Ruprecht-Karls-Universität Heidelberg
Institut für Angewandte Mathematik
Prof. Dr. Jan JOHANNES

Outline of the lecture course

STATISTICS II

Winter semester 2016/17

Preliminary version: February 13, 2017

If you find **errors in the outline**, please send a short note
by email to johannes@math.uni-heidelberg.de

MΛTHEMΛTIKON, Im Neuenheimer Feld 205, 69120 Heidelberg
phone: +49 6221 54.14.190 – fax: +49 6221 54.53.31
email: johannes@math.uni-heidelberg.de
webpage: www.razbaer.eu/ag-johannes/vl/ST2-WS16/

Table of contents

1 Preliminaries **1**
- 1.1 Convergence of random variables . . . 1
- 1.2 Stochastic Landau notation . . . 4

2 M- and Z-estimator **7**
- 2.1 Introduction / motivation / illustration . . . 7
- 2.2 Consistency . . . 11
- 2.3 Asymptotic normality . . . 14
- 2.4 Testing procedures . . . 16

3 Contiguity **19**
- 3.1 Likelihood ratios . . . 19
- 3.2 Contiguity . . . 20

4 Local asymptotic normality (LAN) **23**
- 4.1 Introduction . . . 23
- 4.2 Hellinger-differentiability . . . 25
- 4.3 Limit distributions under alternatives . . . 26
- 4.4 Asymptotic power function . . . 27
- 4.5 Asymptotic relative efficiency . . . 29
- 4.6 Rank tests . . . 29
- 4.7 Asymptotic power of rank tests . . . 32

5 Non-parametric statistics: local smoothing **35**
- 5.1 Non-parametric curve estimation . . . 35
- 5.2 Kernel density estimation . . . 36
- 5.3 Non-parametric regression . . . 40

6 Non-parametric statistics: orthogonal series estimation **43**
- 6.1 Theoretical basics . . . 43
- 6.2 Abstract smoothness condition . . . 47
- 6.3 Approximation by dimension reduction . . . 49
- 6.4 Stochastic process on Hilbert spaces . . . 50
- 6.5 Statistical experiment . . . 52
- 6.6 Orthogonal series estimation . . . 54

7 Minimax optimality **59**
- 7.1 Minimax theory . . . 59
- 7.2 Deriving a lower bound . . . 60
- 7.3 Lower bound based on two hypothesis . . . 61
  - 7.3.1 Examples - lower bound of a maximal Φ-risk . . . 62
- 7.4 Lower bound based on m hypothesis . . . 64
  - 7.4.1 Examples - lower bound of a maximal $\mathbb{H}_{\mathfrak{v}}$-risk . . . 65

Chapter 1

Preliminaries

This chapter presents elements of the PROBABILITY THEORY *along the lines of the lecture course Probability theory II.*

Here and subsequently, for μ a measure on a measurable space $(\Omega, \mathscr{A})$ and $f : \Omega \to \mathbb{R}^k$ a measurable function, μf denotes the integral $\int f d\mu$. In particular, given a probability measure $\mathbb{P}$ and a random variable (r.v.) X distributed according to $\mathbb{P}$ the expectation of $f(X)$ w.r.t. $\mathbb{P}$ is denoted by $\mathbb{P}f$, $\mathbb{E}_{\mathbb{P}} f(X)$ or $\mathbb{E}f(X)$ for short. For example, when applied to the empirical measure $\overline{\mathbb{P}}_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ of a sample $X_1, \ldots, X_n$, the discrete uniform measure on the sample values, this yields $\overline{\mathbb{P}}_n f = \frac{1}{n}\sum_{i=1}^n f(X_i)$. In other words, $\overline{\mathbb{P}}_n f$ is an abbreviation for the average $\frac{1}{n}\sum_{i=1}^n f(X_i)$. Let $\mathcal{M}(\Omega, \mathscr{A}, \mu)$ denote the set of all real-valued Borel-measurable maps on a measure space $(\Omega, \mathscr{A}, \mu)$. Given $f \in \mathcal{M}(\Omega, \mathscr{A}, \mu)$ let $\|f\|_{L^p_\mu} := (\mu|f|^p)^{1/p}$, $p \in [1, \infty)$ and $\|f\|_{L^\infty_\mu} := \inf\{c : \mu(|f| > c) = 0\}$. Thereby, for $p \in [1, \infty]$ we set $L^p_\mu := L^p_\mu(\Omega, \mathscr{A}, \mu) := \{f \in \mathcal{M}(\Omega, \mathscr{A}, \mu) : \|f\|_{L^p_\mu} < \infty\}$ or $L^p := L^p_\mu$ for short. In the sequel, a random vector in $\mathbb{R}^k$ or $\mathbb{R}^k$-valued r.v. is a vector $X = (X^1, \ldots, X^k)$ of real valued r.v.'s defined on a common probability space $(\Omega, \mathscr{A}, \mathbb{P})$. We denote by $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$ the Euclidean norm and inner product on $\mathbb{R}^k$, respectively, i.e, $\|x\| = (\sum_{i=1}^k |x^i|^2)^{1/2}$ and $\langle x, y \rangle = \sum_{i=1}^k x^i y^i$, $x = (x^1, \ldots, x^k), y = (y^1, \ldots, y^k) \in \mathbb{R}^k$. Obviously, for a $\mathbb{R}^k$-valued r.v. $X = (X^1, \ldots, X^k)$, $\|X\|$ is a real valued r.v. and $\|X\| \in L^p$ is equivalent to $X^i \in L^p_{\mathbb{P}}$, i.e., $\|X^i\|_{L^p_{\mathbb{P}}} = (\mathbb{E}|X^i|^p)^{1/p} < \infty$, for each $i \in [\![1, k]\!] := [1, k] \cap \mathbb{Z}$. Moreover, for $x, y \in \mathbb{R}$ we agree on the following notations $\lfloor x \rfloor := \max\{k \in \mathbb{Z} : k \leqslant x\}$ (integer part), $x \vee y = \max(x, y)$ (maximum), $x \wedge y = \min(x, y)$ (minimum), $x^+ = \max(x, 0)$ (positive part), $x^- = \max(-x, 0)$ (negative part) and $|x| = x^- + x^+$ (modulus).

1.1 Convergence of random variables

§1.1.1 **Definition**. Given r.v.'s $X, X_1, X_2, \ldots$ on a probability space $(\Omega, \mathscr{A}, \mathbb{P})$ with values in a metric space $(\mathcal{X}, d)$ equipped with its Borel-σ-field the sequence $(X_n)_{n\in\mathbb{N}}$ converges to X:

(a) *almost surely (a.s.)*, if $\mathbb{P}(\lim_{n\to\infty} d(X_n, X) = 0) = 1$. We write $X_n \overset{n\to\infty}{\longrightarrow} X$ *a.s.*, or briefly, $X_n \overset{a.s.}{\longrightarrow} X$.

(b) *in probability*, if $\lim_{n\to\infty} \mathbb{P}\big(d(X_n, X) \geqslant \varepsilon\big) = 0$ for all $\varepsilon > 0$. We write $X_n \overset{n\to\infty}{\longrightarrow} X$ *in* $\mathbb{P}$, or briefly, $X_n \overset{\mathbb{P}}{\to} X$.

(c) *in distribution*, if $\mathbb{E}_{\mathbb{P}}\big(f(X_n)\big) \overset{n\to\infty}{\longrightarrow} \mathbb{E}_{\mathbb{P}}\big(f(X)\big)$ for any continuous and bounded function $f : \mathcal{X} \to \mathbb{R}$. We write $X_n \overset{n\to\infty}{\longrightarrow} X$ *in distribution*, or briefly, $X_n \overset{d}{\to} X$.

(d) *in L^p or p-th mean*, if $\lim_{n\to\infty} \mathbb{E}|d(X_n, X)|^p = 0$. We write $X_n \overset{n\to\infty}{\longrightarrow} X$ *in* L^p, or briefly, $X_n \overset{L^p}{\longrightarrow} X$. □

§1.1.2 **Remark.** Considering $\mathbb{R}^k$-valued r.v.'s $X, X_1, X_2, \dots$ and the Euclidean norm $\|\cdot\|$ on $\mathbb{R}^k$ convergence of $(X_n)_{n\in\mathbb{N}}$ to X in p-th mean, that is, $\lim_{n\to\infty} \mathbb{E}\,\|X_n - X\|^p = 0$ is equivalent to the convergence of each component in L^p, i.e., $\lim_{n\to\infty} \|X_n^i - X^i\|_{L^p} = 0$, $i \in [\![1, k]\!]$. □

§1.1.3 **Properties.** *Consider r.v.'s $X, X_1, X_2, \dots$ on a probability space $(\Omega, \mathscr{A}, \mathbb{P})$ with values in a metric space $(\mathcal{X}, d)$ equipped with its Borel-σ-field.*

(a) *We have $X_n \xrightarrow{a.s.} X$ if and only if $\sup_{m>n} d(X_m, X_n) \xrightarrow{\mathbb{P}} 0$ if and only if $\sup_{j\geqslant n} d(X_j, X) \xrightarrow{\mathbb{P}} 0$ if and only if $\forall \varepsilon, \delta > 0$, $\exists N(\varepsilon,\delta) \in \mathbb{N}$, $\forall\, n \geqslant N(\varepsilon,\delta)$, $\mathbb{P}\big(\bigcap_{j\geqslant n} \{d(X_j, X) \leqslant \varepsilon\}\big) \geqslant 1 - \delta$.*

(b) *If $X_n \xrightarrow{a.s.} X$, then $X_n \xrightarrow{\mathbb{P}} X$.*

(c) *If $X_n \xrightarrow{a.s.} X$, then $g(X_n) \xrightarrow{a.s.} g(X)$ for any continuous function g.*

(d) *If $X_n \xrightarrow{\mathbb{P}} X$, then $g(X_n) \xrightarrow{\mathbb{P}} g(X)$ for any continuous function g.*

(e) $X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{\mathbb{P}} X \Leftarrow X_n \xrightarrow{L^p} X$ *and* $X_n \xrightarrow{\mathbb{P}} X \Rightarrow X_n \xrightarrow{d} X$ □

§1.1.4 **Properties** (Portemanteau). *Let $X, X_1, X_2, \dots$ be r.v.'s on a probability space $(\Omega, \mathscr{A}, \mathbb{P})$ with values in a metric space $(\mathcal{X}, d)$ equipped with its Borel-σ-field, then the following statements are equivalent:*

(i) $X_n \xrightarrow{d} X$;

(ii) $\liminf_{n\to\infty} \mathbb{P}(X_n \in U) \geqslant \mathbb{P}(X \in U)$ *for all open $U \subset \mathcal{X}$;*

(iii) $\limsup_{n\to\infty} \mathbb{P}(X_n \in F) \leqslant \mathbb{P}(X \in F)$ *for all closed $F \subset \mathcal{X}$;*

(iv) $\lim_{n\to\infty} \mathbb{P}(X_n \in B) = \mathbb{P}(X \in B)$ *for all measurable B with $\mathbb{P}(X \in \partial B) = 0$ where $\overline{B}$, B° and $\partial B = \overline{B} \backslash B^\circ$ is the closure, interior and the boundary of B, respectively.*

§1.1.5 **Property** (Helly-Bray). *For $\mathbb{R}^k$-valued r.v.'s $X, X_1, X_2, \dots$ defining distribution functions for each $x \in \mathbb{R}^k$ by $\mathbb{F}(x) := \mathbb{P}(X \leqslant x)$ and $\mathbb{F}_n(x) := \mathbb{P}(X_n \leqslant x)$, $n \in \mathbb{N}$, are equivalent:* (i) $X_n \xrightarrow{d} X$ *and* (ii) $\lim_{n\to\infty} \mathbb{F}_n(x) = \mathbb{F}(x)$ *for all points of continuity x of $\mathbb{F}$.* □

§1.1.6 **Property** (Continuous mapping theorem). *Let $(\mathcal{X}_1, d_1)$ and $(\mathcal{X}_2, d_2)$ be metric spaces equipped with their Borel-σ-fields and let $\varphi : \mathcal{X}_1 \to \mathcal{X}_2$ be measurable. Denote by U_φ the set of points of discontinuity of φ. If $X, X_1, X_2, \dots$ are $\mathcal{X}_1$-valued r.v.'s with $\mathbb{P}(X \in U_\varphi) = 0$ and $X_n \xrightarrow{d} X$, then $\varphi(X_n) \xrightarrow{d} \varphi(X)$.* □

§1.1.7 **Property** (Slutzky's lemma). *Let $X, X_1, X_2, \dots$ and $Y_1, Y_2, \dots$ be r.v.'s with values in a common metric space $(\mathcal{X}, d)$ satisfying $X_n \xrightarrow{d} X$ and $d(X_n, Y_n) \xrightarrow{\mathbb{P}} 0$. Then $Y_n \xrightarrow{d} X$.* □

§1.1.8 **Remark.** If $X_n \xrightarrow{\mathbb{P}} X$, then $X_n \xrightarrow{d} X$. The converse is false in general. Indeed, if $X, X_1, X_2, \dots$ are independent and identically distributed (i.i.d.) (with nontrivial distribution), then trivially $X_n \xrightarrow{d} X$ but $X_n \not\xrightarrow{\mathbb{P}} X$. □

§1.1.9 **Examples.** Consider $\mathbb{R}^k$-valued r.v.'s $X, X_1, X_2, \dots$ satisfying $X_n \xrightarrow{d} X$.

(i) If $Y_1, Y_2, \dots$ are $\mathbb{R}^k$-valued r.v.'s and $c \in \mathbb{R}^k$ such that $Y_n \xrightarrow{d} c$, then $X_n + Y_n \xrightarrow{d} X + c$.

