UCL Université catholique de Louvain

Institut de statistique, biostatistique
et sciences actuarielles

LINEAR INVERSE PROBLEMS WITH NOISE IN THE OPERATOR:
Minimax-optimal estimation and adaptation

Jan JOHANNES

Syllabus of the lecture course LSTAT3130

Special Topics in Mathematical Statistic

Academic year 2012/2013

last update: September 17, 2014

If you find errors, please send a short note by email to: jan.johannes@uclouvain.be

Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgique

Phone: +32 10 47 34 79 – Fax: +32 10 47 30 32

Email: jan.johannes@uclouvain.be

Table of contents

1 Introduction **1**

2 Theoretical basics and terminology **5**

2.1 Hilbert spaces . . . 5
2.2 Operators on Hilbert spaces . . . 7
2.3 Compact operator . . . 11
2.4 Spectrum of an operator . . . 12
2.5 Eigenvalue and singular value decomposition . . . 13
2.6 Spectral theorem . . . 14
2.7 Functional Calculus . . . 15
2.8 Statistical ill-posed inverse problems . . . 16

3 Minimax optimality and adaptive estimation **21**

3.1 Minimax theory . . . 21
3.2 Deriving a lower bound . . . 23
3.3 Estimation by dimension reduction . . . 29
3.4 Data-driven estimation procedures . . . 30
3.4.1 Lepski's method . . . 31
3.4.2 Model selection by contrast minimisation . . . 35
3.4.3 Combining model selection and Lepski's method . . . 38

4 Gaussian inverse regression **43**

4.1 Gaussian indirect sequence space model . . . 43
4.1.1 Lower bounds . . . 44
4.1.2 Minimax optimal estimation . . . 45
4.1.3 Adaptive estimation . . . 46
4.2 Gaussian indirect regression . . . 47
4.2.1 Lower bounds . . . 48
4.2.2 Minimax optimal estimation . . . 49
4.2.3 Adaptive estimation . . . 50
4.3 Gaussian indirect sequence space model with noise in the operator . . . 50
4.4 Gaussian inverse regression with noise in the operator . . . 50

Chapter 1

Introduction

SHORT SUMMARY

Statistical ill-posed inverse problems are becoming increasingly important in a diverse range of disciplines, including geophysics, astronomy, medicine and economics. Roughly speaking, in all of these applications the observable signal $g = Tf$ is a transformation of the functional parameter of interest f under a linear operator T. Statistical inference on f based on an estimation of g which usually requires an inversion of T is thus called an *inverse problem*. The lecture course focuses on statistical ill-posed inverse problems with noise in the operator where neither the signal g nor the linear operator T are known in advance, although they can be estimated from the data. Our objective in this context is the construction of minimax-optimal fully data-driven estimation procedures of the unknown function f. Special attention is given to four models and their extensions, namely Gaussian inverse regression, density deconvolution, functional linear regression and nonparametric instrumental regression, which lead naturally to statistical ill-posed inverse problems with noise in the operator.

APPLICATIONS

Density deconvolution with unknown error distribution. The biologist who is interested in the density f_X of a gene-expression intensity X, can record in a cDNA microarray the expressed gene intensity X only corrupted by the intensity of a background noise U, that is $Y = X+U$. If the additive measurement error U is independent of X then the density $f_Y = f_X \star f_U$ of Y equals the convolution of f_X and the error density f_U. Consequently, recovering f_X from the estimated density $f_Y = C_{f_U} f_X$ of Y is an inverse problem where C_{f_U} is the convolution operator defined by the error density f_U. In this situation, the density f_X of the random variable X has to be estimated nonparametrically based on an iid. sample from a noisy observation Y of X which is called a density deconvolution problem. There is a vast literature on deconvolution with known error density which leads to a statistical ill-posed inverse problem with known operator. On the other hand, if the error density f_U is estimated from an additional calibration sample of the error U then the deconvolution problem corresponds to a statistical ill-posed inverse problem with noise in the operator. *Functional linear regression.* In climatology, prediction of level of ozone pollution based on continuous measurements of pollutant indicators is often modelled by a functional linear model. In this context a scalar response Y (i.e. the ozone concentration) is modelled in dependence of a random function X (i.e. the daily concentration curve of a pollutant indicator). Typically the dependence is assumed to be linear which finds its expression in a linear normal equation $c_{YX} = \Gamma_{XX}\beta$ where c_{YX} is the cross-correlation between Y and X, and Γ_{XX} is the covariance operator associated to the indicator X. Note that both the cross-correlation function c_{YX} and the covariance operator Γ_{XX} need to be estimated in practice. Consequently, the nonparametric estimation of the functional slope parameter β based on an iid. sample from (Y, X) leads to a statistical ill-posed inverse problem with noise in the operator. *Nonparametric instrumental regression.* An econometrician who wants to analyse an economic relation between a response Y and an endogenous vector X of explanatory variables, might incorporate a vector of exogenous instruments Z. This situation is usually treated by considering a conditional moment equation $r_{Y|Z} = K_{X|Z}\,\varphi$ where $r_{Y|Z} = \mathbb{E}[Y|Z]$ is the conditional expectation function of Y given Z and $K_{X|Z}$ is the conditional expectation operator of X given Z. As these are unknown in practice, inference on φ based on an iid. sample from (Y, X, Z) is a statistical ill-posed inverse problem with noise in the operator.

STATISTICAL ILL-POSED INVERSE PROBLEMS

We study nonparametric estimation of the functional parameter of interest f in an inverse problem, that is, its reconstruction based on an estimation of a linear transformation $g = Tf$. It is important to note that in all the applications discussed above both the signal g and the inherent transformation T are unknown in practice, although they can be estimated from the data. The estimated signal $\widehat{g}_\varepsilon$ and operator $\widehat{T}_\sigma$ respectively given by

$$\widehat{g}_\varepsilon = Tf + \sqrt{\varepsilon}\,\dot{W} \quad \text{and} \quad \widehat{T}_\sigma = T + \sqrt{\sigma}\,\dot{B}. \tag{1.1}$$

are noisy versions of g and T contaminated by additive random errors $\dot{W}$ and $\dot{B}$ with respective noise levels ε and σ. Consequently, a statistical inference on the functional parameter of interest f has to take into account that a random noise is present in both the estimated signal $\dot{W}$ and the estimated operator $\dot{B}$. *Gaussian inverse regression with noise in the operator.* A particularly interesting situation is given by model (1.1) where the random error $\dot{W}$ and $\dot{B}$ are independent Gaussian white noises. This model is particularly useful to characterise the influence of an *a priori* knowledge of the operator T. To this end we will compare three cases: First, the operator T is *fully known* in advance, i.e., the noise level σ is equal to zero. Second, it is *partially known*, that is, the eigenfunctions of T are known in advance but the "observed" eigenvalues of T are contaminated with an additive Gaussian error. Third, the operator T is *unknown*.

MINIMAX-OPTIMAL ESTIMATION

Typical questions in this context are the nonparametric estimation of the functional parameter f on an interval or in a given point, referred to as global or local estimation, respectively. However, these are special cases in a general framework where the accuracy of an estimator $\widehat{f}$ of f given the estimations (1.1) is measured by a distance $\mathfrak{D}_{\text{ist}}(\widehat{f}, f)$. A suitable choice of the distance covers than the global as well as the local estimation problem. Moreover, we call the quantity $\mathbb{E}[\mathfrak{D}^2_{\text{ist}}(\widehat{f}, f)]$ risk of the estimator $\widehat{f}$ of f. It is well-known that in terms of its risk the attainable accuracy of an estimation procedure is essentially determined by the conditions imposed on f and the operator T. Typically, these conditions are expressed in the form $f \in \mathcal{F}$ and $T \in \mathcal{T}$ for suitable chosen classes $\mathcal{F}$ and $\mathcal{T}$. The class $\mathcal{F}$ reflects prior information on the solution f, e.g., its level of smoothness, and the class $\mathcal{T}$ imposes among others conditions on the decay of the eigenvalues of the operator T. The accuracy of $\widehat{f}$ is hence measured by its maximal risk over the classes $\mathcal{F}$ and $\mathcal{T}$, that is,

$$\mathcal{R}_{\varepsilon,\sigma}\big[\widehat{f}\,|\,\mathcal{F}, \mathcal{T}\big] := \sup_{f\in\mathcal{F}}\sup_{T\in\mathcal{T}} \mathbb{E}[\mathfrak{D}^2_{\text{ist}}(\widehat{f}, f)]$$

Moreover, $\widehat{f}$ is called minimax-optimal up to a finite positive constant C if $\mathcal{R}_{\varepsilon,\sigma}\big[\widehat{f}\,|\,\mathcal{F}, \mathcal{T}\big] \leqslant C \inf_{\widetilde{f}} \mathcal{R}_{\varepsilon,\sigma}\big[\widetilde{f}\,|\,\mathcal{F}, \mathcal{T}\big]$ where the infimum is taken over all possible estimators of f. Consequently, minimax-optimality of an estimator $\widehat{f}$ based on observations (1.1) is usually shown by establishing both an upper and a lower bound. More precisely, we search a finite positive constant $\mathcal{R}^\star_\varepsilon(\mathcal{F},\mathcal{T})$ depending only on the noise levels and the classes such that

$$\mathcal{R}_{\varepsilon,\sigma}\big[\widehat{f}\,|\,\mathcal{F},\mathcal{T}\big] \leqslant C_1\mathcal{R}^\star_{\varepsilon,\sigma}\big(\mathcal{F},\mathcal{T}\big) \quad \text{and} \quad \mathcal{R}^\star_{\varepsilon,\sigma}\big(\mathcal{F},\mathcal{T}\big) \leqslant C_2 \inf_{\widetilde{f}} \mathcal{R}_{\varepsilon,\sigma}\big[\widetilde{f}\,|\,\mathcal{F},\mathcal{T}\big]$$

where C_1, C_2 are finite positive constants independent of the noise levels. Moreover, the quantity $\mathcal{R}^\star_{\varepsilon,\sigma}\big(\mathcal{F},\mathcal{T}\big)$ is called the minimax-optimal rate of convergence over the classes $\mathcal{F}$ and $\mathcal{T}$ if it tends to zero as ε and σ tend to zero.

ADAPTIVE ESTIMATION

In many cases the proposed estimation procedures rely on the choice of at least one tuning parameter, which in turn, crucially influences the attainable accuracy of the constructed estimator. In other words, these estimation procedures can attain the minimax rate $\mathcal{R}^\star_{\varepsilon,\sigma}\big(\mathcal{F},\mathcal{T}\big)$ over the classes $\mathcal{F}$ and $\mathcal{T}$ only if the inherent tuning parameters are chosen optimally. This optimal choice, however, follows often from a classical squared-bias-variance compromise and requires a *a priori* knowledge about the classes $\mathcal{F}$ and $\mathcal{T}$, which is usually inaccessible in practice. This motivates its data-driven choice in the context of nonparametric statistics since its very beginning in the fifties of the last century. A demanding challenge is then a fully data driven method to select the tuning parameters in such a way that the resulting data-driven estimator of f still attains the minimax-rate up to a constant over a variety of classes $\mathcal{F}$ and $\mathcal{T}$. The fully data driven estimation procedure is then called *adaptive*.

Chapter 2

Theoretical basics and terminology

2.1 Hilbert spaces

For a detailed and extensive survey on functional analysis we refer the reader, for example, to the series of textbooks by Dunford and Schwartz [1988a,b,c].

§2.1.1 **Hilbert space**. *A normed vector space* $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ *over* $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ *that is complete (in a Cauchy-sense) is called a (real or complex) Hilbert space if there exists an inner product* $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ *on* $\mathbb{H} \times \mathbb{H}$ *with* $\langle h, h \rangle_{\mathbb{H}}^{1/2} = \|h\|_{\mathbb{H}}$ *for all* $h \in \mathbb{H}$.

§2.1.2 **Cauchy-Schwarz inequality**. $|\langle h_1, h_2 \rangle_{\mathbb{H}}|^2 \leqslant \langle h_1, h_1 \rangle_{\mathbb{H}} \cdot \langle h_2, h_2 \rangle_{\mathbb{H}}$ *for all* $h_1, h_2 \in \mathbb{H}$.

§2.1.3 **Examples**. (i) The *Euclidean space* $\mathbb{R}^k$ endowed with the Euclidean inner product $\langle x, y \rangle := y^t x$ and the induced euclidean norm $\|x\| = (x^t x)^{1/2}$ for all $x, y \in \mathbb{R}^k$ is a Hilbert space. More generally, given a strictly positive definite $(k \times k)$-matrix W, $\mathbb{R}^k$ endowed with the weighted inner product $\langle x, y \rangle_W := y^t W x$ for all $x, y \in \mathbb{R}^k$ is also a Hilbert space.

(ii) Given $\mathcal{J} \subset \mathbb{N}$, consider $\ell^2 := \left\{ (x_j)_{j \in \mathcal{J}}, x_j \in \mathbb{K}, \sum_{j \in \mathcal{J}} x_j^2 < \infty \right\}$ where we refer to any sequence $(x_j)_{j \in \mathcal{J}}$ as a whole by omitting its index as for example in «the sequence x» and arithmetic operations on sequences are defined element-wise. The *space of square summable sequences* ℓ^2 endowed with the inner product $\langle x, y \rangle_{\ell^2} := \sum_{j \in \mathcal{J}} y_j \overline{x}_j$ for all $x, y \in \ell^2$ is a Hilbert space.

(iii) Given a strictly positive sequence $\mathfrak{v}$ consider the *weighted norm* $\|x\|_{\mathfrak{v}}^2 := \sum_{j \in \mathcal{J}} \mathfrak{v}_j x_j^2$, we define $\ell^2_{\mathfrak{v}}$ as the completion of ℓ^2 with respect to $\|\cdot\|_{\mathfrak{v}}$ which is a Hilbert space endowed with the inner product $\langle x, y \rangle_{\mathfrak{v}} := \sum_{j \in \mathcal{J}} \mathfrak{v}_j y_j \overline{x}_j$ for all $x, y \in \ell^2_{\mathfrak{v}}$.

(iv) A Hilbert space is given by the *space of* $\mathbb{K}$*-valued functions* defined on $\Omega \subset \mathbb{R}^k$ and *square integrable* with respect to (wrt.) a measure μ: $L^2(\Omega, \mu) := \left\{ f : \Omega \to \mathbb{K}, \int_\Omega |f|^2 d\mu < \infty \right\}$, or in case of the Lebesgue measure $L^2(\Omega)$ for short, which is endowed with the inner product $\langle g, f \rangle_{L^2(\Omega, \mu)} := \int_\Omega f \overline{g} d\mu$ for all $f, g \in L^2(\Omega, \mu)$.

(v) If X is a $\mathbb{R}^k$-valued random variable, then $L^2_X := \left\{ f : \mathbb{R}^k \to \mathbb{R},\ \mathbb{E}|f(X)|^2 < \infty \right\}$ is a Hilbert space endowed with the inner product $\langle f, g \rangle_{L^2_X} := \mathbb{E}[f(X) g(X)]$ for all $f, g \in L^2_X$. □

§2.1.4 **Orthonormal basis**. *A subset* $\mathcal{U}$ *of a Hilbert space* $(\mathbb{H}, \langle \cdot, \cdot \rangle_{\mathbb{H}})$ *is called orthogonal if*

$$\forall u_1, u_2 \in \mathcal{U},\ u_1 \neq u_2 :\ \langle u_1, u_2 \rangle_{\mathbb{H}} = 0$$

and orthonormal system (ONS) if in addition $\|u\|_{\mathbb{H}} = 1$, $\forall u \in \mathbb{U}$. *We say* $\mathcal{U}$ *is an orthonormal basis (ONB) if*

$$\mathcal{U} \subset \mathcal{U}', \quad \mathcal{U}' \text{ is ONS} \quad \Rightarrow \mathcal{U} = \mathcal{U}',$$

i.e., if it is a complete *ONS.*

§2.1.5 **Properties.**

(Pythagorean formula) If $h_1, \dots, h_n \in \mathbb{H}$ are orthogonal, then $\|\sum_{j=1}^n h_j\|_{\mathbb{H}}^2 = \sum_{j=1}^n \|h_j\|_{\mathbb{H}}^2$.

(Bessel's inequality) If $\mathcal{U} \subset \mathbb{H}$ is an ONS, then $\|h\|_{\mathbb{H}}^2 \geqslant \sum_{u \in \mathcal{U}} \langle h, u\rangle_{\mathbb{H}}^2$ for all $h \in \mathbb{H}$.

(Parseval's formula) An ONS $\mathcal{U} \subset \mathbb{H}$ is complete (i.e. ONB) if and only if $\|h\|_{\mathbb{H}}^2 = \sum_{u \in \mathcal{U}} \langle h, u\rangle_{\mathbb{H}}^2$ for all $h \in \mathbb{H}$.

§2.1.6 **Definition.** *Let $\mathcal{U}$ be subset of a Hilbert space $(\mathbb{H}, \langle\cdot,\cdot\rangle_{\mathbb{H}})$, we denote by $\overline{\mathrm{lin}}(\mathcal{U})$ the closure of the linear subspace spanned by the elements of $\mathcal{U}$ and its orthogonal complement in $(\mathbb{H}, \langle\cdot,\cdot\rangle_{\mathbb{H}})$ by $\overline{\mathrm{lin}}(\mathcal{U})^{\perp} := \{h \in \mathbb{H} : \langle h, u\rangle_{\mathbb{H}} = 0, \forall u \in \overline{\mathrm{lin}}(\mathcal{U})\}$ where $\mathbb{H} = \overline{\mathrm{lin}}(\mathcal{U}) \oplus \overline{\mathrm{lin}}(\mathcal{U})^{\perp}$.*

§2.1.7 ***Remark.*** If $\mathcal{U} \subset \mathbb{H}$ is an ONS, then there exists an ONS $\mathcal{V} \subset \mathbb{H}$ such that $\mathbb{H} = \overline{\mathrm{lin}}(\mathcal{U}) \oplus \overline{\mathrm{lin}}(\mathcal{V})$ and for all $h \in \mathbb{H}$ it holds $h = \sum_{u \in \mathcal{U}} \langle h, u\rangle_{\mathbb{H}} u + \sum_{v \in \mathcal{V}} \langle h, v\rangle_{\mathbb{H}} v$ (in a L^2-sense). In particular, if $\mathcal{U}$ is an ONB then $h = \sum_{u \in \mathcal{U}} \langle h, u\rangle_{\mathbb{H}} u$ for all $h \in \mathbb{H}$. □

§2.1.8 **Separable Hilbert space.** *A sequence $(u_j)_{j \in \mathcal{J}}$ in $\mathbb{H}$ is said to be orthonormal and complete (i.e. orthonormal basis) if the subset $\{u_j, j \in \mathcal{J}\}$ is a complete ONS (i.e. ONB). The Hilbert space $\mathbb{H}$ is called separable, if there exists a complete orthonormal sequence.*

§2.1.9 **Examples.** The Hilbert space $(\mathbb{R}^k, \langle\cdot,\cdot\rangle_{\mathfrak{v}})$, $(\ell^2_{\mathfrak{v}}, \langle\cdot,\cdot\rangle_{\mathfrak{v}})$ and $(L^2_{\mathfrak{v}}(\Omega, \mu), \langle\cdot,\cdot\rangle_{\mathfrak{v}})$ with σ-finite measure μ are separable. On the contrary, given $\lambda \in \mathbb{R}$ define the function $f_\lambda : \mathbb{R} \to \mathbb{C}$ with $f_\lambda(x) := e^{i\lambda x}$ and set $\mathcal{H} = \overline{\mathrm{lin}}\{f_\lambda, \lambda \in \mathbb{R}\}$. Observe that $\langle f, g\rangle = \lim_{t\to\infty} \frac{1}{2t} \int_{-t}^{t} f(s) g(s) ds$ defines an inner product on $\mathcal{H}$. The completion of $\mathcal{H}$ with respect to the induced norm $\|f\| = \langle g, g\rangle^{1/2}$ is a Hilbert space which is not separable, since $\|f_\lambda - f_{\lambda'}\| = \sqrt{2}$ for all $\lambda \neq \lambda'$. □

§2.1.10 ***Notations.*** Let $(u_j)_{j \in \mathcal{J}}$ be an ONS with $\mathbb{U} := \overline{\mathrm{lin}}\{u_j, j \in \mathcal{J}\} \subset \mathbb{H}$. For $f \in \mathbb{U}$ we denote by $[f] := ([f]_j)_{j \in \mathcal{J}}$ the sequence of generalised Fourier coefficients $[f]_j := \langle f, u_j\rangle_{\mathbb{H}}$. Given a strictly positive sequence of weights $\mathfrak{v}$, we define a weighted norm $\|\cdot\|_{\mathfrak{v}}$ by $\|f\|_{\mathfrak{v}}^2 := \langle \mathfrak{v}, [f]^2\rangle_{\ell^2} = \sum_{j \in \mathcal{J}} \mathfrak{v}_j [f]_j^2$ and denote by $\mathbb{U}_{\mathfrak{v}}$ the completion of $\mathbb{U}$ with respect to $\|\cdot\|_{\mathfrak{v}}$. If $(u_j)_{j \in \mathcal{J}}$ is complete in $\mathbb{H}$ then let $\mathbb{H}_{\mathfrak{v}}$ be the completion of $\mathbb{H}$ with respect to $\|\cdot\|_{\mathfrak{v}}$. □

§2.1.11 **Class of solutions.** *Given a strictly positive sequence of weights $\mathfrak{a} = (\mathfrak{a}_j)_{j \in \mathcal{J}}$ and an ONB $(u_j)_{j \in \mathcal{J}}$ in $\mathbb{H}$ consider the associated weighted norm $\|\cdot\|_{\mathfrak{a}}$ and the completion $\mathbb{H}_{\mathfrak{a}}$ of $\mathbb{H}$. Let $r > 0$ be a constant. We assume in the following, that the solution f belongs to the ellipsoid*

$$\mathcal{F}_{\mathfrak{a}}^r := \{f \in \mathbb{H}_{\mathfrak{a}} : \|f\|_{\mathfrak{a}}^2 \leqslant r\}.$$

§2.1.12 **Examples.** (i) Consider the real Hilbert space $L^2([0,1])$, then the *trigonometric basis* $\{\psi_j, j \in \mathbb{N}\}$ given for $t \in [0,1]$ by

$$\psi_1(t) := 1, \; \psi_{2k}(t) := \sqrt{2}\cos(2\pi k t), \; \psi_{2k+1}(t) := \sqrt{2}\sin(2\pi k t), \; k = 1, 2, \dots,$$

is orthonormal and complete, i.e. an ONB. Define further a weighted norm $\|\cdot\|_{\mathfrak{v}}$ with respect to the trigonometric basis, that is, $\|h\|_{\mathfrak{v}} := \sum_{j \in \mathbb{N}} \mathfrak{v}_j |\langle h, \psi_j\rangle|^2$. If we set $\mathfrak{v}_1 = 1$, $\mathfrak{v}_{2k} = \mathfrak{v}_{2k+1} = j^{2p}$, $p \in \mathbb{N}$, $k \in \mathbb{N}$, then the weighted ellipsoid $\mathcal{F}_{\mathfrak{v}}^r$ is a subset of the Sobolev space of p-times differentiable periodic functions. Moreover, up to a constant, for any function $h \in L^2_{\mathfrak{v}}([0,1])$, the weighted norm $\|h\|_{\mathfrak{v}}^2$ equals the L^2-norm of the p-th weak derivative $h^{(p)}$ (Tsybakov [2009]). If, on the contrary, $\mathfrak{v}_j = \exp(j^{2p})$, $p > 1$, $j \in \mathbb{N}$, then $L^2_{\mathfrak{v}}([0,1])$ is a class of analytic functions (Kawata [1972]).

(ii) Consider the complex Hilbert space $L^2([0,1))$, then the *exponential basis* $\{e_j, j \in \mathbb{Z}\}$ with

$$e_j(t) := \exp(-i2\pi jt) \text{ for } t \in [0,1) \text{ and } j \in \mathbb{Z},$$

is orthonormal and complete, i.e. an ONB. Define a weighted norm $\|\cdot\|_{\mathfrak{v}}$ with respect to the exponential basis, $\|h\|_{\mathfrak{v}} := \sum_{j\in\mathbb{Z}} \mathfrak{v}_j |\langle h, e_j\rangle|^2$. If we set $\mathfrak{v}_0 = 1$, $\mathfrak{v}_j = |j|^{2p}$, $p \in \mathbb{N}$, $j \in \mathbb{Z}\backslash\{0\}$, then $\mathcal{F}^r_{\mathfrak{v}}$ is contained in the complex Sobolev space of p-times differentiable periodic functions. Moreover, up to a constant, for any function $h \in L^2_{\mathfrak{v}}([0,1))$, the weighted norm $\|h\|^2_{\mathfrak{v}}$ equals the L^2-norm of the p-th weak derivative $h^{(p)}$. If, on the contrary, $\mathfrak{v}_j = \exp(|j|^{2p})$, $p > 1$, $j \in \mathbb{Z}$, then $L^2_{\mathfrak{v}}([0,1))$ is a class of analytic functions. □

2.2 Operators on Hilbert spaces

§2.2.1 **Linear operator**. *A mapping $T : \mathbb{H} \to \mathbb{G}$ between Hilbert spaces $\mathbb{H}$ and $\mathbb{G}$ is called* linear operator *if $T(ah_1 + bh_2) = aTh_1 + bTh_2$ for all $h_1, h_2 \in \mathbb{H}, a, b \in \mathbb{K}$. Its domain will be denoted by $\mathcal{D}(T)$, its range by $\mathcal{R}(T)$ and its null space by $\mathcal{N}(T)$.*

§2.2.2 **Property**. *Let $T : \mathbb{H} \to \mathbb{G}$ be a linear operator, then the following assertions are equivalent:*

(i) T is continuous in zero.

(ii) T is bounded, i.e., $\exists M > 0$ such that $\|Th\|_{\mathbb{G}} \leqslant M\|h\|_{\mathbb{H}}$ for all $h \in \mathbb{H}$.

(iii) T is uniformly continuous.

§2.2.3 **Defintion**. *We denote by $\mathscr{L}(\mathbb{H}, \mathbb{G})$ the class of all bound linear operator $T : \mathbb{H} \to \mathbb{G}$ and in case $\mathbb{H} = \mathbb{G}$ we write $\mathscr{L}(\mathbb{H})$ for short. For $T \in \mathscr{L}(\mathbb{H}, \mathbb{G})$ define its* (uniform) norm *as*

$$\|T\| := \sup_{h\in\mathbb{H}:\|h\|_{\mathbb{H}}\leqslant 1} \|Th\|_{\mathbb{G}}.$$

§2.2.4 **Examples**. (i) Let M be a $(m \times k)$ matrix, then $M \in \mathscr{L}(\mathbb{R}^k, \mathbb{R}^m)$.

(ii) The *integral operator* $T_k : L^2(\Omega_1, \mu_1) \to L^2(\Omega_2, \mu_2)$ defined by

$$[T_k f](\omega_2) := \int_{\Omega_1} h(\omega_1)k(\omega_1, \omega_2)\mu(d\omega_1), \ \omega_2 \in \Omega_2, \ h \in L^2(\Omega_1, \mu_1),$$

belongs to $\mathscr{L}(L^2(\Omega_1, \mu_1), L^2(\Omega_2, \mu_2))$ if $\int_{\Omega_1}\int_{\Omega_2} |k|^2 d\mu_1 d\mu_2 < \infty$.

(iii) Let X and Z be real valued random variables, then the *conditional expectation operator* of X given Z defined by $K_{X|Z}h := \mathbb{E}[h(X)|Z]$ for $h \in L^2_X$ is an element of $\mathscr{L}(L^2_X, L^2_Z)$ with $\|K_{X|Z}\| = 1$.

(iv) Let X be a random function taking its values in a real separable Hilbert space $(\mathbb{H}, \langle\cdot,\cdot\rangle_{\mathbb{H}})$ and having a finite second moment, i.e., $\mathbb{E}\|X\|^2_{\mathbb{H}} < \infty$. We say that X is centred if for all $h \in \mathbb{H}$ the real valued random variable $\langle X, h\rangle_{\mathbb{H}}$ has mean zero. Moreover, the linear operator $\Gamma_{XX} : \mathbb{H} \to \mathbb{H}$ defined by $\langle\Gamma_{XX}h_1, h_2\rangle_{\mathbb{H}} := \mathbb{E}[\langle h_1, X\rangle_{\mathbb{H}}\langle X, h_2\rangle_{\mathbb{H}}]$ for all $h_1, h_2 \in \mathbb{H}$ belongs to $\mathscr{L}(\mathbb{H})$. If the random function X is centred then Γ_{XX} is called the *covariance operator* associated with X.

(v) Let $g \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$, then the *convolution operator* $C_g : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ defined by

$$[C_g h](t) := [g \star h](t) := \int_{\mathbb{R}} g(t-s)h(s)ds, \ t \in \mathbb{R}, \ h \in L^2(\mathbb{R}),$$

belongs to $\mathscr{L}(L^2(\mathbb{R}))$ with $\|C_g\| \leqslant \|g\|_{L^1(\mathbb{R})} := \int_{\mathbb{R}} |g(t)|dt$.

(vi) Let $g \in L^2([0,1))$ and let $\lfloor \cdot \rfloor$ be the floor function, then the *circular convolution operator* $C_g : L^2([0,1)) \to L^2([0,1))$ defined by

$$[C_g h](t) := [g \star h](t) := \int_{[0,1)} g(t-s-\lfloor t-s \rfloor)h(s)ds, \ t \in [0,1), \ h \in L^2([0,1)),$$

belongs to $\mathscr{L}(L^2([0,1)))$.

(vii) Consider ℓ^2, a bounded real sequence $\lambda = (\lambda_j)_{j\in\mathcal{J}}$ and the *multiplication operator* $M_\lambda x := \lambda \cdot x$ for all $x \in \ell^2$, then $M_\lambda \in \mathscr{L}(\ell^2)$ with $\|M_\lambda\| \leqslant \|\lambda\|_\infty := \sup_{j\in\mathcal{J}} |\lambda_j|$. More general given a bounded function $\lambda : \Omega \to \mathbb{R}$ the *multiplication operator* $M_\lambda f := f \cdot \lambda$ for all $f \in L^2(\Omega,\mu)$, belongs to $\mathscr{L}(L^2(\Omega,\mu))$ where $\|M_\lambda\| \leqslant \|\lambda\|_\infty := \sup_{t\in\Omega} |\lambda(t)| < \infty$. In contrary, if $\lambda : \Omega \to \mathbb{R}$ is μ-a.s. finite and non zero, define a dense subset $\mathcal{D}(M_\lambda) := \{f \in L^2(\Omega,\mu) : \lambda \cdot f \in L^2(\Omega,\mu)\}$ of $L^2(\Omega,\mu)$. In this situation the *multiplication operator* $M_\lambda : L^2(\Omega,\mu) \supset \mathcal{D}(M_\lambda) \to L^2(\Omega,\mu)$ is densely defined and self-adjoint. □

§2.2.5 **Linear functional.** *A linear operator $\Phi : \mathbb{H} \supset \mathcal{D}(\Phi) \to \mathbb{K}$ is called linear functional and given an ONB $(u_j)_{j\in\mathcal{J}}$ in $\mathbb{H}$ which belongs to $\mathcal{D}(\Phi)$ we set $[\Phi] = ([\Phi]_j)_{j\in\mathcal{J}}$ with the slight abuse of notations $[\Phi]_j = \Phi(u_j)$. In particular, if $\Phi \in \mathscr{L}(\mathbb{H},\mathbb{K})$, i.e., $\mathcal{D}(\Phi) = \mathbb{H}$, then due to Riesz's theorem there exists a function $\phi \in \mathbb{H}$ such that $\Phi(h) = \langle h, \phi\rangle_{\mathbb{H}}$ for all $h \in \mathbb{H}$, and hence $[\Phi]_j = [\phi]_j$ for all $j \in \mathbb{N}$.*

§2.2.6 **Point evaluation.** Consider an ONB $\{u_j, j \in \mathbb{N}\}$ in the real Hilbert space $L^2([0,1])$. By evaluation at a given point $t_o \in [0,1]$ we mean the linear functional ℓ_{t_o} mapping $h \in L^2([0,1])$ to $h(t_o) := \ell_{t_o}(h) = \sum_{j\in\mathbb{N}} [h]_j u_j(t_o)$. Moreover, a point evaluation of h at t_o is well-defined, if $\sum_{j\in\mathbb{N}} |[h]_j u_j(t_o)| < \infty$. Observe that the point evaluation at t_o is generally not bounded on the subset $\{h \in L^2([0,1] : \sum_{j\in\mathbb{N}} |[h]_j u_j(t_o)| < \infty\}$. Consider the completion $L^2_{\mathfrak{v}}([0,1])$ of $L^2([0,1])$ wrt. a weighted norm $\|\cdot\|_{\mathfrak{v}}$ derived from $\{u_j, j \in \mathbb{N}\}$ and a strictly positive sequence $\mathfrak{v}$. The point evaluation is now well-defined on the set $L^2_{\mathfrak{v}}([0,1])$ if we suppose that $\sum_{j\in\mathbb{N}} |u_j(t_o)|^2 \mathfrak{v}_j^{-1} < \infty$. Consequently, if $\{u_j, j \in \mathbb{N}\}$ is the trigonometric basis (see §2.1.12 (i)), then the condition $\sum_{j\in\mathbb{N}} \mathfrak{v}_j^{-1} < \infty$ is sufficient to guarantee that the point evaluation is well-defined on $L^2_{\mathfrak{v}}([0,1])$. □

§2.2.7 **Class of linear functionals.** *In order to guarantee that the class $\mathcal{F}^r_{\mathfrak{a}}$ of solutions is contained in the domain of a linear functional Φ and that $\Phi(f) = \sum_{j\in\mathcal{J}} [\Phi]_j [f]_j$ for all $f \in \mathcal{F}^r_{\mathfrak{a}}$ with $[\Phi]_j := \Phi(h_j)$, $j \in \mathcal{J}$, it is sufficient that $\sum_{j\in\mathcal{J}} [\Phi]_j^2 \mathfrak{a}_j^{-1} < \infty$. As no confusion can be caused we define $\|\Phi\|^2_{1/\mathfrak{a}} := \sum_{j\in\mathcal{J}} [\Phi]_j^2 \mathfrak{a}_j^{-1}$ and denote by $\mathcal{L}_{1/\mathfrak{a}}$ the set of all linear functionals with $\|\Phi\|^2_{1/\mathfrak{a}} < \infty$.*

§2.2.8 ***Remark.*** We may emphasise that we neither impose that the sequence $[\Phi] = ([\Phi]_j)_{j\in\mathcal{J}}$ tends to zero nor that it is square summable. However, if it is square summable then the entire of $\mathbb{H}$ is the domain of Φ. Moreover, $[\Phi] = [\phi]$ coincides with the sequence of generalised Fourier coefficients of the representer ϕ of Φ given by Riesz's theorem. The assumption $\Phi \in \mathcal{L}_{1/\mathfrak{a}}$ enables us in specific cases to deal with more demanding functionals, such as the pointwise evaluation of the solution. □

§2.2.9 **Adjoint operator**. *If $T \in \mathscr{L}(\mathbb{H}, \mathbb{G})$, then there exists a uniquely determined adjoint operator $T^\star \in \mathscr{L}(\mathbb{G}, \mathbb{H})$ satisfying*

$$\langle Th, g\rangle_{\mathbb{G}} = \langle h, T^\star g\rangle_{\mathbb{H}} \quad \textit{for all } h \in \mathbb{H}, g \in \mathbb{G}.$$

§2.2.10 **Properties**. *Let $S, T \in \mathscr{L}(\mathbb{H}_1, \mathbb{H}_2)$ and $R \in \mathscr{L}(\mathbb{H}_2, \mathbb{H}_3)$. Then we have*
(i) $(S + T)^\star = S^\star + T^\star$, $(RS)^\star = S^\star R^\star$.
(ii) $\|S^\star\| = \|S\|$, $\|SS^\star\| = \|S^\star S\| = \|S\|^2$.
(iii) $\mathcal{N}(S) = \mathcal{R}(S^\star)^\perp$, $\mathcal{N}(S^\star) = \mathcal{R}(S)^\perp$.

