal{F}_n$ in conjunction with consistency and the rate of convergence of least squares estimates over $\mathcal{F}_n$. The next section will provide results referring to this.

2.3. Covering Numbers of $\mathcal{F}_n$

To obtain bounds on the covering numbers of sets of maxima of minima of linear functions we first show a connection from the L_p-ε-covering numbers of sets $\mathcal{G}_1, \mathcal{G}_2, ...$ and the L_p-ε-covering number of their maximum

$$\max\{\mathcal{G}_1, ..., \mathcal{G}_m\} = \Big\{ f : \mathbb{R}^d \to \mathbb{R}; f(x) = \max\{g_1(x), ..., g_m(x)\} \quad (x \in \mathbb{R}^d), \\ \text{for some } g_1 \in \mathcal{G}_1, ..., g_m \in \mathcal{G}_m \Big\}$$

and minimum (defined analogously), respectively.

LEMMA 2.4. *Let $\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_m$ be classes of functions $g : \mathbb{R}^d \to \mathbb{R}$ and suppose we have given n points $x_1^n = (x_1, ..., x_n) \in \mathbb{R}^d \times \cdots \times \mathbb{R}^d$ in $\mathbb{R}^d$. Then*

$$\mathcal{N}_p\left(\varepsilon, \max\left\{\mathcal{G}_1, ..., \mathcal{G}_m\right\}, x_1^n\right) \leq \prod_{i=1}^{m} \mathcal{N}_p\left(\frac{\varepsilon}{m^{1/p}}, \mathcal{G}_i, x_1^n\right) \tag{2.5}$$

and

$$\mathcal{N}_p\left(\varepsilon, \min\left\{\mathcal{G}_1, ..., \mathcal{G}_m\right\}, x_1^n\right) \leq \prod_{i=1}^{m} \mathcal{N}_p\left(\frac{\varepsilon}{m^{1/p}}, \mathcal{G}_i, x_1^n\right) \tag{2.6}$$

hold for all $\varepsilon > 0$.

PROOF. Let $x_1^n = (x_1, ..., x_n)$ and $\varepsilon > 0$ be fixed. Without loss of generality we assume that $\mathcal{N}_p(\varepsilon/(m^{1/p}), \mathcal{G}_i, x_i^n)$ is finite for all $1 \leq i \leq m$. That is, for every set $\mathcal{G}_i$, one can choose a finite set of functions $g_i^1, ..., g_i^{n_i}$, such that for all $g_i \in \mathcal{G}_i$ there exists $j_i = j(g_i) \in \{1, ..., n_i\}$ with

$$\left(\frac{1}{n} \sum_{k=1}^{n} \left| g_i(x_k) - g_i^{j_i}(x_k) \right|^p \right)^{1/p} < \frac{\varepsilon}{m^{1/p}}.$$

Let $g(x) = \max_{i=1,...,m} g_i(x)$ $(x \in \mathbb{R}^d)$ for some $g_i \in \mathcal{G}_i$ $(i = 1, ..., n)$. Choose for every g_i the corresponding function $g_i^{j_i}$ with

$$\left(\frac{1}{n} \sum_{k=1}^{n} \left| g_i(x_k) - g_i^{j_i}(x_k) \right|^p \right)^{1/p} < \frac{\varepsilon}{m^{1/p}}$$

and define the function $h : \mathbb{R}^d \to \mathbb{R}$ by $h(x) := \max_{i=1}^{m} g_i^{j_i}(x)$.

From the triangle inequality for the supremum norm on the real vector space $\mathbb{R}^n$ it can be easily deduced that

$$\Big|\, ||x||_\infty - ||y||_\infty \Big| \leq ||x - y||_\infty$$

holds, for all vectors $x, y \in \mathbb{R}^n$. Therefore, we can also infer that

$$|\max\{a_1, ..., a_n\} - \max\{b_1, ..., b_n\}| \leq \max_{i=1,...,n} |a_i - b_i| \tag{2.7}$$

holds, for positive real numbers $a_1, ..., a_n$, $b_1, ..., b_n \in \mathbb{R}^+$. In fact for arbitrary real numbers $a_1, ..., a_n$, $b_1, ..., b_n \in \mathbb{R}$ we can choose $c = \min\{a_1, ..., a_n, b_1, ..., b_n\}$ so that inequality (2.7) holds for $a_1 + |c|, ..., a_n + |c|$, $b_1 + |c|, ..., b_n + |c| \in \mathbb{R}^+$. Hence the equations

$$\max\{a_1, ..., a_n\} + |c| = \max\{a_1 + |c|, ..., a_n + |c|\}$$

and

$$\max\{b_1, ..., b_n\} + |c| = \max\{b_1 + |c|, ..., b_n + |c|\}$$

imply that inequality (2.7) holds for all real numbers $a_1, ..., a_n, b_1, ..., b_n \in \mathbb{R}$. Hence we obtain that

$$\begin{aligned}
\left(\frac{1}{n}\sum_{k=1}^{n}|g(x_k) - h(x_k)|^p\right)^{1/p} &= \left(\frac{1}{n}\sum_{k=1}^{n}\left|\max_{i=1,...,m} g_i(x_k) - \max_{i=1,...,m} g_i^{j_i}(x_k)\right|^p\right)^{1/p} \\
&\leq \left(\frac{1}{n}\sum_{k=1}^{n}\max_{i=1,...,m}\left|g_i(x_k) - g_i^{j_i}(x_k)\right|^p\right)^{1/p} \\
&\leq \left(\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{m}\left|g_i(x_k) - g_i^{j_i}(x_k)\right|^p\right)^{1/p} \\
&= \left(\sum_{k=1}^{m}\frac{1}{n}\sum_{i=1}^{n}\left|g_i(x_k) - g_i^{j_i}(x_k)\right|^p\right)^{1/p} \\
&< \left(\sum_{k=1}^{m}\varepsilon^p m^{-p/p}\right)^{1/p} = \left(m\varepsilon^p m^{-1}\right)^{1/p} = \varepsilon
\end{aligned}$$

holds, and thus we have shown assertion (2.5) for the L_p-ε-covering number of the maximum function. Moreover, the proof of inequality (2.6) follows directly from (2.5), together with the two simple insights that

$$\min\{\mathcal{G}_1, ..., \mathcal{G}_m\} = \max\{-\mathcal{G}_1, ..., -\mathcal{G}_m\}$$

holds, for arbitrary function spaces $\mathcal{G}_1, ..., \mathcal{G}_m$, and that

$$\mathcal{N}(\varepsilon, \mathcal{G}_i, x_1^n) = \mathcal{N}(\varepsilon, -\mathcal{G}_i, x_1^n), \qquad (i = 1, ..., n),$$

for all $\varepsilon > 0$, and for all $x_1^n \in \mathbb{R}^d \times ... \times \mathbb{R}^d$. □

In the following this lemma will enable us to bound the L_p-ε-covering number of the truncated version of the class of functions $\mathcal{F}_n$. Remind that results from the VC theory in Section 1.4 partially assume the boundedness of the underlying functions, and that the truncation guarantees this sufficient bound.

LEMMA 2.5. *Let $\varepsilon > 0$ and $z_1^n \in \mathbb{R}^d \times ... \times \mathbb{R}^d$. Set $L_n := \max\{L_{1,n}, ..., L_{K_n,n}\}$. Then,*

$$\mathcal{N}_1\left(\varepsilon, T_\beta \mathcal{F}_n, z_1^n\right) \leq 3\left(\frac{6e\beta}{\varepsilon} \cdot K_n L_n\right)^{2(d+2)\left(\sum_{k=1}^{K_n} L_{k,n}\right)}.$$

holds for $\mathcal{F}_n$ defined by (2.1).

PROOF. In the first step we show that we can involve the truncation operator into the class of functions, that is the equality

$$T_\beta \mathcal{F}_n = \left\{ \max_{1\leq k\leq K_n} \min_{1\leq l\leq L_{k,n}} T_\beta\left(a_{k,l} \cdot x + b_{k,l}\right), \text{ for some } a_{k,l} \in \mathbb{R}^d, b_{k,l} \in \mathbb{R} \right\} \tag{2.8}$$

holds. To attain this, we have to verify the equality

$$T_\beta \max_{1\leq i\leq n} z_i = \max_{1\leq i\leq n} T_\beta z_i, \tag{2.9}$$

for real numbers $z_i \in \mathbb{R}$ $(i = 1, ..., n)$. For this purpose, we may assume without loss of generality that $z_1 = \max\{z_1, ..., z_n\}$. For $-\beta < z_1 < \beta$, we get

$$T_\beta \max_{1\leq i\leq n} z_i = T_\beta z_1 = z_1 = \max\{z_1, ..., z_n\} = \max_{1\leq i\leq n} T_\beta z_i,$$

since $\max_{1\leq i\leq n} T_\beta z_i = \max_{1\leq i\leq n} \max\{z_i, -\beta\}$. For $z_1 \geq \beta$, we have

$$T_\beta \max_{1\leq i\leq n} z_i = T_\beta(z_1) = \beta = \max_{1\leq i\leq n} T_\beta(z_i),$$

since $T_\beta(z_i) \leq \beta$ $(i = 1, ..., n)$ and $T_\beta(z_1) = \beta$. Furthermore (2.9) holds obviously in the case $z_1 \leq -\beta$ and hence, we have verified (2.9). Because of

$$\min_{1\leq i\leq n} z_i = -\max_{1\leq i\leq n} z_i \quad \text{and} \quad T_\beta(-z) = -T_\beta(z)$$

it is obvious that we obtain the analogue equation for the minimum, that is

$$T_\beta \min_{1\leq i\leq n} z_i = \min_{1\leq i\leq n} T_\beta z_i.$$

Thus in addition (2.8) holds and hence, Lemma 2.4 implies that it is sufficient to find covering numbers for $T_\beta \mathcal{G}$ where $\mathcal{G}$ is defined by

$$\mathcal{G} = \left\{ g : \mathbb{R}^d \to \mathbb{R};\ g(x) = a \cdot x + b,\ \ (x \in \mathbb{R}^d),\ \text{for some } a \in \mathbb{R}^d, b \in \mathbb{R} \right\},$$

in order to get covering numbers of $T_\beta \mathcal{F}_n$. Obviously $\mathcal{G}$ is a $d+1$ dimensional linear vector space, which by Theorem 1.12 yields

$$V_{\mathcal{G}^+} \leq (d+1) + 1.$$

Before we are able to apply Theorem 1.11 we have to find a bound on the VC dimension of $T_\beta \mathcal{G}^+$. In order to do so, let $(x, y) \in \mathbb{R}^d \times \mathbb{R}$. If $y \leq -\beta$, then (x, y) is contained in every set of $T_\beta \mathcal{G}^+$, if $y > \beta$, then (x, y) is contained in none of them. Thus, if $T_\beta \mathcal{G}^+$ shatters a set of points, then the y-coordinates of these points are

all bounded in absolute value by β, and $\mathcal{G}^+$ also shatters this set of points. Hence we get

$$V_{T_\beta \mathcal{G}^+} \leq V_{\mathcal{G}^+},$$

(cf. equality (10.23) in Györfi et al. (2002)). Thus Lemma 1.9 and Theorem 1.11 imply

$$\mathcal{N}_1\left(\varepsilon, T_\beta \mathcal{G}, z_1^n\right) \leq 3\left(\frac{4e\beta}{\varepsilon} \cdot \log \frac{6e\beta}{\varepsilon}\right)^{(d+2)}.$$

Although the functions of $T_\beta \mathcal{G}$ have the range $[-\beta, \beta]$ and therefore are not bounded below by zero, we are permitted to apply Theorem 1.11, because one can just lift the functions, such that their range is $[0, 2\beta]$. Obviously this class of lifted functions and the original class $T_\beta \mathcal{G}$ have the same covering number.

In addition, Lemma 2.4 implies for $L_n := \max\{L_{1,n}, ..., L_{K_n,n}\}$ that

$$\begin{aligned}
\mathcal{N}_1\left(\varepsilon, T_\beta \mathcal{F}_n, z_1^n\right) &= \mathcal{N}_1\left(\varepsilon, \max_{1\leq k\leq K_n} \min_{1\leq l\leq L_{k,n}} T_\beta \mathcal{G}, z_1^n\right) \\
&\leq \prod_{k=1}^{K_n} \mathcal{N}_1\left(\frac{\varepsilon}{K_n}, \min_{1\leq l\leq L_{k,n}} T_\beta \mathcal{G}, z_1^n\right) \\
&\leq \prod_{k=1}^{K_n} \prod_{l=1}^{L_{k,n}} \mathcal{N}_1\left(\frac{\varepsilon}{K_n \cdot L_n}, T_\beta \mathcal{G}, z_1^n\right) \\
&\leq 3\left(\frac{4e\beta}{\varepsilon} \cdot K_n L_n \cdot \log\left(\frac{6e\beta}{\varepsilon} \cdot K_n L_n\right)\right)^{(d+2)\sum_{k=1}^{K_n} L_{k,n}} \\
&\leq 3\left(\frac{6e\beta}{\varepsilon} \cdot K_n L_n\right)^{2(d+2)\left(\sum_{k=1}^{K_n} L_{k,n}\right)}
\end{aligned}$$

holds for arbitrary $\varepsilon > 0$. □

In view of Chapter 5 we would like to remark that the bound in Lemma 2.5 is a uniform bound, which does not depend on the certain choice of the points $z_1^n \in \mathbb{R}^d \times \cdots \times \mathbb{R}^d$.

Thus we have bounds on the covering numbers of $\mathcal{F}_n$ depending on the underlying parameters $K_n, L_{1,n}, \ldots L_{K_n,n} \in \mathbb{N}$. In the following we will analyse the asymptotics of our maxmin estimate $m_n = T_\beta \tilde{m}_n$ defined by (2.2) and (2.3), using these appraisements for the complexity of the underlying class of functions.

CHAPTER 3

Analysis of Asymptotic Behaviour

In Section 3.1 we shall prove that the estimate introduced in Chapter 2 is strongly universally consistent for all distributions of (X, Y) with $X \in [0,1]^d$ *a.s.* For this purpose we use the results concerning the covering numbers of $\mathcal{F}_n$, we proved in the last chapter. In Section 3.2 we shall derive a rate of convergence of our estimate which is optimal up to a logarithmic factor. The derived rate of convergence holds for all distributions of (X, Y) with $X \in [a,b]^d$ *a.s.* and a (p, C)–smooth regression function m. We do not assume that Y is bounded, because it suffices to suppose a modified Sub-Gaussian condition, that is,

$$\mathbf{E}\left(e^{c\cdot|Y|^2}\right) < \infty,$$

for some constant $c > 0$. The last section describes how one can choose the parameters in dependency of the given set of data and furthermore, provides a rate of convergence for estimates with this data-dependent choice of parameters.

3.1. Universal Consistency

The aim of this section is to prove strong univeral consistency of the maxmin estimate m_n defined by (2.2) and (2.3) or, more precisely, to show that

$$\lim_{n\to\infty} \int |m_n(x) - m(x)|^2 \mu(dx) = 0 \quad a.s.$$

for a certain class of distributions of (X, Y) (cf. Definition 1.2). To obtain the desired consistency we will use the following result from Györfi et al. (2002).

THEOREM 3.1. *Let $\mathcal{G}_n = \mathcal{G}_n(\mathcal{D}_n)$ be a class of functions $f : \mathbb{R}^d \to \mathbb{R}$, and assume that the estimator m_n satisfies*

$$\tilde{m}_n(\cdot) \;=\; \arg\min_{f\in\mathcal{G}_n} \frac{1}{n}\sum_{i=1}^{n} |f(X_i) - Y_i|^2.$$

and

$$m_n(x) = T_{\beta_n}\tilde{m}_n(x),$$

for all $x \in \mathbb{R}^d$.

If

$$\lim_{n\to\infty} \beta_n = \infty,$$

$$\lim_{n\to\infty} \inf_{f\in\mathcal{G}_n, ||f||_\infty \leq \beta_n} \int |f(x) - m(x)|^2 \mu(dx) \quad = \quad 0 \quad a.s., \tag{3.1}$$

$$\lim_{n\to\infty} \sup_{f\in T_{\beta_n}\mathcal{G}_n} \left| \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - T_L Y_i|^2 - \mathbf{E}\left((f(X) - T_L Y)^2\right) \right| \quad = \quad 0 \tag{3.2}$$

a.s., for all $L > 0$, then

$$\lim_{n\to\infty} \int |m_n(x) - m(x)|^2 \mu(dx) = 0 \ a.s.$$

PROOF. The proof can be found in Györfi et al. (2002), Theorem 10.2. □

Since our maxmin estimate is a truncated version of a least squares estimate, we just have to show conditions (3.1) and (3.2), in order to gain the consistency.

THEOREM 3.2. *Let m_n be the esimate defined by (2.2) and (2.3) and set furthermore $L_n = \max\{L_{1,n}, ..., L_{K_n,n}\}$. If the parameters satisfy*

$$\beta_n \to \infty, \ K_n \to \infty, \ L_{k,n} \to \infty \ \ for \ \ k = 1, ..., K_n, \tag{3.3}$$

and in addition

$$\frac{\beta_n^4 \cdot \sum_{k=1}^{K_n} L_{k,n} \cdot \log(\beta_n \cdot K_n \cdot L_n)}{n} \to 0, \tag{3.4}$$

for $n \to \infty$ and if, for some $\delta > 0$,

$$\frac{\beta_n^4}{n^{1-\delta}} \to 0 \tag{3.5}$$

holds, then

$$\lim_{n\to\infty} \int |m_n(x) - m(x)|^2 \mu(dx) = 0 \quad a.s. \ ,$$

for all distributions of (X, Y) with $X \in [0, 1]^d$ a.s. and $\mathbf{E}(Y^2) < \infty$.

From Theorem 3.1 we know that it suffices to verify conditions (3.1) and (3.2) to get the desired result in this setting. For the proof of condition (3.1), we start with the case that the regression function m is Lipschitz continuous with Lipschitz constant $C > 0$, that is

$$|m(x) - m(y)| \leq C \cdot ||x - y||$$

holds, for all $x, y \in \mathbb{R}^d$. Later we will extend this to more general regression functions m by using the denseness of Lipschitz continuous functions in $L_2(\mu)$.

In order to show that Lipschitz continuous functions can be approximated arbitrarily well by maxmin functions we decompose $[0, 1]^d$ into n^d subcubes and

choose certain linear functions such that their maxmin function interpolates the given Lipschitz continuous function in the vertices of the subcubes $1/n \cdot i$, where $i = (i^{(1)}, \dots i^{(d)}) \in \{0, ..., n\}^d$. More detailed, we consider functions of the form

$$x \mapsto m\left(\frac{1}{n} \cdot i\right) + C \cdot C_d \cdot \sum_{k=1}^{d} \delta_d(k) \cdot \left(x^{(k)} - \frac{i^{(k)}}{n}\right), \tag{3.6}$$

where $\delta_d : \{1, ..., d\} \to \{-1, 1\}$. Here, for $d \in \mathbb{N}$, C_d is the constant resulting from the equivalence of the $||\cdot||_1$–norm and the Euclidean–norm, which we will denote by $||\cdot||_2$ in this section, in order to avoid misunderstandings. That is, C_d satisfies

$$\frac{1}{C_d}||x||_1 \leq ||x||_2 \leq C_d||x||_1,$$

for all $x \in \mathbb{R}^d$. Due to the number of possibilities to map $\{1, ..., d\}$ onto $\{-1, 1\}$ there exist 2^d different functions of the form (3.6). It is easy to show that these functions are Lipschitz continuous as well, but with Lipschitz constant $C \cdot C_d^2$. In fact, let $\delta_d : \{1, ..., d\} \to \{-1, 1\}$ be fixed. Then we have, for arbitrary $x = (x^{(1)}, ..., x^{(d)}), y = (y^{(1)}, ..., y^{(d)}) \in \mathbb{R}^d$ and all $i = (i^{(1)}, ..., i^{(d)}) \in \{1, ..., n\}^d$,

$$\begin{aligned}
&\left| m\left(\frac{1}{n} \cdot i\right) + C \cdot C_d \cdot \sum_{k=1}^{d} \delta_d(k) \cdot \left(x^{(k)} - \frac{i^{(k)}}{n}\right) \right. \\
&\qquad \left. - \left(m\left(\frac{1}{n} \cdot i\right) + C \cdot C_d \cdot \sum_{k=1}^{d} \delta_d(k) \cdot \left(y^{(k)} - \frac{i^{(k)}}{n}\right)\right)\right| \\
&= \left| C \cdot C_d \cdot \sum_{k=1}^{d} \delta_d(k) \cdot \left(x^{(k)} - \frac{i^{(k)}}{n}\right) - C \cdot C_d \cdot \sum_{k=1}^{d} \delta_d(k) \cdot \left(y^{(k)} - \frac{i^{(k)}}{n}\right)\right| \\
&= \left| C \cdot C_d \cdot \sum_{k=1}^{d} \left(\delta_d(k) \cdot \left(\left(x^{(k)} - \frac{i^{(k)}}{n}\right) - \left(y^{(k)} - \frac{i^{(k)}}{n}\right)\right)\right)\right| \\
&= \left| C \cdot C_d \cdot \sum_{k=1}^{d} \left(\delta_d(k) \cdot \left(x^{(k)} - y^{(k)}\right)\right)\right| \\
&\leq c \cdot C_d \cdot \sum_{k=1}^{d} |\delta_d(k)| \cdot \left|x^{(k)} - y^{(k)}\right| \\
&= C \cdot C_d \cdot \sum_{k=1}^{d} \left|x^{(k)} - y^{(k)}\right| = C \cdot C_d \cdot ||x - y||_1 \\
&\leq C \cdot C_d^2 \cdot ||x - y||_2.
\end{aligned}$$

The next lemma shows that functions of the shape

$$f_n^d(x) = \max_{i \in \{1,...,n\}^d} \min_{\delta_d} m\left(\frac{1}{n} \cdot i\right) + c \cdot C_d \cdot \sum_{k=1}^{d} \delta_d(k) \cdot \left(x^{(k)} - \frac{i^{(k)}}{n}\right) \tag{3.7}$$

$(x \in [0,1]^d)$, with $\delta_d : \{1,...,d\} \to \{-1,1\}$, approximate the underlying function m arbitrarily close in both the L_2–norm and the $||\cdot||_\infty$–norm.

LEMMA 3.3. *Let $m : \mathbb{R}^d \to \mathbb{R}$ be Lipschitz continuous with Lipschitz constant C and with compact support $[0,1]^d$. Then, for f_n^d defined by (3.7),*

$$\int \left| f_n^d(x) - m(x) \right|^2 \mu(dx) \quad \leq \quad ||f_n^d(x) - m(x)||_{[0,1]^d,\infty} \leq 4C^2 \cdot C_d^6 \cdot \frac{d^2}{n^2}$$

holds for all $n \in \mathbb{N}$.

PROOF. In the first step we show that f_n^d is Lipschitz continuous with Lipschitz constant $C \cdot C_d^2$, too. Firstly note that, for Lipschitz continuous functions $g_1, ..., g_n : \mathbb{R}^d \to \mathbb{R}$ their minimum function

$$g(x) = \min_{1 \leq i \leq n} g_i(x)$$

is also Lipschitz continuous with the same Lipschitz constant, owing to

$$\begin{aligned} |g(x) - g(y)| &= \left| \min_{1 \leq i \leq n} g_i(x) - \min_{1 \leq i \leq n} g_i(y) \right| \\ &\leq \max_{1 \leq i \leq n} |g_i(x) - g_i(y)| \\ &\leq C \cdot ||x - y||_2. \end{aligned}$$

Analogously we get for the maximum function $h(x) = \max_{1 \leq i \leq n} g_i(x)$ that

$$\begin{aligned} |h(x) - h(y)| &= \left| \max_{1 \leq i \leq n} g_i(x) - \max_{1 \leq i \leq n} g_i(y) \right| \\ &\leq \max_{1 \leq i \leq n} |g_i(x) - g_i(y)| \\ &\leq C \cdot ||x - y||_2. \end{aligned}$$

As we have already seen earlier, functions of the form

$$x \mapsto m\left(\frac{1}{n} \cdot i\right) + c \cdot C_d \cdot \sum_{k=1}^{d} \delta_d(k) \cdot \left(x^{(k)} - \frac{i^{(k)}}{n} \right)$$

are Lipschitz continuous with constant $C \cdot C_d^2$, which yields the Lipschitz continuity for f_n^d defined by (3.7), with Lipschitz constant $C \cdot C_d^2$.

Furthermore f_n^d interpolates the function m in the points $1/n \cdot i$ for $i \in \{0,...,n\}^d$, because the Lipschitz continuity of m implies, for arbitrary $i \in \{0,...,n\}^d$,

$$\left| m\left(\frac{1}{n} \cdot i\right) - m(x) \right| \leq C \left|\left| \frac{1}{n} \cdot i - x \right|\right|_2 \leq C \cdot C_d \left|\left| \frac{1}{n} \cdot i - x \right|\right|_1,$$

and hence

$$m\left(\frac{1}{n} \cdot i\right) - m(x) \leq C \cdot C_d \left|\left| \frac{1}{n} \cdot i - x \right|\right|_1,$$

which leads to

$$m\left(\frac{1}{n} \cdot i\right) - C \cdot C_d \left|\left| \frac{1}{n} \cdot i - x \right|\right|_1 \leq m(x).$$

Moreover, from

$$\begin{aligned} -C \cdot C_d \left|\left| \frac{1}{n} \cdot i - x \right|\right|_1 &= -C \cdot C_d \sum_{k=1}^{d} \left| \frac{i_k}{n} - x_k \right| \\ &= C \cdot C_d \cdot \min_{\delta_d : \{1,...,d\} \to \{-1,1\}} \left\{ \sum_{k=1}^{d} \delta_d(k) \cdot \left(x^{(k)} - \frac{i^{(k)}}{n} \right) \right\} \end{aligned}$$

we deduce that $f_n^d(x) \leq m(x)$ holds for all $x \in [0,1]^d$. On the other hand, the definition implies

$$f_{n^d}\left(\frac{1}{n} \cdot i\right) \geq m\left(\frac{1}{n} \cdot i\right),$$

for all $i \in \{0,...,n\}^d$. Hence we obtain equality of the function values in the points $1/n \cdot i$, for $i \in \{0,...,n\}^d$.

