Change point estimation
in noisy Hammerstein integral equations

Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultäten
der Georg-August-Universität zu Göttingen

vorgelegt von
Sophie Hanna Clara Frick geb. Bruns
aus Detmold

Göttingen, 2010

Abstract (deutsch)

In der vorliegenden Arbeit wird das inverse Regressions Model $Y = Hf(X) + \varepsilon$ für verschiedene Klassen nicht linearer Hammerstein Integral Operatoren H betrachtet. Wir diskutieren insbesondere das Problem der Identifizierbarkeit in Abhängikeit des Integralkernes. Vorgestellt werden Schätzer für parametrische Funktionen f mit Unstetigkeiten verschiedener Ordnung, wie beispielsweise stückweise Polynome mit Knicken oder Sprüngen, bzw. Splines mit freien Knoten. Konvergenzraten und asymptotische Normalität der Schätzer werden entwickelt und an einem Datenbeispiel aus der Rheologie illustriert. Eine Erweiterung des Models auf Funktionen f aus Approximationsräumen von parametrischen stückweise stetigen Funktionen wird diskutiert.

Abstract (english)

We consider the inverse regression model $Y = Hf(X) + \varepsilon$ for several classes of nonlinear Hammerstein integral operators H. In particular identifiability depending on the integral kernel is discussed. We introduce estimators for parametric functions f with discontinuities of certain order including piecewise polynomials with kinks or jumps or free-knot splines respectively. We derive rates of convergence and asymptotic normality of these estimators and a data example from rheology illustrates the results. An extension of the model for functions f from approximation spaces of parametric piecewise continuous functions is presented.

Contents

Contents

<div align="right">

Chapter 1

</div>

Introduction

Let $X = (x_1, ..., x_n)$ be a (possibly random) vector of $n \in \mathbb{N}$ design points in an interval (a, b), for $a, b \in \mathbb{R}$. We consider the inverse regression model

$$y_i = H f_0(x_i) + \varepsilon_i \text{ for } i = 1, ..., n, \tag{1.1}$$

where $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)$ denotes the independent identical distributed observation error, which is assumed to be independent of X with mean zero. Further, H denotes a *Hammerstein integral operator* $H : L_2([a, b]) \longrightarrow L_2([a, b])$, defined by

$$f(\cdot) \longmapsto H f(\cdot) := \int_a^b \varphi(\cdot, y) L(f(y), y) \mathrm{d}y. \tag{1.2}$$

Note, that H is in general nonlinear and can be written as a composition $H = \Phi \circ \mathcal{L}$ of a linear integral operator $\Phi : L_2([a, b]) \longrightarrow L_2([a, b])$ and a possibly nonlinear so called, *Niemitzky operator* $\mathcal{L} : L_2([a, b]) \longrightarrow L_2([a, b])$, defined by

$$f(\cdot) \longmapsto \Phi f(\cdot) := \int_a^b \varphi(\cdot, y) f(y) \mathrm{d}y, \tag{1.3}$$

and

$$f(\cdot) \longmapsto \mathcal{L} f(\cdot) := L(f(\cdot), \cdot), \tag{1.4}$$

respectively. We are concerned with reconstructing the unknown function f_0 from the observations $(X, Y) = ((x_1, y_1), ..., (x_n, y_n))$ given by the model in (1.1).

Example. In order to emphasize the relevance of the model in (1.1), we introduce an example from rheology. Here, one is interested in the relaxation behavior of polymers after expansion. This behavior is described by the *relaxation time spectrum*, which is known to be a piecewise linear and continuous function on $[a, b] \in \mathbb{R}^+ \setminus \{0\}$, with two change points, where the slopes of the linear pieces may change. More details about the physical background can be found in [47]. The relaxing process cannot be observed directly, but indirectly as an image under an integral operator H as in (1.2), where Φ
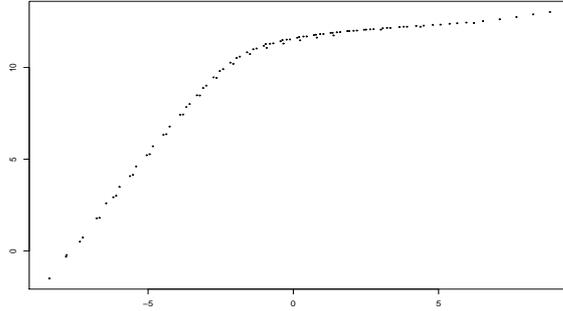
Figure 1.1: The logarithm of design points plotted against the logarithm of the noisy observations.

is as in (1.3), with kernel $\varphi(x, y) = \dfrac{x^2 y^2}{1 + x^2 y^2}$ and $\mathcal{L}$ as in (1.4), with $\mathcal{L}f(y) = y^{-1} e^{cf(y)}$ (with constant $c \neq 0$). Additionally the data are assumed to be disturbed by some noise, as for instance measuring inaccuracy (see Figure 1).

The aim is now to reconstruct the piecewise linear function describing the relaxation time spectrum from this observation (red line in Figure 1.2). Thus, we want to estimate the change points and slopes of this function. This will be done by choosing these parameters such that the image of the corresponding function under the operator H in (1.2) minimizes the sum of the squared distance from the data, that is we use a parametric least squares estimator. As it will follow from the general results in this thesis, this estimator converges with rate $n^{-1/2}$ to the true kink function. Furthermore, we will show that the corresponding estimator for the parameter vector, consisting of slopes and kink locations, is asymptotically distributed according to a multivariate normal distribution. This distribution will be used to calculate $(1 - \alpha)$-confidence bands for the estimated function, i.e. an area around the estimator, which contains the true function with probability $1 - \alpha$ (e.g. the green area in Figure 1.2).

Inverse problems. Motivated by this example, we intend to come up with a general regularization theory for reconstructing piecewise continuous functions from perturbed images of integral operators of type (1.2). These operators are well known to generate *ill-posed problems*. For the concept of *well-* and *ill-posedness* being a basic task in this thesis, we want to give a short introduction, based on the formulation of Kress in [35, Def. 15.1]. According to this definition, an equation

$$A\varphi = f,$$

with an operator $A : U \to V$ from a subset U of a normed space X into a subset V of a normed space Y, is called *well-posed* if A is bijective and the inverse operator
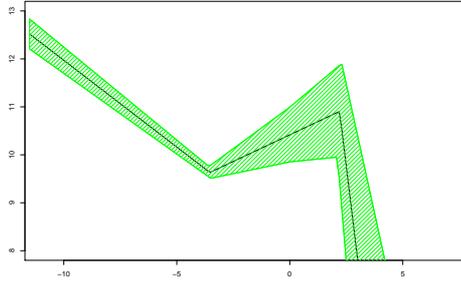
Figure 1.2: 0.98-confidence bands for the estimated relaxation time spectrum

$A^{-1}: V \to U$ is continuous. Otherwise, the equation is called *ill-posed*.

This definition, dating back to the concept of Hadamard (see [29]), covers three types of ill-posedness: *nonexistence* of a solution (A is not surjective), *nonuniqueness* (A is not injective) and *instability* (the solution φ does not depend continuously on the image f). Among these three types, instability is the most delicate to deal with. In this case, small errors in the data may cause arbitrary large errors in the solution and therefore for instance, make computed solutions, based on direct inversion, useless.

However, in general the three types of ill-posedness are not independent. For example stability follows from existence and uniqueness, that is bijectivity of the operator, if A is continuous and U is compact. Furthermore, this example reveals that well-posedness is a property of the operator A together with the solution space X and the data space Y, including their norms.

General problem. As mentioned above, Hammerstein integral operators can generate ill-posed problems. To be more precise, we have to deal with two inverse problems, i.e. the linear one, given by the integral operator in (1.3) and the nonlinear one given by the Niemitzky operator in (1.4). This offers the opportunity to handle these problems separately. To this aim we consider the reduced linear inverse regression model

$$y_i = (\Phi f_0)(x_i) + \varepsilon_i \text{ for } i = 1, ..., n, \tag{1.5}$$

where X and ε are as in (1.1) and $f_0 = \mathcal{L}\tilde{f}_0$. We assume that the operator $\mathcal{L}$ satisfies appropriate conditions, which allow the reconstruction of a function $\tilde{f}_0$ from the image $\mathcal{L}\tilde{f}_0$. In other words, in this thesis we will concentrate on the linear inverse problem, that is the reconstruction of f_0 from (1.5), which due to its ill-posedness requires appropriate regularization. Starting with a simple least squares estimator, we will present three different regularization strategies, depending on the a priori knowledge on the true function f_0:

We begin with the case, where $f_0(y) = f(y, \theta_0)$ is piecewise continuous, with known number of change points, determined by the parameter vector $\theta_0 \in \Theta$, with compact

$\Theta \in \mathbb{R}^d$ for some $d \in \mathbb{N}$. This class covers the polymer example above as a special case. Here the regularization consists in the restriction on the parameter set to be compact. Subsequently, we consider the case, where the number of change points is finite but unknown, where we regularize the least squares estimator by penalizing the number of change points. Finally, we extend this function class to so called *approximation spaces*, which are characterized by a common upper bound for the speed of approximation by piecewise continuous functions, where in addition to the number of change points, we penalize the L_2-norm of the estimated function.

Injectivity. As already mentioned above, injectivity of the operator Φ conditions well-posedness of the corresponding inverse problem and is a key ingredient in the consistency proofs in this thesis. We consider two basic classes of operators in (1.3), namely operators with product kernels $\varphi(x, y) = \phi(xy)$ and operators with convolution kernels $\varphi(x, y) = \phi(x - y)$. In both cases we give conditions, which assure injectivity of the corresponding integral operator.

The requirement on a product kernel is based on a general form of the *Müntz Theorem* (cf. [7]), which extends the original statement about denseness of monomials in the space of continuous functions to L_2-spaces. Moreover, we present two injectivity result for operators with convolution kernel in terms of the kernels Fourier transform. In this context, we additionally introduce an injectivity condition for general symmetric and positive definite kernels (not restricted to any of the above classes), which is based on the theory of native Hilbert spaces. We remark, however, that the asymptotic results of this paper are not restricted to this selection. They are valid for every injective integral operator as in (1.3) with piecewise continuous kernel and Lipschitz continuous image $\Phi f \in L_2([a, b])$.

However, in some applications (as e.g. image inpainting), there often occur integral operators, which are not injective. In such cases, one may use minimum norm solutions, for example, to obtain practicable estimators. This concept will not be treated in this thesis. For more details we refer e.g. to [49].

Known number of change points. For the case, when f_0 is a parametric piecewise continuous function $f_0(y) = f(y, \theta_0)$ with known number of change points, under certain assumptions on error and design, we show an $n^{-1/4}$-convergence rate of the least squares estimator $\hat{f}_n(y) = f(y, \hat{\theta}_n)$. Furthermore, we obtain $n^{-1/2}$-rates for the convergence of the estimated parameter $\hat{\theta}_n$ to the true parameter θ_0 and show that it is asymptotically distributed according to a multivariate normal distribution.

The exact specification of the considered class of functions is crucial: it turns out that it suffices to assume existence and uniform L_2-boundedness of the first derivative of $\theta \mapsto f(y, \theta)$ in the parameter vector $\theta \in \Theta$ for almost every y. This condition allows for a general estimate of the entropy of the class of parametric piecewise continuous functions, which is an important ingredient in the proof of consistency. This property also implies

continuous differentiability of the mapping $\theta \mapsto \Phi f(y, \theta)$, which in turn paves the way to the second order expansion of the expectation of the score function, required for the proof of asymptotic normality.

The applied techniques furthermore cover the case where we consider subclasses of this parametric functions, that allow for dependencies among the parameter components of θ. Such dependencies occur, for example, if f_0 is additionally known to be continuous, that is f_0 has a known number of kinks (as in the introductory example from polymer physics). We show, that the continuity assumption on f_0 improves the convergence rate of the least squares estimate $f(y, \hat{\theta}_n)$, where the improvement depends on the smoothness of the pieces between the kinks. (Here, "pieces" means the restricted function $f(y, \theta) \mid_{y \in [\tau_1, \tau_2)}$ for some interval $[\tau_1, \tau_2) \subset [a, b]$ between two change points τ_1 and τ_2.) For example, a restriction to kink functions with Lipschitz continuous pieces yield convergence rates of $n^{-1/2}$.

For functions with known number of change point, we investigate the asymptotic distribution of the estimated parameter $\hat{\theta}_n$ for calculation of a confidence ellipsoid, which in turn generates confidence bands for $f(y, \hat{\theta}_n)$ (c.f. the introductory example). Because we have to deal with many parameter components, this leads to the non trivial task of constructing simultaneous confidence bands. Remarkably, it turns out, that an *approximative* method, based on the *studentized maximum modulus* statistic, proves to be the best way to determine confidence bands for functions with abrupt changes as jumps or kinks. Here, "best" means not only with respect to the computational effort, but also concerning the width of the resulting bands: the exact confidence ellipsoid for the parameter estimate would lead to more expanded confidence bands in comparison with the bands corresponding to the actually more conservative set, based on the studentized maximum modulus statistic.

Unknown number of change points. When the number of change points of the objective function in (1.5) is not known, we employ penalized least squares for estimation. Here we use the number of jumps as penalty term, which yields a minimization functional, known as *Potts functional*. We show that under the additional assumption of subgaussian tails of the error distribution, the number of change points can be asymptotically estimated correctly with probability one. Thus, this case asymptotically coincides with the case of known number of change points.

Approximation spaces. Finally, we disscuss the extension of the considered function set to approximation spaces $\mathcal{A}^\alpha$ of piecewise continuous functions. These spaces are defined by their approximation properties, that is all functions of an approximation space have common upper bounds (depending on an index α) for the error of approximation by the respective function set. A known example are Besov spaces, which are generated by approximation by trigonometric polynomials (cf. [18, Thm. 9.2, Chap. 7, §9]). In general, characterization of approximation spaces of a given function set is quite involved

and will not be discussed in this work. We refer to the book of deVore and Lorentz [18] instead, which gives a detailed discussion of this problem.

If the function f_0 in (1.5), is contained in such an approximation space, we will employ a least squares estimator, which is penalized by the number of change points and the L_2-norm. The minimization is carried out over the set of piecewise continuous functions with finite but arbitrary number of change points. We will show that this estimator is consistent and under certain source conditions converges up to a rate arbitrarily close to $n^{-1/4}$, provided f_0 can be approximated fast enough. We note, that by the assumption, that f_0 is contained in an approximation space, we leave the domain of parametric problems. This does not only imply more advanced penalization strategies and worse convergence rates.

A further consequence is that now the degree of ill-posedness of the operator has a sustainable impact on the convergence results. This is expressed in the fact that we need f_0 to satisfy certain source conditions, depending on the smoothing properties of the operator. That means, the higher the degree of ill-posedness, the more rigorous is the smoothness assumption on f_0. If we do not impose smoothness assumptions the penalized least squares estimator would still converge, but possibly with arbitrary slow rate (cf. [21, Prop. 3.11]). This is a general difference to the parametric case, where the convergence rates do not depend on the spectral properties of the operator.

Related work. Next we want to compare our results to existing literature. The Hammerstein equation of second kind appeared in the earlier 30s (cf. [30]) as a general model for study of semi-linear boundary-value problems. The kernel $\varphi(x, y)$ typically arises as the Green's function of a differential operator. Generally, if the kernel is positive, then methods of positive operators are applicable to study solutions of Hammerstein equations see, e.g. [2] or [23] and for application of further methods [3] and [50].

There exists a vast literature concerning estimation of change points in direct as well as inverse problems. The discussion of the inverse setting started with Neumann [40], who considered the case, where the function of interest is bounded, has one jump and is Lipschitz continuous elswhere. He estimated the jump location at a rate of $\min(n^{-1/(2\beta+1)}, n^{-1/(\beta+3/2)})$, where $\beta > 0$ depends on the decreasing speed of the Fourier transform of the noise density in his observation model. In connection with *two phase regression*, Hinkley [32] obtained a $n^{-1/2}$ rate for an inverse regression model with convolution kernel $\varphi(x, y) = \mathbf{1}_{[0,\infty)}(x - y)$. Similar results were obtained by Feder [24] for more general regression models. Most recently we published an analysis of the regression model in (1.5) in [9], where the kernel φ is a Lipschitz continuous convolution kernel and the objective functions are step function. The results in this thesis, cover this setting as a special case.

A generalization to piecewise smooth non parametric functions, has been developed by Goldenshluger et al. in [27], where the rates of convergence depend on the Fourier transform of the error density as in the model of Neumann [40], we mentioned above.

Jump penalized least squares, as we used them for estimation in the case of unknown number of change points, were introduced by Potts [44] for a binary spin system to more than two states. Further, Boysen et al. in [10] aim for approximating a regression function by piecewise constant functions. In contrast to the present model in (1.5), they consider the *direct* case, i.e. the case $H = id$, where their framework covers the case, when the true function f_0 is contained in an approximation space, too. Under different error assumptions, they obtain similar rates of convergence, with a slightly less rigorous regularization.

Outline. We start with an introduction of the used notation and assumptions in Chapter 2. In particular we give a rigorous definition of piecewise continuous functions used throughout this thesis. In Chapter 3, we give a detailed classification of integral operators Φ and especially disscuss the question of injectivity. In Chapter 4 we will introduce the least squares estimator and its asymptotic behavior, in the case, where the function of interest is piecewise continuous with known number of jumps. Subsequently, in Chapter 5 we discuss the application of this results in general and for the special example of rheology, mentioned above. In Chapter 6 we consider the case, where the number of jumps is not known and study the asymptotic behavior of the jump-penalized least squares estimator. The case, when the true function is contained in an approximation space of piecewise continuous functions is discussed in Chapter 7. Finally, we collect some technical tools and important results concerning native Hilbert spaces and empirical processes in the Appendix.

Chapter 2

Definitions and assumptions

In this Chapter we are going to fix the notation used throughout this thesis and introduce basic definitions and assumptions. In particular we will establish the class of *parametric piecewise continuous functions*.

2.1 Model and notations

Norms and measures Throughout this work we assume that $a, b \in \mathbb{R}$, $a < b$ and $r, k \in \mathbb{N} \setminus \{0\}$. Moreover, for functions $g, f : [a, b] \to \mathbb{R}$, we denote by $\|f\|_{L_2([a,b])}$ the L_2-norm and by $\langle f, g \rangle_{L_2([a,b])}$ the corresponding inner product. The essential supremum is denoted by $\|f\|_\infty$ and additionally we define the empirical norm and the empirical inner product by

$$\|f\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} f(x_i)^2 \quad \text{and} \quad \langle f, g \rangle_n = \frac{1}{n} \sum_{i=1}^{n} f(x_i) g(x_i),$$

where $x_1, ..., x_n$ are given design points. Accordingly, we define the empirical measure as $P_n := n^{-1} \sum_{i=1}^{n} \delta_{x_i}$. For a vector $Y = (y_1, ..., y_n)$, with a slight abuse of notation, we use the expression $\|Y\|_n$, which means $\|\sum_{i=1}^{n} y_i \mathbf{1}_{x_i}\|_n$. The same holds, if a vector of length n occurs in the empirical inner product.

For vectors $\theta, \theta_1, \theta_2 \in \mathbb{R}^d$, we use the Euclidean norm $|\theta|_2$ and the maximum norm $|\theta|_\infty$ and by $(\theta_1, \theta_2) \subset \mathbb{R}^d$ we denote the segment between θ_1 and θ_2, that is $(\theta_1, \theta_2) := \{\theta \in \mathbb{R}^d \mid \theta = \theta_1 + t(\theta_2 - \theta_1), \text{ for } t \in (0, 1)\}$.

For a measure P we denote the expectation of a measurable function f by

$$\mathbf{E}f = \mathbf{E}_P f := \int f \mathrm{d}P.$$

Convergence in probability and in distribution For $d \in \mathbb{N}$, a sequence of random vectors $X_n \in \mathbb{R}^d$, some constant $c \in \mathbb{R}^d$ and a random vector $X \in \mathbb{R}^d$, we write

$$X_n \stackrel{P}{\longrightarrow} c \quad \text{and} \quad X_n \stackrel{P}{\longrightarrow} X,$$

if X_n converges in probability to c and X, respectively. Furthermore, convergence in distribution is denoted by

$$X_n \xrightarrow{\mathcal{D}} X;$$

if X has a distribution with standard code, such as $N(0,1)$, then also by $X_n \xrightarrow{\mathcal{D}} N(0,1)$.

We say that X_n is *bounded in probability*, denoted as $X_n =: O_P(1)$, if for all $\varepsilon > 0$, there exists some $M > 0$ such that

$$\limsup_{n \to \infty} P(|X_n| > M) < \varepsilon.$$

Moreover, we introduce abbreviations for terms, that are bounded in probability or converge in probability to zero. For a given sequence of random variables R_n we agree upon

$$X_n = o_p(R_n) \quad \text{means: there exists} \quad Y_n \text{ such that } X_n = Y_n R_n \text{ and } Y_n \xrightarrow{P} 0;$$
$$X_n = O_p(R_n) \quad \text{means: there exists} \quad Y_n \text{ such that } X_n = Y_n R_n \text{ and } Y_n = O_P(1).$$

This means, that the sequence X_n converges in probability to zero or is bounded in probability at the "rate" R_n, respectively. For deterministic sequences X_n and R_n, the stochastic O-symbols reduce to the usual o and O from calculus. Note, that for a random variable a_n there may occur the notation $a_n \leq O_p(1)$ (or $a_n \leq o_p(1)$). This precisely means, that there exists a random variable $X_n = O_P(1)$ (or $X_n = o_P(1)$), such that $a_n \leq X_n$, which in turn implies $a_n = O_P(1)$ (or $a_n = o_P(1)$) only if $a_n \geq 0$ for all $n \in \mathbb{N}$.

Some rules of calculus for this symbols are summarized in Section 8.1 in the Appendix.

Assumptions The following assumptions fix up the conditions of the model in (1.5) with respect to the error ε and the design X.

Assumption A. (*Assumptions on the error*) Throughout this thesis we assume, that

A1: the vector $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)$ consists of independent identically distributed random variables with mean zero for every n and $\mathbf{E}(\varepsilon_1^2) = \sigma^2 < \infty$.

In some situations the error is additionally assumed to satisfy the following sub-gaussian condition.

A2: ε satisfies **A1** and there exists some $\alpha > 0$ such that $\mathbf{E}(e^{\varepsilon_1^2/\alpha}) < \infty$.

Assumption B. (*Assumptions on the design*) The design points $x_1, ..., x_n$ are independent of the error terms $\varepsilon_1, ..., \varepsilon_n$. Moreover there exists a function $s : [a,b] \to [s_u, s_l]$ with $0 < s_u < s_l < \infty$ and $\int_a^b s(x)\mathrm{d}x = 1$ such that

$$\frac{i}{n} = \int_a^{x_{(i)}} s(x)\mathrm{d}x + \delta_i$$

with $\nu_n := \max_{i=1,...,n} |\delta_i| = o_p(1)$. Here $x_{(i)}$ denotes the i-th order statistic of $x_1, ..., x_n$.

Assumption **B** covers random designs as well as fixed designs. If the design points $x_1, ..., x_n$ are nonrandom, the $o_p(1)$ term above is to be understood as $o(1)$.

2.2 Piecewise continuous functions

In this section, we introduce the definition of the main object of this thesis, i.e. parametric piecewise continuous functions. To this end, we begin with the definition of parametric functions, which afterwards serve as support-functions for the continuous pieces of the piecewise continuous functions.

Definition 2.2.1. *Assume that $\Psi \subset \mathbb{R}^r$ is compact, with $\overline{\Psi^\circ} = \Psi$ and choose $M > 0$ such that $|\vartheta|_\infty \leq M$ for all $\vartheta \in \Psi$. Let*

$$
\begin{aligned}
\mathfrak{f} : [a,b] \times \Psi &\longrightarrow \mathbb{R} \\
(y, \vartheta) &\longmapsto \mathfrak{f}(y, \vartheta)
\end{aligned}
$$

be a function satisfying the following conditions:

i) $y \mapsto \mathfrak{f}(y, \vartheta)$ is continuous for all $\vartheta \in \Psi$,

ii) $\vartheta \mapsto \mathfrak{f}(\cdot, \vartheta)|_{[\rho_1, \rho_2]}$ is injective for all $a \leq \rho_1 < \rho_2 \leq b$ and the partial derivatives $\frac{\partial}{\partial \vartheta_j} \mathfrak{f}(y, \vartheta)$ exist for $1 \leq j \leq r$ and all $y \in [a,b]$ and they are continuous in $y \in [a,b]$ as well as in $\vartheta \in \Psi$, and

iii) there exists a function $g \in L_2([a,b])$, such that for $j = 1, ..., r$

$$
\left| \frac{\partial}{\partial \vartheta_j} \mathfrak{f}(y, \vartheta) \right| \leq g(y),
$$

for almost every $y \in [a,b]$.

*Then, $\mathcal{F} := \{\mathfrak{f}(\cdot, \vartheta) \mid \vartheta \in \Psi\}$ is called a family of **continuous parametric functions** with parameter domain Ψ.*

For example we consider the family of constant functions

$$
\mathcal{F}_T := \{x \mapsto \mathfrak{f}(x, \vartheta) = \vartheta \mid \vartheta \in \Psi \subset \mathbb{R}, \ |\vartheta| \leq M\}, \tag{2.1}
$$

or the family of linear functions

$$
\mathcal{F}_L := \{x \mapsto \mathfrak{f}(x, \vartheta) = \vartheta_1 + \vartheta_2 x \mid \vartheta \in \Psi \subset \mathbb{R}^2, \ |\vartheta|_\infty \leq M\}. \tag{2.2}
$$

In the following we could also assume that $\mathcal{F}$ is the union of a finite number of families satisfying the conditions of Definition 2.2.1, where the parameter sets Ψ may be different (especially they may have different dimensions r). However, in order to keep things simple, we restrict Definition 2.2.1 to one family of parametric functions, only.

Given a family $\mathcal{F}$, we now define parametric piecewise continuous functions.

Definition 2.2.2. *Assume that $\mathcal{F}$ is a family of continuous parametric functions, as defined in Definition 2.2.1, with parameter domain Ψ. A function $f \in L_2([a,b])$ is called a* **parametric piecewise continuous function** *(pc-function) with k change points, if there exists a partition $a = \tau_0 < \tau_1 < ... < \tau_{k+1} = b$ and parameter vectors $\vartheta^1, .., \vartheta^{k+1} \in \Psi$, such that*

$$f(\cdot, \vartheta^1, \tau_1, ..., \vartheta^k, \tau_k, \vartheta^{k+1}) = \sum_{j=1}^{k+1} \mathfrak{f}(\cdot, \vartheta^j) \boldsymbol{1}_{[\tau_{i-1}, \tau_i)},$$

where $\mathfrak{f} \in \mathcal{F}$. The collection of all parametric piecewise continuous functions with k change points generated by $\mathcal{F}$ is denoted by $\boldsymbol{F}_k[a,b]$ (or shortly by $\boldsymbol{F}_k$).

In general we write

$$\theta := (\vartheta^1, \tau_1, .., \vartheta^k, \tau_k, \vartheta^{k+1})$$

and note that θ lies in the compact parameter set

$$\Theta = (\Psi \times [a,b])^k \times \Psi \subset \mathbb{R}^d,$$

with $d = (k+1)r + k$. Thus

$$\mathbf{F}_k = \{f(\cdot, \theta) \mid \theta \in \Theta\},$$

where $f(\cdot, \theta) := \sum_{i=1}^{k+1} \mathfrak{f}(\cdot, \vartheta^i) \boldsymbol{1}_{[\tau_{i-1}, \tau_i)}$, with $\mathfrak{f} \in \mathcal{F}$.

Accordingly we define

$$\mathbf{F}_\infty[a,b] = \bigcup_{k=1}^{\infty} \mathbf{F}_k[a,b].$$

Thus the families $\mathcal{F}_T$ as in (2.1) and $\mathcal{F}_L$ as in (2.2) generate the set of step functions

$$\mathbf{T}_k := \left\{ f(\cdot, \theta) = \sum_{i=1}^{k+1} \mathfrak{f}(\cdot, \vartheta^i) \boldsymbol{1}_{[\tau_{i-1}, \tau_i)} \,\middle|\, \mathfrak{f} \in \mathcal{F}_T \right\}, \tag{2.3}$$

and the set of piecewise linear functions

$$\mathbf{L}_k := \left\{ f(\cdot, \theta) = \sum_{i=1}^{k+1} \mathfrak{f}(\cdot, \vartheta^i) \boldsymbol{1}_{[\tau_{i-1}, \tau_i)} \,\middle|\, \mathfrak{f} \in \mathcal{F}_L \right\}. \tag{2.4}$$

We call the set $\mathcal{J}(f) := \{\tau_i \mid i \in \{1, ..., k\}$ such that $\vartheta^i \neq \vartheta^{i+1}\}$ *change points* of the function $f \in \mathbf{F}_k$ and denote its cardinality by $\sharp \mathcal{J}(f)$. Hence, pc-functions are continuous on the intervals $[\tau_{l-1}, \tau_l)$, $l = 1, .., k+1$ with continuous continuations on the closed intervals $[\tau_{l-1}, \tau_l]$. We denote the left-side limit in a change point τ_l by

$$f(\tau_l^-) = \lim_{x \nearrow \tau_l} f(x).$$

Moreover, we call a change point *jump*, if $f(\tau_i^-, \theta) \neq f(\tau_i, \theta)$, and *kink*, if $f(\tau_i^-, \theta) = f(\tau_i, \theta)$ and say *kink function* (or *jump function*), if f has kinks (or jumps) in all change points.

Note, that for a function $f \in \mathbf{F}_k$ with less than k change points there possibly are more than one parameter vectors generating the same function. That means, for a certain function $f(\cdot, \theta_0) \in \mathbf{F}_k$ the implication $f(\cdot, \theta) = f(\cdot, \theta_0) \Rightarrow \theta = \theta_0$ is true, if and only if $\sharp \mathcal{J}(f) = k$ (cf. Definition 2.2.1, *ii*). If uniqueness of the parameter vector is required, this implies, that we have to confine ourselves to functions from $\mathbf{F}_k$ with precisely k change points. For example, consider the subset $\tilde{\mathbf{T}}_k \subset \mathbf{T}_k$ (cf. (2.3)) of piecewise constant functions, with precisely k jumps, i.e.

$$\tilde{\mathbf{T}}_k := \{f \in \mathbf{T}_k \mid |f(\tau_i^-, \theta) - f(\tau_i, \theta)| > 0\} \tag{2.5}$$

and the subset $\tilde{\mathbf{L}}_k \subset \mathbf{L}_k$ (cf. (2.4)) of piecewise linear functions with precisely k kinks, i.e.

$$\tilde{\mathbf{L}}_k := \{f \in \mathbf{L}_k \mid \vartheta_1^i = \vartheta_1^{i-1} - (\vartheta_2^{i-1} - \vartheta_2^i)\tau_{i-1}, \text{ and } \vartheta_2^{i-1} \neq \vartheta_2^i, \ i = 2, ..., k+1\}. \tag{2.6}$$

As in the case of kinks there may occur dependencies among the parameter components, such that actually the number of parameters, which determine $f(y, \theta)$ is smaller than the dimension of θ. Therefore we define a so called *reduced parameter vector*.