(ii) If $\Sigma_1, \Sigma_2, \ldots$ are $k \times k$ random matrices and Σ a $k \times k$ matrix such that $\Sigma_n \xrightarrow{d} \Sigma$, then $\Sigma_n X_n \xrightarrow{d} \Sigma X$. If in addition Σ is strictly positive definite, and thus invertible, then $\Sigma_n^{-1} X_n \xrightarrow{d} \Sigma^{-1} X$ and $\Sigma_n^{-1/2} X_n \xrightarrow{d} \Sigma^{-1/2} X$, respectively. □

§1.1.10 **Property** (Law of Large Numbers). *Let $X, X_1, X_2, \ldots$ be i.i.d. $\mathbb{R}^k$-valued r.v.'s with $\mathbb{E}\|X\| < \infty$. Then $\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{a.s.} \mathbb{E}(X)$.* □

§1.1.11 **Property**. *Let X be a $\mathbb{R}^k$-valued r.v. with $\mathbb{E}\|X\|^2 < \infty$. If $b \in \mathbb{R}^k$ and A is a $(k \times k)$-dimensional matrix, then $Y := AX + b$ is a $\mathbb{R}^k$-valued r.v. with $\mathbb{E}\|Y\|^2 < \infty$. If we further denote by $\mu := \mathbb{E}(X)$ and $\Sigma := \mathbb{C}\mathrm{ov}(X) = \mathbb{E}(X-\mu)(X-\mu)^t = \mathbb{E}(XX^t) - \mu\mu^t$ the expectation and covariance matrix of X, respectively, then $\mathbb{E}(Y) = A\mu + b$ and $\mathbb{C}\mathrm{ov}(Y) = A\Sigma A^t$.* □

§1.1.12 **Definition**. A $\mathbb{R}^k$-valued r.v. X with $\mu := \mathbb{E}(X)$ and $\Sigma := \mathbb{C}\mathrm{ov}(X)$ is *multivariate normal distributed*, i.e., $X \sim \mathfrak{N}(\mu, \Sigma)$, if for each $c \in \mathbb{R}^k$ the real valued r.v. $\langle X, c\rangle$ is normal distributed with mean $\langle \mu, c\rangle$ and variance $\langle \Sigma c, c\rangle$, i.e., $\langle X, c\rangle \sim \mathfrak{N}(\langle \mu, c\rangle, \langle \Sigma c, c\rangle)$. If Id_k denotes the k-dimensional identity matrix, then $X \sim \mathfrak{N}(0, \mathrm{Id}_k)$ is called *standard normal distributed*. □

§1.1.13 **Property**. *A r.v. $X = (X^1, \ldots, X^k)$ is standard normal distributed, i.e., $X \sim \mathfrak{N}(0, \mathrm{Id}_k)$ if and only if $X^1, \ldots, X^k$ are independent and identically $\mathfrak{N}(0,1)$-distributed.* □

§1.1.14 **Remark**. In other words, a multivariate $\mathfrak{N}(0, \mathrm{Id}_k)$-distribution equals the product of its marginal $\mathfrak{N}(0,1)$-distributions, or $\mathfrak{N}(0, \mathrm{Id}_k) = \mathfrak{N}^{\otimes k}(0,1) := \prod_{i=1}^k \mathfrak{N}(0,1)$ for short. □

§1.1.15 **Property** (Central Limit Theorem). *Let $X, X_1, X_2, \ldots$ be i.i.d. $\mathbb{R}^k$-valued r.v.'s in L^2, i.e., $\mathbb{E}\|X\|^2 < \infty$. Then $\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \mathbb{E}(X)) \xrightarrow{d} \mathfrak{N}(0, \mathbb{C}\mathrm{ov}(X))$.* □

§1.1.16 **Property** (Lindeberg-Feller CLT). *For each $n \in \mathbb{N}$ let $Y_{n,1}, \ldots, Y_{n,k_n}$ be independent $\mathbb{R}^p$-valued r.v.'s in L^2 such that* (i) $\sum_{i=1}^{k_n} \mathbb{E}\|Y_{n,i}\|^2 \mathbb{1}_{\{\|Y_{n,i}\| > \varepsilon\}} \xrightarrow{n\to\infty} 0$ *for any $\varepsilon > 0$ and* (ii) $\sum_{i=1}^{k_n} \mathbb{C}\mathrm{ov}(Y_{n,i}) \xrightarrow{n\to\infty} \Sigma$. *Then $\sum_{i=1}^{k_n}(Y_{n,i} - \mathbb{E}(Y_{n,i})) \xrightarrow{d} \mathfrak{N}(0, \Sigma)$.* □

§1.1.17 **Example**. Assume i.i.d. $\mathbb{R}^k$-valued r.v.'s $X, X_1, X_2, \ldots$ in L^2 with $\mu = \mathbb{E}(X)$ and $\Sigma = \mathbb{C}\mathrm{ov}(X)$, which is strictly positive definite.

(i) (CLT) $\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \mu) \xrightarrow{d} \mathfrak{N}(0, \Sigma)$,

(ii) (LLN) $\overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{\mathbb{P}} \mu$,

(iii) (LLN) $\frac{1}{n}\sum_{i=1}^n X_i X_i^t \xrightarrow{\mathbb{P}} \mathbb{E}(XX^t)$,

(iv) $\Sigma_n := \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)(X_i - \overline{X}_n)^t = \frac{1}{n}\sum_{i=1}^n X_i X_i^t - \overline{X}_n \overline{X}_n^t \xrightarrow{\mathbb{P}} \mathbb{E}(XX^t) - \mu\mu^t = \mathbb{C}\mathrm{ov}(X) = \Sigma$ (using (ii) and (iii), Slutzky's lemma §1.1.7 and continuous mapping theorem §1.1.6)

(v) $\sqrt{n}\Sigma_n^{-1/2}(\overline{X} - \mu) \xrightarrow{d} \mathfrak{N}(0, \mathrm{Id})$ (using (i), (iv) and Slutzky's lemma §1.1.7 as in the examples §1.1.9 (ii)) □

§1.1.18 **Remark**. A map $\phi : \mathbb{R}^k \to \mathbb{R}^m$, that is defined at least on a neighbourhood of θ_o, is called differentiable at θ_o, if there exists a linear map (matrix) $\dot{\phi}_{\theta_o} := \dot{\phi}(\theta_o) : \mathbb{R}^k \to \mathbb{R}^m$ such

that

$$\lim_{\theta\to\theta_o} \frac{\left\|\phi(\theta) - \phi(\theta_o) - \dot{\phi}_{\theta_o}(\theta - \theta_o)\right\|}{\|\theta - \theta_o\|} = 0.$$

The linear map $x \mapsto \dot{\phi}_{\theta_o} x$ is called *(total) derivative* as opposed to partial derivatives. A sufficient condition for ϕ to be (totally) differentiable is that all partial derivatives $\partial\phi_j(\theta)/\partial\theta_l$ exist for θ in a neighbourhood of θ_o and are continuous at θ_o. □

§1.1.19 **Property** (Delta method). *Let $\phi : \mathbb{R}^k \supset \mathcal{D}_\phi \to \mathbb{R}^m$ be a map defined on a subset $\mathcal{D}_\phi$ of $\mathbb{R}^k$ and differentiable at θ_o. Let $T, T_1, T_2, \ldots$ be r.v.'s taking their values in the domain $\mathcal{D}_\phi$ of ϕ. If $r_n(T_n - \theta_o) \xrightarrow{d} T$ for numbers $r_n \to \infty$, then $r_n(\phi(T_n) - \phi(\theta_o)) \xrightarrow{d} \dot{\phi}_{\theta_o}(T)$. Moreover, the difference between $r_n(\phi(T_n) - \phi(\theta_o))$ and $\dot{\phi}_{\theta_o}(r_n(T_n - \theta_o))$ converges to zero in probability.* □

§1.1.20 **Remark**. Commonly, $\sqrt{n}(T_n - \theta_o) \xrightarrow{d} \mathfrak{N}(\mu, \Sigma)$. Then applying the delta method it follows that $\sqrt{n}(\phi(T_n) - \phi(\theta_o)) \xrightarrow{d} \mathfrak{N}(\dot{\phi}_{\theta_o}\mu, \dot{\phi}_{\theta_o}\Sigma\dot{\phi}^t_{\theta_o})$. □

§1.1.21 **Property** (Markov's inequality). *If X is a $\mathbb{R}^k$-valued r.v. in L^p for some $p \geqslant 1$, then $\mathbb{P}(\|X\| > c) \leqslant c^{-p}\mathbb{E}\,\|X\|^p$.* □

§1.1.22 **Property** (Monotone convergence). *Assume real-valued r.v.'s $X, X_1, X_2, \ldots$ such that $X_1 \leqslant X_2 \leqslant \ldots$ a.s., or $X_n \uparrow$ for short. Then $\mathbb{E}\lim_{n\to\infty} X_n = \lim_{n\to\infty}\mathbb{E}X_n$.* □

§1.1.23 **Property** (Dominated convergence). *Assume real-valued r.v.'s $X, X_1, X_2, \ldots$ such that $X_n \xrightarrow{a.s.} X$. If there exists $Y \in L^1$ with $\sup_{n\geqslant 1}|X_n| \leqslant Y$ a.s., then $\lim_{n\to\infty}\mathbb{E}|X_n - X| = 0$ which in turn implies $X \in L^1$ and $\lim_{n\to\infty}|\mathbb{E}X_n - \mathbb{E}X| = 0$.* □

§1.1.24 **Definition**. Let $(\mathcal{X}, d)$ be a metric space equipped with its Borel-σ-algebra. A sequence of $\mathcal{X}$-valued r.v.'s $(X_n)_{n\in\mathbb{N}}$ is called *(uniformly) tight* (straff) or *bounded in probability*, if, for any $\varepsilon > 0$, there exists a compact set $K_\varepsilon \subset \mathcal{X}$ such that $\mathbb{P}(X_n \in K_\varepsilon) \geqslant 1 - \varepsilon$ for all $n \in \mathbb{N}$. □

§1.1.25 **Remark**. If $(\mathcal{X}, d)$ is Polish, i.e., separable and complete, then every $\mathcal{X}$-valued r.v. X is bounded in probability and thus so is every finite family. □

§1.1.26 **Example**. A sequence $(X_n)_{n\in\mathbb{N}}$ of $\mathbb{R}^k$-valued r.v.'s is bounded in probability, if for any $\varepsilon > 0$, there exists a constant K_ε such that $\mathbb{P}(\|X_n\| > K_\varepsilon) \leqslant \varepsilon$ for all $n \in \mathbb{N}$. □

§1.1.27 **Property** (Prohorov's theorem). *Let $(\mathcal{X}, d)$ be a Polish space equipped with its Borel-σ-algebra and let $X, X_1, X_2, \ldots$ be $\mathcal{X}$-valued r.v.'s.*

(i) *If $X_n \xrightarrow{d} X$, then $(X_n)_{n\in\mathbb{N}}$ is bounded in probability.*

(ii) *If $(X_n)_{n\in\mathbb{N}}$ is bounded in probability, then there exists a sub-sequence $(X_{n_k})_{k\in\mathbb{N}}$ which converges in distribution.*

1.2 Stochastic Landau notation

In the sequel, $X_1, X_2, \ldots$ are r.v.'s on a common probability space $(\Omega, \mathscr{A}, \mathbb{P})$ with values in a metric space $(\mathcal{X}, d)$ equipped with its Borel-σ-algebra. Moreover, $x_1, x_2, \ldots$ belong to $\mathcal{X}$ and $a_1, a_2, \ldots$ are numbers.

§1.2.1 **Notations**. (a) Recall the Landau notations (i) $x_n = o(1)$, if $x_n \overset{n\to\infty}{\longrightarrow} 0$, and (ii) $x_n = O(1)$, if $\sup_{n\in\mathbb{N}} d(x_n, 0) < \infty$, we write analogously (i) $X_n = o_{\mathbb{P}}(1)$, if $X_n \overset{\mathbb{P}}{\to} 0$, and (ii) $X_n = O_{\mathbb{P}}(1)$, if $(X_n)_{n\in\mathbb{N}}$ is bounded in probability.

(b) More generally, given a sequence $(a_n)_{n\in\mathbb{N}}$ of strictly positive numbers, keeping in mind that (i) $x_n = o(a_n)$, if $x_n/a_n = o(1)$, and that (ii) $x_n = O(a_n)$, if $x_n/a_n = O(1)$, we write analogously (i) $X_n = o_{\mathbb{P}}(a_n)$, if $X_n/a_n = o_{\mathbb{P}}(1)$, and (ii) $X_n = O_{\mathbb{P}}(a_n)$, if $X_n/a_n = O_{\mathbb{P}}(1)$.

(c) Assuming a sequence $(A_n)_{n\in\mathbb{N}}$ of strictly positive r.v.'s we write (i) $X_n = o_{\mathbb{P}}(A_n)$, if $X_n/A_n = o_{\mathbb{P}}(1)$, and (ii) $X_n = O_{\mathbb{P}}(A_n)$, if $X_n/A_n = O_{\mathbb{P}}(1)$. □

§1.2.2 **Properties** (Exercise). (a) $o_{\mathbb{P}}(1) + o_{\mathbb{P}}(1) = o_{\mathbb{P}}(1)$ *meaning if* $X_n = o_{\mathbb{P}}(1)$ *and* $Y_n = o_{\mathbb{P}}(1)$ *then* $X_n + Y_n = o_{\mathbb{P}}(1)$*;*

(b) $O_{\mathbb{P}}(1) + o_{\mathbb{P}}(1) = O_{\mathbb{P}}(1)$*;*

(c) $O_{\mathbb{P}}(1) \cdot o_{\mathbb{P}}(1) = o_{\mathbb{P}}(1)$*;*

(d) $(1 + o_{\mathbb{P}}(1))^{-1} = O_{\mathbb{P}}(1)$*;*

(e) $o_{\mathbb{P}}(R_n) = R_n o_{\mathbb{P}}(1)$*;*

(f) $O_{\mathbb{P}}(R_n) = R_n O_{\mathbb{P}}(1)$*;*

(g) $o_{\mathbb{P}}(O_{\mathbb{P}}(1)) = o_{\mathbb{P}}(1)$ *meaning if* $X_n = O_{\mathbb{P}}(1)$ *and* $Y_n = o_{\mathbb{P}}(X_n)$ *then* $Y_n = o_{\mathbb{P}}(1)$. □

Chapter 2

M- and Z-estimator

2.1 Introduction / motivation / illustration

§2.1.1 **Example** *(Linear model).* Describe the dependence of the variation of a real-valued r.v. Y_i (response) on the variation of an explanatory $\mathbb{R}^k$-valued r.v. $X_i = (X_i^1, \dots, X_i^k)^t$ (explanatory variable) by a linear relationship $\mathbb{E}[Y_i|X_i] = \theta_o^1 X_i^1 + \cdots + \theta_o^k X_i^k = X_i^t \theta_o$ or equivalently $Y_i = X_i^t \theta_o + \varepsilon_i$ where ε_i is a random error satisfying $\mathbb{E}[\varepsilon_i|X_i] = 0$. The parameter $\theta_o \in \mathbb{R}^k$ is unknown, and our interest is inference on θ_o. Assuming that $(Y_1, X_1), \dots, (Y_n, X_n)$ form an i.i.d. sample, we write $Y = (Y_1, \dots, Y_n)^t$ and $X^t = (X_1, \dots, X_n)$ for short. Consequently, we have $\mathbb{E}[Y|X] = X\theta_o$. Consider a Least Squares Estimator (LSE) $\widehat{\theta}_o$ satisfying

$$\widehat{\theta}_o \in \operatorname*{arg\,inf}_{\theta \in \mathbb{R}^k} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i^t \theta)^2 = \operatorname*{arg\,inf}_{\theta \in \mathbb{R}^k} \frac{1}{n} \|Y - X\theta\|^2 \tag{2.1}$$

where $\arg\inf$ denotes the set of points attaining the function's smallest value. If $X^t X = \sum_{i=1}^n X_i X_i^t$ is strictly positive definite, and hence, invertible, the unique LSE is given by $\widehat{\theta}_o = (X^t X)^{-1} X^t Y = \left(\frac{1}{n}\sum_{i=1}^n X_i X_i^t\right)^{-1} \frac{1}{n} \sum_{i=1}^n Y_i X_i$. Under "usual" regularity conditions (see §1.1.17) we have $\frac{1}{n}\sum_{i=1}^n X_i X_i^t \xrightarrow{\mathbb{P}} \mathbb{E}(X_1 X_1^t) =: \Omega$ (LLN). If in addition $\mathbb{E}(\varepsilon_i^2|X_i) = \sigma^2$, then $\frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i X_i \xrightarrow{\mathbb{P}} \mathfrak{N}(0, \sigma^2\Omega)$ (CLT). Applying Slutzky's lemma §1.1.7 and the continuous mapping theorem §1.1.6 holds $\sqrt{n}(\widehat{\theta}_o - \theta_o) \xrightarrow{d} \mathfrak{N}(0, \sigma^2(\mathbb{E}(X_1 X_1^t))^{-1})$. A further inference on $\widehat{\theta}_o$ (hypothesis testing, confidence intervals, etc.) might typically based on this asymptotic result. However, the essential assumption is the linear relationship $\theta \mapsto \mathbb{E}[Y|X] = X\theta$. □

§2.1.2 **Example** *(Generalised linear model).* Consider a real r.v. Y_i and a $\mathbb{R}^k$-valued r.v. X obeying $\mathbb{E}[Y_i|X_i] = g(X_i^t \theta_o)$ for a given link function $g : \mathbb{R} \to \mathbb{R}$ and an unknown parameter $\theta_o \in \mathbb{R}^k$. As an illustration consider the effect of a three different drugs on the behaviour of certain animals. Therefore, each drug is given in different dose to certain animals and we count on how many animals an effect occurred. The next table summarises the results of the experiment.

drug	log-dose	effect	no effect	drug	log-dose	effect	no effect
1	1.01	44	6	2	1	18	30
1	0.89	42	7	2	0.71	16	33
1	0.71	24	22	3	1.4	48	2
1	0.58	16	32	3	1.31	43	3
1	0.41	6	44	3	1.18	38	10
2	1.7	48	0	3	1	27	19
2	1.61	47	3	3	0.71	22	24
2	1.48	47	2	3	0.4	7	40
2	1.31	34	14				

Table 1.1: Number of animals exhibit an (no) effect in dependence of the drug's log-dose.