§2.2.11 **Examples**. (i) The adjoint of a $(k \times m)$ matrix M is its $(m \times k)$ transpose matrix M^t.
(ii) Let $T_k \in \mathscr{L}(L^2(\Omega_1, \mu_1), L^2(\Omega_2, \mu_2))$ be an integral operator with kernel k, then its adjoint $T_k^\star = T_{k^\star} \in \mathscr{L}(L^2(\Omega_2, \mu_2), L^2(\Omega_1, \mu_1))$ is an integral operator satisfying

$$[T_{k^\star} g](\omega_1) := \int_{\Omega_2} g(\omega_2) k^\star(\omega_2, \omega_1) \mu_2(d\omega_2), \ \omega_1 \in \Omega_1, \ g \in L^2(\Omega_2, \mu_2),$$

with kernel $k^\star(\omega_2, \omega_1) := \overline{k(\omega_1, \omega_2)}$, $\omega_1 \in \Omega_1, \omega_2 \in \Omega_2$.
(iii) Let $K_{X|Z} \in \mathscr{L}(L^2_X, L^2_Z)$ be the conditional expectation of X given Z, then its adjoint operator $K^\star_{X|Z} = K_{Z|X} \in \mathscr{L}(L^2_Z, L^2_X)$ is the conditional expectation of Z given X satisfying $K_{Z|X} g = \mathbb{E}[g(Z)|X]$ for all $g \in L^2_Z$.
(iv) Let $C_g \in \mathscr{L}(L^2(\mathbb{R}))$ be the *convolution operator*, then its adjoint operator $C_g^\star = C_{g^\star}$ is a convolution operator with $g^\star(t) = \overline{g(-t)}$, $t \in \mathbb{R}$. □

§2.2.12 **Definitions**. *(i) The identity in $\mathscr{L}(\mathbb{H})$ is denoted by $\mathrm{Id}_{\mathbb{H}}$.*
(ii) Let $T \in \mathscr{L}(\mathbb{H}, \mathbb{G})$. Obviously, $T : \mathcal{N}(T)^\perp \to \mathcal{R}(T)$ is bijective and continuous whereas its inverse $T^{-1} : \mathcal{R}(T) \to \mathcal{N}(T)^\perp$ is continuous (i.e. bounded) if and only if $\mathcal{R}(T)$ is closed. In particular, if $T : \mathbb{H} \to \mathbb{G}$ is bijective (invertible) then its inverse $T^{-1} \in \mathscr{L}(\mathbb{G}, \mathbb{H})$ satisfies $\mathrm{Id}_{\mathbb{G}} = TT^{-1}$ and $\mathrm{Id}_{\mathbb{H}} = T^{-1}T$.
(iii) $\Pi \in \mathscr{L}(\mathbb{H})$ is called projection if $\Pi^2 = \Pi$. In particular, for $\Pi \neq 0$ are equivalent:
   (a) Π is an orthogonal projection ($\mathbb{H} = \mathcal{R}(\Pi) \oplus \mathcal{N}(\Pi)$);
   (b) $\|\Pi\| = 1$;
   (c) Π is non-negative.
(iv) $\mathcal{U} \in \mathscr{L}(\mathbb{H}, \mathbb{G})$ is called unitary, if $\mathcal{U}$ is invertible with $\mathcal{U}\mathcal{U}^\star = \mathrm{Id}_{\mathbb{G}}$ and $\mathcal{U}^\star\mathcal{U} = \mathrm{Id}_{\mathbb{H}}$.
(v) $\mathcal{V} \in \mathscr{L}(\mathbb{H}, \mathbb{G})$ is called partial isometry, if $\mathcal{V} : \mathcal{N}(T)^\perp \to \mathcal{R}(T)$ is unitary.
(vi) $T \in \mathscr{L}(\mathbb{H})$ is called self-adjoint, if $T = T^\star$.
(vii) $T \in \mathscr{L}(\mathbb{H})$ is called non-negative or $T \geqslant 0$ for short, if $\langle Th, h\rangle_{\mathbb{H}} \geqslant 0$ for all $h \in \mathbb{H}$.
(viii) $T \in \mathscr{L}(\mathbb{H})$ is called strictly positive or $T > 0$ for short, if $\langle Th, h\rangle_{\mathbb{H}} > 0$ for all $h \in \mathbb{H}\backslash\{0\}$.

§2.2.13 **Examples**. (i) Let $(u_j)_{j \in \mathcal{J}}$ be an ONS in $\mathbb{H}$ with linear span $\mathbb{U} := \overline{\mathrm{lin}}(\{u_j, j \in \mathcal{J}\})$ and for $f \in \mathbb{H}$ recall its sequence $[f] = ([f]_j)_{j \in \mathcal{J}}$ of generalised Fourier coefficients. The associated *generalised Fourier series transform* $\mathcal{U}$ defined for $f \in \mathbb{H}$ as $\mathcal{U} f := [f]$ belongs to $\mathscr{L}(\mathbb{H}, \ell^2)$ and is a partial isometry with adjoint operator $\mathcal{U}^\star x = \sum_{j \in \mathcal{J}} x_j u_j$ for $x \in \ell^2$. Moreover, the orthogonal projection $\Pi_{\mathbb{U}}$ onto $\mathbb{U}$ satisfies $\Pi_{\mathbb{U}} f = \mathcal{U}^\star\mathcal{U} f = \sum_{j \in \mathcal{J}} [f]_j u_j$ for all $f \in \mathbb{H}$. If $(u_j)_{j\in\mathcal{J}}$ is complete (i.e. ONB), then $\mathcal{U}$ is invertible with $\mathcal{U}\mathcal{U}^\star = \mathrm{Id}_{\ell^2}$ and $\mathcal{U}^\star\mathcal{U} = \mathrm{Id}_{\mathbb{H}}$ due to Parseval's formula, and hence $\mathcal{U}$ is unitary.

(ii) Let $\mathcal{F} \in \mathscr{L}(L^2(\mathbb{R}))$ denote the *Fourier-Plancherel transform* satisfying

$$[\mathcal{F}h](t) = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} h(x)e^{-ixt}dx, \quad \forall h \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}).$$

Then $\mathcal{F}$ is unitary with $[\mathcal{F}^\star h](t) = \frac{1}{\sqrt{2\pi}}\int h(x)e^{ixt}dx$ for all $h \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$.

(iii) A *convolution operator* $C_g \in \mathscr{L}(L^2(\mathbb{R}))$ is self-adjoint if g is a real symmetric function.

(iv) A *covariance operator* $\Gamma_{XX} \in \mathscr{L}(\mathbb{H})$ associated with a random function X is self-adjoint and non-negative.

(v) A *multiplication operator* $M_\lambda \in \mathscr{L}(L^2(\Omega,\mu))$ is strictly positive, if the function λ is strictly positive. □

§2.2.14 ***Notations***. Let $(u_k)_{k\in\mathcal{K}}$ and $(v_j)_{j\in\mathcal{J}}$ be an ONB in $\mathbb{H}$ and $\mathbb{G}$ respectively.

- For $T \in \mathscr{L}(\mathbb{H},\mathbb{G})$ denote by $[T]$ the (infinite) matrix with generic entries $[T]_{j,k} := \langle v_j, Tu_k\rangle_{\mathbb{G}}$. For $m \in \mathbb{N}$, let $[T]_{\underline{m}} := ([T]_{j,k})_{1\leqslant j,k,\leqslant m}$ denote the upper $(m\times m)$-sub-matrix of $[T]$. Note that $[T^\star]_{\underline{m}} = [T]^t_{\underline{m}}$.
- Let $\mathbb{U}_m := \overline{\mathrm{lin}}\{u_k, 1 \leqslant k \leqslant m\}$ and $\mathbb{V}_m := \overline{\mathrm{lin}}\{v_j, 1 \leqslant j \leqslant m\}$ denote the subspaces of $\mathbb{H}$ and $\mathbb{G}$ spanned by the m-first functions $\{u_k\}_{k=1}^m$ and $\{v_j\}_{j=1}^m$ respectively. Clearly, if we restrict $\Pi_{\mathbb{V}_m}T\Pi_{\mathbb{U}_m}$ to an operator from $\mathbb{U}_m$ to $\mathbb{V}_m$, then it can be represented by the matrix $[T]_{\underline{m}}$.
- Consider the identity operator $\mathrm{Id} \in \mathscr{L}(\mathbb{H})$, for example, then $[\mathrm{Id}]_{\underline{m}}$ denotes the m-dimensional identity matrix. Observe that $[\Pi_{\mathbb{U}_m}]_{\underline{m}} = [\mathrm{Id}]_{\underline{m}}$.
- For $f \in \mathbb{H}$ with sequence of generalised Fourier coefficients $[f]$. Introduce its upper m-dimensional sub-vector $[f]_{\underline{m}} := ([f]_1,\dots,[f]_m)$, where $[\Pi_{\mathbb{U}_m}f]_{\underline{m}} = [\mathrm{Id}]_{\underline{m}}[f]_{\underline{m}} = [f]_{\underline{m}}$.
- Let $\mathcal{U} \in \mathscr{L}(\mathbb{H},\ell^2)$ denote the generalised Fourier series transform satisfying $\mathcal{U}f = [f]$ and given a real valued sequence $\mathfrak{v}$ let $M_{\mathfrak{v}}$ denote the multiplication operator $M_{\mathfrak{v}}x = \mathfrak{v}\cdot x$ for all $x \in \ell^2$. Define $\nabla_{\mathfrak{v}} := \mathcal{U}^\star M_{\mathfrak{v}}\mathcal{U} : \mathbb{H} \supset \mathcal{D}(\nabla_{\mathfrak{v}}) \to \mathbb{H}$ then $[\nabla_{\mathfrak{v}}]_{\underline{m}}$ is a m-dimensional diagonal matrix with diagonal entries $(\mathfrak{v}_j)_{1\leqslant j\leqslant m}$.
- Keep in mind that $\|x\|$ denotes the Euclidean norm of a vector $x \in \mathbb{R}^m$ and let $\|f\|_{\mathfrak{v}}$ denote a weighted norm with respect to the ONB $(u_j)_{j\geqslant 1}$ in $\mathbb{H}$. In particular, for all $f \in \mathbb{U}_m$ we have $\|f\|^2_{\mathfrak{v}} = [f]^t_{\underline{m}}[\nabla_{\mathfrak{v}}]_{\underline{m}}[f]_{\underline{m}} = \|[\nabla_{\mathfrak{v}}]^{1/2}_{\underline{m}}[f]_{\underline{m}}\|^2$. Furthermore, given a $(m\times m)$-matrix M, let $\|M\|_s := \sup_{\|x\|\leqslant 1}\|Mx\|$ be its spectral norm then $\|\Pi_{\mathbb{V}_m}T\Pi_{\mathbb{U}_m}\| = \|[T]_{\underline{m}}\|_s$. □

§2.2.15 **Property**. *Given an ONB $(u_j)_{j\in\mathbb{N}}$ in $\mathbb{H}$ and two strictly positive sequence $(\mathfrak{a}_j)_{j\in\mathbb{N}}$ and $(\mathfrak{v}_j)_{j\in\mathbb{N}}$ consider the weighted norms $\|\cdot\|_{\mathfrak{a}}$ and $\|\cdot\|_{\mathfrak{v}}$. If the sequence $\mathfrak{a}/\mathfrak{v}$ is non-decreasing, then we have $\|f-\Pi_{\mathbb{U}_m}f\|^2_{\mathfrak{v}} \leqslant r\,(\mathfrak{v}_m/\mathfrak{a}_m)$ for all $f \in \mathcal{F}^r_{\mathfrak{a}}$. On the other hand if $\Phi \in \mathcal{L}_{1/\mathfrak{a}}$ as in §2.2.7 and $\mathfrak{a}$ is non-decreasing, then we have $|\Phi(f-\Pi_{\mathbb{U}_m}f)|^2 \leqslant r\sum_{j=1}^m[\Phi]^2_j\mathfrak{a}_j^{-1}$ for all $f \in \mathcal{F}^r_{\mathfrak{a}}$.*

§2.2.16 **Linear Galerkin approach**. *Let $(u_j)_{j\in\mathbb{N}}$ be an ONB in $\mathbb{H}$, $T \in \mathscr{L}(\mathbb{H})$ and $g \in \mathbb{H}$. An element $f_m \in \mathbb{U}_m = \overline{\mathrm{lin}}\{u_j, 1\leqslant j\leqslant m\}$ satisfying*

$$\|g - Tf_m\|_{\mathbb{H}} \leqslant \|g - T\breve{f}\|_{\mathbb{H}} \quad \textit{for all } \breve{f} \in \mathbb{U}_m$$

is called a Galerkin solution of the equation $g = Tf$.

§2.2.17 **Property**. *Let $T \in \mathscr{L}(\mathbb{H})$ be strictly positive definite, then for all $m \in \mathbb{N}$ the matrix $[T]_{\underline{m}}$ is also strictly positive definite. It follows that the Galerkin solution $f_m \in \mathbb{U}_m$ is uniquely determined by $[f_m]_{\underline{m}} = [T]_{\underline{m}}^{-1}[g]_{\underline{m}}$ and $[f_m]_j = 0$ for all $j > m$.*

§2.2.18 ***Remark***. Consider the orthogonal projections $\Pi_{\mathbb{U}_m} f$ and $\Pi_{\mathbb{U}_m^\perp} f$ of f onto the subspaces $\mathbb{U}_m$ and $\mathbb{U}_m^\perp$ respectively, then the approximation error $\|\Pi_{\mathbb{U}_m} f - f\|_{\mathbb{H}} = \|\Pi_{\mathbb{U}_m^\perp} f\|_{\mathbb{H}}$ converges to zero as $m \to \infty$ by Lebesgue's dominated convergence theorem. On the other hand, the Galerkin solution $f_m \in \mathbb{U}_m$ satisfies $[\Pi_{\mathbb{U}_m} f - f_m]_{\underline{m}} = -[T]_{\underline{m}}^{-1}[T\Pi_{\mathbb{U}_m^\perp} f]_{\underline{m}}$ and, hence does generally not correspond to the orthogonal projection $\Pi_{\mathbb{U}_m} f$. Moreover, the approximation error $\sup_{m\geqslant n}\|f_m - f\|_{\mathbb{H}}$ does generally not converge to zero as $n \to \infty$. However, if $C := \sup_{m\geqslant 1}\sup_{\|f\|_{\mathbb{H}}=1}\|[T]_{\underline{m}}^{-1}[T\Pi_{\mathbb{U}_m^\perp} f]_{\underline{m}}\| < \infty$, then $\|f_m - f\|_{\mathbb{H}} \leqslant (1 + C)\|\Pi_{\mathbb{U}_m^\perp} f\|_{\mathbb{H}}$ which in turn implies $\lim_{n\to\infty}\sup_{m\geqslant n}\|f_m - f\|_{\mathbb{H}} = 0$. Here and subsequently, we will restrict ourselves to classes $\mathcal{F}$ and $\mathcal{T}$ of solutions and operators respectively which ensure the convergence. Obviously, this is a minimal regularity condition for us since we aim to estimate the Galerkin solution. □

2.3 Compact operator

§2.3.1 **Compact operator**. *An operator $K \in \mathscr{L}(\mathbb{H}, \mathbb{G})$ is called compact, if $\{Kh : \|h\|_{\mathbb{H}} \leqslant 1, h \in \mathbb{H}\}$ is relatively compact in $\mathbb{G}$. We denote by $\mathscr{K}(\mathbb{H}, \mathbb{G})$ the subset of all compact operator in $\mathscr{L}(\mathbb{H}, \mathbb{G})$, and we write $\mathscr{K}(\mathbb{H}) = \mathscr{K}(\mathbb{H}, \mathbb{H})$ for short.*

§2.3.2 **Property**. *Let $K \in \mathscr{L}(\mathbb{H}, \mathbb{G})$. If $(K_j)_{j\in\mathbb{N}}$ is a sequence in $\mathscr{L}(\mathbb{H}, \mathbb{G})$ with finite dimensional range and $\lim_{j\to\infty}\|K_j - K\| = 0$, then K is compact. If in addition $\mathbb{G}$ is separable, then the converse holds also true.*

§2.3.3 **Examples**. (i) A *convolution operator* $C_g \in \mathscr{L}(L^2(\mathbb{R}))$ is not compact.

(ii) A *circular convolution operator* $C_g \in \mathscr{L}(L^2([0, 1)))$ is compact. □

§2.3.4 **Nuclear operator**. *An operator $T \in \mathscr{L}(\mathbb{H}, \mathbb{G})$ is called nuclear, if there exist sequences $(h_j)_{j\in\mathbb{N}}$, $(g_j)_{j\in\mathbb{N}}$ and $(\mathfrak{t}_j)_{j\geqslant 1}$ in $\mathbb{H}$, $\mathbb{G}$ and $\mathbb{R}$ respectively with $\|h_j\|_{\mathbb{H}} = \|g_j\|_{\mathbb{G}} = 1$ and $\sum_{j\in\mathbb{N}}|\mathfrak{t}_j| < \infty$ such that $T_n f := \sum_{j=1}^n \mathfrak{t}_j\langle f, h_j\rangle_{\mathbb{H}} g_j$, $f \in \mathbb{H}$, satisfies $\lim_{n\to\infty}\|T_n - T\| = 0$, or $T = \sum_{j\in\mathbb{N}}\mathfrak{t}_j\langle\cdot, h_j\rangle g_j$ for short. We denote by $\mathscr{N}(\mathbb{H}, \mathbb{G})$ the subset of all nuclear operator in $\mathscr{L}(\mathbb{H}, \mathbb{G})$, and we write $\mathscr{N}(\mathbb{H}) := \mathscr{N}(\mathbb{H}, \mathbb{H})$. Furthermore, let $(h_j)_{j\in\mathbb{N}}$ be an ONB in $\mathbb{H}$ and $T \in \mathscr{N}(\mathbb{H})$, then $\operatorname{tr}(T) := \sum_{j\in\mathbb{N}}\langle Th_j, h_j\rangle_{\mathbb{H}}$ denotes the trace of T.*

§2.3.5 ***Remark***. The trace of a nuclear operator does not depend on the choice of the ONB. We have $\mathscr{N}(\mathbb{H}, \mathbb{G}) \subset \mathscr{K}(\mathbb{H}, \mathbb{G}) \subset \mathscr{L}(\mathbb{H}, \mathbb{G})$. Moreover, if $T \in \mathscr{N}(\mathbb{H}, \mathbb{G})$ and $S \in \mathscr{L}(\mathbb{G}, \mathbb{H})$ then $TS \in \mathscr{N}(\mathbb{H})$, $ST \in \mathscr{N}(\mathbb{G})$ and $\operatorname{tr}(TS) = \operatorname{tr}(ST)$. □

§2.3.6 **Example**. Let $\Gamma_{XX} \in \mathscr{L}(\mathbb{H})$ be the *covariance operator* associated with a random function X. If $\mathbb{E}\|X\|_{\mathbb{H}}^2 < \infty$, then $\Gamma_{XX} \in \mathscr{N}(\mathbb{H})$ and $\operatorname{tr}(\Gamma_{XX}) = \mathbb{E}\|X\|_{\mathbb{H}}^2$. □

§2.3.7 **Hilbert-Schmidt operator**. *An operator $T \in \mathscr{L}(\mathbb{H}, \mathbb{G})$ is called Hilbert-Schmidt, if there exists an ONB $(h_j)_{j\in\mathbb{N}}$ in $\mathbb{H}$ such that $\|T\|_{\mathscr{H}}^2 := \sum_{j\in\mathbb{N}}\|Th_j\|_{\mathbb{G}}^2 < \infty$. The number $\|T\|_{\mathscr{H}}$ is called Hilbert-Schmidt norm of T and satisfies $\|T\| \leqslant \|T\|_{\mathscr{H}}$. We denote by $\mathscr{H}(\mathbb{H}, \mathbb{G})$ the subset of all Hilbert-Schmidt operator in $\mathscr{L}(\mathbb{H}, \mathbb{G})$, and we write $\mathscr{H}(\mathbb{H}) := \mathscr{H}(\mathbb{H}, \mathbb{H})$.*

§2.3.8 ***Remark***. The number $\|T\|_{\mathscr{H}}$ does not depend on the choice of the ONB. The space $\mathscr{H}(\mathbb{H}, \mathbb{G})$ endowed with the inner product $\langle T, S\rangle_{\mathscr{H}} := \operatorname{tr}(S^\star T)$, $S, T \in \mathscr{H}(\mathbb{H}, \mathbb{G})$ is a Hilbert space and $\|\cdot\|_{\mathscr{H}}$ the induced norm. Moreover, $\mathscr{H}(\mathbb{H}, \mathbb{G}) \subset \mathscr{K}(\mathbb{H}, \mathbb{G})$. □

§2.3.9 **Examples**. (i) Let $T \in \mathscr{L}(L^2(\Omega_1, \mu_1), L^2(\Omega_2, \mu_2))$. The operator T is Hilbert-Schmidt if and only if it is an *integral operator* $T = T_k$ with square integrable kernel k and it holds $\|T\|_{\mathscr{H}}^2 = \int_{\Omega_1} \int_{\Omega_2} |k(\omega_1, \omega_2)|^2 \mu_1(d\omega_1)\mu_2(d\omega_2)$.

(ii) Consider the *conditional expectation operator* $K_{X|Z} \in \mathscr{L}(L^2_X, L^2_Z)$ of X given Z. Let in addition $p_{X,Z}$, p_X and p_Z be, respectively, the joint and marginal densities of (X, Z), X and Z with respect to a σ-finite measure. In this situation, the operator $K_{X|Z}$ is Hilbert Schmidt if and only if $\mathbb{E}\big[\frac{|p_{XZ}(X,Z)|^2}{|p_X(X)p_Z(Z)|^2}\big] < \infty$. □

2.4 Spectrum of an operator

§2.4.1 **Definitions**. *Let $T \in \mathscr{L}(\mathbb{H})$ be self-adjoint.*

- *Resolvent set of T: $\rho(T) = \{\lambda \in \mathbb{R} : (\lambda - T)^{-1} \in \mathscr{L}(\mathbb{H})\}$.*
- Spectrum of T*: $\sigma(T) = \mathbb{R}\backslash\rho(T) = \sigma_p(T) \cup \sigma_c(T) \cup \sigma_r(T)$.*
  - *$\sigma_p(T) = \{\lambda : \lambda - T$ is not injective$\}$*
  - *$\sigma_c(T) = \{\lambda : \lambda - T$ is injective and not surjective with dense range$\}$*
  - *$\sigma_r(T) = \{\lambda : \lambda - T$ is injective and its range is not dense$\}$*
- *Typically (but not always)*
  - *$\sigma_p(T)$ contains isolated points,*
  - *$\sigma_c(T)$ is a union of intervals,*
  - *$\sigma_r(T)$ is empty.*

An element λ of $\sigma_p(T)$ and $h \in \mathbb{H}\backslash\{0\}$ with $Th = \lambda h$ are called eigenvalue *and* eigenvector *(eigenfunction) respectively.*

§2.4.2 **Properties**. *Let $T \in \mathscr{K}(\mathbb{H})$ be self-adjoint.*

(i) If $\mathbb{H}$ is infinite dimensional, then $0 \in \sigma(T)$.

(ii) The (possibly empty) set $\sigma(T)\backslash\{0\}$ is at most countable.

(iii) Any $\lambda \in \sigma(T)\backslash\{0\}$ is an eigenvalue of T and the associated eigenspace $\mathcal{N}(\lambda - T)$ is finite dimensional.

(iv) In $\sigma(T)$ the only possible accumulation point is zero.

§2.4.3 **Properties**. *For $T \in \mathscr{K}(\mathbb{H})$ self-adjoint, let $\{\lambda_j, j \in \mathcal{J}\} \subset \mathbb{R}$ denote the (possibly empty) countable spectrum of T, i.e., $\sigma(T)\backslash\{0\} = \{\lambda_j, j \in \mathcal{J}\}$.*

(i) If $T \in \mathscr{N}(\mathbb{H})$ is self-adjoint, then the sequence $(\lambda_j)_{j\in\mathcal{J}}$ is absolute summable, i.e. $\sum_{j\in\mathcal{J}} |\lambda_j| < \infty$, and $\operatorname{tr}(T) = \sum_{j\in\mathcal{J}} \lambda_j$.

(ii) If $T \in \mathscr{H}(\mathbb{H})$ is self-adjoint, then the sequence $(\lambda_j)_{j\in\mathcal{J}}$ is square summable, i.e. $\sum_{j\in\mathcal{J}} |\lambda_j|^2 < \infty$, and $\|T\|_{\mathscr{H}}^2 = \sum_{j\in\mathcal{J}} |\lambda_j|^2$.

2.5 Eigenvalue and singular value decomposition

§2.5.1 **Eigenvalue decomposition**. *Let $T \in \mathscr{K}(\mathbb{H})$ be self-adjoint. There exist*

(i) an ordered sequence $\lambda := (\lambda_j)_{j\in\mathcal{J}}$ in $\mathbb{R}\backslash\{0\}$ which has either a finite number of entries or tends to zero, and determines a multiplication operator $M_\lambda \in \mathscr{L}(\ell^2)$,

(ii) a (possibly finite) ONS $\{u_j, j \in \mathcal{J}\}$ in $\mathbb{H}$ with associated generalised Fourier series transform $\mathcal{U}f = [f]$ for all $f \in \mathbb{H}$ which belongs to $\mathscr{L}(\mathbb{H}, \ell^2)$

such that $\mathbb{H} = \mathcal{N}(T) \oplus \overline{\operatorname{lin}}\{u_j, j \in \mathcal{J}\}$ and

$$Tf = \mathcal{U}^\star M_\lambda \mathcal{U} f (= \sum_{j\in\mathcal{J}} \lambda_j \langle f, u_j\rangle_{\mathbb{H}} u_j).$$

For $j \in \mathcal{J}$, λ_j and u_j are, respectively, a non-zero eigenvalue and associated eigenvector of T respectively. $\{(\lambda_j, u_j), j \in \mathcal{J}\}$ is called an eigensystem of T. Moreover, $\|T\| = \sup_{j\in\mathcal{J}} |\lambda_j|$.

§2.5.2 **Example**. Let $\Gamma_{XX} \in \mathscr{N}(\mathbb{H})$ be a strictly positive *covariance operator* associated with a random function X satisfying $\mathbb{E}\|X\|_{\mathbb{H}}^2 < \infty$. In this situation the eigenvectors $\{u_j, j \in \mathbb{N}\}$ of T associated with the strictly positive eigenvalues $\{\lambda_j, j \in \mathbb{N}\}$ form an ONB in $\mathbb{H}$, and hence the corresponding generalised Fourier series transform $\mathcal{U}f = [f]$ is unitary. Furthermore, given the ONB of eigenfunctions the (infinite) matrix representation $[\Gamma_{XX}] = [\nabla_\lambda]$ is diagonal, i.e., for all $m \in \mathbb{N}$, $[\Gamma_{XX}]_{\underline{m}} = [\nabla_\lambda]_{\underline{m}}$ is a m-dimensional diagonal matrix with entries $(\lambda_j)_{1\leqslant j\leqslant m}$. □

§2.5.3 **Singular value decomposition**. *Let $T \in \mathscr{K}(\mathbb{H}, \mathbb{G})$. There exist*

(i) an ordered strictly positive sequence $\mathfrak{s} := (\mathfrak{s}_j)_{j\in\mathcal{J}}$ which has either a finite number of entries or tends to zero, and determines a multiplication operator $M_{\mathfrak{s}} \in \mathscr{L}(\ell^2)$,

(ii) an (possibly finite) ONS $\{u_j, j \in \mathcal{J}\}$ in $\mathbb{H}$ with associated generalised Fourier series transform $\mathcal{U} \in \mathscr{L}(\mathbb{H}, \ell^2)$ (a partial isometry),

(iii) an (possibly finite) ONS $\{v_j, j \in \mathcal{J}\}$ in $\mathbb{G}$ with associated generalised Fourier series transform $\mathcal{V} \in \mathscr{L}(\mathbb{G}, \ell^2)$ (a partial isometry),

such that $\mathbb{H} = \mathcal{N}(T) \oplus \overline{\operatorname{lin}}\{u_j, j \in \mathcal{J}\}$, $\mathbb{G} = \mathcal{N}(T^\star) \oplus \overline{\operatorname{lin}}\{v_j, j \in \mathcal{J}\}$ and

$$Tf = \mathcal{V}^\star M_{\mathfrak{s}} \mathcal{U} f \ (= \sum_{j\in\mathcal{J}} \mathfrak{s}_j \langle f, u_j\rangle_{\mathbb{H}} v_j).$$

In particular, $\{(\mathfrak{s}_j^2, u_j), j \in \mathcal{J}\}$ and $\{(\mathfrak{s}_j^2, v_j), j \in \mathcal{J}\}$ are an eigensystem of $T^\star T$ and $TT^\star$ respectively. The numbers $\{\mathfrak{s}_j, j \in \mathcal{J}\}$ and triplets $\{(\mathfrak{s}_j, u_j, v_j), j \in \mathcal{J}\}$ are, respectively, called singular values and singular system of T.

§2.5.4 **Example**. Let $C_g \in \mathscr{K}(L^2([0,1)))$ be a *circular convolution operator*. Consider the *exponential basis* $\{e_j\}_{j\in\mathbb{Z}}$ in $L^2([0,1))$ and let $[f]_j$, $j \in \mathbb{Z}$, denote the Fourier coefficients of $f \in L^2([0,1))$. Keep in mind that for all $f \in L^2([0,1))$ the convolution theorem states $[g \star f]_j = [g]_j [f]_j$ for all $j \in \mathbb{Z}$.

- $\{(|[g]_j|, e_j, \frac{[g]_j}{|[g]_j|} e_j),\ j \in \mathbb{Z}\}$ is a singular system of the circular convolution operator C_g. Given the exponential basis $\{e_j\}_{j\in\mathbb{Z}}$ and the ONB $\{\frac{[g]_j}{|[g]_j|} e_j, j \in \mathbb{Z}\}$ the (infinite) matrix representation $[C_g] = [\nabla_{|[g]|}]$ is diagonal, i.e., for all $m \in \mathbb{N}$, $[C_g]_{\underline{m}} = [\nabla_{|[g]|}]_{\underline{m}}$ is a m-dimensional diagonal matrix with entries $|[g]_{\underline{m}}| = (|[g]_j|)_{1\leqslant j\leqslant m}$.

- If g is real and even, then $|[g]_j| = [g]_j$ for all $j \in \mathbb{Z}$, which in turn implies that C_g is self-adjoint and $\{([g]_j, e_j),\ j \in \mathbb{Z}\}$ is an eigensystem of C_g. Furthermore, given the exponential basis the (infinite) matrix representation $[C_g] = [\nabla_{[g]}]$ is diagonal, i.e., for all $m \in \mathbb{N}$, $[C_g]_{\underline{m}} = [\nabla_{[g]}]_{\underline{m}}$ is a m-dimensional diagonal matrix with entries $[g]_{\underline{m}} = ([g]_j)_{1 \leqslant j \leqslant m}$. □

2.6 Spectral theorem

§2.6.1 **Spectral theorem**. *If $T \in \mathscr{L}(\mathbb{H})$ is self-adjoint, then T is unitarily equivalent to a multiplication operator, i.e., there exist*

(i) a measurable space (Ω, μ) (σ-finite, if $\mathbb{H}$ is separable),

(ii) an unitary operator $\mathcal{U} \in \mathscr{L}(\mathbb{H}, L^2(\Omega, \mu))$,

(iii) a bounded function $\lambda : \Omega \to \mathbb{R}$ with associated multiplication operator $M_\lambda \in \mathscr{L}(L^2(\Omega, \mu))$

such that $T = \mathcal{U}^\star M_\lambda \mathcal{U}$.

§2.6.2 **Example**. Let $g \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ be a real symmetric function. Consider the associated self-adjoint *convolution operator* $C_g \in \mathscr{L}(L^2(\mathbb{R}))$. Recall that the convolution theorem states $\mathcal{F}(g \star f) = \sqrt{2\pi} \cdot \mathcal{F}g \cdot \mathcal{F}f$ for all $f \in L^2(\mathbb{R})$ where $\mathcal{F}$ denotes the *Fourier-Plancherel transform*. Consequently, the operator C_g is unitarily equivalent to the multiplication operator $M_\lambda \in \mathscr{L}(L^2(\mathbb{R}))$ with $\lambda = \sqrt{2\pi} \cdot [\mathcal{F}g]$, that is $C_g = \mathcal{F}^{-1} M_\lambda \mathcal{F}$. □

§2.6.3 **Spectral theorem Halmos [1963]**. *Let $T : \mathbb{H} \supset \mathcal{D}(T) \to \mathbb{H}$ be a densely-defined self-adjoint operator. There exist*

(i) a measurable space (Ω, μ) (σ-finite, if $\mathbb{H}$ is separable),

(ii) an unitary operator $\mathcal{U} \in \mathscr{L}(\mathbb{H}, L^2(\Omega, \mu))$,

(iii) a function $\lambda : \Omega \to \mathbb{R}$ (μ-a.s. finite and non zero) with associated multiplication operator $M_\lambda : L^2(\Omega, \mu) \supset \mathcal{D}(M_\lambda) \to L^2(\Omega, \mu)$ with $\mathcal{D}(M_\lambda) = \{f \in L^2(\Omega, \mu) : \lambda f \in L^2(\Omega, \mu)\}$

such that $\mathcal{D}(T) = \{h \in \mathbb{H} : \mathcal{U}h \in \mathcal{D}(M_\lambda)\}$ and

(a) for all $f \in \mathcal{D}(M_\lambda)$ we have $M_\lambda f = \lambda \cdot f = \mathcal{U}T\mathcal{U}^\star f$,

(b) for all $h \in \mathcal{D}(T)$ it holds $Th = \mathcal{U}^\star M_\lambda \mathcal{U}h$,

i.e., T is unitarily equivalent to the multiplication operator M_λ.

§2.6.4 **Example**. Let $\Gamma_{XX} \in \mathscr{N}(\mathbb{H})$ be a strictly positive *covariance operator* with eigenvalue decomposition $\Gamma_{XX} = \mathcal{U}^\star M_\lambda \mathcal{U}$ where $\mathcal{U} \in \mathscr{L}(\mathbb{H}, \ell^2)$ is unitary, $M_\lambda \in \mathscr{L}(\ell^2)$ is a multiplication operator and λ a strictly positive summable sequence of eigenvalues. In this situation the range $\mathcal{R}(\Gamma_{XX})$ of Γ_{XX} is dense in $\mathbb{H}$ but not closed. Therefore, there exists an inverse $\Gamma_{XX}^{-1} : \mathcal{R}(\Gamma_{XX}) \to \mathbb{H}$ of Γ_{XX} which is densely-defined and self-adjoint but not continuous. In particular, we have $\mathcal{D}(\Gamma_{XX}^{-1}) = \mathcal{R}(\Gamma_{XX}) = \{h : \lambda^{-1}\mathcal{U}h \in \ell^2\}$ (which is called Picard's condition). Consider the multiplication operator $M_{\lambda^{-1}} : \ell^2 \supset \mathcal{D}(M_{\lambda^{-1}}) \to \ell^2$ with $\mathcal{D}(M_{\lambda^{-1}}) = \{x \in \ell^2 : \lambda^{-1}x \in \ell^2\}$, then $\mathcal{D}(\Gamma_{XX}^{-1}) = \{h \in \mathbb{H} : \mathcal{U}h \in \mathcal{D}(M_{\lambda^{-1}})\}$ and

(a) for all $x \in \mathcal{D}(M_{\lambda^{-1}})$ we have $M_{\lambda^{-1}}x = \lambda^{-1} \cdot x = \mathcal{U}\Gamma_{XX}^{-1}\mathcal{U}^\star x$,

(b) for all $h \in \mathcal{D}(\Gamma_{XX}^{-1})$ it holds $\Gamma_{XX}^{-1}h = \mathcal{U}^\star M_{\lambda^{-1}}\mathcal{U}h$, i.e. Γ_{XX}^{-1} is unitarily equivalent to the multiplication operator $M_{\lambda^{-1}}$. In conclusion we have

$$\Gamma_{XX}h = \mathcal{U}^\star M_\lambda \mathcal{U}h = \sum_{j\in\mathcal{J}} \lambda_j [h]_j u_j, \quad \forall h \in \mathbb{H};$$

$$\Gamma_{XX}^{-1}h = \mathcal{U}^\star M_{\lambda^{-1}} \mathcal{U}h = \sum_{j\in\mathcal{J}} \lambda_j^{-1} [h]_j u_j, \quad \forall h \in \mathbb{H} \text{ such that } \sum_{j\in\mathcal{J}} \frac{[h]_j^2}{\lambda_j^2} < \infty. \qquad \square$$

2.7 Functional Calculus

§2.7.1 **Functional calculus**. *Let $T \in \mathscr{K}(\mathbb{H})$ be self-adjoint with eigensystem $\{(\lambda_j, u_j), j \in \mathcal{J}\}$, i.e., $Th = \mathcal{U}^\star M_\lambda \mathcal{U}h = \sum_{j\in\mathcal{J}} \lambda_j [h]_j u_j$, $h \in \mathbb{H}$, then we have*

$$T^k h = \mathcal{U}^\star M_{\lambda^k} \mathcal{U}h = \sum_{j\in\mathcal{J}} \lambda_j^k [h]_j u_j, \quad \forall k = 1, 2, \dots .$$

Given a function $g : \mathbb{R} \to \mathbb{R}$ define the operator

$$g(T)h := \mathcal{U}^\star M_{g(\lambda)} \mathcal{U}h = \sum_{j\in\mathcal{J}} g(\lambda_j)[h]_j u_j, \quad \forall\, h \in \mathbb{H} \textit{ such that } \sum_{j\in\mathcal{J}} |g(\lambda_j)|^2 |[h]_j|^2 < \infty.$$

which belongs to $\mathscr{L}(\mathbb{H})$ if g is bounded. In this case we have $\|g(T)\| \leqslant \|g\|_\infty$. More generally, let $T \in \mathscr{L}(\mathbb{H})$ be self-adjoint and hence unitary equivalent with multiplication by a bounded function λ in some $L^2(\Omega, \mu)$, that is $Th = U^* M_\lambda U h$, $h \in \mathbb{H}$. Given a function $g : \mathbb{R} \to \mathbb{R}$ define the multiplication operator

$$M_{g(\lambda)} : L^2(\Omega, \mu) \supset \mathcal{D}(M_{g(\lambda)}) \to L^2(\Omega, \mu)$$

with $\mathcal{D}(M_{g(\lambda)}) = \{f \in L^2(\Omega, \mu) : g(\lambda) f \in L^2(\Omega, \mu)\}$ and

$$g(T)h := \mathcal{U}^\star M_{g(\lambda)} \mathcal{U}h, \quad \forall h \in \mathcal{D}(g(T)) := \left\{h \in \mathbb{H} : \mathcal{U}h \in \mathcal{D}(M_{g(\lambda)})\right\}$$

where $g(T) : \mathscr{L}(\mathbb{H}) \supset \mathcal{D}(g(T)) \to \mathscr{L}(\mathbb{H})$. Moreover, if g is bounded then $g(T) \in \mathscr{L}(\mathbb{H})$ with $\|g(T)\| \leqslant \|g\|_\infty$.