Definition 2.2.3. *Let $\mathcal{F}$ be a family of continuous parametric functions as in Definition 2.2.1 and let $\Theta \subset \mathbb{R}^d$ denote the parameter domain of the set $\boldsymbol{F}_k$ of pc-functions generated by $\mathcal{F}$ (cf. Definition 2.2.2). For a subset $\tilde{\boldsymbol{F}}_k \subset \boldsymbol{F}_k$ a parameter set $\tilde{\Theta} \subset \mathbb{R}^{\tilde{d}}$, with $\tilde{d} < d$, is called a* **reduced parameter domain** *of $\tilde{\boldsymbol{F}}_k$, if there exists an injective and continuously differentiable function $h : \tilde{\Theta} \rightarrow \Theta$, such that*

$$\tilde{\boldsymbol{F}}_k = \{f(\cdot, h(\tilde{\theta})) \in \boldsymbol{F}_k \mid \tilde{\theta} \in \tilde{\Theta}\}.$$

For a function $f(\cdot, \theta_0) \in \tilde{\boldsymbol{F}}_k$ with $f(\cdot, \theta_0) = f(\cdot, h(\tilde{\theta}_0))$ for some $\tilde{\theta}_0 \in \tilde{\Theta}$ we call $\tilde{\theta}_0$ the **reduced parameter vector** *of θ_0.*

Note that, if we consider a class of pc-functions $\mathbf{F}_k$ as in Definition 2.2.2, which is generated by a parametric class $\mathcal{F}$ as in Definition 2.2.1, such that for all $\mathfrak{f} \in \mathcal{F}$ it additionally holds, that $(y, \vartheta) \mapsto \mathfrak{f}(y, \vartheta)$ is continuously differentiable, then the condition $f(\tau_i^-, \theta) = f(\tau_i, \theta)$ often implies local existence of a function h as in Definition 2.2.3 by the implicit function theorem. More precisely, if $f(y, \theta_0)$ is a kink function in such a space, the function

$$
\begin{aligned}
F : \Theta &\longrightarrow \mathbb{R}^k \\
\theta &\longmapsto F(\theta) := \big(\mathfrak{f}(\tau_1, \vartheta^1) - \mathfrak{f}(\tau_1, \vartheta^2), ..., \mathfrak{f}(\tau_k, \vartheta^k) - \mathfrak{f}(\tau_k, \vartheta^{k+1})\big)^t
\end{aligned}
$$

vanishes in θ_0. Due to the differentiability of the map $\theta \mapsto F(\theta)$, the implicit function theorem implies, that there exists a function h and a reduced parameter domain $\tilde{\Theta}$ as in Definition 2.2.3, with $\tilde{\Theta} \subset (\Theta_l)_{l \in I} \subset \mathbb{R}^{d-k}$, where $I \subset \{1, ..., d\}$, if the Jacobian $\partial/(\partial\theta_l)_{l \notin I} F(\theta_0)$ is invertible.

This, for example holds for the set $\tilde{\mathbf{L}}_1$ (cf. (2.6) and Example 2.2). There we have $\vartheta_1^2 = \vartheta_1^1 + (\vartheta_2^1 - \vartheta_2^2)\tau_1$ and choosing the reduced parameter vector $\tilde{\theta} = (\vartheta_1^1, \vartheta_2^1, \tau_1, \vartheta_2^2)$ and the function $h(\tilde{\theta}) = (\vartheta_1^1, \vartheta_2^1, \tau_1, \vartheta_1^1 + (\vartheta_2^1 - \vartheta_2^2)\tau_1, \vartheta_2^2)$ satisfies the conditions of Definition 2.2.3.

2.3 Integral operator

Now we want to introduce some further notation and specifications concerning the integral operator Φ in (1.3).

If Φ acts on the set $\mathbf{F}_k \subset L_2([a, b])$, it can be considered as a map acting on the parameter space Θ, by

$$\theta \longmapsto \Phi f(\cdot, \theta) := \int_a^b \varphi(\cdot, y) f(y, \theta) \mathrm{d}y. \tag{2.7}$$

The special classes of integral operators Φ in (1.3), which we are going to discuss in this thesis, have to satisfy the following Assumptions.

Assumption C. (*Assumptions on the integral operator*) The integral operator Φ in (1.3) satisfies the following conditions.

   i) the operator $\Phi : L_2([a, b]) \longrightarrow L_2([a, b])$ is injective,

   ii) the kernel $\varphi : [a, b]^2 \longrightarrow \mathbb{R}$ is piecewise continuous with finite number of jumps and

   iii) the function $\Phi f(\cdot) : [a, b] \longrightarrow \mathbb{R}$ is Lipschitz continuous with uniform Lipschitz constant $c\|f\|_\infty$, such that the constant c only depends on the kernel φ.

Conditions *i)* and *ii)* are essential for the consistency proof for the estimator of f_0 in the following chapters. Condition *iii)* especially will be needed to estimate the L_2-norm of Φf by means of the empirical norm. In Chapter 3 we introduce some special classes of operators satisfying Assumption **C**.

Moreover, we want to mention, that the results of this paper can also be formulated for an operator $\Phi : L_2([a, b]) \to L_2(I)$, with $I \subset \mathbb{R}$ which does not need to coincides with the interval $[a, b]$, but for ease of notation we only discuss the case, where $I = [a, b]$.

<div align="right">

Chapter 3

</div>

The integral operator

This chapter is concerned with the specification of the considered operator classes. We will introduce assumptions on the Niemitzky operator (1.4), which allow to transfer the results for the linear regression model in (1.5) to the nonlinear model in (1.1). Moreover, we will discuss conditions for injectivity of the linear operator in (1.3). Subsequently we will introduce special conditions, which assure that Assumption **C** is satisfied, for two classes of kernels, namely product and convolution kernels.

3.1 Hammerstein integral equations

As mentioned in the introduction, our aim is to estimate f_0 from observations Y as in (1.1). We approach this by solving the linear inverse problem in (1.5), which results in an estimator for $\mathcal{L}f_0$ (here $\mathcal{L}$ denotes the Niemitzky operator in (1.4)). This estimator is then used to reconstruct f_0. The main concern of the upcoming chapters is the linear model in (1.5). Therefore, we shall now consider suitable conditions on $\mathcal{L}$, such that a stable reconstruction of f_0 from an estimator for $\mathcal{L}f_0$ is possible.

Assumption D. For $\mathcal{L}$ as in (1.4) it holds, that

1.) the operator

$$
\begin{aligned}
\mathcal{L} : L_2([a,b]) &\longrightarrow L_2([a,b]) \\
f &\longmapsto \mathcal{L}f(y) := L(f(y), y).
\end{aligned}
$$

is injective,

2.) for every $x \in [-R, R]$, with a constant $R \geq \sup_{f \in \mathbf{F}_k} \|f\|_\infty$ (given by Lemma 8.2.4), the mapping

$$
\begin{aligned}
L(x, \cdot) : [a, b] &\longrightarrow \mathbb{R} \\
y &\longmapsto L(x, y)
\end{aligned}
$$

is continuous and

3.) for every $y \in [a, b]$, the mapping

$$
\begin{aligned}
L(\cdot, y) : [-R, R] &\longrightarrow \mathbb{R} \\
x &\longmapsto L(x, y)
\end{aligned}
$$

is continuously differentiable, with derivative $L_x(x, y) = \frac{\partial}{\partial x} L(x, y)$, that is continuous in $y \in [a, b]$. Furthermore, there exists a constant l, such that for all $x \in [-R, R]$ and almost every $y \in [a, b]$

$$
|L_x(x, y)| \le l.
$$

The introductory example from rheology (cf. Chapter 5) constitutes a special example, where the operator $\mathcal{L}$ satisfies Assumption **D**. In general any composition $\mathcal{L}f(y) = g_1(y)g_2(f(y))$ of functions $g_1 \in C([a, b])$ and $g_2 \in C^1([-R, R])$, such that $|g_1(y)| > 0$ for all $y \in [a, b]$ and g_2 is injective, meets these conditions, too.

For an operator satisfying Assumption **D** we obtain the following

Lemma 3.1.1. *Let $\mathcal{L}$ be an operator satisfying Assumption **D** and let $\boldsymbol{F}_k$ be a set of pc-functions as in Definition 2.2.2. Then, the image set $\tilde{\boldsymbol{F}}_k := \mathcal{L}(\boldsymbol{F}_k)$ satisfies the conditions of Definition 2.2.2.*

Proof. It is straightforward to verify that for a function set $\mathcal{F}$ as in Definition 2.2.1 the set $\mathcal{L}(\mathcal{F})$ again satisfies *i)* and *ii)* in Definition 2.2.1, whenever $\mathcal{L}$ satisfies Assumption **D**. In order to show, that also *iii)* holds, we apply the chain rule, and obtain for all $\mathfrak{f} \in \mathcal{F}$ and $j = 1, ..., r$, $y \in [a, b]$, that

$$
\left| \frac{\partial}{\partial \vartheta_j} L(\mathfrak{f}(y, \vartheta), y) \right| = \left| L_x(\mathfrak{f}(y, \vartheta), y) \frac{\partial}{\partial \vartheta_j} \mathfrak{f}(y, \vartheta)) \right| \le lg(y),
$$

which proves the claim. This means, that for a set $\mathbf{F}_k$ satisfying the conditions of Definition 2.2.2, the corresponding image set $\mathcal{L}(\mathbf{F}_k) =: \tilde{\mathbf{F}}_k$ is a set of pc-functions satisfying the conditions of Definition 2.2.2 as well, with $g(y)$ replaced by $lg(y)$. $\qquad\square$

Consequently, we can transfer the results for the model in (1.5) in Chapter 4, to obtain an estimator $\tilde{f}(y, \hat{\theta}_n)$ of a function $\tilde{f}_0(y) = \tilde{f}(y, \theta_0) = L(f(y, \theta_0), y)$, with $f_0 = f(\cdot, \theta_0) \in \mathbf{F}_k$. Since furthermore, $f(y, \theta_0)$ and $\tilde{f}(y, \theta_0)$ are generated by the same parameter θ_0, this yields an estimator $f(y, \hat{\theta}_n)$ for $f(y, \theta_0)$ simultaneously. So, considered as parametric problem there is no difference between both observation models. Or more precisely, the classes $\mathbf{F}_k$ in Definition 2.2.2 are chosen, such that the model in (1.1) can be formulated in term of the model in (1.5).

We want to emphasize, that it is a necessary condition for a consistent parameter estimate, that the parameter θ_0 is unique (cf. Lemma 4.2.8), that is, for all $\theta \in \Theta$, $f(y, \theta) = f(y, \theta_0)$ implies $\theta = \theta_0$. Due to condition 1.) in Assumption **D**, this is equivalent to $\tilde{f}(y, \theta_0) = \tilde{f}(y, \theta)$ implies $\theta_0 = \theta$.

However, if this "injectivity" condition does not hold, which happens if $f \in \mathbf{F}_k$ has less than k change points, we may obtain a consistent estimator of $\tilde{f}(y, \theta_0)$, whose parameter vector does not converge. So, intending to estimate f_0, a "parametric" argumentation as above fails. In this case, we can apply the following result.

Lemma 3.1.2. *Suppose that $\mathcal{L}$ is an operator satisfying Assumption $\mathbf{D}$. Then, the map*

$$\mathcal{L}|_{\boldsymbol{F}_k} : \boldsymbol{F}_k \longrightarrow \tilde{\boldsymbol{F}}_k$$
$$f \longmapsto \mathcal{L}f,$$

is continuously invertible.

Proof. By condition 1.) the map $\mathcal{L}|_{\mathbf{F}_k}$ is injective. In order to show its continuity, consider a sequence of functions $\{f_n\}_{n \in \mathbb{N}} \subset \mathbf{F}_k$, with $\|f_n - f_0\|_{L_2([a,b])} \to 0$, for $n \to \infty$. By condition 3.) in Assumption $\mathbf{D}$ and the mean value theorem, it holds that

$$|L(f_n(y), y) - L(f_0(y), y)| \leq |(f_n(y) - f_0(y))| \, l$$

for any $y \in [a, b]$, with l as in condition 3.). Hence, we have

$$\|\mathcal{L}f_n - \mathcal{L}f_0\|_{L_2([a,b])} \leq l\|f_n - f_0\|_{L_2([a,b])} \longrightarrow 0, \text{ for } n \to \infty,$$

which means, $\mathcal{L}|_{\mathbf{F}_k} : (\mathbf{F}_k, \| \cdot \|_{L_2([a,b])}) \longrightarrow (\mathcal{L}(\mathbf{F}_k), \| \cdot \|_{L_2([a,b])})$ is continuous. We will see in Subsection 4.2.1, Lemma 4.2.2, that the set $\mathbf{F}_k$ is totally bounded. Since it also contains functions with less than k jumps, it is additionally closed and hence compact. Altogether, this proves that $\mathcal{L}|_{\mathbf{F}_k} : \mathbf{F}_k \to \tilde{\mathbf{F}}_k$ is a homeomorphism, i.e. it is continuously invertible (see [33, Thm. 3.3, Chapter 16]). $\qquad\square$

Hence reconstruction of f_0 from $\mathcal{L}f_0$ under this conditions constitutes a well-posed problem, which finally allows for estimation of f_0 from (1.1) by solving (1.5) and subsequently inverting $\mathcal{L}$.

In Chapter 6, we also consider the case where the number of change points is not known, intending to estimate this number correctly. Then we have to consider the set $\tilde{\mathbf{F}}_\infty = \mathcal{L}(\mathbf{F}_\infty)$. This set, in general, is not compact and thus, we cannot argue as above, that $\mathcal{L} : \mathbf{F}_\infty \to \tilde{\mathbf{F}}_\infty$ is well-posed. But since we are only interested in estimating the number of change points, which is invariant under transformation by $\mathcal{L}$, i.e. $f_0 \in \mathbf{F}_\infty$ and $\mathcal{L}\tilde{f}_0 \in \tilde{\mathbf{F}}_\infty$ always have the same number of change points, considering the model in (1.5) or the model in (1.1) makes again no difference.

Finally, we note, that in general we can *not* transfer the results for the model in (1.5) to the model in (1.1), if f_0 is a function in an approximation space $\mathcal{A}^\alpha$, as in Chapter 7. The reason is again that compactness of the domain of $\mathcal{L}$, in this case $\mathcal{A}^\alpha$, is violated. So the theory in Chapter 7 would yield an estimator for $\mathcal{L}f_0$, but the reconstruction of f_0 itself in turn, leads to a possibly ill-posed problem. So the results of Chapter 7 cannot be transfered to Hammerstein integral equations, unless the operator $\mathcal{L}$ satisfies additional assumptions. For example, we could use the following modification of condition 3.) in Assumption $\mathbf{D}$.

3*.) For every $y \in [a, b]$, the mapping

$$
\begin{aligned}
L(\cdot, y) : \mathbb{R} &\longrightarrow \mathbb{R} \\
x &\longmapsto L(x, y)
\end{aligned}
$$

is continuously differentiable, with derivative $L_x(x, y) = \frac{\partial}{\partial x} L(x, y)$, that is continuous in $y \in [a, b]$. Furthermore, there exist constants $l_u, l_l > 0$, such that for all $x \in [-R, R]$ and almost every $y \in [a, b]$, it holds that

$$
l_l \le |L_x(x, y)| \le l_u.
$$

If this condition is satisfied, the inverse function theorem implies, that there exists a continuously differentiable inverse $L^{-1}(\cdot, y)$, of the map $x \mapsto L(x, y)$, which is defined on the image of $L(\cdot, y) : [-R, R] \mapsto \mathbb{R}$ pointwise for all $y \in [a, b]$. The derivative of L^{-1} can be calculated as

$$
\frac{\partial}{\partial z} L^{-1}(z, y) = L_z^{-1}(z, y) = (L_x(L^{-1}(z, y), y))^{-1} \le l_l^{-1}.
$$

Now, in order to show continuity of the inverse operator $\mathcal{L}^{-1}$, we again consider a converging sequence $f_n \in L_2([a, b])$ with $\|f_n - f_0\|_{L_2([a,b])} \to 0$ for $n \to \infty$ with $\|f_n\|_\infty \le R$ for all $n \in \mathbb{N}$. Then it holds by the mean value theorem, that

$$
\|\mathcal{L}^{-1} f_n - \mathcal{L}^{-1} f_0\|_{L_2([a,b])} = \|L^{-1}(f_n(y), y) - L^{-1}(f_0(y), y)\|_{L_2([a,b])} \le l_l^{-1} \|f_n - f_0\|_{L_2([a,b])}.
$$

Hence condition 3*.) in addition to Assumption **D**, yields well posedness of the inverse problem generated by the Niemitzky operator $\mathcal{L}$ also in the case, where f_0 is contained in an approximation space $\mathcal{A}^\alpha$.

3.2 Injectivity results for integral operators

Reconstruction of the true function f_0 from the observation model in (1.5) requires injectivity of the integral operator Φ (cf. (1.3)) as claimed in Assumption **C**. Since by Lemma 8.2.1, pc-functions are L_2-identifiable, we just need to claim injectivity of the operator in L_2. The following theorems give some conditions on the kernel φ, that assure L_2 injectivity of the corresponding linear integral operator Φ.

Product kernels We start with the discussion of integral operators with product kernels $\varphi(x, y) = \phi(xy)$ with expansion $\phi(xy) = \sum_{i=1}^\infty \alpha_i(xy)^i$. The next theorem establishes a connection between injectivity of an integral operator and the expansion of its kernel. The main argument in the proof is given by the *Full Müntz Theorem* for L_2-spaces, proven by Borwein et al. in [7]:

Lemma 3.2.1. *(Full Müntz-Theorem) Suppose that $I \subset \mathbb{N}$ and that $0 < a < b$. Then, $\mathrm{span}(\{y^i\}_I)$, is dense in $L_2([a,b])$ if and only if*

$$\sum_{i \in I} i^{-1} = \infty.$$

Proof. This follows directly from [7, Thm. 4.2.6]. □

Theorem 3.2.2. *Assume that $0 < a < b$ and that $\varphi(x,y) = \phi(xy)$ is a product kernel defined on $[a,b]^2$. Assume further, that there exists an interval $[\rho_1, \rho_2] \subset [a^2, b^2]$, with $\frac{\rho_1}{a} < \frac{\rho_2}{b}$, such that ϕ has an absolutely converging expansion $\phi(z) = \sum_{j=1}^{\infty} \alpha_j z^j$ for all $z \in [\rho_1, \rho_2]$. If $J := \{j \in \mathbb{N} : \alpha_j \neq 0\}$, then, the operator in (1.3) is injective if*

$$\sum_{j \in J} j^{-1} = \infty. \tag{3.1}$$

If $a^2 = \rho_1$ and $b^2 = \rho_2$, then the converse is true, i.e. (3.1) is a necessary condition.

Proof. By assumption, the expansion of ϕ converges absolutely on the compact interval $[\rho_1, \rho_2]$. Hence it converges uniformly, and integration and summation can be interchanged, which leads to

$$\Phi f(x) = \int_a^b \varphi(x,y) f(y) \mathrm{d}y = \sum_{j=1}^{\infty} x^j \underbrace{\int_a^b \alpha_j y^j f(y) \mathrm{d}y}_{=:c_j},$$

for all $x \in [\frac{\rho_1}{a}, \frac{\rho_2}{b}]$. In order to see that the right hand side of this equation converges absolutely and uniformly, note that

$$\sum_{j=1}^{\infty} \left| x^j \int_a^b \alpha_j y^j f(y) \mathrm{d}y \right| \leq \left| \int_a^b f(y) \mathrm{d}y \right| \sum_{j=1}^{\infty} |\alpha_j| \max(|a|^j, |b|^j) |x|^j.$$

The integral $\left| \int_a^b f(y) \mathrm{d}y \right|$ is finite, because f is an L_2-function (cf. (1.3)). Furthermore, for $x \in [\frac{\rho_1}{a}, \frac{\rho_2}{b}]$, bx as well as ax are contained in $[\rho_1, \rho_2]$. This implies absolute convergence of the sequence $\sum_{j=1}^{\infty} |\alpha_j| \max(|a|^j, |b|^j) |x|^j$ by assumption and we can apply Lemma 8.2.7 and obtain

$$\Phi f(x) = 0 \text{ for all } x \in \left[\tfrac{\rho_1}{a}, \tfrac{\rho_2}{b} \right] \;\Leftrightarrow\; \sum_{j}^{\infty} x^j c_j = 0 \text{ for all } x \in \left[\tfrac{\rho_1}{a}, \tfrac{\rho_2}{b} \right]$$

$$\Leftrightarrow\; c_j = 0 \text{ for all } j \quad \text{(Lemma 8.2.7)}$$

$$\Leftrightarrow\; \int_a^b \alpha_j y^j f(y) \mathrm{d}y = 0 \text{ for all } j \in J. \tag{3.2}$$

Now we can apply the *Full Müntz-Theorem* [7] (see Lemma 3.2.1), which states that the set $\{y^j\}_J$ is dense in $L_2([a,b])$ if and only if $\sum_{j\in J} j^{-1} = \infty$. With help of Gram-Schmidt-orthonormalisation we obtain an orthonormal system of polynomials p_j of degree j, such that we obtain from (3.2) and Lemma 3.2.1

1. $\Phi f(x) \equiv 0 \Leftrightarrow \int_a^b p_j(y)f(y)\mathrm{d}y = 0 \ \forall j \in J$ and

2. $\mathrm{span}(\{p_j\}_{j\in J})$ is dense in $L_2([a,b]) \Leftrightarrow \sum_{j\in J} j^{-1} = \infty$.

Finally, it follows from [35, Thm. 1.28] that $\int_a^b p_j f(y)\mathrm{d}y = 0$ for all $j \in J$ implies $f \equiv 0$ if and only if $\{p_j\}_J$ is dense in $L_2([a,b])$. Together with 1. and 2., this proves the first claim. In order to prove the second claim, observe that $\rho_1 = a^2$ and $\rho_2 = b^2$ implies that (3.2) is valid for all $x \in [a,b]$ and thus the claim follows from the argumentation above, since all implications hold in both directions. □

Positive definite symmetric kernels The next theorem is formulated within the framework of *native Hilbert spaces* $\mathcal{N}_\varphi$, associated with a kernel φ. A short summary on native Hilbert spaces can be found in Section 8.3. For more detailed information we refer to [54].

Theorem 3.2.3. *The integral operator Φ as defined in (1.3), with continuous positive definite symmetric kernel φ (cf. Definition 8.3.3), is injective on a subset $\mathcal{F} \subset L_2([a,b])$ if and only if $\mathcal{N}_\varphi([a,b])$ is dense in $\mathcal{F}$*

Proof. Assume that $\Phi v \equiv 0$ for $v \in \mathcal{F}$. Using Theorem 8.3.4, this is equivalent to

$$0 = (f, \Phi v)_{\mathcal{N}_\varphi([a,b])} = (f, v)_{L_2([a,b])} \text{ for all } f \in \mathcal{N}_\varphi([a,b]).$$

This in turn is equivalent to $v \equiv 0$ if and only if $v \in \overline{\mathcal{N}_\varphi([a,b])}$. □

Note, that the kernel in Theorem 3.2.3 is not restricted to specific classes as for instance product or convolution kernels. In this sense, the range of possible applications of Theorem 3.2.3 covers arbitrary operators with "general" kernels as defined in (1.3).

From a practical point of view, this general approach is only applicable, if the corresponding native space is explicitly known. More precisely, we have to know the L_2-closure of the respective native space. This means, it suffices to show that it contains a set of functions, which is dense in $L_2([a,b])$ (with respect to the L_2-norm) as for example the set of step functions or polynomials or functions from $C^\infty([a,b])$ with compact support.

There is a great amount of kernels with known native Hilbert spaces, given by the *reproducing* kernels of known Hilbert spaces. This is a consequence of [54, Thm 11.11], which states, that any Hilbert space with reproducing kernel coincides with the native space of this kernel. Thus, any integral operator with kernel $\varphi(x,y) = \tilde{\varphi}(x,y)g(y)$, is injective, if $g \in L_2([a,b])$ with $g(y) > 0$ for all $y \in [a,b]$ and $\tilde{\varphi}(x,y)$ is the reproducing kernel of a Hilbert space, which is dense in $\mathcal{F}$ or even in L_2, with respect to the L_2-norm.

For example, we consider a class of integral kernels

$$\varphi(x,y) = \begin{cases} \dfrac{\cosh(x-a)\cosh(b-y)}{\sinh(b-a)}g(y) & a \le x \le y \le b \\[2ex] \dfrac{\cosh(y-a)\cosh(b-x)}{\sinh(b-a)}g(y) & a \le y \le x \le b \end{cases}$$

with g as above. The kernel $\varphi(x,y)g^{-1}(y)$ is the reproducing kernel of the Sobolev space $\mathbf{H}^1([a,b])$ defined as in (3.5) (see [5, Ex. 13]), which in turn, by [54, Thm 10.11] is the native space of this kernel. Since $\mathbf{H}^1([a,b])$ is dense in $L_2([a,b])$ it follows from Theorem 3.2.3, that the integral operator with kernel $\varphi(x,y)g^{-1}$ is injective on $g(y)\mathcal{F}$. Thus, since g is a known positive function, the integral operator with kernel $\varphi(x,y)$ is injective on $\mathcal{F}$.

Finally, Lemma 8.3.5 gives an example, how the native Hilbert space of a positive definite and symmetric convolution kernel can be characterized by means of the Fourier transform of the kernel, which will be applied in Lemma 3.3.2) to deduce injectivity of special integral operators in the following section.

Convolution kernels For integral operators with convolution kernel $\varphi(x,y)$, such that there exists an analytic function $\phi \in L_2(\mathbb{R})$ with $\varphi(x,y) = \phi(x-y)$ for $(x,y) \in [a,b]^2$, as for instance the Gaussian kernel $\phi(z) = (2\pi\sigma^2)^{-1/2}e^{-(z/\sigma)^2/2}$ for some $\sigma > 0$, we have the following injectivity result.

Theorem 3.2.4. *Let Φ be an integral operator as in (1.3), with convolution kernel $\varphi(x,y) = \phi(x-y)$. Assume further, that ϕ is analytic on $\mathbb{R}$ and denote its Fourier transform by $\hat{\phi}$. Then, the operator $\Phi : L_2([a,b]) \to L_2([a,b])$ is injective, if $\hat{\phi}$ vanishes at most on a set with Lebesgue measure zero.*

Proof. For the operator

$$\begin{aligned} \Phi_R : L_2(\mathbb{R}) &\longrightarrow L_2(\mathbb{R}) \\ f &\longmapsto \int_{-\infty}^{\infty}\phi(\cdot-y)f(y)\mathrm{d}y \end{aligned}$$

it holds that $\Phi_R f \equiv 0$ if and only if $\widehat{\Phi_R f} \equiv 0$. Application of [20, Thm. 3.9, Chap. V, §3] yields $\widehat{\Phi_R f} = \hat{\phi}\hat{f}$. For $f \not\equiv 0$, it holds that $\hat{\phi}\hat{f} \not\equiv 0$ if $\hat{\phi}$ vanishes at most on a set with Lebesgue measure zero.

Now consider a function $f \in L_2(\mathbb{R})$ with $\mathrm{supp}(f) \subset [a,b]$. Then $\Phi_R f$ is an analytic function by Lemma 8.2.8, since ϕ is analytic and $[a,b]$ is compact. Hence, $\Phi f = \Phi_R f \mid_{[a,b]} \equiv 0$ implies that $\Phi_R f = 0$ on $\mathbb{R}$ by the identity theorem. Thus, the claim follows from the first part. $\qquad\square$

3.3 Integral operators satisfying Assumption C

In this section we introduce two special classes of kernels, which will be shown to satisfy Assumption **C**. Therefore we introduce the set of functions of *bounded Variation* $BV([a,b])$, i.e.

$$BV([a,b]) := \left\{ f : [a,b] \to \mathbb{R} \mid \sup_{P \in \mathcal{P}} \sum_{i=1}^{n_P - 1} |f(x_{i+1}) - f(x_i)| < \infty \right\},$$

where $\mathcal{P} = \{P = \{x_1, ..., x_{n_P}\} \mid P \text{ is a partition of } [a,b]\}$ is the set of all finite partitions of $[a,b]$.

Again we consider product and convolution kernels.

Assumption C1: (*Product kernels*)

i) It holds that $0 < a < b$ and there exists a piecewise continuous bounded function $\phi \in BV([a^2, b^2])$ with finite number of jumps, such that $\varphi(x,y) = \phi(xy)$.

ii) Furthermore, there exists an interval $[\rho_1, \rho_2] \subset [a^2, b^2]$, with $\frac{\rho_1}{a} < \frac{\rho_2}{b}$, such that ϕ has an absolutely converging expansion

$$\phi(z) = \sum_{j=0}^{\infty} \alpha_j z^j \quad \text{with } \alpha_j \in \mathbb{R} \quad \text{for all } j \in \mathbb{N}, \ z \in [\rho_1, \rho_2]. \tag{3.3}$$

The set $J := \{j \in \mathbb{N} : \alpha_j \neq 0\}$ satisfies the *Müntz*-condition

$$\sum_{j \in J} j^{-1} = \infty.$$

One example of such a kernel occurs in the example from rheology, given in the introduction, which will be discussed in detail in Section 5.2. The Gaussian kernel $\phi(x) = (2\pi\sigma^2)^{-1/2} e^{-(x/\sigma)^2/2}$, mentioned above, is another well known example for a kernel satisfying **C1**.

Assumption C2: (*Convolution kernels*)

i) $\varphi \in L_2(\mathbb{R}^2) \cap C(\mathbb{R}^2)$ is symmetric and positive definite and there exists a function $\phi \in L_2(\mathbb{R})$, such that $\varphi(x,y) = \phi(x-y)$ and $\phi\mid_{[a-b,b-a]} \in BV([a-b, b-a])$.

ii) There exist constants $0 < c_1 \leq c_2$ and $s > 1/2$, such that the Fourier transform $\hat{\phi}(x)$ decays algebraically, i.e.

$$c_1(1 + |x|^2)^{-s} \leq |\hat{\phi}(x)| \leq c_2(1 + |x|^2)^{-s}, \qquad x \in \mathbb{R}.$$

Note that in Assumption **C2**, condition *ii)* implies *i)*, if $s \geq 1$. This is due to the fact, that a kernel ϕ satisfying *ii)* is contained in the Sobolev space

$$H^s(\mathbb{R}) := \left\{ u \in L_2(\mathbb{R}) \,\middle|\, \int (1 + |x|^2)^s |\hat{u}(x)|^2 \mathrm{d}x < \infty \right\} \qquad (3.4)$$

by Lemma 8.3.5 and thus $\phi \mid_{[a-b,b-a]}$ is contained in

$$H^s([a-b,b-a]) := \{ u \in L_2([a,b]) \mid \exists \bar{u} \in H^s(\mathbb{R}) \text{ with } u = \bar{u} \mid_{[a-b,b-a]} \}. \qquad (3.5)$$

For $s \geq 1$, it holds that $H^s([a-b,b-a]) \subset BV([a-b,b-a])$ (see e.g [22, Chap. 5.1, Ex. 1]) and hence $\phi \in BV([a-b,b-a])$.