Without loss of generality, we can choose for $x = (x^{(1)},...,x^{(d)}) \in (0,1)^d$ some $k^{(1)},...,k^{(d)} \in \{0,...,(n-1)/n\}$ such that $k^{(j)} \leq x^{(j)} < k^{(j)} + 1/n$ holds. Thus the Lipschitz continuity of f_n^d and m yields

$$\begin{aligned} |f_n^d(x) - m(x)| &\leq \left| f_n^d(x) - f_n^d(k^{(1)},...,k^{(d)}) \right| \\ &\quad + \left| f_n^d(k^{(1)},...,k^{(d)}) - m(k^{(1)},...,k^{(d)}) \right| \\ &\quad + \left| m(k^{(1)},...,k^{(d)}) - m(x) \right| \\ &\leq C \cdot C_d^2 \cdot ||x - (k^{(1)},...,k^{(d)})||_2 + 0 + C \cdot ||x - (k^{(1)},...,k^{(d)})||_2 \\ &\leq C \cdot C_d^3 \sum_{j=1}^{d} |x^{(j)} - k^{(j)}| + C \cdot C_d \sum_{j=i}^{d} |x^{(j)} - k^{(j)}| \\ &\leq C \cdot C_d^3 \sum_{j=1}^{d} \frac{1}{n} + C \cdot C_d \sum_{j=i}^{d} \frac{1}{n} \\ &\leq 2C \cdot C_d^3 \frac{d}{n}. \end{aligned}$$

Consequently we get

$$\begin{aligned} \int \left| f_n^d(x) - m(x) \right|^2 \mu(dx) &\leq ||f_n^d - m||_{[0,1]^d,\infty}^2 \\ &= \max_{x \in [0,1]^d} |f_n^d(x) - m(x)|^2 \leq \left(2C \cdot C_d^3 \cdot \frac{d}{n} \right)^2 \\ &= \frac{4C^2 \cdot C_d^6 \cdot d^2}{n^2}. \end{aligned}$$

□

Proof of Theorem 3.2. As already mentioned, by Theorem 3.1 it suffices to show conditions (3.1) and (3.2) in order to prove the desired consistency. As for condition (3.1) we start by the observation that $C_0^\infty(\mathbb{R}^d)$ (the set of all infinitely

often continuously differentiable functions on $\mathbb{R}^d$ with compact support) is dense in $L_2(\mu)$. This follows from the denseness of $C_0^\infty(\mathbb{R}^d)$ in $C_0(\mathbb{R}^d)$ and Lemma 1.15. Since all continuously differentiable functions are Lipschitz continuous, the set of all Lipschitz continuous functions is also dense in $L_2(\mu)$. Thus, for a given function $m : \mathbb{R}^d \to \mathbb{R}$ and every $\varepsilon > 0$, there exists a Lipschitz continuous function $g : \mathbb{R}^d \to \mathbb{R}$ such that

$$\int |g(x) - m(x)|^2 \mu(dx) \quad < \quad \varepsilon. \tag{3.8}$$

Furthermore, Lemma 3.3 implies for $K_n \to \infty$ and $L_n \to \infty$ $(n \to \infty)$, and all Lipschitz continuous functions $g : \mathbb{R}^d \to \mathbb{R}$ with compact support $[0,1]^d$, that

$$\lim_{n\to\infty} \inf_{f\in\mathcal{F}_n} \int |f(x) - g(x)|^2 \mu(dx) \leq \lim_{n\to\infty} \inf_{f\in\mathcal{F}_n} ||f - g||_\infty^2 = 0,$$

since $f_n^d \in \mathcal{F}_n$ for sufficiently large $n \in \mathbb{N}$ (or, in other words, for sufficiently large parameters K_n and L_n).

Moreover, f_n^d obviously is bounded in absolute value, because all Lipschitz continuous functions g with compact support are bounded, and by Lemma 3.3 $||f_n^d - g||_\infty$ converges to zero for n tending to infinity. Thus we get

$$\lim_{n\to\infty} \inf_{f\in\mathcal{F}_n, ||f||_\infty \leq \beta_n} \int |f(x) - g(x)|^2 \mu(dx) \leq \lim_{n\to\infty} \inf_{f\in\mathcal{F}_n, ||f||_\infty \leq \beta_n} ||f - g||_\infty^2 = 0$$

almost surely and due to inequality (3.8), this implies condition (3.1).

In order to show (3.2), let $L > 0$ be arbitrary. Without loss of generality we can assume $L < \beta_n$, since $\beta_n \to \infty$ $(n \to \infty)$. Write

$$Z = (X, Y), Z_1 = (X_1, Y_1), ..., Z_n = (X_n, Y_n)$$

and

$$\mathcal{H}_n := \left\{ h : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}, \exists f \in T_{\beta_n}\mathcal{F}_n \text{ mit } h(x,y) = |f(x) - T_L(y)|^2 \right\}.$$

Obviously, for all $h \in \mathcal{H}_n$ and for all $x \in \mathbb{R}^d, y \in \mathbb{R}$, we have

$$\begin{aligned} 0 \leq h(x,y) &= |f^{(h)}(x) - T_L(y)|^2 \leq 2|f^{(h)}(x)|^2 + 2|T_L(y)|^2 \leq 2\beta_n^2 + 2L^2, \\ &\leq 4\beta_n^2 \end{aligned}$$

where $f^{(h)} \in T_{\beta_n}\mathcal{F}_n$ is chosen such that $h(x,y) = |f^{(h)}(x) - T_L(y)|^2$, for all $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$. Hence with Lemma 1.7 we obtain

$$\begin{aligned}
&\mathbf{P}\left\{\sup_{f\in T_{\beta_n}\mathcal{F}_n}\left|\frac{1}{n}\sum_{i=1}^n |f(X_i) - T_L Y_i|^2 - \mathbf{E}\left(|f(X) - T_L Y|^2\right)\right| > \varepsilon\right\} \\
&= \mathbf{P}\left\{\sup_{h\in\mathcal{H}_n}\left|\frac{1}{n}\sum_{i=1}^n h(Z_i) - \mathbf{E}(h(Z))\right| > \varepsilon\right\} \\
&\leq 8\cdot\mathbf{E}\left(\mathcal{N}_1\left(\frac{\varepsilon}{8},\mathcal{H}_n,Z_1^n\right)\right)\cdot e^{-\frac{n\varepsilon^2}{128(4\beta_n^2)^2}} \qquad (3.9)
\end{aligned}$$

for arbitrary $\varepsilon > 0$. In the following we will bound the covering number of $\mathcal{H}_n$ by bounding its packing number and using Lemma 1.9. For this purpose, let $h_i(x,y) = |f_i(x) - T_L y|^2$, $((x,y) \in \mathbb{R}^d \times \mathbb{R})$ for some $f_i \in T_{\beta_n}\mathcal{F}_n$. Then

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n |h_1(Z_i) - h_2(Z_i)| \\
&= \frac{1}{n}\sum_{i=1}^n \left||f_1(X_i) - T_L Y_i|^2 - |f_2(X_i) - T_L Y_i|^2\right| \\
&= \frac{1}{n}\sum_{i=1}^n \Big|\Big((f_1(X_i) - T_L Y_i) - (f_2(X_i) - T_L Y_i)\Big) \\
&\qquad\qquad \cdot\Big((f_1(X_i) - T_L Y_i) + (f_2(X_i) - T_L Y_i)\Big)\Big| \\
&= \frac{1}{n}\sum_{i=1}^n \Big|\Big((f_1(X_i) - f_2(X_i)\Big)\cdot\Big((f_1(X_i) - 2T_L Y_i + f_2(X_i)\Big)\Big| \\
&\leq 4\beta_n \frac{1}{n}\sum_{i=1}^n |(f_1(X_i) - f_2(X_i)|,
\end{aligned}$$

because f_1, f_2 and $T_L Y_i$ are bounded by β_n. Thus, if $\{h_1, ..., h_l\}$ is an $\varepsilon/8$-packing of $\mathcal{H}_n$ on Z_1^n, then $\{f_1, ..., f_l\}$ has to be an $\varepsilon/(8\cdot 4\beta_n)$-packing of $T_{\beta_n}\mathcal{F}_n$ on X_1^n. In terms of packing numbers this means that

$$\mathcal{M}_1\left(\frac{\varepsilon}{8},\mathcal{H}_n,Z_1^n\right) \leq \mathcal{M}_1\left(\frac{\varepsilon}{32\beta_n},T_{\beta_n}\mathcal{F}_n,X_1^n\right)$$

holds. Hence Lemma 2.5 yields

$$\begin{aligned}
\mathcal{M}_1\left(\frac{\varepsilon}{8},\mathcal{H}_n,Z_1^n\right) &\leq \mathcal{M}_1\left(\frac{\varepsilon}{32\beta_n},T_{\beta_n}\mathcal{F}_n,X_1^n\right) \leq \mathcal{N}_1\left(\frac{\varepsilon}{64\beta_n},T_{\beta_n}\mathcal{F}_n,X_1^n\right) \\
&\leq 3\left(\frac{6\cdot 64e\beta_n^2}{\varepsilon}\cdot K_n\cdot L_n\right)^{2(d+2)\sum_{k=1}^{K_n} L_{k,n}} \\
&= 3\left(\frac{384e\beta_n^2}{\varepsilon}\cdot K_n\cdot L_n\right)^{2(d+2)\sum_{k=1}^{K_n} L_{k,n}},
\end{aligned}$$

and therefore inequality (3.9) implies

$$\mathbf{P}\left\{\sup_{f\in T_{\beta_n}\mathcal{F}_n}\left|\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-T_LY_i|^2-\mathbf{E}\left(|f(X)-T_LY|^2\right)\right|>\varepsilon\right\}$$
$$\leq 24\left(\frac{384e\beta_n^2}{\varepsilon}\cdot K_n\cdot L_n\right)^{2(d+2)\sum_{k=1}^{K_n}L_{k,n}}\cdot\exp\left(-\frac{n\varepsilon^2}{2048\beta_n^4}\right).$$

In addition this leads to

$$\begin{aligned}
&\sum_{i=1}^{\infty}\mathbf{P}\left\{\sup_{f\in T_{\beta_n}\mathcal{F}_n}\left|\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-T_LY_i|^2-\mathbf{E}\left(|f(X)-T_LY|^2\right)\right|>\varepsilon\right\}\\
&\leq\sum_{i=1}^{\infty}24\left(\frac{384e\beta_n^2}{\varepsilon}\cdot K_n\cdot L_n\right)^{2(d+2)\sum_{k=1}^{K_n}L_{k,n}}\cdot\exp\left(-\frac{n\varepsilon^2}{2048\beta_n^4}\right)\\
&\leq\sum_{1=1}^{\infty}24\cdot\exp\left(2(d+2)\sum_{k=1}^{K_n}L_{k,n}\cdot\log\left(\frac{384e\beta_n^2}{\varepsilon}\cdot K_n\cdot L_n\right)-\frac{n\varepsilon^2}{2048\beta_n^4}\right)\\
&=\sum_{i=1}^{\infty}24\cdot\exp\left[-n^{\delta}\frac{n^{1-\delta}}{\beta_n^4}\right.\\
&\qquad\left.\cdot\left(\frac{\varepsilon^2}{2048}-\frac{2(d+2)\cdot\beta_n^4}{n}\cdot\sum_{k=1}^{K_n}L_{k,n}\cdot\log\left(\frac{384e\beta_n^2}{\varepsilon}\cdot K_n\cdot L_n\right)\right)\right]\\
&\leq\sum_{i=1}^{\infty}24\cdot e^{-n\delta}<\infty.
\end{aligned}$$

Here the fourth inequality follows from the assumptions (3.3), (3.4) and (3.5) on the parameters. The convergence of the sum results with the comparison test from $|e^{-\delta}|<1$. Now the desired convergence in (3.2) is the direct consequence of the Borel-Cantelli lemma (cf. Lemma 1.18). □

So far we have proved that our estimate is universally strongly consistent, which of course is a desirable property, but not completely satisfactory with regard to applications. Therefore the next section will give us an idea how fast the L_2 error of our estimate is tending to zero.

3.2. Rate of Convergence

In this section we will derive a rate of convergence of our estimate in the case of a (p,C)–smooth regression function. Here we do not have to assume that Y is bounded in absolute value, since the assumption of a modified Sub-Gaussian

condition is sufficient for us. The derived rate of convergence

$$C^{2d/(2p+d)} \left(\frac{\log(n)^3}{n} \right)^{2p/(2p+d)}$$

is optimal (in the minimax sense) up to the logarithmic factor (cf. (1.11)). However it depends on the dimension d of X and hence it is comparitively slow for a large dimension d. We will see in the Chapter 4 how one can circumvent this so-called 'curse of dimensionality' by some structural assumptions on the underlying regression function.

We start with a theorem that gives an upper bound on the expected L_2 error of our estimate.

THEOREM 3.4. *Let $K_n, L_{1,n}, ..., L_{K_n,n} \in \mathbb{N}$, with $K_n \cdot \max\{L_{1,n}, ..., L_{K_n,n}\} \leq n^2$, and set $\beta_n = c_1 \cdot \log(n)$ for some constant $c_1 > 0$. Assume that the distribution of (X, Y) satifies*

$$\mathbf{E}\left(e^{c_2 \cdot |Y|^2} \right) < \infty \tag{3.10}$$

for some constant $c_2 > 0$ and that the regression function m is bounded in absolute value. Then, for the estimate m_n defined by (2.2) and (2.3),

$$\begin{aligned} &\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) \\ &\quad \leq \frac{c_3 \cdot \log(n)^3 \cdot \sum_{k=1}^{K_n} L_{k,n}}{n} \\ &\quad + \mathbf{E}\left(2 \inf_{f \in \mathcal{F}_n} \left(\frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2 - \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2 \right) \right), \end{aligned} \tag{3.11}$$

for some constant $c_3 > 0$, and therefore

$$\begin{aligned} \mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) \quad \leq \quad & \frac{c_3 \cdot \log(n)^3 \cdot \sum_{k=1}^{K_n} L_{k,n}}{n} \\ & + 2 \cdot \inf_{f \in \mathcal{F}_n} \int |f(x) - m(x)|^2 \mu(dx), \end{aligned}$$

where c_3 does not depend on n, β_n or the parameters of the estimate.

The Condition (3.10) is a modified Sub-Gaussian condition. In view of the applications to simulated data in part A.2 of the appendix, we want to remark that (3.10) is satisfied in particular, whenever $\mathbf{P}_{Y|X=x}$ is the normal distribution $\mathcal{N}_{(m(x),\sigma^2)}$ for a bounded regression function m. Moreover, all bounded conditional distributions of Y obviously satisfy Condition (3.10), as well. Therefore this assumption allows us to consider unbounded conditional distributions of Y, such as the normal distribution.

PROOF. In the proof we use the following error decomposition:

$$\begin{aligned}
&\int |m_n(x)-m(x)|^2\mu(dx)\\
&\quad= \Big[\mathbf{E}\Big(|m_n(X)-Y|^2|\mathcal{D}_n\Big)-\mathbf{E}\Big(|m(X)-Y|^2\Big)\\
&\qquad\quad -\mathbf{E}\Big(|m_n(X)-T_{\beta_n}Y|^2|\mathcal{D}_n\Big)-\mathbf{E}\Big(|m_{\beta_n}(X)-T_{\beta_n}Y|^2\Big)\Big]\\
&\qquad+\Bigg[\mathbf{E}\Big(|m_n(X)-T_{\beta_n}Y|^2|\mathcal{D}_n\Big)-\mathbf{E}\Big(|m_{\beta_n}(X)-T_{\beta_n}Y|^2\Big)\\
&\qquad\quad -2\cdot\frac{1}{n}\sum_{i=1}^n\Big(|m_n(X_i)-T_{\beta_n}Y_i|^2-|m_{\beta_n}(X_i)-T_{\beta_n}Y_i|^2\Big)\Bigg]\\
&\qquad+\Bigg[2\cdot\frac{1}{n}\sum_{i=1}^n|m_n(X_i)-T_{\beta_n}Y_i|^2-2\cdot\frac{1}{n}\sum_{i=1}^n|m_{\beta_n}(X_i)-T_{\beta_n}Y_i|^2\\
&\qquad\quad -\left(2\cdot\frac{1}{n}\sum_{i=1}^n|m_n(X_i)-Y_i|^2-2\cdot\frac{1}{n}\sum_{i=1}^n|m(X_i)-Y_i|^2\right)\Bigg]\\
&\qquad+\left[2\left(\frac{1}{n}\sum_{i=1}^n|m_n(X_i)-Y_i|^2-\frac{1}{n}\sum_{i=1}^n|m(X_i)-Y_i|^2\right)\right]\\
&\quad= \sum_{i=1}^4 T_{i,n},
\end{aligned}$$

where $T_{\beta_n}Y$ is the truncated version of Y, and m_{β_n} is the regression function of $T_{\beta_n}Y$, that is,

$$m_{\beta_n}(x)=\mathbf{E}\Big(T_{\beta_n}Y|X=x\Big).$$

We start with bounding $T_{1,n}$. By using $a^2-b^2=(a-b)(a+b)$ we get

$$\begin{aligned}
T_{1,n} &= \mathbf{E}\Big(|m_n(X)-Y|^2-|m_n(X)-T_{\beta_n}Y|^2\Big|\mathcal{D}_n\Big)\\
&\quad -\mathbf{E}\Big(|m(X)-Y|^2-|m_{\beta_n}(X)-T_{\beta_n}Y|^2\Big)\\
&= \mathbf{E}\Big((T_{\beta_n}Y-Y)(2m_n(X)-Y-T_{\beta_n}Y)\Big|\mathcal{D}_n\Big)\\
&\quad +\Big(-\mathbf{E}\Big(\Big(m(X)-m_{\beta_n}(X)+T_{\beta_n}Y-Y\Big)\\
&\qquad\cdot\Big(m(X)+m_{\beta_n}(X)-Y-T_{\beta_n}Y\Big)\Big)\Big)\\
&= T_{5,n}+T_{6,n}.
\end{aligned}$$

The Cauchy-Schwarz inequality and the inequality

$$\mathbf{1}_{\{|Y|>\beta_n\}}\le\frac{\exp(c_2/2\cdot|Y|^2)}{\exp(c_2/2\cdot\beta_n^2)} \tag{3.12}$$

lead to

$$|T_{5,n}| \le \sqrt{\mathbf{E}\big(|T_{\beta_n}Y-Y|^2\big)}\cdot\sqrt{\mathbf{E}\big(|2m_n(X)-Y-T_{\beta_n}Y|^2\big|\mathcal{D}_n\big)}$$

$$\begin{aligned}
&\leq \sqrt{\mathbf{E}\big(|Y|^2 \cdot I_{\{|Y|>\beta_n\}}\big)} \cdot \sqrt{\mathbf{E}\big(2 \cdot |2m_n(X) - T_{\beta_n} Y|^2 + 2 \cdot |Y|^2 \big| \mathcal{D}_n\big)} \\
&\leq \sqrt{\mathbf{E}\left(|Y|^2 \cdot \frac{\exp(c_2/2 \cdot |Y|^2)}{\exp(c_2/2 \cdot \beta_n^2)} \right)} \\
&\qquad \cdot \sqrt{\mathbf{E}\big(2 \cdot |2m_n(X) - T_{\beta_n} Y|^2 \big| \mathcal{D}_n\big) + 2\mathbf{E}\big(|Y|^2\big)} \\
&\leq \sqrt{\mathbf{E}\Big(|Y|^2 \exp(c_2/2 \cdot |Y|^2)\Big)} \exp\left(-\frac{c_2 \cdot \beta_n^2}{4}\right) \sqrt{2(3\beta_n)^2 + 2\mathbf{E}\big(|Y|^2\big)}.
\end{aligned}$$

With $x \leq \exp(x)$, for $x \in \mathbb{R}$, we get

$$|Y|^2 \leq \frac{2}{c_2} \cdot \exp\left(\frac{c_2}{2}|Y|^2\right).$$

Hence $\sqrt{\mathbf{E}\Big(|Y|^2 \cdot \exp\left(c_2/2 \cdot |Y|^2\right)\Big)}$ is bounded by the square root of

$$\mathbf{E}\left(\frac{2}{c_2} \cdot \exp\left(c_2/2 \cdot |Y|^2\right) \cdot \exp(c_2/2 \cdot |Y|^2)\right) \leq \mathbf{E}\left(\frac{2}{c_2} \cdot \exp\left(c_2 \cdot |Y|^2\right)\right) \leq c_4,$$

which is finite by Condition (3.10). Because of

$$\mathbf{E}(|Y|^2) \leq \mathbf{E}\left(\frac{1}{c_2} \cdot \exp\left(c_2 \cdot |Y|^2\right)\right) \leq c_5 < \infty,$$

which results again from (3.10), we obtain for the third term that

$$\sqrt{2(3\beta_n)^2 + 2\mathbf{E}\big(|Y|^2\big)} \leq \sqrt{18\beta_n^2 + c_5}$$

holds, for some constant c_5. With the setting $\beta_n = c_1 \cdot \log(n)$ we have that

$$\begin{aligned}
|T_{5,n}| &\leq \sqrt{c_4}\, \exp\left(\frac{-c_2 \cdot c_1^2}{4} \cdot \log(n)^2\right) \cdot \sqrt{18 \cdot c_1^2 \cdot \log(n)^2 + c_5} \\
&= \sqrt{c_4}\, \left(\exp\left(-\log(n)^2\right)\right)^{c_2 \cdot c_1^2/4} \cdot c_6 \cdot c_1 \cdot \log(n) \\
&\leq \sqrt{c_4} \cdot c_6 \cdot c_1 \exp(-\log(n)^2) \cdot \log(n) \leq \frac{\sqrt{c_4} \cdot c_6 \cdot c_1}{n^2} \cdot \log(n) \\
&\leq c_7 \cdot \frac{\log(n)}{n}
\end{aligned}$$

for sufficiently large constants $c_6, c_7 > 0$. Next we consider $T_{6,n}$. The Cauchy-Schwarz inequality yields

$$\begin{aligned}
|T_{6,n}| \leq\ & \sqrt{2\,\mathbf{E}\Big(|m(X) - m_{\beta_n}(X)|^2\Big) + 2\,\mathbf{E}\Big(|(T_{\beta_n} Y - Y)|^2\Big)} \\
& \cdot \sqrt{\mathbf{E}\left(\Big|m(X) + m_{\beta_n}(X) - Y - T_{\beta_n} Y\Big|^2\right)},
\end{aligned}$$

where we can bound the second factor on the right hand-side in the same way we have bounded the second factor from $T_{5,n}$, since $||m||_\infty$ is bounded by assumption,

and since m_{β_n} obviously is bounded by β_n. Thus we get, for some constant $c_8 > 0$,

$$\sqrt{\mathbf{E}\left(\left|m(X) + m_{\beta_n}(X) - Y - T_{\beta_n} Y\right|^2\right)} \leq c_8 \cdot \log(n).$$

The first term can be bounded with Jensen's inequality, because it implies

$$\mathbf{E}\Big(|m(X) - m_{\beta_n}(X)|^2\Big) \;\leq\; \mathbf{E}\left(\mathbf{E}\Big(|Y - T_{\beta_n} Y|^2 \Big| X\Big)\right) = \mathbf{E}\Big(|Y - T_{\beta_n} Y|^2\Big),$$

which yields

$$|T_{6,n}| \;\leq\; \sqrt{4\mathbf{E}\left(|Y - T_{\beta_n} Y|^2\right)} \cdot c_8 \cdot \log(n).$$

The calculations concerning $T_{5,n}$ furthermore lead to $|T_{6,n}| \leq c_9 \cdot \log(n)/n$, for some constant $c_9 > 0$. Summing up, we have

$$T_{1,n} \leq c_{10} \cdot \frac{\log(n)}{n},$$

for some constant $c_{10} > 0$.