§2.7.2 **Example**. Let M be $(k \times m)$-Matrix. Consider for $y \in \mathbb{R}^k$ and $\alpha > 0$ the *Tikhonov functional*

$$F_\alpha(x) := \|Mx - y\|^2 + \alpha \|x\|^2, \quad \forall x \in \mathbb{R}^k,$$

which has a unique minimiser $x_\alpha := (\alpha \operatorname{Id} + M^t M)^{-1} M^t y \in \mathbb{R}^k$. Define the function $r_\alpha(t) = (\alpha + t)^{-1}$, $t \geqslant 0$, then we have $x_\alpha = r_\alpha(M^t M) M^t y$. Analogously, given a *convolution operator* $C_g \in \mathscr{L}(L^2(\mathbb{R}))$, a function $h \in L^2(\mathbb{R})$ and $\alpha > 0$ the unique minimiser of

$$F_\alpha(f) := \|C_g f - h\|^2 + \alpha \|f\|^2, \quad \forall f \in L^2(\mathbb{R})$$

is given by $f_\alpha = r_\alpha(C_g^\star C_g) C_g^\star h$. Moreover, if $\mathcal{F} \in \mathscr{L}(L^2(\mathbb{R}))$ denotes the *Fourier-Plancherel transform*, then $C_g^\star C_g = \mathcal{F}^{-1} M_\lambda \mathcal{F}$ with $\lambda = 2\pi \cdot |\mathcal{F}g|^2$. Applying the functional calculus we have

$$f_\alpha = r_\alpha(C_g^\star C_g) C_g^\star h = \mathcal{F}^{-1} M_{r_\alpha(\lambda)} \mathcal{F}\, C_g^\star h = \mathcal{F}^{-1} \frac{\sqrt{2\pi} \cdot \overline{\mathcal{F}g} \cdot \mathcal{F}h}{\alpha + 2\pi |\mathcal{F}g|^2}. \qquad \square$$

§2.7.3 **Class of operators**. *Denote by $\mathcal{T}$ the set of all strictly positive operator $T \in \mathscr{K}(\mathbb{H})$. Given an ONB $(u_j)_{j\in\mathcal{J}}$ in $\mathbb{H}$ and a strictly positive sequence $(\mathfrak{b}_j)_{j\in\mathcal{J}}$ let $\|\cdot\|_{\mathfrak{b}^2}$ be the weighted norm given by $\|h\|_{\mathfrak{b}^2} = \sum_{j\in\mathcal{J}} \mathfrak{b}_j^2 |\langle h, u_j\rangle_{\mathbb{H}}|^2$. For all $d \geqslant 1$ define the subset $\mathcal{T}_{\mathfrak{b}}^d$ of $\mathcal{T}$ by*

$$\mathcal{T}_{\mathfrak{b}}^d := \left\{T \in \mathcal{T} : d^{-1}\|h\|_{\mathfrak{b}^2} \leqslant \|Th\|_{\mathbb{H}} \leqslant d\|h\|_{\mathfrak{b}^2}\right\}.$$

We say, T satisfies a link condition, if $T \in \mathcal{T}_{\mathfrak{b}}^d$.

§2.7.4 **Inequality of Heinz**. *If $T \in \mathcal{T}_{\mathfrak{b}}^d$, then for all $|s| \leqslant 1$ holds true*

$$d^{-|s|}\|h\|_{\mathfrak{b}^{2s}} \leqslant \|T^s h\| \leqslant d^{|s|}\|h\|_{\mathfrak{b}^{2s}}$$

§2.7.5 **Bias due to the Galerkin approach Johannes and Schenk [2013]**. *Let $T \in \mathcal{T}_{\mathfrak{b}}^d$ with non-increasing sequence $\mathfrak{b}$, then we have*

$$\begin{aligned}
&\sup_{m\geqslant 1}\left\{\mathfrak{b}_m\|[T]_{\underline{m}}^{-1}\|_s\right\} \leqslant d(d^2+2) \leqslant 3d^3,\\
&\sup_{m\geqslant 1}\|[T]_{\underline{m}}^{-1}[\nabla_{\mathfrak{b}}]_{\underline{m}}\|_s^2 \leqslant d^2(d^2+2)^2 \leqslant 9d^6, \quad \sup_{m\geqslant 1}\|[\nabla_{\mathfrak{b}}]_{\underline{m}}^{1/2}[T]_{\underline{m}}^{-1}[\nabla_{\mathfrak{b}}]_{\underline{m}}^{1/2}\|_s \leqslant d(d^2+2) \leqslant 3d^3,\\
&\sup_{m\geqslant 1}\|[T]_{\underline{m}}[\nabla_{\mathfrak{b}}]_{\underline{m}}^{-1}\|_s^2 \leqslant d^2, \quad \sup_{m\geqslant 1}\|[\nabla_{\mathfrak{b}}]_{\underline{m}}^{-1/2}[T]_{\underline{m}}[\nabla_{\mathfrak{b}}]_{\underline{m}}^{-1/2}\|_s \leqslant d.
\end{aligned}$$

Let in addition $f \in \mathcal{F}_{\mathfrak{a}}^r$ with sequence $\mathfrak{a}$ such that $(\mathfrak{a}/\mathfrak{b})$ is non-decreasing. If f_m denotes a Galerkin solution of $g = Tf$ then for each strictly positive sequence $\mathfrak{v}$ such that $(\mathfrak{v}/\mathfrak{a})$ is non-increasing and for all $m \geqslant 1$ we obtain

$$\begin{aligned}
&\|f - f_m\|_{\mathfrak{v}}^2 \leqslant (d^2+2)^2\, r\, \mathfrak{v}_m \mathfrak{a}_m^{-1} \max\left(1, \mathfrak{b}_m^2 \mathfrak{v}_m^{-1} \max_{1\leqslant j\leqslant m} \mathfrak{v}_j \mathfrak{b}_j^{-2}\right),\\
&\|f_m\|_{\mathfrak{a}}^2 \leqslant r\,(d^2+2)^2, \quad \sup_{m\geqslant 1}\{\mathfrak{b}_m^{-2}\mathfrak{a}_m\,\|T(f-f_m)\|_{\mathbb{H}}^2\} \leqslant d^2 r,\\
&\quad\textit{and}\quad \sup_{m\geqslant 1}\{\mathfrak{b}_m^{-1}\mathfrak{a}_m\,\|T^{1/2}(f-f_m)\|_{\mathbb{H}}^2\} \leqslant d(d^2+2)^2 r.
\end{aligned}$$

Furthermore, for all $m \geqslant 1$ and all $\Phi \in \mathcal{L}_{1/\mathfrak{a}}$ we have

$$|\Phi(f^m - f)|^2 \leqslant (d^2+2)^2\|f\|_{\mathfrak{a}}^2 \max\left\{\sum_{j>m}\frac{[\Phi]_j^2}{\mathfrak{a}_j}, \frac{\mathfrak{b}_m^2}{\mathfrak{a}_m}\sum_{j=1}^m\frac{[\Phi]_j^2}{\mathfrak{b}_j^2}\right\}$$

which implies for all $0 \leqslant s \leqslant 2$ that

$$|\Phi(f^m - f)|^2 \leqslant (d^2+2)^2\|f\|_{\mathfrak{a}}^2 \max\left\{\sum_{j>m}\frac{[\Phi]_j^2}{\mathfrak{a}_j}, \frac{\mathfrak{b}_m^s}{\mathfrak{a}_m}\sum_{j=1}^m\frac{[\Phi]_j^2}{\mathfrak{b}_j^s}\right\}.$$

2.8 Statistical ill-posed inverse problems

Let $T \in \mathscr{L}(\mathbb{H}, \mathbb{G})$ be a linear bounded operator between Hilbert spaces $\mathbb{H}$ and $\mathbb{G}$.

§2.8.1 **Inverse Problem**. *Given $g \in \mathbb{G}$ reconstruct a solution $f \in \mathbb{H}$ of the equation $g = Tf$.*

§2.8.2 **Definition (Hadamard [1932])**. *An inverse problem $g = Tf$ is called **well-posed** if*

- *a solution f exists,*
- *the solution f is unique and*
- *the solution depends continuously on g.*

An inverse problem which is not well-posed is called ***ill-posed****.*

For a broader overview on inverse problems we refer the reader to the monograph by Kress [1989] or Engl et al. [2000].

§2.8.3 **Existence and identification.**
There exists an unique solution of the equation $g = Tf$ if and only if

- *g belongs to the range $\mathcal{R}(T)$ of T (existence),*
- *The operator T is injective, i.e., its null space $\mathcal{N}(T) = \{0\}$ is trivial (identification).*

§2.8.4 ***Remark.*** If there does not exist a solution typically one might consider a least-square solution which exists if and only if $g \in \mathcal{R}(T) \oplus \mathcal{N}(T^\star)$. A least-square solution with minimal norm, if it exists, could be recovered, in case the solution is not unique. Nevertheless, the main issue is often the stability of the inverse problem. More precisely, if the solution does not depend continuously on g, i.e., the inverse T^{-1} of T is not continuous, a reconstruction $f_\varepsilon = T^{-1}\widehat{g}_\varepsilon$ given a noisy version $\widehat{g}_\varepsilon$ of g may be far from the solution f even if the noise level ε tends to zero. □

§2.8.5 **Statistical inverse problem.** *Given a noisy version*

$$\widehat{g}_\varepsilon = g + \sqrt{\varepsilon}\dot{W}$$

of $g \in \mathbb{G}$ the reconstruction of a solution $f \in \mathbb{H}$ of the equation $g = Tf$ is called statistical inverse problem if $\dot{W}$ is a random error with noise level ε. Here and subsequently, we suppose that the random error $\dot{W}$ is a Hilbert-space noise. To be precise, given any ONS $(v_j)_{j\in\mathcal{J}}$ in $\mathbb{G}$ we suppose that $\left\{[\dot{W}]_j := \langle \dot{W}, v_j\rangle_{\mathbb{G}}, j \in \mathcal{J}\right\}$ is a family of real valued random variables and that the observable quantities take the form

$$\langle \widehat{g}_\varepsilon, v_j\rangle_{\mathbb{G}} = \langle g, v_j\rangle_{\mathbb{G}} + \sqrt{\varepsilon}\langle \dot{W}, v_j\rangle_{\mathbb{G}}, \quad j \in \mathcal{J},$$

or, $[\widehat{g}_\varepsilon]_j = [g]_j + \sqrt{\varepsilon}[\dot{W}]_j$, $j \in \mathcal{J}$, in short.

§2.8.6 **Gaussian inverse regression.** Let $T \in \mathscr{K}(\mathbb{H})$ be strictly positive and $\{u_j, j \in \mathbb{N}\}$ be an ONB in $\mathbb{H}$ not necessarily corresponding to the eigenfunctions of T. Our aim is the reconstruction the function $f \in \mathbb{H}$ based on a noisy version $\widehat{g}_\varepsilon = Tf + \sqrt{\varepsilon}\dot{W}$ of $g = Tf \in \mathbb{G}$ where $\dot{W}$ is a Gaussian white noise. Considering the projection onto the ONB $\{u_j, j \in \mathbb{N}\}$ the observable quantities take the form $[\widehat{g}_\varepsilon]_j = [Tf]_j + \sqrt{\varepsilon}[\dot{W}]_j$, $j \in \mathbb{N}$ where the error terms $\left\{[\dot{W}]_j, j \in \mathbb{N}\right\}$ are independent and standard normally distributed. □

§2.8.7 **Indirect sequence space model.** *Assume a statistical inverse problem $\widehat{g}_\varepsilon = Tf + \sqrt{\varepsilon}\dot{W}$ with $T \in \mathcal{K}(\mathbb{H}, \mathbb{G})$ having a singular system $\{(\mathfrak{s}_j, u_j, v_j), j \in \mathcal{J}\}$. Considering the eigenbasis from the inverse problem $g = Tf$ we derive*

$$[g]_j = [Tf]_j = s_j[f]_j, \quad j \in \mathcal{J},$$

and hence the observable quantities take the form

$$[\widehat{g}_\varepsilon]_j = s_j[f]_j + \sqrt{\varepsilon}[\dot{W}]_j, \quad j \in \mathcal{J}.$$

§2.8.8 ***Remark***. If the operator $T = \mathrm{Id}$ is the identity the sequence space model can alternatively be derived from a white noise regression model $\widehat{g}_\varepsilon = f + \sqrt{\varepsilon}\dot{W}$ which has been intensively studied in the literature. For a broad introduction to nonparametric estimation in this model we refer to the textbook by Tsybakov [2009]. □

§2.8.9 **Gaussian sequence space model**. Given a noisy version $\widehat{g}_\varepsilon = f + \sqrt{\varepsilon}\dot{W}$ of $f \in \mathbb{H}$ where $\dot{W}$ is a Gaussian white noise we aim to reconstruct f. In this situation the observable quantities take the form $[\widehat{g}_\varepsilon]_j = [f]_j + \sqrt{\varepsilon}[\dot{W}]_j$, $j \in \mathbb{N}$ where the error terms $\left\{[\dot{W}]_j, j \in \mathbb{N}\right\}$ are independent and standard normally distributed. □

§2.8.10 **Gaussian indirect sequence space model**. Let $T \in \mathscr{K}(\mathbb{H}, \mathbb{G})$ be injective with ordered singular system $\{(\mathfrak{s}_j, u_j, v_j), j \in \mathbb{N}\}$, that is, the sequence $(\mathfrak{s}_j^{-1})_{j\in\mathbb{N}}$ is non-decreasing. Our aim is the reconstruction of a function $f \in \mathbb{H}$ based on a noisy version $\widehat{g}_\varepsilon = Tf + \sqrt{\varepsilon}\dot{W}$ of $g = Tf \in \mathbb{G}$ where $\dot{W}$ is a Gaussian white noise. Considering the projection onto the eigenbasis of T the observable quantities take the form $[\widehat{g}_\varepsilon]_j = \mathfrak{s}_j[f]_j + \sqrt{\varepsilon}[\dot{W}]_j$, $j \in \mathbb{N}$ where the error terms $\left\{[\dot{W}]_j, j \in \mathbb{N}\right\}$ are independent and standard normally distributed. □

§2.8.11 **Circular deconvolution with known error density**. Let X be the circular random variable whose density f_X we are interested in and U an independent additive circular error with unknown density f_U. Denote by Y the contaminated observation and by f_Y its density. We will identify the circle with the unit interval $[0,1)$, for notational convenience. Thus, X and U take their values in $[0,1)$. Let $\lfloor\cdot\rfloor$ be the floor function. Taking into account the circular nature of the data, the model can be written as $Y = X + U - \lfloor X + U \rfloor$ or equivalently $Y = X + U \mod [0,1)$. Then, we have $f_Y = f_X \star f_U$ where $\star$ denotes circular convolution (compare §2.2.4 (vi)) and, hence $f_Y = C_{f_U} f_X$. Suppose further that the error density f_U and thus the operator $C_{f_U} \in \mathscr{K}(L^2([0,1)))$ is know a *priori*. Consider the *exponential basis* $\{e_j\}_{j\in\mathbb{Z}}$ in $L^2([0,1))$ and let $[f]_j$, $j \in \mathbb{Z}$, denote the Fourier coefficients of $f \in L^2([0,1))$. Applying the convolution theorem (compare §2.5.4) we have $[f_Y]_j = [f_U]_j[f_X]_j$ with $[f_Y]_j = \mathbb{E}e_j(-Y)$ for all $j \in \mathbb{Z}$. Assuming an iid. sample $Y_i \sim f_Y$, $i = 1, \dots, n$ we estimate $[f_Y]_j$ by its emprical counterpart $\widehat{[f_Y]}_j = n^{-1}\sum_{i=1}^n e_j(-Y_i)$ for all $j \in \mathbb{Z}$, and hence denote $\widehat{f}_Y = \sum_{j\in\mathbb{Z}} \widehat{[f_Y]}_j e_j$. Given the noisy version $\widehat{f}_Y$ of $f_Y = C_{f_U} f_X$ we aim to reconstruct f_X which is an ill-posed indirect sequence space model where the observable quantities take the form $[\widehat{f}_Y]_j = [f_U]_j[f_X]_j + \frac{1}{\sqrt{n}}[\dot{W}]_j$ with $[\dot{W}]_j := \frac{1}{\sqrt{n}}\sum_{i=1}^n \{e_j(-Y_i) - \mathbb{E}e_j(-Y_i)\}$ for all $j \in \mathbb{Z}$. Note that these error terms are centered but generally not independent and identically distributed. □

§2.8.12 **Statistical inverse problem with noise in the operator**. *A statistical inverse problem $\widehat{g}_\varepsilon = Tf + \sqrt{\varepsilon}\dot{W}$, where in addition a noisy version*

$$\widehat{T}_\sigma = T + \sqrt{\sigma}\dot{B}$$

of the operator $T \in \mathscr{L}(\mathbb{H}, \mathbb{G})$ is observed only, is called statistical inverse problem with noise in the operator if $\dot{B}$ is a random error with noise level σ. Here and subsequently, we suppose that the random error $\dot{B}$ is a Hilbert-space operator noise, that is, given any ONB $(v_j)_{j\in\mathcal{J}}$ and

$(u_k)_{k\in\mathcal{K}}$ in $\mathbb{G}$ and $\mathbb{H}$ respectively we suppose that $\left\{[\dot{B}]_{j,k} := \langle v_j, \dot{B}u_k\rangle_{\mathbb{G}}, k\in\mathcal{K}, j\in\mathcal{J}\right\}$ is a family of real valued random variables and that the observable quantities take the form

$$\langle v_j, \widehat{T}_\sigma u_k\rangle_{\mathbb{G}} = \langle v_j, Tu_k\rangle_{\mathbb{G}} + \sqrt{\sigma}\langle v_j, u_k\rangle_{\mathbb{G}}, \quad j\in\mathcal{J}, k\in\mathcal{K},$$

or, $[\widehat{T}_\sigma]_{j,k} = [T]_{j,k} + \sqrt{\sigma}[\dot{B}]_{j,k}$, $j\in\mathcal{J}, k\in\mathcal{K}$, in short.

§2.8.13 **Gaussian inverse regression with noise in the operator.** Let $T\in\mathscr{K}(\mathbb{H})$ be strictly positive and $\{u_j, j\in\mathbb{N}\}$ be an ONB in $\mathbb{H}$ not necessarily corresponding to the eigenfunctions of T. Our aim is the reconstruction of the function $f\in\mathbb{H}$ based on noisy versions $\widehat{g}_\varepsilon = Tf+\sqrt{\varepsilon}\dot{W}$ and $\widehat{T}_\sigma = T+\sqrt{\sigma}\dot{B}$ of $g = Tf\in\mathbb{G}$ and T, respectively, where $\dot{W}$ and $\dot{B}$ are Gaussian white noise. Considering the projection onto the ONB $\{u_j, j\in\mathbb{N}\}$ the observable quantities take the form $[\widehat{g}_\varepsilon]_j = [g]_j + \sqrt{\varepsilon}[\dot{W}]_j$ and $[\widehat{T}_\sigma]_{j,k} = [T]_{j,k} + \sqrt{\sigma}[\dot{B}]_{j,k}$, for $j,k\in\mathbb{N}$, where the error terms $\left\{[\dot{W}]_j, [\dot{B}]_{j,k}, j,k\in\mathbb{N}\right\}$ are independent and standard normally distributed. □

§2.8.14 **Functional linear regression.** Let X be a random function taking its values in a separable Hilbert space $(\mathbb{H}, \langle\cdot,\cdot\rangle_{\mathbb{H}})$. For convenient notations we assume that the regressor X is centered in the sense that for all $h\in\mathbb{H}$ the real valued random variable $\langle X, h\rangle_{\mathbb{H}}$ has mean zero. The linear relationship between a real random variable Y and the variation of X is expressed by the equation $Y = \langle\beta, X\rangle_{\mathbb{H}} + U$, with the unknown slope parameter $\beta\in\mathbb{H}$ and a real-valued and centered error term U. We suppose that the regressor X has a finite second moment, i.e., $\mathbb{E}\|X\|_{\mathbb{H}}^2 < \infty$, and that X is uncorrelated to the random error U in the sense that $\mathbb{E}(U\langle X, h\rangle_{\mathbb{H}}) = 0$ for all $h\in\mathbb{H}$. Multiplying both sides in the model equation by X and taking the expectation leads to the normal equation $c_{YX} := \mathbb{E}(YX) = \mathbb{E}(\langle\beta, X\rangle_{\mathbb{H}}X) =: \Gamma_{XX}\beta$ where the cross-correlation function c_{YX} belongs to $\mathbb{H}$ and $\Gamma_{XX}\in\mathscr{N}(\mathbb{H})$ denotes the covariance operator associated with the random function X (compare §2.2.4 (iv)). Assuming an iid. sample $\{(Y_i, X_i), i = 1,\dots,n\}$ of (Y, X), it is natural to consider the estimators $\widehat{c}_{YX} := \frac{1}{n}\sum_{i=1}^n Y_iX_i$ and $\widehat{\Gamma}_{XX} := \frac{1}{n}\sum_{i=1}^n\langle\cdot, X_i\rangle_{\mathbb{H}}X_i$ of c_{YX} and Γ_{XX} respectively. Given the noisy versions $\widehat{c}_{YX}$ of $c_{YX} = \Gamma_{XX}\beta$ and $\widehat{\Gamma}_{XX}$ of Γ_{XX} we aim to reconstruct β which is a statistical ill-posed inverse problem with noise in the operator. Moreover, given an ONB $\{u, j\in\mathbb{N}\}$ in $\mathbb{H}$ the observable quantities take the form $[\widehat{c}_{YX}]_j = [\Gamma_{XX}\beta]_j + \frac{1}{\sqrt{n}}[\dot{W}]_j$ and $[\widehat{\Gamma}_{XX}]_{j,k} = [\Gamma_{XX}]_{j,k} + \frac{1}{\sqrt{n}}[\dot{B}]_{j,k}$ with $[\dot{W}]_j := \frac{1}{\sqrt{n}}\sum_{i=1}^n\{Y_i[X_i]_j - \mathbb{E}Y_i[X_i]_j\}$ and $[\dot{B}]_{j,k} := \frac{1}{\sqrt{n}}\sum_{i=1}^n\{[X_i]_j[X_i]_k - \mathbb{E}[X_i]_j[X_i]_k\}$ for all $j,k\in\mathbb{N}$. Note that these error terms are centered but generally not independent or identically distributed. □

§2.8.15 **Nonparametric instrumental regression.** A structural function φ characterises the dependency of a real response Y on the variation of an $\mathbb{R}^p$-valued endogenous explanatory random variable X by $Y = \varphi(X) + U$ where $\mathbb{E}[U|X]\neq 0$ for some error term U. In other words, the structural function equals not the conditional mean function of Y given X. In nonparametric instrumental regression, however, a sample from (Y, X, Z) is available, where Z is an additional $\mathbb{R}^q$-valued random vector of exogenous *instruments* such that $\mathbb{E}[U|Z] = 0$. It is convenient to rewrite the model equations in terms of an operator between Hilbert spaces. Therefore, let us first recall the Hilbert spaces $(L_X^2, \langle\cdot,\cdot\rangle_X)$ and $(L_Z^2, \langle\cdot,\cdot\rangle_Z)$ defined in §2.1.3 (v). Taking the conditional expectation wrt. the instrument Z on both sides in the model equation yields $r_{Y|Z} := \mathbb{E}[Y|Z] = \mathbb{E}[\varphi(X)|Z] =: K_{X|Z}\varphi$ where the regression function $r_{Y|Z}$ belongs to L_Z^2 and $K_{X|Z}$ is the conditional expectation of X given Z assumed to be an element of $\mathscr{K}(L_X^2, L_Z^2)$ (compare §2.2.4 (iii)). Considering an ONB $\{u_j, j\in\mathbb{N}\}$ in L_X^2 and an ONS $\{v_j, j\in\mathbb{N}\}$ in L_Z^2

we have $[r_{Y|Z}]_j = \langle r_{Y|Z}, v_j\rangle_Z = \mathbb{E}[Yv_j(Z)]$ and $[K_{X|Z}]_{j,k} = \langle v_j, K_{X|Z}u_k\rangle_W = \mathbb{E}[u_k(X)v_j(Z)]$ for $j,k \in \mathbb{N}$. Assuming an iid. sample $\{(Y_i, X_i, Z_i), i = 1,\dots,n\}$ of (Y,X,Z), it is natural to consider estimators $\widehat{r}_{Y|Z}$ and $\widehat{K}_{X|Z}$ of $r_{Y|Z}$ and $K_{X|Z}$, respectively, defined by $[\widehat{r}_{Y|Z}]_j := \frac{1}{n}\sum_{i=1}^n Y_i v_j(Z_i)$ and $[\widehat{K}_{X|Z}]_{j,k} := \frac{1}{n}\sum_{i=1}^n u_k(X_i)v_j(Z_i)$ for all $j,k \in \mathbb{N}$. Given the noisy versions $\widehat{r}_{Y|Z}$ of $r_{Y|Z} = K_{X|Z}\varphi$ and $\widehat{K}_{X|Z}$ of $K_{X|Z}$ we aim to reconstruct φ which is a statistical ill-posed inverse problem with noise in the operator. Moreover, the observable quantities take the form $[\widehat{r}_{Y|Z}]_j = [K_{X|Z}\varphi]_j + \frac{1}{\sqrt{n}}[\dot{W}]_j$ and $[\widehat{K}_{X|Z}]_{j,k} = [K_{X|Z}]_{j,k} + \frac{1}{\sqrt{n}}[\dot{B}]_{j,k}$ with $[\dot{W}]_j := \frac{1}{\sqrt{n}}\sum_{i=1}^n \{Y_i v_j(Z_i) - \mathbb{E}Y_i v_j(Z_i)\}$ and $[\dot{B}]_{j,k} := \frac{1}{\sqrt{n}}\sum_{i=1}^n \{u_k(X_i)v_j(Z_i) - \mathbb{E}u_k(X_i)v_j(Z_i)\}$ for all $j,k \in \mathbb{N}$. Note that these error terms are centered but generally not independent and identically distributed. □

§2.8.16 **Indirect sequence space model with noise in the operator**. *Assume a statistical inverse problem $\widehat{g}_\varepsilon = Tf + \sqrt{\varepsilon}\dot{W}$ with noise in the operator $\widehat{T}_\sigma = T + \sqrt{\sigma}\dot{B}$ where the operator $T \in \mathcal{K}(\mathbb{H},\mathbb{G})$ has a known eigenbasis $(v_j)_{j\in\mathcal{J}}$ and $(u_j)_{j\in\mathcal{J}}$ in $\mathbb{G}$ and $\mathbb{H}$ respectively. Considering the eigenbasis from the inverse problem $g = Tf$ we derive*

$$[g]_j = [Tf]_j = [T]_{j,j}[f]_j, \quad j \in \mathcal{J},$$

and hence the observable quantities take the form

$$[\widehat{g}_\varepsilon]_j = [T]_{j,j}[f]_j + \sqrt{\varepsilon}[\dot{W}]_j \quad \textit{and} \quad [\widehat{T}_\sigma]_{j,j} = [T]_{j,j} + \sqrt{\sigma}[\dot{B}]_{j,j}, \quad j \in \mathcal{J}.$$

§2.8.17 **Gaussian indirect sequence space model with noise in the operator**. Let $T \in \mathscr{K}(\mathbb{H},\mathbb{G})$ be injective with singular system $\{(\mathfrak{s}_j, u_j, v_j), j \in \mathcal{J}\}$ with a *priori* known ONB $\{u_j, j \in \mathcal{J}\}$ and ONS $\{v_j, j \in \mathcal{J}\}$. Our aim is the reconstruction of a function $f \in \mathbb{H}$ based on noisy versions $\widehat{g}_\varepsilon = Tf + \sqrt{\varepsilon}\dot{W}$ and $\widehat{T}_\sigma = T + \sqrt{\sigma}\dot{B}$ of $g = Tf \in \mathbb{G}$ and T, respectively, where $\dot{W}$ and $\dot{B}$ are Gaussian white noises. Considering the projection onto the eigenbasis of T the observable quantities take the form $[\widehat{g}_\varepsilon]_j = \mathfrak{s}_j[f]_j + \sqrt{\varepsilon}[\dot{W}]_j$ and $[\widehat{T}_\sigma]_{j,j} = \mathfrak{s}_j + \sqrt{\sigma}[\dot{B}]_{j,j}$, $j \in \mathcal{J}$ where the error terms $\left\{[\dot{W}]_j, [\dot{B}]_{j,j}, j \in \mathcal{J}\right\}$ are independent and standard normally distributed.

§2.8.18 **Circular deconvolution with unknown error density**. Consider the statistical inverse problem §2.8.11. Suppose that we do not know neither the density $f_Y = C_{f_U} f_X$ of the contaminated observations, nor the error density f_U. But we have at our disposal two independent samples of independent and identically distributed random variables $Y_i \sim f_Y$, $i = 1,\dots,n$ and $U_i \sim f_U$, $i = 1,\dots,m$, of size $n \in \mathbb{N}$ and $m \in \mathbb{N}$, respectively. Consider the noisy version $\widehat{f}_Y$ of $f_Y = C_{f_U} f_X$ defined in §2.8.11. In addition we estimate $[f_U]_j$ by its empirical counterpart $\widehat{[f_U]}_j := m^{-1}\sum_{i=1}^m e_j(-U_i)$ for all $j \in \mathbb{Z}$ and, hence denote $\widehat{f_U} = \sum_{j\in\mathbb{Z}} \widehat{[f_U]}_j e_j$. Given the noisy versions $\widehat{f}_Y$ of $f_Y = C_{f_U} f_X$ and $C_{\widehat{f_U}}$ of C_{f_U} we aim to reconstruct f_X which is an ill-posed indirect sequence space model with noise in the operator where the observable quantities take the form $[\widehat{f}_Y]_j = [f_U]_j[f_X]_j + \frac{1}{\sqrt{n}}[\dot{W}]_j$ and $[C_{\widehat{f_U}}]_{j,j} = [f_U]_j + \frac{1}{\sqrt{m}}[\dot{B}]_{j,j}$ with $[\dot{W}]_j := \frac{1}{\sqrt{n}}\sum_{i=1}^n \{e_j(-Y_i) - \mathbb{E}e_j(-Y_i)\}$ and $[\dot{B}]_{j,j} := \frac{1}{\sqrt{m}}\sum_{i=1}^m \{e_j(-U_i) - \mathbb{E}e_j(-U_i)\}$ for all $j \in \mathbb{Z}$. □

Chapter 3

Minimax optimality and adaptive estimation

3.1 Minimax theory: a general approach

Let ε denote a noise level and suppose the function of interest f belongs to a class $\mathcal{F} \subset \mathbb{H}$. Denote further by $\mathcal{P}_{\mathcal{F}} := \{P_{f,\varepsilon}, f \in \mathcal{F}, \varepsilon \in (0,1)\}$ a family of probability measures and let $\mathbb{E}_{f,\varepsilon}$ denote the expectation wrt. to a measure $P_{f,\varepsilon}$ in $\mathcal{P}_{\mathcal{F}}$. Moreover, we assume that the probability measure associated with an observable quantity belongs to $\mathcal{P}_{\mathcal{F}}$.

§3.1.1 **Gaussian sequence space model (GSSM)**. Consider a sequence $([\widehat{g}_\varepsilon]_j)_{j\in\mathbb{N}}$ of Gaussian random variables having mean $([f]_j)_{j\in\mathbb{N}}$ and variance ε for a given ONB $(u_j)_{j\in\mathbb{N}}$ in $\mathbb{H}$ (compare §2.8.9). Our aim is the reconstruction of the function f assuming that it belongs to an ellipsoid $\mathcal{F}_{\mathfrak{a}}^r$ derived from the ONB $(u_j)_{j\in\mathbb{N}}$ and some non-decreasing weight sequence $(\mathfrak{a}_j)_{j\in\mathbb{N}}$ (compare §2.1.11). In this situation $P_{f,\varepsilon}$ denotes the joint distribution of $([\widehat{g}_\varepsilon]_j)_{j\in\mathbb{N}}$ and, consequently $\mathcal{P}_{\mathcal{F}_{\mathfrak{a}}^r}$ is a class of joint distribution of sequences of Gaussian random variables. □

Furthermore, based on the observable quantity we have an estimator $\widetilde{f}$ of f at hand which takes its values in $\mathbb{H}$ but does not necessarily belong to $\mathcal{F}$. We shall measure the accuracy of any estimator $\widetilde{f}$ of f by its distance $\mathfrak{D}_{\text{ist}}(\widetilde{f}, f)$ where $\mathfrak{D}_{\text{ist}}(\cdot,\cdot)$ is a certain semi metric to be specified below. Moreover, we call the quantity $\mathbb{E}_{f,\varepsilon}[\mathfrak{D}_{\text{ist}}^2(\widetilde{f}, f)]$ risk of the estimator $\widetilde{f}$ of f.