In [9] you can find some examples of kernels, which satisfy Assumption **C2**, as e.g. the Laplace kernel $\phi(x) = \frac{1}{2} e^{-|x|}$ or kernels of the type $\phi(x) = (1 - |x|)_+^p$ for $p = 2, 3, \ldots$, where x_+ denotes the positive part of x.

In order to show, that integral operators of type **C1** or **C2** satisfy Assumption **C** we need the following Lemma.

Lemma 3.3.1. *Let $f \in L_\infty([a,b])$ and Φ be an integral operator with kernel $\varphi \in L_\infty([a,b]^2)$, that satisfies one of the following conditions:*

i) $\varphi(x,y) = \phi(xy)$ *is a product kernel with $0 < a < b$ and $\phi \in BV([a^2, b^2])$, or*

ii) $\varphi(x,y) = \phi(x-y)$ *is a convolution kernel, with $\phi \in BV([a-b,b-a])$.*

Then, the map $x \longmapsto \Phi f(x)$ is Lipschitz continuous on $[a,b]$ with uniform Lipschitz constant $c\|f\|_\infty$, such that c only depends on the kernel φ.

Proof. In the following let $x + \delta \in [a,b]$. Under condition *i)* and *ii)* the function ϕ is of bounded variation on $[a^2, b^2]$ and $[a-b,b-a]$ respectively. That means, there exist monotonically increasing and bounded functions ϕ_1, ϕ_2, such that $\phi = \phi_1 - \phi_2$ (cf. [31, Thm. 91.7]) and $\varphi = \varphi_1 - \varphi_2$ respectively. So in both cases we obtain

$$
\begin{aligned}
|\Phi f(x) - \Phi f(x+\delta)| &= \left| \int_a^b (\varphi_1(x,y) - \varphi_1(x+\delta,y) - \varphi_2(x,y) + \varphi_2(x+\delta,y)) f(y) \mathrm{d}y \right| \\
&\leq \|f\|_\infty \int_a^b |\varphi_1(x,y) - \varphi_1(x+\delta,y) - \varphi_2(x,y) + \varphi_2(x+\delta,y)| \, \mathrm{d}y \\
&\leq \|f\|_\infty \left[\int_a^b |\varphi_1(x,y) - \varphi_1(x+\delta,y)| \, \mathrm{d}y \right. \\
&\qquad\qquad \left. + \int_a^b |\varphi_2(x,y) - \varphi_2(x+\delta,y)| \, \mathrm{d}y \right].
\end{aligned}
$$

Since ϕ_1 and ϕ_1 are monotonically increasing functions, we can interchange integration and norm, which, for $i = 1, 2$ and φ satisfying $ii)$, gives

$$
\int_a^b |\phi_i(x-y) - \phi_i(x+\delta-y)| \, \mathrm{d}y = \left| \int_{x-b}^{x-a} \phi_i(u)\mathrm{d}u - \int_{x+\delta-b}^{x+\delta-a} \phi_i(u)\mathrm{d}u \right|
$$
$$
\leq 2|\delta| \|\phi_i\|_\infty.
$$

So we have

$$
|\Phi f(x) - \Phi f(x+\delta)| \leq 2\|f\|_\infty |\delta| (\|\phi_1\|_\infty + \|\phi_2\|_\infty).
$$

If φ is as in $i)$ we have $0 < a < b$ and obtain

$$
\begin{aligned}
\left| \int_a^b \phi_i(xy) - \phi_i((x+\delta)y)\mathrm{d}y \right| &= \left| \int_{ax}^{bx} \frac{1}{x}\phi_i(u)\mathrm{d}u - \int_{a(x+\delta)}^{b(x+\delta)} \frac{1}{x+\delta}\phi_i(u)\mathrm{d}u \right| \\
&\leq \left| \int_{ax}^{bx} \phi_i(u) \left(\frac{1}{x} - \frac{1}{x+\delta} \right) \mathrm{d}u \right| \\
&\quad + \frac{1}{|a|} \left(\left| \int_{ax}^{a(x+\delta)} \phi_i(u)\mathrm{d}u \right| + \left| \int_{b(x+\delta)}^{bx} \phi_i(u)\mathrm{d}u \right| \right) \\
&\leq \|\phi_i\|_\infty \frac{(b-a)}{|x+\delta|}|\delta| + 2\frac{b}{a}\|\phi_i\|_\infty |\delta| \\
&\leq \left(\frac{(b-a)}{a} + 2\frac{b}{a} \right) \|\phi_i\|_\infty |\delta| \\
&\leq 3\frac{b}{a}\|\phi_i\|_\infty |\delta|.
\end{aligned}
$$

Using monotony of ϕ_1 and ϕ_2, again we can interchange norm and integral to get

$$
\begin{aligned}
\left| \int_a^b (\phi_i(xy) - \phi_i((x+\delta)y)) f(y)\mathrm{d}y \right| &\leq \|f\|_\infty \int_a^b |\phi_i(xy) - \phi_i((x+\delta)y)| \, \mathrm{d}y \\
&\leq \|f\|_\infty \left| \int_a^b \phi_i(xy) - \phi_i((x+\delta)y)\mathrm{d}y \right| \\
&\leq 3\frac{b}{a}\|\phi_i\|_\infty \|f\|_\infty |\delta|
\end{aligned}
$$

for $i = 1, 2$. This finally yields

$$
|\Phi f(x) - \Phi f(x+\delta)| \leq 3\frac{b}{a}\|f\|_\infty |\delta| (\|\phi_1\|_\infty + \|\phi_2\|_\infty)
$$

$\square$

Now we are ready to prove the following

Corollary 3.3.2. *Assume that the integral operator* Φ *in* (1.3)

   *a) satisfies Assumption **C1** or*

   *b) satisfies Assumption **C2** or*

   c) has an analytic convolution kernel as in Theorem 3.2.4,

then Φ *satisfies Assumption **C**.*

Proof. From Theorem 3.2.2 and 3.2.4 we directly obtain injectivity for conditions *a)* and *c)*. Lemma 8.3.5 together with Remark 8.3.1 claims, that for a kernel fulfilling Assumption **C2** the corresponding native space $\mathcal{N}_\varphi([a, b])$ coincides with the Sobolev space $H^s([a, b])$ (cf. (3.5)). For this class of kernels Theorem 3.2.3 immediately yields injectivity for condition *b)*. So Assumption, *i)* is satisfied in all three cases.

   Condition *ii)* of Assumption **C** follows directly from Assumption **C1** and **C2** and under condition *c)*, it follows from the fact that the kernel is analytic.

   Finally, application of Lemma 3.3.1 yields uniform Lipschitz continuity, i.e. claim *iii)* of Assumption **C**, for all three cases as well, where we took into account, that analytic functions are contained in $BV([a - b, b - a])$. $\qquad\square$

Known number of change points

In this chapter we consider the case, where the true function f_0 in (1.5) is a parametric piecewise continuous function as in Definition 2.2.2 with known number of change points, i.e. $f_0 \in \mathbf{F}_k$, with $k \in \mathbb{N}$ known.

4.1 Estimate and asymptotic results

In order to reconstruct f_0 from (1.5), we use a least squares estimator, i.e. a function in $\mathbf{F}_k$, whose image under Φ minimizes the empirical distance to the observations. This estimator will be defined in the following Subsection 4.1.1. Subsequently in Subsection 4.1.2 we present some results concerning the asymptotic behavior of this estimator, namely consistency, rates of convergence and asymptotic distribution of the estimated parameters. The proofs of these results are added in Section 4.2.

4.1.1 Estimate

We begin with the definition of the least squares estimator $\hat{f}_n$, such that $\Phi\hat{f}_n$ minimizes the empirical distance to the observations Y in (1.5) with respect to the space $\mathbf{F}_k$. That is, $\hat{f}_n \in \mathbf{F}_k$ and

$$\|\Phi\hat{f}_n - Y\|_n^2 \leq \min_{f \in \mathbf{F}_k} \|\Phi f - Y\|_n^2 + o(n^{-1}). \tag{4.1}$$

From Definition 2.2.2, it follows, that there exists a parameter vector $\hat{\theta}_n \in \Theta$, such that

$$\hat{f}_n(y) = f(y, \hat{\theta}_n) = \sum_{i=1}^{k+1} \mathfrak{f}(y, \hat{\vartheta}^i) \mathbf{1}_{(\hat{\tau}_{i-1}, \hat{\tau}_i)}.$$

Note, that the parameters $\hat{\vartheta}^i$ and $\hat{\tau}_i$ also depend on the index n.

The minimum on the right hand side in (4.1) always exists (see Corollary 4.2.4), but it does not need to be unique. Note, that we do not assume, that the minimum is attained,

but only that the functional above can be minimized up to a term of order $o(n^{-1})$. This allows for numerical approximation of the minimizer and gives an intuition of the needed precision for the asymptotic results to be valid. Furthermore it assures existence of a minimum, even if we minimize over the non-compact set $\mathbf{F}_\infty$, as needed in the case of unknown number of change points in Chapter 6 or for estimation of a function from an approximation space in Chapter 7.

In contrast to these methods, however, compactness of $\mathbf{F}_k$ implies that we do not need any further regularization to stabilize the inverse problem. To be more precise, the restriction of possible estimators to a compact set, actually *is* the regularization, as on a compact set the operator Φ is continuously invertible (see Lemma 4.2.7). This also explains, why the expansion to $\mathbf{F}_\infty$ in Chapter 6 and 7 requires additional penalizing (cf. (6.1) and (7.4)).

4.1.2 Consistency and asymptotic results

Now we investigate the asymptotic behavior of the least squares estimator in (4.1) under the assumption, that the true function $f_0 \in \mathbf{F}_k$ has precisely k change points, that is $\sharp \mathcal{J}(f_0) = k$. Theorem 4.1.1 presents the main asymptotic results, which will be extended in Corollary 4.1.2, for the case where f_0 is known to depend on a reduced parameter as in Definition 2.2.3. Subsequently, in Corollary 4.1.4, we show, that improved results can be obtained, if f_0 is a kink function.

Before approaching the main results we require some more notation. Due to Lemma 8.2.4, *iii)* the function $\theta \mapsto \Phi f(x, \theta)$ as defined in (2.7), is differentiable for almost every $x \in [a, b]$ with gradient

$$D_f(x, \theta) := \frac{\partial}{\partial \theta} \Phi f(x, \theta). \tag{4.2}$$

With this, we define the $d \times d$ matrix V_{f_θ} by

$$(V_{f_\theta})_{i,j} = \int_a^b (D_f(x, \theta))_i (D_f(x, \theta))_j s(x) \mathrm{d}x, \qquad 1 \le i, j \le d, \tag{4.3}$$

where s is as in Assumption **B**.

Theorem 4.1.1. *Suppose that Assumptions $\boldsymbol{A1}$ and $\boldsymbol{B}$ are satisfied and that the integral operator Φ fulfilles Assumption $\boldsymbol{C}$. Let $\hat{f}_n(y) = f(y, \hat{\theta}_n)$ be the least squares estimator of the true function $f_0 = f(\cdot, \theta_0) \in \boldsymbol{F}_k$ as in (4.1), with $\sharp \mathcal{J}(f_0) = k$. If the matrix $V_{f_{\theta_0}}$ is nonsingular, then*

(i) $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N(0, \sigma^2 V_{f_{\theta_0}}^{-1})$,

(ii) $|\theta_0 - \hat{\theta}_n|_2 = O_P(n^{-\frac{1}{2}})$,

(iii) $\|\Phi f_0 - \Phi \hat{f}_n\|_{L_2([a,b])} = O_P(n^{-\frac{1}{2}})$ *and,*

(iv) $\|f_0 - \hat{f}_n\|_{L_2([a,b])} = O_P(n^{-\frac{1}{4}})$.

If f_0 depends on a reduced parameter vector $\tilde{\theta}_0$ as in Definition 2.2.3, the gradient $D_f(x, \tilde{\theta}_0) = \frac{\partial}{\partial \theta} \Phi f(x, h(\tilde{\theta}_0))$ as in (4.2) can be calculated by the chain rule, due to the differentiability of the function h (cf. Definition 2.2.3) and we have the following corollary.

Corollary 4.1.2. *Suppose that the assumptions of Theorem 4.1.1 are satisfied and that the true function $f_0(y) = f(y, h(\tilde{\theta}_0))$ is known to depend on a reduced parameter vector $\tilde{\theta}_0$ as in Definition 2.2.3. Then the results of Theorem 4.1.1 are valid with θ_0 and $\hat{\theta}_n$ substituted by the reduced parameter vectors $\tilde{\theta}_0$ and $\tilde{\theta}_n$.*

Nonsingularity of the covariance matrix $V_{f_{\theta_0}}$ is essential for Theorem 4.1.1, to hold. The next lemma gives a characterization of this property in terms of the partial derivatives $\frac{\partial}{\partial \vartheta^i} f(y, \vartheta_0^i)$, $i = 1, ..., k+1$, for the case, where $f(\cdot, \theta_0)$ has precisely k jumps and Φ satisfies one of the Assumptions **C1** and **C2**. Moreover, it states that $V_{f_{\theta_0}}$ is always singular, if $f(y, \theta_0)$ has a kink in some change point.

Proposition 4.1.3. *Suppose that Assumption **C1** or **C2** is satisfied and $f(\cdot, \theta_0) \in \boldsymbol{F}_k$, with $f(\cdot, \theta_0) = \sum_{i=1}^{k+1} f(\cdot, \vartheta_0^i) \boldsymbol{1}_{[\tau_{0,i-1}, \tau_{0,i})}$ (cf. Definition 2.2.2), has k change points. Then, the matrix $V_{f_{\theta_0}}$ as defined in (4.3) is nonsingular, if and only if $f(\cdot, \theta_0)$ has jumps in all change points and for every $a \in \mathbb{R}^r$ and $i = 1, ..., k+1$,*

$$a^t \frac{\partial}{\partial \vartheta} f(\cdot, \vartheta_0^i) \boldsymbol{1}_{[\tau_{0,i-1}, \tau_{0,i})} \equiv 0 \quad implies \quad a = 0. \tag{4.4}$$

Hence, if the true function f_0 is known to be a kink function, it follows from Proposition 4.1.3, that Theorem 4.1.1 cannot be applied, since the condition of nonsingularity of $V_{f_{\theta_0}}$ is violated. So this case requires restriction to a reduced parameter domain $\tilde{\Theta}$ as in Definition 2.2.3. Then, we obtain the following improved asymptotic results, which depend on the *modulus of continuity* of the considered function class $\mathcal{F}$ (cf. Definition 2.2.1) defined as

$$\nu(\mathcal{F}, \delta) := \sup_{f \in \mathcal{F}} \sup_{|y_1 - y_2| \le \delta} |f(y_1) - f(y_2)|. \tag{4.5}$$

Corollary 4.1.4. *Assume that the conditions of Corollary 4.1.2 are satisfied, but the true function $f_0(y) = f(y, \theta(\tilde{\theta}_0))$ is a kink function and let ν be defined as in (4.5). Then the results of Corollary 4.1.2 are valid, with the improved rate*

$$\|f_0 - \hat{f}_n\|_{L_2([a,b])} = O_P(n^{-\frac{1}{2}} + n^{-\frac{1}{4}} \nu(\mathcal{F}, n^{-\frac{1}{2}})). \tag{4.6}$$

In the discussion of the introductory example from rheology in Subsection 5.2, for instance, we obtain rates of order $n^{-1/2}$ in the case, where $f_0 \in \tilde{L}_2$ (cf. (2.6)) is a piecewise linear function with two kinks, i.e. the case, where $\mathcal{F} = \mathcal{F}_L$ is the set of linear functions with bounded slopes as in (2.2). More general, we obtain rates of $n^{-(1+\alpha)/4}$ if $\mathcal{F}$ is a set of Hölder continuous functions with exponent $0 < \alpha \le 1$.

Remark 4.1.1. (Application to the observation model in (1.1))

As discussed in Section 3.1, Theorem 4.1.1 as well as Corollaries 4.1.2 and 4.1.4 also hold for the model in (1.1), if $\mathcal{L}$ in (1.4) satisfies Assumption **D**. This is formulated in the next

Corollary 4.1.5. *Under the conditions of Theorem 4.1.1, let $\mathcal{L}$ be as in (1.4) satisfy Assumption **D** and $\mathcal{L}\hat{f}_n = \mathcal{L}f(y, \hat{\theta}_n)$ be the least squares estimator of the function $\mathcal{L}f_0 = \mathcal{L}f(y, \theta_0)$ as in (4.1), with $f_0 \in \boldsymbol{F}_k$ and $\sharp\mathcal{J}(f_0) = k$. Furthermore let $H = \Phi \circ \mathcal{L}$ and define $D_{\mathcal{L}f}(x, \theta_0)$ and $V_{\mathcal{L}f_{\theta_0}}$ as in (4.2) and (4.3). If $V_{\mathcal{L}f_{\theta_0}}$ is nonsingular, then*

(i) $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N(0, \sigma^2 V_{\mathcal{L}f_{\theta_0}}^{-1})$,

(ii) $|\theta_0 - \hat{\theta}_n|_2 = O_P(n^{-\frac{1}{2}})$,

(iii) $\|Hf_0 - H\hat{f}_n\|_{L_2([a,b])} = O_P(n^{-\frac{1}{2}})$ *and,*

(iv) $\|f_0 - \hat{f}_n\|_{L_2([a,b])} = O_P(n^{-\frac{1}{4}})$.

Moreover, if f_0 depends on a reduced parameter $\tilde{\theta}_0$ as in Definition 2.2.3, the same holds with θ_0 and $\hat{\theta}_n$ substituted by the reduced parameter vectors $\tilde{\theta}_0$ and $\tilde{\theta}_n$. In particular, if f_0 is a kink function depending on a reduced parameter $\tilde{\theta}_0$, we obtain

(iv)* $\|f_0 - \hat{f}_n\|_{L_2([a,b])} = O_P(n^{-\frac{1}{2}} + n^{-\frac{1}{4}}\nu(\mathcal{F}, n^{-\frac{1}{2}}))$,

with ν as in (4.5), if $V_{\mathcal{L}f_{\tilde{\theta}_0}}$ is nonsingular.

Remark 4.1.2. (Misspecification of the model and spline-regularization)

We will briefly discuss what happens, if the true function f_0 is not, as assumed an element of $\mathbf{F}_k$. In general, under the additional assumption of equidistant design in Assumption **B**, it follows from Lemma 8.2.9, that the minimizer of (4.1) converges to a pc-function $t \in \mathbf{F}_k$ s.t.

$$\|\Phi f_0 - \Phi t\|_{L_2([a,b])} \leq \min_{g \in \mathbf{F}_k} \|\Phi f_0 - \Phi g\|_{L_2([a,b])}. \tag{4.7}$$

By Corollary 4.2.4, the minimum at the right hand side always exists. Lemma 8.2.9 furthermore requires, that it is unique. To check condition (4.7), leads to well known problems of nonlinear approximation theory (e.g. see [11]). Note that, due to injectivity of the operator Φ, the assumption of uniqueness of the minimizer t is equivalent to the assumption of uniqueness $\Phi t \in \Phi(\mathbf{F}_k)$, which minimizes the L_2-distance from Φf_0. It would be of great interest, to relax the assumption of equidistant design as well as improving the asymptotic law (8.5) for $\hat{f}_n$ in Lemma 8.2.9. However, this is beyond the scope of this thesis and not treated here.

If $\mathbf{F}_k$ is the set of piecewise polynomial functions, this offers an interesting connection to distributional asymptotics for splines. Since by the *Curry and Schoenberg Theorem*

(cf. [15, Chap. VIII, (44)]), for fixed change points, we have that the set of piecewise polynomials of degree p is equal to the B-spline space of order p with knots in $\{\tau_0, ..., \tau_{k+1}\}$ with multiplicity p, in the case of jumps and $p-1$ in the case of kinks. Thus, in this case, misspecification of the model, could be considered as spline approximation of f_0 and thus leads to the well known "spline-regularization". Although, here we actually have to deal with *free-knot* splines, i.e. the knots are free parameters and not known in advance, which causes a serious difference to B-splines, starting with the fact that those spaces are not longer linear. It has long been known, that approximation of a function by splines improves dramatically if the knots are free ([46], [16],[12]). However, free knot spline approximation has not been as popular as might be expected by these results. This is due to the so called *lethargy* property (see e.g. [34]) of free not splines, which affects the stable and effective computation of optimal knots. Especially it means, that intending to optimize the locations of free knots, one is usually faced with local optima located on multiple or coalescent knots which correspond to a degenerate case.

Nevertheless, for example Jupp in [34] describes a computational strategy for the minimization problem for approximation with free knot splines. In [1], one can find results about the asymptotic behavior of integrated mean square error (IMSE), by using least squares and BIAS minimizing splines. In this context they determine an optimal density for the knot distribution for a fixed knot number, which minimizes the IMSE. Results similar to the asymptotic law as in Theorem 4.1.1 for the respective estimators do not exist. However, free-knot spline approximation is closely connected to *approximation* spaces, which will be discussed in Chapter 7. For a general discussion on free knot spline spaces as well as some remarks on their approximation spaces, we refer to [15, Chap. XII, §4].

Concerning spline-regularization with fixed knots, in contrast, one can find some interesting results on inverse problems which treat integral equations as in (1.5) as e.g. in [14], [41], [28] or [25]. For detailed informations about splines in general we again refer to de Boor [15].

4.2 Proofs

The proof of Theorem 4.1.1 and the corresponding Corollaries 4.1.2, 4.1.4 and 4.1.5 is separated in three parts. We start by estimating the entropy number of the considered function spaces in Subsection 4.2.1, which yields the basic arguments for the consistency proof of the estimator in (4.1) (Subsection 4.2.2). In Subsection 4.2.3 we give the proof of the asymptotic results in Theorem 4.1.1 and Corollaries 4.1.2, 4.1.4 and 4.1.5. Finally we add the proof of Proposition 4.1.3 separately in Subsection 4.2.4.

The proof of Corollary 4.1.2, under dependency assumptions of the components of θ (cf. Definition 2.2.3), will be given as a draft only. This is due to the fact, that it proceeds analogously with the required derivatives obtained by application of the chain rule. Thus an exact formulation would be more confusing than explicatory. However,

some details may be clarified by the discussion of the application of Corollary 4.1.2 and 4.1.4 to a special example in Section 5.2, where we provide an explicit calculation of the respective derivatives and the resulting covariance matrix $V_{f_{\hat{\theta}}}$.

4.2.1 Entropy results

In order to show consistency of the least squares estimator $\hat{f}_n$ in Equation (4.1), we apply uniform deviation inequalities from empirical process theory. To this end, it is necessary to calculate the *entropy* of the space of interest.

Definition 4.2.1. *Given a measure P, a set of P-measurable functions $\mathcal{G}$ and a real number $\delta > 0$, the δ-**covering number** $\mathbf{N}(\delta, \mathcal{G}, P)$ is defined as the smallest value of N such that there are functions $g_1, ..., g_N$ with*

$$\min_{1 \leq j \leq N} \left(\int (g - g_j)^2 \mathrm{d}P \right)^{\frac{1}{2}} \leq \delta, \quad \text{for all } g \in \mathcal{G}.$$

*Moreover, the δ-**entropy H** and the **entropy integral J** of $\mathcal{G}$ are defined as*

$$\mathbf{H}(\delta, \mathcal{G}, P) = \log \mathbf{N}(\delta, \mathcal{G}, P) \quad \text{and}$$

$$\mathbf{J}(\delta, \mathcal{G}, P) := \max \left(\delta, \int_0^\delta \mathbf{H}^{1/2}(u, \mathcal{G}, P) \mathrm{d}u \right),$$

respectively.

If P is the Lebesgue measure we will write $\mathbf{H}(\delta, \mathcal{G})$ and $\mathbf{N}(\delta, \mathcal{G})$ instead of $\mathbf{H}(\delta, \mathcal{G}, P)$ and $\mathbf{N}(\delta, \mathcal{G}, P)$.

The entropy will be of specific use, with respect to two aspects.

First, in terms of *empirical process theory*, the fact, that the considered function class has finite entropy, as we will show, means that this class is *Donsker*, which loosely speaking names classes of functions for which uniform central limit theorems can be obtained (for details see e.g. [53, Chap. 19]). In this framework, calling $\mathbf{F}_k$ Donsker names the basic property of this class, which finally yields consistency and asymptotic normality of the estimator in (4.1) (see Theorem 4.1.1). In [53] it has already been shown, that parametric classes as $\mathcal{F}$ in Definition 2.2.2 are Donsker, if the respective parameter set is of finite dimension and bounded. But the entropy number calculated there, is not as exact, as it is needed later on in the situation, where we do not know the number of change points (cf. Chapter 6). Therefore, although at this point only finiteness of the entropy is required, we will calculate shaper bounds, in order to resume it afterwards in Chapter 6.

Second, from an *inverse problem* point of view, the knowledge of the entropy of $\mathbf{F}_k$ will help to show its compactness, which in turn is needed to prove continuity of the inverse of the operator (1.3) as well as (1.4). Thus, the entropy number also is a usefull tool to deduce stability of an inverse problem.

Now we start with the estimation of the entropy of the set $\mathbf{F}_k$.

Lemma 4.2.2. *Let $\boldsymbol{F}_k$ and $d = (k+1)r + k$ be as in Definition 2.2.2. Than there exists a constant $T_{\mathcal{F}} > 0$ depending only on the considered function class $\mathcal{F}$ in Definition 2.2.1, such that*

$$\boldsymbol{H}(\delta, \boldsymbol{F}_k) \leq d \log\left(\frac{\sqrt{d}T_{\mathcal{F}} + \delta}{\delta}\right).$$

Proof. For $\theta_1, \theta_2 \in \Theta$, with $\Delta = \theta_1 - \theta_2$ we can use the differentiability stated by Definition 2.2.1, to obtain the first order expansion

$$\mathfrak{f}(x, \vartheta_1^i) - \mathfrak{f}(x, \vartheta_2^i) = (\vartheta_1^i - \vartheta_2^i)^t \frac{\partial \mathfrak{f}}{\partial \vartheta^i}(x, \tilde{\vartheta}^i)$$

for $\tilde{\vartheta}^i \in (\vartheta_1^i, \vartheta_2^i)$. This yields

$$
\begin{aligned}
\|f(\cdot, \theta_1) - f(\cdot, \theta_2)\|_{L_2([a,b])}^2 &\leq \sum_{i=1}^{k+1} \|\mathfrak{f}(\cdot, \vartheta_1^i) - \mathfrak{f}(\cdot, \vartheta_2^i)\|_{L_2([a,b])}^2 \\
&\leq \sup_{\substack{i=1,\ldots,k+1 \\ j=1,\ldots,r}} \left\|\frac{\partial}{\partial \vartheta_j^i} \mathfrak{f}(\cdot, \tilde{\vartheta}^i)\right\|_{L_2([a,b])}^2 \sum_{i=1}^{k+1} |\vartheta_1^i - \vartheta_2^i|_2^2 \\
&\leq |\Delta|_2^2 R^2 \\
&\leq d|\Delta|_\infty^2 R^2,
\end{aligned}
$$

with R as in Lemma 8.2.4. Assuming w.l.o.g., that M in Definition 2.2.2 satisfies $2M \geq b - a$, it follows that

$$\sup_{\theta_1, \theta_2 \in \Theta} |\theta_1 - \theta_2|_\infty \leq 2M,$$

for all $\theta \in \Theta$, independent of the number of change points. Together with the preceding equation, this implies that the the δ-covering number of $\boldsymbol{F}_k$ depends only on the number of balls with radius $\delta(R\sqrt{d})^{-1}$ which are needed to cover a subset of $\mathbb{R}^d$, with diameter bounded by $2M$, which can be estimated by $(4MR\sqrt{d} + \delta)^d/\delta^d$ (cf. [17, Lem. 2.5]). So finally we get

$$\mathbf{H}(\delta, \mathbf{F}_k) \leq d \log\left(\frac{4MR\sqrt{d} + \delta}{\delta}\right),$$

and the claim follows with $T_{\mathcal{F}} = 4MR$. $\qquad\square$

The function class $\mathbf{F}_k$ generates the set

$$\mathbf{G}_k := \{\Phi f \mid \theta \in \Theta \text{ and } f \in \mathbf{F}_k[a,b]\}, \tag{4.8}$$

where Φ is a known integral operator with kernel φ as defined in (1.3). So, we will use the entropy of $\mathbf{F}_k$ to estimate the entropy of $\mathbf{G}_k$ in the following Lemma.

Lemma 4.2.3. *Let $\boldsymbol{G}_k$ be defined as in (4.8). Then there exists a constant $T > 0$ independent of k such that $T \geq \|g\|_{L_2([a,b])}$ for all $g \in \boldsymbol{G}_k$ and for all $\delta > 0$*

$$\boldsymbol{H}(\delta, \boldsymbol{G}_k, P_n) \leq d \log\left(\frac{\sqrt{d}T + \delta}{\delta}\right), \qquad (4.9)$$

with $d = r(k+1) + k$.

Proof. For all $\Delta = \theta_1 - \theta_2$ with $\theta_1, \theta_2 \in \Theta$ we obtain by Hölder's inequality, that

$$
\begin{aligned}
\|\Phi f(\cdot, \theta_1) - \Phi f(\cdot, \theta_2)\|_n^2 &= \frac{1}{n} \sum_{i=1}^n \left(\int_a^b \varphi(x_i, y)(f(y, \theta_1) - f(y, \theta_2)) \mathrm{d}y\right)^2 \\
&\leq \frac{1}{n} \sum_{i=1}^n \|\varphi(x_i, \cdot)\|_{L_2([a,b])}^2 \|f(\cdot, \theta_1) - f(\cdot, \theta_2)\|_{L_2([a,b])}^2 \\
&\leq (b-a)\|\varphi\|_\infty^2 \|f(\cdot, \theta_1) - f(\cdot, \theta_2)\|_{L_2([a,b])}^2.
\end{aligned}
$$

Using Lemma 4.2.2, consequently implies

$$
\begin{aligned}
\mathbf{H}(\delta, \mathbf{G}_k, P_n) &\leq \mathbf{H}\left(\frac{\delta}{\sqrt{b-a}\|\varphi\|_\infty}, \mathbf{F}_k\right) \\
&\leq d \log\left(\frac{\sqrt{d(b-a)}T_\mathcal{F}\|\varphi\|_\infty + \delta}{\delta}\right).
\end{aligned}
$$

Remembering that by definition $T_\mathcal{F} \geq (b-a)R \geq (b-a)\sup_{f \in \mathbf{F}_k}\|f\|_\infty$ (cf. Lemma 8.2.4, i)), choosing $T = \|\varphi\|_\infty \max(\sqrt{b-a}, b-a)T_\mathcal{F} \geq (b-a)\|\varphi\|_\infty \sup_{f \in \mathbf{F}_k}\|f\|_\infty \geq \|g\|_{L_2([a,b])}$ for all $g \in \mathbf{G}_k$ completes the proof. $\square$

Further, we find that by Lemma 4.2.3, the set $(\mathbf{G}_k, \|\cdot\|_n)$ is totally bounded. Since $\mathbf{G}_k$ also contains functions Φf with $f \in \mathbf{F}_k$, such that $\sharp\mathcal{J}(f) < k$, it is additionally closed and hence compact, which directly yields

Corollary 4.2.4. *Let Y be as in (1.5) and $\boldsymbol{G}_k$ as in (4.8). Then, the functional*

$$
\begin{aligned}
\|\cdot\|_n^2 : \quad \boldsymbol{G}_k &\longrightarrow \mathbb{R}^+ \\
\Phi f &\longmapsto \|\Phi f - Y\|_n^2
\end{aligned}
$$

has at least one minimizer in $\boldsymbol{G}_k$.