Now, let us consider $T_{2,n}$, and let $t > 1/n$ be arbitrary. Then

$$\begin{aligned}
&\mathbf{P}\{T_{2,n} > t\} \\
&= \mathbf{P}\left\{\frac{1}{2}\left(\mathbf{E}(|m_n(X) - T_{\beta_n} Y|^2 | \mathcal{D}_n) - \mathbf{E}(|m_{\beta_n}(X) - T_{\beta_n} Y|^2)\right)\right. \\
&\qquad \left. -\frac{1}{n}\sum_{i=1}^n \left(|m_n(X_i) - T_{\beta_n} Y_i|^2 - |m_{\beta_n}(X_i) - T_{\beta_n} Y_i|^2\right) > \frac{t}{2}\right\} \\
&= \mathbf{P}\Bigg\{\mathbf{E}(|m_n(X) - T_{\beta_n} Y|^2 | \mathcal{D}_n) - \mathbf{E}(|m_{\beta_n}(X) - T_{\beta_n} Y|^2) \\
&\qquad -\frac{1}{n}\sum_{i=1}^n \left(|m_n(X_i) - T_{\beta_n} Y_i|^2 - |m_{\beta_n}(X_i) - T_{\beta_n} Y_i|^2\right) \\
&\qquad > \frac{1}{2}\left(t + \mathbf{E}(|m_n(X) - T_{\beta_n} Y|^2 | \mathcal{D}_n) - \mathbf{E}(|m_{\beta_n}(X) - T_{\beta_n} Y|^2)\right)\Bigg\} \\
&\leq \mathbf{P}\left\{\exists f \in T_{\beta_n}\mathcal{F}_n : \mathbf{E}\left(\left|\frac{f(X)}{\beta_n} - \frac{T_{\beta_n} Y}{\beta_n}\right|^2\right) - \mathbf{E}\left(\left|\frac{m_{\beta_n}(X)}{\beta_n} - \frac{T_{\beta_n} Y}{\beta_n}\right|^2\right)\right. \\
&\qquad -\frac{1}{n}\sum_{i=1}^n \left(\left|\frac{f(X_i)}{\beta_n} - \frac{T_{\beta_n} Y_i}{\beta_n}\right|^2 - \left|\frac{m_{\beta_n}(X_i)}{\beta_n} - \frac{T_{\beta_n} Y_i}{\beta_n}\right|^2\right) \\
&\qquad \left. > \frac{1}{2}\left(\frac{t}{\beta_n^2} + \mathbf{E}\left(\left|\frac{f(X)}{\beta_n} - \frac{T_{\beta_n} Y}{\beta_n}\right|^2 \Big| \mathcal{D}_n\right) - \mathbf{E}\left(\left|\frac{m_{\beta_n}(X)}{\beta_n} - \frac{T_{\beta_n} Y}{\beta_n}\right|^2\right)\right)\right\}.
\end{aligned}$$

Thus we can deduce from Theorem 1.17 that

$$\mathbf{P}\{T_{2,n} > t\} \leq 14 \sup_{x_1^n} \mathcal{N}_1\left(\frac{t}{80\beta_n^2}, \left\{\frac{1}{\beta_n} f : f \in T_{\beta_n}\mathcal{F}_n\right\}, x_1^n\right) \cdot \exp\left(-\frac{n}{5136 \cdot \beta_n^2} t\right).$$

holds. Note that the required bound in Theorem 1.17 is equal to 1 in this setting, because obviously we have

$$\left|\frac{f(x)}{\beta_n}\right| \leq 1 \text{ for all } x \in \mathbb{R}^d, \quad \text{and} \quad \left|\frac{T_{\beta_n} Y}{\beta_n}\right| \leq 1 \ a.s.$$

Since moreover the inequality

$$\mathcal{N}_1\left(\delta, \left\{\frac{1}{\beta_n} f : f \in \mathcal{F}\right\}, x_1^n\right) \leq \mathcal{N}_1\left(\delta \cdot \beta_n, \mathcal{F}, x_1^n\right),$$

holds for all $x_1^n = (x_1, ..., x_n) \in \mathbb{R}^d \times ... \times \mathbb{R}^d$ we obtain that

$$\mathbf{P}\{T_{2,n} > t\} \leq 14 \sup_{x_1^n} \mathcal{N}_1\left(\frac{t}{80\beta_n}, T_{\beta_n}\mathcal{F}_n, x_1^n\right) \cdot \exp\left(-\frac{n}{5136 \cdot \beta_n^2} t\right).$$

Furthermore we know from Lemma 2.5 that, with $L_n := \max\{L_{1,n}, ..., L_{K_n,n}\}$, for $1/n < t < 40\beta_n$,

$$\begin{aligned}\mathcal{N}_1\left(\frac{t}{80\beta_n}, T_{\beta_n}\mathcal{F}_n, x_1^n\right) &\leq 3\left(\frac{6e\beta_n \cdot 80\beta_n \cdot K_n L_n}{t}\right)^{2(d+2)\left(\sum_{k=1}^{K_n} L_{k,n}\right)} \\ &\leq n^{c_{11} \cdot \sum_{k=1}^{K_n} L_{k,n}}\end{aligned}$$

holds for some sufficient large $c_{11} > 0$. (This inequality holds also for $t \geq 40\beta_n$, since the right-hand side above does not depend on t and the covering number is decreasing in t.) Using this we get for arbitrary $\varepsilon \geq 1/n$

$$\begin{aligned}\mathbf{E}(T_{2,n}) &\leq \varepsilon + \int_\varepsilon^\infty \mathbf{P}\{T_{2,n} > t\} dt \\ &= \varepsilon + 14 \cdot n^{c_{11}\left(\sum_{k=1}^{K_n} L_{k,n}\right)} \frac{5136\beta_n^2}{n} \cdot \exp\left(-\frac{n}{5136\beta_n^2}\varepsilon\right).\end{aligned}$$

This expression is minimized for

$$\varepsilon = \frac{5136 \cdot \beta_n^2}{n} \log\left(14 \cdot n^{c_{11}\left(\sum_{k=1}^{K_n} L_{k,n}\right)}\right).$$

Thus we see

$$\begin{aligned}\mathbf{E}(T_{2,n}) &\leq \frac{5136 \cdot \beta_n^2}{n} \log\left(14 \cdot n^{c_{11} \cdot \left(\sum_{k=1}^{K_n} L_{k,n}\right)}\right) \\ &\quad + 14 \cdot n^{c_{11} \cdot \left(\sum_{k=1}^{K_n} L_{k,n}\right)} \cdot \frac{5136\beta_n^2}{n} \exp\left(-\log\left(14 \cdot n^{c_{11}\left(\sum_{k=1}^{K_n} L_{k,n}\right)}\right)\right) \\ &= \frac{5136 \cdot \beta_n^2}{n}\left(\log(14) + c_{11} \cdot \left(\sum_{k=1}^{K_n} L_{k,n}\right) \cdot \log(n)\right) + \frac{5136 \cdot \beta_n^2}{n} \\ &= \frac{c_{12} \cdot \log(n)^3 \cdot \sum_{k=1}^{K_n} L_{k,n}}{n},\end{aligned}$$

for some sufficiently large constant $c_{12} > 0$, which does not depend on n, β_n or the parameters of the estimate.

By bounding $T_{3,n}$ similarly to $T_{1,n}$ we also deduce

$$\mathbf{E}(T_{3,n}) \quad \leq \quad c_{13} \cdot \frac{\log(n)}{n}$$

for some constant $c_{13} > 0$, which implies

$$\mathbf{E}\left(\sum_{i=1}^{3} T_{i,n}\right) \quad \leq \quad \frac{c_{14} \cdot \log(n)^3 \cdot \sum_{k=1}^{K_n} L_{k,n}}{n},$$

for a suitable constant $c_{14} > 0$.

We finish the proof by bounding $T_{4,n}$. For this purpose, let A_n be the event, that there exists $i \in \{1, ..., n\}$ such that $|Y_i| > \beta_n$, and let $\mathbf{1}_{A_n}$ be the indicator function of A_n. Then

$$\begin{aligned}
\mathbf{E}(T_{4,n}) &\leq 2\,\mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 \cdot \mathbf{1}_{A_n}\right) \\
&\quad + 2\,\mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 \cdot \mathbf{1}_{A_n^C} - \frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2\right) \\
&= 2\,\mathbf{E}\left(|m_n(X_1) - Y_1|^2 \cdot \mathbf{1}_{A_n}\right) \\
&\quad + 2\,\mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 \cdot \mathbf{1}_{A_n^C} - \frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2\right) \\
&= T_{7,n} + T_{8,n}.
\end{aligned}$$

The Cauchy-Schwarz inequality now shows that

$$\begin{aligned}
\mathbf{E}\left(|m_n(X_1) - Y_1|^2 \cdot I_{A_n}\right) &\leq \sqrt{\mathbf{E}\left((|m_n(X_1) - Y_1|^2)^2\right)} \cdot \sqrt{\mathbf{P}(A_n)} \\
&\leq \sqrt{\mathbf{E}\left((2|m_n(X_1)|^2 + 2|Y_1|^2)^2\right)} \cdot \sqrt{n \cdot \mathbf{P}\{|Y_1| > \beta_n\}} \\
&\leq \sqrt{\mathbf{E}\left(8|m_n(X_1)|^4 + 8|Y_1|^4\right)} \cdot \sqrt{n \cdot \frac{\mathbf{E}\left(\exp(c_2 \cdot |Y_1|^2)\right)}{\exp(c_2 \cdot \beta_n^2)}},
\end{aligned}$$

where the last inequality follows from inequality (3.12). Since $x \leq \exp(x)$ holds for all $x \in \mathbb{R}$ we infer

$$\begin{aligned}
\mathbf{E}\left(|Y|^4\right) &= \mathbf{E}\left(|Y|^2 \cdot |Y|^2\right) \leq \mathbf{E}\left(\frac{2}{c_2} \cdot \exp\left(\frac{c_2}{2} \cdot |Y|^2\right) \cdot \frac{2}{c_2} \cdot \exp\left(\frac{c_2}{2} \cdot |Y|^2\right)\right) \\
&= \frac{4}{c_2^2} \cdot \mathbf{E}\left(\exp\left(c_2 \cdot |Y|^2\right)\right),
\end{aligned}$$

which is finite by condition (3.10). Furthermore $||m_n||_\infty$ is bounded by β_n. Therefore the first factor is bounded by

$$c_{15} \cdot \beta_n^2 = c_{16} \cdot \log(n)^2,$$

for a suitable constant $c_{16} > 0$. The second factor is bounded by $1/n$, since by (3.10), $\mathbf{E}\left(\exp\left(c_2 \cdot |Y_1|^2\right)\right)$ is bounded by some constant $c_{17} < \infty$. Hence

$$\sqrt{n \cdot \frac{\mathbf{E}\left(\exp(c_2 \cdot |Y_1|^2)\right)}{\exp(c_2 \cdot \beta_n^2)}} \quad \leq \quad \sqrt{n} \cdot \frac{\sqrt{c_{17}}}{\sqrt{\exp(c_2 \cdot \beta_n^2)}} \leq \frac{\sqrt{n} \cdot \sqrt{c_{17}}}{\exp((c_2 \cdot c_1^2 \cdot \log(n)^2)/2)}.$$

Since $\exp(-c \cdot \log(n)^2) = \mathcal{O}(n^{-2})$ for $c > 0$, this yields

$$T_{7,n} \quad \leq \quad c_{18} \cdot \frac{\log(n)^2 \sqrt{n}}{n^2} \leq c_{19} \cdot \frac{\log(n)^2}{n}.$$

Furthermore the definition of A_n^C together with $\tilde{m}_n$ defined as in (2.2) implies

$$\begin{aligned} T_{8,n} &\leq 2\,\mathbf{E}\left(\frac{1}{n}\sum_{i=1}^n |\tilde{m}_n(X_i) - Y_i|^2 \cdot \mathbf{1}_{A_n^C} - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right) \\ &\leq 2\,\mathbf{E}\left(\frac{1}{n}\sum_{i=1}^n |\tilde{m}_n(X_i) - Y_i|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right) \\ &\leq 2\,\mathbf{E}\left(\inf_{f \in \mathcal{F}_n} \frac{1}{n}\sum_{i=1}^n |f(X_i) - Y_i|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right), \end{aligned}$$

because $|T_\beta z - y| \leq |z - y|$ holds for $|y| \leq \beta$. Hence

$$\begin{aligned} \mathbf{E}(T_{4,n}) &\leq c_{19} \cdot \frac{\log(n)^2}{n} \\ &\quad +2\,\mathbf{E}\left(\inf_{f \in \mathcal{F}_n} \frac{1}{n}\sum_{i=1}^n |f(X_i) - Y_i|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right), \end{aligned}$$

which completes the proof. □

Together with the approximation result in Lemma 2.2, Theorem 3.4 implies the next corollary, which considers the desired rate of convergence of the maxmin estimate.

COROLLARY 3.5. *Assume that the distribution of (X, Y) has the properties that $X \in [a, b]^d$ a.s. for some $a, b \in \mathbb{R}$, that the modified Sub-Gaussian condition*

$$\mathbf{E}(\exp(c_2 \cdot |Y|^2)) < \infty$$

is fulfilled for some constant $c_2 > 0$, and that the regression function m is (p, C)–smooth for some $0 < p \leq 2$ and $C > 1$.

Then the estimate m_n defined by (2.2) and (2.3) with $\beta_n = c_1 \cdot \log(n)$, for some $c_1 > 0$,

$$K_n = \left\lceil C^{\frac{2d}{2p+d}} \cdot \left(\frac{n}{\log(n)^3}\right)^{d/(2p+d)} \right\rceil \quad \textit{and} \quad L_{k,n} = L_k = 2d + 1, \quad (k = 1, ..., K_n),$$

satisfies

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) \quad \leq \quad c_2 \cdot C^{\frac{2d}{2p+d}} \cdot \left(\frac{\log(n)^3}{n}\right)^{\frac{2p}{2p+d}} \qquad (n \geq 2)$$

for some constant $c_2 > 0$, that does not depend on n, β_n, p or C.

PROOF. In Lemma 2.2 we have seen that it is possible to interpolate a given linear spline function at a fixed given set of data points by maxima of minima of linear functions. Hence we obtain that

$$\begin{aligned}
&\mathbf{E}\left(2 \inf_{f \in \mathcal{F}_n} \left(\frac{1}{n}\sum_{i=1}^{n} |f(X_i) - Y_i|^2 - \frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2\right)\right) \\
&\qquad \leq \mathbf{E}\left(2 \inf_{f \in \mathcal{G}} \left(\frac{1}{n}\sum_{i=1}^{n} |f(X_i) - Y_i|^2 - \frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2\right)\right) \\
&\qquad \leq 2 \cdot \inf_{f \in \mathcal{G}} \int |f(x) - m(x)|^2 \mu(dx),
\end{aligned}$$

where $\mathcal{G}$ is the set of functions which contains all continuous piecewise polynomials of degree 1 with respect to an arbitrary partition Π consisting of K_n rectangulars. Next we increase the right-hand side above by choosing Π such that it consists of equivolume cubes. Now we can apply the approximation result from Lemma 1.16, which together with the (p, C)–smoothness of m and Theorem 3.4 yields

$$\begin{aligned}
\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) &\leq c_3 \cdot \frac{K_n \cdot (2d+1) \cdot \log(n)^3}{n} + c_4 \cdot C^2 \cdot K_n^{-\frac{2p}{d}} \\
&\leq c_5 \cdot C^{\frac{2d}{2p+d}} \cdot \left(\frac{\log(n)^3}{n}\right)^{\frac{2p}{2p+d}},
\end{aligned}$$

for some sufficient large constant $c_5 > 0$, where the last inequaltity results from the choice of K_n. Note that the assumption in Theorem 3.4 concerning the boundedness of the regression function m is obviously satisfied. Since we supposed in this corollary that m is (p, C)–smooth we can deduce from $X \in [a, b]^d$ that m is a continuous function with bounded support. □

We have achieved our aim to compute the rate of convergence of our estimate. Moreover, we can deduce that it is the optimal rate of convergence up to the logarithmic factor. However the parameters of the estimate depend on the smoothness of the regression function, and in most applications there are no a-priori informations concerning the smoothness of the underlying regression function. Hence the next section deals with a data–dependent choice of the parameters.

3.3. Splitting the Sample

In most applications the smoothness of the regression function (measured by (p, C)) is not known in advance and therefore, the parameters of the estimate have to be chosen data–dependent. This can be done for example by *cross–validation*, which

in regression estimation goes back to Clark (1975) and Wahba and Wold (1975) or *complexity regularization*, which was used in regression estimation for the first time in Barron (1991). Another well-known technique to choose the parameters data–dependent is *splitting the sample*, where the estimate is computed for various values of the parameters on a learning sample (consisting, for example, of the first half of the data points) and the parameters are chosen such that the empirical L_2 risk on a testing sample (consisting, for example, of the second half of the data points) is minimized. This idea was already used in 1988 by Devroye, who analysed this method in the context of pattern recognition.

In the following we will use this last method to define an estimate, which is adaptive to the given data. This estimate will have the optimal rate of convergence (up to some logarithmic factor), as well. Here again we do not have to assume that Y is bounded, but it is necessary to require the Sub-Gaussian condition.

We have seen in Chapter 2 that the class of functions, and hence also the estimate, depends on the parameters

$$K_n, L_{1,n}, ..., L_{K_n,n} \in \mathbb{N},$$

where K_n declares the number of minimum functions under the maximum and $L_{i,n}$ declares the number of linear functions under the i-th minimum. Obviously for $L_n = \max\{L_{1,n}, ..., L_{K_n,n}\}$ the class of functions

$$\left\{ f : \mathbb{R}^d \to \mathbb{R} \,:\, f(x) = \max_{k=1,...,K_n} \min_{l=1,...,L_n} \left(a_{k,l} \cdot x + b_{k,l}\right), \quad (x \in \mathbb{R}^d), \right. \tag{3.13}$$
$$\left. \text{for some } a_{k,l} \in \mathbb{R}^d, b_{k,l} \in \mathbb{R} \right\}$$

is a superset of $\mathcal{F}_n$, since for example a minimum function of three linear functions can always be written as a minimum function of four, five or six linear functions by choosing the same linear functions twice, three times or four times, respectively. Hence choosing the class of functions defined by (3.13) instead of $\mathcal{F}_n$ in the previous sections changes nothing about the approximation error (actually even decreases it) and give just slightly different results for the covering numbers and therefore negligible modifications with regard to the estimation error.

In consideration of this fact we choose $\mathcal{Q}_n = \mathbb{N}^2$ as the set of parameters and assume in this section that

$$L_{1,n} = \cdots = L_{K_n,n} = L_n.$$

For $h = (h_1, h_2) \in \mathcal{Q}_n$, we write

$$\mathcal{F}_h = \left\{ f : \mathbb{R}^d \to \mathbb{R} \,:\, f(x) = \max_{k=1,\dots,h_1} \min_{l=1,\dots,h_2} (a_{k,l} \cdot x + b_{k,l}), \quad (x \in \mathbb{R}^d), \right.$$
$$\left. \text{for some } a_{k,l} \in \mathbb{R}^d, b_{k,l} \in \mathbb{R} \right\}$$

throughout this section. Furthermore let $n \in \mathbb{N}$ be the sample size, $n_l \in \mathbb{N}$ the size of the learning data $\mathcal{D}_{n_l} = \{(X_1, Y_1), ..., (X_{n_l}, Y_{n_l})\}$, and $n_t \in \mathbb{N}$ the size of the testing data $\mathcal{D}_{n_t} = \{(X_{n_l+1}, Y_{n_l+1}), ..., (X_n, Y_n)\}$. Then we define for every $h \in \mathcal{Q}_n$

$$m_{n_l}^{(h)}(\cdot) = T_{\beta_n} \arg\min_{f \in \mathcal{F}_h} \frac{1}{n_l} \sum_{i=1}^{n_l} |f(X_i) - Y_i|^2. \tag{3.14}$$

That means for every parameter h we compute a regression estimate by using the principle of least squares over the class of functions $\mathcal{F}_h$ and a subsequent truncation. Afterwards we choose $H \in \mathcal{Q}_n$ such that

$$\frac{1}{n_t} \sum_{i=n_l+1}^{n} |m_{n_l}^{(H)}(X_i) - Y_i|^2 = \min_{h \in \mathcal{Q}_n} \frac{1}{n_t} \sum_{i=n_l+1}^{n} |m_{n_l}^{(h)}(X_i) - Y_i|^2. \tag{3.15}$$

More precisely, we choose an estimate that minimizes the empirical L_2 risk on the testing data over all estimates that we have computed with respect to the learning sample, that is

$$m_n(x) = m_{n_l}^{(H)}(x), \qquad \text{for all } x \in \mathbb{R}^d. \tag{3.16}$$

In order to get a result concerning the rate of convergence of the above defined estimate, we prove the following general theorem:

THEOREM 3.6. *Let $\beta_n = c_1 \cdot \log(n)$ for some constant $c_1 > 0$. Assume that the distribution of (X, Y) satisfies the modified Sub-Gaussian condition*

$$\mathbf{E}\left(e^{c_2 \cdot |Y|^2}\right) < \infty,$$

for some constant $c_2 > 0$, and that the regression function fulfils

$$||m||_\infty \leq L, \qquad \text{for some } L \in \mathbb{R}^+, \text{ with } L \leq \beta_n.$$

Then, for every estimate defined by (3.15) and (3.16) with repect to a family of regression estimates

$$\left(m_n^{(h)}\right)_{h \in \mathcal{Q}_n}, \quad \text{with} \quad ||m_n^{(h)}||_\infty \leq \beta_n \quad \text{for all } h \in \mathcal{Q}_n,$$

where $\mathcal{Q}_n$ is the underlying set of parameters, we get for all $\delta > 0$ that

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx)$$
$$\leq (1+\delta) \min_{h \in \mathcal{Q}_n} \mathbf{E} \int |m_{n_l}^{(h)}(x) - m(x)|^2 \mu(dx) + c_3 \cdot \frac{1 + \log|\mathcal{Q}_n|}{n_t} + c_4 \frac{\log(n)}{n}$$

holds, with $c_3 \geq \beta_n^2(32/\delta + 70 + 39\delta)$ and a sufficiently large constant $c_4 > 0$.

PROOF. We use the following error decomposition

$$\begin{aligned}
&\mathbf{E}\left(\int |m_n(x)-m(x)|^2\mu(dx)\Big|\,\mathcal{D}_{n_l}\right)\\
&\quad=\mathbf{E}\left(\int |m_{n_l}^{(H)}(x)-m(x)|^2\mu(dx)\Big|\,\mathcal{D}_{n_l}\right)\\
&\quad=\Bigg[\mathbf{E}\left(|m_{n_l}^{(H)}(X)-Y|^2\Big|\,\mathcal{D}_{n_l}\right)-\mathbf{E}\left(|m(X)-Y|^2\right)\\
&\qquad-\mathbf{E}\left(|m_{n_l}^{(H)}(X)-Y|^2\Big|\,\mathcal{D}_{n_l}\right)-\mathbf{E}\left(|m_{\beta_n}(X)-T_{\beta_n}Y|^2\right)\Bigg]\\
&\quad+\Bigg[\mathbf{E}\left(|m_{n_l}^{(H)}(X)-T_{\beta_n}Y|^2\Big|\,\mathcal{D}_{n_l}\right)-\mathbf{E}\left(|m_{\beta_n}(X)-T_{\beta_n}Y|^2\right)\\
&\qquad-(1+\delta)\cdot\frac{1}{n_t}\sum_{i=n_l+1}^{n}\left(|m_{n_l}^{(H)}(X_i)-T_{\beta_n}Y_i|^2-|m_{\beta_n}(X_i)-T_{\beta_n}Y_i|^2\right)\Bigg]\\
&\quad+\Bigg[(1+\delta)\cdot\frac{1}{n_t}\sum_{i=n_l+1}^{n}\left(|m_{n_l}^{(H)}(X_i)-T_{\beta_n}Y_i|^2-|m_{\beta_n}(X_i)-T_{\beta_n}Y_i|^2\right)\\
&\qquad-\left((1+\delta)\cdot\frac{1}{n_t}\sum_{i=n_l+1}^{n}\left(|m_{n_l}^{(H)}(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\right)\right)\Bigg]\\
&\quad+\left[(1+\delta)\cdot\frac{1}{n_t}\sum_{i=n_l+1}^{n}\left(|m_{n_l}^{(H)}(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\right)\right]=\sum_{i=1}^{4}T_{i,n},
\end{aligned}$$

where again $T_{\beta_n}Y$ is the truncated version of Y, and m_{β_n} is the regression function of $T_{\beta_n}Y$, that is

$$m_{\beta_n}(x)=\mathbf{E}\left\{T_{\beta_n}Y\big|X=x\right\}.$$

Due to equality (3.15) we can bound the last term $T_{4,n}$ by

$$(1+\delta)\min_{h\in\mathcal{Q}_n}\left(\frac{1}{n_t}\sum_{i=n_l+1}^{n}\left(|m_{n_l}^{(h)}(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\right)\right),$$

and this entails for its conditional expectation

$$\begin{aligned}
\mathbf{E}(T_{4,n}|\mathcal{D}_{n_l}) &\le (1+\delta)\min_{h\in\mathcal{Q}_n}\left(\mathbf{E}\left(|m_{n_l}^{(h)}(X)-Y|^2\Big|\,\mathcal{D}_{n_l}\right)-\mathbf{E}\left(|m(X)-Y|^2\right)\right)\\
&= (1+\delta)\min_{h\in\mathcal{Q}_n}\int |m_{n_l}^{(h)}(x)-m(x)|^2\mu(dx).
\end{aligned}$$

$T_{1,n}$ and $T_{3,n}$ can be bounded analogously to the corresponding terms in the proof of Theorem 3.4, since all relations and assumptions we have used in that proof (such as the Sub-Gaussian condition, $\beta_n=\mathcal{O}(\log(n))$ and the boundedness of m and m_{β_n}) are satisfied in the current settings as well. Thus we have

$$T_{1,n}\le c_5\cdot\frac{\log(n)}{n}\quad\text{und}\quad\mathbf{E}(T_{3,n}|\mathcal{D}_{n_l})\le c_6\cdot\frac{\log(n)}{n},$$

for sufficiently large constants c_5, c_6. Hence it suffices to show

$$\mathbf{E}(T_{2,n}|\mathcal{D}_{n_l}) \leq c_3 \cdot \frac{1+\log(|\mathcal{Q}_n|)}{n_t}$$

to complete this proof. Thus, let $s > 0$ be an arbitrary constant. Then

$$\begin{aligned}
&\mathbf{P}\Big\{T_{2,n} \geq s \mid \mathcal{D}_{n_l}\Big\} \\
&= \mathbf{P}\Bigg\{(1+\delta)\Big(\mathbf{E}\left(|m_{n_l}^{(H)}(X) - T_{\beta_n}Y|^2 \mid \mathcal{D}_{n_l}\right) - \mathbf{E}\left(|m_{\beta_n}(X) - T_{\beta_n}Y|^2\right) \\
&\qquad\qquad -\frac{1}{n_t}\sum_{i=n_l+1}^{n}\left(m_{n_l}^{(H)}(X_i) - T_{\beta_n}Y_i|^2 - |m_{\beta_n}(X_i) - T_{\beta_n}Y_i|^2\right)\Big) \\
&\qquad \geq s + \delta\Big(\mathbf{E}\left(|m_{n_l}^{(H)} - T_{\beta_n}Y|^2|\mathcal{D}_{n_l}\right) - \mathbf{E}\left(|m_{\beta_n}(X) - T_{\beta_n}Y|^2\right)\Big)\Big|\mathcal{D}_{n_l}\Bigg\} \\
&\leq \mathbf{P}\Bigg\{\exists h \in \mathcal{Q}_n : \mathbf{E}\left(|m_{n_l}^{(h)}(X) - T_{\beta_n}Y|^2 \mid \mathcal{D}_{n_l}\right) - \mathbf{E}\left(|m_{\beta_n}(X) - T_{\beta_n}Y|^2\right) \\
&\qquad\qquad -\frac{1}{n_t}\sum_{i=n_l+1}^{n}\left(m_{n_l}^{(h)}(X_i) - T_{\beta_n}Y_i|^2 - |m_{\beta_n}(X_i) - T_{\beta_n}Y_i|^2\right) \\
&\qquad \geq \frac{1}{1+\delta}\Big(s + \delta\cdot\mathbf{E}\left(|m_{n_l}^{(h)} - T_{\beta_n}Y|^2 - |m_{\beta_n}(X) - T_{\beta_n}Y|^2|\mathcal{D}_{n_l}\right)\Big)\Big|\mathcal{D}_{n_l}\Bigg\} \\
&\leq |\mathcal{Q}_n|\cdot\max_{h\in\mathcal{Q}_n}\mathbf{P}\Bigg\{\Big(\mathbf{E}\Big(|m_{n_l}^{(h)}(X) - T_{\beta_n}Y|^2 \mid \mathcal{D}_{n_l}\Big) - \mathbf{E}\left(|m_{\beta_n}(X) - T_{\beta_n}Y|^2\right) \\
&\qquad\qquad -\frac{1}{n_t}\sum_{i=n_l+1}^{n}\left(m_{n_l}^{(h)}(X_i) - T_{\beta_n}Y_i|^2 - |m_{\beta_n}(X_i) - T_{\beta_n}Y_i|^2\right)\Big) \\
&\qquad \geq \frac{1}{1+\delta}\Big(s + \delta\cdot\mathbf{E}\left(|m_{n_l}^{(h)} - T_{\beta_n}Y|^2 - |m_{\beta_n}(X) - T_{\beta_n}Y|^2|\mathcal{D}_{n_l}\right)\Big)\Big|\mathcal{D}_{n,l}\Bigg\}.
\end{aligned}$$