§3.1.2 **Maximal risk**. *The maximal risk of an estimator $\widetilde{f}$ of the function of interest f over the class of solutions $\mathcal{F}$ based on an observable quantity with probability measure $P_{f,\varepsilon} \in \mathcal{P}_{\mathcal{F}}$ is defined by*

$$\mathcal{R}_\varepsilon\big[\widetilde{f}\,|\,\mathcal{P}_{\mathcal{F}}\big] := \sup_{f\in\mathcal{F}} \mathbb{E}_{f,\varepsilon}[\mathfrak{D}_{\text{ist}}^2(\widetilde{f}, f)].$$

Global risk: *Consider the completion $\mathbb{H}_{\mathfrak{v}}$ of $\mathbb{H}$ wrt. a weighted norm $\|\cdot\|_{\mathfrak{v}}$. If $\mathcal{F} \subset \mathbb{H}_{\mathfrak{v}}$ then $\mathfrak{D}_{\text{ist}}(h_1, h_1) := \|h_1 - h_2\|_{\mathfrak{v}}$, $h_1, h_2 \in \mathbb{H}_{\mathfrak{v}}$, defines a global distance. We call $\mathbb{H}_{\mathfrak{v}}$-risk the associated global risk $\mathbb{E}_{f,\varepsilon}\|\widetilde{f} - f\|_{\mathfrak{v}}^2$ and we denote $\mathcal{R}_\varepsilon^{\mathfrak{v}}\big[\widetilde{f}\,|\,\mathcal{P}_{\mathcal{F}}\big] := \sup_{f\in\mathcal{F}} \mathbb{E}_{f,\varepsilon}\|\widetilde{f} - f\|_{\mathfrak{v}}^2$.*

Local risk: *Let Φ be a linear functional and $\mathcal{F} \subset \mathcal{D}(\Phi)$, then $\mathfrak{D}_{\text{ist}}(h_1, h_1) := |\Phi(h_1 - h_2)|$, $h_1, h_2 \in \mathcal{D}(\Phi)$, denotes a local distance. Its associated local risk $\mathbb{E}_{f,\varepsilon}|\Phi(\widetilde{f} - f)|^2$ we call Φ-risk and we set $\mathcal{R}_\varepsilon^{\Phi}\big[\widetilde{f}\,|\,\mathcal{P}_{\mathcal{F}}\big] := \sup_{f\in\mathcal{F}} \mathbb{E}_{f,\varepsilon}|\Phi(\widetilde{f}) - \Phi(f)|^2$.*

§3.1.3 ***Remark***. An advantage of taking a maximal risk instead of a risk is that the former does not depend on the unknown function f. Imagine we would have taken a constant estimator, say $\widetilde{f} = h$, of f. This would be the perfect estimator if by chance $f = h$, but in all other cases this estimator is likely to perform poorly. Therefore it is reasonable to consider the supremum over the whole class of possible functions in order to get consolidated findings. However, considering the maximal risk may be a very pessimistic point of view. □

§3.1.4 **Minimax-optimality**. *Let $\mathcal{R}_\varepsilon\big[\,\cdot\,|\,\mathcal{P}_\mathcal{F}\big]$ be a maximal risk over a class $\mathcal{P}_\mathcal{F}$ of probability measures. If there exist an estimator $\widehat{f}$ and constants $C_- := C_-(\mathcal{P}_\mathcal{F})$, $C_+ := C_+(\mathcal{P}_\mathcal{F})$ and $\mathcal{R}^\star_\varepsilon := \mathcal{R}^\star_\varepsilon\big(\mathcal{P}_\mathcal{F}\big)$, $\varepsilon \in (0,1)$, with $\lim_{\varepsilon\to 0}\mathcal{R}^\star_\varepsilon = 0$, depending on the class $\mathcal{P}_\mathcal{F}$ such that*

Lower bound: the quantity $\mathcal{R}^\star_\varepsilon$ is a lower bound up to the constant C_- of the maximal risk over all possible estimators of f, that is

$$\inf_{\widetilde{f}} \mathcal{R}_\varepsilon\big[\widetilde{f}\,|\,\mathcal{P}_\mathcal{F}\big] \geqslant C_-\,\mathcal{R}^\star_\varepsilon \quad \forall\,\varepsilon \in (0,1)$$

where the infimum is taken over all possible estimators of f;

Upper bound: the quantity $\mathcal{R}^\star_\varepsilon$ is an upper bound up to the constant C_+ of the maximal risk associated with an estimator $\widehat{f}$ of f, that is

$$\mathcal{R}_\varepsilon\big[\widehat{f}\,|\,\mathcal{P}_\mathcal{F}\big] \leqslant C_+\,\mathcal{R}^\star_\varepsilon \quad \forall\,\varepsilon \in (0,1).$$

Then we call $\mathcal{R}^\star_\varepsilon$ minimax-optimal rate of convergence and the estimator $\widehat{f}$ minimax-optimal up to a constant.

§3.1.5 ***Remark***. It is worth noting that a minimax-optimal rate is not unique since every other rate that is equivalent of order is also minimax-optimal. □

§3.1.6 **GSSM (**§3.1.1 **continued)**. We restrict ourselves to the two particular cases of a maximal local Φ-risk and a maximal global $\mathbb{H}_\mathfrak{v}$-risk over the class $\mathcal{P}_{\mathcal{F}^r_\mathfrak{a}}$ which are denoted here and subsequently, respectively, by

$$\mathcal{R}^\Phi_\varepsilon\big[\widetilde{f}\,|\,\mathcal{F}^r_\mathfrak{a}\big] := \sup_{f\in\mathcal{F}^r_\mathfrak{a}} \mathbb{E}_{f,\varepsilon}|\Phi(\widetilde{f}-f)|^2 \quad\text{and}\quad \mathcal{R}^\mathfrak{v}_\varepsilon\big[\widetilde{f}\,|\,\mathcal{F}^r_\mathfrak{a}\big] := \sup_{f\in\mathcal{F}^r_\mathfrak{a}} \mathbb{E}_{f,\varepsilon}\|\widetilde{f}-f\|^2_\mathfrak{v}. \qquad \square$$

Keeping the GSSM (§3.1.1) in mind we will consider in the next chapter a Gaussian inverse regression model where Gaussian observations $\widehat{g}_\varepsilon = Tf + \sqrt{\varepsilon}\dot{W}$ and $\widehat{T}_\sigma = T + \sqrt{\sigma}\dot{B}$ are given. We will first consider the case that the operator T is known *a priori*, that is, the noise level σ is zero, and in a second step we dismiss this information. Obviously, in both cases the distribution $P_{f,T,\varepsilon}$ of $\widehat{g}_\varepsilon$ depends not only on the parameter of interest $f \in \mathcal{F}$ and the noise level $\varepsilon \in (0,1)$, but also on the operator T which plays the role of a nuissance parameter. Consequently, let $\mathcal{P}^T_\mathcal{F}$ denote a class of possible distributions of $[\widehat{g}_\varepsilon]$. In the situation, that T is not known *a priori*, the function of interest f is in general not more identified given only an observation $\widehat{g}_\varepsilon$. However, identification of f can be ensured if we assume, for example, a noisy version $\widehat{T}_\sigma = T + \sqrt{\sigma}\dot{B}$ of T. In this situation, we denote by $P_{f,T,\varepsilon,\sigma}$ the joint distribution of all observable quantities. Moreover, specifying in addition a class of possible operators $\mathcal{T}$ we consider a family of associated probability measures $\mathcal{P}_{\mathcal{F},\mathcal{T}} := \{P_{f,T,\varepsilon,\sigma}, f \in \mathcal{F}, T \in \mathcal{T};\, \varepsilon,\sigma \in (0,1)\}$. If in addition $\dot{W}$ and $\dot{B}$ are independent, then the joint distribution of the observable quantities is given by the product measure $P_{f,T,\varepsilon,\sigma} = P_{f,T,\varepsilon} \otimes P_{T,\sigma}$ where $P_{f,T,\varepsilon}$ and $P_{T,\sigma}$ are the marginal distributions of $\widehat{g}_\varepsilon$ and $\widehat{T}_\sigma$, respectively, and, consequently, $\mathcal{P}_{\mathcal{F},\mathcal{T}} = \{P_{f,T,\varepsilon} \otimes P_{T,\sigma}, f \in \mathcal{F}, T \in \mathcal{T};\, \varepsilon,\sigma \in (0,1)\}$ is a family of product measures. More generally, given a class $\mathcal{F}$ of solutions f and a class Ξ of nuisances parameters ξ we denote by $\mathcal{P}_{\mathcal{F},\Xi} := \{P_{f,\xi,\varepsilon,\sigma}, f \in \mathcal{F}, \xi \in \Xi;\, \varepsilon,\sigma \in (0,1)\}$ a family of probability measures. Let $\mathbb{E}_{f,\xi,\varepsilon,\sigma}$ denote the expectation with respect to a measure $P_{f,\xi,\varepsilon,\sigma}$ in $\mathcal{P}_{\mathcal{T},\Xi}$. In the following we assume that the probability measure associated with an observable quantity belongs to $\mathcal{P}_{\mathcal{F},\Xi}$.

§3.1.7 **Maximal risk**. *The maximal risk of an estimator $\widetilde{f}$ of the function of interest f over a class of solutions $\mathcal{F}$ and a class of nuisances parameters Ξ based on an observable quantity with probability measure in $\mathcal{P}_{\mathcal{F},\Xi}$ is defined by*

$$\mathcal{R}_{\varepsilon,\sigma}\big[\widetilde{f}\,|\,\mathcal{P}_{\mathcal{F},\Xi}\big] := \sup_{f\in\mathcal{F}}\sup_{\xi\in\Xi}\mathbb{E}_{f,\xi,\varepsilon,\sigma}[\mathfrak{D}^2_{ist}(\widetilde{f},f)].$$

Global risk: *Denote by $\mathcal{R}^{\mathfrak{v}}_{\varepsilon,\sigma}\big[\widetilde{f}\,|\,\mathcal{P}_{\mathcal{F},\Xi}\big] := \sup_{f\in\mathcal{F}}\sup_{\xi\in\Xi}\mathbb{E}_{f,\xi,\varepsilon,\sigma}\|\widetilde{f}-f\|^2_{\mathfrak{v}}$ the associated maximal $\mathbb{H}_{\mathfrak{v}}$-risk.*

Local risk: *Denote by $\mathcal{R}^{\Phi}_{\varepsilon,\sigma}\big[\widetilde{f}\,|\,\mathcal{P}_{\mathcal{F},\Xi}\big] := \sup_{f\in\mathcal{F}}\sup_{\xi\in\Xi}\mathbb{E}_{f,\xi,\varepsilon,\sigma}|\Phi(\widetilde{f}-f)|^2$ the associated maximal Φ-risk.*

§3.1.8 ***Remark***. Taking the supremum over the class of nuissance parameters allows us to quantify the additional complexity due to the estimation of the nuissance parameter. Moreover, if there exist an estimator $\widehat{f}$ and constants $C_- := C_-(\mathcal{P}_{\mathcal{F},\Xi})$, $C_+ := C_+(\mathcal{P}_{\mathcal{F},\Xi})$ and $\mathcal{R}^\star_{\varepsilon,\sigma} := \mathcal{R}^\star_{\varepsilon,\sigma}(\mathcal{P}_{\mathcal{F},\Xi})$, $\varepsilon,\sigma\in(0,1)$, with $\lim_{\varepsilon,\sigma\to 0}\mathcal{R}^\star_{\varepsilon,\sigma}=0$, depending only on the class $\mathcal{P}_{\mathcal{F},\Xi}$ such that

Lower bound: $$\inf_{\widetilde{f}}\mathcal{R}_{\varepsilon,\sigma}\big[\widetilde{f}\,|\,\mathcal{P}_{\mathcal{F},\Xi}\big] \geqslant C_-\,\mathcal{R}^\star_{\varepsilon,\sigma}\quad \forall\,\varepsilon,\sigma\in(0,1);$$

Upper bound: $$\mathcal{R}_{\varepsilon,\sigma}\big[\widehat{f}\,|\,\mathcal{P}_{\mathcal{F},\Xi}\big] \leqslant C_+\,\mathcal{R}^\star_{\varepsilon,\sigma}\quad \forall\,\varepsilon\,\sigma\in(0,1).$$

Then we call $\mathcal{R}^\star_{\varepsilon,\sigma}$ minimax-optimal rate of convergence and the estimator $\widehat{f}$ minimax-optimal up to a constant. Typically, we assume first that the nuissance parameter ξ is known *a priori*, and hence $\mathcal{P}_{\mathcal{F},\xi}$ is a class of probability measures associated with the observable quantities. In this situation, we consider the maximal risk $\mathcal{R}_{\varepsilon,\sigma}\big[\,\cdot\,|\,\mathcal{P}_{\mathcal{F},\xi}\big] := \sup_{f\in\mathcal{F}}\mathbb{E}_{f,\xi,\varepsilon,\sigma}[\mathfrak{D}^2_{\mathrm{ist}}(\cdot,f)]$ and we seek a lower bound $\mathcal{R}^\star_{\varepsilon,\sigma}$ up to a constant which depends possibly on the nuissance parameter ξ. However, if the lower bound $\mathcal{R}^\star_{\varepsilon,\sigma}$ is valid uniformly for all nuissance parameters $\xi\in\Xi$, then it is, obviously, also a lower bound of the maximal risk $\mathcal{R}_{\varepsilon,\sigma}\big[\widetilde{f}\,|\,\mathcal{P}_{\mathcal{F},\Xi}\big]$. □

3.2 Deriving a lower bound: a general reduction scheme

For a detailed discussion of several other strategies to derive lower bounds we refer the reader, for example, to the text book by Tsybakov [2009].

§3.2.1 **Definition**. *Let $\mathbb{E}_P$ and $\mathbb{E}_Q$ denote the expectation associated with two probability measures P and Q, which are absolutely continuous wrt. to a σ-finite measure μ, or $P,Q\ll\mu$ for short. We write $dP := dP/d\mu$ and $dQ := dQ/d\mu$.*

- *The Kullback-Leibler divergence between P and Q is defined by*

$$KL(P,Q) := \begin{cases}\int \log\frac{dP}{dQ}dP, & \text{if } P\ll Q;\\ \infty, & \text{otherwise.}\end{cases}$$

- *The Hellinger distance between P and Q is defined by*

$$H(P,Q) := \left(\int(\sqrt{dP}-\sqrt{dQ})^2\right)^{1/2} = \|\sqrt{dP}-\sqrt{dQ}\|_{L^2}$$

which does not depend on the choice of the dominating measure.

- *The Hellinger affinity is given by*

$$\rho(P,Q) := \int \sqrt{dP}\sqrt{dQ} = \langle \sqrt{dP}, \sqrt{dQ}\rangle_{L^2}$$

§3.2.2 **Properties**. *a)* $0 \leqslant H^2(P,Q) \leqslant 2$*; b)* $\rho(P,Q) = 1 - \frac{1}{2}H^2(P,Q)$*; and c)* $H^2(P,Q) \leqslant KL(P,Q)$*.*

Proof of §3.2.2. a) Applying the triangular inequality we obtain

$$0 \leqslant H^2(P,Q) \leqslant 2\{\|\sqrt{dP}\|^2_{L^2} + \|\sqrt{dQ}\|^2_{L^2}\} = 2\{\int dP + \int dQ\} = 2.$$

b) Observe that $H^2(P,Q) = \|\sqrt{dP}\|^2_{L^2} + \|\sqrt{dQ}\|^2_{L^2} + 2\langle\sqrt{dP}, \sqrt{dQ}\rangle_{L^2} = 2 + 2\int\sqrt{dPdQ}$, rewriting the last equality gives the assertion.

c) Suppose $KL(P,Q) < \infty$ and keep in mind that $-\log(x+1) \geqslant -x$ for $x > -1$, then b) implies

$$\begin{aligned} KL(P,Q) &= \int_{dPdQ>0} \log\frac{dP}{dQ}dP = 2\int_{dPdQ>0} dP \log\sqrt{\frac{dP}{dQ}} \\ &= -2\int_{dPdQ>0} dP\log\{\sqrt{\frac{dQ}{dP}} - 1 + 1\} \geqslant -2\int_{dPdQ>0} dP\{\sqrt{\frac{dQ}{dP}} - 1\} \\ &= -2\{\int_{dPdQ>0}\sqrt{dQdP} - 1\} = H^2(P,Q) \quad \square \end{aligned}$$

§3.2.3 **Property**. *Consider a family $\mathcal{P}_{\mathcal{F}}$ of probability measures. Let $\mathbb{E}_P$ and $\mathbb{E}_Q$ be the expectations associated with two probability measures P and Q belonging to $\mathcal{P}_{\mathcal{F}}$, respectively, and let $\widetilde{f}$ be an estimator of the function of interest $f \in \mathcal{F}$. For all $f_1, f_{-1} \in \mathcal{F}$ we have*

$$\mathbb{E}_P[\mathfrak{D}^2_{ist}(\widetilde{f}, f_1)] + \mathbb{E}_Q[\mathfrak{D}^2_{ist}(\widetilde{f}, f_{-1})] \geqslant \frac{1}{2}\mathfrak{D}^2_{ist}(f_1, f_{-1})\,\rho^2(P,Q).$$

Proof of §3.2.3. Consider the Hellinger affinity $\rho(P,Q) = \int\sqrt{dP}\sqrt{dQ}$ then by applying successively a triangular inequality and the Cauchy-Schwarz inequality it follows that

$$\begin{aligned} \rho(P,Q) &\leqslant \int \frac{\mathfrak{D}_{\rm ist}(\widetilde{f}_1, f_1)}{\mathfrak{D}_{\rm ist}(f_1, f_{-1})}\sqrt{dP}\sqrt{dQ} + \int\frac{\mathfrak{D}_{\rm ist}(\widetilde{f}, f_{-1})}{\mathfrak{D}_{\rm ist}(f_1, f_{-1})}\sqrt{dP}\sqrt{dQ} \\ &\leqslant \left(\mathbb{E}_P\frac{\mathfrak{D}^2_{\rm ist}(\widetilde{f}_1, f_1)}{\mathfrak{D}^2_{\rm ist}(f_1, f_{-1})}\mathbb{E}_Q(1)\right)^{1/2} + \left(\mathbb{E}_P(1)\mathbb{E}_Q\frac{\mathfrak{D}^2_{\rm ist}(\widetilde{f}, f_{-1})}{\mathfrak{D}^2_{\rm ist}(f_1, f_{-1})}\right)^{1/2} \end{aligned}$$

Rewriting the last estimate we obtain the result. □

§3.2.4 **Lower bound based on two hypothesis**. *Consider a family $\mathcal{P}_{\mathcal{F}}$ of probability measures. Given a noise level $\varepsilon \in (0,1)$ and $f_1, f_2 \in \mathcal{F}$ with associated probability measures $P_1 := P_{f_1,\varepsilon}$ and $P_2 := P_{f_2,\varepsilon}$ in $\mathcal{P}_{\mathcal{F}}$ such that $H(P_1, P_2) \leqslant 1$ we have*

$$\inf_{\widetilde{f}} \mathcal{R}_\varepsilon\big[\widetilde{f}\,|\,\mathcal{P}_{\mathcal{F}}\big] \geqslant \frac{1}{16}\mathfrak{D}^2_{ist}(f_1, f_2).$$

§1. If $\widetilde{f}$ denotes an estimator of f then we have

$$\sup_{f\in\mathcal{F}} \mathbb{E}_{f,\varepsilon}[\mathfrak{D}^2_{\text{ist}}(\widetilde{f},f)] \geqslant \max\Big\{\mathbb{E}_1[\mathfrak{D}^2_{\text{ist}}(\widetilde{f},f_1)],\mathbb{E}_2[\mathfrak{D}^2_{\text{ist}}(\widetilde{f},f_2)]\Big\}$$
$$\geqslant \frac{1}{2}\Big\{\mathbb{E}_1[\mathfrak{D}^2_{\text{ist}}(\widetilde{f},f_1)]+\mathbb{E}_2[\mathfrak{D}^2_{\text{ist}}(\widetilde{f},f_2)\Big\}.$$

Combining the last estimate with §3.2.3 implies

$$\mathcal{R}_\varepsilon\big[\widetilde{f}\,|\,\mathcal{P}_{\mathcal{F}}\big] \geqslant \frac{1}{4}\,\mathfrak{D}^2_{\text{ist}}(f_1,f_2)\,\rho^2(P_1,P_2).$$

Next, we bound the Hellinger affinity $\rho(P_1,P_2)$ from below. Using the relationship §3.2.2 c) and $H^2(P_1,P_2)\leqslant 1$ we obtain $\rho(P_1,P_2)\geqslant 1/2$, which implies the result. □

§3.2.5 ***Remark***. Let $\mathcal{P}_{\mathcal{F},\Xi}=\{P_{f,\xi,\varepsilon}\otimes P_{\xi,\sigma}, f\in\mathcal{F},\xi\in\Xi;\ \varepsilon,\sigma\in(0,1)\}$ be a family of product measures depending on a function of interest $f\in\mathcal{F}$, a nuissance parameter $\xi\in\Xi$ and noise levels $\varepsilon,\sigma\in(0,1)$. The last assertion allows us to bound from below the maximal risk for each nuissance parameter $\xi\in\Xi$ and noise level $\sigma\in(0,1)$. To be precise, given a noise level $\varepsilon\in(0,1)$ and $f_1,f_2\in\mathcal{F}$ with associated probability measures $P_1:=P_{f_1,\xi,\varepsilon}$ and $P_2:=P_{f_2,\xi,\varepsilon}$ such that $P_1\otimes P_{\xi,\sigma}$ and $P_2\otimes P_{\xi,\sigma}$ belongs to $\mathcal{P}_{\mathcal{F},\Xi}$ we have $\rho(P_1\otimes P_{\xi,\sigma},P_2\otimes P_{\xi,\sigma})=\rho(P_1,P_2)$ due to the independence. Consequently, if $H^2(P_1,P_2)\leqslant 1$, then

$$\sup_{f\in\mathcal{F}} \mathbb{E}_{f,\xi,\varepsilon,\sigma}[\mathfrak{D}^2_{\text{ist}}(\widetilde{f},f)] \geqslant \frac{1}{16}\mathfrak{D}^2_{\text{ist}}(f_1,f_2).$$

It is worth noting that we would derived the same lower bound when assuming the parameter ξ is known in advance. □

§3.2.6 ***Lower bound for a local risk***. Let the class of functions of interest be an ellipsoid $\mathcal{F}^r_{\mathfrak{a}}$. Consider a local Φ-risk associated with a linear functional Φ. In this situation the last assertion states

$$\inf_{\widetilde{f}} \mathcal{R}^\Phi_\varepsilon\big[\widetilde{f}\,|\,\mathcal{P}_{\mathcal{F}^r_{\mathfrak{a}}}\big] \geqslant \frac{1}{16}|\Phi(f_1-f_2)|^2.$$

If we consider furthermore candidates $f_1:=f_*$ and $f_2=-f_*$ for some $f_*\in\mathcal{F}^r_{\mathfrak{a}}$, then trivially $f_1,f_2\in\mathcal{F}^r_{\mathfrak{a}}$ and $|\Phi(f_1-f_2)|^2=4|\Phi(f_*)|^2$ which in turn implies due to the last assertion

$$\inf_{\widetilde{f}} \mathcal{R}^\Phi_\varepsilon\big[\widetilde{f}\,|\,\mathcal{P}_{\mathcal{F}^r_{\mathfrak{a}}}\big] \geqslant \frac{1}{4}|\Phi(f_*)|^2.$$

Often a minimax-optimal lower bound can be found by constructing a candidate $f_*\in\mathcal{F}$ that has the largest possible $|\Phi(f_*)|$-value but $P_{f_*,\varepsilon}$ and $P_{-f_*,\varepsilon}$ are still statistically indistinguishable in the sense that $H(P_{f_*,\varepsilon},P_{-f_*,\varepsilon})\leqslant 1$. □

§3.2.7 **GSSM (§3.1.1 continued) - lower bound of a Φ-risk**. Consider a linear functional Φ in a class $\mathcal{L}_{1/\mathfrak{a}}$ as given in §2.2.7. Define for all $\varepsilon\in(0,1)$ and $m\in\mathbb{N}$,

$$\mathcal{R}^m_\varepsilon(\Phi,\mathfrak{a}) := \max\left(\sum_{j>m}\frac{[\Phi]_j^2}{\mathfrak{a}_j},\max(\mathfrak{a}_m^{-1},\varepsilon)\sum_{j=1}^m[\Phi]_j^2\right);$$
$$\mathcal{R}^\star_\varepsilon(\Phi,\mathfrak{a}) := \min_{m\in\mathbb{N}}\mathcal{R}^m_\varepsilon(\Phi,\mathfrak{a});\quad \text{and}\quad m^\star_\varepsilon := \arg\min_{m\in\mathbb{N}}\mathcal{R}^m_\varepsilon(\Phi,\mathfrak{a}).$$

If, in addition, $\eta := \inf_{\varepsilon\in(0,1)}\big\{\min(\varepsilon\mathfrak{a}_{m_\varepsilon^\star}, (\varepsilon\mathfrak{a}_{m_\varepsilon^\star})^{-1})\big\} > 0$ then for all $\varepsilon \in (0,1)$

$$\inf_{\widetilde{f}} \mathcal{R}_\varepsilon^\Phi\big[\widetilde{f}\,|\,\mathcal{F}_\mathfrak{a}^r\big] \geqslant \frac{\eta}{8}\,\min(2r,1)\,\mathcal{R}_\varepsilon^\star\big(\Phi,\mathfrak{a}\big). \tag{3.1}$$

Proof. We intend to apply the result presented in §3.2.6 to two special choices of $f_\star \in \mathcal{F}_\mathfrak{a}^r$ with $2\|f_\star\|_\mathbb{H}^2 \leqslant \varepsilon$ to be specified below. Furthermore, for each $\theta \in \{-1,1\}$, P_θ is the distribution of the observations $\{[\widehat{g}_\varepsilon]_j, j\in\mathbb{N}\}$ if the parameter of interest equals f_θ. In this situation the observation $[\widehat{g}_\varepsilon]_j$ is distributed according to $\mathcal{N}(\theta[f_\star]_j,\varepsilon)$ for any $j\in\mathbb{N}$. It is easily seen that $[\log dP_1/dP_{-1}]([\widehat{g}_\varepsilon]) = 2\sum_{j\in\mathbb{N}}[\widehat{g}_\varepsilon]_j[f_\star]_j/\varepsilon$, and taking the expectation wrt. P_1 we find

$$KL(P_1,P_{-1}) = \frac{2}{\varepsilon}\sum_{j\in\mathbb{N}}[f_\star]_j^2 = \frac{2}{\varepsilon}\|f_\star\|_\mathbb{H}^2 \leqslant 1.$$

Using the relationship §3.2.2 b) we obtain $H(P_1,P_{-1}) \leqslant 1$, and hence the assumptions of §3.2.4 are satisfied. We will obtain the result by evaluating the bound presented in §3.2.6 for two choices of $f_\star \in \mathcal{F}_\mathfrak{a}^r$ with $2\|f_\star\|_\mathbb{H}^2 \leqslant \varepsilon$ which we will construct in the following. Define $K_\varepsilon^\star := \max(\mathfrak{a}_{m_\varepsilon^\star}^{-1},\varepsilon)$ and $\zeta := \eta\min(r,1/2)$. Observe that $\eta\max(1,(\varepsilon\mathfrak{a}_{m_\varepsilon^\star})^{-1}) \leqslant 1$ and $\eta\max(\varepsilon\mathfrak{a}_{m_\varepsilon^\star},1) \leqslant 1$, and hence $\eta K_\varepsilon^\star\max(\mathfrak{a}_{m_\varepsilon^\star},\varepsilon^{-1}) \leqslant 1$, which we will use below without further reference. On the one hand consider the function $f_\star := (\zeta\alpha_\varepsilon)^{1/2}\sum_{j=1}^{m_\varepsilon^\star}[\Phi]_j u_j$ with $\alpha_\varepsilon := K_\varepsilon^\star(\sum_{j=1}^{m_\varepsilon^\star}[\Phi]_j^2)^{-1}$ which belongs to $\mathcal{F}_\mathfrak{a}^r$. The latter is easily verified recalling that $\mathfrak{a}$ is non-decreasing the definition of ζ, α_ε and η imply

$$\|f_\star\|_\mathfrak{a}^2 = (\zeta\alpha_\varepsilon)\sum_{j=1}^{m_\varepsilon^\star}[\Phi]_j^2\mathfrak{a}_j \leqslant \zeta K_\varepsilon^\star\mathfrak{a}_{m_\varepsilon^\star} \leqslant \zeta\eta^{-1} \leqslant r.$$

It remains to show that $2\varepsilon^{-1}\|f_\star\|_\mathbb{H}^2 \leqslant \varepsilon^{-1}$ which can be seen as follows

$$2\varepsilon^{-1}\|f_\star\|_\mathbb{H}^2 = 2\varepsilon^{-1}(\zeta\alpha_\varepsilon)\sum_{j=1}^{m_\varepsilon^\star}[\Phi]_j^2 = 2\zeta K_\varepsilon^\star\varepsilon^{-1} \leqslant 2\zeta\eta^{-1} \leqslant 1.$$

Obviously, by evaluating the bound presented in §3.2.6 we can conclude

$$\inf_{\widetilde{f}} \mathcal{R}_\varepsilon^\Phi\big[\widetilde{f}\,|\,\mathcal{F}_\mathfrak{a}^r\big] \geqslant \frac{1}{4}|\Phi(f_\star)|^2 = \frac{\zeta}{4}\alpha_\varepsilon\Big|\sum_{j=1}^{m_\varepsilon^\star}[\Phi]_j^2\Big|^2 = \frac{\zeta}{4}\max(\mathfrak{a}_{m_\varepsilon^\star}^{-1},\varepsilon)\sum_{j=1}^{m_\varepsilon^\star}[\Phi]_j^2.$$

On the other hand, consider $f_\star := (\zeta\alpha_\varepsilon)^{1/2}\sum_{j>m_\varepsilon^\star}[\Phi]_j\mathfrak{a}_j^{-1}u_j$ with $\alpha_\varepsilon := (\sum_{j>m_\varepsilon^\star}[\Phi]_j^2\mathfrak{a}_j^{-1})^{-1}$ we conclude from $\eta \leqslant 1$ and $\|f_\star\|_\mathfrak{a}^2 = (\zeta\alpha_\varepsilon)\sum_{j>m_\varepsilon^\star}[\Phi]_j^2\mathfrak{a}_j^{-1} = \zeta \leqslant r$ that $f_\star$ belongs to $\mathcal{F}_\mathfrak{a}^r$. Moreover, we have $2\varepsilon^{-1}\|f_\star\|_\mathbb{H}^2 = (\zeta\alpha_\varepsilon)\sum_{j>m_\varepsilon^\star}[\Phi]_j^2\mathfrak{a}_j^{-2} \leqslant 2\zeta\varepsilon^{-1}\mathfrak{a}_{m_\varepsilon^\star}^{-1} \leqslant 2\zeta\eta^{-1} \leqslant 1$ where we used again that $\mathfrak{a}$ is non-decreasing. By evaluating the bound in §3.2.6 we obtain

$$\inf_{\widetilde{f}} \mathcal{R}_\varepsilon^\Phi\big[\widetilde{f}\,|\,\mathcal{F}_\mathfrak{a}^r\big] \geqslant \frac{1}{4}|\Phi(f_\star)|^2 = \frac{\zeta}{4}\alpha_\varepsilon\Big|\sum_{j>m_\varepsilon^\star}[\Phi]_j^2\mathfrak{a}_j^{-1}\Big|^2 = \frac{\zeta}{4}\sum_{j>m_\varepsilon^\star}[\Phi]_j^2\mathfrak{a}_j^{-1}.$$

Combining the two lower bounds we obtain the assertion. □

§3.2.8 **Lower bound based on two hypothesis**. *Consider a family $\mathcal{P}_{\mathcal{F},\Xi}$ of probability measures. Given noise level $\varepsilon, \sigma \in (0,1)$, nuissance parameters $\xi_1, \xi_2 \in \Xi$ and candidates $f_1, f_2 \in \mathcal{F}$ suppose that there exist associated probability measures $P_{f_1,\xi_1,\varepsilon} \otimes P_{\xi_1,\sigma}$ and $P_{f_2,\xi_2,\varepsilon} \otimes P_{\xi_2,\sigma}$ in $\mathcal{P}_{\mathcal{F},\Xi}$. If in addition $P_{f_1,\xi_1,\varepsilon} = P_{f_2,\xi_2,\varepsilon}$ and $H(P_{\xi_1,\sigma}, P_{\xi_2,\sigma}) \leqslant 1$ then we have*

$$\inf_{\widetilde{f}} \mathcal{R}_{\varepsilon,\sigma}[\widetilde{f} \,|\, \mathcal{P}_{\mathcal{F},\Xi}] \geqslant \frac{1}{16} \mathfrak{D}^2_{ist}(f_1, f_2).$$

Proof of §3.2.8. Following line by line the proof §1 the result is an immediate consequence of §3.2.3 and $\rho(P_{f_1,\xi_1,\varepsilon} \otimes P_{\xi_1,\sigma}, P_{f_2,\xi_2,\varepsilon} \otimes P_{\xi_2,\sigma}) = \rho(P_{\xi_1,\sigma}, P_{\xi_2,\sigma})$ exploiting the independence. □

§3.2.9 ***Remark***. The last assertion allows us often to derive a lower bound depending on the classes $\mathcal{F}$ and Ξ and the noise level σ but not on the noise level ε. Roughly speaking this means that we cover the influence of the estimation of the nuissance parameter. Typically we combine this lower bound with the lower bound obtained in §3.2.5 where the nuissance parameter is assumed to be known in advance. □

§3.2.10 **Assouad's cube technique**. *Consider a family $\mathcal{P}_{\mathcal{F}}$ of probability measures. Suppose there exist distances $\mathfrak{D}^{(j)}_{ist}$, $j = 1, \dots, m$, such that $\mathfrak{D}^2_{ist}(\cdot,\cdot) \geqslant \sum_{j=1}^m |\mathfrak{D}^{(j)}_{ist}(\cdot,\cdot)|^2$. Let $\theta := (\theta_j)_{j=1}^m \in \{-1,1\}^m =: \Theta$ and for each $\theta \in \Theta$ introduce $\theta^{(j)} \in \Theta$ by $\theta^{(j)}_l = \theta_l$ for $j \neq l$ and $\theta^{(j)}_j = -\theta_j$. Given a noise level $\varepsilon \in (0,1)$ for all $\{f_\theta, \theta \in \Theta\} \subset \mathcal{F}$ with associated probability measures $P_\theta := P_{f_\theta,\varepsilon} \in \mathcal{P}_{\mathcal{F}}$, $\theta \in \Theta$, satisfying $H(P_\theta, P_{\theta^{(j)}}) \leqslant 1$ for all $1 \leqslant j \leqslant m$ and $\theta \in \Theta$, we have*

$$\inf_{\widetilde{f}} \mathcal{R}_{\varepsilon}[\widetilde{f} \,|\, \mathcal{P}_{\mathcal{F}}] \geqslant \frac{1}{2^m} \sum_{\theta\in\Theta} \frac{1}{16} \sum_{j=1}^m |\mathfrak{D}^{(j)}_{ist}(f_\theta, f_{\theta^{(j)}})|^2.$$

Proof of §3.2.10. If $\widetilde{f}$ denotes an estimator of f then we have

$$\sup_{f\in\mathcal{F}} \mathbb{E}_{f,\varepsilon} \mathfrak{D}^2_{ist}(\widetilde{f}, f) \geqslant \sup_{\theta\in\Theta} \mathbb{E}_\theta \mathfrak{D}^2_{ist}(\widetilde{f}, f_\theta) \geqslant \frac{1}{2^m} \sum_{\theta\in\Theta} \mathbb{E}_\theta \mathfrak{D}^2_{ist}(\widetilde{f}, f_\theta)$$
$$\geqslant \frac{1}{2^m} \sum_{j=1}^m \sum_{\theta\in\Theta} \mathbb{E}_\theta |\mathfrak{D}^{(j)}_{ist}(\widetilde{f}, f_\theta)|^2 \geqslant \frac{1}{2^m} \sum_{j=1}^m \sum_{\theta\in\Theta} \frac{1}{2} \Big\{ \mathbb{E}_\theta |\mathfrak{D}^{(j)}_{ist}(\widetilde{f}, f_\theta)|^2 + \mathbb{E}_{\theta^{(j)}} |\mathfrak{D}^{(j)}_{ist}(\widetilde{f}, f_{\theta^{(j)}})|^2 \Big\}$$

where we used that $\sum_{\theta\in\Theta} F(\theta) = \sum_{\theta\in\Theta} F(\theta^{(j)})$ which can be seen as follows. Define $\Theta_j^+ := \{\theta \in \Theta | \theta_j = +1\}$ and $\Theta_j^- := \{\theta \in \Theta | \theta_j = -1\}$. It is easily verified that $\Theta_j^+ = \{\theta^{(j)} | \theta \in \Theta_j^-\}$ and $\Theta_j^- = \{\theta^{(j)} | \theta \in \Theta_j^+\}$. Using further $\Theta = \Theta_j^+ \cup \Theta_j^-$ and $\Theta_j^+ \cap \Theta_j^- = \emptyset$ we have

$$\sum_{\theta\in\Theta} F(\theta) = \sum_{\theta\in\Theta_j^+} F(\theta) + \sum_{\theta\in\Theta_j^-} F(\theta) = \sum_{\theta\in\Theta_j^-} F(\theta^{(j)}) + \sum_{\theta\in\Theta_j^+} F(\theta^{(j)}) = \sum_{\theta\in\Theta} F(\theta^{(j)}).$$

The result follows by combining the last estimate with §3.2.3 and $H^2(P_\theta, P_{\theta^{(j)}}) \leqslant 1$. □