4.2.2 Consistency

As claimed above, the entropy results of Subsection 4.2.1 can be used now to show consistency of the least squares estimator in (4.1). First, we show consistency of the image $\Phi\hat{f}_n$.

Theorem 4.2.5. *Let Φ be an operator satisfying Assumption $\boldsymbol{C}$ and $f_0 = f(\cdot, \theta_0) \in F_k$. Furthermore, assume that Assumption $\boldsymbol{A1}$ and $\boldsymbol{B}$ are met. Then, for $\hat{f}_n = f(\cdot, \hat{\theta}_n)$, the least squares estimator in (4.1), it holds that*

$$\|\Phi\hat{f}_n - \Phi f_0\|_n = o_P(1).$$

Proof. Due to Inequality (4.1) we have

$$\|\Phi\hat{f}_n - Y\|_n^2 \leq \|\Phi f_0 - Y\|_n^2 + o(n^{-1}).$$

Inserting $Y = \Phi f_0 + \varepsilon$ leads to

$$\|\Phi\hat{f}_n - \Phi f_0\|_n^2 - 2\langle\Phi\hat{f}_n - \Phi f_0, \varepsilon\rangle_n + \|\varepsilon\|_n^2 \leq \|\varepsilon\|_n^2 + o(n^{-1})$$

which is equivalent to

$$
\begin{aligned}
\|\Phi\hat{f}_n - \Phi f_0\|_n^2 &\leq 2\langle\Phi\hat{f}_n - \Phi f_0, \varepsilon\rangle_n + o(n^{-1}) \\
&= 2(\langle\Phi\hat{f}_n, \varepsilon\rangle_n - \langle\Phi f_0, \varepsilon\rangle_n) + o(n^{-1}) \\
&\leq 4\sup_{g \in \mathbf{G}_k} |\langle g, \varepsilon\rangle_n| + o(n^{-1}).
\end{aligned}
$$

Lemma 4.2.3 gives boundedness of the entropy $\boldsymbol{H}(\delta, \mathbf{G}_k, P_n)$ uniformly in n, for all $\delta > 0$ and so $n^{-1}\boldsymbol{H}(\delta, \mathbf{G}_k, P_n) \to 0$ as $n \to \infty$. With this result it follows from Lemma 8.4.2, that $\sup_{g \in \mathbf{G}_k} |\langle g, \varepsilon\rangle_n| = o_P(1)$. $\qquad\square$

As stated by Assumption $\mathbf{C}$ *iii)* the functions Φf are uniform Lipschitz continuous for all $f \in \mathbf{F}_k$. This directly implies convergence in L_2-norm:

Corollary 4.2.6. *Under the assumptions of Theorem 4.2.5 one has*

$$\|\Phi\hat{f}_n - \Phi f_0\|_{L_2([a,b])} = o_P(1).$$

Proof. Since the design in Definition (1.5) is assumed to satisfy Assumption $\mathbf{B}$ the claim follows directly from Theorem 4.2.5 and Corollary 8.2.6. $\qquad\square$

The next lemma states stability of the inverse problem of reconstructing f_0 from Φf_0, i.e. if $\Phi\hat{f}_n$ is close to Φf_0, than $\hat{f}_n$ is close to f_0, too.

Lemma 4.2.7. *Under the assumptions of Theorem 4.2.5 it holds that*

$$\|\Phi\hat{f}_n - \Phi f_0\|_{L_2([a,b])} = o(1) \quad implies \quad \|f(\cdot, \theta_0) - f(\cdot, \hat{\theta}_n)\|_{L_2([a,b])} = o(1).$$

Proof. The operator $\Phi : (\mathbf{F}_k, \|\cdot\|_{L_2([a,b])}) \longrightarrow (L_2([a,b]), \|\cdot\|_{L_2([a,b])})$ is linear and bounded and hence continuous. According to Assumption $\mathbf{C}$, *i)* it is injective and it follows from Lemma 4.2.2, that the set $(\mathbf{F}_k, \|\cdot\|_{L_2([a,b])})$ is totally bounded. Since it also contains functions with less than k change points, it is additionally closed and therefore compact. Hence $\Phi : \mathbf{F}_k \longrightarrow \{\Phi f \in L_2([a,b]) : f \in \mathbf{F}_k\}$ is a bijective continuous mapping from a compact set to a Hausdorff space, that means it is a homeomorphism (see [33, Thm. 3.3, Chapter 16]). $\qquad\square$

The next lemma states, that convergence of a sequence of functions $f_n = f(\cdot, \theta_n) \in \mathbf{F}_k$ to a function $f_0 = f(\cdot, \theta_0) \in \mathbf{F}_k$ implies convergence of its parameter vector θ_n to θ_0, if f_0 has precisely k change points. This is a necessary condition, since otherwise there may by more than one parameter vectors, which generate the same limit function.

Lemma 4.2.8. *Assume that $f_0 = f(\cdot, \theta_0) \in \mathbf{F}_k$ with $\sharp \mathcal{J}(f_0) = k$ and let $\{f(\cdot, \theta_n)\}_{n \in \mathbb{N}}$ be a sequence in $\mathbf{F}_k$. Then*

$$\|f(\cdot, \theta_0) - f(\cdot, \theta_n)\|_{L_2([a,b])} = o(1) \quad implies \quad |\theta_0 - \theta_n|_\infty = o(1).$$

Proof. Due to the definition of $\mathcal{J}(\cdot)$ in Subsection 2.2, the assumption $\sharp \mathcal{J}(f_0) = k$ implies that $f(\cdot, \theta_0)$ has precisely k change points. That means, $f(\cdot, \theta_0) \equiv f(\cdot, \theta)$ implies $\theta = \theta_0$, i.e. for all $\theta_0 \neq \theta \in \Theta$ we have $\|f(\cdot, \theta_0) - f(\cdot, \theta)\|_{L_2([a,b])} > 0$. Now assume that $\|f(\cdot, \theta_0) - f(\cdot, \theta_n)\|_{L_2([a,b])} = o(1)$ but that there exist a subsequence $\{\theta_{k_n}\}_{n \in \mathbb{N}}$ and a constant $c_1 > 0$, such that $|\theta_0 - \theta_{k_n}|_\infty > c_1$ for all $n \in \mathbb{N}$. Since Θ is compact, we can choose a further subsequence of this subsequence, which converges to some $\hat{\theta} \in \Theta$. W.l.o.g we assume $\lim_{n \to \infty} |\hat{\theta} - \theta_{k_n}|_\infty = 0$. By construction $|\theta_0 - \hat{\theta}|_\infty > c_1$ and so uniqueness of θ_0 implies $\|f(\cdot, \theta_0) - f(\cdot, \hat{\theta})\|_{L_2([a,b])} > c_2 > 0$ for some constant c_2. Continuity of the map $\theta \mapsto \|f(\cdot, \theta) - f(\cdot, \theta_0)\|_{L_2([a,b])}$, stated by Lemma 8.2.4, implies existence of some $n_0 \in \mathbb{N}$, such that for all $n \geq n_0$ we have

$$\|f(\cdot, \theta_0) - f(\cdot, \theta_{k_n})\|_{L_2([a,b])} > \frac{1}{2} c_2 > 0.$$

This is a contradiction to $\|f(\cdot, \theta_0) - f(\cdot, \theta_n)\|_{L_2([a,b])} = o(1)$ and the claim follows. $\square$

Corollary 4.2.9. *Under the assumptions of Theorem 4.2.5 it holds, that*

$$\|f(\cdot, \theta_0) - f(\cdot, \hat{\theta}_n)\|_{L_2([a,b])} = o_P(1).$$

Moreover, if the true function f_0 has exactly k change points it also holds, that

$$|\theta_0 - \hat{\theta}_n|_\infty = o_P(1).$$

Proof. This follows from Theorem 4.2.5 by application of Lemma 4.2.7 and 4.2.8 together with Lemma 8.1.1. $\square$

4.2.3 Asymptotic normality

In this subsection we investigate the rates of convergence of the least squares estimator $\hat{f}_n(y) = f(y, \hat{\theta}_n)$ in (4.1) and especially show asymptotic normality of the corresponding parameter estimate $\hat{\theta}_n$, which finally leads to the proof Theorem 4.1.1. To this end, we focus on the stochastic process $\|Y - \Phi \hat{f}_n\|_n^2 = n^{-1} \sum_{i=1}^n (y_i - \Phi \hat{f}_n(x_i))^2$ for the random observations (Y, X) as in (1.5), which henceforth we write as the empirical expectation

$$\mathbb{E}_n m(\cdot, \cdot, \theta) := n^{-1} \sum_{i=1}^n m(x_i, y_i, \theta),$$

(cf. Chapter 2), with m defined as

$$m(x, y, \theta) := (y - \Phi f(x, \theta))^2. \tag{4.10}$$

By definition of the least squares estimator, we have, that $\hat{\theta}_n$ is the minimizer of the map

$$\theta \longmapsto \mathbb{E}_n m(\cdot, \cdot, \theta),$$

if y is a random variable satisfying $y = \Phi f_0(x) + \varepsilon_1$ as in Model (1.5), with $\mathbf{E}\varepsilon_1 = 0$ and $\mathbf{E}\varepsilon_1^2 = \sigma^2$ and the expectation of $m(\cdot, \cdot, \theta)$ can be calculated as

$$
\begin{aligned}
\mathbf{E}m(\cdot, \cdot, \theta) &= \mathbf{E}(\Phi f(\cdot, \theta_0) - \Phi f(\cdot, \hat{\theta}_n))^2 + \sigma^2 \\
&= \mathbf{E}(\Phi f(\cdot, \theta_0) - \Phi f(\cdot, \hat{\theta}_n))^2 + \mathbf{E}m(\cdot, \cdot, \theta_0).
\end{aligned} \tag{4.11}
$$

By Lemma 8.2.4, the function $\theta \mapsto m(x, y, \theta)$ is almost everywhere differentiable with derivative $\partial/\partial\theta\, m(x, y, \theta) =: \dot{m}(x, y, \theta) = 2(\Phi f(x, \theta) - y)D_f(x, \theta)$, with $D_f(x, \theta)$ as in (4.2), such that

$$\mathbf{E}\dot{m}(\cdot, \cdot, \theta_0)\dot{m}(\cdot, \cdot, \theta_0)^t = 4\sigma^2 \mathbf{E}D_f(\cdot, \theta_0)D_f(\cdot, \theta_0)^t = \sigma^2 4V_{f_{\theta_0}}. \tag{4.12}$$

In general, for proving asymptotic normality of the parameter estimator, empirical process theory requires, that the function $\theta \mapsto m(x, y, \theta)$ is twice differentiable, in order to obtain a second order expansion of this function. But according to [53, Thm. 5.23], rather a second order expansion of the expectation $\mathbf{E}m(\cdot, \cdot, \theta)$, instead of the function m itself, is needed. In the case at hand, the function m in (4.10) is just once differentiable (a.e.) by Lemma 8.2.4. So we have to use another way to obtain a second order expansion as needed in [53], than by using the second derivative. This will be the main topic of this subsection, which aims at the application of a modified version of the mentioned result in [53]. The only difference between this modified version, i.e. Theorem 4.2.10 and [53, Thm. 5.23] is, that the assumption of twice differentiability of $\theta \mapsto \mathbf{E}m(\cdot, \cdot, \theta)$ is weakened here, following an explanatory note from the author in connection with this theorem. The conditions of Theorem 4.2.10 meet all requirements of the proof, given there, i.e. it follows by the proof of [53, Thm. 5.23] without any change.

Theorem 4.2.10. *For each θ in an open subset of Euclidean space let $(x, y) \mapsto m(x, y, \theta)$ be a measurable function such that $\theta \mapsto m(x, y, \theta)$ is differentiable at θ_0 for $\mathbf{P}$-almost every (x, y) and such that, for every θ_1 and θ_2 in a neighborhood of θ_0 and a measurable function $\dot{m}$ with $\mathbf{E}\dot{m}^2 < \infty$*

$$|m(x, y, \theta_1) - m(x, y, \theta_2)| \le \dot{m}(x, y)|\theta_1 - \theta_2|_\infty. \tag{4.13}$$

Furthermore, assume that the map $\theta \mapsto \mathbf{E}m(\cdot, \cdot, \theta)$ admits an expansion

$$\mathbf{E}m(\cdot, \cdot, \theta) = \mathbf{E}m(\cdot, \cdot, \theta_0) + \frac{1}{2}(\theta - \theta_0)^t V(\theta - \theta_0) + r(|\theta_0 - \theta|_\infty).$$

at a point of minimum θ_0 with nonsingular symmetric matrix V and remainder term r, such that

$$\lim_{|\theta_0 - \theta|_\infty \to 0} \frac{r(|\theta_0 - \theta|)_\infty}{|\theta_0 - \theta|_\infty^2} = 0.$$

If $\mathbb{E}_n m(\cdot, \cdot, \hat{\theta}_n) \leq \inf_\theta \mathbb{E}_n m(\cdot, \cdot, \theta) + o_P(n^{-1})$ and $\hat{\theta}_n \xrightarrow{P} \theta_0$, then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V^{-1} \frac{1}{\sqrt{n}} \sum_{1=1}^{n} \dot{m}(x_i, y_i, \theta_0) + o_P(1).$$

In particular, the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and covariance matrix $V^{-1} \boldsymbol{E} \dot{m}(\cdot, \cdot, \theta_0) \dot{m}(\cdot, \cdot, \theta_0)^t V^{-1}$.

Proof. Along the lines of the proof of [53, Thm. 2.23]. $\qquad \square$

Now, first we want to show the Lipschitz property (4.13) of $\theta \mapsto m(x, y, \theta)$ for m defined as in (4.10).

Lemma 4.2.11. *Let Assumption $\boldsymbol{A1}$ and $\boldsymbol{B}$ be satisfied and Φ be an integral operator with piecewise continuous kernel operating on the set $\boldsymbol{F}_k$. Then, for the function $m(x, y, \theta)$ in (4.10) and for every θ_1 and θ_2 in Θ one has*

$$|m(x, y, \theta_1) - m(x, y, \theta_2)| \leq \dot{m}(y, x)|\theta_1 - \theta_2|_\infty, \tag{4.14}$$

with a measurable function $\dot{m}$ with $\boldsymbol{E} \dot{m}^2 < \infty$.

Proof. By Lemma 8.2.4 the function $\theta \mapsto \Phi f(x, \theta)$ is differentiable for almost every $x \in [a, b]$, with derivative $D_f(x, \theta)$ as in (8.3). From the mean value theorem it follows for some $\tilde{\theta} \in (\theta_1, \theta_2)$, that

$$
\begin{aligned}
|m(x, y, \theta_1) - m(x, y, \theta_2)| &\leq |2(y - \Phi f(x, \tilde{\theta})) D_f(x, \tilde{\theta})(\theta_1 - \theta_2)| \\
&\leq |2(y + C))|Cd|\theta_1 - \theta_2|_\infty,
\end{aligned}
$$

where we took into account, that for the constants C and R in Lemma 8.2.4, *vi*) and *i*) together with (8.3), it holds that

$$C \geq (b - a)\|\varphi\|_\infty R \geq \|\Phi f\|_\infty$$

for all $f \in \boldsymbol{F}_k$.

Defining $\dot{m}(x, y) = \infty$ if x lies in the null set, where $\Phi f(x, \theta)$ is not differentiable and $\dot{m}(x, y) = dC2|y + C|$ else, implies the Lipschitz condition (4.14). Remembering that $y = \Phi f(x, \theta_0) + \varepsilon_1 \leq C + |\varepsilon_1|$, we obtain

$$\dot{m}(x, y) \leq 2d(|\varepsilon_1| + 2C)C,$$

almost everywhere. Since $\boldsymbol{E} \varepsilon_1^2 < \infty$, and hence $\boldsymbol{E}|\varepsilon_1| < \infty$, this finally yields $\boldsymbol{E} \dot{m}^2 \leq \infty$. $\qquad \square$

In the proof of the next lemma, we derive a second order expansion by using differentiability of $\theta \mapsto \mathbf{E}m(\cdot, \cdot, \theta)$.

Lemma 4.2.12. *Assume that the conditions of Lemma 4.2.11 are satisfied. For the least squares estimator $\hat{\theta}_n$ in (4.1) of the true parameter θ_0, define*

$$\Delta_n := (\hat{\theta}_n - \theta_0)$$

and the $d \times d$ matrix $V_{f_\theta} = \boldsymbol{E}D_f(\cdot, \theta)D_f(\cdot, \theta)^t$ (cf. (4.2) and (4.3)), for any $\theta \in \Theta$. Then

$$\boldsymbol{E}m(\cdot, \cdot, \hat{\theta}_n) = \boldsymbol{E}m(\cdot, \cdot, \theta_0) + \Delta_n^t V_{f_{\theta_0}} \Delta_n + h(|\Delta_n|_\infty),$$

with remainder term h, such that

$$\lim_{|\Delta_n|_\infty \to 0} \frac{h(|\Delta_n|_\infty)}{|\Delta_n|_\infty^2} = 0.$$

Proof. As in (4.11), we have

$$\mathbf{E}m(\cdot, \cdot, \hat{\theta}_n) = \sigma^2 + \mathbf{E}(\Phi f(\cdot, \theta_0) - \Phi f(\cdot, \hat{\theta}_n))^2 = \mathbf{E}m(\cdot, \cdot, \theta_0) + \mathbf{E}(\Phi f(\cdot, \theta_0) - \Phi f(\cdot, \hat{\theta}_n))^2.$$

By Lemma 8.2.4, the map $\theta \mapsto \Phi f(x, \theta)$ is almost everywhere differentiable with derivative $D_f(x, \theta)$. Thus, it follows from the mean value theorem, that for some $\tilde{\theta}_n \in (\theta_0, \hat{\theta}_n)$

$$
\begin{aligned}
\mathbf{E}(\Phi f(\cdot, \theta_0) - \Phi f(\cdot, \hat{\theta}_n))^2 &= \mathbf{E}(\Delta_n^t D_f(\cdot, \tilde{\theta}_n))^2 \\
&= \mathbf{E}(\Delta_n^t D_f(\cdot, \theta_0))^2 + |\Delta_n|_\infty^2 O(|\mathbf{E}(D_f(\cdot, \theta_0) - D_f(\cdot, \tilde{\theta}_n))|_2) \\
&\quad + |\Delta_n|_\infty^2 O(|\mathbf{E}(D_f(\cdot, \theta_0) - D_f(\cdot, \tilde{\theta}_n))|_2^2).
\end{aligned}
$$

Continuity of $\theta \mapsto |\mathbf{E}(D_f(\cdot, \theta))|_2$ stated by Lemma 8.2.4, v), together with [53, Lem. 2.12] yields $\lim_{|\Delta_n|_\infty \to 0} |\mathbf{E}(D_f(\cdot, \theta_0) - D_f(\cdot, \tilde{\theta}_n))|_2 = 0$ and the claim follows. $\square$

The preceding lemmata show that the conditions of Theorem 4.2.10 are satisfied. Hence we are ready to proof the asymptotic results presented in Theorem 4.1.1 and Corollaries 4.1.2, 4.1.4 and 4.1.5 in Section 4.1.2.

Theorem 4.1.1

Proof. It follows from definition of $\hat{\theta}_n$ in (4.1) as well as from Lemma 4.2.9, 4.2.11 and 4.2.12, that the conditions of Theorem 4.2.10 are fulfilled. According to this theorem, together with (4.12), the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and covariance matrix

$$(2V_{f_{\theta_0}})^{-1}\mathbf{E}[\dot{m}(\cdot, \cdot, \theta_0)\dot{m}(\cdot, \cdot, \theta_0)^t](2V_{f_{\theta_0}})^{-1} = \sigma^2 V_{f_{\theta_0}}^{-1},$$

which proves (i).

By [53, Cor. 5.53], the Lipschitz condition from Lemma 4.2.11 and the expansion in Lemma 4.2.12 yield (ii).

Again using the gradient $D_f(x, \theta)$ in (4.2) (cf. 8.2.4), for almost every $x \in [a, b]$, we get a first order expansion

$$\Phi f(x, \theta_0 + \Delta_n) = \Phi f(x, \theta_0) + \Delta_n^t D_f(x, \tilde{\theta}), \tag{4.15}$$

with $\tilde{\theta} \in (\theta_0, \hat{\theta}_n)$. Using this expansion and taking into account, that $|\Delta_n|_\infty \leq |\Delta_n|_2$, we get

$$
\begin{aligned}
\|\Phi f(x, \theta_0 + \Delta_n) - \Phi f(x, \theta_0)\|_{L_2([a,b])} &\leq |\Delta_n|_\infty \sup_{\theta \in \Theta} \|D_f(\cdot, \theta)\|_\infty (b - a) \\
&\leq (b - a) C O_P(n^{-\frac{1}{2}}) = O_P(n^{-\frac{1}{2}})
\end{aligned}
$$

where $C \geq \sup_{\theta \in \Theta, i=1,...d} \|(D_f(\cdot, \theta))_i\|_\infty$ as in Lemma 8.2.4. Now (iii) follows from (ii).

For the proof of (iv) recall that the maps $\vartheta^i \mapsto \mathfrak{f}(x, \vartheta^i)$, $i = 1, ..., k+1$ are continuously differentiable (cf. Definition 2.2.1). This, together with Lemma 8.2.4 and skipping the indices 0 and n for the parameter components ϑ^i and τ_i, the mean value theorem yields

$$
\begin{aligned}
\|f_0 - \hat{f}_n\|_{L_2([a,b])}^2 &= \sum_{i=1}^{k+1} \int_{\min(\tau_{i-1}, \hat{\tau}_{i-1})}^{\max(\tau_i, \hat{\tau}_i)} \left[(\vartheta^i - \hat{\vartheta}^i)^t \frac{\partial}{\partial \vartheta^i} \mathfrak{f}(y, \tilde{\vartheta}^i) \right]^2 \mathrm{d}y \\
&\quad + \sum_{i=1}^{k} \left[\int_{\tau_i}^{\hat{\tau}_i} (\mathfrak{f}(y, \vartheta^{i+1}) - \mathfrak{f}(y, \hat{\vartheta}^i))^2 \mathbf{1}_{\tau_i > \hat{\tau}_i} \mathrm{d}y \right. \\
&\quad \left. - \int_{\tau_i}^{\hat{\tau}_i} (\mathfrak{f}(y, \vartheta^i) - \mathfrak{f}(y, \hat{\vartheta}^{i+1}))^2 \mathbf{1}_{\tau_i < \hat{\tau}_i} \mathrm{d}y \right] \\
&\leq |\Delta|_\infty^2 (k+1) r R^2 + \sum_{i=1}^{k} 4 R^2 |\tau_i - \hat{\tau}_i| \\
&= O(|\Delta_n|_\infty) = O_P(n^{-\frac{1}{2}}). \tag{4.16}
\end{aligned}
$$

Where R is defined as in Lemma 8.2.4 and $\tilde{\vartheta}^i \in (\vartheta^i, \hat{\vartheta}^i)$ for $i = 1, ..., k+1$. $\qquad\square$

Corollary 4.1.2

Proof. Due to the differentiability of h in Definition 2.2.3 the needed derivatives in the proof of Theorem 4.1.1 can be calculated by application of the chain rule. Hence Corollary 4.1.2 follows analogously to the proofs in Subsections 4.2.1, 4.2.2 and 4.2.3 by substituting the required derivatives respectively. Note that for the same reason the technical results in the Appendix as in particular Lemma 8.2.4 also apply to $f(y, h(\tilde{\theta}))$. $\qquad\square$

Corollary 4.1.4

Proof. Statements (i) - (iv) from Theorem 4.1.1 are valid for the reduced parameter vectors $\tilde{\theta}_0$ and $\tilde{\theta}_n$ by Corollary 4.1.2. In order to show (4.6), we skip the dependencies of the parameter components, for the sake of simplicity and consider the pieces $\mathfrak{f}(y, \vartheta^i)$ instead of $f(y, \vartheta^i(\tilde{\theta}))$ for all $i = 1, ..., k + 1$, keeping in mind, that for all occurring derivatives we actually need to apply the chain rule.

Now f has a kink in τ_i for all $i = 1, ..., k$. W.l.o.g. we assume, that $\tau_i > \hat{\tau}_i$, then we have

$$\int_{\tau_i}^{\hat{\tau}_i} \left(\mathfrak{f}(y, \vartheta^{i+1}) - \mathfrak{f}(y, \hat{\vartheta}^i) \right)^2 dy \leq \int_{\tau_i}^{\hat{\tau}_i} \left(|\mathfrak{f}(y, \vartheta^{i+1}) - \mathfrak{f}(\tau_i, \vartheta^{i+1})| \right.$$

$$\left. + |\mathfrak{f}(\tau_i, \vartheta^{i+1}) - \mathfrak{f}(\tau_i, \vartheta^i)| + |\mathfrak{f}(\tau_i, \vartheta^i) - \mathfrak{f}(\tau_i, \hat{\vartheta}^i)| + |\mathfrak{f}(\tau_i, \hat{\vartheta}^i) - \mathfrak{f}(y, \hat{\vartheta}^i)| \right)^2 dy.$$

Again using the differentiability of the map $\vartheta^i \mapsto \mathfrak{f}(y, \vartheta^i)$ as in the proof of Theorem 4.1.1 (iv), we obtain $|\mathfrak{f}(\tau_i, \vartheta^i) - \mathfrak{f}(\tau_i, \hat{\vartheta}^i)| = O(|\vartheta^i - \hat{\vartheta}^i|)$. The term $|\mathfrak{f}(\tau_i, \vartheta^{i+1}) - \mathfrak{f}(\tau_i, \vartheta^i)|$ vanishes because there is a kink in τ_i. Finally, remembering the definition of the modulus of continuity ν in (4.5), we get

$$\sup_{y \in [\tau_i, \hat{\tau}_i]} (|\mathfrak{f}(y, \vartheta^{i+1}) - \mathfrak{f}(\tau_i, \vartheta^{i+1})|, |\mathfrak{f}(\tau_i, \hat{\vartheta}^i) - \mathfrak{f}(y, \hat{\vartheta}^i)|) = \nu(\mathcal{F}, |\tau_i - \hat{\tau}_i|)$$

and thus, it follows from (ii), that

$$\int_{\tau_i}^{\hat{\tau}_i} \left(\mathfrak{f}(y, \vartheta^{i+1}) - \mathfrak{f}(y, \hat{\vartheta}^i) \right)^2 dy = O(|\tau_i - \hat{\tau}_i|)(\nu(\mathcal{F}, |\tau_i - \hat{\tau}_i|) + |\vartheta^i - \hat{\vartheta}^i|)^2$$

$$= O_P(n^{-\frac{1}{2}}(\nu(\mathcal{F}, n^{-\frac{1}{2}})^2 + n^{-1})).$$

Since this holds for all $i = 1, ..., k$, together with (4.16), this proves (4.6). $\square$

Corollary 4.1.5

Proof. (of Corollary 4.1.5) As described in Section 3.1, the function $\mathcal{L}f_0$ is contained in a pc-function set $\tilde{\mathbf{F}}_k$ as in Definition 2.2.2 with $\sharp \mathcal{J}(\mathcal{L}f_0) = k$, because $\mathcal{L}$ satisfies Assumption **D**. Hence application of Theorem 4.1.1 implies (i), (ii) and (iii). Then (iv) follows from (ii) analogously to the proof of (iv) in Theorem 4.1.1. The second part of the Corollary follows for the same reasons from Corollaries 4.1.2 and 4.1.4, where $(iv)^*$ again follows from (ii) as in the proof of (4.6) in Corollary 4.1.4. $\square$

4.2.4 Nonsingularity of V_θ

In order to prove Proposition 4.1.3 we need the following

Lemma 4.2.13. *Suppose that $0 \not\equiv g \in L_2([a,b])$ and $a = y_0 < y_1 < ... < y_n < y_{n+1} = b$. Let further $\varphi(x,y)$ be an integral kernel fulfilling Assumption **C1** or **C2**. For $\alpha, \beta_1, .., \beta_n \in \mathbb{R}$ it holds*

$$\alpha \int_a^b \varphi(\cdot, y) g(y) \mathrm{d}y + \sum_{i=1}^n \beta_i \varphi(\cdot, y_i) \equiv 0$$

if and only if $\alpha = 0 = \beta_i$, for all $i = 1, ..., n$.