In order to get bounds on this probability we consider, for a fixed $h \in \mathcal{Q}_n$ the random variables $Z, Z_1, ..., Z_{n_t}$, which are defined by

$$Z = |m_{n_l}^{(h)}(X) - T_{\beta_n}Y|^2 - |m_{\beta_n}(X) - T_{\beta_n}Y|^2,$$

and

$$Z_i = |m_{n_l}^{(h)}(X_{n_l+i}) - T_{\beta_n}Y_{n_l+i}|^2 - |m_{\beta_n}(X_{n_l+i}) - T_{\beta_n}Y_{n_l+i}|^2, \quad (i = 1, ..., n_t).$$

For these random variables we obtain

$$\begin{aligned}
\sigma^2 &= \mathbf{Var}(Z|\mathcal{D}_{n_l}) \\
&\leq \mathbf{E}(Z^2|\mathcal{D}_{n_l}) \\
&= \mathbf{E}\Bigg(\Big|\Big(m_{n_l}^{(h)}(X) - T_{\beta_n}Y\Big) - (m_{\beta_n}(X) - T_{\beta_n}Y)\Big|^2 \\
&\qquad \times\Big|\Big(m_{n_l}^{(h)}(X) - T_{\beta_n}Y\Big) + (m_{\beta_n}(X) - T_{\beta_n}Y)\Big|^2\Big|\mathcal{D}_{n_l}\Bigg),
\end{aligned}$$

where the last equality is an application of the binomial theorem. Futhermore every single term in the second factor of the right-hand side of the above inequality is bounded by β_n by assumption. This implies

$$\begin{aligned}\sigma^2 &= \mathbf{Var}(Z|\mathcal{D}_{n_l}) \\ &\leq 16\beta_n^2 \int |m_{n_l}^{(h)}(x) - m_{\beta_n}(x)|^2 \mu(dx) \\ &= 16\beta_n^2\, \mathbf{E}(Z|\mathcal{D}_{n_l}).\end{aligned}$$

Now we can rewrite the probability above and get

$$\begin{aligned}&\mathbf{P}\Bigg\{ \Big(\mathbf{E}\Big(|m_{n_l}^{(h)}(X) - T_{\beta_n}Y|^2 \mid \mathcal{D}_{n_l}\Big) - \mathbf{E}\left(|m_{\beta_n}(X) - T_{\beta_n}Y|^2\right) \\ &\qquad -\frac{1}{n_t}\sum_{i=n_l+1}^{n}\Big(m_{n_l}^{(h)}(X_i) - T_{\beta_n}Y_i|^2 - |m_{\beta_n}(X_i) - T_{\beta_n}Y_i|^2\Big)\Big) \\ &\quad \geq \left(\frac{1}{1+\delta}\right)\Big(s + \delta\cdot\mathbf{E}\left(|m_{n_l}^{(h)} - T_{\beta_n}Y|^2 - |m_{\beta_n}(X) - T_{\beta_n}Y|^2|\mathcal{D}_{n_l}\right)\Big)\Big|\mathcal{D}_{n_l}\Bigg\} \\ &= \mathbf{P}\left\{\mathbf{E}\left(Z|\mathcal{Q}_n\right) - \frac{1}{n_t}\sum_{i=n_l+1}^{n} Z_i \geq \frac{1}{1+\delta}\Big(s + \delta\cdot\mathbf{E}\left(Z|\mathcal{D}_{n,l}\right)\Big)\Big|\mathcal{D}_{n_l}\right\} \\ &\leq \mathbf{P}\left\{\mathbf{E}\left(Z|\mathcal{Q}_n\right) - \frac{1}{n_t}\sum_{i=n_l+1}^{n} Z_i \geq \frac{1}{1+\delta}\Big(s + \delta\cdot\frac{\sigma^2}{16\beta_n}\Big)\Big|\mathcal{D}_{n_l}\right\} \\ &\leq \exp\left(-n_t\frac{\left(\frac{1}{1+\delta}\left(s+\delta\frac{\sigma^2}{16\beta_n}\right)\right)^2}{2\sigma^2 + \frac{2}{3}\frac{8\beta_n^2}{1+\delta}\left(s+\delta\frac{\sigma^2}{16\beta_n}\right)}\right).\end{aligned}$$

The last inequality is a direct consequence of Bernstein's inequality. Note that we do not need the factor 2 in the exponential term owing to the absence of the absolute value inside the probability. The following calculation permits the desired bounding of the above probability. Due to $\delta > 0$ we obtain

$$\begin{aligned}&\frac{1}{(1+\delta)^2}\,\frac{\left(s+\delta\frac{\sigma^2}{16\beta_n}\right)^2}{2\sigma^2 + \frac{2}{3}\frac{8\beta_n^2}{1+\delta}\left(s+\delta\frac{\sigma^2}{16\beta_n}\right)} \\ &\qquad \geq \frac{1}{(1+\delta)^2}\frac{\left(s+\delta\frac{\sigma^2}{16\beta_n}\right)^2}{2\frac{16\beta_n}{\delta}\left(\frac{\delta}{16\beta_n}\sigma^2 + s\right) + \frac{2}{3}\frac{8\beta_n^2}{1+\delta}\left(s+\delta\frac{\sigma^2}{16\beta_n}\right)} \\ &\qquad = \frac{1}{(1+\delta)^2}\frac{s+\delta\frac{\sigma^2}{16\beta_n}}{\frac{32\beta_n}{\delta} + \frac{16\beta_n^2}{3(1+\delta)}} \geq \frac{1}{(1+\delta)^2}\frac{s}{\frac{32\beta_n}{\delta} + \frac{6\beta_n^2}{1+\delta}} \\ &\qquad = \frac{s}{\frac{32\beta_n}{\delta} + 64\beta_n + 32\beta_n\delta + 6\beta_n^2 + 6\beta_n^2\delta} \geq \frac{s}{\beta_n^2\left(32/\delta + 70 + 39\delta\right)} \geq \frac{s}{c_3},\end{aligned}$$

where $c_3 \geq \beta_n^2(32/\delta + 70 + 39\delta)$. Combining the previous inequalities yields

$$\mathbf{P}\left\{T_{2,n} \geq s \middle| \mathcal{D}_{n_l}\right\} \leq |\mathcal{Q}_n| \cdot \exp\left(-n_t \frac{s}{c_3}\right).$$

For $u > 0$ this leads to

$$\begin{aligned}\mathbf{E}\left(T_{2,n}\middle|\mathcal{D}_{n_l}\right) &\leq u + \int_u^\infty \mathbf{P}\left\{T_{2,n} > s \middle| \mathcal{D}_{n_l}\right\} ds \\ &\leq u + \frac{|\mathcal{Q}_n| \cdot c_3}{n_t} \cdot \exp\left(-\frac{n_t u}{c_3}\right),\end{aligned}$$

and hence, with $u = c_3 \cdot \log(|\mathcal{Q}_n|)/n_t$ to

$$\mathbf{E}\left(T_{2,n}\middle|\mathcal{D}_{n_l}\right) \leq c_3 \, \frac{1 + \log|\mathcal{Q}_n|}{n_t},$$

which implies the assertion and completes this proof. □

In view of Theorem 3.6 it is now easy to obtain the rate of convergence of the estimate defined by (3.14) – (3.16). However we still have to make some restrictions concerning the smoothness of the regression function, similarly as we did in the Corollary 3.5, where we considered the rate of convergence of the maxmin estimate with a certain choice of parameters.

COROLLARY 3.7. *Suppose that the distribution of (X, Y) has the properties that $X \in [a, b]^d$ a.s. for some $a, b \in \mathbb{R}$, that the modified Sub-Gaussian condition*

$$\mathbf{E}\left(e^{c_2 \cdot |Y|^2}\right) < \infty,$$

is fulfilled for some constant $c_2 > 0$ and that the regression function m is (p, C)–smooth, for some $0 < p \leq 2$ and $C > 1$. Furthermore choose the set of parameters $\mathcal{Q}_n$ in such a way that for $n \geq 2$

$$\log|\mathcal{Q}_n| \leq c_3 \cdot \log(n).$$

Then the estimate m_n defined by (3.14) – (3.16) with $\beta_n = c_1 \cdot \log(n)$, for some constant $c_1 > 0$ and with $n_l = \lceil \frac{n}{2} \rceil$ and $n_t = n - n_l$, satisfies

$$\begin{aligned}&\mathbf{E}\int |m_n(x) - m(x)|^2 \mu(dx) \\ &\quad\leq c_4 \cdot C^{(2d)/(2p+d)} \left(\frac{\log(n)^3}{n}\right)^{(2p)/(2p+d)} + c_5 \frac{\log(n)^3}{n} \qquad (n \geq 2),\end{aligned}$$

for constants c_4, c_5 chosen sufficiently large.

PROOF. Obviously the assumptions from Theorem 3.6 are satisfied. Particularly the boundedness of m can be deduced again from its (p, C)–smoothness. Thus

we obtain with Theorem 3.6 that for $\delta = 1$

$$\begin{aligned}
&\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) \\
&\leq\ 2 \min_{h \in \mathcal{Q}_n} \mathbf{E} \int |m_{n_l}^{(h)}(x) - m(x)|^2 \mu(dx) + c_7 \cdot \frac{1 + \log |\mathcal{Q}_n|}{n_t} + c_8 \frac{\log(n)}{n} \\
&\leq\ 2 \min_{h \in \mathcal{Q}_n} \mathbf{E} \int |m_{n_l}^{(h)}(x) - m(x)|^2 \mu(dx) + c_9 \cdot \frac{\log(n) \cdot \beta_n^2}{n_t} + c_8 \frac{\log(n)}{n} \\
&\leq\ 2 \min_{h \in \mathcal{Q}_n} \mathbf{E} \int |m_{n_l}^{(h)}(x) - m(x)|^2 \mu(dx) + c_{10} \cdot \frac{\log(n)^3}{n_t} + c_8 \frac{\log(n)}{n},
\end{aligned}$$

holds for sufficiently large constants $c_7, ..., c_{10}$. Here the second inequality follows from the lower bound on c_7 in Theorem 3.6, that is $c_7 \geq \beta_n^2(32/\delta + 70 + 39\delta)$, and the requirement $\log |Q_n| \leq c_3 \cdot \log(n)$. Corollary 3.5 implies with K_{n_l} chosen as in that corollary

$$\begin{aligned}
\min_{h \in \mathcal{Q}_n} \mathbf{E} \int |m_{n_l}^{(h)}(x) - m(x)|^2 \mu(dx) &\leq \mathbf{E} \int |m_{n_l}^{(K_{n_l}, 2d+1)}(x) - m(x)|^2 \mu(dx) \\
&\leq c_5 \cdot C^{(2d)/(2p+d)} \left(\frac{\log(n_l)^3}{n_l} \right)^{(2p)/(2p+d)},
\end{aligned}$$

for sufficiently large $c_5 > 0$, which in turn implies the assertion, since we have chosen $n_l = \lceil \frac{n}{2} \rceil$ and $n_t = n - n_l$. □

With regard to (1.11) is easy to see that Corollary 3.7 implies that the corresponding estimate with data–dependent parameter choice also has the optimal rate of convergence (up to some logarithmic factor).

In this chapter we have analysed the asymptotic behaviour of the maxmin estimate. Firstly we have shown that this estimate is strongly universally consistent for all distributions of (X, Y) which satisfy with $X \in [0, 1]^d$ $a.s.$ After that we have derived an upper bound for the expected L_2 error of our estimate under the assumption of the modified Sub-Gaussian condition and, accordingly, we obtained a rate of convergence result for distributions of (X, Y) which satisfy that $X \in [a, b]^d$ and that the belonging regression function m is (p, C)–smooth. Thirdly we have shown that for the estimate with data-dependent choice of parameters, a similar rate of convergence holds under the same assumptions on the distribution of (X, Y).

However the above rates of convergence are not completely satisfactory in the case of large dimension d of the predictor variable X. In the next chapter we will see that it is possible to circumvent this 'curse of dimensionality' by assuming that the regression function has a particular structur. More precisely, we shall see that under the assumptions of single index models, our estimate will attain the one-dimensional rate of convergence, even in the case of a large dimension d.

CHAPTER 4

Dimension Reduction

As already discussed in Section 1.3, the lower minimax rate of convergence for the estimation of a (p, C)–smooth regression function is $n^{-2p/(2p+d)}$. Therefore, even regression estimates with an optimal rate of convergence converge to zero quite slowly if the dimension d of the predictor variable X is large. The only way to achieve a rate of convergence which is independent of the dimension d, or in other words, to achieve the one–dimensional rate of convergence in the case of d–dimensional X, is to impose restrictions on the regression function.

In this chapter we shall present results in terms of *single index models* and *projection pursuit*, which are based on structural assumptions on the regression function. Section 4.1 considers the maxmin estimate from Chapter 2 and provides its rate of convergence in single index models. In Section 4.2 we use an estimate different from the estimate defined by (2.2) and (2.3), in order to discuss the rate of convergence in the setting of projection pursuit.

4.1. Single Index Models

For so–called single index models, one assumes that the regression function m can be written as

$$m(x) = \bar{m}(\alpha \cdot x), \qquad (x \in \mathbb{R}^d) \tag{4.1}$$

where $\bar{m} : \mathbb{R} \to \mathbb{R}$ and $\alpha \in \mathbb{R}^d$. Furthermore, we actually will consider regression functions of the form (4.1), with (p, C)–smooth $\bar{m}$. These structural assumption certainly are quite restrictive. In particular functions defined by (4.1) cannot approximate all measurable functions arbitrarily closely and therefore of course not even all (p, C)–smooth functions.

On the other hand a function of the form (4.1) changes only in one direction α_i, $1 \leq i \leq d$ and moreover, the behaviour in this direction can be described by an (p, C)–smooth function, which makes the estimate relatively easy to interpret. The next corollary asserts the rate of convergence of the estimate from Chapter 2 in single index models.

COROLLARY 4.1. *Assume that the distribution of (X, Y) has the properties that $X \in [a, b]^d$ a.s. for some $a, b \in \mathbb{R}$ and that the modified Sub-Gaussian condition*

$$\mathbf{E}(\exp(c_2 \cdot |Y|^2)) < \infty$$

is fulfilled for some constant $c_2 > 0$. Furthermore assume that the regression function m satisfies

$$m(x) = \bar{m}(\alpha \cdot x) \qquad (x \in \mathbb{R}^d),$$

where $\bar{m} : \mathbb{R} \to \mathbb{R}$ is (p, C)–smooth with $0 < p \leq 2$, $C > 1$ and $\alpha \in \mathbb{R}^d$ with $||\alpha|| = 1$. Then, for the estimate m_n defined by (2.2) and (2.3) with $\beta_n = c_1 \cdot \log(n)$, for some $c_1 > 0$,

$$K_n = \left\lceil C^{\frac{2}{2p+1}} \cdot \left(\frac{n}{\log(n)^3}\right)^{1/(2p+1)} \right\rceil \quad \text{and} \quad L_{k,n} = L_k = 3 \quad (k = 1, ..., K_n),$$

we have

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) \quad \leq \quad c_3 \cdot C^{\frac{2}{2p+1}} \cdot \left(\frac{\log(n)^3}{n}\right)^{\frac{2p}{2p+1}}, \qquad (n \geq 2)$$

for a sufficiently large constant c_3.

PROOF. In the current settings, Theorem 3.4 holds obviously and therefore, we have for $c_4 > 0$ sufficiently large,

$$\begin{aligned}
&\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) \\
&\quad \leq \frac{c_4 \cdot \log(n)^3 \cdot \sum_{k=1}^{K_n} L_{k,n}}{n} \\
&\quad\quad + \mathbf{E}\left(2 \inf_{f \in \mathcal{F}_n} \left(\frac{1}{n} \sum_{i=1}^n |f(X_i) - Y_i|^2 - \frac{1}{n} \sum_{i=1}^n |m(X_i) - Y_i|^2\right)\right).
\end{aligned}$$

By the assumptions on the regression function, the second term on the right-hand side is equal to

$$\mathbf{E}\left(2 \inf_{f \in \mathcal{F}_n} \left(\frac{1}{n} \sum_{i=1}^n |f(X_i) - Y_i|^2 - \frac{1}{n} \sum_{i=1}^n |\bar{m}(\alpha \cdot X_i) - Y_i|^2\right)\right),$$

and with the notation

$$\mathcal{F}_n^1 := \left\{ \max_{k=1,...,K_n} \min_{l=1,...,L_k} a_{k,l} \cdot x + b_{k,l}, \text{ for some } a_{k,l}, b_{k,l} \in \mathbb{R} \right\} \tag{4.2}$$

this expected value is less than or equal to

$$\mathbf{E}\left(2 \inf_{h \in \mathcal{F}_n^1} \left(\frac{1}{n} \sum_{i=1}^n |h(\alpha \cdot X_i) - Y_i|^2 - \frac{1}{n} \sum_{i=1}^n |\bar{m}(\alpha \cdot X_i) - Y_i|^2\right)\right),$$

because

$$f(x) = h(\alpha \cdot x), \qquad (x \in \mathbb{R}^d)$$

is contained in $\mathcal{F}_n$ for every function $h \in \mathcal{F}_n^1$ and every vector $\alpha \in \mathbb{R}^d$. Moreover suppose that $\mathcal{G}$ is the set of all continuous piecewise linear functions $g : \mathbb{R} \to \mathbb{R}$ with respect to a partition of $[\hat{a}, \hat{b}]$ consisting of K_n intervals. Then together with Lemma 2.2 this yields

$$\begin{aligned}
&\mathbf{E}\left(2 \inf_{f\in\mathcal{F}_n}\left(\frac{1}{n}\sum_{i=1}^n |f(X_i)-Y_i|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i)-Y_i|^2\right)\right)\\
&\quad\le \mathbf{E}\left(2 \inf_{h\in\mathcal{G}}\left(\frac{1}{n}\sum_{i=1}^n |h(\alpha\cdot X_i)-Y_i|^2 - \frac{1}{n}\sum_{i=1}^n |\bar{m}(\alpha\cdot X_i)-Y_i|^2\right)\right)\\
&\quad\le 2\cdot \inf_{h\in\mathcal{G}} \int |h(\alpha\cdot x)-\bar{m}(\alpha\cdot x)|^2 \mu(dx)\\
&\quad\le 2\cdot \inf_{h\in\mathcal{G}}\left(\max_{x\in[a,b]^d} |h(\alpha\cdot x)-\bar{m}(\alpha\cdot x)|^2\right)\\
&\quad\le 2\cdot \inf_{h\in\mathcal{G}}\left(\max_{x\in[\hat{a},\hat{b}]} |h(x)-\bar{m}(x)|^2\right).
\end{aligned}$$

Here $[\hat{a}, \hat{b}]$ is chosen such that $\alpha\cdot x \in [\hat{a}, \hat{b}]$ for $x \in [a, b]^d$. Apparently the choice of an equidistant partition increases this upper bound and therefore, the approximation result from Lemma 1.16 implies, for some sufficiently large constant c_5

$$\mathbf{E}\left(2 \inf_{f\in\mathcal{F}_n}\left(\frac{1}{n}\sum_{i=1}^n |f(X_i)-Y_i|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i)-Y_i|^2\right)\right) \;\le\; c_5\cdot C^2\cdot K_n^{-2p}.$$

Summarising the above arguments leads to

$$\begin{aligned}
\mathbf{E}\int |m_n(x)-m(x)|^2\mu(dx) &\le \frac{c_4\cdot\log(n)^3\cdot\sum_{k=1}^{K_n} L_{k,n}}{n} + c_5\cdot C^2\cdot K_n^{-2p}\\
&\le \frac{c_6}{n}\,\log(n)^3\cdot 3C^{2/(2p+1)}\left(\frac{n}{\log(n)^3}\right)^{1/(2p+1)}\\
&\quad + c_5\cdot C^2\cdot\left(C^{2/(2p+1)}\left(\frac{n}{\log(n)^3}\right)^{1/(2p+1)}\right)^{-2p}\\
&\le c_3\cdot C^{2/(2p+1)}\cdot\left(\frac{\log(n)^3}{n}\right)^{\frac{2p}{2p+1}},
\end{aligned}$$

for $n \ge 2$. Here the second inequality follows from the choice of the parameters K_n and $L_{1,n}, ..., L_{K_n,n}$ for a suitable constant $c_6 > 0$. □

Thus, our maxmin estimate indeed achieves the one–dimensional rate of convergence in single index models and therefore it circumvent the curse of dimensionality in this setting. Note that an adaptive parameter choice via splitting the sample is also possible in the setting of single index models. In an analogous manner as in Section 3.3 one can prove that the maxmin estimate with data-dependent parameter choice also achieves the one–dimensional rate of convergence under the

assumptions of the single index model. However, we have already discussed the relatively strong restrictions of single index models and hence we consider a more general setting in the next section.

4.2. Projection Pursuit

The idea of projection pursuit, which was proposed by Friedman and Tukey (1974) and by Friedman and Stuetzle (1981), is to assume that the regression function is of the form

$$m(x) = \sum_{j=1}^{K} m_j(\alpha_j \cdot x), \qquad (x \in \mathbb{R}^d) \tag{4.3}$$

where $\alpha_j \in \mathbb{R}^d$ and $m_j : \mathbb{R} \to \mathbb{R}$. That is, the regression function is a sum of univariate functions, which are applied to the projection of x onto $\alpha_j \in \mathbb{R}^d$.

Obviously projection pursuit is more general than single index models. Actually we will see in Lemma 5.4, that every square integrable function can be approximated arbitrarily closely (with respect to the L_2 norm) by functions defined by (4.3) . However, it is much more difficult to fit functions of the form (4.3) to a set of data than fitting a function of the form (4.1) to the data. Nevertheless, from the theoretical point of view it is of course reasonable to analyse estimates in the context of projection pursuit, although we cannot use the maxmin estimate from Chapter 2.

In the case of projection pursuit we will analyse the estimate, defined by

$$m_n : \mathbb{R}^d \to \mathbb{R}, \quad m_n(x) = \sum_{k=1}^{K} f_k(x), \text{ with } f_1 \in T_{\beta_n}\mathcal{F}_n, ..., f_K \in T_{\beta_n}\mathcal{F}_n, \tag{4.4}$$

such that

$$\frac{1}{n}\sum_{i=1}^{n}\left|\sum_{k=1}^{K} f_k(X_i) - Y_i\right|^2 = \min_{g \in \bigoplus_{i=1}^{K} T_{\beta_n}\mathcal{F}_n} \frac{1}{n}\sum_{i=1}^{n} |g(X_i) - Y_i|^2. \tag{4.5}$$

Here $\bigoplus_{k=1}^{K} T_{\beta_n}\mathcal{F}_n$ denotes the class of functions given by

$$\bigoplus_{k=1}^{K} T_{\beta_n}\mathcal{F}_n = \left\{ g : \mathbb{R}^d \to \mathbb{R}, \; g(x) = \sum_{k=1}^{K} g_k(x), \; (x \in \mathbb{R}^d), \right. \tag{4.6}$$

$$\left. \text{for some } g_k \in T_{\beta_n}\mathcal{F}_n, 1 \le k \le K \right\}.$$

Thus, the so–defined estimate fits a sum of truncated maxmin functions to the data by using the principle of least squares. Next we will see an result concerning the rate of convergence of this estimate under standard smoothness conditions.

THEOREM 4.2. *Let $\beta_n = c_1 \log(n)$ for some $c_1 > 0$. Suppose that the distribution of (X, Y) satifies $X \in [a,b]^d$ a.s. for some $a, b \in \mathbb{R}$ and that the modified Sub-Gaussian condition*

$$\mathbf{E}(\exp(c_2 \cdot |Y|^2)) < \infty,$$

is fulfilled, for some constant $c_2 > 0$. Furthermore assume that the regression function m satisfies

$$m(x) = \sum_{j=1}^{K} m_j(\alpha_j \cdot x), \qquad (x \in \mathbb{R}^d)$$

for some (p, C)–smooth functions $m_j : \mathbb{R} \to \mathbb{R}$ and some $\alpha_j \in \mathbb{R}^d$. Then the estimate m_n defined by (4.4) and (4.5) with

$$K_n = \left\lceil C^{\frac{2}{2p+1}} \cdot \left(\frac{n}{\log(n)^3} \right)^{1/(2p+1)} \right\rceil \quad \text{and} \quad L_{k,n} = 2d+1, \ (k = 1, ..., K_n),$$

satisfies

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) \leq c_3 \cdot C^{\frac{2}{2p+1}} \cdot \left(\frac{\log(n)^3}{n} \right)^{\frac{2p}{2p+1}},$$

for a sufficiently large constant c_3.