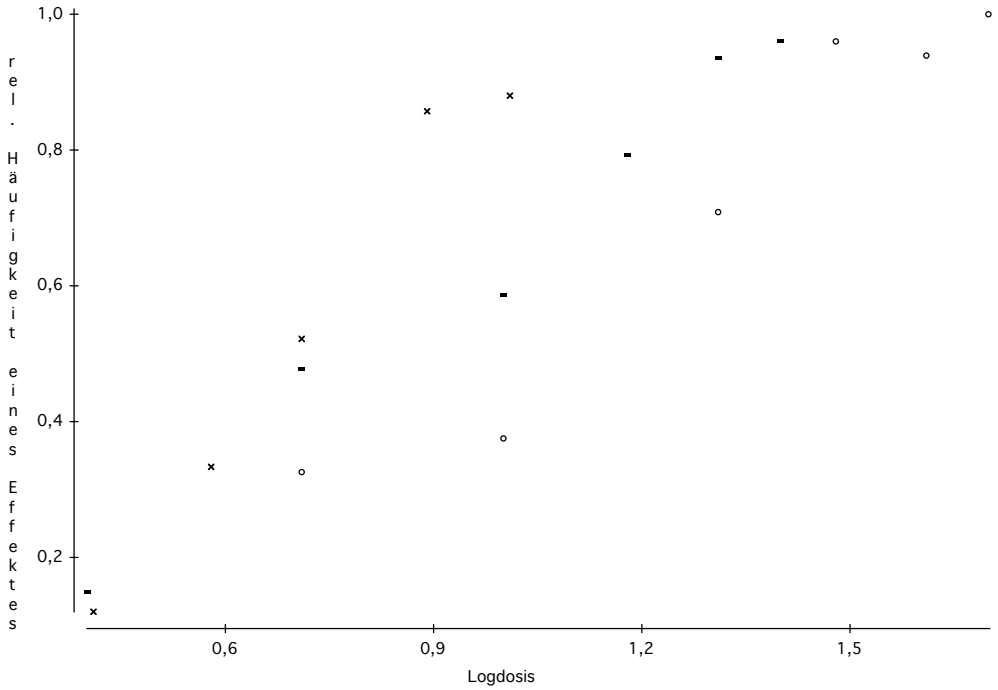
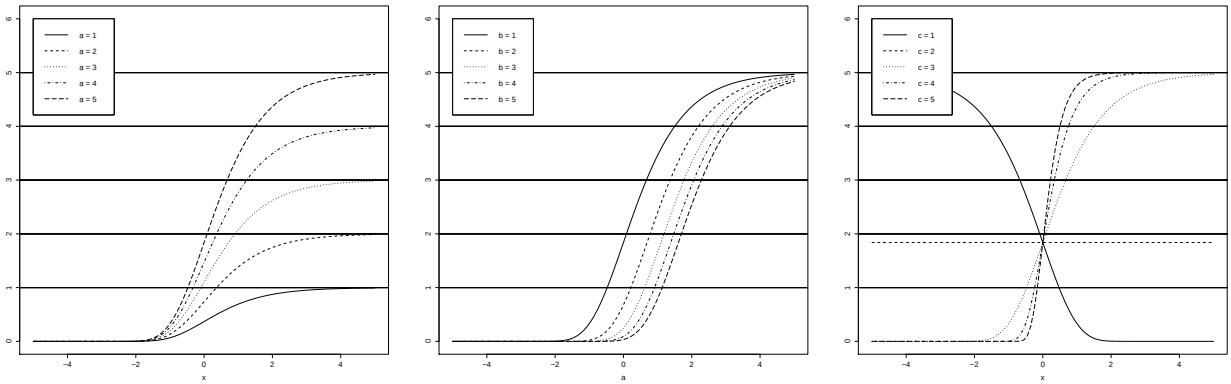


Figure 1.1: Relative frequency of the effects in dependence of the log-dose, drug 1: x; 2: ○; 3: -.

Let Y_{jk} denote the counts of an effect among n_{jk} animals applying a log-dose X_{jk}, $j \in [\![1, J_k]\!]$ of the drug $k \in [\![1, K]\!]$. Assuming an "independent and identical" behaviour of the n_{jk} animals it seems reasonable to model Y_{jk} as Binomial-$\mathfrak{Bin}(n_{jk}, \pi_{jk})$-distributed r.v. with unknown percentage $\pi_{jk} \in (0, 1)$. Typically, it is assumed that $n_{jk}\pi_{jk} = \mathbb{E}[Y_{jk}|X_{jk}] = g(\theta_o^k + \theta_o^0 x_{jk})$ where $\theta_o^1, \ldots, \theta_o^k$ is a drug specific factor and θ_o^0 is a common effect of the log-dose for all drugs. The model is called "Probit" if g is the distribution function of a standard-normal distribution while it is called "Logit" if $g = \frac{e^x}{1+e^x}$ is the logit-distribution function. Keeping in mind example §2.1.1 a LSE $\widehat{\theta}_o \in \arg\inf_{\theta \in \mathbb{R}^k} \sum_{k=1}^{K} \sum_{j=1}^{J_k} (Y_{jk} - g(\theta_o^k + \theta_o^0 x_{jk}))^2$ might be considered. □

§2.1.3 **Example** *(Non-linear regression).* Consider a real r.v. Y_i and a $\mathbb{R}^k$-valued r.v. X obeying $\mathbb{E}[Y_i|X_i] = g(X, \theta_o)$ for a given link function $g : \mathbb{R}^k \times \mathbb{R}^p \to \mathbb{R}$ and an unknown parameter $\theta_o \in \mathbb{R}^p$. The next figure shows, for example, the widely used Gompertz function $g(x) = a \exp(-b \exp(x \log(c)))$.



As an illustration consider the following data of a reaction rate of a catalytic isomerisation of

n-pentane into an isopentane given the partial pressure of hydrogen, n-pentane, and isopentane (see Carr [1960]). Isomerisation is a chemical process where a complex chemical product is transformed into basic elements. The reaction rate depends on several factors as for example, the partial pressure and the concentration of a catalyser (hydrogen).

Reaction	Partial pressure			Reaction	Partial pressure		
rate	hydrogen	n-pentane	isopentane	rate	hydrogen	n-pentane	isopentane
3,541	205,8	90,9	37,1	5,686	297,3	142,2	10,5
2,397	404,8	92,9	36,3	1,193	314	146,7	157,1
6,694	209,7	174,9	49,4	2,648	305,7	142	86
4,722	401,6	187,2	44,9	3,303	300,1	143,7	90,2
0,593	224,9	92,7	116,3	3,054	305,4	141,1	87,4
0,268	402,6	102,2	128,9	3,302	305,2	141,5	87
2,797	212,7	186,9	134,4	1,271	300,1	83	66,4
2,451	406,2	192,6	134,9	11,648	106,6	209,6	33
3,196	133,3	140,8	87,6	2,002	417,2	83,9	32,9
2,021	470,9	144,2	86,9	9,604	251	294,4	41,5
0,896	300	68,3	81,7	7,754	250,3	148	14,7
5,084	301,6	214,6	101,7	11,59	145,1	291	50,2

Table 1.3: Isomerisation reaction rate of an n-pentane into an isopentane.

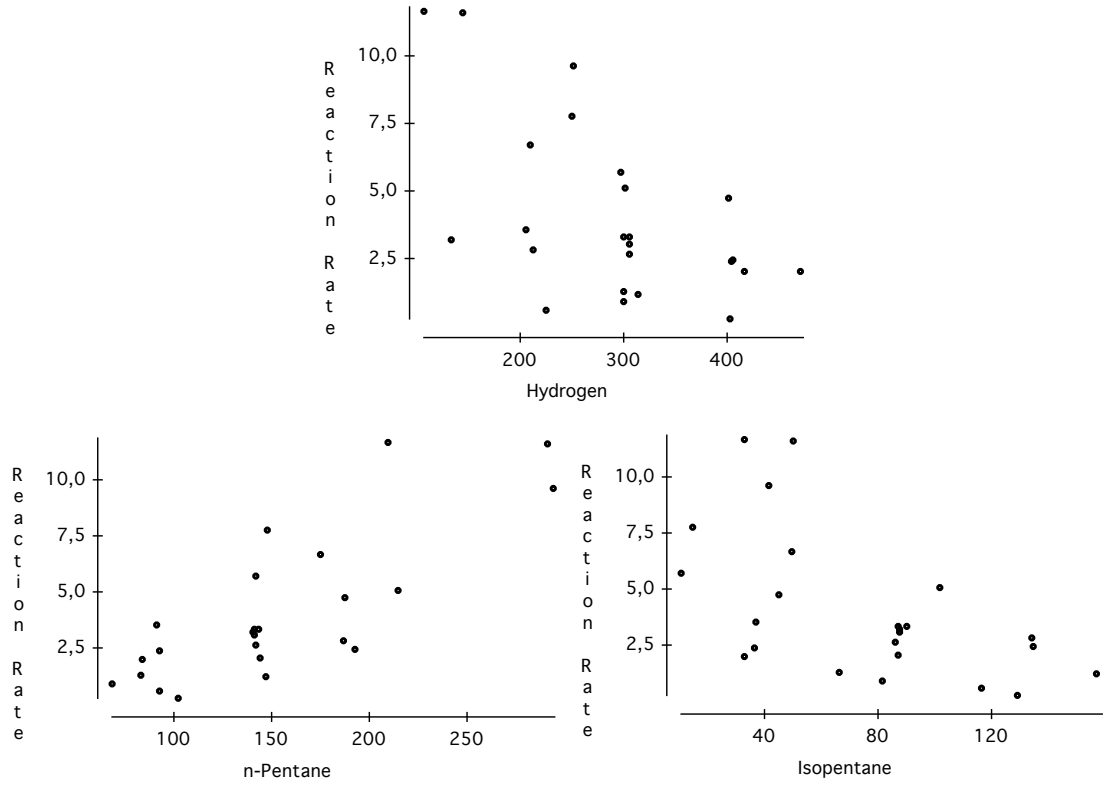


Figure 1.3: Reaction rate in dependence of the partial hydrogen, n-pentane and isopentane pressure.

A commonly used modelling for a reaction rate Y is the Hougen-Watson model where a special

case is given by

$$\mathbb{E}[Y_i|X_i^1, X_i^2, X_i^3] = \frac{\theta_o^1\theta_o^3(X_i^2 - X_i^3/1.632)}{1 + \theta_o^2 X_i^1 + \theta_o^3 X_i^2 + \theta_o^4 X_i^3}, \quad i \in [\![1, n]\!], \tag{2.2}$$

where X_i^1, X_i^2 and X_i^3 is the partial pressure of hydrogen, isopentane and n-pentane, respectively, and $\theta_o^1, \ldots, \theta_o^4$ are unknown parameters. Aiming inference on θ_o we might again consider a LSE $\widehat{\theta}_o \in \arg\inf_{\theta\in\mathbb{R}^k} \sum_{i=1}^n (Y_i - g(X_i, \theta_o))^2$. □

§2.1.4 **Example** *(Quantile regression).* Consider a $\mathbb{R}^p$-valued r.v. Y_i and a $\mathbb{R}^k$-valued r.v. X obeying $Y_i = X_i^t\theta_o + \varepsilon_i$ with $\mathbb{P}(\varepsilon_i \leqslant 0|X) = \alpha$ for a given value $\alpha \in (0, 1)$ for the quantile or equivalently $\mathbb{P}(Y_i \leqslant X_i^t\theta_o|X_i) = \alpha$ meaning that the conditional-α-quantile of Y_i given X_i equals $X_i^t\theta_o$. Keeping in mind that q_α is a α-quantile of Z if $\mathbb{P}(Z \leqslant q_\alpha) = \alpha$. Define[1] $\tau_\alpha(z) := (1-\alpha)z^- + \alpha z^+$ where $\tau_\alpha(z) = (1-\alpha)|z|$ if $z \leqslant 0$ and $\tau_\alpha(z) = \alpha z$ otherwise. Under regularity conditions the function $q \mapsto \mathbb{E}(\tau_\alpha(Z - q))$ attains its minimum at the value $q = q_\alpha$. Roughly, we have

$$\begin{aligned}\frac{\partial}{\partial q}\mathbb{E}(\tau_\alpha(Z-q)) &= (1-\alpha)\frac{\partial}{\partial q}\int_{-\infty}^{q}(q-z)f(z)dz + \alpha\frac{\partial}{\partial q}\int_q^\infty (z-q)f(z)dz \\ &= (1-\alpha)\int_{-\infty}^q f(z)dz - \alpha\int_q^\infty f(z)dz \\ &= (1-\alpha)\mathbb{P}(Z \leqslant q) - \alpha\mathbb{P}(Z > q) = \mathbb{P}(Z \leqslant q) - \alpha.\end{aligned}$$

Consequently, the α-quantile satisfies $0 = \frac{\partial}{\partial q}\mathbb{E}(\tau_\alpha(Z-q))\big|_{q=q_\alpha}$. Thereby, given an i.i.d. sample $(Y_1, X_1), \ldots, (Y_n, X_n)$ a reasonable estimator of θ_o is $\widehat{\theta}_o \in \arg\inf_{\theta\in\mathbb{R}^k} \sum_{i=1}^n \tau_\alpha(Y_i - X_i^t\theta)$. □