§3.2.11 ***Remark***. Consider a family $\mathcal{P}_{\mathcal{F},\Xi} = \{P_{f,\xi,\varepsilon} \otimes \mathcal{P}_{\xi,\sigma}, f \in \mathcal{F}, \xi \in \Xi; \varepsilon, \sigma \in (0,1)\}$ of product measures. The last assertion allows us to bound from below the maximal risk for each nuissance parameter $\xi \in \Xi$ and noise level $\sigma \in (0,1)$. To be precise, given a noise level $\varepsilon \in (0,1)$ and $f_\theta \in \mathcal{F}$, $\theta \in \Theta$, let the associated probability measures $P_\theta := P_{f_\theta,\xi,\varepsilon}$ be such that

$P_\theta \otimes P_{\xi,\sigma} \in \mathcal{P}_{\mathcal{F},\Xi}$, $\theta \in \Theta$. If in addition $H(P_\theta, P_{\theta^{(j)}}) \leqslant 1$ for all $1 \leqslant j \leqslant m$ and $\theta \in \Theta$, then we have

$$\inf_{\widetilde{f}} \mathcal{R}_{\varepsilon,\sigma}\big[\widetilde{f}\,\big|\,\mathcal{P}_{\mathcal{F},\Xi}\big] \geqslant \frac{1}{2^m} \sum_{\theta\in\Theta} \frac{1}{16} \sum_{j=1}^{m} |\mathfrak{D}^{(j)}_{\text{ist}}(f_\theta, f_{\theta^{(j)}})|^2.$$

It is worth noting that we would derived the same lower bound when assuming the parameter ξ is known in advance. Based on the last assertion in circular deconvolution with known error density (see §2.8.11), for example, we will derive a lower bound which involves the Y-sample size n. This lower bound, however, we will combine in circular deconvolution with unknown error density (see §2.8.18) with a lower bound employing the result stated in §3.2.8 which depends only on the U-sample size m and hence covers the influence of the estimation of the error density. □

§3.2.12 ***Lower bound of a global risk.*** Let the class of functions of interest be an ellipsoid $\mathcal{F}^r_{\mathfrak{a}}$. Consider a global $\mathbb{H}_{\mathfrak{v}}$-risk with weighted norm $\|\cdot\|_{\mathfrak{v}}$ derived from the ONB associated with $\mathcal{F}^r_{\mathfrak{a}}$ and some weight sequence $(\mathfrak{v}_j)_{j\in\mathcal{J}}$. In this situation the last assertion states

$$\inf_{\widetilde{f}} \mathcal{R}^{\mathfrak{v}}_{\varepsilon}\big[\widetilde{f}\,\big|\,\mathcal{P}_{\mathcal{F}^r_{\mathfrak{a}}}\big] \geqslant \frac{1}{2^m} \sum_{\theta\in\Theta} \frac{1}{16} \sum_{j=1}^{m} \mathfrak{v}_j |[f_\theta]_j - [f_{\theta^{(j)}}]_j|^2.$$

If we consider furthermore candidates $f_\theta := \sum_{j=1}^m \theta_j [f_*]_j u_j$, $\theta \in \Theta$, for some $f_* \in \mathcal{F}^r_{\mathfrak{a}}$, then it is easily verified that $\{f_\theta, \theta \in \Theta\} \subset \mathcal{F}^r_{\mathfrak{a}}$ and $\sum_{j=1}^m \mathfrak{v}_j |[f_\theta]_j - [f_{\theta^{(j)}}]_j|^2 = 4 \sum_{j=1}^m \mathfrak{v}_j |[f_*]_j|^2 = 4\|\Pi_{\mathbb{U}_m} f_*\|^2_{\mathfrak{v}}$ which in turn implies by applying the last assertion

$$\inf_{\widetilde{f}} \mathcal{R}^{\mathfrak{v}}_{\varepsilon}\big[\widetilde{f}\,\big|\,\mathcal{P}_{\mathcal{F}}\big] \geqslant \frac{1}{2^m} \sum_{\theta\in\Theta} \frac{1}{4} \|\Pi_{\mathbb{U}_m} f_*\|^2_{\mathfrak{v}} = \frac{1}{4} \|\Pi_{\mathbb{U}_m} f_*\|^2_{\mathfrak{v}}.$$

Often a minimax-optimal lower bound can be found by choosing the parameter m and the function f_* that have the largest possible $\|\Pi_{\mathbb{U}_m} f_*\|^2_{\mathfrak{v}}$-value although that the associated $P_{f_\theta,\varepsilon} \in \mathcal{P}_{\mathcal{F}}$, $\theta \in \Theta$, are still statistically indistinguishable in the sense that $H(P_{f_\theta,\varepsilon}, P_{f_{\theta^{(j)}},\varepsilon}) \leqslant 1$ for all $1 \leqslant j \leqslant m$ and $\theta \in \Theta$. □

§3.2.13 **GSSM (§3.1.1 continued) - lower bound of a $\mathbb{H}_{\mathfrak{v}}$-risk.** Let $\mathfrak{a}$ and $\mathfrak{v}$ be weight sequences such that $\mathfrak{v}\mathfrak{a}^{-1}$ is non-increasing. Define for all $\varepsilon \in (0,1)$ and $m \in \mathbb{N}$,

$$\mathcal{R}^m_\varepsilon\big(\mathfrak{v},\mathfrak{a}\big) := \max(\mathfrak{v}_m \mathfrak{a}_m^{-1}, \sum_{j=1}^m \varepsilon \mathfrak{v}_j); \quad \mathcal{R}^\star_\varepsilon := \mathcal{R}^\star_\varepsilon\big(\mathfrak{v},\mathfrak{a}\big) := \min_{m\in\mathbb{N}} \mathcal{R}^m_\varepsilon\big(\mathfrak{v},\mathfrak{a}\big); \quad \text{and}$$

$$m^\star_\varepsilon := \arg\min_{m\in\mathbb{N}} \mathcal{R}^m_\varepsilon\big(\mathfrak{v},\mathfrak{a}\big)$$

If, in addition, $\eta := \inf_{\varepsilon\in(0,1)} \Big\{(\mathcal{R}^\star_\varepsilon)^{-1} \min(\mathfrak{v}_{m^\star_\varepsilon} \mathfrak{a}^{-1}_{m^\star_\varepsilon}, \sum_{j=1}^{m^\star_\varepsilon} \varepsilon \mathfrak{v}_j)\Big\} > 0$ then for all $\varepsilon \in (0,1)$

$$\inf_{\widetilde{f}} \mathcal{R}^{\mathfrak{v}}_{\varepsilon}\big[\widetilde{f}\,\big|\,\mathcal{F}^r_{\mathfrak{a}}\big] \geqslant \frac{\eta}{8}\ \min(2r,1)\ \mathcal{R}^\star_\varepsilon\big(\mathfrak{v},\mathfrak{a}\big). \tag{3.2}$$

Proof. We intend to apply the result presented in §3.2.12. Let $\alpha_\varepsilon := (\sum_{j=1}^{m_\varepsilon^\star} \varepsilon\mathfrak{v}_j)^{-1}\mathcal{R}_\varepsilon^\star \leqslant \eta^{-1}$ and $\zeta := \eta\min(r, 1/2)$. Define the function $f_\star := (\varepsilon\zeta\alpha_\varepsilon)^{1/2}\sum_{j=1}^{m_\varepsilon^\star} u_j$ which belongs to $\mathcal{F}_{\mathfrak{a}}^r$. The latter is easily verified recalling that $\mathfrak{v}^{-1}\mathfrak{a}$ is non-decreasing the definition of ζ, α_ε and η imply

$$\|f_\star\|_{\mathfrak{a}} = (\varepsilon\zeta\alpha_\varepsilon)\sum_{j=1}^{m_\varepsilon^\star}\mathfrak{a}_j \leqslant \zeta\frac{\mathfrak{a}_{m_\varepsilon^\star}}{\mathfrak{v}_{m_\varepsilon^\star}}\alpha_\varepsilon\sum_{j=1}^{m_\varepsilon^\star}\varepsilon\mathfrak{v}_j = \zeta\frac{\mathfrak{a}_{m_\varepsilon^\star}}{\mathfrak{v}_{m_\varepsilon^\star}}\mathcal{R}_\varepsilon^\star \leqslant \zeta\eta^{-1} \leqslant r.$$

Furthermore, for a fixed θ, P_θ is the distribution of the observations $\{[\widehat{g}_\varepsilon]_j, j \in \mathbb{N}\}$ if the parameter of interest equals f_θ. In this situation the observation $[\widehat{g}_\varepsilon]_j$ is distributed according to $\mathcal{N}(\theta_j[f_\star]_j, \varepsilon)$ for any $j \in \mathbb{N}$. Considering P_θ and $P_{\theta^{(j)}}$ it is easily seen that $[\log dP_\theta/dP_{\theta^{(j)}}]([\widehat{g}_\varepsilon]) = 2[\widehat{g}_\varepsilon]_j\theta_j[f_\star]_j/\varepsilon$, and taking the expectation wrt. P_θ we find

$$KL(P_\theta, P_{\theta^{(j)}}) = \frac{2}{\varepsilon}\,[f_\star]_j^2 = 2\zeta\alpha_\varepsilon \leqslant 1.$$

Using the relationship §3.2.2 b) we obtain $H(P_\theta, P_{\theta^{(j)}}) \leqslant 1$ for all $1 \leqslant j \leqslant m$ and $\theta \in \Theta$. Thereby, the assumptions of §3.2.10 are satisfied and hence we can apply the bound presented in §3.2.12. Combining this lower bound and $\|\Pi_{\mathbb{U}_m} f_\star\|_{\mathfrak{v}}^2 = \zeta\alpha_\varepsilon\sum_{j=1}^{m_\varepsilon^\star}\varepsilon\mathfrak{v}_j = \zeta\mathcal{R}_\varepsilon^\star$ implies the assertion. □

3.3 Estimation by dimension reduction

Throughout this note we will estimate the function of interest $f \in \mathbb{H}$ using a dimension reduction. To be more precise, let $(u_j)_{j\in\mathcal{J}}$ be an ONB in $\mathbb{H}$. With respect to this basis, for $f \in \mathbb{H}$ we consider the expansion $f = \sum_{j\in\mathcal{J}}[f]_j u_j$ of f where $[f] = ([f]_j)_{j\in\mathcal{J}}$ is the sequence of its generalised Fourier coefficients. Given a dimension parameter $m \in \mathbb{N}$ we have the subspace $\mathbb{U}_m = \overline{\mathrm{lin}}\,\{u_1, \dots, u_m\}$ spanned by the first m basis functions at our disposal. Let $f_m \in \mathbb{U}_m$ be a theoretical approximation of f with associated approximation error $\mathrm{bias}_m := \sup_{k\geqslant m}\mathfrak{D}_{\mathrm{ist}}(f_k, f)$ tending to zero as $m \to \infty$. We restrict ourselves to two particular cases. On the one hand we consider the orthogonal projection $f_m = \Pi_{\mathbb{U}_m} f$ of f onto $\mathbb{U}_m$ where $\mathrm{bias}_m = o(1)$ as $m \to \infty$ for all $f \in \mathbb{H}$. On the other hand f_m denotes a Galerkin solution in $\mathbb{U}_m$. In this case $\mathrm{bias}_m = o(1)$ as $m \to \infty$ holds true only under additional regularity conditions (compare Remark §2.2.18). In both cases the approximation f_m is uniquely determined by the vector of coefficients $[f_m]_{\underline{m}}$ since $[f_m]_j = 0$ for all $j > m$. Here and subsequently the observations will allow us to construct an estimator $\widehat{[f_m]}_{\underline{m}}$ of the coefficient vector $[f_m]_{\underline{m}}$. Having this estimator at hand, we will consider the emprirical counterpart $\widehat{f}_m := \sum_{j=1}^m \widehat{[f_m]}_j u_j \in \mathbb{H}$ of f_m as a possible estimator of f.

§3.3.1 **Definition**. *The estimator $\widehat{f}_m$ is called orthogonal series estimator (OSE), if $f_m = \Pi_{\mathbb{U}_m} f$ is the orthogonal projection of f onto $\mathbb{U}_m$ while it is called Galerkin estimator (GE), if f_m is a Galerkin solution in $\mathbb{U}_m$.*

We shall measure the performance of the estimator $\widehat{f}_m$ within a minimax framework and hence consider its maximal risk $\mathcal{R}_{\varepsilon,\sigma}\big[\widehat{f}_m \,|\, \mathcal{P}_{\mathcal{F},\Xi}\big] := \sup_{f\in\mathcal{F}}\sup_{\xi\in\Xi}\mathbb{E}_{f,\xi,\varepsilon,\sigma}[\mathfrak{D}_{\mathrm{ist}}^2(\widehat{f}_m, f)]$ wrt. a $\mathfrak{D}_{\mathrm{ist}}$-function over certain classes $\mathcal{F}$ and Ξ of solutions and nuisances parameters respectively. We will show in the next chapter that $\widehat{f}_m$ can attain optimal rates of convergence over a variety of classes $\mathcal{F}$ and Ξ provided the dimension parameter is chosen appropriately.

§3.3.2 **GSSM (§3.1.1 continued) - minimax optimality**. *Preliminaries.* Keep in mind that given observations as in §3.1.1 for each $j \in \mathbb{N}$, $\widehat{[f]}_j := [\widehat{g}_\varepsilon]_j$ is the unique best unbiased estimator of $[f]_j$ due to Lehman-Scheffé's Theorem. Consequently, introducing a dimension parameter $m \in \mathbb{N}$ we consider the orthogonal series estimator (OSE) $\widehat{f}_m := \sum_{j=1}^m \widehat{[f]}_j u_j$. Observe that $[\widehat{f}]_{\underline{m}} = [\widehat{g}_\varepsilon]_{\underline{m}}$, where trivially $[\widehat{f}]_{\underline{m}} \sim \mathcal{N}([f]_{\underline{m}}, \varepsilon[\mathrm{Id}]_{\underline{m}})$ and, hence $\mathbb{E}\|\widehat{f}_m - f_m\|_{\mathfrak{v}}^2 = 2\varepsilon \sum_{j=1}^m \mathfrak{v}_j$ and $[\Phi]_{\underline{m}}^t [\widehat{f}]_{\underline{m}} \sim \mathcal{N}([\Phi]_{\underline{m}}^t [f]_{\underline{m}}, \varepsilon[\Phi]_{\underline{m}}^t [\Phi]_{\underline{m}})$. We exploit these properties in the following proofs.

Local Φ-risk: Let $\mathfrak{a}$ be a non-decreasing sequence and let $\Phi \in \mathcal{L}_{1/\mathfrak{a}}$. Consider the OSE $\widehat{f}_{m_\varepsilon^\star}$ with dimension parameter $m_\varepsilon^\star$ defined as in §3.2.7. For all $\varepsilon \in (0,1)$ we have

$$\mathcal{R}_\varepsilon^\Phi\big[\widehat{f}_{m_\varepsilon^\star} \,|\, \mathcal{F}_\mathfrak{a}^r\big] \leqslant (1+r)\, \mathcal{R}_\varepsilon^\star\big(\Phi, \mathfrak{a}\big).$$

Proof. For all $m \in \mathbb{N}$ we have

$$\begin{aligned}\mathbb{E}|\Phi(\widehat{f}_m) - \Phi(f)|^2 &= \mathbb{V}\mathrm{ar}(\Phi(\widehat{f}_m)) + |\mathbb{E}\Phi(\widehat{f}_m) - \Phi(f)|^2 \\ &= \varepsilon \sum_{j=1}^m [\Phi]_j^2 + |\Phi(\Pi_{\mathbb{U}_m^\perp} f)|^2 = \varepsilon \sum_{j=1}^m [\Phi]_j^2 + |\sum_{j>m} [\Phi]_j [f]_j|^2\end{aligned}$$

and $|\sum_{j>m}[\Phi]_j[f]_j|^2 \leqslant r \sum_{j\geqslant m} [\Phi]_j^2 \mathfrak{a}_j^{-1}$ due to §2.2.15, since $f \in \mathcal{F}_\mathfrak{a}^r$. Combining the upper bounds and the definition of $\mathcal{R}_\varepsilon^\star\big(\Phi,\mathfrak{a}\big)$ given in §3.2.7 implies the result.

Global $\mathbb{H}_\mathfrak{v}$-risk: Let $\mathfrak{v}$ be a weight sequence such that $\mathfrak{v}\mathfrak{a}^{-1}$ is non-decreasing. Consider the OSE $\widehat{f}_{m_\varepsilon^\star}$ with dimension parameter $m_\varepsilon^\star$ as defined in §3.2.13. For all $\varepsilon \in (0,1)$ we have

$$\mathcal{R}_\varepsilon^\mathfrak{v}\big[\widehat{f}_{m_\varepsilon^\star} \,|\, \mathcal{F}_\mathfrak{a}^r\big] \leqslant (2+r)\, \mathcal{R}_\varepsilon^\star\big(\mathfrak{v}, \mathfrak{a}\big).$$

Proof. From the Pythagorean formula we obtain for all $m \in \mathbb{N}$

$$\mathbb{E}\|\widehat{f}_m - f\|_\mathfrak{v}^2 = \mathbb{E}\|\widehat{f}_m - \Pi_{\mathbb{U}_m} f\|_\mathfrak{v}^2 + \|\Pi_{\mathbb{U}_m^\perp} f\|_\mathfrak{v}^2 = 2\varepsilon \sum_{j=1}^m \mathfrak{v}_j + \|\Pi_{\mathbb{U}_m^\perp} f\|_\mathfrak{v}^2.$$

The result follows by exploiting §2.2.15, that is, $\|\Pi_{\mathbb{U}_m^\perp} f\|_\mathfrak{v}^2 \leqslant r\mathfrak{a}_m^{-1}\mathfrak{v}_m$ for all $f \in \mathcal{F}_\mathfrak{a}^r$, and the definition of $\mathcal{R}_\varepsilon^\star\big(\mathfrak{v},\mathfrak{a}\big)$ given in §3.2.13.

Since the last two upper bounds coincide up to a constant with the corresponding lower bounds given in §3.1 and §3.2 it follows that the rates $\mathcal{R}_\varepsilon^\star\big(\Phi,\mathfrak{a}\big)$ and $\mathcal{R}_\varepsilon^\star\big(\mathfrak{v},\mathfrak{a}\big)$ are minimax-optimal and the OSE is minimax-optimal up to a constant. □

The optimal performance of the estimator $\widehat{f}_m$ depends crucially on the choice of the tuning parameter m, which in turn, relies strongly on a *priori* knowledge of the sets $\mathcal{F}$ and Ξ. However, this information is widely inaccessible in practice. Therefore, we will propose in the sequel a fully data-driven procedure for the selection of the dimension parameter.

3.4 Data-driven estimation procedures

Given a dimension parameter $m \in \mathbb{N}$ let $\widehat{f}_m \in \mathbb{U}_m$ be an estimator of the unknown solution $f \in \mathbb{H}$. Moreover, let $f_m \in \mathbb{U}_m$ be a theoretical approximation of f where the approximation error $\operatorname{bias}_m := \sup_{k\geqslant m} \mathfrak{D}_{\text{ist}}(f_k, f)$ tends to zero as m increases. In what follows we construct a data-driven procedure to choose the dimension parameter m based first on Lepski's method and second a model selection approach. Hence, the symbols for the contrast, penalty and the selected dimension parameter may denote different objects according to the specific case.

3.4.1 Lepski's method

Let us consider a family $\mathcal{P}_{\mathcal{F}}$ of probability measures. Given an integer M and a subsequence of non-negative and non-decreasing penalties $(\operatorname{pen}_1, \dots, \operatorname{pen}_{\mathrm{M}})$ define the contrast

$$\Upsilon_m := \max_{m\leqslant k\leqslant M} \Big\{\mathfrak{D}^2_{\text{ist}}(\widehat{f}_k, \widehat{f}_m) - \operatorname{pen}_k\Big\}, \quad 1 \leqslant m \leqslant \mathrm{M}. \tag{3.3}$$

Observe that $(\Upsilon_1, \dots, \Upsilon_{\mathrm{M}})$ is monotonically decreasing with $\Upsilon_{\mathrm{M}} \leqslant 0$. Following Lepski's method we select the dimension parameter $\widetilde{m}$ among the random collection of admissible values $\{1, \dots, \mathrm{M}\}$, that is

$$\widetilde{m} := \min \{1 \leqslant m \leqslant \mathrm{M} : \Upsilon_m \leqslant 0\}. \tag{3.4}$$

The data-driven estimator of f is now given by $\widehat{f}_{\widetilde{m}}$ and below we derive an upper bound for its maximal risk $\mathcal{R}_\varepsilon\big[\widehat{f}_{\widetilde{m}} \,|\, \mathcal{P}_{\mathcal{F}}\big] = \sup_{f\in\mathcal{F}} \mathbb{E}_{f,\varepsilon}[\mathfrak{D}^2_{\text{ist}}(\widehat{f}_{\widetilde{m}}, f)]$ over certain classes of solutions $\mathcal{F}$. The construction of the penalty sequence pen and the upper bound M given below is motivated by the key arguments used in the proof of the risk bound which we present first. Moreover, both pen and M will depend, among others, on the noise level ε, however, for sake of simplicity we will omit an additional subscript. The key argument for our reasoning is the next assertion.

§3.4.1 **Lemma (key argument).** *For all* $1 \leqslant m \leqslant \mathrm{M}$ *we have*

$$\mathfrak{D}^2_{\text{ist}}(\widehat{f}_{\widetilde{m}}, f) \leqslant 10 \max(\operatorname{bias}^2_m, \operatorname{pen}_m) + 4 \max_{m\leqslant k\leqslant \mathrm{M}} \Big(\mathfrak{D}^2_{\text{ist}}(\widehat{f}_k, f_k) - \operatorname{pen}_k\Big)_+ + 2 \operatorname{pen}_{\mathrm{M}} \mathbb{1}_{\{\widetilde{m}>m\}}$$

where $(a)_+ = \max(a, 0)$.

Proof of §3.4.1. Given a dimension parameter $1 \leqslant m \leqslant \mathrm{M}$ let us introduce the events $\{\widetilde{m} \leqslant m\}$ and $\{\widetilde{m} > m\}$. By making successively use of the definition (3.3) of Υ and (3.4) of $\widetilde{m}$ we have

$$\begin{aligned}
\mathfrak{D}^2_{\text{ist}}(\widehat{f}_{\widetilde{m}}, f)\mathbb{1}_{\{\widetilde{m}\leqslant m\}} &\leqslant 2\Big(\mathfrak{D}^2_{\text{ist}}(\widehat{f}_m, \widehat{f}_{\widetilde{m}}) + \mathfrak{D}^2_{\text{ist}}(\widehat{f}_m, f)\Big)\mathbb{1}_{\{\widetilde{m}\leqslant m\}} \\
&= 2\Big(\Big\{\mathfrak{D}^2_{\text{ist}}(\widehat{f}_m, \widehat{f}_{\widetilde{m}}) - \operatorname{pen}_m\Big\} + \operatorname{pen}_m + \mathfrak{D}^2_{\text{ist}}(\widehat{f}_m, f)\Big)\mathbb{1}_{\{\widetilde{m}\leqslant m\}} \\
&\leqslant 2\Big\{\Upsilon_{\widetilde{m}} + \operatorname{pen}_m + \mathfrak{D}^2_{\text{ist}}(\widehat{f}_m, f)\Big\}\mathbb{1}_{\{\widetilde{m}\leqslant m\}} \\
&\leqslant 2\Big\{\operatorname{pen}_m + \mathfrak{D}^2_{\text{ist}}(\widehat{f}_m, f)\Big\}\mathbb{1}_{\{\widetilde{m}\leqslant m\}} \\
&\leqslant 2\Big\{3\operatorname{pen}_m + 2\Big(\mathfrak{D}^2_{\text{ist}}(\widehat{f}_m, f_m) - \operatorname{pen}_m\Big)_+ + 2\mathfrak{D}^2_{\text{ist}}(f_m, f)\Big\}\mathbb{1}_{\{\widetilde{m}\leqslant m\}} \\
&\leqslant 10 \max(\operatorname{bias}^2_m, \operatorname{pen}_m)\mathbb{1}_{\{\widetilde{m}\leqslant m\}} + 4\Big(\mathfrak{D}^2_{\text{ist}}(\widehat{f}_m, f_m) - \operatorname{pen}_m\Big)_+\mathbb{1}_{\{\widetilde{m}\leqslant m\}}
\end{aligned}$$

where the last estimate uses the definition of bias, only. On the other hand on the event $\{\widetilde{m} > m\}$ the monotony of bias and pen implies

$$\begin{aligned}
\mathfrak{D}^2_{\text{ist}}(\widehat{f}_{\widetilde{m}}, f)\mathbb{1}_{\{\widetilde{m}>m\}} &= \sum_{k=m+1}^{\mathrm{M}} \mathfrak{D}^2_{\text{ist}}(\widehat{f}_k, f)\mathbb{1}_{\{\widetilde{m}=k\}} \\
&\leqslant 2 \sum_{k=m+1}^{\mathrm{M}} \Big\{\mathfrak{D}^2_{\text{ist}}(\widehat{f}_k, f_k) + \mathfrak{D}^2_{\text{ist}}(f_k, f)\Big\}\mathbb{1}_{\{\widetilde{m}=k\}} \\
&\leqslant 2 \sum_{k=m+1}^{\mathrm{M}} \Big\{\mathfrak{D}^2_{\text{ist}}(\widehat{f}_k, f_k) + \mathrm{bias}_k^2\Big\}\mathbb{1}_{\{\widetilde{m}=k\}} \\
&\leqslant 2 \sum_{k=m+1}^{\mathrm{M}} \mathfrak{D}^2_{\text{ist}}(\widehat{f}_k, f_k)\mathbb{1}_{\{\widetilde{m}=k\}} + 2\,\mathrm{bias}_m^2\,\mathbb{1}_{\{\widetilde{m}>m\}} \\
&\leqslant 2 \max_{m+1\leqslant k\leqslant \mathrm{M}} \Big(\mathfrak{D}^2_{\text{ist}}(\widehat{f}_k, f_k)| - \mathrm{pen}_k\Big)_+ \mathbb{1}_{\{\widetilde{m}>m\}} \\
&\qquad + 2\,\mathrm{pen}_{\mathrm{M}}\,\mathbb{1}_{\{\widetilde{m}>m\}} + 2\,\mathrm{bias}_m^2\,\mathbb{1}_{\{\widetilde{m}>m\}}
\end{aligned}$$

Combining the bounds on the two events $\{\widetilde{m} \leqslant m\}$ and $\{\widetilde{m} > m\}$ implies the result. □

Let $m^\diamond$ realise a penalty-squared-bias compromise among the collection of admissible values $\{1, \dots, \mathrm{M}\}$ which we specify below. Due to the last assertion we have for all $P_{f,\varepsilon} \in \mathcal{P}_{\mathcal{F}}$

$$\mathbb{E}_{f,\varepsilon}\mathfrak{D}^2_{\text{ist}}(\widehat{f}_{\widetilde{m}}, f) \leqslant 10\max(\mathrm{bias}^2_{m^\diamond}, \mathrm{pen}_{m^\diamond}) + 4\mathbb{E}_{f,\varepsilon}\Big\{\max_{m^\diamond\leqslant k\leqslant \mathrm{M}}\Big(\mathfrak{D}^2_{\text{ist}}(\widehat{f}_k, f_k) - \mathrm{pen}_k\Big)_+\Big\} + 2\,\mathrm{pen}_{\mathrm{M}}\,P_{f,\varepsilon}\big(\widetilde{m} > m^\diamond\big). \quad (3.5)$$

Keeping in mind that $m^\diamond$ should mimic the value of the optimal variance-squared-bias trade-off, we wish the upper bound M to be as large as possible. In contrast, in order to control both remainder terms, the second and third right hand side (rhs.) term in (3.5), we a forced to use a rather small upper bound M. However, we bound the second remainder term by imposing the following assumption, which though holds true for a wide range of classes $\mathcal{F}$ under reasonable model assumptions.

§3.4.2 **Assumption**. *There exists a constant $C_1 := C_1(\mathcal{P}_{\mathcal{F}})$ possibly depending on the family $\mathcal{P}_{\mathcal{F}}$ such that*

$$\sup_{f\in\mathcal{F}} \mathrm{pen}_{\mathrm{M}}\,P_{f,\varepsilon}\big(\widetilde{m} > m^\diamond\big) \leqslant C_1\,\varepsilon, \quad \textit{for all } \varepsilon \in (0,1).$$

§3.4.3 ***Remark***. If in addition $16\,\mathrm{bias}^2_{m^\diamond} \leqslant \mathrm{pen}_{m^\diamond}$, which in turn by monotony implies that $16\,\mathrm{bias}^2_k \leqslant \mathrm{pen}_k$, $m^\diamond \leqslant k \leqslant \mathrm{M}$, then we have $\{\widetilde{m} > m^\diamond\} \subset \bigcup_{k=m^\diamond}^{\mathrm{M}}\Big\{16\mathfrak{D}^2_{\text{ist}}(\widehat{f}_k, f_k) > \mathrm{pen}_k\Big\}$. To be more precise, we observe that $\{\widetilde{m} > m^\diamond\} = \{\Upsilon_{m^\diamond} > 0\} = \bigcup_{k=m^\diamond}^{\mathrm{M}}\Big\{\mathfrak{D}^2_{\text{ist}}(\widehat{f}_k, \widehat{f}_m) > \mathrm{pen}_k\Big\}$ which implies that $\{\widetilde{m} > m^\diamond\} \subset \bigcup_{k=m^\diamond}^{\mathrm{M}}\Big\{\mathfrak{D}^2_{\text{ist}}(\widehat{f}_k, f_k) > \mathrm{pen}_k/8 - \mathrm{bias}_k^2\Big\}$ by making use of the subadditivity of $\mathfrak{D}_{\text{ist}}$ and the definition of bias. The claimed result follows now from the additional assumption $16\,\mathrm{bias}^2_{m^\diamond} \leqslant \mathrm{pen}_{m^\diamond}$. □

On the other hand, in view of the first remainder term, the second rhs. term in (3.5), the penalty term pen_m should, roughly speaking, provide an upper bound for the estimator's variation which allows to establish a concentration inequality for the corresponding empirical process. Considering the first remainder term we impose the following assumption, and in the next chapter we present sufficient model conditions to ensure this assumption.

§3.4.4 **Assumption**. *There exists a constant $C_2 := C_2(\mathcal{P}_{\mathcal{F}})$ possibly depending on the family $\mathcal{P}_{\mathcal{F}}$ such that*

$$\sup_{f\in\mathcal{F}} \mathbb{E}_{f,\varepsilon}\Big\{\max_{m^\diamond\leqslant k\leqslant \mathrm{M}}\Big(\mathfrak{D}^2_{ist}(\widehat{f}_k, f_k) - \mathrm{pen}_k\Big)_+\Big\} \leqslant C_2\,\varepsilon, \quad \textit{for all } \varepsilon\in(0,1).$$

The next assertion provides an upper bound for the maximal risk over the family $\mathcal{P}_{\mathcal{F}}$ of probability measures of the estimator $\widehat{f}_{\widetilde{m}}$ with data-driven choice $\widetilde{m}$ given by (3.4).

§3.4.5 **Proposition**. *If Assumption §3.4.2 and §3.4.4 hold true, then we have*

$$\mathcal{R}_\varepsilon\big[\widehat{f}_{\widetilde{m}}\,|\,\mathcal{P}_{\mathcal{F}}\big] \leqslant 10 \sup_{f\in\mathcal{F}} \max\{\mathrm{bias}^2_{m^\diamond}, \mathrm{pen}_{m^\diamond}\} + (2C_1 + 4C_2)\,\varepsilon, \quad \forall\,\varepsilon\in(0,1).$$

Proof of §3.4.5. Keeping in mind the upper bound given in (3.5), the results follows by employing Assumption §3.4.2 and §3.4.4. □

§3.4.6 ***Remark***. The first rhs. term in the last upper risk-bound is strongly reminiscent of a variance-squared-bias decomposition of the risk for the estimator $\widehat{f}_{m^\diamond}$ with dimension parameter $m^\diamond$. Indeed, in many cases the penalty term $\mathrm{pen}_{m^\diamond}$ is in the same order as the variance of the estimator $\widehat{f}_{m^\diamond}$. Consequently, in this situation the upper risk bound of the data-driven estimator is essentially given by $\mathcal{R}_\varepsilon\big[\widehat{f}_{m^\diamond}\,|\,\mathcal{P}_{\mathcal{F}}\big]$. Moreover, by balancing penalty and squared-bias the value $m^\diamond$ realises the optimal trade-off between variance and squared-bias which in turn in many cases means that $\mathcal{R}_\varepsilon\big[\widehat{f}_{m^\diamond}\,|\,\mathcal{P}_{\mathcal{F}}\big]$ is of optimal order. □

§3.4.7 **GSSM (§3.1.1 continued) - local Φ-risk - adaptive estimation**. *Preliminaries.* Let us first state some elementary inequalities for Gaussian random variables. If U is standard normally distributed, then for all $\eta > 0$ and $\zeta \geqslant 1$ we have

$$P(U > \eta) \leqslant (2\pi\eta^2)^{-1/2}\exp(-\eta^2/2) \tag{3.6}$$
$$\mathbb{E}\big(U^2 - 2\zeta(1-\ln\varepsilon)\big)_+ \leqslant \varepsilon^\zeta, \quad \text{for all } \varepsilon\in(0,1). \tag{3.7}$$

Rewriting the elementary inequality $\eta\mathbb{E}\mathbb{1}_{\{U_1\geqslant\eta\}} \leqslant \mathbb{E}U_1\mathbb{1}_{\{U_1\geqslant\eta\}} = (2\pi)^{-1/2}\exp(-\eta^2/2)$ we obtain (3.6). Moreover, keeping in mind the identity $\mathbb{E}(X)_+ = \int_0^\infty P(X\geqslant x)dx$ from (3.6), $2\exp(-\zeta)\leqslant 1$, $2\pi\{2\zeta(1-\ln\varepsilon)+x\}\geqslant 4$ for all $x\geqslant 0$, $\zeta\geqslant 1$, $0<\varepsilon<1$, and using the symmetry of a centered normal distribution we obtain (3.7) as follows

$$\mathbb{E}\big(U^2 - 2\zeta(1-\ln\varepsilon)\big)_+ = \int_0^\infty P(U^2 \geqslant 2\zeta(1-\ln\varepsilon)+x)dx$$
$$\leqslant \int_0^\infty \exp(-\zeta(1-\ln\varepsilon)-x/2)dx \leqslant \varepsilon^\zeta.$$

Consider the OSE $\widehat{f}_m := \sum_{j=1}^m [\widehat{g}_\varepsilon]_j u_j$ with $[\widehat{g}_\varepsilon]_{\underline{m}} \sim \mathcal{N}([f]_{\underline{m}}, \varepsilon[\mathrm{Id}]_{\underline{m}})$. Consequently, $|\Phi(\widehat{f}_m - f_m)|^2 = \varepsilon U^2\sum_{j=1}^m[\Phi]_j^2$ with $U\sim\mathcal{N}(0,1)$.