Proof. We are going to discuss the two kinds of kernels separately.
Assumption C2: Define the set of real *Schwartz functions* as

$$\mathcal{S}(\mathbb{R}) := \{ f \in C^\infty(\mathbb{R}) : \lim_{|x| \to \infty} |x^n f^{(m)}(x)| = 0 \quad \forall n, m = 0, 1, 2, ...\}.$$

By [9, Lem. 5.1], the native space of a kernel satisfying **C2**, contains the subset of all functions from $\mathcal{S}(\mathbb{R})$ with support in $[a, b]$,

$$\{ f \in \mathcal{S}(\mathbb{R}) : \mathrm{supp}(f) \subset [a, b] \} \subset \mathcal{N}_\varphi([a, b]).$$

In particular, this implies, that for any interval $[c, d] \subset [a, b]$ there exists some test function $\psi \in \mathcal{N}_\varphi([a, b])$ satisfying $\mathrm{supp}(\psi) = [c, d]$. Thus, taking into account that $\|g\|_{L_2([a,b])} \neq 0$, we can choose $\rho, \delta_i \in \mathcal{S}(\mathbb{R})$ such that

$$\rho \quad \text{satisfies} \quad : \quad \begin{cases} \int_a^b \rho(y) g(y) \mathrm{d}y \neq 0 & \text{and} \\ \rho(y) = 0 & \text{if } y \in \{y_1, ..., y_n\} \cup \mathbb{R}\backslash[a, b] \end{cases}$$

$$\text{and} \quad \delta_i \quad \text{satisfies} \quad : \quad \begin{cases} \delta_i(y_i) \neq 0 & \text{and} \\ \delta_i(y) = 0 & \text{if } y \in \mathbb{R}\backslash(y_{i-1}, y_{i+1}), \end{cases}$$

for all $i = 1, ..., n$. As elements of the native space, ρ and all δ_i for $i = 1, ..., k$, can be interpreted as function by

$$\rho(y) = \langle \rho, \varphi(\cdot, y) \rangle_{\mathcal{N}_\varphi([a,b])},$$

and

$$\delta_i(y) = \langle \delta_i, \varphi(\cdot, y) \rangle_{\mathcal{N}_\varphi([a,b])},$$

as shown in Lemma 8.3.2. Applying the inner product of the native space $\langle \rho, \cdot \rangle_{\mathcal{N}_\varphi([a,b])}$ to both sides of the equation

$$\alpha \int_a^b \varphi(\cdot, y) g(y) \mathrm{d}y + \sum_{i=1}^n \beta_i \varphi(\cdot, y_i) \equiv 0$$

yields

$$\alpha \int_a^b \langle \rho, \varphi(\cdot, y) \rangle_{\mathcal{N}_\varphi([a,b])} g(y) \mathrm{d}y + \sum_{i=1}^n \beta_i \langle \rho, \varphi(\cdot, y_i) \rangle_{\mathcal{N}_\varphi([a,b])} = 0.$$

Hence

$$\alpha \int_a^b \rho(y) g(y) \mathrm{d}y = 0$$

and therefore $\alpha = 0$, since $\int_a^b \rho(y) g(y) \mathrm{d}y \neq 0$ by definition. This implies

$$\sum_{i=1}^n \beta_i \varphi(\cdot, y_i) \equiv 0$$

applying the inner product $\langle \delta_i, \cdot \rangle_{\mathcal{N}_\varphi([a,b])}$, we obtain

$$\beta_i \delta_i(y_i) = 0 \quad \Rightarrow \quad \beta_i = 0 \quad \forall \quad i = 1, ..., n.$$

Assumption C1: Using the expansion in Assumption **C1** gives

$$0 = \alpha \int_a^b \varphi(x, y) g(y) \mathrm{d}y + \sum_{i=1}^n \beta_i \varphi(x, y_i) = \sum_{j \in J} \alpha_j x^j \left[\alpha \int_a^b y^j g(y) \mathrm{d}y + \sum_{i=1}^n \beta_i y_i^j \right],$$

for all $x \in [\frac{\rho_1}{a}, \frac{\rho_2}{b}]$. By Lemma 8.2.7 this is equivalent to

$$\alpha \int_a^b y^j g(y) \mathrm{d}y + \sum_{i=1}^n \beta_i y_i^j = 0 \quad \forall j \in J. \tag{4.17}$$

By the Müntz Theorem for $C([a, b])$-functions in [8, Thm. 6.2], the condition $\sum_{j \in J} j^{-1} = \infty$ implies, that the set of monomials $\{y^j\}_{j \in J}$ is dense in $C([a, b])$, with respect to the supremums norm. Since $\mathcal{S}(\mathbb{R}) \subset C([a, b])$, the functions ρ and δ_i from the first part of the proof, can also be interpreted as uniform limit of linear combinations $\rho(y) = \sum_{j \in J} \mu_j y^j$ and $\delta_i = \sum_{j \in J} \nu_j^i y^j$, for $i = 1, ..., n$, with appropriate coefficients μ_j and ν_j^i. Hence, together with equation (4.17), we obtain

$$\begin{aligned} 0 &= \sum_{j \in J} \mu_j \alpha \int_a^b y^j g(y) \mathrm{d}y + \sum_{j \in J} \mu_j \sum_{i=1}^n \beta_i y_i^j \\ . &= \alpha \int_a^b \rho(y) g(y) \mathrm{d}y + \sum_{i=1}^n \beta_i \rho(y_i) \\ &= \alpha \int_a^b \rho(y) g(y) \mathrm{d}y \end{aligned}$$

and thus $\alpha = 0$. Consequently, (4.17) yields

$$\sum_{i=1}^{n} \beta_i y_i^j = 0 \quad \forall j \in J \tag{4.18}$$

and hence

$$\sum_{j \in J} \nu_j^l \sum_{i=1}^{n} \beta_i y_i^j = \sum_{i=1}^{n} \beta_i \delta_l(y_i) = 0 \quad \forall l = 1, ..., n,$$

which again implies $\beta_i = 0$ for all $i = 1, ..., 0$. $\qquad\qquad\square$

Proposition 4.1.3

Proof. Assume that $h \in \mathbb{R}^d$. The matrix $V_{f_{\theta_0}}$ in (4.3) is nonsingular, if

$$h^t V_{f_{\theta_0}} h = \int_a^b (h^t D_f(x, \theta_0))^2 s(x) \mathrm{d}x = 0 \quad \Leftrightarrow \quad h = 0.$$

So assume that $h^t V_{f_{\theta_0}} h = 0$. This implies $\|h^t D_f(x, \theta_0)\|_{L_2([a,b])} = 0$, because $s \geq s_l > 0$, by Assumption **B**. It follows from (8.3) together with the assumptions on the kernel φ and Lemma 3.3.1, that the components $x \mapsto (D_f(x, \theta_0))_i$ are at least left- or right-continuous for all $x \in [a, b]$. Thus $h^t D_f(\cdot, \theta_0) \equiv 0$ and equivalently

$$\int_a^b \varphi(\cdot, y) \sum_{i \neq 0 \bmod (r+1)} h_i \frac{\partial}{\partial \theta_i} f(y, \theta_0) \mathrm{d}y$$
$$+ \sum_{i = 0 \bmod (r+1)} h_i \varphi(\cdot, \tau_i)(f(\tau_i^-, \theta_0) - f(\tau_i, \theta_0)) \equiv 0. \tag{4.19}$$

If $f(\cdot, \theta_0)$ has a kink in τ_i for some $i = 0 \bmod (r+1)$, it holds that $f(\tau_i^-, \theta_0) - f(\tau_i, \theta_0) = 0$ and the second term in (4.19) vanishes for all $h \in \mathbb{R}^d$, with $h_j = 0$ for all $j \neq i$ and h_i arbitrary. Thus the matrix $V_{f_{\theta_0}}$ is singular if $f(\cdot, \theta_0)$ has a kink in some change point, which shows the first part of Proposition 4.1.3.

If in contrast, $f(\cdot, \theta_0)$ has exactly k jumps, it holds that $f(\tau_i^-, \theta_0) - f(\tau_i, \theta_0) \neq 0$ for all $i = 0 \bmod (r+1)$ and Lemma 4.2.13 yields $h_i = 0$ for all $i = 0 \bmod r + 1$ and moreover

$$\int_a^b \varphi(\cdot, y) \sum_{i \neq 0 \bmod (r+1)} h_i \frac{\partial}{\partial \theta_i} f(y, \theta_0) \mathrm{d}y \equiv 0.$$

By Corollary 3.3.2, the integral operator associated with the kernel φ is injective and thus the preceding equation yields

$$\sum_{i \neq 0 \bmod (r+1)} h_i \frac{\partial}{\partial \theta_i} f(\cdot, \theta_0) \equiv 0 \qquad \text{a.e.}$$

and thus,

$$\sum_{j=1}^{r} h_{(i-1)(r+1)+j} \frac{\partial}{\partial \vartheta_j^i} \mathfrak{f}(\cdot, \vartheta_0^i) \mathbf{1}_{[\tau_{0,i-1}, \tau_{0,i})} \equiv 0, \tag{4.20}$$

for $i = 1, ..., k+1$, where we took into account, that every function $\mathfrak{f} \in \mathcal{F}$ (cf. Definition 2.2.1) is continuous. That means $h_{(i-1)(r+1)+j} = 0$, for $i = 1, ..., k+1$, $j = 1, ..., r$, if the partial derivatives $\frac{\partial}{\partial \vartheta_j^i} \mathfrak{f}(\cdot, \vartheta_0^i) \mathbf{1}_{[\tau_{0,i-1}, \tau_{0,i})}$, $j = 1, ..., r$ are linear independent for all $i = 1, ..., k+1$, as (4.4) claims. So we have $h = 0$ and thus nonsingularity of $V_{f_{\theta_0}}$, if (4.4) is satisfied. If conversely (4.4) is not satisfied, there exists $h \neq 0$ such that (4.20) holds, which implies that $V_{f_{\theta_0}}$ is singular. Altogether, this finally proves the claim. $\qquad \square$

Chapter 5

Application

5.1 Confidence bands

Statement (i) in Theorem 4.1.1 implies that the quadratic form

$$n\sigma^{-2}(\hat{\theta}_n - \theta_0)V_{f_{\theta_0}}(\hat{\theta}_n - \theta_0)^t$$

is asymptotically distributed according to a χ^2-distribution with d degrees of freedom. This is still true if σ and $V_{f_{\theta_0}}$ are replaced by consistent estimators $\hat{\sigma}_n$ and $V_{f_{\hat{\theta}_n}}$, respectively. Hence we are now able to determine a $(1 - \alpha)$-confidence ellipsoid for $\hat{\theta}_n$ in $\mathbb{R}^d$ by

$$\frac{n}{\hat{\sigma}_n^2}(\hat{\theta}_n - \theta)(V_{f_{\hat{\theta}_n}})(\hat{\theta}_n - \theta)^t \leq \chi_d^2(1 - \alpha). \tag{5.1}$$

Here $\chi_d^2(1 - \alpha)$ denotes the $(1 - \alpha)$-quantile of the χ^2-distribution with d degrees of freedom. By maximizing and minimizing $f(y, \theta)$ over θ inside this confidence ellipsoid, we obtain simultaneous confidence bands for $\hat{f}_n$. As far as the calculation of this confidence bands is concerned, we have to account for the fact that (depending on d) a possibly great amount of parameters has to be dealt with. This complicates the calculation of confidence bands for $\hat{f}_n$. So we choose one of the common methods for approximate confidence sets, such as Bonferroni, Scheffée or studentized maximum modulus statistics. For computation of these statistics and corresponding confidence sets, we refer to the book of Miller [39]. As our simulations show (cf. Figure 5.1), the studentized statistic proves to be the most advantageous in the case of discontinuities. To make this obvious, we calculated confidence bands for the most simple example, namely for the estimator of a step function with one jump, using the approximative statistics just mentioned and the exact χ^2 statistic in comparison. Therefore we generated $n = 100$ noisy observations of the image Φh, of the step function

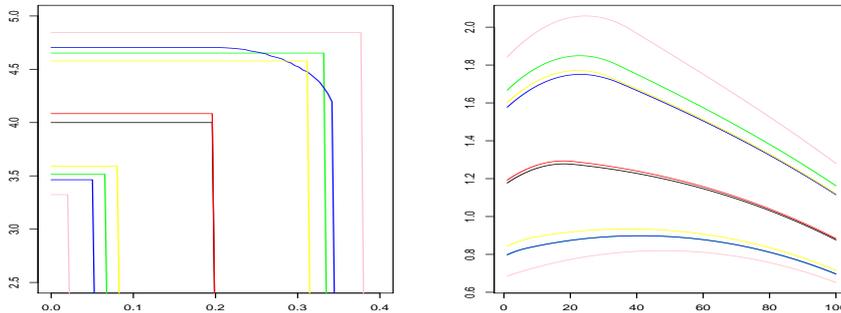$$h(y) = 4\mathbf{1}_{y<0.2} + \mathbf{1}_{y \geq 0.2}$$

Figure 5.1: Left: confidence bands for the estimator $\hat{h}_n$ (red) of the true step function h (black) using Scheffée (pink), Bonferroni (green), studentized maximum modulus (yellow) and Chi-squared (blue) statistics. Right: respective confidence bands for the estimation of the operator image.

with integral operator Φ as in Definition 1.3, with $a = 0$, $b = 1$ and Laplace kernel $\varphi(x, y) = 0.5e^{-|x-y|}$, that is

$$\Phi h(x) = \int_0^1 0.5e^{-|x-y|}4\mathbf{1}_{y<0.2} + \mathbf{1}_{y\geq 0.2}\mathrm{d}y + \varepsilon.$$

with normal distributed error ε, with variance $\sigma^2 = 0.05$ and mean zero. We calculated the least squares estimator $\hat{h}_n$ defined by (4.1) and, using the covariance matrix V_{f_θ} in (4.3), obtained 0.95-confidence bands for h and Φh.

Figure 5.1 shows, that the studentized statistic (yellow) leads to less conservative confidence sets for the parameter estimate, than Bonferroni (green) and Scheffée (pink). This is due to the fact, that in all three cases, the confidence sets are cuboids, which side lengths depend on the quantiles of the distribution of the used statistic. This so called *critical value*, is allways lowest for the studentized statistic, as long as the number of parameters is lower than the number of observations (cf. [39]). So the confidence cuboids for $\hat{\theta}_n$ are nested, with the cuboid of the studentized maximum modulus statistic as the least conservative of them.

The comparison of the studentized statistic and the exact Chi-squared statistic does not lead to similar obvious results. The reason is that the boundaries of the respective confidence sets for $\hat{\theta}_n$, namely cuboid and ellipsoid respectively, intersect. Hence for further comparison, we consider the surface area of the corresponding bands, generated by the exact and the approximative method, as a further criterion. Remarkably it is the approximate and not the exact method, which yields the narrower bands. Figure 5.1 shows, that the bands from the studentized statistic predominates even the exact bands obtained from the elliptic confidence set. Since furthermore the confidence sets based on the studentized maximum modulus statistic are the most straightforward to

calculate, it seems to be altogether the best choice for computation of confidence bands of pc-functions.

5.2 Example

In this subsection we are going to discuss an application of Corollary 4.1.4 to a problem from rheology. The aim here is the determination of the so called *relaxation time spectrum* (see [47]). The relaxation time spectrum is a characteristic quantity used in rheology which describes the viscoelastic properties of polymer solutions and polymer melts. Given this spectrum, it is very easy to convert one material function into another one. Additionally, many theories are based on the spectrum or provide predictions about its character (see for example [26]). Unfortunately, the relaxation time spectrum is not directly accessible by experiment. It is only possible to infer the spectrum from noisy observations of their image under a nonlinear integral operator defined as follows (see [47]).

Definition 5.2.1. *Let* $0 < a < 1 < b < \infty$ *and* $c \neq 0$.

$$H : L_2([a, b]) \longrightarrow L_2([a, b])$$
$$f \longmapsto Hf(x) := \int_a^b \frac{x^2 y}{1 + x^2 y^2} e^{cf(y)} \mathrm{d}y.$$

So, we have to deal with a Hammerstein integral $H = \Phi \circ \mathcal{L}$, with

$$\Phi f(y) = \int_a^b \frac{x^2 y^2}{1 + x^2 y^2} f(y) \mathrm{d}y$$

and

$$\mathcal{L}f(y) := y^{-1} e^{cf(y)}.$$

Note, that the exponential operator $\mathcal{L}$ and the linear operator Φ obviously satisfy Assumption **D** in Section 3.1 and Assumption **C1**, respectively.

The function f describing the relaxation time spectrum, is known to have the interpretation $f(\cdot, \theta) = \tilde{f}(\log(\cdot), \theta)$, such that $\tilde{f}(\cdot, \theta)$ is continuous and piecewise linear with two kinks (see [45]). This means that $\tilde{f}$ is an element of $\tilde{\mathbf{L}}_2$ as defined in (2.6) with reduced parameter vector $\tilde{\theta} = (\vartheta_1^1, \vartheta_2^1, \tau_1, \vartheta_2^2, \tau_2, \vartheta_2^3)$ (cf. Definition 2.2.3). For simplicity we rename $\tilde{\theta}$ as $\theta = (b_0, b_1, \tau_1, b_2, \tau_2, b_3)$. Then we have

$$\tilde{\mathbf{L}}_2 = \{\tilde{f} \in L_2([\log(a), \log(b)]) \mid \tilde{f}(y, \theta) = b_0 + b_1 y + b_2(y - \tau_1)_+ + b_3(y - \tau_2)_+, \ \theta \in \Theta\},$$

where Θ is assumed to be compact. Then, the true function $f_0(y) = f(y, \theta_0)$, we intend to estimate, is an element of the set

$$\mathbf{L}_{\log} := \{f(y, \theta) = \tilde{f}(\log(y), \theta) \mid \tilde{f} \in \tilde{\mathbf{L}}_2\},$$
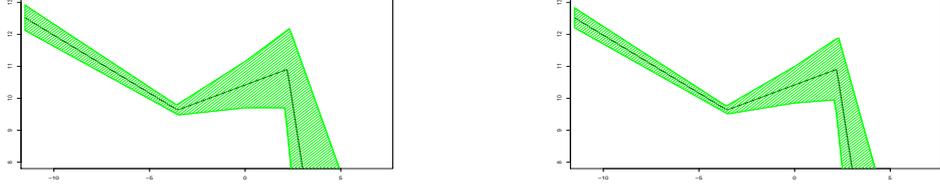
Figure 5.2: 0.95- and 0.80-confidence bands for the estimated function $\hat{f}_n$ plotted against $\log(y)$

which obviously satisfies the conditions of Definition 2.2.2.

In [47] it is assumed, that the observation model coincides with (1.1), where Assumptions **A1** and **B** on error and design are fulfilled. In the considered case this is equivalent to the model in (1.5) with f_0 substituted by $\mathcal{L}f_0$, namely

$$y_i = Hf(x_i, \theta_0) + \varepsilon = \Phi\mathcal{L}f(x_i, \theta_0) + \varepsilon_i \text{ for } i =, ..., n.$$

Figure 1 shows $n = 98$ noisy observations from the operator in Definition 5.2.1 acting on the unknown function $f_0 \in \mathbf{L}_{\log}$. Hence, as described in Section 3.1 and Corollary 4.1.5 in Remark 4.1.1, the results for (1.5) from Chapter 4, can be applied to gain an estimator for θ_0 and therewith for $f(y, \theta_0)$. More precisely, for the least squares estimator $\mathcal{L}f(y, \hat{\theta}_n) = \mathcal{L}\tilde{f}(\log(y), \hat{\theta}_n)$ of the function $\mathcal{L}f(y, \theta_0) = \mathcal{L}\tilde{f}(\log(y), \theta_0)$ as defined in (4.1), application of Corollary 4.1.5 yields

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N(0, \sigma^2 V_{\mathcal{L}f_{\theta_0}}^{-1}), \tag{5.2}$$

where $\sigma^2 = \mathbf{E}(\varepsilon^2)$ and $V_{\mathcal{L}f_{\theta_0}} \in \mathbb{R}^{6 \times 6}$ as in (4.3) is taking the form

$$(V_{\mathcal{L}f_{\theta_0}})_{ij} = \int_a^b (D_{\mathcal{L}f})_i(x, \theta_0)(D_{\mathcal{L}f})_j(x, \theta_0)\mathrm{d}x, \tag{5.3}$$

for $i, j = 1, ..., 6$, with gradient $D_{\mathcal{L}f}$ as defined in (4.2). Applying the chain rule, we have

$$(D_{\mathcal{L}f}(x, \theta_0))_i = \frac{\partial}{\partial \theta_i} \Phi\mathcal{L}f(y, \theta_0)(x) = \frac{\partial}{\partial \theta_i} Hf(y, \theta_0)(x)$$

$$= c \int_a^b \frac{x^2 y}{1 + x^2 y^2} e^{cf(y, \theta_0)}(df(y, \theta_0))_i \mathrm{d}y, \tag{5.4}$$

for $i = 1, ..., 6$, where

$$df(y, \theta) = \begin{pmatrix} 1 \\ \log(y) \\ -b_2 \mathbf{1}_{[e^{\tau_1}, b]} \\ (\log(y) - \tau_1)\mathbf{1}_{[e^{\tau_1}, b]} \\ -b_3 \mathbf{1}_{[e^{\tau_2}, b]} \\ (\log(y) - \tau_2)\mathbf{1}_{[e^{\tau_2}, b]} \end{pmatrix}.$$
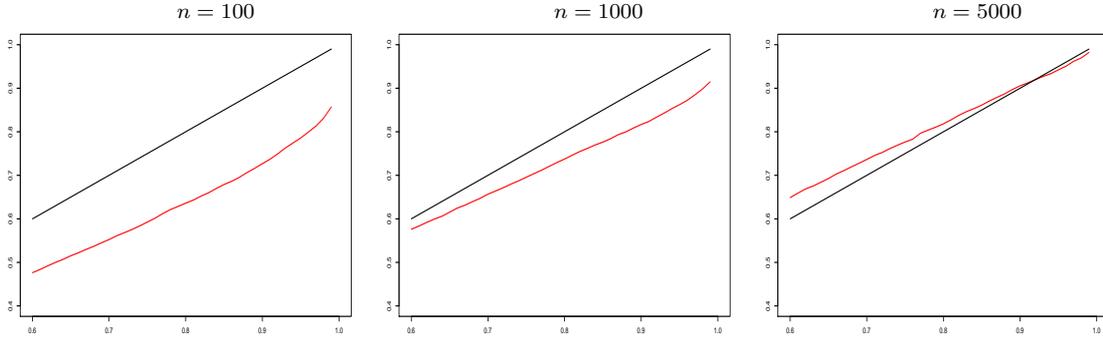
Figure 5.3: Empirical coverage probability of confidence bands for the estimated kink function for observations with $\sigma^2 = 0.01$, different sample sizes $n = 100$, $n = 1000$ and $n = 5000$, each with 10^4 simulations (red lines). The x-axis shows the nominal and the y-axis the empirical coverage probability.

First of all, we have to check, that the matrix $V_{\mathcal{L}f_{\theta_0}}$ in (5.3) is non singular. To this end, analogously to the proof of Proposition 4.1.3, we have to show, that the components of $D_{\mathcal{L}f}(x, \theta_0)$ in (5.4) are linearly independent. Because of the injectivity of the operator Φ (see Corollary 3.3.2) this follows, if the components of df are linear independent, i.e. if for $\theta_0 = (b_0^0, b_1^0, \tau_1^0, b_2^0, \tau_2^0, b_3^0)$,

$$\alpha_1 - \alpha_2 \log(y) + [\alpha_3(\log(y) - \tau_1^0) - \alpha_4 b_2^0]\mathbf{1}_{[e^{\tau_1^0}, b]}$$
$$+ [\alpha_5(\log(y) - \tau_2^0) - \alpha_6 b_3^0]\mathbf{1}_{[e^{\tau_2^0}, b]} \equiv 0 \quad \forall y \in [a, b]$$

implies $\alpha_i = 0$ for $i = 1, ..., 6$. Remembering that $b_2^0 \neq 0 \neq b_3^0$ and $\tau_1^0 < \tau_2^0$, since otherwise f_0 has less than two kinks, this is obviously the case.

Moreover, since f_0 is a kink function, Corollary 4.1.5, $(iv)^*$ yields the improved rate

$$\|f_0 - \hat{f}_n\|_{L_2([a,b])} = O_P(n^{-\frac{1}{2}}). \tag{5.5}$$

Note that the above rate in (5.5) follows from the fact, that the logarithm (on the compact interval $[a, b] \in \mathbb{R}^+$) as well as functions in $\mathcal{F}_L$ (cf. (2.4)) are Lipschitz continuous and hence the modulus of continuity $\nu(\mathcal{F}_L, n^{-1/2})$ (cf. (4.5)) is of order $O(n^{1/2})$.

Figure 5.2 shows the estimated kink function with 95%- and 80%- confidence bands, calculated by using a studentized maximum modulus statistic as discussed in Subsection 5.1. (For details see [39, p.70ff]), from the associated data of relaxation time spectrum (see [47]).

We evaluated the speed of convergence, i.e. quality of the approximation by the asymptotic law from Theorem 5.2 in this special example, by performing a simulation

study (see Figure 5.3). Here we used the Operator from Definition 5.2.1 operating on the space of kink functions with one kink. For $n = 100$ the confidence bands are very anti-conservative. For increasing n the empirical coverage approaches the nominal coverage. From $n = 5000$ the procedure begins to give reasonably useful results.

Chapter 6

Unknown number of change points

In this chapter we consider the case, where the true function f_0 in (1.5) has a finite but unknown number of change points, i.e. the case, where $f_0 \in \mathbf{F}_\infty$. In this case it obviously does not suffice to apply a least squares estimator as in (4.1), since then the number of change points is likely to diverge. To avoid this, we additionally penalize the number of change points. The resulting minimization functional is called *Potts functional*. For an application of this functional in a direct problem, i.e. when $\Phi = id$ in (1.3), where the estimated functions are step functions, we refer to [10].

6.1 Estimation and asymptotic results

We consider the minimizer $\hat{f}_{\lambda_n}$ of the *Potts functional*:

$$\|\Phi \hat{f}_{\lambda_n} - Y\|_n^2 + \lambda_n \sharp \mathcal{J}(\hat{f}_{\lambda_n}) \leq \min_{f \in \mathbf{F}_\infty} \|\Phi f - Y\|_n^2 + \lambda_n \sharp \mathcal{J}(f) + o(n^{-1}) \qquad (6.1)$$

where λ_n is a sequence of positive parameters converging to zero and for the true function f_0, the number of change points $\sharp \mathcal{J}(f_0)$ is assumed to be nonzero (otherwise take $\sharp \mathcal{J}(f) + 1$ in (6.1) instead for technical reasons).

The next Theorem states, that for a large range of regularization parameters $(\lambda_n)_{n \in \mathbb{N}}$ (see 6.1), the correct number of change points is estimated with probability tending to one. That means, for large enough n, the estimators $\hat{f}_n$ in (4.1) and $\hat{f}_{\lambda_n}$ in (6.1) coincide.

Theorem 6.1.1. *Suppose that Assumptions **A2**, **B** and **C** are satisfied. Let $f_0 \in \mathbf{F}_\infty$ and choose $\{\lambda_n\}_{n \in \mathbb{N}}$ such that*

$$\lambda_n \longrightarrow 0 \quad and \quad \lambda_n n^{\frac{1}{1+\epsilon}} \longrightarrow \infty,$$

for some $\epsilon > 0$. Then, the minimizer $\hat{f}_{\lambda_n}$ of (6.1) satisfies

$$P\left(\sharp \mathcal{J}(\hat{f}_{\lambda_n}) = \sharp \mathcal{J}(f_0)\right) \longrightarrow 1.$$

6.2 Proof

From (6.1) we obtain the basic inequality

$$\|\Phi\hat{f}_{\lambda_n} - \Phi f_0\|_n^2 \leq 2\langle \Phi f_0 - \Phi\hat{f}_{\lambda_n}, \varepsilon\rangle_n + \lambda_n(\sharp\mathcal{J}(f_0) - \sharp\mathcal{J}(\hat{f}_{\lambda_n})) + o(n^{-1}). \qquad (6.2)$$

Hence, as in Chapter 4, we have to consider the behavior of the empirical process $\langle \Phi f_0 - \Phi\hat{f}_{\lambda_n}, \varepsilon\rangle_n$. Again, we gain a bound for this process, by calculating the entropy of the respective function space. This will be done in Subsection 6.2.1. Subsequently, in Subsection 6.2.2 we introduce the proof of Theorem 6.1.1.

6.2.1 Entropy results

We aim to calculate the entropy number of the function class

$$\mathbf{G}_\infty := \{\Phi f \in L_2([a,b]) \mid f \in \mathbf{F}_\infty\}.$$

To this end we use a result from empirical process theory in [51], which is added in the Appendix in Section 8.4, Lemma 8.4.1.

Lemma 6.2.1. *Suppose that Assumptions $\mathbf{A}$ and $\mathbf{A1}$ are satisfied. Then, for all $\Phi f \in \mathbf{G}_\infty = \{\Phi f \in L_2([a,b]) \mid f \in \mathbf{F}_\infty\}$, there exists a constant T_1, independent of $\sharp\mathcal{J}(f)$, such that*

$$|\langle \Phi f, \varepsilon\rangle_n| = O_P(n^{-\frac{1}{2}})\sqrt{\sharp\mathcal{J}(f)}\|\Phi f\|_n \left(1 + \log\left(\frac{T_1\sqrt{\sharp\mathcal{J}(f)} + \|\Phi f\|_n}{\|\Phi f\|_n}\right)\right). \qquad (6.3)$$

Proof. For a fixed number of change points k, we find from Lemma 4.2.3 for $\mathbf{G}_k$ as in (4.8), that

$$H(\delta, \mathbf{G}_k, P_n) \leq d\log\left(\frac{T\sqrt{d} + \delta}{\delta}\right),$$

with $d = (k+1)r + k$ and a constant T, independent of k, that satisfies $T \geq \|g\|_{L_2([a,b])}$, for all $g \in \mathbf{G}_k$. Taking into account, that $e(u^{-1}\sqrt{d}T + 1) > e$ and hence $\log((T\sqrt{d} + u)u^{-1})^{\frac{1}{2}} < \log(e(T\sqrt{d} + u)u^{-1})^{\frac{1}{2}} < \log(e(T\sqrt{d} + u)u^{-1})$, we obtain

$$
\begin{aligned}
\mathcal{J}(\delta, \mathbf{G}_k, P_n) \;\; &:= \;\; \int_0^\delta \mathbf{H}^{1/2}(u, \mathbf{G}_k, P_n)\mathrm{d}u \\
&\leq \;\; \sqrt{d} \int_0^\delta 1 + \log\left(\frac{T\sqrt{d}+u}{u}\right)\mathrm{d}u \\
&= \;\; \sqrt{d}\left[u + (T\sqrt{d}+u)\log(T\sqrt{d}+u) - u\log(u)\right]_0^\delta \\
&= \;\; \sqrt{d}\Big((T\sqrt{d}(\log(T\sqrt{d}+\delta) - \log(T\sqrt{d})) \\
&\qquad\qquad + \delta(1 + \log(T\sqrt{d}+\delta) - \log(\delta))\Big) \\
&\leq \;\; 2c_1\sqrt{k}\delta(1 + \log(\tfrac{Tc_1\sqrt{k}+\delta}{\delta})),
\end{aligned}
$$

where $c_1 = \sqrt{2r+1}$ is independent of k. Here we used the mean value theorem to obtain $\log(T\sqrt{d}+\delta) - \log(T\sqrt{d}) = \delta(T\sqrt{d}+\tilde{\delta})^{-1} \leq \delta/(T\sqrt{d})$ for some $\tilde{\delta} \in (0, \delta)$. By Theorem 8.4.1, there exists a constant c_2 only depending on the constant α in the sub-gaussian error condition **A1**, such that

$$
\sqrt{n}\rho \geq c_2\left(\int_0^\delta \mathbf{H}^{1/2}(u, \mathbf{G}_k, P_n)\mathrm{d}u \vee \delta\right)
$$

implies

$$
P\Big(\sup_{f\in\mathbf{G}_k, \|f\|_n\leq\delta} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\Big| \geq \rho\Big) \leq c_2\exp\Big(-\frac{n\rho^2}{c_2^2\delta^2}\Big).
$$

Consequently for all $t \geq 2c_1c_2n^{-\frac{1}{2}}$, with $T_1 = Tc_1$ it holds

$$
\begin{aligned}
P\left(\sup_{f\in\mathbf{G}_k, \|f\|_n\leq\delta} |\langle f, \varepsilon\rangle_n| \geq t\sqrt{k}\delta\left(1 + \log\left(\frac{T_1\sqrt{k}+\delta}{\delta}\right)\right)\right) \\
\leq c_2\exp\left(-\frac{nt^2k(1+\log(\frac{T_1\sqrt{k}+\delta}{\delta}))^2}{c_2^2}\right) \leq c_2\exp\left(-\frac{nt^2}{c_2^2}\right),
\end{aligned}
$$

where the right hand side is independent of k. Thus

$$
\sup_{f\in\mathbf{G}_k, \|f\|_n\leq\delta} \frac{|\langle f, \varepsilon\rangle_n|}{\sqrt{k}\delta\left(1 + \log\left(\frac{T_1\sqrt{k}+\delta}{\delta}\right)\right)} = O_P(n^{-\frac{1}{2}}),
$$

holds uniformly for all k. For all $g \in \mathbf{G}_\infty$, this implies

$$\frac{|\langle g, \varepsilon \rangle_n|}{\sqrt{\sharp J(g)}\|g\|_n \left(1 + \log\left(\frac{T_1\sqrt{\sharp J(g)}+\|g\|_n}{\|g\|_n}\right)\right)}$$

$$\leq \sup_{\substack{f \in \mathbf{G}_{\sharp J(g)} \\ \|f\|_n \leq \|g\|_n}} \frac{|\langle f, \varepsilon \rangle_n|}{\sqrt{\sharp J(g)}\|g\|_n \left(1 + \log\left(\frac{T_1\sqrt{\sharp J(g)}+\|g\|_n}{\|g\|_n}\right)\right)} = O_P(n^{-\frac{1}{2}}),$$

and the claim follows. $\qquad\square$

Lemma 6.2.2. *Suppose that the Assumptions of Lemma 6.2.1 are satisfied. Then*

$$|\langle \Phi f, \varepsilon \rangle_n| = O_P(n^{-\frac{1}{2}})\|\Phi f\|_n^{1-\epsilon}(\sharp\mathcal{J}(f))^{\frac{1}{2}(1+2\epsilon)},$$

for any $\epsilon > 0$.