§2.1.5 **Example** *(Generalised Method of Moments).* Given a r.v. Z and functions $h_1, \ldots, h_J$ let θ_o be a parameter of interest satisfying $\mathbb{E}[h_j(Z, \theta_o)] = 0$ for $j \in [\![1, J]\!]$ or $\mathbb{E}[H(Z, \theta_o)] = 0$ where $H(Z, \theta_o) = (h_1(Z, \theta_o), \ldots, h_J(Z, \theta_o))^t$ for short. Supposing an i.i.d. sample $Z_1, \ldots, Z_n$ of Z an estimator $\widehat{\theta}_o$ is called a moment estimator if $\frac{1}{n}\sum_{i=1}^n h_j(Z_i, \widehat{\theta}_o) = 0$ for $j \in [\![1, J]\!]$, or $\frac{1}{n}\sum_{i=1}^n H(Z_i, \widehat{\theta}_o) = 0$ for short. Since $\widehat{\theta}_o$ does often not exist or is not unique setting $M_n(\theta) := (\frac{1}{n}\sum_{i=1}^n H(Z_i, \theta))^t W_n (\frac{1}{n}\sum_{i=1}^n H(Z_i, \theta))$ for a given weighting matrix W_n any estimator $\widehat{\theta}_o \in \arg\inf_{\theta\in\Theta} M_n(\theta)$ is called a Generalised Method of Moments (GMM) estimator. □

§2.1.6 **Definition** *(Statistical experiment).* Let $\mathbb{P}_\Theta := \{\mathbb{P}_\theta, \theta \in \Theta\}$ be a family of probability measures on a measurable space $(\mathcal{X}, \mathscr{B})$. The set of indices Θ is called parameter space. If X is a r.v. taking values in $(\mathcal{X}, \mathscr{B})$ with distribution $\mathbb{P}_\theta$ for some $\theta \in \Theta$, i.e. $X \sim \mathbb{P}_\theta$, then we write $X \circlearrowright \mathbb{P}_\Theta$. The triple $(\mathcal{X}, \mathscr{B}, \mathbb{P}_\Theta)$ is called a *statistical experiment* or *statistical model*. If the r.v.'s $X_1, \ldots, X_n$ form an *independent and identically distributed* (i.i.d.) sample of $X \sim \mathbb{P}$, then $\mathbb{P}^{\otimes n} = \otimes_{j=1}^n \mathbb{P}$ denotes its joint product probability measure on the product measure space $(\mathcal{X}^n, \mathscr{B}^{\otimes n})$. We write $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathbb{P}$ or $(X_1, \ldots, X_n) \sim \mathbb{P}^{\otimes n}$ for short. More generally, if $\mathbb{P}_\Theta^n = \{\mathbb{P}_\theta^n, \theta \in \Theta\}$ denotes a family of probability measures on $(\mathcal{X}^n, \mathscr{B}^{\otimes n})$ we write $(X_1, \ldots, X_n) \circlearrowright \mathbb{P}_\Theta^n$ if $(X_1, \ldots, X_n) \sim \mathbb{P}_\theta^n$ for some $\theta \in \Theta$. A statistical model $(\mathcal{X}, \mathscr{B}, \mathbb{P}_\Theta)$ or the family of probability measures $\mathbb{P}_\Theta$ is called *dominated*, if there exists a σ-finite measure μ on $\mathscr{B}$ such that for each $\theta \in \Theta$ the probability measure $\mathbb{P}_\theta$ is absolutely

[1] We use the notation $z^+ = \max(z, 0) = z \vee 0$ and $z^- = (-z) \vee 0$.

continuous w.r.t. μ, i.e. $\mathbb{P}_\theta \ll \mu$. We write $\mathbb{P}_\Theta \ll \mu$ for short. The Radon-Nikodym density $L_\theta(x) := [\mathrm{d}\mathbb{P}_\theta/\mathrm{d}\mu](x)$ and its logarithm $\ell_\theta(x) := \log(L_\theta(x))$ parametrised by $\theta \in \Theta$ is called *likelihood* and *log-likelihood* function, respectively. Assuming $(X_1, \dots, X_n) \wr \mathbb{P}_\Theta^{\otimes n}$, that is, $X_1, \dots, X_n$ form an i.i.d. sample of $X \wr \mathbb{P}_\Theta$, its (joint) likelihood and log-likelihood fulfils $L_\theta(x_1, \dots, x_n) = \prod_{i=1}^n L_\theta(x_i)$ and $\ell_\theta(x_1, \dots, x_n) = \sum_{i=1}^n \ell_\theta(x_i)$, respectively. □

§2.1.7 **Example** *(Maximum-Likelihood-Estimator (MLE)).* Let $X \wr \mathbb{P}_\Theta \ll \mu$. Consider the likelihood $L_\theta(x)$ and log-likelihood $\ell_\theta(x)$ as a function of θ parametrised by x. An estimator $\widehat{\theta} := \widehat{\theta}(X)$ is called Maximum-Likelihood-Estimator (MLE) for θ, if $L_{\widehat{\theta}(x)}(x) = \sup_{\theta\in\Theta} L_\theta(x)$ or equivalently $\ell_{\widehat{\theta}(x)}(x) = \sup_{\theta\in\Theta} \ell_\theta(x)$ for μ-a.e. $x \in \mathcal{X}$. Consequently, based on an i.i.d. sample $X_1, \dots, X_n$ of $X \wr \mathbb{P}_\Theta$ or $(X_1, \dots, X_n) \wr \mathbb{P}_\Theta^{\otimes n}$ the MLE satisfies $\widehat{\theta} \in \arg\sup_{\theta\in\Theta} \overline{\mathbb{P}}_n \ell_\theta$ by using that $\overline{\mathbb{P}}_n \ell_\theta = \frac{1}{n}\sum_{i=1}^n \ell_\theta(X_i)$. □

§2.1.8 **Remark.** Keep in mind that $\mathbb{P}f$ and $\overline{\mathbb{P}}_n f$ denotes the integral $\int f d\mathbb{P}$ and $\int f d\overline{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^n f(X_i)$ w.r.t. $\mathbb{P}$ and the empirical measure $\overline{\mathbb{P}}_n = \frac{1}{n}\delta_{X_i}$ of a sample $X_1, \dots, X_n$, respectively. In all the examples the estimator is characterised either by $\widehat{\theta} \in \arg\sup_{\theta\in\Theta} \overline{\mathbb{P}}_n m_\theta$ for a given real-valued function m_θ of the data or $\widehat{\theta}$ is a zero of the mapping $\theta \mapsto \overline{\mathbb{P}}_n \psi_\theta$ for a given $\mathbb{R}^p$-valued function ψ_θ of the data. Obviously, rather than maximising a criterion function we might search for a zero of the associated normal or estimating equations. □

§2.1.9 **Definition** *(M- and Z-estimator).* We call $\widehat{\theta}$ an *M-estimator*, if $\widehat{\theta}$ maximises a criterion function $M_n(\theta)$ over the parameter space Θ or more generally, if it is a near maximum, that is, $M_n(\widehat{\theta}_n) \geqslant \sup_{\theta\in\Theta} M_n(\theta) - o_{\mathbb{P}}(1)$. We call $\widehat{\theta}$ a *Z-estimator*, if it's a zero of a normal or estimating equation $\Psi_n(\theta)$ or more generally, if it is a near zero, that is, $\Psi_n(\widehat{\theta}_n) = o_{\mathbb{P}}(1)$. □

§2.1.10 **Example.** Consider $(X_1, \dots, X_n) \wr \mathbb{P}_\Theta^n$. Given a function $m_\theta : \mathcal{X} \to \overline{\mathbb{R}}$ any $\widehat{\theta} = \widehat{\theta}(X_1, \dots, X_n)$ (nearly) maximising the map $\theta \mapsto M_n(\theta) := \overline{\mathbb{P}}_n m_\theta = \frac{1}{n}\sum_{i=1}^n m_\theta(X_i)$ is an M-estimator. On the other hand given a function $\psi_\theta : \mathcal{X} \to \mathbb{R}^k$, any $\widehat{\theta} = \widehat{\theta}(X_1, \dots, X_n)$ being a (near) zero of the map $\theta \mapsto \Psi_n(\theta) := \overline{\mathbb{P}}_n \psi_\theta = \frac{1}{n}\sum_{i=1}^n \psi_\theta(X_i)$ is a Z-estimator. □

2.2 Consistency

Here and subsequently, let (Θ, d) be a metric space. An estimator $\widehat{\theta}_n$ of θ_o is called consistent if the sequence $(\widehat{\theta}_n)_{n\in\mathbb{N}}$ converges in probability to θ_o, i.e. $d(\widehat{\theta}_n, \theta_o) = o_{\mathbb{P}}(1)$. For instance, by the LLN the sample mean $\overline{X}_n$ is consistent for the population mean $\mathbb{E}X$.

Consider an M-estimator $\widehat{\theta}_n$ for a random criterion function $M_n(\theta)$. Suppose there is a deterministic criterion function $M(\theta)$ such that $M_n(\theta) \xrightarrow{\mathbb{P}} M(\theta)$ holds point-wise for each $\theta \in \Theta$. For example, due to the LLN $M_n(\theta) = \overline{\mathbb{P}}_n m_\theta \xrightarrow{\mathbb{P}} \mathbb{P}m_\theta = M(\theta)$ provided $\mathbb{P}m_\theta$ exists. The hope is that a maximiser of M_n then converges to the maximising value of M. However, in general point-wise convergence will not be sufficient.

§2.2.1 **Theorem.** *Consider real-valued random functions M_n on Θ, $n \in \mathbb{N}$, and a deterministic real-valued function M on Θ such that for any $\varepsilon > 0$*

(i) $\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| = o_{\mathbb{P}}(1)$ *(uniform convergence in probability);*

(ii) $\sup_{\theta:d(\theta,\theta_o)\geqslant\varepsilon} M(\theta) < M(\theta_o)$ *(identification).*

Any sequence of estimators $(\widehat{\theta}_n)_{n\in\mathbb{N}}$ *of* θ_o *with* $M_n(\widehat{\theta}_n) \geqslant M_n(\theta_o) - o_{\mathbb{P}}(1)$ *is consistent, i.e., converges in probability to* θ_o*.*

Proof of Theorem §2.2.1 is given in the lecture. □

§2.2.2 **Theorem**. *Consider* $\mathbb{R}^k$*-valued random functions* Ψ_n*,* $n \in \mathbb{N}$*, and a deterministic* $\mathbb{R}^k$*-valued function* Ψ *such that for any* $\varepsilon > 0$

(i) $\sup_{\theta\in\Theta} \|\Psi_n(\theta) - \Psi(\theta)\| = o_{\mathbb{P}}(1)$ *(uniform convergence in probability);*

(ii) $\inf_{\theta:d(\theta,\theta_o)\geqslant\varepsilon} \|\Psi(\theta)\| > 0 = \|\Psi(\theta_o)\|$ *(identification).*

Any sequence of estimators $(\widehat{\theta}_n)_{n\in\mathbb{N}}$ *of* θ_o *such that* $\Psi_n(\widehat{\theta}_n) = o_{\mathbb{P}}(1)$ *is consistent, i.e., converges in probability to* θ_o*.*

Proof of Theorem §2.2.2 is given in the lecture. □

§2.2.3 **Lemma**. *Assume that* (i) Θ *is compact,* (ii) $M(\theta_o) > M(\theta)$*, for all* $\theta \neq \theta_o$*, and* (iii) $\theta \mapsto M(\theta)$ *is continuous. Then, it holds* $\sup_{\theta:d(\theta,\theta_o)\geqslant\varepsilon} M(\theta) < M(\theta_o)$ *for all* $\varepsilon > 0$*.*

Proof of Lemma §2.2.3 is left as an exercise. □

§2.2.4 **Example** *(MLE, §2.1.7 continued)*. Keep in mind that a MLE $\widehat{\theta}$ maximises the map $\theta \mapsto \overline{\mathbb{P}}_n\ell_\theta = \frac{1}{n}\sum_{i=1}^n \ell_\theta(X_i)$ or equivalently $\theta \mapsto \overline{\mathbb{P}}_n\ell_\theta - \overline{\mathbb{P}}_n\ell_{\theta_o} = \overline{\mathbb{P}}_n \log(d\mathbb{P}_\theta/d\mathbb{P}_{\theta_o}) = \overline{\mathbb{P}}_n\ell_{\theta,\theta_o}$ where $L_{\theta,\theta_o}(x) := [d\mathbb{P}_\theta/d\mathbb{P}_{\theta_o}](x) = [d\mathbb{P}_\theta/d\mu](x)/[d\mathbb{P}_{\theta_o}/d\mu](x) = L_\theta(x)/L_{\theta_o}(x)$ assuming $\mathbb{P}_\theta \ll \mathbb{P}_{\theta_o}$. Given $\ell_{\theta,\theta_o} := \log(L_{\theta,\theta_o}) = \log(d\mathbb{P}_\theta/d\mathbb{P}_{\theta_o})$ considering $M_n(\theta) := \overline{\mathbb{P}}_n\ell_{\theta,\theta_o}$ and $M(\theta) := \mathbb{P}_{\theta_o}\ell_{\theta,\theta_o}$ we have $M_n(\theta) = M(\theta) + o_{\mathbb{P}_{\theta_o}}(1)$ for all $\theta \in \Theta$. The quantity $KL(\mathbb{P}_{\theta_o}, \mathbb{P}_\theta) = \mathbb{P}_{\theta_o} \log(d\mathbb{P}_{\theta_o}/d\mathbb{P}_\theta) = -\mathbb{P}_{\theta_o}\ell_{\theta,\theta_o}$ is called Kullback-Leibler-divergence of $\mathbb{P}_{\theta_o}$ and $\mathbb{P}_\theta$. Assume here and subsequently that the parameter θ is identifiable, that is, from $\mathbb{P}_{\theta_1} = \mathbb{P}_{\theta_2}$ follows $\theta_1 = \theta_2$. Identifiability is a natural condition since it is a necessary condition for the existence of a consistent estimator. However, if θ is identifiable then $M(\theta) = -KL(\mathbb{P}_{\theta_o}, \mathbb{P}_\theta)$ attains its maximum uniquely at θ_o. Precisely, keeping in mind that $M(\theta_o) = \mathbb{P}_{\theta_o} \log(1) = 0$ it holds $M(\theta) < 0$ for each $\theta \neq \theta_o$. Indeed, employing $\log x \leqslant 2(\sqrt{x} - 1)$ for each $x \geqslant 0$ we have

$$\mathbb{P}_{\theta_o}\ell_{\theta,\theta_o} \leqslant 2\mathbb{P}_{\theta_o}(\sqrt{L_{\theta,\theta_o}} - 1) = 2\langle\sqrt{L_\theta}, \sqrt{L_{\theta_o}}\rangle_{L^2_\mu} - 2 = -\left\|\sqrt{L_\theta} - \sqrt{L_{\theta_o}}\right\|^2_{L^2_\mu}$$

where the right hand side equals zero, if $\theta = \theta_o$, and it is strictly negative, otherwise. The quantity $H(\mathbb{P}_\theta, \mathbb{P}_{\theta_o}) := \left\|\sqrt{L_\theta} - \sqrt{L_{\theta_o}}\right\|_{L^2_\mu}$ is called *Hellinger-distance* between $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta_o}$, which does not depend on the choice of the dominating measure. However, assuming in addition that Θ is compact and $\theta \mapsto \mathbb{P}_{\theta_o}\ell_{\theta,\theta_o}$ is continuous then employing Lemma §2.2.3 the condition (ii) of Theorem §2.2.1 is satisfied. □

§2.2.5 **Proposition**. *If the following conditions*

(i) (Θ, d) *is a compact metric space,*

(ii) $\theta \mapsto M(\theta)$ *is continuous and* $M_n(\theta) = M(\theta) + o_{\mathbb{P}}(1)$ *for all* $\theta \in \Theta$*,*

(iii) $\lim_{\delta\downarrow 0} \limsup_{n\to\infty} \mathbb{P}\big(\sup_{\theta_1,\theta_2:d(\theta_1,\theta_2)\leqslant\delta} |M_n(\theta_1) - M_n(\theta_2)| \geqslant \varepsilon\big) = 0$ *for all* $\varepsilon > 0$*,*

hold, then $\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| = o_{\mathbb{P}}(1)$.