PROOF. We use the error decomposition

$$\begin{aligned}
&\int |m_n(x) - m(x)|^2 \mu(dx) \\
&= \Bigg[\mathbf{E}\{|m_n(X) - Y|^2 | \mathcal{D}_n\} - \mathbf{E}\{|m(X) - Y|^2\} \\
&\qquad - \mathbf{E}\{|m_n(X) - T_\beta Y|^2 | \mathcal{D}_n\} - \mathbf{E}\{|m_\beta(X) - T_\beta Y|^2\} \Bigg] \\
&\quad + \Bigg[\mathbf{E}\{|m_n(X) - T_\beta Y|^2 | \mathcal{D}_n\} - \mathbf{E}\{|m_\beta(X) - T_\beta Y|^2\} \\
&\qquad - 2 \cdot \frac{1}{n} \sum_{i=1}^{n} \left(|m_n(X_i) - T_\beta Y_i|^2 - |m_\beta(X_i) - T_\beta Y_i|^2 \right) \Bigg] \\
&\quad + \Bigg[2 \cdot \frac{1}{n} \sum_{i=1}^{n} |m_n(X_i) - T_\beta Y_i|^2 - 2 \cdot \frac{1}{n} \sum_{i=1}^{n} |m_\beta(X_i) - T_\beta Y_i|^2 \\
&\qquad - \left(2 \cdot \frac{1}{n} \sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 - 2 \cdot \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2 \right) \Bigg] \\
&\quad + \left[2 \left(\frac{1}{n} \sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 - \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2 \right) \right] = \sum_{i=1}^{4} T_{i,n},
\end{aligned}$$

where again $T_{\beta_n}Y$ is the truncated version of Y, and m_{β_n} is the corresponding regression function. The definition of the estimate implies immediately its boundedness, that is $||m_n||_\infty \leq K \cdot \beta_n$. Analogously to the proof of Theorem 3.4, we obtain

$$T_{1,n} \leq c_4 \cdot \frac{\log(n)}{n}\,, \qquad \mathbf{E}(T_{3,n}) \leq c_5 \cdot \frac{\log(n)}{n}$$

and

$$\begin{aligned}\mathbf{E}(T_{4,n}) \;&\leq\; 2 \cdot \mathbf{E}\left(\inf_{f \in \bigoplus_{k=1}^K T_{\beta_n}\mathcal{F}_n}\left(\frac{1}{n}\sum_{i=1}^n |f(X_i)-Y_i|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i)-Y_i|^2\right)\right)\\ &\quad + c_6 \cdot \frac{\log(n)}{n}.\end{aligned}$$

The definition (4.6) of the function space $\bigoplus_{k=1}^K T_{\beta_n}\mathcal{F}_n$ in connection with the assumptions on the regression function adds up to

$$\begin{aligned}&\mathbf{E}\left(\inf_{f \in \bigoplus_{k=1}^K T_{\beta_n}\mathcal{F}_n}\left(\frac{1}{n}\sum_{i=1}^n |f(X_i)-Y_i|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i)-Y_i|^2\right)\right)\\ &= \mathbf{E}\Bigg(\inf_{g_1,\ldots,g_K \in T_{\beta_n}\mathcal{F}_n}\Bigg(\frac{1}{n}\sum_{i=1}^n \left|\sum_{k=1}^K g_k(X_i) - Y_i\right|^2\\ &\qquad\qquad -\frac{1}{n}\sum_{i=1}^n \left|\sum_{k=1}^K m_k(\alpha_k \cdot X_i) - Y_i\right|^2\Bigg)\Bigg).\end{aligned}$$

We have already seen in the previous section that every function of the form

$$f(x) = h(\alpha \cdot x) \qquad (x \in \mathbb{R}^d),$$

with $\alpha \in \mathbb{R}^d$ and $h \in \mathcal{F}_n^1$ (as defined in 4.2) is contained in $\mathcal{F}_n$. Furthermore, the choice of K_n and $L_{1,n}, ..., L_{k,n}$ permits the use of Lemma 2.2. Hence the above term can be bounded by

$$\begin{aligned}\mathbf{E}(T_{4,n}) \;&\leq\; 2 \cdot \mathbf{E}\left(\inf_{h_1,\ldots,h_K \in T_{\beta_n}\mathcal{F}_n^1}\left(\frac{1}{n}\sum_{i=1}^n \left|\sum_{k=1}^K h_k(\alpha_k \cdot X_i) - Y_i\right|^2\right.\right.\\ &\qquad\qquad \left.\left. -\frac{1}{n}\sum_{i=1}^n \left|\sum_{k=1}^K m_k(\alpha_k \cdot X_i) - Y_i\right|^2\right)\right)\\ &\leq\; 2 \cdot \mathbf{E}\left(\inf_{h_1,\ldots,h_K \in T_{\beta_n}\mathcal{G}}\left(\frac{1}{n}\sum_{i=1}^n \left|\sum_{k=1}^K h_k(\alpha_k \cdot X_i) - Y_i\right|^2\right.\right.\\ &\qquad\qquad \left.\left. -\frac{1}{n}\sum_{i=1}^n \left|\sum_{k=1}^K m_k(\alpha_k \cdot X_i) - Y_i\right|^2\right)\right)\\ &\leq\; 2 \cdot \inf_{h_1,\ldots,h_K \in T_{\beta_n}\mathcal{G}} \int \left|\sum_{k=1}^K h_k(\alpha_k \cdot x) - \sum_{k=1}^K m_k(\alpha_k \cdot x)\right|^2 \mu(dx),\end{aligned}$$

where $\mathcal{G}$ is the class of all continuous, piecewise linear functions with respect to a partition of $[\hat{a},\hat{b}]$ in K_n intervals, and $[\hat{a},\hat{b}]$ is chosen in such a way that $\alpha_k \cdot x \in [\hat{a},\hat{b}]$, for all $x \in [a,b]^d$ and $k = 1,...,K$. Obviously this implies

$$\begin{aligned}
\mathbf{E}(T_{4,n}) &\leq 2 \cdot \inf_{h_1,...,h_K \in T_{\beta_n}\mathcal{G}} \int \left| \sum_{k=1}^{K} (h_k(\alpha_k \cdot x) - m_k(\alpha_k \cdot x)) \right|^2 \mu(dx) \\
&\leq 2 \cdot \inf_{h_1,...,h_K \in T_{\beta_n}\mathcal{G}} K \cdot \int \sum_{k=1}^{K} |h_k(\alpha_k \cdot x) - m_k(\alpha_k \cdot x)|^2 \, \mu(dx) \\
&\leq 2 \cdot \inf_{h_1,...,h_K \in T_{\beta_n}\mathcal{G}} K \int K \max_{k=1,...,K} \left(|h_k(\alpha_k \cdot x) - m_k(\alpha_k \cdot x)|^2 \right) \mu(dx) \\
&\leq 2 \cdot \inf_{h_1,...,h_K \in T_{\beta_n}\mathcal{G}} K^2 \cdot \max_{x \in [a,b]^d} \max_{k=1,...,K} |h_k(\alpha_k \cdot x) - m_k(\alpha_k \cdot x)|^2 \\
&\leq 2 \cdot \inf_{h_1,...,h_K \in T_{\beta_n}\mathcal{G}} K^2 \cdot \max_{x \in [\hat{a},\hat{b}]} \max_{k=1,...,K} |h_k(x) - m_k(x)|^2 .
\end{aligned}$$

Now note that $|\min\{a,k\} - b| \leq |a - b|$ for $a \in \mathbb{R}^+$ and $b \in [0,k]$ and that the functions m_k are bounded due to their (p,C)–smoothness and their compact support. Therefore choosing $L > 0$ such that $||m_k||_\infty < L$, for all $k = 1,...,K$, leads to

$$\begin{aligned}
&\inf_{h_1 \in T_{\beta_n}\mathcal{G}, ..., h_K \in T_{\beta_n}\mathcal{G}} K^2 \cdot \max_{x \in [\hat{a},\hat{b}]} \max_{k=1,...,K} |h_k(x) - m_k(x)|^2 \\
&\qquad \leq \inf_{h_1,...,h_K \in \mathcal{G}} K^2 \cdot \max_{x \in [\hat{a},\hat{b}]} \max_{k=1,...,K} |h_k(x) - m_k(x)|^2 ,
\end{aligned}$$

for $\beta_n > L$, owing to the definition of T_{β_n}. Together with the approximation result from Lemma 1.16 this implies

$$\mathbf{E}(T_{4,n}) \leq 2 \cdot K^2 \cdot c_5 \cdot C^2 \cdot K_n^{-2p},$$

for sufficiently large $c_5 > 0$.

To finish this proof we consider $T_{2,n}$. Similarly to the proof of Theorem 3.4 we get

$$\mathbf{P}\{T_{2,n} > t\} \leq 14 \cdot \sup_{x_1^n} \mathcal{N}_1 \left(\frac{t}{80K\beta_n}, \bigoplus_{k=1}^{K} T_{\beta_n}\mathcal{F}_n, x_1^n \right) \cdot \exp\left(-\frac{n \cdot t}{5136 K^2 \beta_n^2} \right). \quad (4.7)$$

Here we have taken into account that functions in $\bigoplus_{k=1}^{K} T_{\beta_n}\mathcal{F}_n$ are bounded by $K \cdot \beta_n$ instead of β_n as in Theorem 3.4. Furthermore, from Lemma 1.13 we infer

$$\mathcal{N}_1 \left(\varepsilon, \bigoplus_{k=1}^{K} T_{\beta_n}\mathcal{F}_n, z_1^n \right) \leq \mathcal{N}_1 \left(\frac{\varepsilon}{K}, T_{\beta_n}\mathcal{F}_n, z_1^n \right)^K ,$$

which together with Lemma 2.5 leads to

$$\mathcal{N}_1 \left(\varepsilon, \bigoplus_{k=1}^{K} T_{\beta_n}\mathcal{F}_n, z_1^n \right) \leq 3^K \left(\frac{6e \cdot \beta_n \cdot K}{\varepsilon} \cdot K_n \cdot L_n \right)^{K \cdot 2(d+2) \sum_{k=1}^{K_n} L_{k.n}} ,$$

for $0 < \varepsilon < \beta_n/2$. In order to find an upper bound for (4.7), we consider $1/n < t$, and obtain accordingly

$$\begin{aligned}
&\mathcal{N}_1\left(\frac{t}{80K\cdot\beta_n}, \bigoplus_{k=1}^{K} T_{\beta_n}\mathcal{F}_n, x_1^n\right) \\
&\qquad \leq 3^K\left(\frac{6e\cdot\beta_n\cdot K\cdot 80K\cdot\beta_n}{t}\cdot K_n\cdot L_n\right)^{K\cdot 2(d+2)\sum_{k=1}^{K_n} L_{k.n}} \\
&\qquad \leq 3^K\left(480\beta_n^2\cdot K^2\cdot K_n\cdot L_n\cdot n\right)^{2K\cdot(d+2)\sum_{k=1}^{K_n} L_{k,n}} \\
&\qquad \leq n^{c_6\cdot K\cdot\sum_{k=1}^{K_n} L_{k,n}},
\end{aligned}$$

for sufficiently a large constant $c_6 > 0$. Hence, in view of (4.7), for $\varepsilon > 1/n$, this yields

$$\begin{aligned}
\mathbf{E}(T_{2,n}) &\leq \varepsilon + \int_{\varepsilon}^{\infty} \mathbf{P}\{T_{2,n} > t\}dt \\
&= \varepsilon + \int_{\varepsilon}^{\infty} 14\cdot n^{c_6\cdot K\cdot\sum_{k=1}^{K_n} L_{k,n}}\cdot\exp\left(-\frac{n}{5136K^2\cdot\beta_n^2}\cdot t\right)dt \\
&= \varepsilon + 14\cdot n^{c_6\cdot K\cdot\sum_{k=1}^{K_n} L_{k,n}}\cdot\frac{5136K^2\cdot\beta_n^2}{n}\cdot\exp\left(-\frac{n}{5136K^2\cdot\beta_n^2}\cdot\varepsilon\right),
\end{aligned}$$

which is minimized for

$$\varepsilon = \frac{5136K^2\cdot\beta_n^2}{n}\cdot\log\left(14\cdot n^{c_6\cdot K\cdot\sum_{k=1}^{K_n} L_{k,n}}\right).$$

More precisely, this choice implies, together with the settings $L_{i,n} = 2d+1$ for $(i = 1, ..., K_n)$ and $\beta_n = c_1\cdot\log(n)$ from the theorem, that

$$\begin{aligned}
\mathbf{E}(T_{2,n}) &\leq \frac{5136K^2\cdot\beta_n^2}{n}\cdot\log\left(14\cdot n^{c_6\cdot K\cdot K_n(2d+1)}\right) \\
&\quad +14\cdot n^{c_6\cdot K\cdot K_n(2d+1)}\cdot\frac{5136K^2\cdot\beta_n^2}{n}\exp\left(-\log\left(14\cdot n^{c_6\cdot K\cdot K_n(2d+1)}\right)\right) \\
&\leq \frac{c_7\cdot K^2\cdot\log(n)^2}{n}\cdot c_8\cdot K\cdot K_n(2d+1)\cdot\log(n) + \frac{5136K^2\cdot\log(n)^2}{n} \\
&\leq \frac{c_9\cdot K^3\cdot K_n\cdot\log(n)^3}{n},
\end{aligned}$$

where $c_9 > 0$ is chosen sufficiently large. Thus, we have

$$\mathbf{E}\int|m_n(x) - m(x)|^2\mu(dx) \leq \frac{c_9\cdot K^3\cdot K_n\cdot\log(n)^3}{n} + 2\cdot K^2\cdot c_5\cdot C^2\cdot K_n^{-2p},$$

which together with the choice of K_n reduces to

$$\begin{aligned}
\mathbf{E}\int|m_n(x) - m(x)|^2\mu(dx) &\leq c_9\cdot K^3\cdot C^{2/(2p+1)}\left(\frac{\log(n)^3}{n}\right)^{2p/(2p+1)} \\
&\quad +c_{10}\cdot K\cdot C^2\cdot C^{-4p/(2p+1)}\cdot\left(\frac{\log(n)^3}{n}\right)^{2p/(2p+1)}
\end{aligned}$$

$$\leq \quad c_3 \cdot K^3 \cdot C^{\frac{2}{2p+1}} \cdot \left(\frac{\log(n)^3}{n} \right)^{\frac{2p}{2p+1}}$$

for c_3 large enough. □

Summarising the above, it can be stated that the estimate defined in Chapter 2 can achieve the one–dimensional rate of convergence under the assumptions of a single index model. Actually we can moreover deduce the same rate of convergence for the corresponding estimate with data–dependent parameter choice in a similar manner as in Section 3.3.

Furthermore, we have defined an estimate which, in context of projection pursuit, also achieves the one–dimensional rate of convergence. This estimate is obviously closely related to the estimate from Chapter 2, even though it is not obvious, how to compute this estimate in applications, since it is defined with respect to a class of functions consisting of sums of maxima of minima of linear functions.

In the appendix we shall describe briefly how the estimate defined by (2.2) and (2.3) can be calculated for given sets of data. Note that an algorithm that can solve the minimisation problem in (2.2) (at least approximately) is not automatically able to solve the optimisation problem in (4.5). Just to the contrary, solving the minimisation problem in (4.5), and therefore the computation of the estimate considered in Theorem 4.2 is not possible in practice. Instead one can try to construct a similar estimate by using a stepwise approach. This will be done in the next chapter, where we provide an estimate which, on the one hand is a sum of functions from $\mathcal{F}_n$ and, on the other hand can be calculated by the use of a so–called greedy algorithm combined with the algorithm we will use for solving problem (2.2).

CHAPTER 5

L_2 Boosting

In this chapter we present an L_2 boosting estimate for a regression function. The used method fits repeatedly a function from a fixed function space to the residuals of the data and the estimate is a weighted sum of the fitted functions. In this context, the number of iteration steps which relates to the parameter is chosen by splitting the sample. In Section 5.1 we obtain a general bound on the L_2 error of the so–defined estimates with respect to arbitrary classes of functions $f : \mathbb{R}^d \to \mathbb{R}$. The provided bound depends on a *uniform* bound on the covering number of the underlying class of function, and it holds again without assuming the boundedness of Y.

Section 5.2 provides a rate of convergence for an L_2 boosting estimate, which has $T_{\beta_n}\mathcal{F}_{2,2}$ as underlying class of functions. The achieved rate of convergence does not depend on the dimension of X, and it holds for all regression functions, having certain smoothness properties referring to their Fourier transform.

5.1. A general L_2 Boosting Result

In pattern recognition one of the main achievements in the last decade was fitting linear combinations of (weak) classifiers to the data. Particularly the AdaBoost algorithm for classifiers, which was introduced by Freund and Schapire in 1996, attracted a great deal of attention, both in machine learning and in statistics. Its success can be traced back to the good performance of the algorithm in many different settings. The awareness that AdaBoost can be considered as a gradient descent optimisation technique (cf. Breiman (1998)) brought up the idea of using boosting methods in other connections than classification as well.

In particular, the idea of L_2 boosting in regression estimation goes back to Friedman (2001). He developed boosting methods in the context of regression estimation via optimisation using the squared error loss function. In 2006, Bühlmann proved the consistency of L_2 boosting for high-dimensional linear models. In the same year Barron, Cohen, Dahmen and De Vore (2006) developed a universally consistent estimate by applying Greedy algorithms to certain function spaces $\mathcal{F}$. Furthermore

for bounded data, they derived the rate of convergence

$$\left(\frac{\log(n)}{n}\right)^{1/2}$$

for this estimate, by performing a data dependent choice of the number iteration steps by complexity regularization. This rate holds for all regression functions m which admit an expansion $m = \sum_{f\in\mathcal{F}} c_f f$ where $\mathcal{F}$ is the underlying class of functions, and the sequence (c_f) of coefficients is absolutely summable.

In this dissertation we consider an algorithm similar to the relaxed greedy algorithm in Barron et al. (2006). More precisely, for a given class of functions $\mathcal{F}$ and a set of data $\mathcal{D}_n = \mathcal{D}_{n_l} \cup \mathcal{D}_{n_t} = \{(X_1, Y_1), ..., (X_n, Y_n)\}$, we define a sequence of estimates by

$$m_{n_l,1} = \tilde{m}_{n_l,1} = 0, \quad \text{and} \quad m_{n_l,k+1} = T_{\beta_n} \circ \tilde{m}_{n_l,k+1}, \tag{5.1}$$

where, for $k > 1$,

$$\tilde{m}_{n_l,k+1} = \left(1 - \frac{2}{k+1}\right) \cdot \tilde{m}_{n_l,k} + f_{n_l,k}, \tag{5.2}$$

and $f_{n_l,k}$ is chosen in such a way that

$$f_{n_l,k}(\cdot) = \arg\min_{f\in\mathcal{F}} \frac{1}{n_l} \sum_{i=1}^{n_l} \left| Y_i - \left(1 - \frac{2}{k+1}\right) \cdot \tilde{m}_{n_1,k}(X_i) - f(X_i) \right|^2. \tag{5.3}$$

Here one has to pay attention to the part (5.3) of the definition. In order to obtain $m_{n_l,k+1}$, we minimize with respect to $\tilde{m}_{n_l,k}$ rather than $m_{n_l,k}$. Thus, during the computation only the sequence $(\tilde{m}_{n_l,k})$ is relevant, and the truncation is carried out afterwards to obtain a bounded estimate in every iteration step. Moreover, (5.3) exhibits that this method fits repeatedly a function from the class $\mathcal{F}$ to the residuals of the data. However, (5.2) shows that $m_{n_l,k}$ is a weighted sum of functions from $\mathcal{F}$, and that the used weights only depend on k.

The parameter k of the estimate is chosen by minimizing the empirical L_2 risk on the testing sample $\mathcal{D}_{n_t} = \{(X_{n_l+1}, Y_{n_l+1}), ..., (X_n, Y_n)\}$, that is,

$$m_n(\cdot) = m_{n_l,k^*}(\cdot), \tag{5.4}$$

where k^* satisfies

$$k^* = \arg\min_{k\in\{1,...,n\}} \frac{1}{n_t} \sum_{i=n_l+1}^{n} |m_{n_l,k}(X_i) - Y_i|^2, \tag{5.5}$$

or in other words, k is chosen by splitting the sample (cf. Section 3.3).

Now, in order to establish our first theorem for L_2 boosting estimates, we introduce a certain class of functions, which is implicitly generated by the definition of the estimate.

For a given class $\mathcal{F}$ of functions $f : \mathbb{R}^d \to \mathbb{R}$, and for fixed $N \in \mathbb{N}$, we denote by $\mathcal{H}_N = \mathcal{H}_N^{\mathcal{F}}$ the class of functions of the form

$$h : \mathbb{R}^d \to \mathbb{R};\ h(x) = \alpha_1^h g_1(x) + ... + \alpha_N^h g_N(x),$$

with $\alpha_i^h \geq 0$ and $g_i \in \mathcal{F}$, $i \in \{1, ..., N\}$, satisfying

$$\left(\frac{2}{l}\sum_{i=1}^{N} \alpha_i^h\right) \cdot g_j \in \mathcal{F}, \tag{5.6}$$

for all $j \in \{1, \ldots, N\}$ and $l \in \{1, \ldots, k\}$, and

$$||g_j||_\infty \leq 1, \tag{5.7}$$

for all $j \in \{1, \ldots, N\}$.

The following result is an extension of Theorem 3.1 in Barron et al. (2006) to unbounded Y. Moreover, it uses splitting the sample instead of complexity regularisation in order to determine the number of iterartion steps. In order to state this extension we need a further notation, since we use an upper bound on the covering numbers of the underlying class of function $\mathcal{F}$, that is

$$\mathcal{N}_1\left(\varepsilon, \mathcal{F}, z_1^n\right) \leq \mathcal{N}_1\left(\varepsilon, \mathcal{F}\right), \tag{5.8}$$

for all $\varepsilon > 0$ and $z_1^n \in \mathbb{R}^d \times \cdots \times \mathbb{R}^d$. To be more precise, the upper bound $\mathcal{N}_1\left(\varepsilon, \mathcal{F}\right)$ is independent of the certain choice of the points z_1^n and we will refer to it as *uniform* bound on the covering number of $\mathcal{F}$.

THEOREM 5.1. *Let $\mathcal{F}$ be an arbitrary class of functions $f : \mathbb{R}^d \to \mathbb{R}$. Suppose that the distribution of (X, Y) satisfies*

$$\mathbf{E}\left(\exp\left(c_2 \cdot |Y|^2\right)\right) < \infty, \tag{5.9}$$

for some constant $c_2 > 0$, and that the regression function m is bounded in absolute value by some constant. Then, the estimate m_n defined by (5.1) - (5.5) with $\beta_n = c_1 \cdot \log(n)$, satisfies

$$\begin{aligned}
&\mathbf{E}\int |m_n(x) - m(x)|^2 \mu(dx) \\
&\quad \leq \min_{k \in \{1,\ldots,n\}} \left[c_3 \left(\frac{k \cdot \log(n)^2}{n_l} \cdot \log\left(\mathcal{N}_1\left(\frac{1}{80\beta_n \cdot k \cdot n_l}, \mathcal{F} \right) \right) \right) \right. \\
&\qquad \left. + \inf_{N \in \mathbb{N}} \inf_{h \in \mathcal{H}_N} \left(16 \frac{(\alpha_1^h + \cdots + \alpha_N^h)^2}{k} + 4 \int |h(x) - m(x)|^2 \mu(dx) \right) \right] \\
&\qquad + c_4 \frac{\log(n)^3}{n_t},
\end{aligned}$$

for sufficiently large constants c_3, c_4, which do not depend on n, β_n or k.

Please note that Chapter 2 provides precisely such a uniform bound on the covering numbers of the class $\mathcal{F}_n$. The obtained bound in Lemma 2.5 is obviously independent of the points x_1^n, and hence we may apply this theorem to maxmin functions in the next section.

For the proof of the above theorem we need a deterministic lemma, which is closely related to Theorem 2.4 in Barron et al. (2006). For this purpose, let $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ and define $m_{n,k}$ recursively as

$$m_{n,1} = 0, \quad \text{and} \quad m_{n,k+1} = \left(1 - \frac{2}{k+1}\right) \cdot m_{n,k} + f_{n,k}, \tag{5.10}$$

where

$$f_{n,k}(\cdot) = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \left| y_i - \left(1 - \frac{2}{k+1}\right) \cdot m_{n_1,k}(x_i) - f(x_i) \right|^2. \tag{5.11}$$

Thus, obviously the definition of the sequence $m_{n,k}$ resembles the definition of the sequence of estimates defined by (5.1) - (5.5), but without truncation or splitting the sample.

LEMMA 5.2. *Let $m_{n,k}$ be defined by (5.10) and (5.11). Then, for any $N \in \mathbb{N}$, $g_1, \ldots, g_N \in \mathcal{F}$ and $\alpha_1, \ldots, \alpha_N > 0$, satisfying*

$$\left(\frac{2}{l} \sum_{i=1}^{N} \alpha_i\right) \cdot g_j \in \mathcal{F}, \qquad \textit{for all} \quad j \in \{1, \ldots, N\}, l \in \{1, \ldots, k\}, \tag{5.12}$$

$$\textit{and} \quad ||g_j||_\infty \leq 1, \qquad \textit{for all} \quad j \in \{1, \ldots, N\}, \tag{5.13}$$

we have

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} |y_i - m_{n,k}(x_i)|^2 \quad \leq \quad & \frac{1}{n} \sum_{i=1}^{n} |y_i - (\alpha_1 g_1 + \cdots + \alpha_N g_N)(x_i)|^2 \\ & + 4 \cdot \frac{\left(\sum_{i=1}^{N} \alpha_i\right)^2}{k}. \end{aligned}$$

The proof is a straightforward modification of the proof of the corresponding theorem from Barron et al. (2006), but for the sake of completeness it is given anyway.