Technical assertions. For technical reasons and without loss of generality we assume that $[\Phi]_1^2 = 1$, which can always be ensured by reordering and rescaling, except for the trivial case $\Phi \equiv 0$. Thereby, $(1 - \ln \varepsilon)\varepsilon[\Phi]_1^2 \leqslant 1$ for all $0 < \varepsilon < 1$ which in turn ensures the existence of $\mathrm{M} := \max\Big\{1 \leqslant m \leqslant \lfloor \varepsilon^{-1} \rfloor : (1 - \ln \varepsilon)\, \varepsilon \sum_{j=1}^m [\Phi]_j^2 \leqslant 1\Big\}$ because the defining set is not empty. Consider further a sequence of penalties given by $\mathrm{pen}_k := 64(1 - \ln \varepsilon)\varepsilon \sum_{j=1}^k [\Phi]_j^2$, $1 \leqslant k \leqslant \mathrm{M}$, which is obviously non-negative and non-decreasing with $\mathrm{pen}_{\mathrm{M}} \leqslant 64$. From (3.7) and $\varepsilon \sum_{j=1}^m [\Phi]_j^2 \leqslant 1$ for all $1 \leqslant m \leqslant \mathrm{M}$, we obtain further

$$\mathbb{E}_{f,\varepsilon}\Big\{\max_{m^\diamond \leqslant k \leqslant \mathrm{M}} \Big(|\Phi(\widehat{f}_k - f_k)|^2 - \mathrm{pen}_k\Big)_+\Big\} \leqslant \sum_{m=1}^{\mathrm{M}} \mathbb{E}_{f,\varepsilon}\Big(|\Phi(\widehat{f}_m - f_m)|^2 - \mathrm{pen}_m\Big)_+$$
$$= \sum_{m=1}^{\mathrm{M}} \varepsilon \sum_{j=1}^{m} [\Phi]_j^2\ \mathbb{E}\big(U^2 - 64(1 - \ln \varepsilon)\big)_+ \leqslant \mathrm{M}\varepsilon^{32} \leqslant \varepsilon^{31}. \quad (3.8)$$

On the other hand, if $16\,\mathrm{bias}_{m^\diamond}^2 \leqslant \mathrm{pen}_{m^\diamond}$ (cf. Remark §3.4.3) then due to (3.6), $\mathrm{pen}_{\mathrm{M}} \leqslant 64$ and $128(2\pi 4(1 - \ln \varepsilon))^{-1/2} \exp(-2) \leqslant 4$ for all $0 < \varepsilon < 1$, we have

$$\mathrm{pen}_{\mathrm{M}}\, P_{f,\varepsilon}\big(\widetilde{m} > m^\diamond\big) \leqslant 64 \sum_{k=m^\diamond}^{\mathrm{M}} P_{f,\varepsilon}\Big(16|\Phi(\widehat{f}_k - f_k)|^2 > \mathrm{pen}_k\Big)$$
$$= 64\mathrm{M}P\big(U^2 > 4(1 - \ln \varepsilon)\big) \leqslant 128\,\mathrm{M}(2\pi 4(1 - \ln \varepsilon))^{-1/2} \exp(-4(1 - \ln \varepsilon)/2)$$
$$\leqslant 4\mathrm{M}\varepsilon^2 \leqslant 4\varepsilon. \quad (3.9)$$

The definition of $\mathcal{R}_x^m(\Phi, \mathfrak{a})$, $x \in (0, 1)$, given in §3.2.7, the bound $\mathrm{bias}_m^2 \leqslant r \sum_{j \geqslant m} \mathfrak{a}_j^{-1} [\Phi]_j^2$ for all $f \in \mathcal{F}_{\mathfrak{a}}^r$ (see §2.2.15) and $\mathrm{pen}_m = 64(1 - \ln \varepsilon)\varepsilon \sum_{j=1}^m [\Phi]_j^2$ imply together

$$\sup_{f \in \mathcal{F}_{\mathfrak{a}}^r} \max\{\mathrm{bias}_m^2, \mathrm{pen}_m\} \leqslant \max(r, 64)\mathcal{R}_{(1-\ln \varepsilon)\varepsilon}^m(\Phi, \mathfrak{a}). \quad (3.10)$$

Upper Φ-risk bound. Combining the estimates (3.8)-(3.10) we derive now an upper bound for the maximal Φ-risk, $\mathcal{R}_\varepsilon^\Phi\big[\widehat{f}_{\widetilde{m}} \,|\, \mathcal{F}_{\mathfrak{a}}^r\big]$, for the data-driven OSE $\widehat{f}_{\widetilde{m}} := \sum_{j=1}^{\widetilde{m}} [\widehat{g}_\varepsilon]_j u_j$ with dimension parameter $\widetilde{m}$ selected as in (3.4), where we use the contrast

$$\Upsilon_m := \max_{m \leqslant k \leqslant M} \Big\{|\Phi(\widehat{f}_k - \widehat{f}_m)|^2 - \mathrm{pen}_k\Big\}, \qquad 1 \leqslant m \leqslant \mathrm{M}.$$

Precisely, due to (3.9) and (3.8), respectively, the Assumption §3.4.2 and §3.4.4 are satisified with $C_1 = 4$ and $C_2 = 1$. Therefore, we can apply Proposition §3.4.5 which together with (3.8) leads to

$$\mathcal{R}_\varepsilon^\Phi\big[\widehat{f}_{\widetilde{m}} \,|\, \mathcal{F}_{\mathfrak{a}}^r\big] \leqslant 10 \max(64, r)\mathcal{R}_{(1-\ln \varepsilon)\varepsilon}^{m^\diamond}(\Phi, \mathfrak{a}) + 9\varepsilon.$$

If in addition $m^\diamond = \arg\min_{m \in \mathbb{N}} \mathcal{R}_{(1-\ln \varepsilon)\varepsilon}^m(\Phi, \mathfrak{a})$ satisfies $m^\diamond \leqslant \mathrm{M}$ and $16\,\mathrm{bias}_{m^\diamond} \leqslant \mathrm{pen}_{m^\diamond}$ then

$$\mathcal{R}_\varepsilon^\Phi\big[\widehat{f}_{\widetilde{m}} \,|\, \mathcal{F}_{\mathfrak{a}}^r\big] \leqslant C(r)\mathcal{R}_{(1-\ln \varepsilon)\varepsilon}^\star(\Phi, \mathfrak{a}).$$

The upper bound in the last display features a logarithmic factor when compared to the minimax rate of convergence $\mathcal{R}_\varepsilon^\star\big(\Phi, \mathfrak{a}\big)$ which possibly results in a deterioration of the rate. Therefore, the completely data-driven estimator is optimal or nearly optimal in the minimax sense simultaneously over a variety of solution sets $\mathcal{F}_{\mathfrak{a}}^{r}$. The appearance of the logarithmic factor within the rate is a known fact in the context of local estimation (cf. Laurent et al. [2008] who consider model selection). Brown and Low [1996] show that it is unavoidable in the context of non-parametric Gaussian regression and, hence it is widely considered as an acceptable price for adaptation. We shall emphasize that - as usual in non-parametric estimation- in comparison to a global tuning parameter selection, a local one might be favorable in terms of the attainable accuracy of the estimator. As only local features are of interest, it is likely that compared to the overall performance, certain areas might be estimated more accurately - even with the logarithmic factor. For an upper bound of a *global* $\mathbb{H}$*-risk* of the completely data-driven estimator relying on Lespki's method we refer the interest reader to Birgé [2001]. □

3.4.2 Model selection by contrast minimisation

Let us consider a family $\mathcal{P}_{\mathcal{F}}$ of probability measures and a weighted norm as a global distance, that is $\mathfrak{D}_{\text{ist}}(h_1, h_2) = \|h_1 - h_2\|_{\mathfrak{v}}$ for $h_1, h_2 \in \mathbb{H}$ associated with an ONB $(u_j)_{j\in\mathbb{N}}$ and a sequence $\mathfrak{v}$. In the sequel we assume an estimator $\widehat{[f]} := (\widehat{[f]}_j)_{j\in\mathbb{N}}$ of the sequence of generalised Fourier coefficients $[f] = \mathcal{U}f$. Having the estimated sequence at hand, $\widehat{f}_m := \sum_{j=1}^{m} \widehat{[f]}_j u_j \in \mathbb{U}_m$ is an orthogonal series estimator of the unknown solution $f \in \mathcal{F}$ and $f_m = \Pi_{\mathbb{U}_m} f$ is the orthogonal projection of f onto $\mathbb{U}_m$. Keeping in mind that $\mathcal{U}^\star x = \sum_{j\in\mathbb{N}} x_j u_j$ is the adjoint of the generalised Fourier series transform U, we define for $g \in \mathbb{H}$ the contrast

$$\Upsilon(g) := \|g\|_{\mathfrak{v}}^2 - 2\langle g, U^\star\widehat{[f]}\rangle_{\mathfrak{v}} = \|g\|_{\mathfrak{v}}^2 - 2\sum_{j\in\mathbb{N}} \mathfrak{v}_j [g]_j \widehat{[f]}_j. \tag{3.11}$$

Note that we can compute

$$\Upsilon(\widehat{f}_m) = -\|\widehat{f}_m\|_{\mathfrak{v}}^2 = -\sum_{j=1}^{m} \mathfrak{v}_j |\widehat{[f]}_j|^2.$$

Moreover, for all $m \in \mathbb{N}$ the estimator $\widehat{f}_m$ minimises the contrast function Υ, that is

$$\Upsilon_m := \Upsilon(\widehat{f}_m) = \min_{h\in\mathbb{U}_m} \Upsilon(h). \tag{3.12}$$

Indeed, since $\langle h, U^\star\widehat{[f]}\rangle_{\mathfrak{v}} = \langle h, \widehat{f}_m\rangle_{\mathfrak{v}}$ for all $h \in \mathbb{U}_m$ it follows that $\Upsilon(h) = \|h - \widehat{f}_m\|_{\mathfrak{v}}^2 - \|\widehat{f}_m\|_{\mathfrak{v}}^2$ which in turn implies the assertion. Given an integer M and a subsequence of non-negative and non-decreasing penalties $(\mathrm{pen}_1, \dots, \mathrm{pen}_{\mathrm{M}})$ we select the dimension parameter $\widetilde{m}$ among the collection of admissible values $\{1, \dots, \mathrm{M}\}$ as minimiser of a penalised contrast criterion. To be precise, setting $\arg\min_{m\in A}\{a_m\} := \min\{m \in A : a_m \leqslant a_{m'}, \forall\, m' \in A\}$ for a sequence $(a_m)_{m\geqslant 1}$ with minimal value in $A \subset \mathbb{N}$, we define

$$\widetilde{m} := \mathop{\arg\min}_{1\leqslant m\leqslant \mathrm{M}} \{\Upsilon_m + \mathrm{pen}_m\}. \tag{3.13}$$

The data-driven estimator of f is now given by $\widehat{f}_{\widetilde{m}}$ and below we derive an upper bound for its maximal risk $\mathcal{R}_{\varepsilon}^{\mathfrak{v}}\big[\widehat{f}_{\widetilde{m}} \,|\, \mathcal{P}_{\mathcal{F}}\big] = \sup_{f\in\mathcal{F}} \mathbb{E}_{f,\varepsilon}\|\widehat{f}_{\widetilde{m}} - f\|_{\mathfrak{v}}^2$ over the class $\mathcal{P}_{\mathcal{F}}$. The key argument for our reasoning is the next assertion.

§3.4.8 **Lemma (key argument)**. *For all* $1 \leqslant m \leqslant \mathrm{M}$ *we have*

$$\|\widehat{f}_{\widetilde{m}} - f\|_{\mathfrak{v}}^2 \leqslant 7\max(\mathrm{bias}_m^2, \mathrm{pen}_m) + 8 \max_{m \leqslant k \leqslant \mathrm{M}} \Big(\|\widehat{f}_k - f_k\|_{\mathfrak{v}}^2 - \frac{1}{4}\,\mathrm{pen}_k \Big)_+ .$$

Proof of §3.4.8. Applying successively the definition of $\widetilde{m}$ given in (3.13) and the property (3.12) of the contrast function Υ it follows that

$$\Upsilon(\widehat{f}_{\widetilde{m}}) + \mathrm{pen}_{\widetilde{m}} \leqslant \Upsilon(\widehat{f}_m) + \mathrm{pen}_m \leqslant \Upsilon(f_m) + \mathrm{pen}_m, \quad \forall 1 \leqslant m \leqslant \mathrm{M},$$

which in particular implies that

$$\begin{aligned} \|\widehat{f}_{\widetilde{m}}\|_{\mathfrak{v}}^2 - \|f_m\|_{\mathfrak{v}}^2 &\leqslant 2\Big\{ \langle \widehat{f}_{\widetilde{m}}, U^\star\widehat{[f]}\rangle_{\mathfrak{v}} - \langle f_m, U^\star\widehat{[f]}\rangle_{\mathfrak{v}} \Big\} + \mathrm{pen}_m - \mathrm{pen}_{\widehat{m}} \\ &= 2\langle \widehat{f}_{\widetilde{m}} - f_m, U^\star\widehat{[f]}\rangle_{\mathfrak{v}} + \mathrm{pen}_m - \mathrm{pen}_{\widetilde{m}} \end{aligned}$$

Rewriting the last equality we conclude that

$$\begin{aligned} \|\widehat{f}_{\widetilde{m}} - f\|_{\mathfrak{v}}^2 &= \|f - f_m\|_{\mathfrak{v}}^2 + \|\widehat{f}_{\widetilde{m}}\|_{\mathfrak{v}}^2 - \|f_m\|_{\mathfrak{v}}^2 - 2\langle \widehat{f}_{\widetilde{m}} - f_m, f\rangle_{\mathfrak{v}} \\ &\leqslant \|f - f_m\|_{\mathfrak{v}}^2 + \mathrm{pen}_m - \mathrm{pen}_{\widetilde{m}} + 2\langle \widehat{f}_{\widetilde{m}} - f_m, U^\star\widehat{[f]} - f\rangle_{\mathfrak{v}} \end{aligned}$$

Consider an unit ball $\overline{\mathcal{B}}_m := \{h \in \mathbb{U}_m : \|h\|_{\mathfrak{v}} \leqslant 1\}$ in $\mathbb{U}_m$ and let $\widetilde{m} \vee m := \max(\widetilde{m}, m)$. Keep in mind for $\tau > 0$ and $g \in \mathbb{U}_m$ the elementary inequality

$$2|\langle f, g\rangle_{\mathfrak{v}}| \leqslant 2\|f\|_{\mathfrak{v}} \sup_{t \in \overline{\mathcal{B}}_m} |\langle t, g\rangle_{\mathfrak{v}}| \leqslant \tau\|f\|_{\mathfrak{v}}^2 + \frac{1}{\tau} \sup_{t \in \overline{\mathcal{B}}_m} |\langle t, g\rangle_{\mathfrak{v}}|^2 = \tau\|f\|_{\mathfrak{v}}^2 + \frac{1}{\tau}\|g\|_{\mathfrak{v}}^2.$$

Let $V_m^2 := \sup_{t \in \overline{\mathcal{B}}_m} |\langle t, U^\star\widehat{[f]} - f\rangle_{\mathfrak{v}}|^2$, $m \geqslant 1$, then from $\widehat{f}_{\widetilde{m}} - f_m \in \mathbb{U}_{\widetilde{m} \vee m} \subset \mathbb{U}_{\mathrm{M}}$ we obtain

$$\|\widehat{f}_{\widetilde{m}} - f\|_{\mathfrak{v}}^2 \leqslant \|f - f_m\|_{\mathfrak{v}}^2 + \mathrm{pen}_m - \mathrm{pen}_{\widetilde{m}} + \tau\|\widehat{f}_{\widetilde{m}} - f_m\|_{\mathfrak{v}}^2 + \frac{1}{\tau} V_{\widetilde{m} \vee m}^2.$$

Noting that $\mathrm{pen}_{m \vee m'} \leqslant \mathrm{pen}_m + \mathrm{pen}_{m'}$ and $\|\widehat{f}_{\widetilde{m}} - f_m\|_{\mathfrak{v}}^2 \leqslant 2\|\widehat{f}_{\widetilde{m}} - f\|_{\mathfrak{v}}^2 + 2\|f_m - f\|_{\mathfrak{v}}^2$, we get, for $\tau = 1/4$ and for all $1 \leqslant m \leqslant \mathrm{M}$ that

$$\begin{aligned} \frac{1}{2}\|\widehat{f}_{\widetilde{m}} - f\|_{\mathfrak{v}}^2 &\leqslant \frac{3}{2}\|f - f_m\|_{\mathfrak{v}}^2 + \mathrm{pen}_m - \mathrm{pen}_{\widetilde{m}} + \mathrm{pen}_{\widetilde{m} \vee m} + 4\big(V_{\widetilde{m} \vee m}^2 - \frac{1}{4}\,\mathrm{pen}_{\widetilde{m} \vee m}\big)_+ \\ &\leqslant \frac{3}{2}\|f - f_m\|_{\mathfrak{v}}^2 + 2\,\mathrm{pen}_m + 4 \max_{m \leqslant k \leqslant \mathrm{M}} \big(V_k^2 - \frac{1}{4}\,\mathrm{pen}_k\big)_+ . \end{aligned}$$

Combining the last bound and the elementary identity $V_m^2 = \|\widehat{f}_m - f_m\|_{\mathfrak{v}}^2$ imply the assertion. □

Compare the last assertion with Lemma §3.4.1. Obviously in order to prove an upper risk bound for a data-driven estimator based on a model selection approach we can follow line by line the development given in case of Lepski's method. Consequently, let $m^\diamond$ realise a penalty-squared-bias compromise among the collection of admissible values $\{1, \dots, \mathrm{M}\}$. Due to the last assertion we have for all $P_{f,\varepsilon} \in \mathcal{P}_{\mathcal{F}}$

$$\mathbb{E}_{f,\varepsilon}\|\widehat{f}_{\widetilde{m}} - f\|_{\mathfrak{v}}^2 \leqslant 7\max(\mathrm{bias}_{m^\diamond}^2, \mathrm{pen}_{m^\diamond}) + 8\mathbb{E}_{f,\varepsilon}\Big\{ \max_{m^\diamond \leqslant k \leqslant \mathrm{M}} \Big(\|\widehat{f}_k - f_k\|_{\mathfrak{v}}^2 - \frac{1}{4}\,\mathrm{pen}_k \Big)_+ \Big\}. \tag{3.14}$$

Next, we bound the remainder term (the second rhs. term).

§3.4.9 **Assumption**. *There exists a constant $C := C(\mathcal{P}_{\mathcal{F}})$ possibly depending on the family $\mathcal{P}_{\mathcal{F}}$ such that*

$$\sup_{f\in\mathcal{F}} \mathbb{E}_{f,\varepsilon}\Big\{\max_{m^\diamond\leqslant k\leqslant \mathrm{M}}\Big(\|\widehat{f}_k - f_k\|_{\mathfrak{v}}^2 - \frac{1}{4}\,\mathrm{pen}_k\Big)_+\Big\} \leqslant C\,\varepsilon, \qquad \textit{for all } \varepsilon\in(0,1).$$

The next assertion provides an upper bound for the maximal risk over the family $\mathcal{P}_{\mathcal{F}}$ of probability measures of the estimator $\widehat{f}_{\widetilde{m}}$ with data-driven choice $\widetilde{m}$ given by (3.13).

§3.4.10 **Proposition**. *If there exists a constant $C := C(\mathcal{P}_{\mathcal{F}})$ possibly depending on the class $\mathcal{P}_{\mathcal{F}}$ such that Assumption §3.4.4 holds true, then we have*

$$\mathcal{R}_\varepsilon\big[\widehat{f}_{\widetilde{m}}\,|\,\mathcal{P}_{\mathcal{F}}\big] \leqslant 7\sup_{f\in\mathcal{F}}\max\{\mathrm{bias}^2_{m^\diamond},\mathrm{pen}_{m^\diamond}\} + 8\,C\,\varepsilon, \quad \forall\,\varepsilon\in(0,1).$$

Proof of §3.4.10. Keeping in mind the upper bound given in (3.14), the results follows by employing Assumption §3.4.9. □

§3.4.11 ***Remark***. Compare the last assertion with §3.4.5 considering Lepski's method. In both cases we make use of Assumption §3.4.4, however Lepski's method requires in addition Assumption §3.4.2. Moreover, the Remark §3.4.6 applies also here. □

§3.4.12 **GSSM (§3.1.1 continued) - global $\mathbb{H}_{\mathfrak{v}}$-risk - adaptive estimation**. *Preliminaries.* The assertion (3.6) presents a tail bound of a Gaussian random variable. There exist several results for an analogous tail bound for sums of independent squared Gaussian random variables and we present next a version which is due to Birgé [2001] and can be shown following the lines of the proof of Lemma 1 in Laurent and Massart [2000]. Let $\{U_i, 1\leqslant i\leqslant m\}$ be independent and standard normally distributed random variables. Given $\mathfrak{a}_1,\dots,\mathfrak{a}_m$ non-negative we set $|\mathfrak{a}|_\infty := \max_{1\leqslant i\leqslant m}\mathfrak{a}_i$ and $|\mathfrak{a}|_1 := \sum_{i=1}^m \mathfrak{a}_i$. Let $S_m := \sum_{i=1}^m \mathfrak{a}_i U_i^2$, then for all $\eta>0$ we have

$$P(S_m - \mathbb{E}S_m \geqslant (3/2)\eta|\mathfrak{a}|_1) \leqslant \exp(-\eta(\eta\wedge 1)|\mathfrak{a}|_1/|\mathfrak{a}|_\infty). \tag{3.15}$$

From the last tail bound, $\mathbb{E}S_m = |\mathfrak{a}|_1$ and $\{\zeta + 2x/(3|\mathfrak{a}|_1)\}\geqslant 1$ for all $x\geqslant 0$ and $\zeta\geqslant 1$, we conclude that for all $\zeta\geqslant 1$ holds

$$\begin{aligned}\mathbb{E}(S_m - (1+(3/2)\zeta)|\mathfrak{a}|_1)_+ &= \int_0^\infty P(S_m - \mathbb{E}S_m \geqslant (3/2)\{\zeta + 2x/(3|\mathfrak{a}|_1)\}|\mathfrak{a}|_1)dx\\ &\leqslant \int_0^\infty \exp\left(-\{\zeta+2x/(3|\mathfrak{a}|_1)\}|\mathfrak{a}|_1/|\mathfrak{a}|_\infty\right)dx = (3/2)|\mathfrak{a}|_\infty\exp(-\zeta|\mathfrak{a}|_1/|\mathfrak{a}|_\infty).\end{aligned} \tag{3.16}$$

Consider the OSE $\widehat{f}_m := \sum_{j=1}^m [\widehat{g}_\varepsilon]_j u_j$ with $[\widehat{g}_\varepsilon]_{\underline{m}} \sim \mathcal{N}([f]_{\underline{m}}, \varepsilon[\mathrm{Id}]_{\underline{m}})$ which in turn implies that $\|\widehat{f}_m - f_m\|_{\mathfrak{v}}^2 = \varepsilon\sum_{j=1}^m \mathfrak{v}_j U_j^2$ where $U_j\sim\mathcal{N}(0,1)$, $1\leqslant j\leqslant m$, are independent.

Technical assertions. For technical reasons we assume that $\mathfrak{v}_1 = 1$, which in turn ensures the existence of $M := \max\Big\{1\leqslant m\leqslant \lfloor\varepsilon^{-1}\rfloor : \varepsilon\sum_{j=1}^m \mathfrak{v}_j\leqslant 1\Big\}$ because the defining set is not empty. Consider further $\mathrm{pen}_k = 10\varepsilon\sum_{j=1}^k\mathfrak{v}_j$, $1\leqslant k\leqslant \mathrm{M}$, as sequence of penalties which is obviously non-negative and non-decreasing with $\mathrm{pen}_{\mathrm{M}}\leqslant 10$. Set further $|\mathfrak{v}_{\underline{m}}|_\infty := \max_{1\leqslant j\leqslant m}\mathfrak{v}_j$ and $|\mathfrak{v}_{\underline{m}}|_1 := \sum_{j=1}^m\mathfrak{v}_j$. If the sequence $\mathfrak{v}$ satisfies $\Sigma_{\mathfrak{v}} := \sum_{m=1}^\infty |\mathfrak{v}_{\underline{m}}|_\infty\exp(-|\mathfrak{v}_{\underline{m}}|_1/|\mathfrak{v}_{\underline{m}}|_\infty) <$

∞, then from (3.16) with $S_m := \sum_{j=1}^m \mathfrak{v}_j U_j^2$ we obtain

$$\mathbb{E}_{f,\varepsilon}\Big\{\max_{m^\diamond\leqslant k\leqslant \mathrm{M}}\Big(\|\widehat{f}_k - f_k\|_{\mathfrak{v}}^2 - \frac{1}{4}\,\mathrm{pen}_k\Big)_+\Big\} \leqslant \sum_{m=1}^{\mathrm{M}} \mathbb{E}_{f,\varepsilon}\Big(\|\widehat{f}_m - f_m\|_{\mathfrak{v}}^2 - \frac{1}{4}\,\mathrm{pen}_m\Big)_+$$
$$= \varepsilon\sum_{m=1}^{\mathrm{M}} \mathbb{E}(S_m - (5/2)|\mathfrak{v}_{\underline{m}}|_1)_+ \leqslant (3/2)\varepsilon\sum_{m=1}^{\mathrm{M}} |\mathfrak{v}_{\underline{m}}|_\infty\ \exp(-|\mathfrak{v}_{\underline{m}}|_1/|\mathfrak{v}_{\underline{m}}|_\infty) \leqslant (3/2)\Sigma_{\mathfrak{v}}\,\varepsilon. \tag{3.17}$$

From the definition of $\mathcal{R}_x^m(\mathfrak{v},\mathfrak{a})$, $x\in(0,1)$, given in §3.2.13, the estimate $\mathrm{bias}_m^2 \leqslant r\mathfrak{v}_m\mathfrak{a}_m^{-1}$ for all $f\in\mathcal{F}_{\mathfrak{a}}^r$ (see §2.2.15) and $\mathrm{pen}_m = 10\varepsilon\sum_{j=1}^k \mathfrak{v}_j$ we conclude

$$\sup_{f\in\mathcal{F}_{\mathfrak{a}}^r}\max\{\mathrm{bias}_m^2,\mathrm{pen}_m\} \leqslant \max(r,10)\mathcal{R}_\varepsilon^m\big(\mathfrak{v},\mathfrak{a}\big). \tag{3.18}$$

Upper $\mathbb{H}_{\mathfrak{v}}$-risk bound. Combining the estimates (3.17) and (3.18) we derive now an upper bound for the maximal $\mathbb{H}_{\mathfrak{v}}$-risk, $\mathcal{R}_\varepsilon^{\mathfrak{v}}\big[\widehat{f}_{\widetilde{m}}\,|\,\mathcal{F}_{\mathfrak{a}}^r\big]$, for the data-driven OSE $\widehat{f}_{\widetilde{m}} := \sum_{j=1}^{\widetilde{m}}[\widehat{g}_\varepsilon]_j u_j$ with dimension parameter $\widetilde{m}$ selected as in (3.4), where we use the contrast

$$\Upsilon(f) := \|f\|_{\mathfrak{v}}^2 - 2\langle f, U^\star[\widehat{g}_\varepsilon]\rangle_{\mathfrak{v}} = \|f\|_{\mathfrak{v}}^2 - 2\sum_{j\in\mathbb{N}}\mathfrak{v}_j[f]_j[\widehat{g}_\varepsilon]_j.$$

Precisely, due to (3.17) the Assumption §3.4.9 holds with $C = (3/2)\Sigma_{\mathfrak{v}}$. Therefore, we can apply Proposition §3.4.10 which together with (3.18) leads to

$$\mathcal{R}_\varepsilon^{\mathfrak{v}}\big[\widehat{f}_{\widetilde{m}}\,|\,\mathcal{F}_{\mathfrak{a}}^r\big] \leqslant 7\max(r,10)\mathcal{R}_\varepsilon^{m^\diamond}(\mathfrak{v},\mathfrak{a}) + 12\Sigma_{\mathfrak{v}}\varepsilon.$$

If in addition $m^\diamond = \arg\min_{m\in\mathbb{N}}\mathcal{R}_\varepsilon^m(\mathfrak{v},\mathfrak{a})$ satisfies $m^\diamond\leqslant \mathrm{M}$ then we have

$$\mathcal{R}_\varepsilon^{\mathfrak{v}}\big[\widehat{f}_{\widetilde{m}}\,|\,\mathcal{F}_{\mathfrak{a}}^r\big] \leqslant C(r,\mathfrak{v})\ \mathcal{R}_\varepsilon^\star\big(\mathfrak{v},\mathfrak{a}\big).$$

Since $\mathcal{R}_\varepsilon^\star\big(\mathfrak{v},\mathfrak{a}\big)$ is the minimax-optimal rate pf convergence the upper bound in the last display establishes the minimax-optimality up to a constant of the completely data-driven estimator. Therefore, in contrast to a local Φ-risk (see §3.4.7) we do not face a deterioration of the rate in case of a global $\mathbb{H}_{\mathfrak{v}}$-risk. We shall further emphasise the special case of a $\mathbb{H}$-risk, that is, the sequence $\mathfrak{v}\equiv 1$ is constant one. In this situation we have $|\mathfrak{v}_{\underline{m}}|_\infty = 1$ and $|\mathfrak{v}_{\underline{m}}|_1 = m$, and consequently $\Sigma_{\mathfrak{v}} = \sum_{m=1}^\infty \exp(-m) < \infty$. Furthermore, it is easily seen that $M = \lfloor\varepsilon^{-1}\rfloor$ and $m^\diamond \leqslant M$ for all strictly increasing and unbounded sequences $\mathfrak{a}$. An upper bound of a *local Φ-risk* of a completely data-driven estimator based on a model selection approach can be found, for example, in Laurent et al. [2008]. □

3.4.3 Combining model selection and Lepski's method

The upper bound of the data-driven estimator based on a model selection approach presented in the last section heavily relies on the property (3.12), that the estimator minimises the contrast function. However, in case of a Galerkin estimator, for example, we did not succeed to write the estimator as a minimiser of a contrast. Therefore, the next selection method combines model selection and Lepski's's method which is inspired by a bandwidth selection method in kernel density estimation proposed recently in Goldenshluger and Lepski [2011]. We shall emphasize, that with regard to our illustration GSSM the proof of the local Φ-risk in case of Lepski's method as well as the global $\mathbb{H}_{\mathfrak{v}}$-risk in case of a model selection approach are rather elementary, however, the corresponding result of a global $\mathbb{H}_{\mathfrak{v}}$-risk in case of Lepski's method as well as a local Φ-risk in case of a model selection approach are much more involved. Interestingly, combining both approaches allows us to present a unified approach for both types of risks and their proof shares their elementary nature. To be more precise we consider a penalised contrast criterion as in (3.13), that is,

$$\widetilde{m} := \arg\min_{1 \leqslant m \leqslant \mathrm{M}} \{\Upsilon_m + \mathrm{pen}_m\}.$$

Given a family $\mathcal{P}_{\mathcal{F}}$ of probability measures and a maximal risk $\mathcal{R}_{\varepsilon}\big[\widehat{f}_{\widetilde{m}} \,|\, \mathcal{P}_{\mathcal{F}}\big]$ based on a distance $\mathfrak{D}_{\mathrm{ist}}$ we use Lepski's contrast function defined in (3.3), that is,

$$\Upsilon_m := \max_{m \leqslant k \leqslant \mathrm{M}} \Big\{ \mathfrak{D}^2_{\mathrm{ist}}(\widehat{f}_k, \widehat{f}_m) - \mathrm{pen}_k \Big\}, \quad 1 \leqslant m \leqslant \mathrm{M}.$$

Next we will show an analogue to the key argument §3.4.1 and §3.4.8 in case of Lepski's method and the model selection approach respectively.

§3.4.13 **Lemma (key argument)**. *If the subsequence* $(\mathrm{pen}_1, \dots, \mathrm{pen}_{\mathrm{M}})$ *is non-decreasing, then for all* $1 \leqslant m \leqslant \mathrm{M}$ *we have*

$$\mathfrak{D}^2_{\mathrm{ist}}(\widehat{f}_{\widetilde{m}}, f) \leqslant 85 \max(\mathrm{bias}^2_m, \mathrm{pen}_m) + 42 \max_{m \leqslant k \leqslant \mathrm{M}} \Big(\mathfrak{D}^2_{\mathrm{ist}}(\widehat{f}_k, f_k) - \frac{1}{6}\mathrm{pen}_k \Big)_+ .$$

Proof of §3.4.13. From the definition of $\widetilde{m}$ we deduce for all $1 \leqslant m \leqslant \mathrm{M}$ that

$$\begin{aligned}\mathfrak{D}^2_{\mathrm{ist}}(\widehat{f}_{\widetilde{m}}, f) &\leqslant 3\Big\{\Upsilon_m + \mathrm{pen}_{\widetilde{m}} + \Upsilon_{\widetilde{m}} + \mathrm{pen}_m + \mathfrak{D}^2_{\mathrm{ist}}(\widehat{f}_m, f)\Big\} \\ &\leqslant 6\{\Upsilon_m + \mathrm{pen}_m\} + 3\mathfrak{D}^2_{\mathrm{ist}}(\widehat{f}_m, f). \qquad (3.19)\end{aligned}$$

Firstly, applying an elementary triangular inequality allows us to write

$$\mathfrak{D}^2_{\mathrm{ist}}(\widehat{f}_{\widetilde{m}}, f) \leqslant \frac{1}{3}\mathrm{pen}_m + 2\,\mathrm{bias}^2_m + 2 \max_{m \leqslant k \leqslant \mathrm{M}} \Big(\mathfrak{D}^2_{\mathrm{ist}}(\widehat{f}_k, f_k) - \frac{1}{6}\mathrm{pen}_k \Big)_+$$

for all $1 \leqslant m \leqslant \mathrm{M}$. Secondly, since $(\mathrm{pen}_1, \dots, \mathrm{pen}_{\mathrm{M}})$ is non-decreasing and $4\,\mathrm{bias}^2_m \geqslant \max_{m \leqslant k \leqslant \mathrm{M}} \mathfrak{D}^2_{\mathrm{ist}}(f_k, f_m)$, $1 \leqslant m \leqslant \mathrm{M}$, it is easily verified that

$$\Upsilon_m \leqslant 6 \sup_{m \leqslant k \leqslant \mathrm{M}} \Big(\mathfrak{D}^2_{\mathrm{ist}}(\widehat{f}_k, f_k) - \frac{1}{6}\mathrm{pen}_k \Big)_+ + 12\,\mathrm{bias}^2_m .$$

Combining the last two inequalities and (3.19), we obtain the result. □

Compare the last assertion with Lemma §3.4.8. Obviously in order to prove an upper risk bound for a data-driven estimator combining model selection and Lepski's method we can follow line by line the model selection development, for example. Consequently, let $m^{\diamond}$ realise a penalty-squared-bias compromise among the collection of admissible values $\{1, \dots, \mathrm{M}\}$. Due to the last assertion we have for all $P_{f,\varepsilon} \in \mathcal{P}_{\mathcal{F}}$

$$\mathbb{E}_{f,\varepsilon}\|\widehat{f}_{\widetilde{m}} - f\|^2_{\mathfrak{v}} \leqslant 85 \max(\mathrm{bias}^2_{m^{\diamond}}, \mathrm{pen}_{m^{\diamond}}) + 42\mathbb{E}_{f,\varepsilon}\Big\{ \max_{m^{\diamond} \leqslant k \leqslant \mathrm{M}} \Big(\|\widehat{f}_k - f_k\|^2_{\mathfrak{v}} - \frac{1}{6}\mathrm{pen}_k \Big)_+ \Big\}. \qquad (3.20)$$

Next, we bound the remainder term the second rhs. term by imposing the following assumption.

§3.4.14 **Assumption**. *There exists a constant $C := C(\mathcal{P}_{\mathcal{F}})$ possibly depending on the family $\mathcal{P}_{\mathcal{F}}$ such that*

$$\sup_{f\in\mathcal{F}} \mathbb{E}_{f,\varepsilon}\Big\{\max_{m^\diamond\leqslant k\leqslant \mathrm{M}}\Big(\|\widehat{f}_k - f_k\|_{\mathfrak{v}}^2 - \frac{1}{6}\,\mathrm{pen}_k\Big)_+\Big\} \leqslant C\,\varepsilon, \quad \textit{for all } \varepsilon\in(0,1).$$

The next assertion provides an upper bound for the maximal risk over the family $\mathcal{P}_{\mathcal{F}}$ of probability measures of the estimator $\widehat{f}_{\widetilde{m}}$ with data-driven choice $\widetilde{m}$ given by (3.13).