Proof of Proposition §2.2.5 is given in the lecture. □

§2.2.6 **Example**. Let $(X_1, \dots, X_n) \sim \mathbb{P}^{\otimes n}$ and let $m_\theta : \mathcal{X} \to \mathbb{R}$ be a function belonging to $L^1_{\mathbb{P}}$ for all $\theta \in \Theta$. Consider $M_n(\theta) = \overline{\mathbb{P}}_n m_\theta = \frac{1}{n}\sum_{i=1}^n m_\theta(X_i)$ and $M(\theta) = \mathbb{P}m_\theta = \mathbb{E}m_\theta(X)$ where due to the LLN $M_n(\theta) = M(\theta) + o_{\mathbb{P}}(1)$ for each $\theta \in \Theta$. Suppose in addition

(a) (Θ, d) is a compact metric space,

(b) $\theta \mapsto m_\theta(x)$ is continuous for all x,

(c) there is $H \in L^1_{\mathbb{P}}$ with $\sup_{\theta\in\Theta} |m_\theta(x)| \leqslant |H(x)|$ for all x, or equivalently, $\sup_{\theta\in\Theta} |m_\theta(X)|$ belongs to $L^1_{\mathbb{P}}$.

hold. It follows (I) $\theta \mapsto \mathbb{P}m_\theta = M(\theta)$ is continuous and (II) $\sup_{\theta\in\Theta}|M_n(\theta) - M(\theta)| = o_{\mathbb{P}}(1)$. Indeed, by employing dominated convergence (b) and (c) imply together (I). Consider (II). Setting $\Delta^n_\delta := \sup\{|M_n(\theta_1) - M_n(\theta_2)| : d(\theta_1, \theta_2) \leqslant \delta\}$ we show below for all $\varepsilon, \eta > 0$ exists $\delta > 0$ such that $\limsup_{n\to\infty} \mathbb{P}\big(\Delta^n_\delta \geqslant \varepsilon\big) \leqslant \eta$ which in turn by Proposition §2.2.5 implies the claim (II) and, whence condition (i) of Theorem §2.2.1 is satisfied. Given $\varepsilon > 0$ and $\eta > 0$ from (b) and (c) by applying dominated convergence there is $\delta > 0$ such that $\rho_\delta := \mathbb{E}(\sup\{|m_{\theta_1}(X) - m_{\theta_2}(X)|, d(\theta_1, \theta_2) \leqslant \delta\}) \leqslant \eta\varepsilon$ which in turn implies $\mathbb{E}(\Delta^n_\delta) \leqslant \rho_\delta \leqslant \eta\varepsilon$. Employing Markov's inequality the last estimate implies the claim (II).

If in addition $M(\theta_o) > M(\theta)$, for all $\theta \neq \theta_o$, then due to Lemma §2.2.3 also the condition (ii) of Theorem §2.2.1 holds true. Consequently, in this situation any estimator $\widehat{\theta}_n$ of θ_o with $M_n(\widehat{\theta}_n) \geqslant M_n(\theta_o) - o_{\mathbb{P}}(1)$ is consistent, i.e., converges in probability to θ_o. □

§2.2.7 **Remark**. If Θ is not compact we eventually might choose $\Theta_o \subset \Theta$ compact with $\theta_o \in \Theta_o$ satisfying $\sup_{\theta\in\Theta\setminus\Theta_o} M_n(\theta) < M_n\theta_o$ and the conditions (b) and (c) in Example §2.2.6 where Θ is replaced by Θ_o. Then still $\sup_{\theta\in\Theta} \|M_n(\theta) - M(\theta)\| = o_{\mathbb{P}}(1)$ holds true. □

§2.2.8 **Example**. Let $(X_1, \dots, X_n) \sim \mathbb{P}^{\otimes n}$ and for each $\theta \in \Theta$ let $\psi_\theta : \mathcal{X} \to \mathbb{R}^k$ be a function such that the real-valued r.v. $\|\psi_\theta(X)\|$ belongs to $L^1_{\mathbb{P}}$. Keeping in mind that more generally for any function ψ_θ taking values in a separable normed vector space there exists $\mathbb{E}\psi_\theta(X) = \mathbb{P}\psi_\theta$ whenever $\mathbb{E}\|\psi_\theta(X)\| < \infty$. Consider $\Psi_n(\theta) = \overline{\mathbb{P}}_n\psi_\theta = \frac{1}{n}\sum_{i=1}^n \psi_\theta(X_i)$ and $\Psi(\theta) = \mathbb{P}\psi_\theta = \mathbb{E}\psi_\theta(X)$ where due to the LLN $\Psi_n(\theta) = \Psi(\theta) + o_{\mathbb{P}}(1)$ for each $\theta \in \Theta$. Suppose in addition

(a) (Θ, d) is a compact metric space,

(b) $\theta \mapsto \psi_\theta(x)$ is continuous for all x,

(c) $\sup_{\theta\in\Theta} \|\psi_\theta(X)\|$ belongs to $L^1_{\mathbb{P}}$.

hold. It follows (i) $\theta \mapsto \mathbb{P}\psi_\theta = \Psi(\theta)$ is continuous and (ii) $\sup_{\theta\in\Theta} \|\Psi_n(\theta) - \Psi(\theta)\| = o_{\mathbb{P}}(1)$, i.e. condition (i) of Theorem §2.2.2 is satisfied. If in addition $\|\Psi(\theta_o)\| = 0 < \|\Psi(\theta)\|$, for all $\theta \neq \theta_o$, then due to Lemma §2.2.3 also the condition (ii) of Theorem §2.2.2 holds true. Consequently, in this situation any estimator $\widehat{\theta}_n$ of θ_o with $\Psi_n(\widehat{\theta}_n) = o_{\mathbb{P}}(1)$ is consistent, i.e., converges in probability to θ_o. □

§2.2.9 **Remark**. The conditions (i) and (ii) of Theorem §2.2.2 being sufficient to ensure consistency might be weakened in specific situations as we see next. □

§2.2.10 **Proposition**. *Let $\Theta \subset \mathbb{R}$ and $\Psi_n(\theta) = \Psi(\theta) + o_{\mathbb{P}}(1)$ for all $\theta \in \Theta$ where Ψ is a deterministic function. Assume, either*

(i) *$\theta \mapsto \Psi_n(\theta)$ is continuous and has exactly one zero $\widehat{\theta}_n$, or*

(ii) $\theta \mapsto \Psi_n(\theta)$ *is non-decreasing with* $\Psi_n(\widehat{\theta}_n) = o_{\mathbb{P}}(1)$.

Let θ_o *be a point such that* $\Psi(\theta_o - \varepsilon) < 0 < \Psi(\theta_o + \varepsilon)$ *for every* $\varepsilon > 0$. *Then* $\widehat{\theta}_n = \theta_o + o_{\mathbb{P}}(1)$.

Proof of Proposition §2.2.10 is given in the lecture. □

§2.2.11 **Example**. Let $(X_1, \dots, X_n) \sim \mathbb{P}^{\otimes n}$. The sample median $\widehat{\theta}_n$ is a (near) zero of the map $\theta \mapsto \Psi_n(\theta) = \frac{1}{n}\sum_{i=1}^n \operatorname{sign}(X_i - \theta)$ where $\operatorname{sign}(x) := \mathbb{1}_{\{x \geqslant 0\}} - \mathbb{1}_{\{x \leqslant 0\}}$. Considering $\Psi(\theta) = \mathbb{E}\operatorname{sign}(X - \theta) = \mathbb{P}(X > \theta) - \mathbb{P}(X < \theta)$ we have obviously $\Psi_n(\theta) = \Psi(\theta) + o_{\mathbb{P}}(1)$ for each $\theta \in \Theta$. Keeping in mind that $\theta \mapsto \Psi_n(\theta)$ is non-increasing from Proposition §2.2.10 follows consistency of the sample median $\widehat{\theta}_n$, i.e., $\widehat{\theta}_n = \theta_o + o_{\mathbb{P}}(1)$, if for any $\varepsilon > 0$ in addition $\Psi(\theta_o - \varepsilon) > 0 > \Psi(\theta_o + \varepsilon)$ or equivalently $\mathbb{P}(X < \theta_o - \varepsilon) < 1/2 < \mathbb{P}(X < \theta_o + \varepsilon)$. In other words, the sample median $\widehat{\theta}_n$ is a consistent estimator of the population median, if it is unique. □

2.3 Asymptotic normality

Consider $(X_1, \dots, X_n) \sim \mathbb{P}^{\otimes n}$, $\Psi_n(\theta) = \frac{1}{n}\sum_{i=1}^n \psi_\theta(X_i) = \overline{\mathbb{P}}_n\psi_\theta$ and $\Psi(\theta) = \mathbb{P}\psi_\theta$ for $\theta \in \Theta$. Let $\widehat{\theta}_n$ be a zero of $\Psi_n(\theta)$, i.e., $\widehat{\theta}_n$ is a Z-estimator. Assume in addition that $\widehat{\theta}_n = \theta_o + o_{\mathbb{P}}(1)$ where θ_o is a zero of $\Psi(\theta)$. Heuristically, consider a Taylor expansion of a real-valued $\Psi_n(\cdot)$ around $\theta_o \in \Theta \subset \mathbb{R}$, that is, $0 = \Psi_n(\widehat{\theta}_n) = \Psi_n(\theta_o) + (\widehat{\theta}_n - \theta_o)\dot{\Psi}_n(\theta_o) + \frac{1}{2}(\widehat{\theta}_n - \theta_o)\ddot{\Psi}_n(\widetilde{\theta}_n)$ for some $\widetilde{\theta}_n$ between θ_o and $\widehat{\theta}_n$. Thus, rewriting the last identity $\sqrt{n}(\widehat{\theta}_n - \theta_o) = -\sqrt{n}\Psi_n(\theta_o)\big(\dot{\Psi}_n(\theta_o) + \frac{1}{2}(\widehat{\theta}_n - \theta_o)^2\ddot{\Psi}_n(\widetilde{\theta}_n)\big)^{-1}$. If ψ_{θ_o} belongs to $L^2_{\mathbb{P}}$, then due to the CLT it holds $-\sqrt{n}(\Psi_n(\theta_o) - \Psi(\theta_o)) = -\sqrt{n}(\overline{\mathbb{P}}_n\psi_{\theta_o} - \mathbb{P}\psi_{\theta_o}) \xrightarrow{d} \mathfrak{N}(0, \mathbb{P}\psi^2_{\theta_o})$. If moreover $\dot{\psi}_{\theta_o} \in L^1_{\mathbb{P}}$, then by the LLN $\dot{\Psi}_n(\theta_o) = \overline{\mathbb{P}}_n\dot{\psi}_{\theta_o} = \mathbb{P}\dot{\psi}_{\theta_o} + o_{\mathbb{P}}(1)$. If in addition $\ddot{\Psi}_n(\widetilde{\theta}_n) = O_{\mathbb{P}}(1)$ then employing Slutzky's lemma §1.1.7 it follows $\sqrt{n}(\widehat{\theta}_n - \theta_o) \xrightarrow{d} \mathfrak{N}(0, (\mathbb{P}\dot{\psi}_{\theta_o})^{-2}\mathbb{P}\psi^2_{\theta_o})$. In the sequel, θ is a vector and $\Psi(\cdot)$ vector-valued. Consequently, $\dot{\Psi}(\theta_o)$ is matrix and we denote by $\|\dot{\Psi}(\theta_o)\|_F$ its Frobenius norm, where $\|M\|_F := \big(\sum_{j=1}^J \sum_{k=1}^K M^2_{jk}\big)^{1/2}$ for any $J \times K$ matrix $M = (M_{jk})_{j,k}$.

§2.3.1 **Theorem**. *Let the following conditions*

(i) $\theta \mapsto \Psi_n(\theta)$ *is differentiable in a neighbourhood* U *of* $\theta_o \in \Theta^\circ$

(ii) $\dot{\Psi}_n(\theta) := \frac{\partial}{\partial\theta}\Psi_n(\theta)$ *satisfies* $\sup_{\theta\in U}\|\dot{\Psi}_n(\theta) - \dot{\Psi}(\theta)\| = o_{\mathbb{P}}(1)$ *for some continuous deterministic function* $\dot{\Psi}(\theta)$ *with invertible* $\dot{\Psi}(\theta_o)$,

(iii) $\sqrt{n}\Psi_n(\theta_o) \xrightarrow{d} \mathfrak{N}(0, \Omega_o)$ *(CLT),*

hold true. If in addition $\widehat{\theta}_n = \theta_o + o_{\mathbb{P}}(1)$ *with* $\Psi_n(\widehat{\theta}_n) = o_{\mathbb{P}}(n^{-1/2})$ *then* $\sqrt{n}(\widehat{\theta}_n - \theta_o) + \sqrt{n}|\dot{\Psi}(\theta_o)|^{-1}\Psi_n(\theta_o) = o_{\mathbb{P}}(1)$, *and hence* $\sqrt{n}(\widehat{\theta}_n - \theta_o) \xrightarrow{d} \mathfrak{N}(0, |\dot{\Psi}(\theta_o)|^{-1}\Omega_o|\dot{\Psi}(\theta_o)|^{-1})$.

Proof of Theorem §2.3.1 is given in the lecture. □

§2.3.2 **Theorem**. *Let the following conditions*

(i) $\theta \mapsto M_n(\theta)$ *is twice differentiable in a neighbourhood* U *of* $\theta_o \in \Theta^\circ$

(ii) $\ddot{M}_n(\theta) := \frac{\partial^2}{\partial^2\theta}M_n(\theta)$ *satisfies* $\sup_{\theta\in U}\|\ddot{M}_n(\theta) - \ddot{M}(\theta)\| = o_{\mathbb{P}}(1)$ *for some continuous deterministic function* $\ddot{M}(\theta)$ *with invertible* $\ddot{M}(\theta_o)$,

(iii) $\dot{M}_n(\theta) := \frac{\partial}{\partial\theta} M_n(\theta)$ *fulfils* $\sqrt{n}\dot{M}_n(\theta_o) \xrightarrow{d} \mathfrak{N}(0, \Omega_o)$ ***(CLT),***

hold true. If in addition $\widehat{\theta}_n = \theta_o + o_{\mathbb{P}}(1)$ *with* $\widehat{\theta}_n = \sup_{\theta\in\Theta} M_n(\theta)$ *then* $\sqrt{n}(\widehat{\theta}_n - \theta_o) \xrightarrow{d} \mathfrak{N}(0, |\ddot{M}(\theta_o)|^{-1}\Omega_o|\ddot{M}(\theta_o)|^{-1})$.