Proof. By construction of T_1 in the proof of Lemma 6.2.1 it holds $T_1 \geq \|\Phi f\|_n$ for all $f \in \mathbf{F}_\infty$. Since furthermore $\sharp\mathcal{J}(f) \geq 1$, we have

$$1 + \log\left(\frac{T_1\sqrt{\sharp\mathcal{J}(f)} + \|\Phi f\|_n}{\|\Phi f\|_n}\right) \leq 1 + \log\left(\frac{2T_1\sqrt{\sharp\mathcal{J}(f)}}{\|\Phi f\|_n}\right) \leq c\left(1 + .\log\left(\frac{\sqrt{\sharp\mathcal{J}(f)}}{\|\Phi f\|_n}\right)\right),$$

with $c = \log(2T_1) \geq 1$. Together with Theorem 6.2.1 this implies

$$|\langle \Phi f, \varepsilon \rangle_n| = O_P(n^{-\frac{1}{2}})\sqrt{\sharp\mathcal{J}(f)}\|\Phi f\|_n \left(1 + \log\left(\frac{\sqrt{\sharp\mathcal{J}(f)}}{\|\Phi f\|_n}\right)\right). \tag{6.4}$$

Now observe, that $\sqrt{x}(1 + \log(x)) \leq cx^{1/2+\epsilon}$ for $x \geq 1$, $\epsilon \geq 0$ and $c \geq \max(\epsilon^{-1}, 1)$. Moreover, if c is large enough and $x > 0$, then $x(1 + \log(x^{-1}) \leq cx^{1-\epsilon}$. Combining these results, remembering, that $\sharp\mathcal{J}(f) \geq 1$ and $\|\Phi f\|_n \leq T_1$, it follows, that

$$\sharp\mathcal{J}(f)^{\frac{1}{2}}\|\Phi f\|_n \left(1 + \log(\sharp\mathcal{J}(f)^{\frac{1}{2}}\|\Phi f\|_n^{-1})\right)$$
$$\leq \sharp\mathcal{J}(f)^{\frac{1}{2}}\|\Phi f\|_n \left(1 + \log(\sharp\mathcal{J}(f)^{\frac{1}{2}})\right) + \sharp\mathcal{J}(f)^{\frac{1}{2}}\|\Phi f\|_n \left(1 + \log(\|\Phi f\|_n^{-1})\right)$$

and thus

$$\begin{aligned}
\sharp\mathcal{J}(f)^{\frac{1}{2}}\|\Phi f\|_n \left(1 + \log(\sharp\mathcal{J}(f)^{\frac{1}{2}}\|\Phi f\|_n^{-1})\right) &\leq \|\Phi f\|_n\sharp\mathcal{J}(f)^{\frac{1}{2}+\epsilon} + \sharp\mathcal{J}(f)^{\frac{1}{2}}\|\Phi f\|_n^{1-\epsilon} \\
&\leq (\|\Phi f\|_n^\epsilon + 1)\sharp\mathcal{J}(f)^{\frac{1}{2}+\epsilon}\|\Phi f\|_n^{1-\epsilon} \\
&\leq (T_1^\epsilon + 1)\sharp\mathcal{J}(f)^{\frac{1}{2}+\epsilon}\|\Phi f\|_n^{1-\epsilon}.
\end{aligned}$$

This, together with (6.4), proves the claim. $\qquad\square$

6.2.2 Consistency

Before we start to proof Theorem 6.1.1 the following statement is needed.

Lemma 6.2.3. *Let $f_0 \in \boldsymbol{F}_\infty$ and $\{f_n\}_{n \in \mathbb{N}}$ be a sequence in $\boldsymbol{F}_{\sharp \mathcal{J}(f_0)}$, with*

$$\|f_0 - f_n\|_{L_2([a,b])} = o_P(1),$$

then

$$\lim_{n \to \infty} P(\sharp \mathcal{J}(f_0) = \sharp \mathcal{J}(f_n)) = 1.$$

Proof. From the assumption $\|f_0 - f_n\|_{L_2([a,b])} = o_P(1)$ it follows that

$$\lim_{n \to \infty} P(\|f_0 - f_n\|_{L_2([a,b])} \leq \rho) = 1 \text{ for all } \rho > 0$$

and the claim follows by application of Lemma 8.1.1, $i)$, if additionally there exists some $\bar{\rho} > 0$ such that $\lim_{n \to \infty} P(\sharp \mathcal{J}(f_0) = \sharp \mathcal{J}(f_n) \mid \|f_0 - f_n\|_{L_2([a,b])} \leq \bar{\rho}) = 1$. Thus, we have to show, that there exist some $\bar{\rho} > 0$, such that

$$\sharp \mathcal{J}(f_0) = \sharp \mathcal{J}(f) \text{ for all } f \in \mathbf{F}_{\sharp \mathcal{J}(f_0)} \text{ with } \|f_0 - f\|_{L_2([a,b])} \leq \bar{\rho}. \tag{6.5}$$

To this end, w.l.o.g. let $\sharp \mathcal{J}(f_0) = 1$. Now we assume, that the contrary of (6.5) holds, that is for all $\rho > 0$ there exists $f_\rho \in \mathbf{F}_{\sharp \mathcal{J}(f_0)}$, such that

$$\|f_0 - f_\rho\|_{L_2([a,b])} \leq \rho \text{ and } \sharp \mathcal{J}(f_\rho) = 0 < \sharp \mathcal{J}(f_0).$$

This means that $f_\rho \in \mathbf{F}_0$ for all $\rho > 0$. Since ρ is arbitrary we have $\lim_{\rho \to 0} \|f_0 - f_\rho\|_{L_2([a,b])} = 0$, which implies that $f_\rho \in \mathbf{F}_0$ converges to f_0, if ρ converges to zero. As shown in the proof of Lemma 4.2.7, the set $\mathbf{F}_0$ is compact with respect to the L_2-norm and thus the limit function of f_ρ has to be contained in $\mathbf{F}_0$, which leads to the contradiction

$$f_0 \in \mathbf{F}_0.$$

$\square$

Now we are prepared for the proof of

Theorem 6.1.1

Proof. Throughout the proof w.l.o.g we assume, that $\epsilon \leq 1$. From Lemma 6.2.2 and (6.2), it follows that

$$
\begin{aligned}
\|\Phi \hat{f}_{\lambda_n} - \Phi f_0\|_n^2 \leq{} & O_P(n^{-\frac{1}{2}}) \|\Phi \hat{f}_{\lambda_n} - \Phi f_0\|_n^{1-\frac{1}{2}\epsilon} (\sharp \mathcal{J}(\hat{f}_{\lambda_n} - f_0))^{\frac{1}{2}(1+\epsilon)} \\
& + \lambda_n (\sharp \mathcal{J}(f_0) - \sharp \mathcal{J}(\hat{f}_{\lambda_n})) + o(n^{-1}) \\
\leq{} & O_P(n^{-\frac{1}{2}}) \|\Phi \hat{f}_{\lambda_n} - \Phi f_0\|_n^{1-\frac{1}{2}\epsilon} \sharp \mathcal{J}(\hat{f}_{\lambda_n})^{\frac{1}{2}(1+\epsilon)} \\
& - \lambda_n \sharp \mathcal{J}(\hat{f}_{\lambda_n}) + \lambda_n \sharp \mathcal{J}(f_0),
\end{aligned}
\tag{6.6}
$$

where we took into account, that λ_n is assumed to converge slower than n^{-1} and that we have $\sharp\mathcal{J}(f_0) < \infty$, which implies, that $\sharp\mathcal{J}(\hat{f}_{\lambda_n} - f_0) = O_P(\sharp\mathcal{J}(\hat{f}_{\lambda n}))$.

Choosing $f \equiv 0$ on the right hand side of Equation (6.1) implies $\lambda_n\sharp\mathcal{J}(\hat{f}_{\lambda_n}) \leq \|Y\|_n^2 = O_P(1)$ and hence, we have

$$\sharp\mathcal{J}(\hat{f}_{\lambda_n}) = O_P(\lambda_n^{-1}). \tag{6.7}$$

We assumed that $\lambda_n^{-1}n^{-1/(1+\epsilon)} \to 0$, for $n \to \infty$, which gives

$$n^{-1} = o(\lambda_n^{1+\epsilon}). \tag{6.8}$$

By Lemma 8.2.4 we have that $\sup_{f\in\mathbf{F}_\infty}\|f\|_\infty \leq R$ and thus

$$\sup_{f\in\mathbf{F}_\infty}\|\Phi f\|_n \leq \|\varphi\|_\infty R \leq \infty \tag{6.9}$$

Inserting (6.9), (6.7) and (6.8) into (6.6), we obtain

$$
\begin{aligned}
\|\Phi\hat{f}_{\lambda_n} - \Phi f_0\|_n^2 &\leq o_P(\lambda_n^{\frac{1+\epsilon}{2}})O_P(\lambda_n^{\frac{1-\epsilon}{2}})\|\Phi\hat{f}_{\lambda_n} - \Phi f_0\|_n^{1-\frac{1}{2}\epsilon}\sharp\mathcal{J}(\hat{f}_n) \\
&\quad -\lambda_n\sharp\mathcal{J}(\hat{f}_{\lambda_n}) + \lambda_n\sharp\mathcal{J}(f_0) \\
&= (o_p(\lambda_n) - \lambda_n)\sharp\mathcal{J}(\hat{f}_{\lambda_n}) + \lambda_n\sharp\mathcal{J}(f_0). \tag{6.10}
\end{aligned}
$$

Since $\sharp\mathcal{J}(f_0)$ is bounded and by Lemma 8.1.1, *ii)* we have $(o_P(1) - 1)\sharp\mathcal{J}(\hat{f}_n) \leq O_P(\lambda_n)$. This implies

$$\|\Phi\hat{f}_{\lambda_n} - \Phi f_0\|_n^2 = O_P(\lambda_n)$$

and with Corollary 8.2.6,

$$\|\Phi\hat{f}_{\lambda_n} - \Phi f\|_{L_2([a,b])}^2 = O_P(\lambda_n) + o_P(1) = o_P(1). \tag{6.11}$$

Again considering (6.10) we find, that

$$0 \leq (o_p(\lambda_n) - \lambda_n)\sharp\mathcal{J}(\hat{f}_{\lambda_n}) + \lambda_n\sharp\mathcal{J}(f_0),$$

which means

$$(1 + o_P(1))\sharp\mathcal{J}(\hat{f}_{\lambda_n}) \leq \sharp\mathcal{J}(f_0).$$

Because $\sharp\mathcal{J}(f)$ and $\sharp\mathcal{J}(\hat{f}_{\lambda_n})$ are integers, this implies $P(\sharp\mathcal{J}(\hat{f}_{\lambda_n}) \leq \sharp\mathcal{J}(f_0)) \to 1$ by application of Lemma 8.1.1, *iii)*. From Lemma 4.2.7 together with (6.11), we obtain $\|f_0 - \hat{f}_{\lambda_n}\|_{L_2([a,b])} = o_P(1)$, if $\sharp\mathcal{J}(\hat{f}_{\lambda_n}) \leq \sharp\mathcal{J}(f_0)$, i.e. $f_0, \hat{f}_{\lambda_n} \in \mathbf{F}_{\sharp\mathcal{J}(f_0)}$. That means, for all $\rho > 0$ it holds, that

$$\lim_{n\to\infty} P\left(\|f_0 - \hat{f}_{\lambda_n}\|_{L_2([a,b])} \leq \rho \mid \sharp\mathcal{J}(\hat{f}_{\lambda_n}) \leq \sharp\mathcal{J}(f_0)\right) = 1$$

and Lemma 8.1.1, *i)* yields

$$\|f_0 - \hat{f}_{\lambda_n}\|_{L_2([a,b])} = o_P(1).$$

Finally using Lemma 6.2.3 this implies $\lim_{n\to\infty} P\left(\sharp\mathcal{J}(f_0) = \sharp\mathcal{J}(\hat{f}_{\lambda_n})\right) = 1$, which is the claim. $\qquad\square$

Chapter 7

Approximation spaces

In this chapter, we will discuss what happens if the true function f_0 in (1.5) is not a pc-function but can be approximated by those sufficiently well. To be more precise, f_0 is assumed to be contained in an *approximation space* $\mathcal{A}^\alpha(\mathbf{F}_\infty)$. Such a space is a subset of the closure of $\mathbf{F}_\infty$, whose elements can be approximated by pc-functions at a certain rate, indicated by α. They have been studied in various contexts, as by Besov [6], who studied approximation spaces of trigonometric functions (cf. [18, Thm. 9.1, Chap. 7, §9]), by Butzer and Scherer [13] and Peetre and Spaar [42], who pointed out relations to the interpolation of operators, or by Lorenz [37], who considered the computation of metric entropy of sets of functions. More details about approximation spaces can be found in [18].

For estimation we will use a least squares minimizer, which is additionally penalized by the number of jumps and the L_2-norm of the objective function. We will show, that in our observation model (1.5), this minimization yields a sequence of pc-functions, which converges to the true function f_0 contained in an approximation space $\mathcal{A}^\alpha$. That means, it is possible to solve the ill-posed problem of reconstructing $f_0 \in \mathcal{A}^\alpha$ from the observation model in (1.5), by approximation with functions from $\mathbf{F}_\infty$.

Under appropriate source conditions we furthermore determine how fast this sequence converges. The direct problem, i.e. $\Phi = id$ in (1.5), has been studied by Boysen, Kempe et al. in [10], where they considered the approximation space of step functions. They showed, that in the direct case, penalizing the number of jumps suffices to obtain a converging sequence of step functions. Moreover they obtain convergence rates of order $O_P(\gamma_n^{\alpha/(2\alpha+1)})$, where the regularization parameter γ_n is chosen such that, $\gamma_n \to 0$ for $n \to \infty$ and $\gamma_n n \log(n)^{-1} \to \infty$.

Considering the inverse problem in (1.5), we obtain (under a source condition) rates of order $O_P(\lambda_n^{\alpha/2(2\alpha+1)})$ for the indirect problem and $O_P(\lambda_n^{\alpha/(2\alpha+1)})$ for the direct problem, where the regularization parameter λ_n satisfies $\lambda_n \to 0$ for $n \to \infty$ and $\lambda_n n^{1/(1+\epsilon)} \to \infty$.

7.1 Definition

In order to define the approximation space with respect to a given set of pc-functions we have to introduce some more assumptions on the function sets $\mathbf{F}_k$ and $\mathbf{F}_\infty$.

Assumption E. For all $k \in \mathbb{N}$, the set $\tilde{\mathbf{F}}_k$ (with the corresponding parameter set $\Theta \subset \mathbb{R}^{(k+1)r+k}$) fulfills the conditions of Definition 2.2.2, where additionally it holds, that

- $\Psi = \mathbb{R}^r$ (cf. Definition 2.2.1),

- $0 \in \tilde{\mathbf{F}}_k$, $\tilde{\mathbf{F}}_0 := \{0\}$,

- $a\tilde{\mathbf{F}}_k = \tilde{\mathbf{F}}_k$ for each $a \neq 0$,

- $\tilde{\mathbf{F}}_k + \tilde{\mathbf{F}}_k = \tilde{\mathbf{F}}_{2k}$.

We define $\tilde{\mathbf{F}}_\infty = \bigcup_{k=0}^\infty \tilde{\mathbf{F}}_k$ respectively.

Note that under these conditions $\tilde{\mathbf{F}}_\infty$ becomes a linear space. They are satisfied for example if the function class $\mathcal{F}$, which generates $\tilde{\mathbf{F}}_\infty$ (cf. Definition 2.2.1), consists of polynomials of arbitrary but fix degree, such as step functions or linear functions, and the respective pc-function spaces $\mathbf{F}_k$ are free knot splines (cf. Remark 4.1.2).

Definition 7.1.1. *Let $\tilde{\boldsymbol{F}}_k$ be a set of pc-functions, which satisfies Assumption $\boldsymbol{E}$. Then, we define the approximation error $\Lambda_k(f)$, by*

$$\Lambda_k(f) := \inf\{\|g - f\|_{L_2([a,b])} : g \in \tilde{\mathbf{F}}_k[a,b]\}.$$

Further, for $\alpha > 0$ the approximation space $\mathcal{A}^\alpha = \mathcal{A}^\alpha(\tilde{\boldsymbol{F}}_\infty)$, is defined as

$$\mathcal{A}^\alpha := \{f \in L_2([a,b]) : \sup_{k \geq 1} k^\alpha \Lambda_k(f) < \infty\}. \tag{7.1}$$

In the following section, we will consider the bounded subset $\tilde{\mathbf{F}}_{\infty,C} \subset \tilde{\mathbf{F}}_\infty$, defined as $\tilde{\mathbf{F}}_{\infty,C} := \bigcup_{k=1}^\infty \tilde{\mathbf{F}}_{k,C}$, with

$$\tilde{\mathbf{F}}_{k,C} := \{f(\cdot, \theta) \in \tilde{\mathbf{F}}_k[a,b] \mid |\theta|_\infty < C\},$$

where the constant $C < \infty$ has to be chosen such that the true function f_0 is contained in the closure $\overline{\tilde{\mathbf{F}}_{\infty,C}}$. Note, that by this restriction, $\tilde{\mathbf{F}}_{k,C}$ again satisfy the conditions of Definition 2.2.2, such that we are in a similar situation as in the first part of the thesis. Especially it follows from Lemma 8.2.4, that here we also have a uniform bound

$$\sup_{f \in \tilde{\mathbf{F}}_{\infty,C}} \|f\|_\infty < \infty \tag{7.2}$$

and thus, there exists a constant $K > 0$, such that

$$\sup_{f \in \tilde{\mathbf{F}}_{\infty,C}} \|\Phi f\|_\infty \leq (b-a)\|\varphi\|_\infty \sup_{f \in \tilde{\mathbf{F}}_{\infty,C}} \|f\|_\infty := K < \infty. \tag{7.3}$$

7.2 Estimate and asymptotic results

We now aim for estimating a function $f_0 \in \mathcal{A}^\alpha$ from noisy observations given by the model in (1.5). To this end, we again employ the penalized least squares estimator in (6.1), that is, we intend to approximate f_0 by functions from $\mathbf{F}_\infty$. The penalization of the number of change points however, is not enough to ensure continuity of the resulting regularization operator. But as we will show, we obtain a consistent estimator $\hat{f}_n \in \tilde{F}_{\infty,C}$, as minimizer of the following extension of the Potts functional.

$$\|\Phi\hat{f}_n - Y\|_n^2 + \lambda_n \sharp \mathcal{J}(\hat{f}_n) + \mu_n \|\hat{f}_n\|_{L_2([a,b])}^2$$
$$\leq \min_{g \in \tilde{\mathbf{F}}_{\infty,C}} \|\Phi g - Y\|_n^2 + \lambda_n \sharp \mathcal{J}(g) + \mu_n \|g\|_{L_2([a,b])}^2 + o(n^{-1}). \quad (7.4)$$

We use $\hat{f}_n$ as abbreviation for $\hat{f}_{\lambda_n,\mu_n}$, so we have to keep in mind that the minimizer depends on the two regularization parameters λ_n and μ_n.

For the estimator $\hat{f}_n$ we obtain the following asymptotic results.

Theorem 7.2.1. *Let $f_0 \in \mathcal{A}^\alpha$, $\hat{f}_n$ be as in (7.4), λ_n such that $\lambda_n \to 0$ and $\lambda_n n^{1/(1+\epsilon)} \to \infty$ for some $\epsilon > 0$, and μ_n, such that $\mu_n \to 0$ and $\mu_n^{-1} \lambda_n^{2\alpha/(2\alpha+1)} \to 0$ for $n \to \infty$. The operator Φ defined in (1.3) satisfies Assumption **C** and the observation model (1.5) satisfies Assumption **A**, **A1** and **B**, with $\nu_n = O_P(\mu_n)$. Then*

$$\|\Phi f_0 - \Phi\hat{f}_n\|_{L_2([a,b])} = O_P(\mu_n^{\frac{1}{2}}),$$

and

$$\|\hat{f}_n - f_0\|_{L_2([a,b])} = o_p(1).$$

Proof. Since $f_0 \in \mathcal{A}^\alpha$ and since the constant C is chosen, such that $f_0 \in \overline{\tilde{\mathbf{F}}_{\infty,C}}$, we can find a sequence of functions $f_{k_n} \in \tilde{\mathbf{F}}_{k_n,C}$, with $k_n := \lfloor \lambda_n^{-1/(2\alpha+1)} \rfloor$, such that

$$\|f_{k_n} - f_0\|_{L_2([a,b])}^2 \leq O\left(k_n^{-2\alpha}\right). \quad (7.5)$$

We use this rate to find an upper bound for the rate of convergence of $\Phi\hat{f}_n$ and $\hat{f}_n$. To this end, we first set $g = f_{k_n}$ in (7.4) and find

$$\|\Phi\hat{f}_n - \Phi f_0 - \varepsilon\|_n^2 + \lambda_n \sharp \mathcal{J}(\hat{f}_n) + \mu_n \|\hat{f}_n\|_{L_2([a,b])}^2$$
$$\leq \|\Phi f_{k_n} - \Phi f_0 - \varepsilon\|_n^2 + \lambda_n k_n + \mu_n \|f_{k_n}\|_{L_2([a,b])}^2 + o(n^{-1}),$$

and therefore

$$\|\Phi\hat{f}_n - \Phi f_0\|_n^2 \leq \|\Phi f_{k_n} - \Phi f_0\|_n^2 + 2|\langle \Phi\hat{f}_n - \Phi f_{k_n}, \varepsilon \rangle_n|$$
$$+ \lambda_n(k_n - \sharp \mathcal{J}(\hat{f}_n)) + \mu_n(\|f_{k_n}\|_{L_2([a,b])}^2 - \|\hat{f}_n\|_{L_2([a,b])}^2) + o(n^{-1}). \quad (7.6)$$

By Hölder's inequality and keeping in mind that the kernel φ of the operator Φ is bounded, the first term of the right hand side can be estimated by

$$\|\Phi f_{k_n} - \Phi f_0\|_n^2 \le (b-a)\|\varphi\|_\infty^2 \|f_{k_n} - f_0\|_{L_2([a,b])}^2 = O(k_n^{-2\alpha}). \tag{7.7}$$

Moreover, setting $g \equiv 0$ in (7.4), we obtain

$$\lambda_n \sharp \mathcal{J}(\hat{f}_n) \le \|\Phi f_0\|_n^2 + \|\varepsilon\|_n^2 \le O_P(1).$$

Hence we have $\sharp \mathcal{J}(\hat{f}_n) = O_P(\lambda_n^{-1})$. Remembering that by definition $\sharp \mathcal{J}(f_{k_n}) \le k_n = \lfloor \lambda_n^{-1/(2\alpha+1)} \rfloor = O(\lambda_n^{-1})$, this means

$$\sharp \mathcal{J}(\hat{f}_n - f_{k_n}) \le \sharp \mathcal{J}(\hat{f}_n) + k_n = O_P(\lambda_n^{-1}). \tag{7.8}$$

By assumption, it holds that $n^{-1/(1+\epsilon)}\lambda_n^{-1} \to 0$ for $n \to \infty$. This yields the estimate

$$n^{-1} = o(\lambda_n^{1+\epsilon}). \tag{7.9}$$

Since $\hat{\theta}_n$ and θ_{k_n} are bounded by C, we can apply Corollary 6.2.2 with ϵ substituted by $\epsilon/2$, which yields

$$2|\langle \Phi \hat{f}_n - \Phi f_{k_n}, \varepsilon \rangle_n| \le (\sharp \mathcal{J}(\hat{f}_n) + k_n))^{\frac{1+\epsilon}{2}} O_P(n^{-\frac{1}{2}}) \|\Phi \hat{f}_n - \Phi f_{k_n}\|_n^{\frac{2-\epsilon}{2}}.$$

With $\|\Phi \hat{f}_n - \Phi f_{k_n}\|_n \le 2\sup_{f \in \tilde{\mathbf{F}}_{\infty,C}} \|\Phi f\|_\infty \le 2K$ from (7.3) and inserting (7.8) and (7.9), this implies

$$2|\langle \Phi \hat{f}_n - \Phi f_{k_n}, \varepsilon \rangle_n| \le (\sharp \mathcal{J}(\hat{f}_n) + k_n))(2K)^{\frac{2-\epsilon}{2}} o_P(\lambda_n^{\frac{1+\epsilon}{2}}) O(\lambda_n^{\frac{1-\epsilon}{2}})$$
$$= o_P(\lambda_n)\sharp \mathcal{J}(\hat{f}_n) + o_P(k_n\lambda_n). \tag{7.10}$$

Applying (7.7) and (7.10) to (7.6) now we obtain

$$\begin{aligned}
\|\Phi \hat{f}_n - \Phi f_0\|_n^2 &\le o_P(\lambda_n)\sharp \mathcal{J}(\hat{f}_n) - \lambda_n \sharp \mathcal{J}(\hat{f}_n) \\
&\quad + \mu_n(\|f_{k_n}\|_{L_2([a,b])}^2 - \|\hat{f}_n\|_{L_2([a,b])}^2) + O_P(k_n^{-2\alpha} + k_n\lambda_n + n^{-1}) \\
&= \sharp \mathcal{J}(\hat{f}_n)(o_P(\lambda_n) - \lambda_n) + \mu_n(\|f_{k_n}\|_{L_2([a,b])}^2 - \|\hat{f}_n\|_{L_2([a,b])}^2) + O_P(\lambda_n^{\frac{2\alpha}{2\alpha+1}}),
\end{aligned}$$

where we took into account, that $k_n = O(\lambda_n^{-1/(2\alpha+1)})$ and $n^{-1} = O(\lambda_n)$.

From Lemma 8.1.1, $ii)$ we obtain $\sharp \mathcal{J}(\hat{f}_n)(o_P(\lambda_n) - \lambda_n) \le O_P(\lambda_n^{2\alpha/(2\alpha+1)})$. Hence we can skip that term in the above equation and obtain

$$\|\Phi \hat{f}_n - \Phi f_0\|_n^2 \le O_P(\lambda_n^{\frac{2\alpha}{2\alpha+1}}) + \mu_n(\|f_{k_n}\|_{L_2([a,b])}^2 - \|\hat{f}_n\|_{L_2([a,b])}^2) = O_P(\mu_n), \tag{7.11}$$

since $\|f_{k_n}\|_{L_2([a,b])}^2$ and $\|\hat{f}_n\|_{L_2([a,b])}^2$ are bounded by (7.2) and μ_n dominates $\lambda_n^{\frac{2\alpha}{2\alpha+2}}$. With Corollary 8.2.6 and taking into account, that by assumption of the theorem $\nu_n = O_P(\mu_n)$, for ν as in Assumption **B**, this results in

$$\|\Phi \hat{f}_n - \Phi f_0\|_{L_2([a,b])}^2 = O_P(\mu_n),$$

and proves the first claim. This can now be used to show the second one, since for all $g \in L_2([a,b])$ the first claim implies

$$|\langle \Phi \hat{f}_n - \Phi f_0, g \rangle_{L_2([a,b])}| = |\langle \hat{f}_n - f_0, \Phi^* g \rangle_{L_2([a,b])}| = o_P(1). \tag{7.12}$$

The operator $\Phi : L_2([a,b]) \to L_2([a,b])$ is injective by Assumption **C**, *i*). Thus, the range of its adjoint operator $\Phi^*(L_2([a,b]))$ is dense in $L_2([a,b])$ (see e.g. [35, Thm. 15.8]). Therefore (7.12) yields $|\langle \hat{f}_n - f_0, \bar{g} \rangle_{L_2([a,b])}| = o_P(1)$ for all $\bar{g} \in L_2([a,b])$, which we call *weak convergence in probability*, denoted as

$$\hat{f}_n \xrightarrow{P} f_0.$$

From (7.11), we get

$$0 \leq O_P(\lambda_n^{\frac{2\alpha}{2\alpha+1}}) + \mu_n(\|f_{k_n}\|_{L_2([a,b])}^2 - \|\hat{f}_n\|_{L_2([a,b])}^2)$$

and taking into account, that $\mu_n^{-1}\lambda_n^{2\alpha/(2\alpha+1)} = o(1)$ by assumption of the theorem, this gives

$$\|\hat{f}_n\|_{L_2([a,b])}^2 - \|f_{k_n}\|_{L_2([a,b])}^2 = O_P(\mu_n^{-1}\lambda_n^{\frac{2\alpha}{2\alpha+1}}) = o_P(1). \tag{7.13}$$

Finally (7.5) and (7.13), together with the weak convergence of $\hat{f}_n$ yield

$$
\begin{aligned}
\|\hat{f}_n - f_0\|_{L_2([a,b])}^2 &= \|\hat{f}_n\|_{L_2([a,b])}^2 - 2\langle \hat{f}_n, f_0 \rangle_{L_2([a,b])} + \|f_0\|_{L_2([a,b])}^2 \\
&\leq 2\langle f_0 - \hat{f}_n, f_0 \rangle_{L_2([a,b])} + \|\hat{f}_n\|_{L_2([a,b])}^2 - \|f_0\|_{L_2([a,b])}^2 \\
&\leq 2\langle f_0 - \hat{f}_n, f_0 \rangle_{L_2([a,b])} + \|\hat{f}_n\|_{L_2([a,b])}^2 - \|f_{k_n}\|_{L_2([a,b])}^2 \\
&\quad + O(\|f_0 - f_{k_n}\|_{L_2([a,b])}) \\
&= o_P(1).
\end{aligned}
\tag{7.14}
$$

$\square$

The next theorem shows, that the result of Theorem 7.2.1 can be improved, if we have some a prioriinformation on the smoothing properties of the operator Φ, more precisely, if we know, that $f_0 \in \Phi^*(L_2([a,b]))$.