PROOF. Let $j \in \{1, \ldots, N\}$, and write

$$\beta_k = \frac{2}{k} \cdot \sum_{i=1}^{N} \alpha_i.$$

Since $\beta_k \cdot g_j \in \mathcal{F}$, we infer from the definition of the estimate that

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} |y_i - m_{n,k}(x_i)|^2 \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \left| y_i - \left(1-\frac{2}{k}\right)\cdot m_{n,k-1}(x_i) - \beta_k \cdot g_j(x_i) \right|^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \left| \left(1-\frac{2}{k}\right)\cdot (y_i - m_{n,k-1}(x_i)) + \frac{2}{k}\cdot y_i - \beta_k \cdot g_j(x_i) \right|^2 \\
&= \left(1-\frac{2}{k}\right)^2 \cdot \frac{1}{n}\sum_{i=1}^{n} |y_i - m_{n,k-1}(x_i)|^2 \\
&\quad +2\left(1-\frac{2}{k}\right)\cdot \frac{1}{n}\sum_{i=1}^{n} (y_i - m_{n,k-1}(x_i)) \cdot \left(\frac{2}{k}\cdot y_i - \beta_k \cdot g_j(x_i)\right) \\
&\quad +\frac{1}{n}\sum_{i=1}^{n} \left(\frac{2}{k}\left(y_i - \sum_{l=1}^{N}\alpha_l g_l(x_i)\right) + \frac{2}{k}\cdot\sum_{l=1}^{N}\alpha_l g_l(x_i) - \beta_k \cdot g_j(x_i)\right)^2 \\
&\leq \left(1-\frac{2}{k}\right)^2 \cdot \frac{1}{n}\sum_{i=1}^{n} |y_i - m_{n,k-1}(x_i)|^2 \\
&\quad +2\left(1-\frac{2}{k}\right)\cdot \frac{1}{n}\sum_{i=1}^{n} (y_i - m_{n,k-1}(x_i)) \cdot \left(\frac{2}{k}\cdot y_i - \beta_k \cdot g_j(x_i)\right) \\
&\quad +\left(\frac{2}{k}\right)^2 \frac{1}{n}\sum_{i=1}^{n} \left(y_i - \sum_{l=1}^{N}\alpha_l g_l(x_i)\right)^2 \\
&\quad +\frac{4}{k}\frac{1}{n}\sum_{i=1}^{n} \left(y_i - \sum_{l=1}^{N}\alpha_l g_l(x_i)\right)\cdot\left(\frac{2}{k}\cdot\sum_{l=1}^{N}\alpha_l g_l(x_i) - \beta_k \cdot g_j(x_i)\right) \\
&\quad +\left(\frac{2}{k}\right)^2 \frac{1}{n}\sum_{i=1}^{n} \left(\sum_{l=1}^{N}\alpha_l g_l(x_i)\right)^2 \\
&\quad -2\beta_k \cdot \frac{2}{k}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{l=1}^{N}\alpha_l g_l(x_i)\right)\cdot g_j(x_i) + \beta_k^2 \\
&=: L_j
\end{aligned}
$$

Furthermore, from $\alpha_j \geq 0$ and $\sum_{j=1}^{N}(2/k)\cdot\alpha_j = \beta_k$ we can conclude that

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} |y_i - m_{n,k}(x_i)|^2 \;&\leq\; \sum_{j=1}^{N} \frac{2\cdot\alpha_j}{k\cdot\beta_k}\cdot L_j \\
&= \left(1-\frac{2}{k}\right)^2 \cdot \frac{1}{n}\sum_{i=1}^{n} |y_i - m_{n,k-1}(x_i)|^2 \\
&\quad +2\;\left(1-\frac{2}{k}\right)\cdot \frac{1}{n}\sum_{i=1}^{n} (y_i - m_{n,k-1}(x_i)) \cdot \left(\frac{2}{k}\cdot y_i - \frac{2}{k}\sum_{j=1}^{N}\alpha_j g_j(x_i)\right)
\end{aligned}
$$

$$
\begin{aligned}
&+\left(\frac{2}{k}\right)^2 \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{l=1}^{N}\alpha_l g_l(x_i)\right)^2 - \left(\frac{2}{k}\right)^2 \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{l=1}^{N}\alpha_l g_l(x_i)\right)^2 + \beta_k^2 \\
&\leq \left(1-\frac{2}{k}\right)^2 \frac{1}{n}\sum_{i=1}^{n}(y_i - m_{n,k-1}(x_i))^2 + \left(\frac{2}{k}\right)^2 \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{l=1}^{N}\alpha_l g_l(x_i)\right)^2 \\
&\quad + \left(1-\frac{2}{k}\right)\cdot\frac{2}{k}\cdot\frac{1}{n}\sum_{i=1}^{n} 2\cdot(y_i - m_{n,k-1}(x_i))\left(y_i - \sum_{l=1}^{N}\alpha_l g_l(x_i)\right) \\
&\quad - \left(\frac{2}{k}\right)^2 \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{l=1}^{N}\alpha_l g_l(x_i)\right)^2 + \beta_k^2.
\end{aligned}
$$

Using $2\cdot a\cdot b \leq a^2 + b^2$, we obtain

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}|y_i - m_{n,k}(x_i)|^2 \\
&\leq \left(1-\frac{2}{k}\right)\cdot\frac{1}{n}\sum_{i=1}^{n}(y_i - m_{n,k-1}(x_i))^2 + \frac{2}{k}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{l=1}^{N}\alpha_l g_l(x_i)\right)^2 \\
&\quad + \beta_k^2 - \left(\frac{2}{k}\right)^2 \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{l=1}^{N}\alpha_l g_l(x_i)\right)^2,
\end{aligned}
$$

which is equivalent to

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}|y_i - m_{n,k}(x_i)|^2 - \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{l=1}^{N}\alpha_l g_l(x_i)\right)^2 \\
&\leq \left(1-\frac{2}{k}\right)\cdot\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - m_{n,k-1}(x_i))^2 - \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{l=1}^{N}\alpha_l g_l(x_i)\right)^2\right) \\
&\quad + \frac{4}{k^2}\cdot\left(\left(\sum_{j=1}^{N}\alpha_j\right)^2 - \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{l=1}^{N}\alpha_l g_l(x_i)\right)^2\right).
\end{aligned}
$$

We use this representation, in order to show that, for $k \geq 2$,

$$
\begin{aligned}
a_k &= \frac{1}{n}\sum_{i=1}^{n}|y_i - m_{n,k}(x_i)|^2 - \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{l=1}^{N}\alpha_l g_l(x_i)\right)^2 \\
&\leq 4\frac{\left(\sum_{j=1}^{N}\alpha_j\right)^2}{k} = 4\frac{M}{k}
\end{aligned}
$$

holds, with $M := \left(\sum_{j=1}^{N}\alpha_j\right)^2$. For $k = 2$, the above inequality amounts to

$$
\frac{1}{n}\sum_{i=1}^{n}|y_i - m_{n,2}(x_i)|^2 - \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{l=1}^{N}\alpha_l g_l(x_i)\right)^2
$$

$$\leq \left(\sum_{j=1}^{N} \alpha_j\right)^2 - \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{l=1}^{N} \alpha_l g_l(x_i)\right)^2,$$

which obviously implies

$$a_2 \leq \left(\sum_{j=1}^{N} \alpha_j\right)^2 \;<\; 2\left(\sum_{j=1}^{N} \alpha_j\right)^2 \;=\; 4\,\frac{M}{2}.$$

Furthermore, from

$$a_2 \;\leq\; 2M, \quad \text{and} \quad a_k \;\leq\; \left(1-\frac{2}{k}\right) a_{k-1} + \frac{4}{k^2}\,M,$$

for $k > 2$, we can infer inductively that

$$a_n \;\leq\; \frac{4M}{n} \tag{5.14}$$

holds for all $n \in \mathbb{N}$. More precisely, this can be deduced from

$$\begin{aligned} \frac{(n-1)^2 - n(n-2)}{n^2} \;=\; \frac{1}{n^2} \;>\; 0 \quad &\Longleftrightarrow \quad \frac{n-2}{n} \;\leq\; \frac{(n-1)^2}{n^2} \\ &\Longleftrightarrow \quad \left(1-\frac{2}{n}\right)\frac{4M}{n-1} \;\leq\; \frac{n-1}{n^2}\cdot 4M. \end{aligned}$$

Since this transformation, together with the assumption that a_{n-1} satisfies (5.14), leads obviously to

$$\left(1-\frac{2}{n}\right) a_{n-1} \leq \frac{n-1}{n^2}\cdot 4M$$

which, on the other hand, is equivalent to

$$\left(1-\frac{2}{n}\right) a_{n-1} + \frac{4M}{n^2} \leq \frac{4M}{n},$$

we obtain that (5.14) holds, for all $n \geq 2$. □

Proof of Theorem 5.1. As already mentioned in the definition, the estimate is defined by splitting the sample. Since by hypothesis, the modified Sub-Gaussian condition is fulfilled and since the regression function is bounded, that is, $||m||_\infty \leq L$, for some constant $L > 0$, we are in the position to apply Theorem 3.6. We choose the set of parameters as $\mathcal{Q}_n = \{1, ..., n\}$. Since our sequence of estimates is obviously bounded by β_n we obtain, for any $\delta > 0$, that

$$\begin{aligned} &\mathbf{E}\int |m_n(x) - m(x)|^2 \mu(dx) \\ &\qquad \leq (1+\delta) \min_{k\in\{1,\ldots,n\}} \mathbf{E}\int |m_{n_l,k}(x) - m(x)|^2 \mu(dx) \\ &\qquad\quad + c_5 \cdot \frac{1 + \log|\mathcal{Q}_n|}{n_t} + c_6 \frac{\log(n)}{n} \end{aligned}$$

holds, for $c_5 \geq \beta_n^2(32/\delta + 70 + 39\delta)$, and for a suitable constant $c_6 > 0$. Thus, for $\delta = 1$, the expectation of the L_2 error is bounded by

$$2 \min_{k \in \{1,\ldots,n\}} \mathbf{E} \int |m_{n_l,k}(x) - m(x)|^2 \mu(dx) + 141\beta_n^2 \frac{1 + \log(n)}{n_t} + c_6 \frac{\log(n)}{n},$$

since apparently $|\mathcal{Q}_n|$ is bounded by n. Now, for $k \in \{1, ..., n\}$, we use the error decomposition

$$\begin{aligned}
&\int |m_{n_l,k}(x) - m(x)|^2 \mu(dx) \\
&\quad = \Big[\mathbf{E}\Big(|m_{n_l,k}(X) - Y|^2 | \mathcal{D}_{n_l}\Big) - \mathbf{E}\Big(|m(X) - Y|^2\Big) \\
&\qquad -\mathbf{E}\Big(|m_{n_l,k}(X) - T_{\beta_n} Y|^2 | \mathcal{D}_{n_l}\Big) - \mathbf{E}\Big(|m_{\beta_n}(X) - T_{\beta_n} Y|^2\Big)\Big] \\
&\quad + \Bigg[\mathbf{E}\Big(|m_{n_l,k}(X) - T_{\beta_n} Y|^2 | \mathcal{D}_{n_l}\Big) - \mathbf{E}\Big(|m_{\beta_n}(X) - T_{\beta_n} Y|^2\Big) \\
&\qquad -2 \cdot \frac{1}{n_l} \sum_{i=1}^{n_l} \Big(|m_{n_l,k}(X_i) - T_{\beta_n} Y_i|^2 - |m_{\beta_n}(X_i) - T_{\beta_n} Y_i|^2\Big)\Bigg] \\
&\quad + \Bigg[2 \cdot \frac{1}{n_l} \sum_{i=1}^{n_l} |m_{n_l,k}(X_i) - T_{\beta_n} Y_i|^2 - 2 \cdot \frac{1}{n_l} \sum_{i=1}^{n_l} |m_{\beta_n}(X_i) - T_{\beta_n} Y_i|^2 \\
&\qquad - \left(2 \cdot \frac{1}{n_l} \sum_{i=1}^{n_l} |m_{n_l,k}(X_i) - Y_i|^2 - 2 \cdot \frac{1}{n_l} \sum_{i=1}^{n_l} |m(X_i) - Y_i|^2\right)\Bigg] \\
&\quad + \left[2 \left(\frac{1}{n_l} \sum_{i=1}^{n_l} |m_{n_l,k}(X_i) - Y_i|^2 - \frac{1}{n_l} \sum_{i=1}^{n_l} |m(X_i) - Y_i|^2\right)\right] \\
&\quad = \sum_{i=1}^{4} T_{i,n},
\end{aligned}$$

where again $T_{\beta_n} Y$ is the truncated version of Y, and m_{β_n} is the regression function of $T_{\beta_n} Y$.

Both terms, $T_{1,n}$ and $T_{3,n}$, can be bounded in the same way as their corresponding terms in the proof of Thereom 3.4. Hence we have

$$T_{1,n} \leq c_7 \cdot \frac{\log n}{n} \quad \text{and} \quad \mathbf{E}(T_{3,n}) \leq c_7 \cdot \frac{\log n}{n},$$

for a sufficiently large constant $c_7 > 0$.

Next we consider $T_{4,n}$. Let A_{n_l} be the event, that there exists $i \in \{1, ..., n_l\}$ such that $|Y_i| > \beta_n$. Then, we have

$$\begin{aligned}
\mathbf{E}(T_{4,n}) &\leq 2\,\mathbf{E}\left(\frac{1}{n_l}\sum_{i=1}^{n_l}|m_{n_l,k}(X_i) - Y_i|^2 \cdot \mathbf{1}_{A_{n_l}}\right) \\
&\quad + 2\,\mathbf{E}\left(\frac{1}{n_l}\sum_{i=1}^{n_l}|m_{n_l,k}(X_i) - Y_i|^2 \cdot \mathbf{1}_{A_{n_l}^C} - \frac{1}{n_l}\sum_{i=1}^{n_l}|m(X_i) - Y_i|^2\right) \\
&= 2\,\mathbf{E}\left(|m_{n_l,k}(X_1) - Y_1|^2 \cdot \mathbf{1}_{A_{n_l}}\right) \\
&\quad + 2\,\mathbf{E}\left(\frac{1}{n_l}\sum_{i=1}^{n_l}|m_{n_l,k}(X_i) - Y_i|^2 \cdot \mathbf{1}_{A_{n_l}^C} - \frac{1}{n_l}\sum_{i=1}^{n_l}|m(X_i) - Y_i|^2\right) \\
&= T_{7,n} + T_{8,n}.
\end{aligned}$$

With the Cauchy-Schwarz inequality, we see that $T_{7,n}$ satisfies the inequality

$$\begin{aligned}
\frac{1}{2}\,T_{7,n} &\leq \sqrt{\mathbf{E}\left((|m_{n_l,k}(X_1) - Y_1|^2)^2\right)} \cdot \sqrt{\mathbf{P}(A_{n_l})} \\
&\leq \sqrt{\mathbf{E}\left((2|m_{n_l,k}(X_1)|^2 + 2|Y_1|^2)^2\right)} \cdot \sqrt{n_l \cdot \mathbf{P}\{|Y_1| > \beta_n\}} \\
&\leq \sqrt{\mathbf{E}\left(8|m_{n_l,k}(X_1)|^4 + 8|Y_1|^4\right)} \cdot \sqrt{n_l \cdot \frac{\mathbf{E}\left(\exp(c_2 \cdot |Y_1|^2)\right)}{\exp(c_2 \cdot \beta_n^2)}},
\end{aligned}$$

where the last inequality results directly from

$$\mathbf{1}_{\{|Y|>\beta_n\}} \leq \frac{\exp(c_2 \cdot |Y|^2)}{\exp(c_2 \cdot \beta_n^2)}.$$

Since $x \leq \exp(x)$ holds for all $x \in \mathbb{R}$, we get

$$\begin{aligned}
\mathbf{E}\left(|Y|^4\right) &= \mathbf{E}\left(|Y|^2 \cdot |Y|^2\right) \leq \mathbf{E}\left(\frac{2}{c_2} \cdot \exp\left(\frac{c_2}{2} \cdot |Y|^2\right) \cdot \frac{2}{c_2} \cdot \exp\left(\frac{c_2}{2} \cdot |Y|^2\right)\right) \\
&= \frac{4}{c_2^2} \cdot \mathbf{E}\left(\exp\left(c_2 \cdot |Y|^2\right)\right),
\end{aligned}$$

which is finite by the assumption (5.9). Furthermore, $||m_{n_l,k}||_\infty$ is bounded by β_n, which implies that the first factor is bounded by

$$c_8 \cdot \beta_n^2 = c_9 \cdot \log(n)^2,$$

for some constant $c_9 > 0$. On the other hand, the second factor is bounded by $c_{10} \cdot \sqrt{n_l}/n^2$ for a suitable constant $c_{10} > 0$. Since condition (5.9) implies that

$$\begin{aligned}
\sqrt{n_l \cdot \frac{\mathbf{E}\left(\exp(c_2 \cdot |Y_1|^2)\right)}{\exp(c_2 \cdot \beta_n^2)}} &\leq \sqrt{n_l} \cdot \frac{\sqrt{c_{11}}}{\sqrt{\exp(c_2 \cdot \beta_n^2)}} \\
&\leq \sqrt{n_l}\sqrt{c_{11}} \cdot \exp\left(-\frac{c_{12} \cdot \log(n)^2}{2}\right)
\end{aligned}$$

holds, verifying the correctness of $\exp(-c_{12} \cdot \log(n)^2) = \mathcal{O}(n^{-2})$ establishes this bound. From the above, we deduce that

$$T_{7,n} \leq c_{13} \cdot \frac{\log(n)^2 \sqrt{n_l}}{n^2} \leq c_{14} \cdot \frac{\log(n)}{n}. \tag{5.15}$$

With the definition of A_{n_l}, and $\tilde{m}_{n_l,k}$ defined by (5.2), it follows that

$$\begin{aligned} T_{8,n} &\leq 2\,\mathbf{E}\left(\frac{1}{n_l}\sum_{i=1}^{n_l}|\tilde{m}_{n_l,k}(X_i)-Y_i|^2 \cdot I_{A_{n_l}^C} - \frac{1}{n_l}\sum_{i=1}^{n_l}|m(X_i)-Y_i|^2\right) \\ &\leq 2\,\mathbf{E}\left(\frac{1}{n_l}\sum_{i=1}^{n_l}|\tilde{m}_{n_l,k}(X_i)-Y_i|^2 - \frac{1}{n_l}\sum_{i=1}^{n_l}|m(X_i)-Y_i|^2\right). \end{aligned}$$

Obviously, the functions contained in $\mathcal{H}_N$ satisfy the assumptions of Lemma 5.2, and moreover the sequence of estimates $\tilde{m}_{n_l,k}$ is exactly of the form (5.10) and (5.11). Consequently, for arbitrary $N \in \mathbb{N}$ and $h \in \mathcal{H}_N$, Lemma 5.2 yields

$$\begin{aligned} T_{8,n} &\leq 2\,\mathbf{E}\left(4\frac{(\alpha_1^h+\cdots+\alpha_N^h)^2}{k} + \frac{1}{n_l}\sum_{i=1}^{n_l}|h(X_i)-Y_i|^2 - \frac{1}{n_l}\sum_{i=1}^{n_l}|m(X_i)-Y_i|^2\right) \\ &= 8\frac{(\alpha_1^h+\cdots+\alpha_N^h)^2}{k} + 2\,\mathbf{E}\left(\frac{1}{n_l}\sum_{i=1}^{n_l}|h(X_i)-Y_i|^2 - \frac{1}{n_l}\sum_{i=1}^{n_l}|m(X_i)-Y_i|^2\right) \\ &= 8\frac{(\alpha_1^h+\cdots+\alpha_N^h)^2}{k} + 2\,\left(\mathbf{E}(|h(X)-Y|^2) - \mathbf{E}(|m(X)-Y)\right) \\ &= 8\frac{(\alpha_1^h+\cdots+\alpha_N^h)^2}{k} + 2\int |h(x)-m(x)|^2\mu(dx), \end{aligned}$$

which, together with (5.15), leads to

$$\begin{aligned} \mathbf{E}(T_{4,n}) &\leq c_{14} \cdot \frac{\log(n)}{n} \\ &\quad + \inf_{N\in\mathbb{N}}\inf_{h\in\mathcal{H}_N}\left(8\frac{(\alpha_1^h+\cdots+\alpha_N^h)^2}{k} + 2\int |h(x)-m(x)|^2\mu(dx)\right). \end{aligned}$$

Now, the last part of the proof considers $T_{2,n}$. In order to obtain bounds on the expectation of $T_{2,n}$ we need conclusions for the covering numbers of $\mathcal{F}$. With the similar notation as in (4.6), that is, $\bigoplus_{k=1}^K \mathcal{F}$ is defined by

$$\left\{g : \mathbb{R}^d \to \mathbb{R},\ g(x) = \sum_{k=1}^K g_k(x),\ (x \in \mathbb{R}^d), \text{for some } g_k \in \mathcal{F}, 1 \leq k \leq K\right\},$$

obviously $m_{n_l,k} \in T_\beta(\bigoplus_{i=1}^k \mathcal{F})$ holds. Furthermore, it is easy to see that

$$\mathcal{N}(\varepsilon, T_\beta\mathcal{G}, z_1^n) \leq \mathcal{N}(\varepsilon, \mathcal{G}, z_1^n) \tag{5.16}$$

holds, for an arbitrary class of functions $\mathcal{G}$ of real functions on $\mathbb{R}^d$. Since, whenever $g_1, ..., g_N$ is an L_p-ε-cover of $\mathcal{G}$ on z_1^n, then $T_\beta g_1, ..., T_\beta g_N$ is an L_p-ε-cover of $T_\beta\mathcal{G}$

on z_1^n. Hence, in particular,

$$\mathcal{N}\left(\varepsilon, T_\beta \bigoplus_{i=1}^{k} \mathcal{F}, z_1^n\right) \leq \mathcal{N}\left(\varepsilon, \bigoplus_{i=1}^{k} \mathcal{F}, z_1^n\right)$$

holds for all $\varepsilon > 0$. From this and Lemma 1.13, we can deduce that

$$\mathcal{N}\left(\varepsilon, T_\beta \bigoplus_{i=1}^{k} \mathcal{F}, z_1^n\right) \quad \leq \quad \mathcal{N}\left(\frac{\varepsilon}{k}, \mathcal{F}, z_1^n\right)^k.$$

holds for all $\varepsilon > 0$, too. Thus we have, for arbitrary $t > 1/n$,

$$\begin{aligned}
&\mathbf{P}\{T_{2,n} > t\} \\
&\leq \mathbf{P}\left\{\exists f \in T_{\beta_n} \bigoplus_{i=1}^{k} \mathcal{F} : \mathbf{E}\left(\left|\frac{f(X)}{\beta_n} - \frac{T_{\beta_n} Y}{\beta_n}\right|^2\right) - \mathbf{E}\left(\left|\frac{m_{\beta_n}(X)}{\beta_n} - \frac{T_{\beta_n} Y}{\beta_n}\right|^2\right)\right. \\
&\qquad - \frac{1}{n_l} \sum_{i=1}^{n_l} \left(\left|\frac{f(X_i)}{\beta_n} - \frac{T_{\beta_n} Y_i}{\beta_n}\right|^2 - \left|\frac{m_{\beta_n}(X_i)}{\beta_n} - \frac{T_{\beta_n} Y_i}{\beta_n}\right|^2\right) \\
&\qquad \left. > \frac{1}{2}\left(\frac{t}{\beta_n^2} + \mathbf{E}\left(\left|\frac{f(X)}{\beta_n} - \frac{T_{\beta_n} Y}{\beta_n}\right|^2\right) - \mathbf{E}\left(\left|\frac{m_{\beta_n}(X)}{\beta_n} - \frac{T_{\beta_n} Y}{\beta_n}\right|^2\right)\right)\right\}.
\end{aligned}$$

Thus, we can infer from Theorem 1.17, and from the inequality

$$\mathcal{N}_1\left(\delta, \left\{\frac{1}{\beta_n} f : f \in \mathcal{F}\right\}, z_1^n\right) \leq \mathcal{N}_1\left(\delta \cdot \beta_n, \mathcal{F}, z_1^n\right)$$

that, for $z_1^n = (z_1, ..., z_n) \in \mathbb{R}^d \times ... \times \mathbb{R}^d$,

$$\begin{aligned}
\mathbf{P}\{T_{2,n} > t\} \quad &\leq \quad 14 \sup_{z_1^n} \mathcal{N}_1\left(\frac{t}{80\beta_n}, T_{\beta_n} \bigoplus_{i=1}^{k} \mathcal{F}, z_1^n\right) \cdot \exp\left(-\frac{n_l}{5136 \cdot \beta_n^2} t\right) \\
&\leq \quad 14 \sup_{z_1^n} \mathcal{N}_1\left(\frac{t}{80\beta_n \cdot k}, \mathcal{F}, z_1^n\right)^k \cdot \exp\left(-\frac{n_l}{5136 \cdot \beta_n^2} t\right).
\end{aligned}$$

Now, with the uniform bound on the covering number of $\mathcal{F}$, that is,

$$\mathcal{N}_1\left(\varepsilon, \mathcal{F}, z_1^n\right) \leq \mathcal{N}_1\left(\varepsilon, \mathcal{F}\right),$$

for all $z_1^n \in (\mathbb{R}^d)^n$, and $\varepsilon > 0$, we obtain that

$$\mathbf{P}\{T_{2,n} > t\} \quad \leq \quad 14 \cdot \mathcal{N}_1\left(\frac{1}{80\beta_n \cdot k \cdot n_l}, \mathcal{F}\right)^k \cdot \exp\left(-\frac{n_l}{5136 \cdot \beta_n^2} t\right)$$

holds for $1/n_l < t$. Thus, we have for arbitrary $\varepsilon \geq 1/n$,

$$\begin{aligned}
\mathbf{E}(T_{2,n}) \quad &\leq \quad \varepsilon + \int_\varepsilon^\infty \mathbf{P}\{T_{2,n} > t\} dt \\
&= \quad \varepsilon + 14 \cdot \mathcal{N}_1\left(\frac{1}{80\beta_n \cdot k \cdot n_l}, \mathcal{F}\right)^k \cdot \frac{5136\beta_n^2}{n_l} \cdot \exp\left(-\frac{n_l}{5136\beta_n^2} \varepsilon\right),
\end{aligned}$$

which is minimal for

$$\varepsilon \quad = \quad \frac{5136 \cdot \beta_n^2}{n_l} \log\left(14 \cdot \mathcal{N}_1\left(\frac{1}{80\beta_n \cdot k \cdot n_l}, \mathcal{F}\right)^k\right).$$

To be more precise, we obtain

$$\begin{aligned}
\mathbf{E}(T_{2,n}) &\leq \frac{5136 \cdot \beta_n^2}{n_l} \log \left(14 \cdot \mathcal{N}_1 \left(\frac{1}{80\beta_n \cdot k \cdot n_l}, \mathcal{F} \right)^k \right) \\
&\quad + 14 \cdot \mathcal{N}_1 \left(\frac{1}{80\beta_n \cdot k \cdot n_l}, \mathcal{F} \right)^k \\
&\quad \cdot \frac{5136\beta_n^2}{n_l} \left(14 \cdot \mathcal{N}_1 \left(\frac{1}{80\beta_n \cdot k \cdot n_l}, \mathcal{F} \right)^k \right)^{-1} \\
&= \frac{5136 \cdot \beta_n^2}{n_l} \left(\log(14) + k \cdot \log \left(\mathcal{N}_1 \left(\frac{1}{80\beta_n \cdot k \cdot n_l}, \mathcal{F} \right) \right) + 1 \right) \\
&\leq \frac{c_{16} \cdot \log(n)^2 \cdot k \cdot \log \left(\mathcal{N}_1 \left(\frac{1}{80\beta_n \cdot k \cdot n_l}, \mathcal{F} \right) \right)}{n_l},
\end{aligned}$$

for some sufficiently large constant $c_{16} > 0$, which does not depend on n, β_n or k. Thus, we can deduce that

$$\begin{aligned}
&\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) \\
&\leq 2 \min_{k \in \{1,\ldots,n\}} \left(c_3 \left(\frac{k \cdot \log(n)^2}{n_l} \cdot \log \left(\mathcal{N}_1 \left(\frac{1}{80\beta_n \cdot k \cdot n_l}, \mathcal{F} \right) \right) \right) \right. \\
&\qquad \left. + \inf_{N \in \mathbb{N}} \inf_{h \in \mathcal{H}_N} \left(8 \frac{(\alpha_1^h + \cdots + \alpha_N^h)^2}{k} + 2 \int |h(x) - m(x)|^2 \mu(dx) \right) \right) \\
&\quad + 141\beta^2 \, \frac{1 + \log(n)}{n_t} + c_4 \frac{\log n}{n}
\end{aligned}$$

holds, for sufficiently large constants $c_3, c_4 > 0$, and therefore we have proved the desired result. □

We want to remark that Theorem 1.11 immediately leads to the conclusion that, for a class $\mathcal{F}$ of bounded functions $f : \mathbb{R}^d \to \mathbb{R}$, the first term in the minimum can be replaced by

$$c_3 \cdot \frac{k \cdot \log(n)^3 \cdot V_{\mathcal{F}^+}}{n_l}.$$

Sometimes it is much easier to obtain bounds on the VC–dimension of a certain function space rather than obtain a uniform bound on the covering numbers. Therefore, in some cases this might be a helpful bound, too.