Proof of Theorem §2.3.2 is given in the lecture. □

§2.3.3 **Example**. Let $(X_1, \ldots, X_n) \sim \mathbb{P}^{\otimes n}$ and let $m_\theta : \mathcal{X} \to \mathbb{R}$ be a function belonging to $L^1_{\mathbb{P}}$ for all $\theta \in \Theta$. Consider $M_n(\theta) = \overline{\mathbb{P}}_n m_\theta = \frac{1}{n}\sum_{i=1}^n m_\theta(X_i)$ and $M(\theta) = \mathbb{P}m_\theta = \mathbb{E}m_\theta(X)$ where due to the LLN $M_n(\theta) = M(\theta) + o_{\mathbb{P}}(1)$ for each $\theta \in \Theta$. Suppose in addition that

(a) $\theta \mapsto m_\theta(x)$ is twice continuously differentiable in a neighbourhood U of $\theta_o \in \Theta^\circ$,

(b) $\dot{m}_\theta := \frac{\partial}{\partial\theta} m_\theta$ belongs to $L^2_{\mathbb{P}}$, fulfils $\mathbb{P}\dot{m}_{\theta_o} = 0$ and ensures the existence of $\mathbb{P}\dot{m}_{\theta_o}\dot{m}^t_{\theta_o}$,

(c) $\ddot{m}_\theta := \frac{\partial^2}{\partial^2\theta} m_\theta$ satisfies $\sup_{\theta\in U} \|\ddot{m}_\theta\|_F \in L^1_{\mathbb{P}}$ and $\mathbb{P}\ddot{m}_\theta$ is strictly negative definite,

hold true. If $\widehat{\theta}_n = \theta_o + o_{\mathbb{P}}(1)$ then $\sqrt{n}(\widehat{\theta}_n - \theta_o) \xrightarrow{d} \mathfrak{N}(0, (\mathbb{P}\ddot{m}_\theta)^{-1}(\mathbb{P}\dot{m}_{\theta_o}\dot{m}^t_{\theta_o})(\mathbb{P}\ddot{m}_\theta)^{-1})$. Indeed, the claim follows from Theorem §2.3.2 if the conditions (i)-(iii) are satisfied, where (i) follows directly from (a). Moreover, following Example §2.2.6, (b) implies the condition (ii) and due to the CLT the condition (iii) follows from (c). We have shown $\sqrt{n}(\widehat{\theta}_n - \theta_o) \xrightarrow{d} \mathfrak{N}(0, H_o^{-1}\Omega_o H_o^{-1})$ where $H_o := \mathbb{P}\ddot{m}_{\theta_o}$ and $\Omega_o := \mathbb{P}\dot{m}_{\theta_o}\dot{m}^t_{\theta_o}$. Thereby, if one wants to use the asymptotic distribution to conduct inference then estimators of H_o and Ω_o are needed. A typical approach to obtain these estimators is as follows. First replacing $\mathbb{P}$ by $\mathbb{P}_n$, the quantity $\widehat{H}(\theta) = \mathbb{P}_n\ddot{m}_\theta$ and $\widehat{\Omega}(\theta) = \mathbb{P}_n\dot{m}_\theta\dot{m}^t_\theta$ is just an empirical counterpart of $H(\theta) = \mathbb{P}\ddot{m}_\theta$ and $\Omega(\theta) = \mathbb{P}\dot{m}_\theta\dot{m}^t_\theta$, respectively. Secondly, replace θ_o by its estimator $\widehat{\theta}_n$ we obtain $\widehat{H}_n := \widehat{H}(\widehat{\theta}_n)$ and $\widehat{\Omega}_n := \widehat{\Omega}(\widehat{\theta}_n)$ as estimator of $H_o := H(\theta_o)$ and $\Omega_o := \Omega(\theta_o)$, respectively. If in addition to (a)-(c) the conditions

(d) (Θ, d) is a compact metric space,

(e) $\sup_{\theta\in U} \|\dot{m}_\theta\|$ belongs to $L^2_{\mathbb{P}}$,

are satisfied, then $\sup_{\theta\in U}\|\widehat{H}(\theta) - H(\theta)\|_F = o_{\mathbb{P}}(1)$ and $\sup_{\theta\in U}\|\widehat{\Omega}(\theta) - \Omega(\theta)\|_F = o_{\mathbb{P}}(1)$ following line by line the arguments in Example §2.2.8. From these uniform convergences and $\widehat{\theta}_n = \theta_o + o_{\mathbb{P}}(1)$ follows $\widehat{H}_n = \widehat{H}(\widehat{\theta}_n) = H(\theta_o) + o_{\mathbb{P}}(1)$ and $\widehat{\Omega}_n = \widehat{\Omega}(\widehat{\theta}_n) = \Omega(\theta_o) + o_{\mathbb{P}}(1)$ which in turn implies $\widehat{V}_n := \widehat{H}_n^{-1}\widehat{\Omega}_n\widehat{H}_n^{-1} = H_o^{-1}\Omega_o H_o^{-1} + o_{\mathbb{P}}(1)$. Consequently, by applying Slutzky's lemma §1.1.7 we have $\sqrt{n}\widehat{V}_n^{-1/2}(\widehat{\theta}_n - \theta_o) \xrightarrow{d} \mathfrak{N}(0, \mathrm{Id})$. □

§2.3.4 **Example** *(MLE, §2.2.4 continued)*. Consider the MLE $\widehat{\theta}_n$ which maximises the (joint) log-likelihood $\theta \mapsto \overline{\mathbb{P}}_n\ell_\theta = \frac{1}{n}\sum_{i=1}^n \ell_\theta(X_i)$ given a sample $(X_1, \ldots, X_n) \bigcirc\!\!\!\!\sim \mathbb{P}_\Theta^{\otimes n}$ with $\mathbb{P}_\Theta \ll \mu$. If the following conditions

(a) (Θ, d) is a compact metric space,

(b) the parameter θ is identifiable, i.e., $\theta_1 \neq \theta_2$ implies $\mathbb{P}_{\theta_1} \neq \mathbb{P}_{\theta_2}$

(c) the map $\theta \mapsto \ell_\theta(x)$ is continuous for all x,

(d) $\sup_{\theta\in\Theta} |\ell_\theta|$ belongs to $L^1_{\mathbb{P}_{\theta_o}}$

hold true, then combining the arguments in Example §2.2.4 and §2.2.6 the assumptions of Theorem §2.2.1 are satisfied, which in turn implies consistency of the MLE $\widehat{\theta}_n = \theta_o + o_{\mathbb{P}_{\theta_o}}(1)$. As shown in the lecture course Statistik 1 if in addition the following conditions

(e) the map $\theta \mapsto \ell_\theta(x)$ is twice continuously differentiable in a neighbourhood U of $\theta_o \in \Theta^\circ$,

(f) $\dot{\ell}_\theta := \frac{\partial}{\partial\theta}\ell_\theta$ satisfies $\sup_{\theta\in U}\|\dot{\ell}_\theta\| \in L^2_{\mathbb{P}_{\theta_o}}$ and $\ddot{\ell}_\theta := \frac{\partial^2}{\partial^2\theta}\ell_\theta$ fulfils $\sup_{\theta\in U}\|\ddot{\ell}_\theta\|_F \in L^1_{\mathbb{P}_{\theta_o}}$,

(g) the Fisher-information matrix $\mathcal{I}_{\theta_o} := \mathbb{P}_{\theta_o}\dot{\ell}_{\theta_o}\dot{\ell}^t_{\theta_o}$ is strictly positive definite,

are fulfilled, then the identity $\mathcal{I}_{\theta_o} = -\mathbb{P}_{\theta_o}\ddot{\ell}_{\theta_o}$ holds true and the MLE satisfies $\sqrt{n}(\widehat{\theta}_n - \theta_o) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathcal{I}^{-1}_{\theta_o}\dot{\ell}_{\theta_o}(X_i) + o_{\mathbb{P}_{\theta_o}}(1)$ and, consequently, $\sqrt{n}(\widehat{\theta}_n - \theta_o) \xrightarrow{d} \mathfrak{N}(0, \mathcal{I}^{-1}_{\theta_o})$. □

§2.3.5 **Remark**. The conditions (e) and (f) in the last example can be weakened replacing differentiability by Hellinger-differentiability. More precisely, keeping in mind the *Hellinger-distance* $H(\mathbb{P}_\theta, \mathbb{P}_{\theta_o}) = \big\|\sqrt{L_\theta} - \sqrt{L_{\theta_o}}\big\|_{L^2_\mu}$ the family $\mathbb{P}_\Theta$ is called *Hellinger-differentiable* in $\theta_o \in \Theta^\circ \subset \mathbb{R}^k$ if there exists a map $x \mapsto \dot{\ell}_{\theta_o}(x) \in \mathbb{R}^k$ such that

$$\int_{\mathcal{X}} \Big|\sqrt{L_\theta(x)} - \sqrt{L_{\theta_o}(x)} - \tfrac{1}{2}\langle\dot{\ell}_{\theta_o}(x), \theta - \theta_o\rangle\sqrt{L_{\theta_o}(x)}\Big|^2 \mu(dx)$$
$$= \Big\|\sqrt{L_\theta} - \sqrt{L_{\theta_o}} - \tfrac{1}{2}\langle\dot{\ell}_{\theta_o}, \theta - \theta_o\rangle\sqrt{L_{\theta_o}}\Big\|^2_{L^2_\mu} = o\big(\|\theta - \theta_o\|^2\big).$$

The map $\theta \mapsto \dot{\ell}_{\theta_o}(x)$ is called score function. We note that Hellinger-differentiability implies $\langle\dot{\ell}_{\theta_o}, \theta - \theta_o\rangle\sqrt{L_{\theta_o}} \in L^2_\mu$ where $\sqrt{L_{\theta_o}} \in L^2_\mu$ using $\|\sqrt{L_{\theta_o}}\|^2_{L^2_\mu} = \int L_{\theta_o}(x)\mu(dx) = 1 < \infty$ and hence $\mathbb{P}_{\theta_o}|\langle\dot{\ell}_{\theta_o}, \theta - \theta_o\rangle|^2 = \|\langle\dot{\ell}_{\theta_o}, \theta - \theta_o\rangle\sqrt{L_{\theta_o}}\|^2_{L^2_\mu} < \infty$ which in turn implies $\dot{\ell}_{\theta_o} \in L^2_{\mathbb{P}_{\theta_o}}$. Thereby, the Fisher-information matrix $\mathcal{I}_{\theta_o} = \mathbb{P}_{\theta_o}\dot{\ell}_{\theta_o}\dot{\ell}^t_{\theta_o}$ is well-defined. Note that, the score function and the Fisher-information matrix are independent of the dominating measure μ. □

2.4 Testing procedures

Consider a parameter of interest $\theta_o \in \Theta$. Given a map $A : \Theta \to \mathbb{R}^p$ we eventually want to test a hypothesis $H_0 : A(\theta_o) = 0$ against the alternative $H_1 : A(\theta_o) \neq 0$. Typical examples include $A(\theta_o) = \theta_o - \theta_*$ for a given value θ_*, or more generally, linear hypothesis $A(\theta_o) = M\theta_o - a_*$ for a given value a_* and matrix M which covers in particular testing of the j-th coordinate of $\theta_o = (\theta^1_o, \ldots, \theta^k_o)$, i.e., $A(\theta_o) = \theta^j_o - a^j_*$. Under regularity conditions it seems reasonable to assume an estimator $\widehat{\theta}_n$ of θ_o having the property $\sqrt{n}(A(\widehat{\theta}_n) - A(\theta_o)) \xrightarrow{d} \mathfrak{N}(0, \Sigma)$ with invertible asymptotic covariance matrix Σ. If we have in addition an estimator $\widehat{\Sigma}_n = \Sigma + o_{\mathbb{P}}(1)$ at hand, then under the hypothesis H_0 a *Wald test* exploits the property $\widehat{W}_n := nA(\widehat{\theta}_n)^t\widehat{\Sigma}^{-1}_n A(\widehat{\theta}_n) \xrightarrow{d} \chi^2_p$ where χ^2_p is a Chi-square-distribution with p degrees of freedom. Precisely, a *Wald test* rejects the hypothesis $H_0 : A(\theta_o) = 0$ if $\widehat{W}_n$ exceeds the 1-α-Quantile $\chi^2_{p,1-\alpha}$ of a χ^2_p-distribution. Obviously, the Wald test does exactly meets the asymptotic level α, i.e., $\lim_{n\to\infty}\mathbb{P}(\widehat{W}_n \geqslant \chi^2_{p,1-\alpha}) = \mathbb{P}(W \geqslant \chi^2_{p,1-\alpha}) = \alpha$ where $W \sim \chi^2_p$. However, the behaviour of the test statistic $\widehat{W}_n$ under the alternative H_1 is still an open questions, which we intent to study in the next sections.

§2.4.1 **Example** *(§2.3.3 continued)*. Consider a sample $X_1 \ldots, X_n \sim \mathbb{P}^{\otimes n}$ and functions $m_\theta : \mathcal{X} \to \mathbb{R}$ belonging to $L^1_{\mathbb{P}}$ for all $\theta \in \Theta \subset \mathbb{R}^k$. For each $\theta \in \Theta$ let $M_n(\theta) = \overline{\mathbb{P}}_n m_\theta = \frac{1}{n}\sum_{i=1}^n m_\theta(X_i)$ and $M(\theta) = \mathbb{P}m_\theta$. Under the conditions (a)-(e) given in Example §2.3.3 an M-estimator $\widehat{\theta}_n := \arg\max_{\theta\in\Theta} M_n(\theta)$ satisfies $\sqrt{n}(\widehat{\theta}_n - \theta_o) \xrightarrow{d} \mathfrak{N}(0, H_o^{-1}\Omega_o H_o^{-1})$. Moreover, we have access to estimators $\widehat{H}_n = H_o + o_{\mathbb{P}}(1)$ and $\widehat{\Omega}_n = \Omega_o + o_{\mathbb{P}}(1)$. Let A be continuously differentiable in a neighbourhood of θ_o then applying the delta method §1.1.19 we obtain

$\sqrt{n}(A(\widehat{\theta}_n) - A(\theta_o)) \xrightarrow{d} \mathfrak{N}(0, \Sigma_o)$ with $\Sigma_o := \dot{A}_{\theta_o} H_o^{-1} \Omega_o H_o^{-1} \dot{A}_{\theta_o}^t$. From $\dot{A}_{\widehat{\theta}_n} = \dot{A}_{\theta_o} + o_{\mathbb{P}}(1)$ follows $\widehat{\Sigma}_n := \dot{A}_{\widehat{\theta}_n} \widehat{H}_n^{-1} \widehat{\Omega}_n \widehat{H}_n^{-1} \dot{A}_{\widehat{\theta}_n}^t = \Sigma + o_{\mathbb{P}}(1)$ and, thus $\sqrt{n}\widehat{\Sigma}_n^{-1/2}(A(\widehat{\theta}_n) - A(\theta_o)) \xrightarrow{d} \mathfrak{N}(0, \mathrm{Id})$ which under H_0 implies $\widehat{W}_n := nA(\widehat{\theta}_n)^t \widehat{\Sigma}_n^{-1} A(\widehat{\theta}_n) \xrightarrow{d} \chi_p^2$. □

Chapter 3

Contiguity

Motivation: Considering a parameter of interest $\theta_o \in \Theta$ we eventually want to test a hypothesis $H_0 : \theta_o \in \Theta_0$ against the alternative $H_1 : \theta_o \in \Theta_1 = \Theta \backslash \Theta_0$. Under regularity conditions we may have at hand an estimator $\widehat{\theta}_n$ of θ_o with the property $\sqrt{n}(\widehat{\theta}_n - \theta_o) \xrightarrow{d} \mathfrak{N}(0, \mathcal{I}_{\theta_o}^{-1})$. Typically based on $\widehat{\theta}_n$ we can construct a test statistic T_n with known asymptotic distribution under H_0 such that the associated test does not exceed asymptotically the given level on the hypothesis H_0. However, we like to invest also its power on the alternative which depends on the specific value of $\theta \in \Theta_1$ commonly getting closer and closer to the hypothesis as the sample size increases.