Theorem 7.2.2. *Assume that the conditions of Theorem 7.2.1 are satisfied, but now we have that $\nu_n = O_P(\mu_n^2)$ and $\mu_n = O(\lambda_n^{\frac{\alpha}{2\alpha+1}})$. If additionally $f_0 = \Phi^* g_0$ for some $g_0 \in L_2([a,b])$, one has*

$$\|\Phi \hat{f}_n - \Phi f_0\|_{L_2([a,b])} = O_P(\lambda_n^{\frac{\alpha}{2\alpha+1}})$$

and

$$\|\hat{f}_n - f_0\|_{L_2([a,b])} = O_P(\lambda_n^{\frac{\alpha}{2(2\alpha+1)}}).$$

Proof. If $f_0 = \Phi^* g_0$ for some $g_0 \in L_2([a,b])$ this yields

$$
\begin{aligned}
0 &\leq \|\hat{f}_n - f_0\|_{L_2([a,b])}^2 \\
&\leq 2\langle \Phi f_0 - \Phi \hat{f}_n, g_0 \rangle_{L_2([a,b])} + \|\hat{f}_n\|_{L_2([a,b])}^2 - \|f_{k_n}\|_{L_2([a,b])}^2 \\
&\quad + O(\|f_0 - f_{k_n}\|_{L_2([a,b])}).
\end{aligned}
\tag{7.15}
$$

Inserting (7.5) with $k_n = \lfloor \lambda_n^{-1/(2\alpha+1)} \rfloor$, together with Hölder's inequality, this gives

$$
\begin{aligned}
\|f_{k_n}\|_{L_2([a,b])}^2 - \|\hat{f}_n\|_{L_2([a,b])}^2 &\leq c\|\Phi f_0 - \Phi \hat{f}_n\|_{L_2([a,b])} + O(\|f_0 - f_{k_n}\|_{L_2([a,b])}) \\
&= c\|\Phi f_0 - \Phi \hat{f}_n\|_{L_2([a,b])} + O_P(\lambda_n^{\frac{\alpha}{2\alpha+1}}),
\end{aligned}
\tag{7.16}
$$

where $c \geq 2\|g_0\|_{L_2([a,b])}$. From Lemma 8.2.6 it follows, that

$$
\begin{aligned}
\|\Phi \hat{f}_n - \Phi f_0\|_{L_2([a,b])}^2 &\leq s_u \|\Phi \hat{f}_n - \Phi f_0\|_n^2 + O_P(\nu_n + n^{-1}) \\
&= s_u \|\Phi \hat{f}_n - \Phi f_0\|_n^2 + O_P(\lambda_n^{\frac{2\alpha}{2\alpha+1}}),
\end{aligned}
\tag{7.17}
$$

with s_u and ν_n as in Lemma 8.2.6, taking into account that $\nu_n = O_P(\mu_n^2) = O_P(\lambda_n^{\frac{2\alpha}{2\alpha+1}})$ by assumption of the theorem.

Recalling (7.11) from the proof of Theorem 7.2.1 and applying (7.16) and (7.17), we have

$$
\begin{aligned}
\|\Phi \hat{f}_n - \Phi f_0\|_{L_2([a,b])}^2 &\leq O_P(\lambda_n^{\frac{2\alpha}{2\alpha+1}}) + \mu_n s_u (\|f_{k_n}\|_{L_2([a,b])}^2 + \|\hat{f}_n\|_{L_2([a,b])}^2) \\
&\leq O_P(\lambda_n^{\frac{2\alpha}{2\alpha+1}}) + \mu_n c s_u \|\Phi f_0 - \Phi \hat{f}_n\|_{L_2([a,b])} + O_P(\mu_n \lambda_n^{\frac{\alpha}{2\alpha+1}}) \\
&= O_P(\lambda_n^{\frac{2\alpha}{2\alpha+1}}) + \mu_n c s_u \|\Phi f_0 - \Phi \hat{f}_n\|_{L_2([a,b])},
\end{aligned}
$$

using $\mu_n = O(\lambda_n^{\frac{\alpha}{2\alpha+1}})$. This in turn implies

$$
\left(\|\Phi \hat{f}_n - \Phi f_0\|_{L_2([a,b])} - \tfrac{1}{2} c s_u \mu_n \right)^2 \leq O_P(\lambda_n^{\frac{2\alpha}{2\alpha+1}}) + \tfrac{1}{4}(s_u c \mu_n)^2 = O_P(\lambda_n^{\frac{2\alpha}{2\alpha+1}})
$$

and finally the first claim

$$
\|\Phi \hat{f}_n - \Phi f_0\|_{L_2([a,b])} \leq O_P(\lambda_n^{\frac{\alpha}{2\alpha+1}}) + O(\mu_n) = O_P(\lambda_n^{\frac{\alpha}{2\alpha+1}}).
$$

To prove the second claim, recall (7.13)

$$
\|\hat{f}_n\|_{L_2([a,b])}^2 - \|f_{k_n}\|_{L_2([a,b])}^2 \leq O_p(\mu_n^{-1} \lambda_n^{\frac{2\alpha}{2\alpha+1}}) = O_P(\lambda_n^{\frac{\alpha}{2\alpha+1}}).
$$

Combining this result, with the first claim, (7.5) with $k_n = O(\lambda_n^{-1/(2\alpha+1)})$ and (7.15), yields

$$
\begin{aligned}
\|\hat{f}_n - f_0\|_{L_2([a,b])}^2 &\leq c\|\Phi f_0 - \Phi \hat{f}_n\|_{L_2([a,b])} + \|\hat{f}_n\|_{L_2([a,b])}^2 - \|f_{k_n}\|_{L_2([a,b])}^2 \\
&\quad + O(\|f_0 - f_{k_n}\|_{L_2([a,b])}) \\
&= O_P(\lambda_n^{\frac{\alpha}{2\alpha+1}}),
\end{aligned}
$$

which finishes the proof. $\qquad\square$

Example 7.2.3. (Approximation spaces of step functions) We consider the particular instance of approximation spaces of step functions. As it was shown in [10, Ex. 2], the class of Hölder continuous functions with exponent α ($0 \leq \alpha \leq 1$), is contained in $\mathcal{A}^{\alpha}(\tilde{T}_{\infty})$, where $\tilde{\mathbf{T}}_{\infty}$ is generated by the sets of step functions $\mathbf{T}_k$ from (2.3) by means of Assumption **E**. Thus, under the conditions of Theorem 7.2.2 for the estimator $\hat{f}_n \in \mathbf{T}_{\infty}$ in (7.4), we have

$$\|f_0 - \hat{f}_n\|_{L_2([a,b])} = O_P(\lambda_n^{\frac{\alpha}{2(2\alpha+1)}}).$$

Choosing α sufficiently large and λ_n close to n^{-1}, then yields convergence rates, arbitrary close to $O_P(n^{-1/4})$, which is the rate in the case where $f_0 \in \mathbf{T}_k$ for some finite $k \in \mathbb{N}$ (see Theorem 4.1.1).

Example 7.2.4. (Spline approximation) We consider the set $\tilde{\mathbf{P}}_{q,k}$ of piecewise polynomials of degree q with k kinks, which satisfy Assumption **E** and the corresponding set $\tilde{\mathbf{P}}_{q,\infty} = \bigcup_{k=1}^{\infty} \tilde{\mathbf{P}}_{q,k}$. We know from [15, Thm. XII. (34)], that any function $f \in C^p[0,1]$, $0 < p < q$ can be approximated by piecewise polynomials of order q with k kinks, with fixed kink locations (i.e. B-splines as described in Example 4.1.2), at a rate of $O(k^{-p})$. Approximation with piecewise polynomials with free kink locations has to be at least as fast as this rate. Thus $C^p[0,1]$ is contained in the corresponding approximation space $\mathcal{A}^p = \mathcal{A}^p(\tilde{\mathbf{P}}_{q,\infty})$ of piecewise polynomials and Theorem 7.2.2 leads to

$$\|\Phi f_0 - \Phi \hat{f}_n\|_{L_2([a,b])} = O_P(\lambda_n^{\frac{p}{2p+1}})$$

and

$$\|f_0 - \hat{f}_n\|_{L_2([a,b])} = O_P(\lambda_n^{\frac{p}{2(2p+1)}}),$$

for the penalized least squares estimator $\hat{f}_n \in \tilde{\mathbf{P}}_{q,\infty}$ from (7.4).

Note, that these rates cannot be assumed to be the best possible, since approximation by piecewise polynomials with *fixed* kink locations cannot be expected to yield the best upper bound for the convergence rates for approximation by piecewise polynomials with *free* kink locations.

Indeed, there can be found better results. For comparison we want to mention the paper of Cardot [14], who considered the same model as in (1.5) with $f_0 \in C^p[0,1]$ with similar assumptions on error and design. The spline fitting of f_0 is realized by minimizing least squares with local roughness penalties and rates of order $n^{-p/(2p+1)}$ are obtained for the direct problem (with respect to the empirical norm), which is the limit of the rates above for $\epsilon \to 0$ in Theorem 7.2.1. Convergence rates for the indirect problem are not discussed in [14].

Chapter 8

Appendix

8.1 The symbols O_P and o_P

There are many rules of calculus for the o_P and O_P symbols, which are applied without comment (for details we refer to [53]). Here we mention some basic examples:

$$
\begin{aligned}
o_P(1) + o_P(1) &= o_P(1) \\
o_P(1) + O_P(1) &= O_P(1) \\
o_P(1)O_P(1) &= o_P(1) \\
(1 - o_P(1))^{-1} &= O_P(1) \\
o_P(R_n) &= R_n o_P(1) \\
O_P(R_n) &= R_n O_P(1) \\
o_P(O_P(1)) &= o_P(1).
\end{aligned}
$$

Furthermore, we often use the fact that

$$
X_n = O_P(1) \text{ implies } X_n^p = O_P(1) \text{ for } p > 0.
$$

This is a straightforward consequence of the definition of boundedness in probability, using that $f(x) = x^p$ is monotone for $p > 0$. The same holds for o_P instead of O_P.

Finally, there are some more rules, which are more complex, given in the following

Lemma 8.1.1. *Suppose that (Ω, Σ, P) is a probability space and that $n \in \mathbb{N}$.*

i) If $A_n, B_n \subset \Sigma$ are sequences of events, then

$$
\lim_{n \to \infty} P(A_n) = 1 \wedge \lim_{n \to \infty} P(B_n | A_n) = 1 \text{ implies } \lim_{n \to \infty} P(B_n) = 1.
$$

ii) For an arbitrary sequence $\lambda_n \in \mathbb{R} \setminus \{0\}$, it holds that

$$
o_P(1) - 1 \leq O_P(\lambda_n).
$$

iii) If $a_n : \Omega \to \mathbb{N}$ is a sequence of random variables and $b \in \mathbb{N}$, then

$$a_n(1 + o_P(1)) \leq b \text{ for all } n \in \mathbb{N} \text{ implies } \lim_{n \to \infty} P(a_n \leq b) = 1.$$

Proof. It follows from the conditions in *i)*, that $\lim_{n \to \infty} P(B_n|A_n)P(A_n) = 1$ as well as $\lim_{n \to \infty} P(A_n^c) = 0$. Hence we obtain the total probability

$$\lim_{n \to \infty} P(B_n) = \lim_{n \to \infty} P(B_n|A_n)P(A_n) + \lim_{n \to \infty} P(B_n|A_n^c)P(A_n^c) = 1.$$

In order to show *ii)*, consider a random variable $X_n = o_P(1)$. If there is a subsequence $k_n \in \mathbb{N}$ with $X_{k_n} - 1 \geq 0$ it holds, that

$$
\begin{aligned}
P(|(X_{k_n} - 1)\lambda_{k_n}^{-1}| \geq 1) &= P((X_{k_n} - 1)|\lambda_{k_n}^{-1}| \geq 1) \\
&= P(X_{k_n} \geq |\lambda_{k_n}| + 1) \leq P(X_{k_n} \geq 1) \longrightarrow 0,
\end{aligned}
$$

as $n \to \infty$, since $X_n = o_P(1)$. So $X_{k_n} - 1 = O_P(\lambda_{k_n})$. If $X_n - 1 < 0 = O_P(\lambda_n)$, then *ii)* also follows.

Finally we want to prove *iii)*. From the assumption $a_n(1 + o_P(1)) \leq b$, we obtain

$$a_n(1 + X_n) \leq b, \tag{8.1}$$

with some random variable $X_n = o_P(1)$. Furthermore, the definition of $o_P(1)$ yields

$$\lim_{n \to \infty} P(|X_n| \leq (2b + 1)^{-1}) = 1. \tag{8.2}$$

Applying $|X_n| \leq (2b + 1)^{-1}$ to (8.1) gives

$$a_n \leq (1 + X_n)^{-1}b \leq |1 - |X_n||^{-1}b \leq (2b + 1)(2b)^{-1}b = b + \tfrac{1}{2},$$

which means, that $a_n \leq b$, since a_n and b are integers. Thus, we have

$$P(a_n \leq b \mid |X_n| \leq (2b + 1)^{-1}) = 1$$

and together with (8.2), the claim follows from *i)*. $\qquad\qquad\square$

8.2 Technical tools

The following lemmata are essential for the proofs in this thesis. They are added separately because of their technical character. They mainly collect some properties of the functions $f \in \mathbf{F}_\infty$ as in Definition 2.2.2 and $\Phi f \in L_2([a, b])$ with Φ satisfies Assumption **C**.

Lemma 8.2.1. *Let $f, g \in \boldsymbol{F}_\infty(a, b)$. If $\|f - g\|_{L_2([a,b])} = 0$, then f and g are pointwise identical pc-functions.*

Proof. By Definition 2.2.2 the functions $g, f \in \mathbf{F}_\infty$ have finite number of changepoints and they are right-continuous with existing left-sided limits in any $y \in [a, b]$. Thus the same holds for the square of the difference $(f - g)^2$, which means that $\|f - g\|_{L_2([a,b])}$ implies $(f(y) - g(y))^2 \equiv 0$ for all $y \in [a, b]$. $\square$

Lemma 8.2.2. *Assume that $g(\cdot, \theta) \in L_2([a, b])$ and that $\theta \mapsto g(y, \theta)$ is continuous in θ_0 for almost every $y \in [a, b]$. Let Φ be an integral operator as defined in (1.3), with kernel $\varphi \in L_\infty([a, b]^2)$. Assume further that there exists a function $\bar{g} \in L_2[a, b]$ and an open neighbourhood $U(\theta_0)$, such that for all $y \in [a, b]$*

$$|g(y, \theta)| \leq \bar{g}(y) \quad \forall \theta \in U(\theta_0).$$

Then, the map $\theta \mapsto \Phi g(x, \theta)$ is continuous in θ_0 for all $x \in [a, b]$.

Proof. Let $\{\theta_n\}_{n \in \mathbb{N}}$ be a sequence in Θ converging to θ_0, and $\{g(y, \theta_n)\}_{n \in \mathbb{N}}$ the sequence of corresponding functions. Since θ_n converges to θ_0, w.l.o.g. we assume that $\theta_n \in U(\theta_0)$. This means that for all $x \in [a, b]$ and almost every $y \in [a, b]$ one has

$$\lim_{n \to \infty} \varphi(x, y) g(y, \theta_n) = \varphi(x, y) g(y, \theta_0).$$

Furthermore, by assumption $\sup_{\theta_n \in U(\theta_0)} |g(y, \theta_n)| < \bar{g}(y)$, with $\bar{g} \in L_2([a, b])$ and by dominated convergence (e.g. [4, Thm. 15.1 and 15.6]) we find

$$
\begin{aligned}
\lim_{\theta_n \to \theta_0} (\Phi g(x, \theta_n) - \Phi g(x, \theta_0)) &= \lim_{\theta_n \to \theta_0} \int_a^b \varphi(x, y)(g(y, \theta_n) - g(y, \theta_0)) \mathrm{d}y \\
&= \int_a^b \lim_{\theta_n \to \theta_0} \varphi(x, y)(g(y, \theta_n) - g(y, \theta_0)) \mathrm{d}y \\
&= 0,
\end{aligned}
$$

for all $x \in [a, b]$. $\square$

Lemma 8.2.3. *Let $\vartheta \in \mathbb{R}^r$ and $g(y, \vartheta)$ be a function, which is continuous in every $y \in [a, b] \subset \mathbb{R}$ and additionally continuous in ϑ^0 for all $y \in [a, b]$. Then, for $a \leq \tau_1 < \tau_2 \leq b$ the map*

$$\theta = (\tau_1, \vartheta_1, ..., \vartheta_r, \tau_2) \longmapsto \tilde{g}(y, \theta) := g(y, \vartheta) \mathbf{1}_{[\tau_1, \tau_2)}(y)$$

is continuous in $\theta^0 = (\tau_1, \vartheta^0, \tau_2)$, for all $y \in [a, b] \setminus \{\tau_1, \tau_2\}$.

   If furthermore there exists a neighborhood $B(\vartheta^0) \subset \Theta$, independent of $y \in [a, b]$ such that $\vartheta \mapsto g(y, \vartheta)$ is continuously differentiable for all $\vartheta \in B(\vartheta^0)$ for all $y \in [a, b]$, the map

$$(y, \vartheta) \longmapsto g(y, \vartheta)$$

is continuous in $[a, b] \times B(\vartheta^0)$.

Proof. Since $\vartheta \mapsto g(y, \vartheta)$ is continuous in ϑ^0, we can find for every $\varepsilon > 0$ a neighborhood $B_\varepsilon(\vartheta_0) \subset \mathbb{R}^r$, such that for all Δ_ϑ, with $\vartheta^0 + \Delta_\vartheta \in B_\varepsilon(\vartheta^0)$

$$|g(y, \vartheta^0 + \Delta_\vartheta) - g(y, \vartheta^0)| < \varepsilon.$$

If $y \notin \{\tau_1, \tau_2\}$ there exist open neighborhoods $U(\tau_1)$ and $U(\tau_2)$ in $[a, b]$, that do not contain y. Then for all $\Delta = (\delta_1, \Delta_\vartheta, \delta_2) \in \mathbb{R}^{r+2}$, such that $\theta^0 + \Delta \in B_1 := U(\tau_1) \times B_\varepsilon(\vartheta_0) \times U(\tau_2)$ we have

$$\begin{aligned}
|\tilde{g}(y, \theta^0 + \Delta) - \tilde{g}(y, \theta^0)| &= |g(y, \vartheta^0 + \Delta_\vartheta)\mathbf{1}_{[\tau_1 + \delta_1, \tau_2 + \delta_2)}(y) - g(y, \vartheta^0)\mathbf{1}_{[\tau_1, \tau_2)}(y)| \\
&= |g(y, \vartheta^0 + \Delta)\mathbf{1}_{[\tau_1, \tau_2)}(y) - g(y, \vartheta^0)\mathbf{1}_{[\tau_1, \tau_2)}(y)| < \varepsilon,
\end{aligned}$$

which proves the first claim.

If additionally $\vartheta \mapsto g(y, \vartheta)$ is differentiable in $\vartheta \in B(\vartheta^0)$ for every $y \in [a, b]$, we have for all $\vartheta^0 + \Delta \in B(\vartheta^0)$, that

$$g(y, \vartheta^0 + \Delta) - g(y, \vartheta^0) = \sum_{i=1}^r \Delta_i^t \frac{\partial}{\partial \vartheta_i} g(y, \vartheta^0) + r(\Delta),$$

with

$$\lim_{\Delta \to 0} \frac{r(\Delta)}{|\Delta|_r} = 0.$$

Furthermore $y \mapsto g(y, \vartheta)$ is continuous on $[a, b]$ and thus for all $\varepsilon > 0$ we can find a neighborhood $U_{\frac{1}{3}\varepsilon}(y)$, such that for all $y + \delta \in U_{\frac{1}{3}\varepsilon}(y)$ it holds

$$|g(y, \vartheta^0) - g(y + \delta, \vartheta^0)| < \frac{1}{3}\varepsilon.$$

If $\sup_{i=1,\dots,r} \left| \frac{\partial}{\partial \vartheta_i} g(y, \theta^0) \right| = c > 0$ define

$$B_{\frac{1}{3}\varepsilon}(\vartheta^0) = \{\vartheta^0 + \Delta \in B(\vartheta^0) \mid |\Delta|_\infty \leq \tfrac{\varepsilon}{3rc} \text{ and } |r(\Delta)| < \tfrac{1}{3}\varepsilon\}$$

and finally $B_2 := U_{\frac{1}{3}\varepsilon}(y) \times B_{\frac{1}{3}\varepsilon}(\vartheta^0)$. Then, it holds for all $(y + \delta, \vartheta^0 + \Delta) \in B_2$

$$\begin{aligned}
|g(y + \delta, \vartheta^0 + \Delta) - g(y, \vartheta^0)| &\leq |g(y + \delta, \vartheta^0) - g(y, \vartheta^0)| + |\Delta|_\infty c + |r(\Delta)| \\
&\leq \frac{1}{3}\varepsilon + \frac{1}{3}\varepsilon + \frac{1}{3}\varepsilon.
\end{aligned}$$

If $\sup_{i=1,\dots,r} \left| \frac{\partial}{\partial \vartheta_i} g(y, \vartheta^0) \right| = 0$, a slight change of the neighborhood of ϑ^0, i.e. taking $B_{\frac{1}{3}\varepsilon}(\vartheta^0) = \{\vartheta^0 + \Delta \in B(\vartheta^0) \mid |r(\Delta)| < \tfrac{1}{3}\varepsilon\}$ instead, will do it. $\square$

The next Lemma presents a collection of properties of $\Phi f(x, \theta)$ implied by the Assumption on the objective function $f \in \mathbf{F}_k$ in Definition 2.2.2.

Lemma 8.2.4. *Let Φ be defined as in (1.3) and satisfy Assumption $\boldsymbol{C}$. Moreover assume, that s is a density fulfilling Assumption $\boldsymbol{B}$ and that $f \in \boldsymbol{F}_k$ (cf. Definition 2.2.2). Then,*

   i) there is a constant $R > 0$, independent of k, such that

$$\sup_{\substack{i=1,\ldots,k+1, \\ j=1,\ldots r,\, \theta \in \Theta}} \left(\|f(\cdot, \theta)\|_{\infty}, \|\frac{\partial}{\partial \vartheta_j^i} f(\cdot, \theta)\|_{L_2([a,b])} \right) \le R,$$

   ii) the map $\theta \mapsto \Phi f(x, \theta)$ is continuous for all $x \in [a, b]$,

   iii) the map $\theta \mapsto \Phi f(x, \theta)$ is differentiable for almost every $x \in [a, b]$ with gradient $D_f(x, \theta) = \frac{\partial}{\partial \theta} \Phi f(x, \theta)$ as in (4.2),

   iv) the map $x \mapsto \Phi f(x, \theta)$ is Lipschitz continuous for all $\theta \in \Theta$, with uniform Lipschitz constant $c_1 R$, with constant c_1 depending only on the kernel φ,

   v) the map

$$|\boldsymbol{E} D_f(x, \cdot)|_2 \quad : \quad \Theta \longrightarrow \mathbb{R}$$
$$\theta \longmapsto \left| \int_a^b D_f(x, \theta) s(x) \mathrm{d}x \right|_2$$

   is continuous,

   vi) there is a constant $C > 0$, such that

$$\sup_{\theta \in \Theta,\, i=1,\ldots d} \|(D_f(\cdot, \theta))_i\|_{\infty} \le C.$$

Proof. Due to Definition 2.2.1 *ii)* and the second claim of Lemma 8.2.3, for every $\mathfrak{f} \in \mathcal{F}$, the mapping $(y, \vartheta) \mapsto \mathfrak{f}(y, \vartheta)$ is continuous on the compact set $[a, b] \times \Psi$ and therefore we have a uniform bound for all $\mathfrak{f} \in \mathcal{F}$. Since $f \in \mathbf{F}_k$ is generated by functions in $\mathcal{F}$ this implies that there exists a constant $c_1 \ge 0$, such that

$$\sup_{f \in \mathbf{F}_k} \|f\|_{\infty} = \sup_{\mathfrak{f} \in \mathcal{F}} \|\mathfrak{f}\|_{\infty} \le c_1.$$

Similar, with $g \in L_2([a, b])$ as in Definitions 2.2.1, *iii)*, we have

$$\sup_{\substack{f \in \mathbf{F}_k \\ }} \sup_{\substack{i=1,\ldots,k+1, \\ j=1,\ldots,r}} \left\| \frac{\partial}{\partial \vartheta_j^i} f(\cdot, \theta) \right\|_{L_2([a,b])} = \sup_{\substack{\mathfrak{f} \in \mathcal{F} \\ j=1,\ldots,r}} \left\| \frac{\partial}{\partial \vartheta_j} \mathfrak{f}(\cdot, \vartheta) \right\|_{L_2([a,b])} \le \|g\|_{L_2([a,b])} \le c_2.$$

Choosing $R = \max(c_1, c_2)$ proves the first claim.

To show the second claim, remember that $\theta \mapsto f(y, \theta)$ is continuous in θ for every $y \in [a, b] \setminus \{\tau_0, ..., \tau_{k+1}\}$ by Lemma 8.2.3 and is uniformly bounded by R (cf. i)). Hence ii) follows from Lemma 8.2.2.

The bound R from i) for $f \in \mathbf{F}_k[a, b]$ together with Assumption **C** proves iv).

Using the bound from i) for the partial derivatives in addition with Definitions 2.2.1 and 2.2.2, yields the conditions for application of [20, Thm 5.7, Chap. IV], which claims existence of the partial derivatives

$$\frac{\partial}{\partial \vartheta_j^i} \Phi f(x, \theta) = \int_{\tau_{i-1}}^{\tau_i} \varphi(x, y) \frac{\partial}{\partial \vartheta_j^i} f(y, \theta) \mathrm{d}y = \left(\Phi \frac{\partial}{\partial \vartheta_j^i} \mathfrak{f}(y, \vartheta^i) \mathbf{1}_{[\tau_{i-1}, \tau_i)} \right)(x).$$

Furthermore, $\Phi f(x, \theta)$ is differentiable in τ_i for $i = 1, \ldots, k$ by the fundamental theorem of calculus, with

$$\frac{\partial}{\partial \tau_i} \Phi f(x, \theta) = \varphi(x, \tau_i)(f(\tau_i^-, \theta) - f(\tau_i, \theta))$$

and we obtain the gradient $D = \frac{\partial}{\partial \theta} \Phi f(x, \theta)$, calculated as

$$D_i = (D_f(x, \theta))_i = \begin{cases} \displaystyle\int_a^b \varphi(x, y) \frac{\partial}{\partial \theta_i} f(y, \theta) \mathrm{d}y & i \neq 0 \bmod (r+1), \\[2ex] \varphi(x, \tau_{\frac{i}{r+1}})(f(\tau_{\frac{i}{r+1}}^-, \theta) - f(\tau_{\frac{i}{r+1}}, \theta)) & i = 0 \bmod (r+1). \end{cases} \tag{8.3}$$

In order to prove iii), we have to show, that the components of $D_f(x, \theta)$ are continuous in θ. Lemma 8.2.3 implies continuity of $\theta \mapsto \frac{\partial}{\partial \vartheta_j^i} f(y, \theta)$ for all $i = 1, ..., k+1$, $j = 1, ..., r$ and $y \in [a, b] \setminus \{\tau_0, ...\tau_{k+1}\}$. Thus, the integral $\int_{\tau_{i-1}}^{\tau_i} \varphi(x, y) \frac{\partial}{\partial \vartheta_j^i} f(y, \theta) \mathrm{d}y$ is also continuous in θ by Lemma 8.2.2. The map

$$\theta \mapsto \varphi(x, \tau_i)(f(\tau_i^-, \theta) - f(\tau_i, \theta)) = \varphi(x, \tau_i)(\mathfrak{f}(\tau_i, \vartheta^{i-1}) - \mathfrak{f}(\tau_i, \vartheta^i))$$

is continuous for almost every $x \in [a, b]$, since as argued above, the pieces $\mathfrak{f}(y, \vartheta^i)$ are continuous in $(y, \vartheta^i) \in [a, b] \times \Psi$. Additionally it was claimed, that the kernel $\varphi(x, y)$ is discontinuous only in a null set. This in turn means that $\Phi f(x, \theta)$ is continuously partial differentiable in θ_i, $i = 1, \ldots, d$ for almost every $x \in [a, b]$ and hence differentiable in θ for almost every $x \in [a, b]$, which proves iii).

Statement vi) follows from (8.3) together with the bound in i), by

$$\sup_{\theta \in \Theta, \, i=1, ...d} \|(D_f(x, \theta))_i\|_\infty \leq \|\varphi\|_\infty R \max(2, (b-a)) =: C.$$

Finally, we again use (8.3) to show v). As already shown, the components $(D_f(x, \theta))_i$ for $i = 1, \ldots, k$, are continuous in θ, for almost every $x \in [a, b]$. So, by dominated convergence in the same way as in the proof of Lemma 8.2.2 we obtain continuity of the map $\theta \mapsto \mathbf{E}(D_f(\cdot, \theta))_i = \int_a^b (D_f(x, \theta))_i s(x) \mathrm{d}x$ for any $i = 1, ..., d$. That means, the Euclidean norm $|\mathbf{E}D_f(\cdot, \theta)|_2$ is a sum of continuous functions and hence continuous in θ, as v) claims. $\qquad \square$

We often need to compare empirical norm and L_2-norm of some function. The following result proves to be useful for that purpose.