In the following, we shall apply the result from this section to a class of maxmin functions, in order to derive the rate of convergence of the corresponding L_2 boosting estimate.

5.2. L_2 Boosting with Maxmin Functions

In the previous section, we have seen a bound on the L_2 error of a boosting estimate that only depends on the covering number of the underlying class of functions. However, it is evident that, especially in the context of boosting estimates, not every class of functions is senseful and that the corresponding estimates might not even be computable.

In this section we consider $T_{\beta_n}\mathcal{F}_{2,2}$ as underlying class of functions and therefore, the estimate of interest is defined by (5.1) - (5.5), with $\mathcal{F}$ replaced by $T_{\beta_n}\mathcal{F}_{2,2}$. For the so–defined estimate we will derive a rate of convergence, which does not depend on the dimension d of the observation variable X, and hence circumvents the curse of dimensionality without the restrictions on the structure of the regression function m made in Chapter 4. Moreover, the computability of this L_2 boosting estimate secured, since we can use similar methods as used for the computation of the maxmin estimate from Chapter 2.

However, to derive a reasonable rate of convergence, of course we still have to make certain smoothness assumptions. In these settings this means that we consider functions $f \in L_1(\mathbb{R}^d)$, which satisfy

$$f(x) = f(0) + \int \left(e^{i(\omega \cdot x)} - 1\right) \hat{F}(\omega)\, d\omega, \tag{5.17}$$

almost surely, where $\hat{F}$ is the Fourier transform of f, that is,

$$\hat{F}(\omega) = \frac{1}{(2\pi)^{d/2}} \int e^{-i(\omega \cdot x)} f(x)\, dx \quad (\omega \in \mathbb{R}^d).$$

Furthermore we assume

$$\int ||\omega|| \cdot |\hat{F}(\omega)|\, d\omega \leq C, \quad 0 < C < \infty. \tag{5.18}$$

In the sequel, the class of functions $f : \mathbb{R}^d \to \mathbb{R}$ satisfying (5.17) and (5.18), will be denoted by $\mathcal{F}_C$.

As already mentioned, the corresponding smoothness assumptions in Barron et al. (2006), is that the regression function m has an expansion $m = \sum_{f \in \mathcal{F}} c_f f$,with an absolutely summable sequence (c_f). They discussed their smoothness conditions in the case that the resulting estimate is a neural network, and in this special situation their conditions turn out to be very similar to our smoothness assumptions on m.

Furthermore note that, with Condition (5.18), the assumption $m \in \mathcal{F}_C$ implies directly that m must have a Fourier transform with finite first moment (cf. Györfi et al. (2002), p. 317). Therefore under this assumption m has to be continuously differentiable and consequently bounded, if its support is bounded.

COROLLARY 5.3. *Suppose that the distribution of* (X, Y) *satisfies*

$$\mathbf{E}\left(\exp\left(c_2 \cdot |Y|^2\right)\right) < \infty,$$

for some constant $c_2 > 0$*, that* $X \in [-a, a]^d$ *a.s. for some* $a \in \mathbb{R}^+$*, and that the regression function is bounded in absolute value by some constant less than or equal to* β_n *and that it satisfies* $m \in \mathcal{F}_C$*, for some* $0 < C < \infty$*.*

Let $\beta_n = c_1 \cdot \log(n)$*, with* $c_1 > 0$ *is chosen in such a way that* $\beta_n \geq 6\sqrt{d} \cdot C \cdot a$ *for* $n \geq 2$*. Then, the estimate* m_n *defined by (5.1) - (5.5), with* $\mathcal{F} = T_{\beta_n}\mathcal{F}_{2,2}$*, and with* $n_l = \lceil \frac{n}{2}, \rceil$ *satisfies that*

$$\mathbf{E}\int |m_n(x) - m(x)|\mu(dx) = c_3 \cdot C^2 \left(\frac{\log(n)^3}{n}\right)^{1/2} \qquad (n \geq 2)$$

for some sufficiently large constant $c_3 > 0$*, that does not depend on* n, k *or* C*.*

In order to prove this corollary, we need the following Lemma 16.8 from Györfi et al. (2002) provides an approximation result for neural networks.

LEMMA 5.4. *Let* σ *be a squashing function. Then, for every probability measure* μ *on* $\mathbb{R}^d$*, every measurable* $f \in \mathcal{F}_C$ *and* $k \geq 1$*, there exists a neural network* f_k *in*

$$\left\{\sum_{i=1}^{k} c_i \sigma(a_i \cdot x + b_i) + c_0;\ k \in \mathbb{N}, a_i \in \mathbb{R}^d, b_i, c_i \in \mathbb{R}\right\} \tag{5.19}$$

such that

$$\int_{S_r} (f(x) - f_k(x))^2 \mu(dx) \leq \frac{(2rC)^2}{k}.$$

The coefficients of the linear combination in (5.19) may be chosen so that

$$\sum_{i=0}^{k} |c_i| \leq 3rC + f(0).$$

PROOF. For a proof we refer to the corresponding proof in Györfi et al. (2002). □

Here, $S_r = \{x \in \mathbb{R}^d : ||x|| \leq r\}$ denotes the closed Euclidean ball in $\mathbb{R}^d$, centered at 0 with radius r. Furthermore, a squashing function simply is a nondecreasing function $\sigma : \mathbb{R} \to [0, 1]$ which satisfies $\lim_{x\to-\infty} \sigma(x) = 0$ and $\lim_{x\to\infty} \sigma(x) = 1$.

In the proof of Corollary 5.3 we shall see a close connection between the class of functions defined in (5.19), and the class $\mathcal{H}_{k+1}^{T_{\beta_n}\mathcal{F}_{2,2}}$. This relationship enables us to use the above lemma, in order to get the desired rate of convergence.

PROOF OF COROLLARY 5.3. Obviously the assumptions of Theorem 5.1 are satisfied. Hence we can deduce from the setting $n_l = \lceil n/2 \rceil$ that

$$\begin{aligned}
&\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) \\
&\quad \leq \min_{k \in \{1,\ldots,n\}} \left(c_4 \left(\frac{k \cdot \log(n)^2}{n} \cdot \log \left(\mathcal{N}_1 \left(\frac{1}{80 \beta_n \cdot k \cdot n}, T_{\beta_n} \mathcal{F}_{2,2} \right) \right) \right) \right. \\
&\qquad \left. + \inf_{N \in \mathbb{N}} \inf_{h \in \mathcal{H}_N} \left(16 \frac{(\alpha_1^h + \cdots + \alpha_N^h)^2}{k} + 4 \int |h(x) - m(x)|^2 \mu(dx) \right) \right) \\
&\qquad + c_5 \frac{\log(n)^3}{n},
\end{aligned}$$

holds, for suitable constants c_4 and c_5. Since Lemma 2.5 implies that

$$\begin{aligned}
\mathcal{N}_1 \left(\frac{1}{80 \beta_n \cdot k \cdot n}, T_{\beta_n} \mathcal{F}_{2,2} \right) &\leq 3 \left(6 e \beta_n \cdot 80 \beta_n \cdot k \cdot n \cdot 2 \cdot 2 \right)^{2(d+2) \cdot 2 \cdot 2} \\
&= 3 (1920 \cdot e \cdot \beta_n^2 \cdot k \cdot n)^{8(d+2)},
\end{aligned}$$

we obtain

$$\begin{aligned}
\log \left(\mathcal{N}_1 \left(\frac{1}{80 \beta_n \cdot k \cdot n}, T_{\beta_n} \mathcal{F}_{2,2} \right) \right) &\leq c_6 \cdot \log(\log(n)^2 \cdot k \cdot n) \\
&\leq c_7 \cdot \log(n),
\end{aligned}$$

for sufficiently large constants c_6, c_7, and this in turn leads to

$$\begin{aligned}
&\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) \\
&\quad \leq \min_{k \in \{1,\ldots,n\}} \left(c_8 \left(\frac{k \cdot \log(n)^3}{n} \right) \right. \\
&\qquad \left. + \inf_{N \in \mathbb{N}} \inf_{h \in \mathcal{H}_N} \left(16 \frac{(\alpha_1^h + \cdots + \alpha_N^h)^2}{k} + 4 \int |h(x) - m(x)|^2 \mu(dx) \right) \right)
\end{aligned}$$

for a suitable constant c_8. Since the above inequality involves the minimum over $k \in \{1, ..., n\}$, we get an upper bound if we choose

$$k = \left(\frac{n}{\log(n)^3} \right)^{1/2}.$$

Then we obtain that

$$\begin{aligned}
\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) \quad \leq \quad & c_8 \left(\frac{\log(n)^3}{n} \right)^{1/2} \\
& + \inf_{N \in \mathbb{N}} \inf_{h \in \mathcal{H}_N} \left(16 (\alpha_1^h + \cdots + \alpha_N^h)^2 \left(\frac{\log(n)^3}{n} \right)^{1/2} \right. \\
& \left. + 4 \int |h(x) - m(x)|^2 \mu(dx) \right)
\end{aligned}$$

holds, for a sufficiently large constant $c_8 > 0$, that does not depend on n, β_n or k.

Hence, in order to complete the proof, it suffices to find a bound on the infimum over $h \in \mathcal{H}_N$ in the above inequality. As already mentioned, we will use Lemma 5.4 to derive such a bound. However, in order to apply this lemma we need a connection between the class of functions

$$\left\{ \sum_{i=1}^{k} c_i \sigma(a_i \cdot x + b_i) + c_0;\ k \in \mathbb{N}, a_i \in \mathbb{R}^d, b_i, c_i \in \mathbb{R} \right\},$$

for an arbitrary squashing function σ, and the class of functions $\mathcal{H}_N$, we are considering here.

First, it is quite easy to see that the so–called ramp squasher σ^*, defined by

$$\sigma^*(x) = \begin{cases} 0, & x < 0, \\ x, & 0 \le x \le 1, \\ 1, & x > 1, \end{cases} \tag{5.20}$$

is a squashing function. Secondly functions of the form

$$\sum_{i=1}^{k} c_i \sigma^*(a_i \cdot x + b_i)$$

are elements of $\mathcal{H}_k$. Indeed, for arbitrary $a_i \in \mathbb{R}^d$, and $b_i \in \mathbb{R}$ we have that

$$\begin{aligned} \sigma^*(a_i \cdot x + b_i) &= \begin{cases} 0, & a_i \cdot x < -b_i, \\ a_i \cdot x + b_i, & -b_i \le a_i \cdot x \le 1 - b_i, \\ 1, & a_i \cdot x > 1 - b_i, \end{cases} \\ &= \max\left\{ 0\ ,\ \min\left\{ a_i \cdot x + b_i\ ,\ 1 \right\} \right\} := f_i^+ \in \mathcal{F}_{2,2}, \end{aligned}$$

with $||f_i^+||_\infty \le 1$, and that

$$\begin{aligned} -\sigma^*(a_i \cdot x + b_i) &= \begin{cases} 0, & a_i \cdot x < -b_i, \\ -(a_i \cdot x + b_i), & -b_i \le a_i \cdot x \le 1 - b_i, \\ -1, & a_i \cdot x > 1 - b_i, \end{cases} \\ &= \max\left\{ -1\ ,\ \min\left\{ -(a_i \cdot x + b_i)\ ,\ 0 \right\} \right\} := f_i^- \in \mathcal{F}_{2,2}, \end{aligned}$$

with $||f_i^-||_\infty \le 1$, as well. Therefore condition (5.7) is obviously satisfied, and we can rewrite

$$\sum_{i=1}^{k} c_i \sigma^*(a_i \cdot x + b_i),$$

by using the algebraic sign of the c_i to choose whether f_i^+ or f_i^-, as

$$|c_1| \cdot f_1^{sign(c_1)} + |c_2| \cdot f_2^{sign(c_2)} + \cdots + |c_k| \cdot f_k^{sign(c_k)}.$$

Now it is easy to show that $\sum_{i=1}^{k} c_i \sigma^*(a_i \cdot x + b_i) \in \mathcal{H}_k$. Note that the correctness of condition (5.6) can be deduced from the fact that multiplication of a function from

$\mathcal{F}_{2,2}$ with a positive factor still yields a function from $\mathcal{F}_{2,2}$. If β_n is large enough this is still true for $T_{\beta_n}\mathcal{F}_{2,2}$ since the boundedness of the regression function and the boundedness of the weights in Lemma 5.4 imply that the truncation does not affect these functions at all.

We have moreover assumed that $X \in [-a,a]^d$ *a.s.* and therefore we obtain with $r = \sqrt{d} \cdot a$ that $X \in S_r = \{x \in \mathbb{R}^d, ||x|| \leq r\}$ *a.s.* Thus from Lemma 5.4, and from the assumptions $N = k+1$ and $\beta_n > 3rC + m(0)$, we infer

$$\begin{aligned}
&\inf_{h \in \mathcal{H}_N} \left(16(\alpha_1^h + \cdots + \alpha_N^h)^2 \cdot \left(\frac{\log(n)^3}{n} \right)^{1/2} + 4 \int |h(x) - m(x)|^2 \mu(dx) \right) \Bigg) \\
&\quad \leq \; 16 \cdot (3rC + m(0))^2 \cdot \left(\frac{\log(n)^3}{n} \right)^{1/2} + 4 \cdot (2rC)^2 \cdot \left(\frac{\log(n)^3}{n} \right)^{1/2} \\
&\quad \leq \; c_6 \cdot C^2 \cdot \left(\frac{\log(n)^3}{n} \right)^{1/2},
\end{aligned}$$

for a suitable chosen constant c_6, that does not depend an C, n or k. □

We want to remark that the rate of convergence in Corollary 5.3 holds generally for $\mathcal{F} = T_{\beta_n}\mathcal{F}_{m,n}$, with $m, n \geq 2$. This can be deduced from the inclusion

$$\mathcal{F}_{m_1,n_1} \subset \mathcal{F}_{m_2,n_2}, \quad \text{for } m_1 \leq m_2 \text{ and } n_1 \leq n_2,$$

since this containment ensures that Lemma 5.4 is still applicable. Furthermore it is easy to see that the necessary uniform bound on the covering number of $T_\beta \mathcal{F}_{m,n}$ is of the size $\mathcal{O}(n)$ too, and that we can therefore infer the rate of convergence for all estimates defined by (5.1) - (5.5) with $\mathcal{F}$ chosen as $T_{\beta_n}\mathcal{F}_{m,n}$ for some $m, n \geq 2$.

In this Chapter we extended the result of Barron et al. (2006) to unbounded Y. Furthermore, we considered an explicit L_2 boosting estimate and proved that its rate of convergence does not depend on the dimension d of the observation variable X.

Even though we are able to compute the estimate defined by (5.1) - (5.5), with $\mathcal{F} = T_{\beta_n}\mathcal{F}_{2,2}$ we are unable to present any applications yet. The implementation of this estimate is still in progress. However, the next chapter provides applications of the estimate presented in Chapter 2 to simulated data, and briefly describes the algorithm belonging to that estimate. The computation of the L_2 boosting estimate will be done in a similar way by using additionally its stepwise definition.

Appendix

A.1. The Algorithm

We have seen that both the maxmin estimate presented in Chapter 2 and the L_2 boosting estimate from the preceding chapter have promising theoretical properties. However, even a theoretical brilliant estimate makes no sense if its computation is too hard, or even worse if it cannot be computed at all. Hence as a matter of course, it is very important to provide algorithms for the computation of new estimates, and to observe the behaviour of estimates in practical applications, or at least in simulation studies.

This appendix deals with the computation of the maxmin estimate, and with its performance in a simulation study. Above all, we would like to emphasize that the development of the algorithm used for the computation of the maxmin estimate, was predominantly made by Adil Bagirov, and that it was not the aim of this dissertation. However, for the sake of completeness, we want to provide a brief insight into this optimisation part, and refer to Bagirov, Clausen and Kohler (2007) for a detailed discussion of the implementation.

Since we consider least squares estimates in this thesis, it is evident that the computation of the estimate defined by (2.2) and (2.3) in fact is an optimisation problem or, to be more precise, a minimisation problem, which can be formulated as follows

$$\text{minimise } F(a,b) = \frac{1}{n}\sum_{i=1}^{n}\left|\left(\max_{k=1,\dots,K}\min_{l=1,\dots,L_k}(a_{k,l}\cdot x_i + b_{k,l})\right) - y_i\right|^2 \tag{5.1}$$

for given (fixed) $x_1,\dots,x_n \in \mathbb{R}^d$, $y_1,\dots,y_n \in \mathbb{R}$, with respect to

$$a = (a_{1,1},\dots,a_{1,L_1},\dots,a_{K,1},\dots,a_{K,L_K}) \in \mathbb{R}^{d\times p},$$

and

$$b = (b_{1,1},\dots,b_{1,L_1},\dots,b_{K,1},\dots,b_{K,L_K}) \in \mathbb{R}^{p},$$

where $p = \sum_{k=1}^{K} L_k$. Unfortunately we cannot solve this problem exactly, since continuous piecewise linear functions typically are nonsmooth and nonconvex. Therefore also the function F in (5.1) usually is nonsmooth and nonconvex. In general, such functions have many local minima. Especially, the number of local minima of

F increases drastically as the number of maxima and minima functions increases. Unfortunately, most of this local minimisers do not provide a good approximation of the regression function, or even of the data points. Hence one is interested in global solutions of (5.1), or at least in finding a minimiser which is close to the global one. Classical methods of global optimisation are not effective for minimising such functions, since they are very time consuming and cannot solve this problem in a reasonable time. Since the function to be minimised is moreover a quite complicated nonsmooth function, the calculation even of only one subgradient of such a function is a difficult task.

The discrete gradient method from Bagirov (2002) allows an approximation of subgradients of the function F, and therefore an approximative computation of the estimate. This method requires a couple of properties of F. It can be seen in Bagirov, Clausen and Kohler (2007) that F is a semismooth quasidifferentiable function, whose subdifferential and superdifferential are polytopes. Therefore it is possible to approximate its subgradients. For the definition of semismoothness we refer to Mifflin (1977), and the definition of quasidifferentiable functions goes back to Demyanov and Rubinov (1995). Moreover, it can be shown that F is also piecewise partially separable (for the definition we refer to Bagirov and Ugon (2006)), and thus we can apply the improved discrete gradient method described in Bagirov and Ugon (2006), and Bagirov, Ghosh and Webb (2006).

Now, for each number of minima functions we start with a small number of maxima functions, and we increase their number stepwise until a further increase does not improve the approximation of the data anymore (with respect to some tolerance). Following these ideas, we get a set of piecewise functions. Using these functions on a testing set and choosing the best of them, we obtain a global solution, or at least a solution close to a global one. Furthermore, it should be mentioned that the data dependent choice of the parameters (via splitting the sample) is included in the implementation and that the implementation of the estimate was realized in both Fortran and R.

A.2. Application to Simulated Data

In order to compare the estimates proposed in this dissertation with other nonparametric regression estimates, we made a small simulation study. Here, we define the underlying random vector (X, Y) by

$$Y = m(X) + \sigma \cdot \varepsilon,$$

where ε is standard normally distributed and independent of X and $\sigma \geq 0$, and where X is uniformly distributed on $[-2,2]^d$. For the noise level σ we use three different values: 0, 0.5 and 1, and we generate data sets of two different sample sizes, namely $n = 500$ and $n = 5000$.

For the univariate case, that is $d = 1$, we compare our estimate with kernel estimates (with Gaussian kernel) (cf. Chapter 5 in Györfi et al. (2002)), local linear kernel estimates (cf. Section 5.4 in Györfi et al. (2002)), smoothing splines (cf. Chapter 20 in Györfi et al. (2002)), neural networks and regression trees (as implemented in the freely available statistics software R), by applying every one of these six estimates to samples of the above distributions. Since for $d > 1$, not all of these estimates are easily applicable in R, we compare our estimate only with neural networks and regression trees (again by applying each of these three estimates to samples of the above distributions), for $d > 1$. In all cases we choose the smoothing parameter of the estimates by splitting the sample, where for each simulation, the size of the training sample and the testing sample is $n/2$.

In order to compute the L_2 errors of the estimates, we use Monte Carlo integration, that is, we approximate

$$\int |m_l(x) - m(x)|^2 \mu(dx) = \mathbf{E}\left(|m_l(X) - m(X)|^2 | \mathcal{D}_l\right)$$

by

$$\frac{1}{N} \sum_{j=1}^{N} |m_l(\tilde{X}_j) - m(\tilde{X}_j)|^2,$$

where the random variables $\tilde{X}_1, \tilde{X}_2, \ldots$ are independent and identically distributed, with distribution μ, and moreover independent of $\mathcal{D}_l$. In the sequel we use $N = 3000$. Since this error is a random variable itself, we repeat the experiment 25 times with independent realizations of the sample, and report the mean and the standard deviation of the Monte Carlo estimates of the L_2 error.

Firstly we consider the case $d = 1$, and we examine the following four different regression functions:

- $m_1(x) = 2 \cdot \max\left\{1, \min\{3 + 2 \cdot x, 3 - 8 \cdot x\}\right\}$,
- $m_2(x) = \begin{cases} 1 & , x \leq 0, \\ 3 & , \text{ else}, \end{cases}$
- $m_3(x) = \begin{cases} 10 \cdot \sqrt{-x} \cdot \sin(8 \cdot \pi \cdot x), & -0.25 \leq x < 0, \\ 0, & \text{else}, \end{cases}$
- $m_4(x) = 3 \cdot \sin(\pi \cdot x/2)$.

Since there exists no 'typical' regression function, in terms of a type of functions that appears in most regression estimation problems, we tried to choose as differing functions as possible in this simulations, in order to analyse the behaviour of the maxmin estimate.

The choice of the underlying regression functions obviously is not an easy task, since different focuses lead to very different choices. However, if one wants to make such a comparison to established estimates, one has to make a selection. Figure 9 sketches the four considered univariate regression functions.

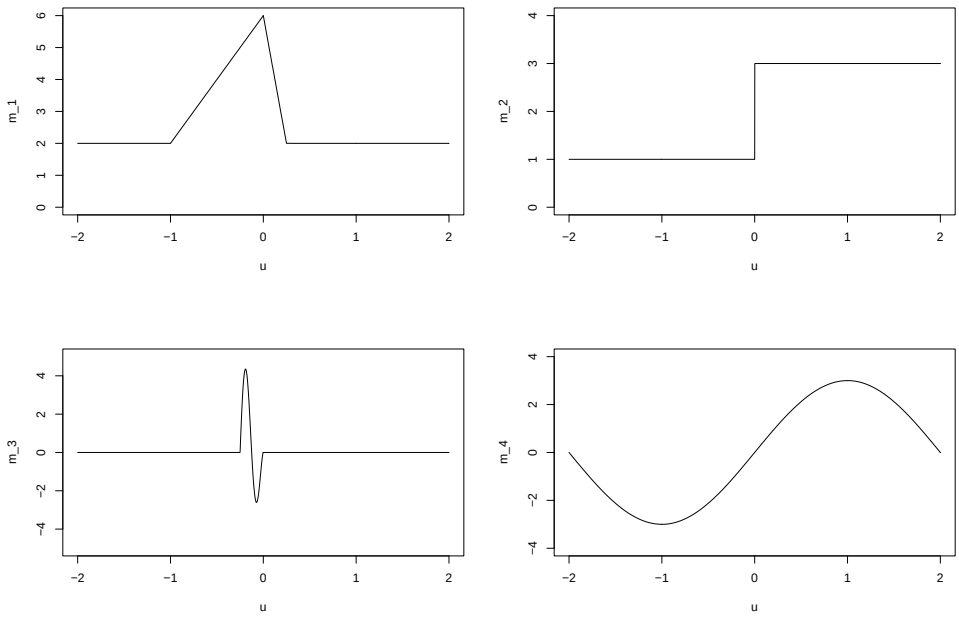

FIGURE 9. The four univariate regression functions.

Figure 10 shows these function, together with our maxmin estimate applied to a sample with variance $\sigma = 0.2$ and sample size $n = 500$.