Here and subsequently, we restrict our attention to two sequences $(\mathbb{P}_n)_{n\in\mathbb{N}}$ and $(\mathbb{Q}_n)_{n\in\mathbb{N}}$ of probability measures. We aim to obtain the limiting distribution of a sequence $(T_n)_{n\in\mathbb{N}}$ of (test) statistics under $\mathbb{Q}_n$ if its limiting distribution under $\mathbb{P}_n$ is known.

3.1 Likelihood ratios

§3.1.1 **Definition**. Let $\mathbb{P}$ and $\mathbb{Q}$ be measures on a common measurable space $(\Omega, \mathscr{A})$. We say $\mathbb{Q}$ is *absolutely continuous* w.r.t. $\mathbb{P}$, if for any $A \in \mathscr{A}$ with $\mathbb{P}(A) = 0$ follows $\mathbb{Q}(A) = 0$. Write $\mathbb{Q} \ll \mathbb{P}$. If $\Omega = \Omega_{\mathbb{P}} \cup \Omega_{\mathbb{Q}}$ with $\Omega_{\mathbb{P}} \cap \Omega_{\mathbb{Q}} = \emptyset$ and $\mathbb{Q}(\Omega_{\mathbb{P}}) = \mathbb{P}(\Omega_{\mathbb{Q}}) = 0$, then $\mathbb{P}$ and $\mathbb{Q}$ are called *orthogonal* or *singular*. Write $\mathbb{Q} \perp \mathbb{P}$. □

§3.1.2 **Remark**. Keep in mind that generally $\mathbb{P}$ and $\mathbb{Q}$ need to be neither absolutely continuous nor singular. Assuming densities q and p w.r.t. some measure μ, we may consider $\Omega_{\mathbb{P}} = \{p > 0\}$ and $\Omega_{\mathbb{Q}} = \{q > 0\}$ where $\mathbb{P}(\Omega \backslash \Omega_p) = \mathbb{P}\mathbb{1}_{\{p=0\}} = 0$. Keep in mind if $\mu(\Omega_p \cap \Omega_q) > 0$ then $\Omega_p \cap \Omega_q$ receives positive measure from both $\mathbb{P}$ and $\mathbb{Q}$. The measure $\mathbb{Q}$ can be written as the sum $\mathbb{Q} = \mathbb{Q}^a + \mathbb{Q}^\perp$ of the measures $\mathbb{Q}^a(A) = \mathbb{Q}(A \cap \{p > 0\})$ and $\mathbb{Q}^\perp(A) = \mathbb{Q}(A \cap \{p = 0\})$ which is called Lebesgue decomposition of $\mathbb{Q}$ w.r.t. $\mathbb{P}$. Where $\mathbb{Q}^a \ll \mathbb{P}$ and $\mathbb{Q}^\perp$ are called absolutely continuous part and the orthogonal (or singular) part of $\mathbb{Q}$ w.r.t. $\mathbb{P}$, respectively. Obviously, the function q/p is a density of $\mathbb{Q}^a$ w.r.t. $\mathbb{P}$ and we denote it $\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P}$ (not: $\mathrm{d}\mathbb{Q}^a/\mathrm{d}\mathbb{P}$!). The density $\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P}$ is only $\mathbb{P}$-almost surely unique by definition. We note that $\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P}$ and the Lebesgue decomposition are independent of the dominating measure. Here and subsequently, we consider $\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P}$ and $\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{Q}$ as r.v.'s on $(\Omega, \mathscr{A})$ with values in $(\mathbb{R}, \mathscr{B})$. □

§3.1.3 **Lemma**. *Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures with densities p and q w.r.t. a measure μ. Then for the measure $\mathbb{Q}^a(A) := \mathbb{Q}\mathbb{1}_A\mathbb{1}_{\{p>0\}}$ and $\mathbb{Q}^\perp(A) := \mathbb{Q}\mathbb{1}_A\mathbb{1}_{\{p=0\}}$*

(i) $\mathbb{Q} = \mathbb{Q}^a + \mathbb{Q}^\perp$, $\mathbb{Q}^a \ll \mathbb{P}$, $\mathbb{Q}^\perp \perp \mathbb{P}$.

(ii) $\mathbb{Q}^a(A) = \mathbb{Q}^a\mathbb{1}_A = \mathbb{P}(\frac{q}{p}\mathbb{1}_A)$ *for every measurable set A.*

(iii) $\mathbb{Q} \ll \mathbb{P}$ *if and only if* $\mathbb{Q}\mathbb{1}_{\{p=0\}} = 0$ *if and only if* $\mathbb{P}\frac{q}{p} = 1$.

Proof of Lemma §3.1.3 is given in the lecture. □

§3.1.4 **Remark**. For each measurable function $f \geqslant 0$ it holds generally

$$\mathbb{Q}f \geqslant \mathbb{Q}f\mathbb{1}_{\{p>0\}} = \mu f q \mathbb{1}_{\{p>0\}} = \mathbb{P}f\tfrac{q}{p}.$$

In particular, for any f identity holds if and only if $\mathbb{Q} \ll \mathbb{P}$. □

3.2 Contiguity

Consider two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on a common measure space $(\Omega, \mathscr{A})$ and let X be a $\mathbb{R}^k$-valued r.v. on Ω. If $\mathbb{Q} \ll \mathbb{P}$, then the $\mathbb{Q}$-law of X, i.e., its induced probability measure $\mathbb{Q}^X$ on $\mathbb{R}^k$, can be calculated from the $\mathbb{P}$-law of the random vector $(X, V) := (X, \mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P})$, i.e., its induced probability measure $\mathbb{P}^{(X,V)}$ on $(\mathbb{R}^{k+1}, \mathscr{B}^{\otimes(k+1)})$, through the formula

$$\mathbb{Q}^X f = \mathbb{E}_{\mathbb{Q}} f(X) = \mathbb{E}_{\mathbb{P}} f(X)\tfrac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} = \mathbb{P}^{(X,V)}[f{\otimes}\mathrm{id}] \quad \text{setting} \quad [f{\otimes}\mathrm{id}](x, v) = f(x)v.$$

Obviously, this relationship could also be expressed as

$$\mathbb{Q}(X \in B) = \mathbb{Q}^X \mathbb{1}_B = \mathbb{E}_{\mathbb{P}} \mathbb{1}_B(X)\tfrac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} = \mathbb{P}^{(X,V)}[\mathbb{1}_B{\otimes}\mathrm{id}]$$

which is only valid under the assumption $\mathbb{Q} \ll \mathbb{P}$, since a part of $\mathbb{Q}$ orthogonal to $\mathbb{P}$ can't be recovered.

We introduce next an asymptotic version of absolute continuity. For $n \in \mathbb{N}$ let $\mathbb{Q}_n$ and $\mathbb{P}_n$ be probability measures on a measurable space $(\Omega_n, \mathscr{A}_n)$. Given for each $n \in \mathbb{N}$ a r.v. X_n defined on $(\Omega_n, \mathscr{A}_n)$ we aim to derive conditions which allow to calculate the $\mathbb{Q}_n$-limit of X_n from a suitable $\mathbb{P}_n$-limit of $(X_n, V_n) := (X_n, \mathrm{d}\mathbb{Q}_n/\mathrm{d}\mathbb{P}_n)$.

§3.2.1 **Definition**. Let $\mathbb{P}_n$ and $\mathbb{Q}_n$ be measures on a common measurable space $(\Omega_n, \mathscr{A}_n)$, $n \in \mathbb{N}$. The sequence $(\mathbb{Q}_n)_{n\in\mathbb{N}}$ is called *contiguous* w.r.t. $(\mathbb{P}_n)_{n\in\mathbb{N}}$, if for any $A_n \in \mathscr{A}_n$, $n \in \mathbb{N}$, with $\lim_{n\to\infty} \mathbb{P}_n(A_n) = 0$ follows $\lim_{n\to\infty} \mathbb{Q}_n(A_n) = 0$. Write $\mathbb{Q}_n \triangleleft \mathbb{P}_n$. The sequences $(\mathbb{Q}_n)_{n\in\mathbb{N}}$ and $(\mathbb{P}_n)_{n\in\mathbb{N}}$ are called *mutually contiguous* if both $\mathbb{Q}_n \triangleleft \mathbb{P}_n$ and $\mathbb{P}_n \triangleleft \mathbb{Q}_n$. Write $\mathbb{Q}_n \triangleleft\triangleright \mathbb{P}_n$. □

Next we characterise contiguity in terms of the asymptotic behaviour of the sequence of likelihood ratios $(\mathrm{d}\mathbb{Q}_n/\mathrm{d}\mathbb{P}_n)_{n\in\mathbb{N}}$ and $(\mathrm{d}\mathbb{P}_n/\mathrm{d}\mathbb{Q}_n)_{n\in\mathbb{N}}$. Keeping in mind that for each $n \in \mathbb{N}$ both likelihood ratios $\mathrm{d}\mathbb{Q}_n/\mathrm{d}\mathbb{P}_n$ and $\mathrm{d}\mathbb{P}_n/\mathrm{d}\mathbb{Q}_n$ are non-negative and satisfy $\mathbb{P}_n\frac{\mathrm{d}\mathbb{Q}_n}{\mathrm{d}\mathbb{P}_n} \leqslant 1$ and $\mathbb{Q}_n\frac{\mathrm{d}\mathbb{P}_n}{\mathrm{d}\mathbb{Q}_n} \leqslant 1$. Employing Markov's inequality §1.1.21 for any $K > 0$ and all $n \in \mathbb{N}$ we have $\mathbb{Q}_n(\mathrm{d}\mathbb{P}_n/\mathrm{d}\mathbb{Q}_n \geqslant K) \leqslant K^{-1}\mathbb{Q}_n\frac{\mathrm{d}\mathbb{P}_n}{\mathrm{d}\mathbb{Q}_n} \leqslant K^{-1}$ and $\mathbb{P}_n(\mathrm{d}\mathbb{Q}_n/\mathrm{d}\mathbb{P}_n \geqslant K) \leqslant K^{-1}$, whence both sequences $(\mathrm{d}\mathbb{Q}_n/\mathrm{d}\mathbb{P}_n)_{n\in\mathbb{N}}$ and $(\mathrm{d}\mathbb{P}_n/\mathrm{d}\mathbb{Q}_n)_{n\in\mathbb{N}}$ are uniformly tight. Consequently, due to Prohorov's theorem §1.1.27 (ii) along a sub-sequence both, $(\mathrm{d}\mathbb{Q}_n/\mathrm{d}\mathbb{P}_n)_{n\in\mathbb{N}}$ and $(\mathrm{d}\mathbb{P}_n/\mathrm{d}\mathbb{Q}_n)_{n\in\mathbb{N}}$, converge in distribution. In analogy to Lemma §3.1.3 (iii) where absolute continuity is shown to be equal to $\mathbb{Q}\mathbb{1}_{\{p=0\}} = 0$ and $\mathbb{P}\frac{q}{p} = 1$ we establish below the equality of contiguity and "each weak limit point of $\mathrm{d}\mathbb{P}_n/\mathrm{d}\mathbb{Q}_n$ under $\mathbb{Q}_n$ gives mass zero to zero" and "each weak limit point of $\mathrm{d}\mathbb{Q}_n/\mathrm{d}\mathbb{P}_n$ under $\mathbb{P}_n$ has mean one". However, the next lemma gathers preliminary results used in the proofs.

§3.2.2 **Lemma**. *For each $n \in \mathbb{N}$ let X_n and Y_n be $(\mathbb{R}^k, \mathscr{B}^{\otimes k})$-valued r.v.'s on a common probability space $(\Omega_n, \mathscr{A}_n, \mathbb{P}_n)$.*

(i) *If there is a $\mathbb{R}^k$-valued r.v. X and a constant $c \in \mathbb{R}^k$ such that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$ then $(X_n, Y_n) \xrightarrow{d} (X, c)$.*

(ii) $X_n \xrightarrow{d} X$ *holds if and only if* $\liminf_{n\to\infty} \mathbb{E} f(X_n) \geqslant \mathbb{E} f(X)$ *for any non-negative and continuous function* f *(not necessarily bounded).*

Proof of Lemma §3.2.2 is left as an exercise. □

§3.2.3 **Lemma**. *Let* $\mathbb{P}_n$ *and* $\mathbb{Q}_n$ *be measures on a common measurable space* $(\Omega_n, \mathscr{A}_n)$, $n \in \mathbb{N}$. *The following statements are equivalent:*

(i) $\mathbb{Q}_n \triangleleft \mathbb{P}_n$,

(ii) *if* $U_n := \mathrm{d}\mathbb{P}_n/\mathrm{d}\mathbb{Q}_n \xrightarrow{d} U$ *under* $\mathbb{Q}_n$, *i.e.,* $\mathbb{Q}_n^{U_n} \xrightarrow{d} \mathbb{Q}^U$, *along a sub-sequence, then* $\mathbb{Q}(U > 0) = \mathbb{E}_{\mathbb{Q}} \mathbb{1}_{\{U>0\}} = 1$,

(iii) *if* $V_n := \mathrm{d}\mathbb{Q}_n/\mathrm{d}\mathbb{P}_n \xrightarrow{d} V$ *under* $\mathbb{P}_n$, *i.e.,* $\mathbb{P}_n^{V_n} \xrightarrow{d} \mathbb{P}^V$, *along a sub-sequence, then* $\mathbb{E}V = 1$,

(iv) *for any r.v.* $T_n : \Omega_n \to \mathbb{R}^k$, $n \in \mathbb{N}$, *with* $T_n \xrightarrow{\mathbb{P}_n} 0$ *follows* $T_n \xrightarrow{\mathbb{Q}_n} 0$.