§3.4.15 **Proposition**. *If there exists a constant $C := C(\mathcal{P}_{\mathcal{F}})$ possibly depending on the class $\mathcal{P}_{\mathcal{F}}$ such that Assumption §3.4.4 holds true, then we have*

$$\mathcal{R}_\varepsilon\big[\widehat{f}_{\widetilde{m}}\,|\,\mathcal{P}_{\mathcal{F}}\big] \leqslant 85\sup_{f\in\mathcal{F}}\max\{\mathrm{bias}^2_{m^\diamond},\mathrm{pen}_{m^\diamond}\} + 42\,C\,\varepsilon, \quad \forall\,\varepsilon\in(0,1).$$

Proof of §3.4.15. Keeping in mind the upper bound given in (3.20), the results follows by employing Assumption §3.4.14. □

§3.4.16 **GSSM (§3.1.1 continued) - adaptive estimation**. Consider first a *local Φ-risk* and consequently, the contrast

$$\Upsilon_m := \max_{m\leqslant k\leqslant\mathrm{M}}\Big\{|\Phi(\widehat{f}_k - \widehat{f}_m)|^2 - \mathrm{pen}_k\Big\}, \quad 1\leqslant m\leqslant \mathrm{M},$$

where we set $\mathrm{M} := \max\Big\{1\leqslant m\leqslant\lfloor\varepsilon^{-1}\rfloor : (1-\ln\varepsilon)\,\varepsilon\,\sum_{j=1}^m[\Phi]_j^2\leqslant 1\Big\}$ and $\mathrm{pen}_k = 24(1-\ln\varepsilon)\varepsilon\sum_{j=1}^k[\Phi]_j^2$, $1\leqslant k\leqslant\mathrm{M}$ (compare with the development in §3.4.7). As in case of (3.10) it is easily verified that $\sup_{f\in\mathcal{F}_{\mathfrak{a}}^r}\max\{\mathrm{bias}^2_m,\mathrm{pen}_m\}\leqslant\max(r,24)\mathcal{R}^m_{(1-\ln\varepsilon)\varepsilon}(\Phi,\mathfrak{a})$. Moreover, following line by line the proof of (3.8) it is straightforward to see that Assumption §3.4.14 holds with $C = 1$. Therefore, we can apply Proposition §3.4.15 which in turn leads to the bound

$$\mathcal{R}^\Phi_\varepsilon\big[\widehat{f}_{\widetilde{m}}\,|\,\mathcal{F}^r_{\mathfrak{a}}\big]\leqslant 85\max(24,r)\mathcal{R}^{m^\diamond}_{(1-\ln\varepsilon)\varepsilon}(\Phi,\mathfrak{a}) + 42\varepsilon.$$

If in addition $m^\diamond = \arg\min_{m\in\mathbb{N}}\mathcal{R}^m_{(1-\ln\varepsilon)\varepsilon}(\Phi,\mathfrak{a})$ satisfies $m^\diamond\leqslant\mathrm{M}$ then

$$\mathcal{R}^\Phi_\varepsilon\big[\widehat{f}_{\widetilde{m}}\,|\,\mathcal{F}^r_{\mathfrak{a}}\big]\leqslant C(r)\mathcal{R}^\star_{(1-\ln\varepsilon)\varepsilon}(\Phi,\mathfrak{a}).$$

Observe that the upper bound in the last display features a logarithmic factor when compared to the minimax rate of convergence $\mathcal{R}^\star_\varepsilon(\Phi,\mathfrak{a})$ (as in §3.4.1 using Lepski's method). The completely data-driven estimator is consequently optimal or nearly optimal in the minimax sense simultaneously over a variety of solution sets $\mathcal{F}^r_{\mathfrak{a}}$.

Imposing a *global $\mathbb{H}_{\mathfrak{v}}$-risk* we use the contrast

$$\Upsilon_m := \max_{m\leqslant k\leqslant\mathrm{M}}\Big\{\|\widehat{f}_k - \widehat{f}_m\|^2_{\mathfrak{v}} - \mathrm{pen}_k\Big\}, \quad 1\leqslant m\leqslant\mathrm{M},$$

where we set $\mathrm{M} := \max\Big\{1\leqslant m\leqslant\lfloor\varepsilon^{-1}\rfloor : \lfloor\varepsilon^{-1}\rfloor : \varepsilon\,\sum_{j=1}^m\mathfrak{v}_j\leqslant 1\Big\}$ and $\mathrm{pen}_k = 15\varepsilon\sum_{j=1}^k\mathfrak{v}_j$, $1\leqslant k\leqslant\mathrm{M}$ (compare with the development in §3.4.12). As in case of (3.18) it is easily verified that $\sup_{f\in\mathcal{F}^r_{\mathfrak{a}}}\max\{\mathrm{bias}^2_m,\mathrm{pen}_m\}\leqslant\max(r,15)\mathcal{R}^m_\varepsilon\big(\mathfrak{v},\mathfrak{a}\big)$. If the sequence $\mathfrak{v}$ satisfies $\Sigma_{\mathfrak{v}} := \sum_{m=1}^\infty|\mathfrak{v}_{\underline{m}}|_\infty\exp(-|\mathfrak{v}_{\underline{m}}|_1/|\mathfrak{v}_{\underline{m}}|_\infty) < \infty$, then following line by line the proof of (3.17) it is straightforward to see that Assumption §3.4.14 holds with $C = (3/2)\Sigma_{\mathfrak{v}}\varepsilon$. Therefore, we can apply Proposition §3.4.15 which in turn leads to the bound

$$\mathcal{R}_\varepsilon^{\mathfrak{v}}\big[\widehat{f}_{\widetilde{m}}\,|\,\mathcal{F}_{\mathfrak{a}}^r\big] \leqslant 85\max(15,r)\mathcal{R}_\varepsilon^{m^\diamond}(\mathfrak{v},\mathfrak{a}) + 63\Sigma_{\mathfrak{v}}\varepsilon.$$

If in addition $m^\diamond = \underset{m\in\mathbb{N}}{\arg\min}\,\mathcal{R}_\varepsilon^m\big(\mathfrak{v},\mathfrak{a}\big)$ satisfies $m^\diamond \leqslant \mathrm{M}$ then

$$\mathcal{R}_\varepsilon^{\Phi}\big[\widehat{f}_{\widetilde{m}}\,|\,\mathcal{F}_{\mathfrak{a}}^r\big] \leqslant C(r)\mathcal{R}_\varepsilon^{\star}\big(\mathfrak{v},\mathfrak{a}\big)$$

which establishes the minimax-optimality up to a constant of the completely data-driven estimator (as in §3.4.12 using a model selection approach). □

Let us now consider a family of measures $\mathcal{P}_{\mathcal{F},\Xi}$ indexed by a set of solutions $\mathcal{F}$ and a set of nuissance parameters Ξ. The presence of a nuissance parameter is often reflected in the variance of the estimator $\widehat{f}_m$ which in turn in order to ensure Assumption §3.4.4 and §3.4.2 a force us to use a subsequence of penalty terms $\mathrm{pen}_1^\xi,\dots,\mathrm{pen}_{M_\xi}^\xi$ and an upper bound M_ξ depending on the nuissance parameter ξ. In this situation, however, the observations will allow us to construct an estimator $\widehat{\xi}$ of the nuissance parameter ξ. Having this estimator at hand, we will consider empirical counterparts of M_ξ and pen_k^ξ given by $\widehat{\mathrm{M}} := \mathrm{M}_{\widehat{\xi}}$ and $\widehat{\mathrm{pen}}_k := \mathrm{pen}_k^{\widehat{\xi}}$, $1 \leqslant k \leqslant \widehat{\mathrm{M}}$, respectively. It is worth to note that in the remainder of this section the upper bound $\widehat{\mathrm{M}}$ and hence the collection of admissible values $\{1,\dots,\widehat{\mathrm{M}}\}$ as well as the penalty sequence $(\widehat{\mathrm{pen}}_k)_k$ are random. Moreover, all quantities will depend also on the noise levels ε and σ and we will omit additional subscripts for simplicity. Nervertheless, the random sequence of penalties is still assumed to be non-negative and non-decreasing and we consider a penalised

$$\widehat{m} := \arg\min_{1\leqslant m\leqslant\widehat{\mathrm{M}}}\big\{\widehat{\Upsilon}_m + \widehat{\mathrm{pen}}_m\big\}. \tag{3.21}$$

Imposing a maximal risk $\mathcal{R}_{\varepsilon,\sigma}\big[\,\cdot\,|\,\mathcal{P}_{\mathcal{F},\Xi}\big] = \sup_{f\in\mathcal{F}}\sup_{\xi\in\Xi}\mathbb{E}_{f,\xi,\varepsilon,\sigma}[\mathfrak{D}_{\text{ist}}^2(\cdot,f)]$ based on a distance $\mathfrak{D}_{\text{ist}}$ we use again Lepski's contrast function defined in (3.3), i.e.,

$$\widehat{\Upsilon}_m := \max_{m\leqslant k\leqslant\widehat{\mathrm{M}}}\Big\{\mathfrak{D}_{\text{ist}}^2(\widehat{f}_k,\widehat{f}_m) - \widehat{\mathrm{pen}}_k\Big\},\quad 1\leqslant m\leqslant\widehat{\mathrm{M}}\,.$$

The data-driven estimator of f is now given by $\widehat{f}_{\widehat{m}}$ and below we derive an upper bound for its maximal risk $\mathcal{R}_{\varepsilon,\sigma}\big[\widehat{f}_{\widehat{m}}\,|\,\mathcal{P}_{\mathcal{F},\Xi}\big]$ over certain classes of solutions $\mathcal{F}$ and nuisances parameters Ξ. The key argument for our reasoning is again Lemma §3.4.13 which allows us to conclude that for all $1\leqslant m\leqslant\widehat{\mathrm{M}}$

$$\mathfrak{D}_{\text{ist}}^2(\widehat{f}_{\widehat{m}},f) \leqslant 85\max(\mathrm{bias}_m^2,\widehat{\mathrm{pen}}_m) + 42\max_{m\leqslant k\leqslant\widehat{\mathrm{M}}}\left(\mathfrak{D}_{\text{ist}}^2(\widehat{f}_k,f_k) - \frac{1}{6}\,\widehat{\mathrm{pen}}_k\right)_+$$

provided that the subsequence $\widehat{\mathrm{pen}}_1,\dots,\widehat{\mathrm{pen}}_{\widehat{\mathrm{M}}}$ is non-negative and non-decreasing. Keeping the last bound in mind we decompose the risk with respect to an event on which the quantities $\widehat{\mathrm{pen}}_m = \mathrm{pen}_m^{\widehat{\xi}}$ and $\widehat{\mathrm{M}} = \mathrm{M}_{\widehat{\xi}}$ are close to certain theoretical counterparts pen_m^ξ, M_ξ^- and M_ξ^+ where pen_k^ξ, $1\leqslant k\leqslant\mathrm{M}_\xi^+$, are non-negativ and non-decreasing for all $\xi\in\Xi$. More precisely, define for all $\xi\in\Xi$ the event

$$\Omega_\xi := \Big\{\mathrm{pen}_k^\xi \leqslant \widehat{\mathrm{pen}}_k \leqslant c_1\,\mathrm{pen}_k^\xi,\ \forall\,1\leqslant k\leqslant\mathrm{M}_\xi^+\Big\} \cap \Big\{\mathrm{M}_\xi^- \leqslant \widehat{\mathrm{M}} \leqslant \mathrm{M}_\xi^+\Big\} \tag{3.22}$$

for some finite numerical constant c_1 and denote its complement by Ω_ξ^c which allow us to decompose the risk

$$\mathcal{R}_{\varepsilon,\sigma}\big[\widehat{f}_{\widehat{m}}\,|\,\mathcal{P}_{\mathcal{F},\Xi}\big] = \sup_{f\in\mathcal{F}}\sup_{\xi\in\Xi}\mathbb{E}_{f,\xi,\varepsilon,\sigma}[\mathbb{1}_{\Omega_\xi}\mathfrak{D}^2_{\text{ist}}(\widehat{f}_{\widehat{m}},f)] + \sup_{f\in\mathcal{F}}\sup_{\xi\in\Xi}\mathbb{E}_{f,\xi,\varepsilon,\sigma}[\mathbb{1}_{\Omega_\xi^c}\mathfrak{D}^2_{\text{ist}}(\widehat{f}_{\widehat{m}},f)]. \quad (3.23)$$

Consider the first rhs. term. We apply Lemma §3.4.1 which on the event Ω_ξ leads to

$$\mathfrak{D}^2_{\text{ist}}(\widehat{f}_{\widehat{m}},f)\mathbb{1}_{\Omega_\xi} \leqslant 85\,c_1\max(\text{bias}^2_{m_\xi^\diamond},\text{pen}^\xi_{m_\xi^\diamond}) + 42\max_{m_\xi^\diamond\leqslant k\leqslant \mathrm{M}_\xi^+}\Big(\mathfrak{D}^2_{\text{ist}}(\widehat{f}_k,f_k)-\text{pen}^\xi_k\Big)_+ \quad (3.24)$$

where $m_\xi^\diamond$ realises a penalty-squared-bias compromise among the collection of admissible values $\{1,\dots,\mathrm{M}_\xi^-\}$. On the other hand, in view of the remainder term, the second rhs. term in (3.24), the penalty term pen^ξ_k should, roughly speaking, provide an upper bound for the estimator's variation which allows to establish a concentration inequality for the corresponding empirical process. Considering the remainder term we impose the following assumption.

§3.4.17 **Assumption**. *There exist finite constants $K^I_{\varepsilon,\sigma}(\mathcal{P}_{\mathcal{F},\Xi})$ such that for all $\varepsilon,\sigma\in(0,1)$*

$$42\sup_{f\in\mathcal{F}}\sup_{\xi\in\Xi}\mathbb{E}_{f,\xi,\varepsilon,\sigma}\left\{\max_{m_\xi^\diamond\leqslant k\leqslant \mathrm{M}_\xi^+}\Big(\mathfrak{D}^2_{ist}(\widehat{f}_k,f_k)-\text{pen}^\xi_k\Big)_+\right\}\leqslant K^I_{\varepsilon,\sigma}(\mathcal{P}_{\mathcal{F},\Xi}).$$

Under Assumption §3.4.17 we bound the first rhs. term in (3.23) by

$$\sup_{f\in\mathcal{F}}\sup_{\xi\in\Xi}\mathbb{E}_{f,\xi,\varepsilon,\sigma}[\mathfrak{D}^2_{\text{ist}}(\widehat{f}_{\widehat{m}},f)\mathbb{1}_{\Omega_\xi}]\leqslant 85\,c_1\sup_{f\in\mathcal{F}}\sup_{\xi\in\Xi}\max\{\text{bias}^2_{m_\xi^\diamond},\text{pen}^\xi_{m_\xi^\diamond}\}+K^I_{\varepsilon,\sigma}(\mathcal{P}_{\mathcal{F},\Xi}).$$

It remains to consider the second rhs. term in (3.23) which bound imposing the next assumption.

§3.4.18 **Assumption**. *There exist finite constants $K^{II}_{\varepsilon,\sigma}(\mathcal{P}_{\mathcal{F},\Xi})$ such that for all $\varepsilon,\sigma\in(0,1)$*

$$\sup_{f\in\mathcal{F}}\sup_{\xi\in\Xi}\mathbb{E}_{f,\xi,\varepsilon,\sigma}[\mathfrak{D}^2_{ist}(\widehat{f}_{\widehat{m}},f)\mathbb{1}_{\Omega_\xi^c}]\leqslant K^{II}_{\varepsilon,\sigma}(\mathcal{P}_{\mathcal{F},\Xi}).$$

The next assertion provides an upper bound for the maximal risk over the classes $\mathcal{F}$ and Ξ of the estimator $\widehat{f}_{\widehat{m}}$ with data-driven choice $\widehat{m}$ given by (3.21).

§3.4.19 **Proposition**. *If Assumption §3.4.17 and §3.4.18 hold true, then for all $\varepsilon,\sigma\in(0,1)$*

$$\mathcal{R}_{\varepsilon,\sigma}\big[\widehat{f}_{\widehat{m}}\,|\,\mathcal{P}_{\mathcal{F},\Xi}\big]\leqslant 85\,c_1\sup_{f\in\mathcal{F}}\sup_{\xi\in\Xi}\max\{\text{bias}^2_{m_\xi^\diamond},\text{pen}^\xi_{m_\xi^\diamond}\}+K^I_{\varepsilon,\sigma}(\mathcal{P}_{\mathcal{F},\Xi})+K^{II}_{\varepsilon,\sigma}(\mathcal{P}_{\mathcal{F},\Xi}).$$

Proof of §3.4.19. Keeping in mind the risk decomposition (3.23) and the upper bound given in (3.24), the results follows by employing Assumption §3.4.17 and §3.4.18. □

In view of Assumption §3.4.17 and §3.4.18, we aime to construct $\widehat{\mathrm{M}}$ and $\widehat{\text{pen}}_m$ such that they behave similarly to their theoretical counterparts with sufficiently high probability so as not to deteriorate the estimators risk. Roughly speaking, we will construct them such that the upper bounds $K^I_{\varepsilon,\sigma}(\mathcal{P}_{\mathcal{F},\Xi})$ and $K^{II}_{\varepsilon,\sigma}(\mathcal{P}_{\mathcal{F},\Xi})$ of the remainder terms are negligeble with respect to the first rhs. in the last upper risk-bound which aimes to reflect an optimal variance-squared-bias decomposition. In the sequel we will derive analytic expressions for the upper bounds $K^I_{\varepsilon,\sigma}(\mathcal{P}_{\mathcal{F},\Xi})$ and $K^{II}_{\varepsilon,\sigma}(\mathcal{P}_{\mathcal{F},\Xi})$ which obviously depend on the statistical model.

Chapter 4

Gaussian inverse regression

Let $T \in \mathscr{K}(\mathbb{H}, \mathbb{G})$ be an operator mapping a separable Hilbert spaces $\mathbb{H}$ to a Hilbert space $\mathbb{G}$. Consider a statistical inverse problem $\widehat{g}_\varepsilon = Tf + \sqrt{\varepsilon}\dot{W}$ with noise in the operator $\widehat{T}_\sigma = T + \sqrt{\sigma}\dot{B}$ where we assume in this chapter that the error terms $\dot{W}$ and $\dot{B}$ are independent Gaussian white noises. Given an ONB $(u_j)_{j\in\mathbb{N}}$ in $\mathbb{H}$ and an ONS $(v_j)_{j\in\mathbb{N}}$ in $\mathbb{G}$ the observable quantities take the form $\langle v_j, \widehat{g}_\varepsilon\rangle_{\mathbb{G}} = \langle v_j, g\rangle_{\mathbb{G}} + \sqrt{\sigma}\langle v_j, \dot{W}\rangle_{\mathbb{G}}$ and $\langle v_j, \widehat{T}_\sigma u_k\rangle_{\mathbb{G}} = \langle v_j, Tu_k\rangle_{\mathbb{G}} + \sqrt{\sigma}\langle v_j, \dot{B}u_k\rangle_{\mathbb{G}}$, or

$$[\widehat{g}_\varepsilon]_j = [g]_j + \sqrt{\varepsilon}[\dot{W}]_j \quad \text{and} \quad [\widehat{T}_\sigma]_{j,k} = [T]_{j,k} + \sqrt{\sigma}[\dot{B}]_{j,k}, \quad j,k \in \mathbb{N},$$

in short, where the error terms $\{[\dot{W}]_j, [\dot{B}]_{j,k},\ j,k \in \mathbb{N}\}$ are independent and standard normally distributed. Our aim is the reconstruction of the function of interest f which we assume to belong to an ellipsoid $\mathcal{F}_{\mathfrak{a}}^r$ constructed with respect to the ONB $(u_j)_{j\in\mathbb{N}}$ and a non-decreasing sequence $(\mathfrak{a}_j)_{j\in\mathbb{N}}$ of weights (compare §2.1.11). Furthermore, we will assume that the operator T satisfies a link condition, that is $(T^*T)^{1/2} \in \mathcal{T}_{\mathfrak{b}}^d$, with respect to the ONB $(u_j)_{j\in\mathbb{N}}$ and a strictly positive non-increasing sequence $\mathfrak{b}$ (see §2.7.3).
This section serves in two ways. On the one hand, it allows us to introduce the key arguments to derive the minimax-theory for the models circular deconvolution (CD), functional linear regression (FLR) and nonparametric instrumental regression (NIR). On the other hand comparing the minimax-optimal rates under different model assumptions we can characterise the influence of both noise levels ε and σ. The section is organised as follows, in the first part we study statistical inverse problem with known operator T, where we distinguish two cases. First we assume the singular value decompositon of T is known which allows us to formalise the statistical inverse problem as an indirect sequence space model while in the second case we do not rely on the singular value decomposition. The second part is concerned with statistical inverse problems with noise in the operator, where again we consider two cases. First we assume the eigenfunctions of T are known in advance, and its eigenvalues have to be estimated. This model corresponds to an indirect sequence space model with noise in the operator.

4.1 Gaussian indirect sequence space model

Let us introduce the subset $\mathscr{S}_{u,v}(\mathbb{H}, \mathbb{G})$ of $\mathscr{K}(\mathbb{H}, \mathbb{G})$ containing all compact linear operators having $(u_j)_{j\in\mathbb{N}}$ and $(v_j)_{j\in\mathbb{N}}$ as eigenfunctions. Consequently, $T \in \mathscr{S}_{u,v}(\mathbb{H}, \mathbb{G})$ admits a singular system $\{(\mathfrak{s}_j, u_j, v_j), j \in \mathbb{N}\}$ where $(\mathfrak{s}_j)_{j\in\mathbb{N}}$ denotes its non-negative sequence of singular values which tends to zero. In other words, its (infinite) matrix representation $[T] = [\nabla_{\mathfrak{s}}]$ is diagonal, i.e., for all $m \in \mathbb{N}$, $[T]_{\underline{m}} = [\nabla_{\mathfrak{s}}]_{\underline{m}}$ is a m-dimensional diagonal matrix with entries $\mathfrak{s}_{\underline{m}} = (\mathfrak{s}_j)_{1\leqslant j\leqslant m}$. Finally, observe that the link condition $(T^*T)^{1/2} \in \mathcal{S}_{\mathfrak{b}}^d$ is satisfied if and only if $d^{-1} \leqslant \mathfrak{s}_j\mathfrak{b}_j^{-1} \leqslant d$, for all $j \in \mathbb{N}$. Let us denote by $\mathcal{S}_{\mathfrak{b}}^d := \{T \in \mathscr{S}_{u,v}(\mathbb{H}, \mathbb{G}) : (T^*T)^{1/2} \in \mathcal{T}_{\mathfrak{b}}^d\}$ the subset of $\mathscr{S}_{u,v}(\mathbb{H}, \mathbb{G})$ including only operators satisfying the link condition. Note that each element of $\mathcal{S}_{\mathfrak{b}}^d$ is injective. Given $T \in \mathcal{S}_{\mathfrak{b}}^d$ our aim is the reconstruction of a function $f \in \mathbb{H}$

based on a noisy version $\widehat{g}_\varepsilon = Tf + \sqrt{\varepsilon}\dot{W}$ of $g = Tf \in \mathbb{G}$ where $\dot{W}$ is a Gaussian white noise. Considering the projection onto the eigenbasis of T the observable quantities take the form

$$[\widehat{g}_\varepsilon]_j = \mathfrak{s}_j[f]_j + \sqrt{\varepsilon}[\dot{W}]_j, \quad j \in \mathbb{N} \tag{4.1}$$

where the error terms $\{[\dot{W}]_j, j \in \mathbb{N}\}$ are independent and standard normally distributed and where the sequence $\mathfrak{s}$ of singular values of T satisfies $[T]_{j,j} = \mathfrak{s}_j$, $j \in \mathbb{N}$. Moreover, let $\mathbb{E}_{f,T,\varepsilon}$ denote the expectation with respect to the probability measure associated with the observable quantities. Given $T \in \mathcal{S}^d_{\mathfrak{b}}$ the accuracy of any estimator $\widetilde{f}$ is measured by its maximal risk over the class $\mathcal{F}^r_{\mathfrak{a}}$, that is, $\sup_{f\in\mathcal{F}^r_{\mathfrak{a}}} \mathbb{E}_{f,T,\varepsilon}\big[\mathfrak{D}^2_{\text{ist}}(\widetilde{f}, f)\big]$. We restrict ourselves to the two particular cases of a maximal local Φ-risk and a maximal global $\mathbb{H}_{\mathfrak{v}}$-risk, respectively, given by

$$\mathcal{R}_\varepsilon\big[\widetilde{f}|\Phi, \mathcal{F}^r_{\mathfrak{a}}, T\big] := \sup_{f\in\mathcal{F}^r_{\mathfrak{a}}} \mathbb{E}_{f,T,\varepsilon}|\Phi(\widetilde{f} - f)|^2$$

$$\text{and} \quad \mathcal{R}_\varepsilon\big[\widetilde{f}|\mathfrak{v}, \mathcal{F}^r_{\mathfrak{a}}, T\big] := \sup_{f\in\mathcal{F}^r_{\mathfrak{a}}} \mathbb{E}_{f,T,\varepsilon}\|\widetilde{f} - f\|^2_{\mathfrak{v}}.$$

Let us summarize the conditions on the sequences $\mathfrak{a}$, $\mathfrak{b}$ and $\mathfrak{v}$ as well as the linear functional Φ.

§4.1.1 **Assumption.** *Let $\Phi \in \mathcal{L}_{1/\mathfrak{a}}$ and let $\mathfrak{a}$, $\mathfrak{b}$ and $\mathfrak{v}$ be strictly positive sequences of weights such that $1/\mathfrak{a}$, $\mathfrak{b}$ and $\mathfrak{v}/\mathfrak{a}$ are non-increasing with limit zero. For technical reasons and without loss of generality we assume that $[\Phi]^2_1 = 1$, $\mathfrak{a}_1 = 1$, $\mathfrak{b} = 1$ and $\mathfrak{v}_1 = 1$.*

4.1.1 Lower bounds

§4.1.2 **Lower bound of a Φ-risk.** *Under Assumption §4.1.1 define for $\varepsilon \in (0,1)$ and $m \in \mathbb{N}$,*

$$\mathcal{R}^m_\varepsilon(\Phi, \mathfrak{a}, \mathfrak{b}) := \max\left(\sum_{j>m} \frac{[\Phi]^2_j}{\mathfrak{a}_j}, \max(\mathfrak{b}^2_m \mathfrak{a}^{-1}_m, \varepsilon) \sum_{j=1}^m \frac{[\Phi]^2_j}{\mathfrak{b}^2_m}\right);$$

$$\mathcal{R}^\star_\varepsilon := \mathcal{R}^\star_\varepsilon(\Phi, \mathfrak{a}, \mathfrak{b}) := \min_{m\in\mathbb{N}} \mathcal{R}^m_\varepsilon(\Phi, \mathfrak{a}, \mathfrak{b})$$

$$\textit{and} \quad m^\star_\varepsilon := m^\star_\varepsilon(\Phi, \mathfrak{a}, \mathfrak{b}) := \arg\min_{m\in\mathbb{N}} \mathcal{R}^m_\varepsilon(\Phi, \mathfrak{a}, \mathfrak{b}).$$

If, in addition, $\eta := \inf_{\varepsilon\in(0,1)} \{\min(\varepsilon\mathfrak{a}_{m^\star_\varepsilon}\mathfrak{b}^{-2}_{m^\star_\varepsilon}, (\varepsilon\mathfrak{a}_{m^\star_\varepsilon})^{-1}\mathfrak{b}^2_{m^\star_\varepsilon})\} > 0$ then for all $\varepsilon \in (0,1)$

$$\inf_{\widetilde{f}} \mathcal{R}_\varepsilon\big[\widetilde{f}|\Phi, \mathcal{F}^r_{\mathfrak{a}}, T\big] \geqslant \frac{\eta}{8d^2} \min(2rd^2, 1)\, \mathcal{R}^\star_\varepsilon, \qquad \textit{for all } T \in \mathcal{S}^d_{\mathfrak{b}}.$$

Proof of §4.1.2. The proof follows line by line the proof given in §3.2.7. We start with the observation that $[\log dP_1/dP_{-1}]([\widehat{g}_\varepsilon]) = 2\sum_{j\in\mathbb{N}}[\widehat{g}_\varepsilon]_j[f_*]_j\mathfrak{s}_j/\varepsilon$ and hence $KL(P_1, P_{-1}) = 2\|f_*\|^2_{\mathfrak{s}^2}/\varepsilon$. Since $T \in \mathcal{S}^d_{\mathfrak{b}}$ it follows that $\|f_*\|^2_{\mathfrak{s}^2} \leqslant d^2\|f_*\|^2_{\mathfrak{b}^2}$. Thereby, for all $f_* \in \mathcal{F}^r_{\mathfrak{a}}$ with $2d^2\|f_*\|^2_{\mathfrak{b}^2} \leqslant \varepsilon$ we have $KL(P_1, P_{-1}) \leqslant 1$ and hence the assumptions of §3.2.4 are satisfied. The result is again obtained by evaluating the bound presented in §3.2.6 for two choices of $f_* \in \mathcal{F}^r_{\mathfrak{a}}$ with $2d^2\|f_*\|^2_{\mathfrak{s}^2} \leqslant \varepsilon$ constructed as in §3.2.7. To be precise, consider $f_* := (\zeta\alpha_\varepsilon)^{1/2}\sum_{j=1}^{m^\star_\varepsilon} \mathfrak{b}^{-2}_j[\Phi]_j u_j$ with $\alpha_\varepsilon := K^*_\varepsilon(\sum_{j=1}^{m^\star_\varepsilon} \mathfrak{b}^{-2}_j[\Phi]^2_j)^{-1}$, $K^*_\varepsilon := \max(\mathfrak{b}^2_{m^\star_\varepsilon}\mathfrak{a}^{-1}_{m^\star_\varepsilon}, \varepsilon)$ and $\zeta := \eta\min(r, 2^{-1}d^{-2})$. It is easily verified that $\|f_*\|^2_{\mathfrak{a}} \leqslant r$, $2d^2\varepsilon^{-1}\|f_*\|^2_{\mathfrak{b}^2} \leqslant 1$ and $|\Phi(f_*)|^2 = \zeta\max(\mathfrak{b}^2_{m^\star_\varepsilon}\mathfrak{a}^{-1}_{m^\star_\varepsilon}, \varepsilon)\sum_{j=1}^{m^\star_\varepsilon} \mathfrak{b}^{-2}_j[\Phi]^2_j$, which together leads to

$$\inf_{\widetilde{f}} \mathcal{R}_\varepsilon\big[\widetilde{f}|\Phi, \mathcal{F}^r_{\mathfrak{a}}, T\big] \geqslant \frac{\zeta}{4}\max(\mathfrak{b}^2_{m^\star_\varepsilon}\mathfrak{a}^{-1}_{m^\star_\varepsilon}, \varepsilon)\sum_{j=1}^{m^\star_\varepsilon} \mathfrak{b}^{-2}_j[\Phi]^2_j.$$

On the other hand, consider $f_* := (\zeta\alpha_\varepsilon)^{1/2}\sum_{j>m_\varepsilon^\star}[\Phi]_j\mathfrak{a}_j^{-1}u_j$ with $\alpha_\varepsilon := (\sum_{j>m_\varepsilon^\star}[\Phi]_j^2\mathfrak{a}_j^{-1})^{-1}$ as in §3.2.7 where we have shown that $\|f_*\|_{\mathfrak{a}}^2 \leqslant r$ and

$$\inf_{\widetilde{f}} \mathcal{R}_\varepsilon\big[\widetilde{f}|\Phi, \mathcal{F}_{\mathfrak{a}}^r, T\big] \geqslant \frac{\zeta}{4}\sum_{j>m_\varepsilon^\star}[\Phi]_j^2\mathfrak{a}_j^{-1}.$$

It remains only to check that $2d^2\varepsilon^{-1}\|f_*\|_{\mathfrak{b}^2}^2 \leqslant 1$ which is straightforward to see. Combining the two lower bounds we obtain the assertion. □

§4.1.3 **Lower bound of a $\mathbb{H}_{\mathfrak{v}}$-risk.** *Under Assumption §4.1.1 define for $\varepsilon \in (0,1)$ and $m \in \mathbb{N}$,*

$$\mathcal{R}_\varepsilon^m\big(\mathfrak{v},\mathfrak{a},\mathfrak{b}\big) := \max(\mathfrak{v}_m\mathfrak{a}_m^{-1}, \sum_{j=1}^m \varepsilon\mathfrak{v}_j\mathfrak{b}_j^{-2}); \quad \mathcal{R}_\varepsilon^\star := \mathcal{R}_\varepsilon^\star\big(\mathfrak{v},\mathfrak{a},\mathfrak{b}\big) := \min_{m\in\mathbb{N}}\mathcal{R}_\varepsilon^m\big(\mathfrak{v},\mathfrak{a},\mathfrak{b}\big);$$

$$\text{and} \quad m_\varepsilon^\star := m_\varepsilon^\star\big(\mathfrak{v},\mathfrak{a},\mathfrak{b}\big) := \arg\min_{m\in\mathbb{N}}\mathcal{R}_\varepsilon^m\big(\mathfrak{v},\mathfrak{a},\mathfrak{b}\big).$$

If, in addition, $\eta := \inf_{\varepsilon\in(0,1)}\Big\{(\mathcal{R}_\varepsilon^\star)^{-1}\min(\mathfrak{v}_{m_\varepsilon^\star}\mathfrak{a}_{m_\varepsilon^\star}^{-1}, \sum_{j=1}^{m_\varepsilon^\star}\varepsilon\mathfrak{v}_j\mathfrak{b}_j^{-2})\Big\} > 0$ then for all $\varepsilon \in (0,1)$

$$\inf_{\widetilde{f}} \mathcal{R}_\varepsilon\big[\widetilde{f}|\mathfrak{v}, \mathcal{F}_{\mathfrak{a}}^r, T\big] \geqslant \frac{\eta}{8d^2}\,\min(2rd^2,1)\,\mathcal{R}_\varepsilon^\star, \quad \text{for all } T \in \mathcal{S}_{\mathfrak{b}}^d.$$

Proof of §4.1.3. The proof follows line by line the proof of §3.2.13 where we consider $f_* := (\varepsilon\zeta\alpha_\varepsilon)^{1/2}\sum_{j=1}^{m_\varepsilon^\star}\mathfrak{b}_j^{-2}u_j$ with $\alpha_\varepsilon := (\sum_{j=1}^{m_\varepsilon^\star}\varepsilon\mathfrak{b}_j^{-2}\mathfrak{v}_j)^{-1}\mathcal{R}_\varepsilon^\star$ and $\zeta := \eta\min(r, 2^{-1}d^{-2})$ which belongs to $\mathcal{F}_{\mathfrak{a}}^r$ by a straightforward adaptation of the arguments given in the proof of §3.2.13. Furthermore we have $[\log dP_\theta/dP_{\theta^{(j)}}]([\widehat{g}_\varepsilon]) = 2[\widehat{g}_\varepsilon]_j[f_*]_j\mathfrak{s}_j/\varepsilon$ and $KL(P_\theta, P_{\theta^{(j)}}) = 2[f_*]_j^2\mathfrak{s}_j^2/\varepsilon$ which in turn implies $H(P_\theta, P_{\theta^{(j)}}) \leqslant 1$ for all $1 \leqslant j \leqslant m$ and $\theta \in \Theta$ since $2\mathfrak{s}_j^2[f_*]_j^2 \leqslant 2d^2\mathfrak{b}_j^2[f_*]_j^2 \leqslant \varepsilon$ for all $T \in \mathcal{S}_{\mathfrak{b}}^d$. Thereby, the assumptions of §3.2.10 are satisfied and hence we can apply the bound presented in §3.2.12. Combining this lower bound and $\|\Pi_{\mathbb{U}_m}f_*\|_{\mathfrak{v}}^2 = \zeta\alpha_\varepsilon\sum_{j=1}^{m_\varepsilon^\star}\varepsilon\mathfrak{b}_j^{-2}\mathfrak{v}_j = \zeta\mathcal{R}_\varepsilon^\star$ implies the assertion. □

4.1.2 Minimax optimal estimation

Keep in mind that given observations as in (4.1) for each $j \in \mathbb{N}$, $\widehat{[f]}_j := \mathfrak{s}_j^{-1}[\widehat{g}_\varepsilon]_j$ is the unique best unbiased estimator of $[f]_j$ due to Lehman-Scheffé's Theorem. Consequently, introducing a dimension parameter $m \in \mathbb{N}$ we consider the OSE $\widehat{f}_m := \sum_{j=1}^m\mathfrak{s}_j^{-1}\widehat{[f]}_ju_j$. Observe that $[\widehat{f}]_{\underline{m}} = [\nabla_{1/\mathfrak{s}}]_{\underline{m}}[\widehat{g}_\varepsilon]_{\underline{m}}$, where trivially $[\widehat{f}]_{\underline{m}} \sim \mathcal{N}([f]_{\underline{m}}, \varepsilon[\nabla_{\mathfrak{s}^{-2}}]_{\underline{m}})$ and, hence $\mathbb{E}\|\widehat{f}_m - f_m\|_{\mathfrak{v}}^2 = 2\varepsilon\sum_{j=1}^m\mathfrak{v}_j\mathfrak{s}_j^{-2}$ and $\mathbb{E}|\Phi(\widehat{f}_m - f_m)|^2 = \varepsilon\sum_{j=1}^m[\Phi]_j^2\mathfrak{s}_j^{-2}$. We exploit these properties in the following proofs.