Lemma 8.2.5. *Suppose that s, ν_n and δ_i, for $i = 1, ..., n$ satisfy Assumption $\boldsymbol{B}$ and $\mathcal{F} \subset L_2([a, b])$ is a class of Lipschitz continuous functions, with uniform Lipschitz constant κ. Then it holds uniformly for all $f \in \mathcal{F}$, that*

$$\int_a^b f(x)s(x)dx = \frac{1}{n}\sum_{i=1}^n f(x_i) + O_P(\nu_n + n^{-1}).$$

Proof. W.l.o.g we assume that $a = 0$ and $b = 1$. Let $S(x) = \int_0^x s(y)\mathrm{d}y$, where s is as in Assumption $\boldsymbol{B}$. Note that S is strictly monotone and the inverse S^{-1} is well defined on $[0, 1]$. For $0 \le c \le d \le 1$ we have that

$$d - c = S(S^{-1}(d)) - S(S^{-1}(c)) = \int_{S^{-1}(c)}^{S^{-1}(d)} s(y)\mathrm{d}y \ge s_l(S^{-1}(d) - S^{-1}(c)).$$

Hence S^{-1} is Lipschitz continuous and so is $f \circ S^{-1}$ for any $f \in \mathcal{F}$ (with constant κ/s_l). By Assumption $\boldsymbol{B}$ we have $S^{-1}(i/n - \delta_i) = x_{(i)}$ with $\nu_n = \max_{i=1,...,n} |\delta_i|$, $i = 1, ..., n$. Consequently,

$$n\int_{(i-1)/n}^{i/n} f(S^{-1}(y))\mathrm{d}y = f(x_{(i)}) + n\int_{(i-1)/n}^{i/n} f(S^{-1}(y)) - f(S^{-1}(i/n))\mathrm{d}y$$

$$+ n\int_{(i-1)/n}^{i/n} f(S^{-1}(i/n)) - f(S^{-1}(i/n - \delta_i))\mathrm{d}y$$

and by Lipschitz continuity of $f \circ S^{-1}$ we obtain

$$\left| \int_{(i-1)/n}^{i/n} f(S^{-1}(y))\mathrm{d}y - n^{-1}f(x_{(i)}) \right| \le \kappa/s_l(n^{-1} + \nu_n).$$

Since $s_l, \kappa < \infty$, substituting $y = S^{-1}(u)$, this implies

$$\left| \frac{1}{n}\sum_{i=1}^n f(x_i) - \int_0^1 f(y)s(y)\mathrm{d}y \right| = \left| \frac{1}{n}\sum_{i=1}^n f(x_i) - \int_0^1 f(S^{-1}(u))\mathrm{d}u \right|$$

$$\le \kappa/s_l(n^{-1} + \nu_n),$$

which proves the claim, since κ and s_l do not depend on $f \in \mathcal{F}$. $\qquad\square$

Corollary 8.2.6. *Suppose that Assumption $\boldsymbol{C}$ and Assumption $\boldsymbol{B}$ are met. Let $\boldsymbol{F} \subset L_\infty([a, b])$ be a class of functions, such that there exists a constant $c_F \ge \|f\|_\infty$ for all $f \in \boldsymbol{F}$. Then,*

$$O_P(\nu_n + n^{-1}) + s_l\|\Phi f\|_n^2 \le \|\Phi f\|_{L_2([a,b])}^2 \le s_u\|\Phi f\|_n^2 + O_P(\nu_n + n^{-1}),$$

*with constants s_l, s_u depending only on the design density (cf. Assumption **B**) (i.e. this holds uniformly for all $f \in \mathbf{F}$). Especially this implies, that uniformly for all $f \in \mathbf{F}_\infty$, with $\mathbf{F}_\infty$ as in Definition 2.2.2, we have*

$$o_P(1) + s_l \|\Phi f\|_n^2 \leq \|\Phi f\|_{L_2([a,b])}^2 \leq s_u \|\Phi f\|_n^2 + o_P(1).$$

Proof. By Assumption **C**, the functions $\Phi f(x)$ are Lipschitz continuous for all $f \in \mathbf{F}$, with uniform Lipschitz constant $c_\Phi \|f\|_\infty = c_\Phi c_F$, with constant c_Φ only depending on the operator Φ. Hence the functions $(\Phi f)^2$, $f \in \mathbf{F}$ are also Lipschitz continuous with uniform Lipschitz constant and application of Lemma 8.2.5 yields

$$\sup_{f \in \mathbf{F}} \left| \frac{1}{n} \sum_{i=1}^{n} (\Phi f(x_i))^2 - \int_a^b (\Phi f(y))^2 s(y) \mathrm{d}y \right| = O(n^{-1} + \nu_n).$$

Finally, recalling $s_l \leq s(x) \leq s_u$ for all $y \in [a, b]$ (Assumption **B**), this yields the first claim. The second claim follows from the assumption $\nu_n = o_P(1)$, together with Lemma 8.2.4, which yields a uniform bound for all $f \in \mathbf{F}_\infty$. $\qquad\square$

Lemma 8.2.7. *Assume that $\rho_1 < x < \rho_2$ and let $\{c_n\}_{n \in \mathbb{N}}$ be a sequence of real numbers, such that the series $\sum_{n=0}^{\infty} c_n x^n$ converges absolutely on the interval $[\rho_1, \rho_2]$. Then,*

$$\sum_{n=0}^{\infty} c_n x^n = 0 \Leftrightarrow c_n = 0 \qquad \forall n.$$

Proof. Due to the absolute convergence, we obtain by differentiation

$$\sum_{n=0}^{\infty} c_n x^n = 0 \;\Rightarrow\; \sum_{n=1}^{\infty} n c_n x^{n-1} = 0$$
$$\Leftrightarrow \sum_{n=0}^{\infty} (n+1) c_{n+1} x^n = 0.$$

Comparing the coefficients yields $(n+1)c_{n+1} = c_n$ for all $n \in \mathbb{N}$. If $0 \in [\rho_1, \rho_2]$ applying $x = 0$ yields $c_0 = 0$ and by induction the assertion. If $0 \notin [\rho_1, \rho_2]$, multiplying the second equation above with x and subsequently comparing the coefficients again, gives $n c_n = c_n$. That means $c_n = 0$ for all $n \geq 1$ and therefore $c_0 = 0$. $\qquad\square$

Lemma 8.2.8. *Suppose that $\phi \in L_2(\mathbb{R})$ is an analytic function and $f \in L_2(\mathbb{R})$ has compact support* $\mathrm{supp} f = [a, b] \subset \mathbb{R}$. *Then the convolution*

$$x \longmapsto \int_{-\infty}^{\infty} \phi(x - y) f(y) \mathrm{d}y$$

is analytic.

Proof. The aim is to show that for all $x_0 \in \mathbb{R}$ there exists an open neighborhood $U(x_0)$ and a sequence $(a_n)_{n \in \mathbb{N}}$ in $\mathbb{R}$ such that

$$\int_{-\infty}^{\infty} \phi(x-y)f(y)\mathrm{d}y = \sum_{n=0}^{\infty} a_n(x-x_0)^n$$

for all $x \in U(x_0)$.

Therefore, w.l.o.g we assume $x_0 = 0$. Since ϕ is analytic and $[a,b]$ compact, there exists a finite number of open neighborhoods $U(x_1), ..., U(x_l)$, with $x_1, ..., x_l \in (-b, -a)$, such that for all $i = 1, ..., l$, it holds that

$$[-b, -a] \subsetneq \bigcup_{i=1}^{l} U(x_i) \qquad \text{and} \qquad \phi(z) = \sum_{n=0}^{\infty} a_n^i(z-x_i)^n \text{ for all } z \in U(x_i).$$

So we can choose a partition $a = \tau_0 < \tau_1 < ... < \tau_l = b$ such that $x_i \in [-\tau_i, -\tau_{i-1}]$ and $[-\tau_i, -\tau_{i-1}] \subsetneq U(x_i)$ for all $i = 1, ..., l$.. Hence, there exists a neigborhood $U(0)$, such that for all $x \in U(0)$

$$x - y \in U(x_i) \text{ for all } y \in [\tau_{i-1}, \tau_i],$$

for $i = 1, ..., l$. For any $x \in U(0)$ and all $i = 1, ..., l$ this means that

$$\begin{aligned}
\phi(x - x_i - y) &= \sum_{n=1}^{\infty} a_n^i(x - x_i - y)^n \\
&= \sum_{n=1}^{\infty} a_n^i \sum_{k=0}^{n} \binom{n}{k} x^k (-1)^{n-k} (y + x_i)^{n-k} \\
&= \sum_{k=0}^{\infty} x^k \sum_{n=k}^{\infty} a_n^i \binom{n}{k} (-1)^{n-k} (y + x_i)^{n-k} \qquad (8.4)
\end{aligned}$$

converges absolutely and uniformly for all y in the compact interval $[\tau_{i-1}, \tau_i]$. For the last equation we took into account that

$$\begin{aligned}
\sum_{n=1}^{\infty} \sum_{k=0}^{n} \left| a_n^i \binom{n}{k} x^k (-1)^{n-k} (y + x_i)^{n-k} \right| &\leq \sum_{n=1}^{\infty} |a_n^i| (|x| + |y + x_i|)^n \\
&= \sum_{n=1}^{\infty} |a_n^i| \big| |x| + |y + x_i| + x_i - x_i \big|^n,
\end{aligned}$$

which converges, since by construction, it holds that $|y + x_i| + x_i \in [-\tau_{i-1}, -\tau_i]$ for all $y \in [\tau_i, \tau_{i-1}]$ and hence $|x| + |y + x_i| + x_i \in U(x_i)$. So we can interchange summations in the (8.4).

Finally, using that the convergence in (8.4) is uniformly in $y \in [\tau_{i-1}, \tau_i]$ we can interchange summation and integration and get

$$
\begin{aligned}
\int_a^b \phi(x-y)f(y)\mathrm{d}y &= \sum_{i=1}^l \int_{\tau_{i-1}}^{\tau_i} \sum_{k=0}^\infty x^k \sum_{n=k}^\infty a_n^i \binom{n}{k} (-1)^{n-k} (y+x_i)^{n-k} f(y)\mathrm{d}y \\
&= \sum_{i=1}^l \sum_{k=0}^\infty x^k \underbrace{\sum_{n=k}^\infty a_n^i \binom{n}{k} (-1)^{n-k} \int_{\tau_{i-1}}^{\tau_i} (y+x_i)^{n-k} f(y)\mathrm{d}y}_{=:b_k^i},
\end{aligned}
$$

for all $x \in U(0)$, which proves the claim with $a_k = \sum_{i=1}^l b_k^i$.

$\square$

Lemma 8.2.9. *Assume that $f \in L_2([a,b])$ with $\|f\|_\infty \leq R$ (R as in Lemma 8.2.4), and that there exists a unique $t \in \mathbf{F}_k$, such that (4.7) in Remark 4.1.2 is satisfied. Furthermore, Assumption **B** holds with $s \equiv 1$, i.e. $x_{(i)} = i/n + \delta_i$. For $\hat{f}_n$, the least squares estimator from (4.1), it holds that*

$$
\|\Phi t - \Phi \hat{f}_n\|_{L_2([a,b])} = o_P(1)
$$

and

$$
\|t - \hat{f}_n\|_{L_2([a,b])} = o_P(1). \tag{8.5}
$$

Proof. By definition of $\hat{f}_n$ we obtain

$$
\|\Phi f + \varepsilon - \Phi \hat{f}_n\|_n^2 \leq \|\Phi f + \varepsilon - \Phi t\|_n^2 + o_P(n^{-1})
$$

which yields

$$
\|\Phi f - \Phi \hat{f}_n\|_n^2 \leq \|\Phi f - \Phi t\|_n^2 + 2\langle \Phi \hat{f}_n - \Phi t, \varepsilon \rangle_n + o_P(n^{-1}).
$$

Since $\hat{f}_n - t \in \mathbf{F}_k$, application of Lemma 6.2.2 gives an upper bound for the empirical process , i.e. $|\langle \Phi \hat{f}_n - \Phi t, \varepsilon \rangle_n| = o_P(1)$. So we have

$$
\|\Phi f - \Phi \hat{f}_n\|_n^2 \leq \|\Phi f - \Phi t\|_n^2 + o_p(1). \tag{8.6}
$$

Due to Corollary 8.2.6, remembering that $s_u = 1 = s_l$, in this case, we obtain from (8.6), together with the minimization property of t (4.7), that

$$
\|\Phi f - \Phi t\|_{L_2([a,b])}^2 \leq \|\Phi f - \Phi \hat{f}_n\|_{L_2([a,b])}^2 \leq \|\Phi f - \Phi t\|_{L_2([a,b])}^2 + o_P(1).
$$

Consequently we find

$$
\left| \|\Phi f - \Phi t\|_{L_2([a,b])}^2 - \|\Phi f - \Phi \hat{f}_n\|_{L_2([a,b])}^2 \right| = o_P(1). \tag{8.7}
$$

Now assume that $\Phi \hat{f}_n$ does not converge to Φt in probability. Then, we can choose a subsequence $(\Phi \hat{f}_{k_n})_{n \in \mathbb{N}}$ and $c, \delta_1 > 0$, such that $P(\|\Phi t - \Phi \hat{f}_{k_n}\|_2 \geq \delta_1) > c$ for all $n \in \mathbb{N}$. Since $\|\Phi f - \Phi t\|_2$ is the unique minimum, that means $P(\|\Phi f - \Phi \hat{f}_{k_n}\|_2 - \|\Phi f - \Phi t\|_2 > \delta_2) > c$ for some $\delta_2 > 0$. This, however, is a contradiction to (8.7), which proves the first claim, that is

$$\|\Phi t - \Phi \hat{f}_n\|_{L_2([a,b])} = \|\Phi(t - \hat{f}_n)\|_{L_2([a,b])} = o_P(1).$$

According to the proof of Lemma 4.2.7, $\Phi : \mathbf{F}_{2k} \longrightarrow \Phi(\mathbf{F}_{2k})$ is a homeomorphism. Therefore it has a continuous inverse, which yields convergence of $\hat{f}_n$ to t in probability and hence the second claim. $\qquad \square$

8.3 Native Hilbert spaces

In order to prove injectivity of the operator Φ in (1.3) under the conditions of Theorem 3.2.3 as well as for operators satisfying Assumption **C2** we use results from the theory of native Hilbert spaces. We will present the main definitions and some statements needed in the proofs of these results. To find an overview on native spaces we suggest e.g. the book of Wendland (2005) [54, Chap. 10]. A classical reference concerning reproducing kernel Hilbert spaces is furthermore Meschkowski (1962) [38] and the very detailed book of Berlinet/Thomas-Agnan (2004) [5].

Definition 8.3.1. *Let $\mathcal{H}$ be a real Hilbert space of functions $f : \Omega \to \mathbb{R}$. A function $\varphi : \Omega \times \Omega \to \mathbb{R}$ is called a **reproducing kernel** for $\mathcal{H}$ if*

$$\varphi(y, \cdot) \in \mathcal{H} \quad \text{for all } y \in \Omega$$

and

$$f(y) = \langle f, \varphi(y, \cdot) \rangle_{\mathcal{H}} \quad \text{for all } f \in \mathcal{H} \quad \text{and } y \in \Omega.$$

*If in turn a symmetric positive definite function $\varphi : \Omega \times \Omega \to \mathbb{R}$ is the reproducing kernel of a real Hilbert space $\mathcal{H}$ of real valued functions in Ω, then $\mathcal{H}$ is called the **native space** of φ.*

Lemma 8.3.2. *Every positive definite function φ on a domain Ω has a unique native space $\mathcal{N}_\varphi(\Omega)$. It is the closure of the space*

$$\mathcal{F}_\varphi(\Omega) := \left\{ \sum_{i=1}^{M} \alpha_i \varphi(\cdot, x_i) \mid \alpha_i \in \mathbb{R}, x_i \in \Omega, M \in \mathbb{N} \right\},$$

equipped with the inner product

$$\left\langle \sum_{j=1}^{N} \alpha_j \varphi(\cdot, x_j), \sum_{k=1}^{M} \beta_j \varphi(\cdot, y_k) \right\rangle_{\mathcal{N}_\varphi} := \sum_{j=1}^{N} \sum_{k=1}^{M} \varphi(x_j, y_k).$$

The elements of $\mathcal{N}_\varphi(\Omega)$ can be interpreted as functions by

$$f(x) = \langle f, \varphi(\cdot, x)\rangle_{\mathcal{N}_\varphi}$$

Proof. [48, Thm. 8]. □

The next Lemma draws a connection between the native space of a kernel φ and the image of the corresponding integral operator $\Phi : L_2(\Omega) \to L_2(\Omega)$, with compact $\Omega \subset \mathbb{R}^d$, defined by

$$\Phi v(x) := \int_\Omega \varphi(x, y)v(y)\mathrm{d}y, \qquad v \in L_2(\Omega), \quad x \in \Omega.$$

Therefore, we use the following definition.

Definition 8.3.3. *A continuous kernel $\varphi : \Omega \times \Omega \to \mathbb{C}$ is called **positive definite** on $\Omega \subseteq \mathbb{R}^d$ if for all $N \in \mathbb{N}$, all pairwise distinct $X = \{x_1, ..., x_N\} \subseteq \Omega$, and all $\alpha \in \mathbb{C} \setminus \{0\}$ we have*

$$\sum_{j=1}^N \sum_{k=1}^N \alpha_j \bar{\alpha}_k \varphi(x_j, x_k) > 0.$$

Lemma 8.3.4. *Suppose that φ is a continuous, symmetric and positive definite kernel defined on a compact set $\Omega \subset \mathbb{R}^d$. Then the integral operator Φ maps $L_2(\Omega)$ continuously into the native space $\mathcal{N}_\varphi(\Omega)$. It is the adjoint of the embedding operator of the native space $\mathcal{N}_\varphi(\Omega)$ into $L_2(\Omega)$, i.e. it satisfies*

$$(f, v)_{L_2(\Omega)} = (f, \Phi v)_{\mathcal{N}_\varphi(\Omega)}, \qquad f \in \mathcal{N}_\varphi(\Omega), \quad v \in L_2(\Omega).$$

The range of Φ is dense in $\mathcal{N}_\varphi(\Omega)$.

Proof. [54, Prop. 10.28]. □

Lemma 8.3.5. *Suppose that $\phi \in L_1(\mathbb{R}^d) \cap C(\mathbb{R}^d)$ satisfies*

$$c_1(1 + |x|_2^2)^{-s} \leq |\hat{\phi}(x)| \leq c_2(1 + |x|_2^2)^{-s}, \qquad x \in \mathbb{R}^d$$

with $s > d/2$ and two positive constants $c_1 \leq c_2$. Then the native space $\mathcal{N}_\varphi(\mathbb{R}^d)$ corresponding to $\varphi(x, y) = \phi(x - y)$ coincides with the Sobolev space $H^s(\mathbb{R}^d)$, and the native space norm and the Sobolev norm are equivalent.

Proof. [54, Cor. 10.13]. □

Remark 8.3.1. For $d = 1$, Lemma 8.3.5 together with [36, Thm. 9.1] implies, that under the same decaying assumptions on the Fourier transform as in Lemma 8.3.5, it follows that $\mathcal{N}_\varphi([a, b]) = H^s([a, b])$ with equivalent norms.

8.4 Empirical Process Theory

In this section we introduce the main results from empirical process theory as used in the proofs of Lemma 6.2.2 and 4.2.5. There is a large amount of literature on this theory especially on inequalities of the type of 8.4.1. Some references are Pollard [43], van der Vaart and Wellner [52], van der Vaart [53], van de Geer [51], [17], and Devroye and Lugosi [19] to mention a few.

The cited results are taken from van de Geer [51]. Note that the error condition used in that book is weaker than Assumption **A** and **A1**, since it claims existence of $0 < C_0,\ \sigma_0 < \infty$ such that

$$\lim_{n \to \infty} \max_{i=1...n} C_0^2 \mathbf{E}(e^{\varepsilon_i^2/C_0^2}) \le \sigma_0^2.$$

In the following theorems $\mathcal{G}$ denotes some arbitrary function space.

Lemma 8.4.1. *Suppose that Assumptions **A** and **A1** are satisfied. Assume further that $\mathcal{G}$ is a class of functions, with $\sup_{g \in \mathcal{G}} \|g\|_n \le R$ and there exists a constant c depending on Assumption **A1** only, such that for all $\delta > 0$ satisfying*

$$\sqrt{n}\delta \ge c\left(\int_0^R \boldsymbol{H}^{1/2}(u, \mathcal{G}, Q_n) du \vee R \right) \tag{8.8}$$

we have that

$$P\left(\sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i g(x_i) \right| \ge \delta \right) \le C e^{-\frac{n\delta^2}{C^2 R^2}}. \tag{8.9}$$

Proof. See [51, Lem. 3.2]. $\qquad\square$

Lemma 8.4.2. *Assume that $\varepsilon_1, ..., \varepsilon_n$ are i.i.d. with mean zero and $\boldsymbol{E}(\varepsilon_1^2) = \sigma^2 < \infty$. Set $\mathcal{G}_n(R) = \{ g \in \mathcal{G} : \|g\|_n \le R \}$ and suppose that*

$$\frac{1}{n}\boldsymbol{H}(\delta, \mathcal{G}_n(R), P_n) \longrightarrow 0 \quad \text{for all} \quad \delta > 0,\ R > 0.$$

Then

$$\sup_{g \in \mathcal{G}_n(R)} |\langle \varepsilon, g \rangle_n| = \sup_{g \in \mathcal{G}_n(R)} \left| \sum_{i=1}^n \varepsilon_i g(x_i) \right| = o_P(1)$$

for every $R > 0$.

Proof. This follows directly from the proof of [51, Thm. 4.8]. $\qquad\square$

Bibliography

[1] G. G. Agarwal and W. J. Studden. Asymptotic integrated mean square error using least squares and bias minimizing splines. *Ann. Statist.*, 8(6):1307–1325, 1980.

[2] H. Amann. Über die Existenz und iterative Berechnung einer Lösung der Hammerstein'schen Gleichung. *Aequationes Math.*, 1:242–266, 1968.

[3] H. Amann. Über die Existenz und Eindeutigkeit einer Lösung der Hammerstein'schen Gleichung in Banachräumen. *J. Math. Mech.*, 19:143–153, 1969/1970.

[4] H. Bauer. *Maß- und Integrationstheorie*. de Gruyter Lehrbuch. [de Gruyter Textbook]. Walter de Gruyter & Co., Berlin, second edition, 1992.

[5] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic Publishers, Boston, MA, 2004. With a preface by Persi Diaconis.

[6] O. V. Besov. Investigation of a class of function spaces in connection with imbedding and extension theorems. *Trudy. Mat. Inst. Steklov.*, 60:42–81, 1961.

[7] P. Borwein and T. Erdélyi. *Polynomials and polynomial inequalities*, volume 161 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1995.

[8] P. Borwein and T. Erdélyi. Generalizations of Müntz's theorem via a Remez-type inequality for Müntz spaces. *J. Amer. Math. Soc.*, 10(2):327–349, 1997.

[9] L. Boysen, S. B., and A. Munk. Jump estimation in inverse regression. *Electron. J. Statist.*, 3:1322–1359, 2009.

[10] L. Boysen, A. Kempe, V. Liebscher, A. Munk, and O. Wittich. Consistencies and rates of convergence of jump-penalized least squares estimators. *Ann. Statist.*, 37(1):157–183, 2009.

[11] D. Braess. *Nonlinear approximation theory*, volume 7 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1986.

[12] H. G. Burchard. Splines (with optimal knots) are better. *Applicable Anal.*, 3:309–319, 1973/74.

[13] P. L. Butzer and K. Scherer. Jackson and Bernstein-type inequalities for families of commutative operators in Banach spaces. *J. Approximation Theory*, 5:308–342, 1972. Collection of articles dedicated to J. L. Walsh on his 75th birthday, III.

[14] H. Cardot. Spatially adaptive splines for statistical linear inverse problems. *J. Multivariate Anal.*, 81(1):100–119, 2002.

[15] C. De Boor. *A practical guide to splines*, volume 27 of *Applied Mathematical Sciences*. Springer-Verlag, New York, revised edition, 2001.

[16] C. De Boor and J. R. Rice. Least squares cubic spline approximation i - fixed knots. Technical report, CSD TR 21, Purdue University Report, 1968.

[17] E. del Barrio, P. Deheuvels, and S. van de Geer. *Lectures on empirical processes*. EMS Series of Lectures in Mathematics. European Mathematical Society (EMS), Zürich, 2007. Theory and statistical applications, With a preface by Juan A. Cuesta Albertos and Carlos Matrán.

[18] R. A. DeVore and G. G. Lorentz. *Constructive approximation*, volume 303 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1993.

[19] L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer Series in Statistics. Springer-Verlag, New York, 2001.

[20] J. Elstrodt. *Maß- und Integrationstheorie*. Springer-Lehrbuch. [Springer Textbook]. Springer-Verlag, Berlin, fourth edition, 2005. Grundwissen Mathematik. [Basic Knowledge in Mathematics].

[21] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1996.

[22] L. C. Evans and R. F. Gariepy. *Measure theory and fine properties of functions*. Studies in Advanced Mathematics. CRC Press, Boca Raton, FL, 1992.

[23] F. Faraci and V. Moroz. Solutions of Hammerstein integral equations via a variational principle. *J. Integral Equations Appl.*, 15(4):385–402, 2003.

[24] P. Feder. On asymptotic distribution theory in segmented regression problems - identified case. *Ann. Statist*, 3:49–83, 1975.

[25] E. G. Fergyanto, H. Hiroomi, and K. Yasuhiro. Two-step b-splines regularization method for solving an ill-posed problem of impact-force reconstruction. *Journal of Sound and Vibration*, 297(1-2):200 – 214, 2006.

[26] J. D. Ferry. *Viscoelastic properties of polymers [by] John D. Ferry.* Wiley New York,, 2d ed. edition, 1970.

[27] A. Goldenshluger, A. Tsybakov, and A. Zeevi. Optimal change-point estimation from indirect observations. *Ann. Statist.*, 34(1):350–372, 2006.

[28] A. I. Grebénnikov. Spline-regularization method for solving inverse problems: theory and algorithms. *Mat. Model.*, 14(5):35–50, 2002. Second International Conference OFEA'2001 "Optimization of Finite Element Approximation, Splines and Wavelets" (Russian) (St. Petersburg, 2001).

[29] J. Hadamard. *Lectures on Cauchy's problem in linear partial differential equations.* New Haven, 1923.

[30] A. Hammerstein. Nichtlineare Integralgleichungen nebst Anwendungen. *Acta Math.*, 54(1):117–176, 1930.

[31] H. Heuser. *Lehrbuch der Analysis. Teil 2.* Mathematische Leitfäden. [Mathematical Textbooks]. B. G. Teubner, Stuttgart, sixth edition, 1991.

[32] D. V. Hinkley. Inference about the intersection in two-phase regression. *Biometrika*, 56(3):495–504, 1969.

[33] T. Hohage. Lecture notes on inverse problems. Georg-August-Universität Göttingen.

[34] D. L. B. Jupp. Approximation to data by splines with free knots. *SIAM J. Numer. Anal.*, 15(2):328–343, 1978.

[35] R. Kress. *Linear integral equations*, volume 82 of *Applied Mathematical Sciences*. Springer-Verlag, New York, second edition, 1999.

[36] J.-L. Lions and E. Magenes. *Non-homogeneous boundary value problems and applications. Vol. I.* Springer-Verlag, New York, 1972. Translated from the French by P. Kenneth, Die Grundlehren der mathematischen Wissenschaften, Band 181.

[37] G. G. Lorentz. Metric entropy and approximation. *Bull. Amer. Math. Soc.*, 72:903–937, 1966.

[38] H. Meschkowski. *Hilbertsche Räume mit Kernfunktion.* Die Grundlehren der mathematischen Wissenschaften, Bd. 113. Springer-Verlag, Berlin, 1962.

[39] R. G. Miller, Jr. *Simultaneous statistical inference.* McGraw-Hill Book Co., New York, 1966.

[40] M. H. Neumann. Optimal change-point estimation in inverse problems. *Scand. J. Statist.*, 24(4):503–521, 1997.

[41] F. O'Sullivan. A statistical perspective on ill-posed inverse problems. *Statist. Sci.*, 1(4):502–527, 1986. With comments and a rejoinder by the author.

[42] J. Peetre and G. Sparr. Interpolation of normed abelian groups. *Ann. Mat. Pura Appl. (4)*, 92:217–262, 1972.

[43] D. Pollard. *Some generalized order-disorder transitions*. Springer Series in Statistics. Springer-Verlag, New York, 1984.

[44] R. Potts. Some generalized order-disorder transitions. *Proc. Camb. Phil. Soc.*, 48:106–109, 1952.

[45] J. Prince E. Rouse. A theory of the linear viscoelastic properties of dilute solutions of coiling polymers. *The Journal of Chemical Physics*, 21(7):1272–1280, 1953.

[46] J. R. Rice. *The approximation of functions. Vol. 2: Nonlinear and multivariate theory*. Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont., 1969.

[47] T. Roths, D. Maier, C. Friedrich, M. Marth, and J. Honerkamp. Determination of the relaxation time spectrum from dynamic moduli using an edge preserving regularization method. *Rheologica Acta*, 2000.

[48] R. Schaback. Native Hilbert spaces for radial basis functions. I. In *New developments in approximation theory (Dortmund, 1998)*, volume 132 of *Internat. Ser. Numer. Math.*, pages 255–282. Birkhäuser, Basel, 1999.

[49] O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, and F. Lenzen. *Variational methods in imaging*, volume 167 of *Applied Mathematical Sciences*. Springer, New York, 2009.

[50] H. R. Thieme. On a class of Hammerstein integral equations. *Manuscripta Math.*, 29(1):49–84, 1979.

[51] S. A. van de Geer. *Applications of empirical process theory*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2000.

[52] A. Van der Vaart and J. Wellner. *Weak Convergence an Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

[53] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[54] H. Wendland. *Scattered data approximation*, volume 17 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2005.

Curriculum Vitae

Sophie Hanna Clara Frick geb. Bruns.

Born 30. October 1979 in Detmold.
Married, 1 Child, German.

08/1986 – 06/1999	Schooling *Abitur* at the *Gymnasium Leopoldinum Detmold*.
10/1999 – 03/2001	Graduate studies in mathematics *Faculty of Mathematics, University of Göttingen*.
04/2001	"Vordiplom".
04/2001 – 09/2006	Graduate studies in theology *Faculty of Theology, University of Göttingen* and mathematics *Faculty of Mathematics, University of Göttingen*.
06/2006	"Erste Staatsprüfung für das Lehramt an Gymnasien". Examination thesis: *Gruppeneigenschaften von elliptischen Kurven* supervised by *Prof. Dr. Ina Kersten*.
10/2006 – 12/2010	Post graduate studies in mathematics *Institute for Mathematical Stochastics,* *University of Göttingen* supervised by *Prof. Dr. Axel Munk*.
10/2006 – 09/2007	Research assistent *Institute for Mathematical Stochastics,* *University of Göttingen*.
07/2007 – 12/2010	Member of the DFG Graduiertenkolleg 1023 *Identifikation in mathematischen Modellen:* *Synergie stochastischer und numerischer Modelle*.
07/2010 – 10/2010	Parental leave.