In Tables 1 to 4, we report the error values for the maxmin estimate and the other five univariate regression estimates, which are applied to the simulated data as described above. In the tables we use the following abbreviations:

- kernel estimates with the Gaussian kernel (ker–est.)
- local linear kernel estimates (llk–est.)
- smoothing splines (s–splines)
- neural networks (nn–est.)
- regression trees (reg–trees)

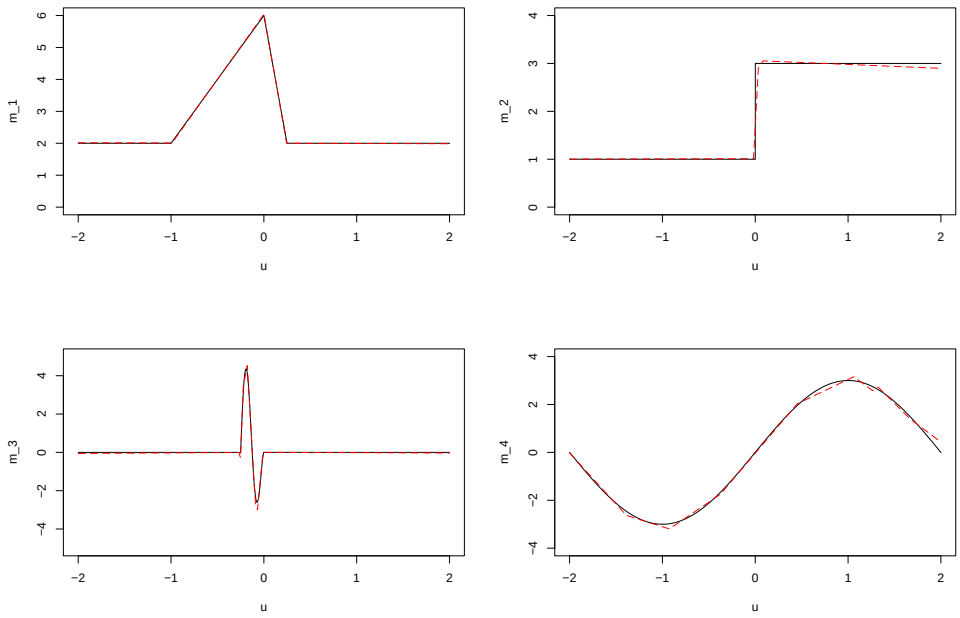

FIGURE 10. The four regression functions (solid lines) and the maxmin estimate (dash lines). ($\sigma = 0.2$, $n = 500$).

n	σ	ker–est.	llk–est.	s–splines	nn–est.	reg–trees	maxmin est.
500	0	0.0022	0.0005	0.0001	0.0020	0.0347	**0.0000**
		(0.0017)	(0.0004)	(0.0001)	(0.0004)	(0.0062)	(0.0000)
	0.5	0.0288	0.0278	0.0242	0.0161	0.0798	**0.0093**
		(0.0075)	(0.0078)	(0.0065)	(0.0039)	(0.0123)	(0.0048)
	1	0.0741	0.0816	0.0760	0.0438	0.2204	**0.0408**
		(0.0268)	(0.0389)	(0.0327)	(0.0206)	(0.0445)	(0.0254)
5000	0	0.0003	0.0003	**0.0000**	0.0006	0.0009	**0.0000**
		(0.0000)	(0.0000)	(0.0000)	(0.0002)	(0.0001)	(0.0000)
	0.5	0.0044	0.0043	0.0038	0.0030	0.0105	**0.0007**
		(0.0011)	(0.0009)	(0.0007)	(0.0008)	(0.0017)	(0.0005)
	1	0.0131	0.0121	0.0118	0.0091	0.1358	**0.0028**
		(0.0032)	(0.0036)	(0.0030)	(0.0020)	(0.0232)	(0.0015)

TABLE 1. Mean (standard deviation) of the L_2 error for the associated estimates. Regression function: m_1.

n	σ	ker–est.	llk–est.	s–splines	nn–est.	reg–trees	maxmin est.
500	0	0.0078	0.0096	0.0072	0.0110	0.0087	**0.0045**
		(0.0486)	(0.0047)	(0.0051)	(0.0047)	(0.0108)	(0.0046)
	0.5	0.0365	0.0396	0.0375	0.0165	0.0608	**0.0156**
		(0.0100)	(0.0087)	(0.0083)	(0.0052)	(0.0153)	(0.0110)
	1	0.0684	0.0806	0.0746	**0.0288**	0.2260	0.0431
		(0.0160)	(0.0171)	(0.0170)	(0.0184)	(0.0489)	(0.0240)
5000	0	0.0058	0.0074	0.0026	0.0040	0.0009	**0.0007**
		(0.0011)	(0.0013)	(0.0007)	(0.0009)	(0.0018)	(0.0011)
	0.5	0.0106	0.0119	0.0110	0.0051	0.0033	**0.0013**
		(0.0013)	(0.0013)	(0.0011)	(0.0009)	(0.0032)	(0.0008)
	1	0.0219	0.0241	0.0226	0.0076	0.1539	**0.0041**
		(0.0039)	(0.0039)	(0.0039)	(0.0021)	(0.0203)	(0.0022)

TABLE 2. Mean (standard deviation) of the L_2 error for the associated estimates. Regression function: m_2.

n	σ	ker–est.	llk–est.	s–splines	nn–est.	reg–trees	maxmin est.
500	0	0.0539	0.0450	0.0052	**0.0081**	0.1241	0.0234
		(0.0502)	(0.0402)	(0.0041)	(0.0064)	(0.0610)	(0.0585)
	0.5	0.0879	0.0922	0.0748	**0.0214**	0.1761	0.0255
		(0.0238)	(0.0383)	(0.0183)	(0.0101)	(0.0477)	(0.0178)
	1	0.2450	0.2749	0.2426	**0.0814**	0.3506	0.1201
		(0.0644)	(0.0735)	(0.0645)	(0.0490)	(0.0657)	(0.0556)
5000	0	0.0151	0.0175	0.0002	0.0010	0.0066	**0.0006**
		(0.0025)	(0.0054)	(0.0002)	(0.0003)	(0.0019)	(0.0001)
	0.5	0.0202	0.0220	0.0095	0.0030	0.0344	**0.0022**
		(0.0036)	(0.0060)	(0.0014)	(0.0008)	(0.0070)	(0.0006)
	1	0.0351	0.0357	0.0286	0.0080	0.1875	**0.0068**
		(0.0044)	(0.0047)	(0.0040)	(0.0041)	(0.0173)	(0.0021)

TABLE 3. Mean (standard deviation) of the L_2 error for the associated estimates. Regression function: m_3.

n	σ	ker–est.	llk–est.	s–splines	nn–est.	reg–trees	maxmin est.
500	0	0.0010	**0.0000**	**0.0000**	**0.0000**	**0.0000**	0.0041
		(0.0003)	(0.0000)	(0.0000)	(0.0000)	(0.0492)	(0.0125)
	0.5	0.0188	0.0084	**0.0072**	0.0129	0.0813	0.0207
		(0.0058)	(0.0027)	(0.0034)	(0.0060)	(0.0113)	(0.0069)
	1	0.0622	**0.0316**	0.0318	0.0564	0.2157	0.0634
		(0.0260)	(0.0157)	(0.0161)	(0.0321)	(0.0404)	(0.0192)
5000	0	0.0001	**0.0000**	**0.0000**	**0.0000**	0.0005	0.0037
		(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0017)
	0.5	0.0031	0.0014	**0.0010**	0.0014	0.0190	0.0061
		(0.0006)	(0.0004)	(0.0004)	(0.0005)	(0.0017)	(0.0014)
	1	0.0085	0.0040	**0.0034**	0.0049	0.0533	0.0113
		(0.0022)	(0.0018)	(0.0014)	(0.0023)	(0.0082)	(0.0035)

TABLE 4. Mean (standard deviation) of the L_2 error for the associated estimates. Regression function: m_4.

From these tables we can infer that, in the case of the distributions considered above, the maxmin estimate outperforms the other estimates if the sample size is large and the regression function is not globally smooth, such as the fourth regression function.

Next, we consider the case $d = 2$ and the following three regression functions:

- $m_5(u_1, u_2) = u_1 \cdot \sin(u_1^2) - u_2 \cdot \sin(u_2^2)$,
- $m_6(u_1, u_2) = \frac{4}{1+4\cdot u_1^2+4\cdot u_2^2}$,
- $m_7(u_1, u_2) = 6 - 2 \cdot \min\left\{3, 4 \cdot u_1^2 + 4 \cdot |u_2|\right\}$.

Figures 11, 12 and 13 show the three bivariate regression functions, together with the maxmin estimate, which is applied to a sample with variance $\sigma = 0.2$ and sample size $n = 5000$.

In Table 5 we compare our maxmin estimate with regression trees and neural networks. In the same way as above, we report the error values for the maxmin estimate, and the other two bivariate regression estimates, which are applied to the simulated data as described above. Here, our estimate most of the time is better than regression trees, and sometimes better and sometimes worse than neural networks.

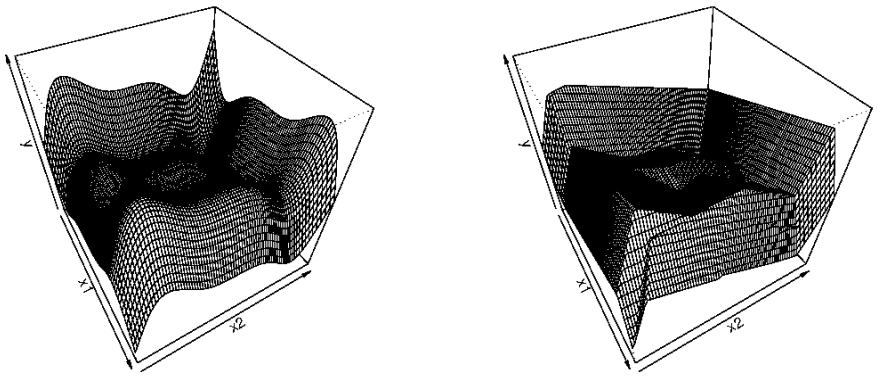

FIGURE 11. The regression function m_5 (left hand) and the maxmin estimate (right hand). ($\sigma = 0.2$, $n = 5000$).

	n	500			5000		
	σ	0	0.5	1	0	0.5	1
m_5	nn-est.	**0.0001** (0.0000)	**0.0657** (0.0206)	0.2897 (0.1105)	**0.0000** (0.0000)	**0.0049** (0.0021)	**0.02284** (0.0103)
	reg–trees	0.3718 (0.0551)	0.4128 (0.0458)	0.5610 (0.0922)	0.0613 (0.0070)	0.1002 (0.0088)	0.1872 (0.0169)
	maxmin est.	0.0796 (0.0170)	0.1449 (0.0310)	**0.2280** (0.0490)	0.0593 (0.0090)	0.0700 (0.0064)	0.0889 (0.0104)
m_6	nn-est.	**0.0015** (0.0006)	0.0822 (0.0211)	0.2026 (0.0438)	**0.0001** (0.0000)	**0.0110** (0.0034)	0.0339 (0.0076)
	reg–trees	0.0817 (0.0202)	**0.0123** (0.0261)	0.2062 (0.0621)	0.0083 (0.0006)	0.0312 (0.0041)	0.0607 (0.0073)
	maxmin est.	0.0134 (0.0040)	0.0540 (0.0135)	**0.1543** (0.0629)	0.0066 (0.0018)	0.0137 (0.0015)	**0.0293** (0.0048)
m_7	nn–est.	**0.0298** (0.0108)	0.1874 (0.0617)	0.4884 (0.1198)	**0.0078** (0.0011)	0.0253 (0.0033)	0.0699 (0.0112)
	reg–trees	0.3034 (0.1547)	0.3175 (0.1967)	0.3757 (0.1820)	0.0484 (0.0071)	0.0610 (0.0081)	0.0902 (0.0166)
	maxmin est.	0.0325 (0.0087)	**0.0868** (0.0321)	**0.1734** (0.0660)	0.0136 (0.0036)	**0.0176** (0.0046)	**0.0260** (0.0055)

TABLE 5. Mean (standard deviation) of the L_2 error.

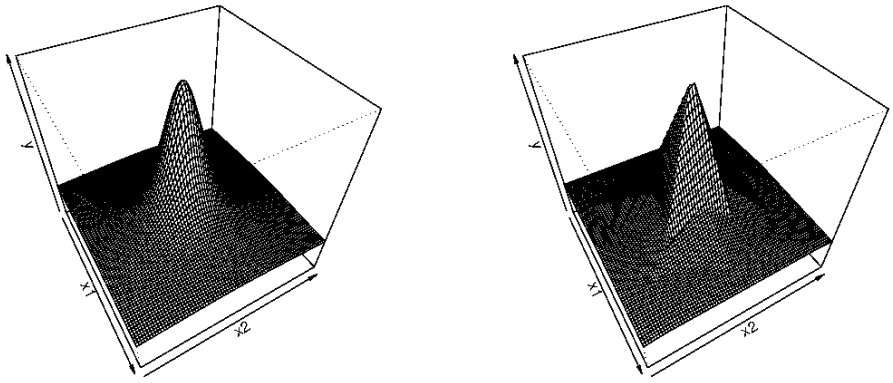

FIGURE 12. The regression function m_6 (left hand) and the maxmin estimate (right hand). ($\sigma = 0.2$, $n = 5000$).

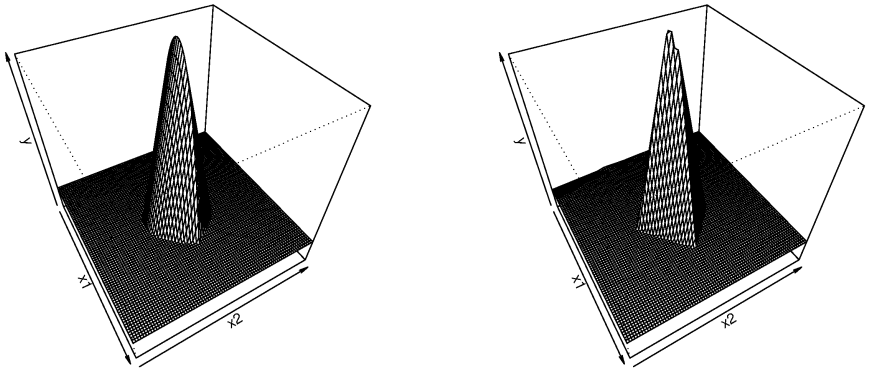

FIGURE 13. The regression function m_7 (left hand) and the maxmin estimate (right hand). ($\sigma = 0.2$, $n = 5000$).

Finally, we consider the case $d = 10$, where we used the following four regression functions for our simulations:

- $m_8(u_1, ..., u_{10}) = \sum_{j=1}^{10} (-1)^{j-1} \cdot u_j \cdot \sin(u_j^2)$,
- $m_9(u_1, ..., u_{10}) = m_7(u_1, u_2)$,
- $m_{10}(u_1, ..., u_{10}) = m_6(u_1 + ... + u_5, u_6 + ... + u_{10})$,
- $m_{11}(u_1, ..., u_{10}) = m_2(u_1 + ... + u_{10})$.

We compare our maxmin estimate again with regression trees and neural networks. In Table 6 we report the error values for the maxmin estimate and the other two multivariate regression estimates.

	n	500			5000		
	σ	0	0.5	1	0	0.5	1
m_8	nn–est.	5.5527 (0.1840)	5.4825 (0.2261)	5.6506 (0.2479)	4.7018 (0.1304)	4.6583 (0.1361)	4.7093 (0.1071)
	reg–trees	5.6535 (0.1817)	5.6297 (0.2013)	5.7852 (0.2513)	5.0029 (0.1515)	4.9726 (0.1431)	5.0189 (0.1139)
	maxmin est.	**4.4715** (0.1884)	**4.4842** (0.1593)	**4.5392** (0.1532)	**3.7220** (0.1526)	**3.7106** (0.1403)	**3.7852** (0.1250)
m_9	nn–est.	**0.0265** (0.0081)	**0.1790** (0.0531)	0.4805 (0.0917)	**0.0079** (0.0014)	**0.0247** (0.0023)	**0.0680** (0.0097)
	reg–trees	0.3011 (0.1826)	0.2980 (0.1073)	**0.3756** (0.2008)	0.0477 (0.0071)	0.0587 (0.0078)	0.0901 (0.0131)
	maxmin est.	0.6216 (0.1049)	0.8003 (0.1255)	0.9121 (0.0928)	0.0279 (0.0133)	0.0521 (0.0085)	0.1471 (0.0358)
m_{10}	nn–est.	0.2064 (0.0231)	0.2122 (0.0147)	0.2284 (0.0284)	0.2018 (0.0116)	0.1982 (0.0185)	0.2061 (0.0190)
	reg–trees	0.2033 (0.0226)	**0.2024** (0.0134)	**0.2053** (0.0263)	0.2028 (0.0116)	0.1987 (0.0186)	0.2039 (0.0190)
	maxmin est.	**0.1893** (0.0215)	0.2577 (0.0697)	0.2944 (0.0757)	**0.0236** (0.0035)	**0.0502** (0.0066)	**0.1135** (0.0232)
m_{11}	nn–est.	0.8902 (0.0180)	0.9057 (0.0286)	0.9270 (0.0381)	0.8711 (0.0126)	0.8766 (0.0126)	0.8738 (0.0139)
	reg–trees	0.9659 (0.0244)	0.9745 (0.0281)	1.0037 (0.0231)	0.9006 (0.0132)	0.9064 (0.0122)	0.9107 (0.0144)
	maxmin est.	**0.0732** (0.0338)	**0.2037** (0.1014)	**0.4585** (0.1099)	**0.0152** (0.0028)	**0.0258** (0.0057)	**0.0552** (0.0181)

TABLE 6. Mean (standard deviation) of the L_2 error for the associated estimates. The regression function is m_8, m_9, m_{10} or m_{11}, respectively.

Here none of the estimates is able to estimate m_8 well. The two other methods outperform our estimate, for m_9, which is a very simple function depending in fact only of two of the components of the predictor variable, but our estimate

clearly outperforms the other estimates, in case of $n = 5000$ and m_{10} and for $n \in \{500, 5000\}$ in case of m_{11}.

In summary, we can state that our estimate certainly performs well in comparison with the established estimates. Especially in view of the variety of the underlying problems and the fact that one usually has no a priori infomation about the shape of the underlying regression function, it is always helpful to have a variety of suitable estimates.

Moreover, this application to simulted data should have suggested that regression function estimation with maxmin functions is at least a good alternative to the established methods, and might sometimes be the better choice in applications.

Bibliography

[1] A.M. Bagirov. Minimization methods for one class of nonsmooth functions and calculation of semi-equilibrium prices," In: A. Eberhard et al. (eds.) Progress in Optimization: Contribution from Australia, Kluwer Academic Publishers: Dordrecht, pp. 147-175, 1999.

[2] A.M. Bagirov. A method for minimization of quasidifferentiable functions. *Optimization Methods and Software*, **17**:31–60, 2002.

[3] A.M. Bagirov, A.M. Rubinov, N.V. Soukhoroukova and J.Yearwood. Unsupervised and Supervised Data Classification Via Nonsmooth and Global Optimization. *Sociedad de Estadística e Investigación Operativa Top*, **11**:1–93, 2003.

[4] A.M. Bagirov and J. Ugon. Piecewise partially separable functions and a derivative-free method for large-scale nonsmooth optimization. *Journal of Global Optimization*, **35**:163–195, 2006.

[5] A.M. Bagirov, M. Ghosh and D. Webb. A derivative-free method for linearly constrained nonsmooth optimization. *Journal of Industrial and Management Optimization.* **2(3)**:319–338, 2006.

[6] A.M. Bagirov, B. Karasozen and M. Sezer. Discrete gradient method: a derivative free method for nonsmooth optimization. *Journal of Optimization Theory and Applications*, accepted for publication, 2007.

[7] A.M. Bagirov, C. Clausen and M. Kohler. An algorithm for the estimation of a regression function by continuous piecewise linear functions. *Computational Optimization and Applications*, accepted for publication, 2007.

[8] A.R. Barron. Complexity regularization with application to artificial neural networks. In *Nonparametric Functional Estimation and Related Topics*, Roussas, G., editor, 561–576. NATO ASI Series, Kluwer Academic Publishers, Dortrecht, 1991.

[9] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory* **39**:930–944, 1993.

[10] A. R. Barron. Approximation and estimation bounds for neural networks. *Neural Networks* **14**:115–133, 1994.

[11] A.R. Barron, A. Cohen, W. Dahmen and R. De Vore. Approximation and learning by greedy algorithms. *Annals of Statistics* **36**:64–94, 2006.

[12] S.G. Bartels, L. Kuntz and S. Sholtes. Continuous selections of linear functions and nonsmooth critical point theory. *Nonlinear Analysis, TMA.* **24**:385–407, 1995.

[13] G. Beliakov and M. Kohler. Estimation of regression functions by Lipschitz continuous functions. 2005.

[14] S.N. Bernstein. *The Theory of Probabilities.* Gastehizdat Publishing House, Moscow. 1946.

[15] P. Billingsley. *Probability and Measure*. John Wiley, New York, 1995.

[16] L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **65**:181–237, 1983.

[17] L. Breiman, J. H. Friedman,R. H. Olshen and C. J. Stone. *Classification and regression trees.* Wadsworth, Belmont, CA, 1984.

[18] L. Breiman. Arcing classifiers (with discussion). *Annals of Statistics* **26**:801–849, 1998.

[19] L. Breiman. Prediction games and arcing algorithms. *Neural Computation* **11**:1493–1517, 1999.

[20] P. Bühlmann. Boosting for high-dimensional linear models. *Annals of Statistics* **34**:559–583, 2006.

[21] R.M. Clark. A calibration curve for radiocarbon dates. *Antiquity*, **49**:251–266, 1975.

[22] V.F. Demyanov and A.M. Rubinov, *Constructive Nonsmooth Analysis*. Peter Lang, Frankfurt am Main, 1995.

[23] L. Devroye and T.J. Wagner. Distribution-free consistency results in nonparametric discrimination and regression finction estimation. *Annals of Statistics*, **8**:231–239, 1980.

[24] L. Devroye. Any discrimination rule can have an arbitrarily bad probability of error for finite sample size. *IEEE Transactions on Pattern Analysis and Maschine Intelligence*, **4**:154–157, 1982.

[25] L. Devroye. Automatic pattern recognition: A study of the probability of error. *IEEE Transactions on Pattern Analysis and Maschine Intelligence*, **10**:530–543, 1988.

[26] L. Devroye, L. Györfi, A. Krzyżak and G. Lugosi. On the strong universal consistency of nearest neighbor regression function estimates. *Annals of Statistics*, **22**:1371–1385, 1994.

[27] R. Dudley. Central limit theorems for empirical measures. *Annals of Probability*, **6**:899–929, 1978.

[28] J. Elstrodt. *Maß- und Integrationstheorie*, Springer Verlag, Heidelberg, 1996.

[29] Y. Freund and R. Schapire. *Experiments with a new boosting algorithm.* In Maschine Learning: Proc. Thirteenth International Conference 148–156. Morgan Kaufmann, San Francisco, 1996.

[30] J. Friedman and J. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, **23**:881–889, (1974).

[31] J. Friedman and W. Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, **76**:817–823, 1981.

[32] J.H. Friedman. Multivariate adaptive regression splines (with discussion). *Annals of Statistics*, **19**:1–141, 1991.

[33] J. Friedman, T. Hastie and R. Tibschirani. Additive logistic regression: A statistical view of boosting (with discussion). *Annals of Statistics* **28**:337–407, 2000.

[34] J. Friedman. Greedy function approximation: A gradient boosting maschine. *Annals of Statistics* **29**:1189–1232, 2001.

[35] U. Grenander. *Abstract Inference*. Wiley, New York, 1981.

[36] L. Györfi, M. Kohler, A. Krzyżak and H. Walk. A *Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics, Heidelberg, 2002.

[37] T. Hastie, R. Tibshirani and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag, New York, 2001.

[38] D. Haussler. Decision theoretic generalizations of PAC model for neural net and other learning applications. *Information and Computation*, **100**:78–150, 1992.

[39] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of American Statistical Association*. **58**:13–30, 1963.

[40] I.A. Ibragimov and R.Z. Khasminskii. On nonparametric estimation of regression. *Doklady Akademii Nauk SSSR*, **252**:780–784, 1980.

[41] I.A. Ibragimov and R.Z. Khasminskii. *Statistical Estimation: Asymtotic Theory*. Springer Verlag, New York, 1981.

[42] I.A. Ibragimov and R.Z. Khasminskii. On the bounds for quality of nanparametric regression finction estimation. *Theory of Probability and its Applications*, **27**:81–94, 1982.

[43] M. Kohler. Nonparametric estimation of piecewise smooth regression functions. *Statistics & Probability Letters* **43**:49–55, 1999.

[44] M. Kohler. Inequalities for uniform deviations of averages from expectations with applications to nonparametric regression. *Journal of Statistical Planning and Inference* **89**:1–23, 2000.

[45] M. Kohler. Nonparametric regression with additional measurements errors in the dependent variable. *Journal of Statistical Planning and Inference* **136**:3339–3361, 2006.

[46] A. Kolmogorov and V. Tikhomirov. ε–entropy and ε–capacity of sets in function spaces. *Translations of the American Mathematical Society*, **17**:277–364, (1961).

[47] W.S. Lee, P.L. Bartlett and R.C. Williamson. Efficient agnostic Learning of neural networks with bounded fan-in. *IEEE Transactions on Information Theory*. **22**:2118–2132, 1996.

[48] G. Lugosi K. and Zeger. Nonparametric estimation via empirical risk minimization. *IEEE Transactions on Information Theory*, **41**:677-687, 1995.

[49] R. Mifflin. Semismooth and semiconvex functions in constrained optimization. *SIAM Journal on Control and Optimization*, **15**:957–972, 1977.

[50] A. Nobel. *On uniform laws of averages*. PhD Thesis, Department of Statistics, Stanford University, Stanford, CA, 1992.

[51] D. Nolan and D. Pollard. U-processes: Rates of convergence. *Annals of Statistics*, **15**:780–799, 1987.

[52] D. Pollard. *Convergence of Stochastic Processes*. Springer Verlag, New York, 1984.

[53] D. Pollard. *Empirical Processes: Theory and Applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, Insitute of Mathematical Statistics, Hayward, CA, 1990.

[54] The R Project for Statistical Computing, available on: www.r-project.org.

[55] L. Schumaker. *Spline functions: Basic Theory.* Wiley, New York, 1981.

[56] C.J. Steele. *Combinatirial entropy and uniform limit laws.* PhD Thesis, Stanford University, Stanford, CA, 1975.

[57] C.J. Stone. Consistent nonparametric regression. *Annals of Statistics*, **5**:595–645, 1977.

[58] C.J. Stone. Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, **10**:1040–1053, 1982.

[59] C.J. Stone. Additive regression and other nonparametric models. *Annals of Statistics*, **13**:689–705, 1985.

[60] C.J. Stone. The use of polynomial splines and their tensor products in multivariate function estimation. *Annals of Statistics*, **22**:118–184, 1994.

[61] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, **16**:264–280, 1971.

[62] S. van de Geer. *Empirical Processes in M–Estimation.* Cambridge University Press, 2000.

[63] G. Wahba and S. Wold. A completely automatic French curve: fitting spline functions by cross-validation. *Communications in Statistics - Theory and Methods*, **4**:1–17, 1975.