§4.1.4 **Upper bound of a Φ-risk.** *Under Assumption §4.1.1 consider the OSE $\widehat{f}_{m_\varepsilon^\star}$ with dimension parameter $m_\varepsilon^\star$ and the rate $\mathcal{R}_\varepsilon^\star := \mathcal{R}_\varepsilon^\star\big(\Phi,\mathfrak{a},\mathfrak{b}\big)$ defined as in §4.1.2. We have*

$$\mathcal{R}_\varepsilon\big[\widehat{f}_{m_\varepsilon^\star}|\Phi, \mathcal{F}_{\mathfrak{a}}^r, T\big] \leqslant (1+r)\,\mathcal{R}_\varepsilon^\star, \quad \text{for all } T \in \mathcal{S}_{\mathfrak{b}}^d \text{ and } \varepsilon \in (0,1).$$

Proof of §4.1.4. The proof follows line by line the proof of §3.3.2 and we ommit the details. □

§4.1.5 **Upper bound of a $\mathbb{H}_{\mathfrak{v}}$-risk.** *Under Assumption §4.1.1 consider the OSE $\widehat{f}_{m_\varepsilon^\star}$ with dimension parameter $m_\varepsilon^\star$ and the rate $\mathcal{R}_\varepsilon^\star := \mathcal{R}_\varepsilon^\star\big(\mathfrak{v},\mathfrak{a},\mathfrak{b}\big)$ defined as in §4.1.3. We have*

$$\mathcal{R}_\varepsilon\big[\widehat{f}_{m_\varepsilon^\star}|\mathfrak{v}, \mathcal{F}_{\mathfrak{a}}^r, T\big] \leqslant (2+r)\,\mathcal{R}_\varepsilon^\star, \quad \text{for all } T \in \mathcal{S}_{\mathfrak{b}}^d \text{ and } \varepsilon \in (0,1).$$

Proof of §4.1.5. The proof follows line by line the proof of §3.3.2 and we ommit the details. □

4.1.3 Adaptive estimation

By a combination of a model selection approach and Lepski's method we determine the dimension parameter among a collection of admissible values by minimising a penalised contrast function. To be precise we consider a penalised contrast as in (3.13), that is,

$$\widetilde{m} := \arg\min_{1\leqslant m\leqslant \mathrm{M}} \{\Upsilon_m + \mathrm{pen}_m\}$$

and Lepski's contrast function defined in (3.3), that is,

$$\Upsilon_m := \max_{m\leqslant k\leqslant \mathrm{M}} \Big\{\mathfrak{D}^2_{\text{ist}}(\widehat{f}_k, \widehat{f}_m) - \mathrm{pen}_k\Big\}, \quad 1\leqslant m\leqslant \mathrm{M}.$$

Consider first a local Φ-risk and consequently, the contrast

$$\Upsilon_m = \max_{m\leqslant k\leqslant \mathrm{M}} \Big\{|\Phi(\widehat{f}_k - \widehat{f}_m)|^2 - \mathrm{pen}_k\Big\}, \quad 1\leqslant m\leqslant \mathrm{M}.$$

Given $T\in\mathcal{S}^d_{\mathfrak{b}}$ with associated sequence of singular values $(\mathfrak{s}_j)_{j\in\mathbb{N}}$ we set

$$\mathrm{M} := \max\Big\{1\leqslant m\leqslant \lfloor\varepsilon^{-1}\rfloor : (1-\ln\varepsilon)\,\varepsilon\sum_{j=1}^m [\Phi]_j^2\mathfrak{s}_j^{-2}\leqslant 1\Big\}$$

$$\text{and}\qquad \mathrm{pen}_k = 24(1-\ln\varepsilon)\varepsilon\sum_{j=1}^k [\Phi]_j^2\mathfrak{s}_j^{-2}, \quad 1\leqslant k\leqslant \mathrm{M}$$

(compare with the development in §3.4.16).

§4.1.6 **Upper bound of a Φ-risk.** *Under Assumption §4.1.1 consider the fully data-driven OSE $\widehat{f}_{\widetilde{m}}$ with data-driven dimension parameter $\widetilde{m}$ and the rate $\mathcal{R}^\star_{(1-\ln\varepsilon)\,\varepsilon} := \mathcal{R}^\star_{(1-\ln\varepsilon)\,\varepsilon}\big(\Phi,\mathfrak{a},\mathfrak{b}\big)$ defined as in §4.1.2. If in addition $m^\diamond = \arg\min_{m\in\mathbb{N}} \mathcal{R}^m_{(1-\ln\varepsilon)\varepsilon}\big(\Phi,\mathfrak{a},\mathfrak{b}\big)$ satisfies $m^\diamond\leqslant\mathrm{M}$ then*

$$\mathcal{R}_\varepsilon\big[\widehat{f}_{\widetilde{m}}|\Phi,\mathcal{F}^r_{\mathfrak{a}},T\big]\leqslant (1+r)\,\mathcal{R}^\star_{(1-\ln\varepsilon)\,\varepsilon}, \qquad \textit{for all } T\in\mathcal{S}^d_{\mathfrak{b}} \textit{ and } \varepsilon\in(0,1).$$

Proof of §4.1.6. As in case of (3.10) it is easily verified that $\sup_{f\in\mathcal{F}^r_{\mathfrak{a}}}\max\{\mathrm{bias}^2_m,\mathrm{pen}_m\}\leqslant \max(r,24d^2)\mathcal{R}^m_{(1-\ln\varepsilon)\varepsilon}\big(\Phi,\mathfrak{a},\mathfrak{b}\big)$ for all $T\in\mathcal{S}^d_{\mathfrak{b}}$. Moreover, following line by line the proof of (3.8) it is straightforward to see that Assumption §3.4.14 holds with $C=1$. Therefore, we can apply Proposition §3.4.15 which in turn leads to the bound

$$\mathcal{R}_\varepsilon\big[\widehat{f}_{\widetilde{m}}|\Phi,\mathcal{F}^r_{\mathfrak{a}},T\big]\leqslant 85\max(24d^2,r)\mathcal{R}^{m^\diamond}_{(1-\ln\varepsilon)\varepsilon}\big(\Phi,\mathfrak{a},\mathfrak{b}\big)+42\varepsilon$$

since $m^\diamond = \arg\min_{m\in\mathbb{N}} \mathcal{R}^m_{(1-\ln\varepsilon)\varepsilon}\big(\Phi,\mathfrak{a},\mathfrak{b}\big)$ satisfies $m^\diamond\leqslant\mathrm{M}$. Exploiting further the definition $\mathcal{R}^\star_{(1-\ln\varepsilon)\,\varepsilon}\big(\Phi,\mathfrak{a},\mathfrak{b}\big) = \min_{m\in\mathbb{N}}\mathcal{R}^m_{(1-\ln\varepsilon)\varepsilon}\big(\Phi,\mathfrak{a},\mathfrak{b}\big)$ it follows that

$$\mathcal{R}_\varepsilon\big[\widehat{f}_{\widetilde{m}}|\Phi,\mathcal{F}^r_{\mathfrak{a}},T\big]\leqslant C(d,r)\,\mathcal{R}^\star_{(1-\ln\varepsilon)\,\varepsilon}.$$

Observe that the upper bound in the last display features a logarithmic factor when compared to the minimax rate of convergence $\mathcal{R}^\star_\varepsilon\big(\Phi,\mathfrak{a},\mathfrak{b}\big)$. The completely data-driven estimator is consequently optimal or nearly optimal in the minimax sense simultaneously over a variety of solution sets $\mathcal{F}^r_{\mathfrak{a}}$ and for all $T\in\mathcal{S}^d_{\mathfrak{b}}$. □

Consider next a global $\mathbb{H}_{\mathfrak{v}}$-risk and consequently, the contrast

$$\Upsilon_m = \max_{m\leqslant k\leqslant \mathrm{M}} \Big\{ \|\widehat{f}_k - \widehat{f}_m\|_{\mathfrak{v}}^2 - \mathrm{pen}_k \Big\}, \quad 1 \leqslant m \leqslant \mathrm{M}.$$

Given $T \in \mathcal{S}_{\mathfrak{b}}^d$ with associated sequence of singular values $(\mathfrak{s}_j)_{j\in\mathbb{N}}$ we set

$$\mathrm{M} := \max\left\{ 1 \leqslant m \leqslant \lfloor \varepsilon^{-1} \rfloor : \lfloor \varepsilon^{-1} \rfloor : \varepsilon \sum_{j=1}^{m} \mathfrak{v}_j \mathfrak{s}_j^{-2} \leqslant 1 \right\}$$

$$\text{and} \quad 15\varepsilon \sum_{j=1}^{k} \mathfrak{v}_j \mathfrak{s}_j^{-2}, \quad 1 \leqslant k \leqslant \mathrm{M}$$

(compare with the development in §3.4.16). Set further $|(\mathfrak{v}\mathfrak{s}^{-2})_{\underline{m}}|_\infty := \max_{1\leqslant j\leqslant m} \mathfrak{v}_j \mathfrak{s}_j^{-2}$, $|(\mathfrak{v}\mathfrak{s}^{-2})_{\underline{m}}|_1 := \sum_{j=1}^m \mathfrak{v}_j\mathfrak{s}_j^{-2}$ and $\Sigma_{\mathfrak{v},\mathfrak{s}} := \sum_{m=1}^\infty |(\mathfrak{v}\mathfrak{s}^{-2})_{\underline{m}}|_\infty \exp(-|(\mathfrak{v}\mathfrak{s}^{-2})_{\underline{m}}|_1 / |(\mathfrak{v}\mathfrak{s}^{-2})_{\underline{m}}|_\infty)$.

§4.1.7 **Upper bound of a $\mathbb{H}_{\mathfrak{v}}$-risk.** *Under Assumption §4.1.1 consider the fully data-driven OSE $\widehat{f}_{\widetilde{m}}$ with data-driven dimension parameter $\widetilde{m}$ and the rate $\mathcal{R}_\varepsilon^\star := \mathcal{R}_\varepsilon^\star(\mathfrak{v},\mathfrak{a},\mathfrak{b})$ defined as in §4.1.3. Suppose in addition that $m^\diamond = \arg\min_{m\in\mathbb{N}} \mathcal{R}_\varepsilon^m(\mathfrak{v},\mathfrak{a},\mathfrak{b})$ satisfies $m^\diamond \leqslant \mathrm{M}$ and that the sequence of singular values $\mathfrak{s}$ associated with $T \in \mathcal{S}_{\mathfrak{b}}^d$ satisfies $\Sigma_{\mathfrak{v},\mathfrak{s}} < \infty$ then*

$$\mathcal{R}_\varepsilon\big[\widehat{f}_{\widetilde{m}} | \mathfrak{v}, \mathcal{F}_{\mathfrak{a}}^r, T\big] \leqslant 85 \max(r, 15d^2)\, \mathcal{R}_\varepsilon^\star + 63\Sigma_{\mathfrak{v},\mathfrak{s}}\, \varepsilon, \qquad \textit{for all } \ \varepsilon \in (0,1).$$

Proof of §4.1.6. As in case of (3.18) it is easily verified that $\sup_{f\in\mathcal{F}_{\mathfrak{a}}^r} \max\{\mathrm{bias}_m^2, \mathrm{pen}_m\} \leqslant \max(r, 15d^2)\mathcal{R}_\varepsilon^m(\mathfrak{v},\mathfrak{a},\mathfrak{b})$. Since the sequence $\mathfrak{v}/\mathfrak{s}^2$ satisfies $\Sigma_{\mathfrak{v},\mathfrak{s}} < \infty$, then from (3.16) with $S_m := \sum_{j=1}^m \mathfrak{v}_j\mathfrak{s}_j^{-2} U_j^2$ we obtain following line by line the proof of (3.17) that Assumption §3.4.14 holds with $C = (3/2)\Sigma_{\mathfrak{v},\mathfrak{s}}\, \varepsilon$. Therefore, we can apply Proposition §3.4.15 which in turn leads to the bound

$$\mathcal{R}_\varepsilon\big[\widehat{f}_{\widetilde{m}} | \mathfrak{v}, \mathcal{F}_{\mathfrak{a}}^r, T\big] \leqslant 85 \max(15, r) \mathcal{R}_\varepsilon^{m^\diamond}(\mathfrak{v},\mathfrak{a},\mathfrak{b}) + 63\Sigma_{\mathfrak{v},\mathfrak{s}}\varepsilon,$$

since $m^\diamond = \underset{m\in\mathbb{N}}{\arg\min}\, \mathcal{R}_\varepsilon^m(\mathfrak{v},\mathfrak{a},\mathfrak{b})$ satisfies $m^\diamond \leqslant \mathrm{M}$. The last bound together with the definition $\mathcal{R}_\varepsilon^\star(\mathfrak{v},\mathfrak{a},\mathfrak{b}) = \min_{m\in\mathbb{N}} \mathcal{R}_\varepsilon^m(\mathfrak{v},\mathfrak{a},\mathfrak{b})$ implies the assertion. □

4.2 Gaussian indirect regression

Let $T \in \mathscr{K}(\mathbb{H})$ be strictly positive and $\{u_j, j\in\mathbb{N}\}$ be an ONB in $\mathbb{H}$ not necessarily corresponding to the eigenfunctions of T. Our aim is the reconstruction the function $f \in \mathbb{H}$ based on a noisy version $\widehat{g}_\varepsilon = Tf + \sqrt{\varepsilon}\dot{W}$ of $g = Tf \in \mathbb{G}$ where $\dot{W}$ is a Gaussian white noise. Considering the projection onto the ONB $\{u_j, j\in\mathbb{N}\}$ the observable quantities take the form

$$[\widehat{g}_\varepsilon]_j = [Tf]_j + \sqrt{\varepsilon}[\dot{W}]_j, \quad j \in \mathbb{N} \tag{4.2}$$

where the error terms $\Big\{[\dot{W}]_j, j\in\mathbb{N}\Big\}$ are independent and standard normally distributed. We assume throughout this section that $T \in \mathcal{T}_{\mathfrak{b}}^d$ satisfies a link condition (see §2.7.3) derived from the ONB $\{u_j, j\in\mathbb{N}\}$ and a non-increasing sequence $\mathfrak{b}$. Moreover, let us recall that $\mathbb{E}_{f,T,\varepsilon}$ denote the expectation with respect to the probability measure associated with the observable quantities. Given $T \in \mathcal{T}_{\mathfrak{b}}^d$ the accuracy of any estimator $\widetilde{f}$ is measured by its maximal risk over the class $\mathcal{F}_{\mathfrak{a}}^r$, that is, $\sup_{f\in\mathcal{F}_{\mathfrak{a}}^r}\mathbb{E}_{f,T,\varepsilon}\big[\mathfrak{D}^2_{\text{ist}}(\widetilde{f},f)\big]$. We restrict ourselves again to the two particular cases of a maximal local Φ-risk, $\mathcal{R}_\varepsilon\big[\widetilde{f}|\Phi,\mathcal{F}_{\mathfrak{a}}^r,T\big]$, and a maximal global $\mathbb{H}_{\mathfrak{v}}$-risk, $\mathcal{R}_\varepsilon\big[\widetilde{f}|\mathfrak{v},\mathcal{F}_{\mathfrak{a}}^r,T\big]$. Let us summarize the conditions on the sequences $\mathfrak{a}$, $\mathfrak{b}$ and $\mathfrak{v}$ as well as the linear functional Φ.

§4.2.1 **Assumption.** *Let $\Phi\in\mathcal{L}_{1/\mathfrak{a}}$ and let $\mathfrak{a}$, $\mathfrak{b}$ and $\mathfrak{v}$ be strictly positive sequences of weights such that $1/\mathfrak{a}$, $\mathfrak{b}$ and $\mathfrak{v}/\mathfrak{a}$ are non-increasing with limit zero. For technical reasons and without loss of generality we assume that $[\Phi]_1^2=1$, $\mathfrak{a}_1=1$, $\mathfrak{b}=1$ and $\mathfrak{v}_1=1$.*

4.2.1 Lower bounds

§4.2.2 **Lower bound of a Φ-risk.** *Under Assumption §4.2.1 define for $\varepsilon\in(0,1)$ and $m\in\mathbb{N}$,*

$$\mathcal{R}_\varepsilon^m(\Phi,\mathfrak{a},\mathfrak{b}):=\max\left(\sum_{j>m}\frac{[\Phi]_j^2}{\mathfrak{a}_j},\max(\mathfrak{b}_m^2\mathfrak{a}_m^{-1},\varepsilon)\sum_{j=1}^m\frac{[\Phi]_j^2}{\mathfrak{b}_m^2}\right);$$
$$\mathcal{R}_\varepsilon^\star:=\mathcal{R}_\varepsilon^\star(\Phi,\mathfrak{a},\mathfrak{b}):=\min_{m\in\mathbb{N}}\mathcal{R}_\varepsilon^m(\Phi,\mathfrak{a},\mathfrak{b})$$
$$\text{and}\quad m_\varepsilon^\star:=m_\varepsilon^\star(\Phi,\mathfrak{a},\mathfrak{b}):=\arg\min_{m\in\mathbb{N}}\mathcal{R}_\varepsilon^m(\Phi,\mathfrak{a},\mathfrak{b}).$$

If, in addition, $\eta:=\inf_{\varepsilon\in(0,1)}\big\{\min(\varepsilon\mathfrak{a}_{m_\varepsilon^\star}\mathfrak{b}_{m_\varepsilon^\star}^{-2},(\varepsilon\mathfrak{a}_{m_\varepsilon^\star})^{-1}\mathfrak{b}_{m_\varepsilon^\star}^2)\big\}>0$ then for all $\varepsilon\in(0,1)$

$$\inf_{\widetilde{f}}\mathcal{R}_\varepsilon\big[\widetilde{f}|\Phi,\mathcal{F}_{\mathfrak{a}}^r,T\big]\geqslant\frac{\eta}{8d^2}\min(2rd^2,1)\,\mathcal{R}_\varepsilon^\star,\qquad\text{for all }\ T\in\mathcal{T}_{\mathfrak{b}}^d.$$

Proof of §4.2.2. Keeping in mind that $[\log dP_1/dP_{-1}]([\widehat{g}_\varepsilon])=2\sum_{j\in\mathbb{N}}[\widehat{g}_\varepsilon]_j[Tf_*]_j/\varepsilon$, and hence $KL(P_1,P_{-1})=2\|Tf_*\|_{\mathbb{H}}^2/\varepsilon\leqslant2d^2\|f_*\|_{\mathfrak{b}^2}^2/\varepsilon$, since $T\in\mathcal{T}_{\mathfrak{b}}^d$ (compare §2.7.3), the proof follows line by line the proof of §4.1.2, and we omit the details. □

§4.2.3 **Lower bound of a global $\mathbb{H}_{\mathfrak{v}}$-risk.** *Let $\mathbb{H}_{\mathfrak{v}}$, $\mathcal{T}_{\mathfrak{b}}^d$ and $\mathcal{F}_{\mathfrak{a}}^r$ be derived from the same ONB and weight sequences $\mathfrak{v}$, $\mathfrak{b}$ and $\mathfrak{a}$ such that $\mathfrak{b}$ and $\mathfrak{v}\mathfrak{a}^{-1}$ are non-increasing. Define for all $\varepsilon\in(0,1)$ and $m\in\mathbb{N}$,*

$$\mathcal{R}_{\mathfrak{v}}^m\big[\varepsilon;\mathcal{F}_{\mathfrak{a}}^r,\mathcal{T}_{\mathfrak{b}}^d\big]:=\max(\mathfrak{v}_m\mathfrak{a}_m^{-1},\sum_{j=1}^m\varepsilon\mathfrak{v}_j\mathfrak{b}_j^{-2});\quad m_\varepsilon^\star:=\arg\min_{m\in\mathbb{N}}\mathcal{R}_{\mathfrak{v}}^m\big[\varepsilon;\mathcal{F}_{\mathfrak{a}}^r,\mathcal{T}_{\mathfrak{b}}^d\big];$$
$$\text{and}\quad\mathcal{R}_\varepsilon^\star:=\mathcal{R}_{\mathfrak{v}}^\star\big[\varepsilon;\mathcal{F}_{\mathfrak{a}}^r,\mathcal{T}_{\mathfrak{b}}^d\big]:=\min_{m\in\mathbb{N}}\mathcal{R}_{\mathfrak{v}}^m\big[\varepsilon;\mathcal{F}_{\mathfrak{a}}^r,\mathcal{T}_{\mathfrak{b}}^d\big]$$

If, in addition, $\eta:=\inf_{\varepsilon\in(0,1)}\{(\mathcal{R}_\varepsilon^\star)^{-1}\min(\mathfrak{v}_{m_\varepsilon^\star}\mathfrak{a}_{m_\varepsilon^\star}^{-1},\sum_{j=1}^{m_\varepsilon^\star}\varepsilon\mathfrak{v}_j\mathfrak{b}_j^{-2})\}>0$ then for all $\varepsilon\in(0,1)$

$$\inf_{\widetilde{f}}\mathcal{R}_\varepsilon^{\mathfrak{v}}\big[\widetilde{f}\,|\,\mathcal{F},\Xi\big]\geqslant\frac{\eta}{8d^2}\min(2rd^2,1)\mathcal{R}_{\mathfrak{v}}^\star\big[\varepsilon;\mathcal{F}_{\mathfrak{a}}^r,\mathcal{T}_{\mathfrak{b}}^d\big],\quad\forall\,T\in\mathcal{T}_{\mathfrak{b}}^d.$$

§2. The proof follows line by line the proof of §**??** where we set $f_*:=(\varepsilon\zeta\alpha_\varepsilon)^{1/2}\sum_{j=1}^{m_\varepsilon^\star}\mathfrak{b}_j^{-2}u_j$ with $\alpha_\varepsilon:=\mathcal{R}_\varepsilon^\star(\sum_{j=1}^{m_\varepsilon^\star}\varepsilon\mathfrak{b}_j^{-2}\mathfrak{v}_j)^{-1}$ and $\zeta:=\eta\min(r,2^{-1}d^{-2})$ which belongs to $\mathcal{F}_{\mathfrak{a}}^r$ by a straightforward adaptation of the arguments given in §**??**. Furthermore after some calulus we obtain $[\log dP_\theta/dP_{\theta^{(j)}}]([\widehat{g}_\varepsilon])=(2/\varepsilon)\theta_j[f_*]_j\sum_{k\in\mathbb{N}}([\widehat{g}_\varepsilon]_k-[Tf_\theta]_k)[Tu_j]_k+(2/\varepsilon)\|Tu_j\|_{\mathbb{H}}^2[f_*]_j^2$ and hence $KL(P_\theta,P_{\theta^{(j)}})=2[f_*]_j^2\|Tu_j\|_{\mathbb{H}}^2/\varepsilon\leqslant2d^2[f_*]_j^2\mathfrak{b}_j^2/\varepsilon$ since $T\in\mathcal{T}_{\mathfrak{b}}^d$ (compare §2.7.3). Furthermore we have $2d^2[f_*]_j^2\mathfrak{b}_j^2 \leqslant \varepsilon$ which in turn implies $H(P_\theta, P_{\theta^{(j)}}) \leqslant 1$ for all $1 \leqslant j \leqslant m$ and $\theta \in \Theta$. Thereby, the assumptions of §3.2.10 are satisfied and hence we can apply the bound presented in §3.2.12. Combining this lower bound and $\|\Pi_{\mathbb{U}_m} f_*\|_{\mathfrak{v}}^2 = \zeta\alpha_\varepsilon \sum_{j=1}^{m_\varepsilon^\star} \varepsilon\mathfrak{b}_j^{-2}\mathfrak{v}_j = \zeta\mathcal{R}_\varepsilon^\star$ implies the assertion. □

4.2.2 Minimax optimal estimation

§4.2.4 **Definition**. *Given observations as in (4.2) and a dimension parameter $m \in \mathbb{N}$ consider $\widehat{[f]}_{\underline{m}} := [T]_{\underline{m}}^{-1}[\widehat{g}_\varepsilon]_{\underline{m}}$ and the Galerkin estimator (GE) $\widehat{f}_m := \sum_{j=1}^m \widehat{[f]}_j u_j$ as an estimator of $[f]_{\underline{m}}$ and f respectively.*

§4.2.5 ***Remark***. Since $T \in \mathscr{K}(\mathbb{H})$ is strictly positive it follows that $[T]_{\underline{m}}^{-1}[g]_{\underline{m}} = [f^m]_{\underline{m}}$ determines the unique Galerkin solution $f^m \in \mathbb{H}_{\underline{m}}$ of $g = Tf$. Observe that $\widehat{[f]}_{\underline{m}} = [T]_{\underline{m}}^{-1}[\widehat{g}_\varepsilon]_{\underline{m}}$ satisfies $\mathbb{E}\widehat{[f]}_{\underline{m}} = [T]_{\underline{m}}^{-1}\mathbb{E}[\widehat{g}_\varepsilon]_{\underline{m}} = [T]_{\underline{m}}^{-1}[g]_{\underline{m}} = [f^m]_{\underline{m}}$ and trivially $\widehat{[f]}_{\underline{m}} \sim \mathcal{N}([f^m]_{\underline{m}}, \varepsilon[T]_{\underline{m}}^{-2})$ which in turn implies $\mathbb{E}\|\widehat{[f]}_{\underline{m}} - [f]_{\underline{m}}\|_{\mathfrak{v}}^2 = 2\varepsilon\,\mathrm{tr}([\nabla_{\mathfrak{v}}]_{\underline{m}}^{1/2}[T]_{\underline{m}}^{-2}[\nabla_{\mathfrak{v}}]_{\underline{m}}^{1/2}) + \|\Pi_{\mathbb{H}_{\underline{m}}}(f - f^m)\|_{\mathfrak{v}}^2$ and $[\Phi]_{\underline{m}}^t\widehat{[f]}_{\underline{m}} \sim \mathcal{N}([\Phi]_{\underline{m}}^t[f^m]_{\underline{m}}, \varepsilon[\Phi]_{\underline{m}}^t[T]_{\underline{m}}^{-2}[\Phi]_{\underline{m}})$. □

§4.2.6 **Upper bounds**. *Let the class of functions of interest be an ellipsoid $\mathcal{F}_{\mathfrak{a}}^r$ derived from some ONB $(u_j)_{j\in\mathbb{N}}$. For $m \in \mathbb{N}$ consider the GE $\widehat{f}_m := \sum_{j=1}^m \widehat{[f]}_j u_j$.*

Local Φ-risk: Let $\mathfrak{a}$ be a non decreasing and let $\Phi \in \mathcal{L}_{1/\mathfrak{a}}$. Consider the GE $\widehat{f}_{m_\varepsilon^\star}$ with dimension parameter $m_\varepsilon^\star$ defined as in §4.2.2. For all $\varepsilon \in (0,1)$ we have

$$\mathcal{R}_\varepsilon^\Phi\big[\widehat{f}_{m_\varepsilon^\star}\,|\,\mathcal{F}_{\mathfrak{a}}^r\big] \leqslant (d^2 + r)(d^2+2)^2\,\mathcal{R}_\Phi^\star\big[\varepsilon; \mathcal{F}_{\mathfrak{a}}^r, \mathcal{T}_{\mathfrak{b}}^d\big]; \quad \forall\, T \in \mathcal{T}_{\mathfrak{b}}^d.$$

Global $\mathbb{H}_{\mathfrak{v}}$-risk: Let $\mathfrak{v}$ be a weight sequence such that $\mathfrak{v}^{-1}\mathfrak{a}$ and $\mathfrak{v}\mathfrak{b}^{-2}$ are non-decreasing. Consider the GE $\widehat{f}_{m_\varepsilon^\star}$ with dimension parameter $m_\varepsilon^\star$ as defined in §4.2.3. For all $\varepsilon \in (0,1)$

$$\mathcal{R}_\varepsilon^{\mathfrak{v}}\big[\widehat{f}_{m_\varepsilon^\star}\,|\,\mathcal{F}_{\mathfrak{a}}^r\big] \leqslant (2d^2 + r)(d^2+2)^2\,\mathcal{R}_{\mathfrak{v}}^\star\big[\varepsilon; \mathcal{F}_{\mathfrak{a}}^r, \mathcal{T}_{\mathfrak{b}}^d\big]; \quad \forall\, T \in \mathcal{T}_{\mathfrak{b}}^d.$$

§3. Let $\mathbb{U}_m = \overline{\mathrm{lin}}\,\{u_j, 1 \leqslant j \leqslant m\}$, $m \in \mathbb{N}$.

Local risk: Keeping in mind §4.2.5 and §2.7.5 we have for all $f \in \mathcal{F}_{\mathfrak{a}}^r$,

$$\begin{aligned}
\mathbb{E}|\Phi(\widehat{f}_{m_\varepsilon^\star}) - \Phi(f)|^2 &= \mathbb{V}\mathrm{ar}(\Phi(\widehat{f}_{m_\varepsilon^\star})) + |\mathbb{E}\Phi(\widehat{f}_{m_\varepsilon^\star}) - \Phi(f)|^2 \\
&= \varepsilon[\Phi]_{\underline{m_\varepsilon^\star}}^t[T]_{\underline{m_\varepsilon^\star}}^{-2}[\Phi]_{\underline{m_\varepsilon^\star}} + |\Phi(f_{m_\varepsilon^\star} - f)|^2 \\
&\leqslant \varepsilon\|[T]_{\underline{m_\varepsilon^\star}}^{-1}[\nabla_{\mathfrak{b}}]_{\underline{m_\varepsilon^\star}}\|_s^2\|[\nabla_{\mathfrak{b}}]_{\underline{m_\varepsilon^\star}}^{-1}[\Phi]_{\underline{m_\varepsilon^\star}}\|^2 + (d^2+2)^2 r \max\left\{\sum_{j>m_\varepsilon^\star}\frac{[\Phi]_j^2}{\mathfrak{a}_j}, \frac{\mathfrak{b}_{m_\varepsilon^\star}^2}{\mathfrak{a}_{m_\varepsilon^\star}}\sum_{j=1}^{m_\varepsilon^\star}\frac{[\Phi]_j^2}{\mathfrak{b}_j^2}\right\} \\
&\leqslant \varepsilon d^2(d^2+2)^2\sum_{j=1}^{m_\varepsilon^\star}\frac{[\Phi]_j^2}{\mathfrak{b}_j^2} + (d^2+2)^2 r \max\left\{\sum_{j>m_\varepsilon^\star}\frac{[\Phi]_j^2}{\mathfrak{a}_j}, \frac{\mathfrak{b}_{m_\varepsilon^\star}^2}{\mathfrak{a}_{m_\varepsilon^\star}}\sum_{j=1}^{m_\varepsilon^\star}\frac{[\Phi]_j^2}{\mathfrak{b}_j^2}\right\}
\end{aligned}$$

by applying the Cauchy-Schwarz inequality. Combining the upper bound and the definition of $\mathcal{R}_\Phi^\star\big[\varepsilon; \mathcal{F}_{\mathfrak{a}}^r, \mathcal{T}_{\mathfrak{b}}^d\big]$ given in §4.2.2 implies the result.

Global risk: We use the Pythagorean formula and keeping in mind §4.2.5 we obtain

$$\begin{aligned}
&\mathbb{E}\|\widehat{f}_{m_\varepsilon^\star} - f\|_{\mathfrak{v}}^2 = \mathbb{E}\|\widehat{f}_{m_\varepsilon^\star} - \Pi_{\mathbb{U}_{m_\varepsilon^\star}} f\|_{\mathfrak{v}}^2 + \|\Pi_{\mathbb{U}_{m_\varepsilon^\star}^\perp} f\|_{\mathfrak{v}}^2 \\
&= 2\varepsilon\,\mathrm{tr}([\nabla_{\mathfrak{v}}]_{\underline{m_\varepsilon^\star}}^{1/2}[T]_{\underline{m_\varepsilon^\star}}^{-2}[\nabla_{\mathfrak{v}}]_{\underline{m_\varepsilon^\star}}^{1/2}) + \|\Pi_{\mathbb{U}_{m_\varepsilon^\star}}(f - f_{m_\varepsilon^\star})\|_{\mathfrak{v}}^2 + \|\Pi_{\mathbb{U}_{m_\varepsilon^\star}^\perp} f\|_{\mathfrak{v}}^2 \\
&= 2\varepsilon\,\mathrm{tr}([T]_{\underline{m_\varepsilon^\star}}^{-1}[\nabla_{\mathfrak{b}}]_{\underline{m_\varepsilon^\star}}^{1-1}[\nabla_{\mathfrak{v}}]_{\underline{m_\varepsilon^\star}}[\nabla_{\mathfrak{b}}]_{\underline{m_\varepsilon^\star}}^{-1+1}[T]_{\underline{m_\varepsilon^\star}}^{-1}) + \|f - f_{m_\varepsilon^\star}\|_{\mathfrak{v}}^2 \\
&\leqslant 2\varepsilon\,\mathrm{tr}([\nabla_{\mathfrak{b}}]_{\underline{m_\varepsilon^\star}}^{-1}[\nabla_{\mathfrak{v}}]_{\underline{m_\varepsilon^\star}}[\nabla_{\mathfrak{b}}]_{\underline{m_\varepsilon^\star}}^{-1})\|[T]_{\underline{m_\varepsilon^\star}}^{-1}[\nabla_{\mathfrak{b}}]_{\underline{m_\varepsilon^\star}}\|_s^2 + \|f - f_{m_\varepsilon^\star}\|_{\mathfrak{v}}^2 \\
&\leqslant 2d^2(d^2+2)^2\varepsilon\sum_{j=1}^{m_\varepsilon^\star}\frac{\mathfrak{v}_j}{\mathfrak{b}_j^2} + (d^2+2)^2\, r\, \mathfrak{v}_{m_\varepsilon^\star}\mathfrak{a}_{m_\varepsilon^\star}^{-1}\max\big(1, \mathfrak{b}_{m_\varepsilon^\star}^2\mathfrak{v}_{m_\varepsilon^\star}^{-1}\max_{1\leqslant j\leqslant m_\varepsilon^\star}\mathfrak{v}_j\mathfrak{b}_j^{-2}\big) \\
&\qquad\qquad = 2d^2(d^2+2)^2\varepsilon\sum_{j=1}^{m_\varepsilon^\star}\frac{\mathfrak{v}_j}{\mathfrak{b}_j^2} + (d^2+2)^2\, r\, \mathfrak{v}_{m_\varepsilon^\star}\mathfrak{a}_{m_\varepsilon^\star}^{-1}
\end{aligned}$$

where we used that $\mathfrak{v}\mathfrak{b}^{-2}$ is non-decreasing. The result follows by exploiting §2.2.15 and the definition of $\mathcal{R}_{\mathfrak{v}}^\star\big[\,\varepsilon; \mathcal{F}_{\mathfrak{a}}^r\big]$ given in §4.2.3. □

4.2.3 Adaptive estimation

4.3 Gaussian indirect sequence space model with noise in the operator

4.4 Gaussian inverse regression with noise in the operator

Bibliography

L. Birgé. An alternative point of view on lepski's method. In Monogr., editor, *State of the art in probability and statistics*, volume 36 of *IMS Lecture Notes*, pages 113–133. (Leiden 1999), 2001.

L. D. Brown and M. G. Low. A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Stat.*, 24(6):2524–2535, 1996. doi: 10.1214/aos/1032181166.

N. Dunford and J. T. Schwartz. *Linear Operators, Part I: General Theory*. Wiley Classics Library. John Wiley & Sons Ltd, New York, 1988a.

N. Dunford and J. T. Schwartz. *Linear operators. Part II: Spectral theory, self adjoint operators in Hilbert space.* Wiley Classics Library. John Wiley & Sons Ltd, New York, 1988b.

N. Dunford and J. T. Schwartz. *Linear operators. Part III, Spectral Operators.* Wiley Classics Library. John Wiley & Sons Ltd, New York, 1988c.

H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems.* Kluwer Academic, Dordrecht, 2000.

A. Goldenshluger and O. Lepski. Bandwidth selection in kernel density estimation: Oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, 39:1608–1632, 2011.

J. Hadamard. *Le Problème de Cauchy et les Équations aux Dérivées Partielles Linéaires Hyperboliques*. Paris, Hermann, 1932.

P. R. Halmos. What does the spectral theorem say? *Amer. Math. Monthly*, 70:241–247, 1963.

J. Johannes and R. Schenk. On rate optimal local estimation in functional linear regression. *Electronic Journal of Statistics*, 7:191–216, 2013.

T. Kawata. *Fourier analysis in probability theory.* Academic Press, New York, 1972.

R. Kress. *Linear integral equations*, volume 82 of *Applied Mathematical Sciences.* Springer, New York, NY, 2 edition, 1989.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Stat.*, 28(5):1302–1338, 2000. doi: 10.1214/aos/1015957395.

B. Laurent, C. Ludena, and C. Prieur. Adaptive estimation of linear functionals by model selection. *Electronic Journal of Statistics*, 2(993-1020), 2008.

A. B. Tsybakov. *Introduction to nonparametric estimation.* Springer Series in Statistics. New York, NY: Springer. xii, 214 p., 2009.