Proof of Lemma §3.2.3 is given in the lecture. □

§3.2.4 **Example**. Let $\mathbb{P}_n$ and $\mathbb{Q}_n$ be measures on a common measurable space $(\Omega_n, \mathscr{A}_n)$, $n \in \mathbb{N}$ satisfying $\mathrm{d}\mathbb{P}_n/\mathrm{d}\mathbb{Q}_n \xrightarrow{d} U = \exp(W)$ under $\mathbb{Q}_n$ with $W \sim \mathfrak{N}(\mu, \sigma^2)$, then $\mathbb{Q}_n \triangleleft \mathbb{P}_n$. Indeed, from $U = \exp(W) > 0$ a.s. and hence $\mathbb{E}\mathbb{1}_{\{U>0\}} = 1$ follows the claim employing Lemma §3.2.3 (ii). Furthermore, $\mathbb{Q}_n \triangleleft\triangleright \mathbb{P}_n$ holds if and only if $\mu = -\frac{1}{2}\sigma^2$. By using Lemma §3.2.3 (iii) with switched roles of $\mathbb{Q}_n$ and $\mathbb{P}_n$ we have $\mathbb{P}_n \triangleleft \mathbb{Q}_n$ if and only if $1 = \mathbb{E}U = \mathbb{E}\exp(W) = \int \exp(w) \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(w-\mu)^2}{2\sigma^2}) dw = \exp(\frac{(\mu+\sigma^2)^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2})$ which in turn implies the claim. □

§3.2.5 **Theorem**. *Let* $\mathbb{P}_n$ *and* $\mathbb{Q}_n$ *be probability measures, and* X_n *be a* $(\mathbb{R}^k, \mathscr{B}^{\otimes k})$*-valued r.v. on a common measurable space* $(\Omega_n, \mathscr{A}_n)$, $n \in \mathbb{N}$. *Suppose that* $\mathbb{Q}_n \triangleleft \mathbb{P}_n$ *and for* $V_n := \mathrm{d}\mathbb{Q}_n/\mathrm{d}\mathbb{P}_n$ *assume that* $(X_n, V_n) \xrightarrow{d} (X, V)$ *under* $\mathbb{P}_n$, *i.e.,* $\mathbb{P}_n^{(X_n,V_n)} \xrightarrow{d} \mathbb{P}^{(X,V)}$. *Considering the map* $\mathscr{B}^{\otimes k} \ni B \mapsto \mathbb{Q}^X(B) := \mathbb{P}^{X,V}[\mathbb{1}_B \otimes \mathrm{id}] = \mathbb{E}_{\mathbb{P}} \mathbb{1}_B(X)V$, *then* $\mathbb{Q}^X$ *defines a probability measure on* $(\mathbb{R}^k, \mathscr{B}^{\otimes k})$ *satisfying* $\mathbb{Q}^X f = \mathbb{P}^{X,V}[f \otimes \mathrm{id}] = \mathbb{E}_{\mathbb{P}} f(X)V$ *for any* $\mathbb{Q}^X$*-integrable function* f *and* $X_n \xrightarrow{d} \mathbb{Q}^X$ *under* $\mathbb{Q}_n$, *i.e.,* $\mathbb{Q}_n^{X_n} \xrightarrow{d} \mathbb{Q}^X$.

Proof of Theorem §3.2.5 is given in the lecture. □

§3.2.6 **Example** *(Le Cam's third lemma)*. Let $\mathbb{P}_n$ and $\mathbb{Q}_n$ be probability measures, and X_n be a $(\mathbb{R}^k, \mathscr{B}^{\otimes k})$-valued r.v. on a common measurable space $(\Omega_n, \mathscr{A}_n)$, $n \in \mathbb{N}$. Setting $W_n := \log(\mathrm{d}\mathbb{Q}_n/\mathrm{d}\mathbb{P}_n)$ suppose that $\mathbb{P}_n^{(X_n,W_n)} \xrightarrow{d} \mathbb{P}^{X,W}$ where (X, W) is jointly normal distributed with marginals $X \sim \mathfrak{N}_k(\mu, \Sigma)$ and $W \sim \mathfrak{N}(-\sigma^2/2, \sigma^2)$. In other words setting $\tau := \mathbb{C}\mathrm{ov}_{\mathbb{P}}(X, W) = \mathbb{E}_{\mathbb{P}}(X - \mu)(W + \sigma^2/2)$ we assume that

$$\mathbb{P}_n^{(X_n,W_n)} \xrightarrow{d} \mathbb{P}^{(X,W)} = \mathfrak{N}_{k+1}\left(\begin{pmatrix} \mu \\ -\frac{\sigma^2}{2} \end{pmatrix}, \begin{pmatrix} \Sigma & \tau \\ \tau^t & \sigma^2 \end{pmatrix} \right). \tag{3.1}$$

Then, $X_n \xrightarrow{d} \mathfrak{N}_k(\mu + \tau, \Sigma)$ under $\mathbb{Q}_n$, that is, $\mathbb{Q}_n^{X_n} \xrightarrow{d} \mathfrak{N}_k(\mu + \tau, \Sigma)$. Indeed, by the continuous mapping theorem §1.1.6 from (3.1) for $V_n := \exp(W_n) = \mathrm{d}\mathbb{Q}_n/\mathrm{d}\mathbb{P}_n$ follows $\mathbb{P}_n^{(X_n,V_n)} \xrightarrow{d} \mathbb{P}^{(X,V)}$ with $V = \exp(W)$. Since $\mathbb{E}_{\mathbb{P}} V = 1$ following the arguments in Example §3.2.4 we have $\mathbb{Q}_n \triangleleft\triangleright \mathbb{P}_n$ and thus, from Theorem §3.2.5 follows $\mathbb{Q}_n^{X_n} \xrightarrow{d} \mathbb{Q}^X$ with $\mathbb{Q}^X f = \mathbb{P}^{X,V}[f \otimes \mathrm{id}] = \mathbb{E}_{\mathbb{P}} f(X)V$. Thereby, it remains to show that $\mathbb{Q}^X = \mathfrak{N}_k(\mu + \tau, \Sigma)$. Keep in mind that for each

$t \in \mathbb{R}$ the characteristic function $\psi_Z(t) := \mathbb{E}\exp(i\langle t, Z\rangle)$ of $Z \sim \mathfrak{N}(\nu, \Gamma)$ satisfies $\psi_z(t) = \exp(i\langle t, \nu\rangle - \frac{1}{2}\langle \Gamma t, t\rangle)$. Considering the characteristic functions ψ_X and $\psi_{(X,W)}$ of $\mathbb{Q}^X$ and $\mathbb{P}^{(X,W)}$, respectively, the elementary identity $\psi(t) = \mathbb{E}_{\mathbb{Q}}\exp(i\langle t, X\rangle) = \mathbb{E}_{\mathbb{P}}\exp(i\langle t, X\rangle)V = \mathbb{E}_{\mathbb{P}}\exp(i\langle t, X\rangle + W) = \psi_{(X,W)}(t, -i)$ holds for each $t \in \mathbb{R}$. Exploiting (3.1) it is easily seen that $\psi_{(X,W)}(t, -i) = \exp(i\langle t, \mu+\tau\rangle - \frac{1}{2}\langle \Sigma t, t\rangle)$ holds for all $t \in \mathbb{R}$, whence $\mathbb{Q}^X = \mathfrak{N}_k(\mu+\tau, \Sigma)$, which shows the claim. □

Chapter 4

Local asymptotic normality (LAN)

4.1 Introduction

For each $n \in \mathbb{N}$ let $(\Omega_n, \mathscr{A}_n, \mathbb{P}^n_\Theta = \{\mathbb{P}^n_\theta, \theta \in \Theta\})$ be a statistical experiment. Typically, we may think of i.i.d. r.v.'s $X, X_1, X_2, \ldots$ taking values in some measurable space $(\mathcal{X}, \mathscr{B})$ and satisfying $X \circlearrowright \mathbb{P}_\Theta$ for some parametrised familly $\mathbb{P}_\Theta = \{\mathbb{P}_\theta, \theta \in \Theta\}$ of probability measures on $(\mathcal{X}, \mathscr{B})$. In this situation, $(X_1, \ldots, X_n) \circlearrowright \mathbb{P}^{\otimes n}_\Theta$ where $\mathbb{P}^{\otimes n}_\Theta = \{\mathbb{P}^{\otimes n}_\theta, \theta \in \Theta\}$ is a family of product measures on $(\mathcal{X}^n, \mathscr{B}^{\otimes n})$, and thus, $\Omega_n = \mathcal{X}^n$, $\mathscr{A}_n = \mathscr{B}^{\otimes n}$ and $\mathbb{P}^n_\Theta = \mathbb{P}^{\otimes n}_\Theta$.

Aim: Approximate $(\Omega_n, \mathscr{A}_n, \mathbb{P}^n_\Theta)$ in a certain sense by a Gaussian location model after suitable reparametrisation.

§4.1.1 **Definition**. Consider on $(\mathbb{R}^k, \mathscr{B}^{\otimes k})$ the family $\mathfrak{N}(\mathbb{R}^k, \Sigma) := \{\mathfrak{N}(h, \Sigma), h \in \mathbb{R}^k\}$ of multivariate normal distributions with common covariance matrix Σ and varying mean $h \in \mathbb{R}^k$. Noting that for each $h \in \mathbb{R}^k$ the likelihood L_h of $\mathfrak{N}(h, \Sigma)$ w.r.t. the Lebesgue measure satisfies $L_h(x) = L_0(x - h)$ for all $x \in \mathbb{R}^k$ the statistical experiment $(\mathbb{R}^k, \mathscr{B}^{\otimes k}, \mathfrak{N}(\mathbb{R}^k, \Sigma))$ is called a *Gaussian location model*. □

Consider a localised reparametrisation centred around a fixed value θ_o of the parameter which is in the sequel regarded as known.

§4.1.2 **Definition**. Consider a sequence of statistical experiments $(\Omega_n, \mathscr{A}_n, \mathbb{P}^n_\Theta)$, $n \in \mathbb{N}$, with common parameter set $\Theta \subseteq \mathbb{R}^k$. Given a *localising rate* $(\delta_n)_{n\in\mathbb{N}}$ with $\delta_n = o(1)$ for each $n \in \mathbb{N}$ define a *local parameter set* $\Theta^n_o := \{\delta_n^{-1}(\theta - \theta_o) : \theta \in \Theta\} \subseteq \mathbb{R}^k$. For each $\theta \in \Theta$ and associated *local parameter* $h = \delta_n^{-1}(\theta - \theta_o) \in \Theta^n_o$ rewriting $\mathbb{P}^n_\theta$ as $\mathbb{P}^n_{\theta_o + \delta_n h}$ we obtain a sequence of *localised statistical experiment* $(\Omega_n, \mathscr{A}_n, \mathbb{P}^n_{\delta_n\Theta^n_o + \theta_o} := \{\mathbb{P}^n_{\theta_o+\delta_n h}, h \in \Theta^n_o\})$, $n \in \mathbb{N}$. □

§4.1.3 **Remark**. In the sequel we eventually take the local parameter set Θ^n_o equal to $\mathbb{R}^k$ which is not correct if the parameter set Θ is a strict subset of $\mathbb{R}^k$. However, if θ_o is an inner point of Θ, which is assumed throughout this section, then for each $h \in \mathbb{R}^k$ the parameter $\theta = \theta_o + \delta_n h$ belongs to Θ for every sufficiently large n. In other words, the local parameter set Θ^n_o converges to the whole of $\mathbb{R}^k$ as $n \to \infty$, i.e., $\cup_{n\in\mathbb{N}}\Theta^n_o = \mathbb{R}^k$. Thereby, we tactically may either define the probability measure $\mathbb{P}_{\theta_o+\delta_n h}$ arbitrarily if $\theta_o + \delta_n h$ does not belong to Θ, or assume that n is sufficiently large. □

Aim: Show, for large n, that the localised statistical experiment $(\Omega_n, \mathscr{A}_n, \mathbb{P}_{\delta_n\mathbb{R}^k+\theta_o})$ and the Gaussian location model $(\mathbb{R}^k, \mathscr{B}^{\otimes k}, \mathfrak{N}(\mathbb{R}^k, \mathcal{I}^{-1}_{\theta_o}))$ are similar in statistical properties whenever the original experiments, i.e., $\theta \mapsto \mathbb{P}_\theta$, are "smooth".

§4.1.4 **Heuristic**. Consider a μ-dominated family $\mathbb{P}_\Theta$ on $(\mathcal{X}, \mathscr{B})$, i.e., $\mathbb{P}_\Theta \ll \mu$, with $\Theta \subseteq \mathbb{R}$ and likelihood function L_θ of $\mathbb{P}_\theta$ w.r.t. μ. Assume that for all $x \in \mathcal{X}$, the map $\theta \mapsto \ell_\theta(x) =$

$\log(L_\theta(x))$ is twice differentiable with derivatives $\dot\ell_\theta(x)$ and $\ddot\ell_\theta(x)$. For every fixed x a "Taylor expansion of the log of the likelihood-ratio" leads to $\log([L_{\theta+h}/L_\theta](x)) = h\dot\ell_\theta(x) + \frac{1}{2}h^2\ddot\ell_\theta(x) + o_x(h^2)$ where the remainder term depends on x. Consequently, assuming a product experiment $(\mathcal{X}^n, \mathscr{B}^{\otimes n}, \mathbb{P}_\Theta^{\otimes n})$ eventually it holds $\log([L^n_{\theta+h/\sqrt{n}}/L^n_\theta]) = h\sqrt{n}\,\overline{\mathbb{P}}_n\dot\ell_\theta + \frac{1}{2}h^2\,\overline{\mathbb{P}}_n\ddot\ell_\theta + R_n$ where the score $\dot\ell_\theta$ has mean zero, i.e., $\mathbb{P}_\theta\dot\ell_\theta = 0$, and the Fisher information $\mathcal{I}_\theta$ equals $-\mathbb{P}_\theta\ddot\ell_\theta = \mathbb{P}_\theta(\dot\ell_\theta)^2$. Setting $\mathcal{Z}^n_\theta := \sqrt{n}\mathcal{I}_\theta^{-1}\,\overline{\mathbb{P}}_n\dot\ell_\theta$ from the Central Limit Theorem §1.1.15 follows $\mathcal{Z}^n_\theta \xrightarrow{d} \mathfrak{N}(0, \mathcal{I}_\theta^{-1})$ under $\mathbb{P}_\theta^{\otimes n}$ while due to the Law of Large Numbers §1.1.10 it holds $\overline{\mathbb{P}}_n\ddot\ell_\theta = -\mathcal{I}_\theta + o_{\mathbb{P}_\theta^{\otimes n}}(1)$. If in addition the remainder term is negligible, i.e., $R_n = o_{\mathbb{P}_\theta^{\otimes n}}(1)$, then the log of the likelihood-ratio permits an expansion

$$\log(\mathrm{d}\mathbb{P}^{\otimes n}_{\theta+h/\sqrt{n}}/\mathrm{d}\mathbb{P}^{\otimes n}_\theta) = h\mathcal{I}_\theta\mathcal{Z}^n_\theta - \tfrac{1}{2}h^2\mathcal{I}_\theta + o_{\mathbb{P}_\theta^{\otimes n}}(1)$$

which in the limit equals the log of the likelihood-ratio in a Gaussian location model. If the likelihood process permits such an expansion in a neighbourhood of θ we call the sequence of experiments "local asymptotic normal". □

§4.1.5 **Definition**. A sequence of statistical experiments $(\Omega_n, \mathscr{A}_n, \mathbb{P}^n_\mathcal{H})_{n\in\mathbb{N}}$ converges to a *limit experiment* $(\Omega, \mathscr{A}, \mathbb{P}_\mathcal{H})$ if for any finite subset $\mathcal{I} \subset \mathcal{H}$ and each $h_o \in \mathcal{H}$ holds weak convergence of the finite dimensional distributions $(\mathrm{d}\mathbb{P}^n_h/\mathrm{d}\mathbb{P}^n_{h_o}, h \in \mathcal{I}) \xrightarrow{d} (\mathrm{d}\mathbb{P}_h/\mathrm{d}\mathbb{P}_{h_o}, h \in \mathcal{I})$ under $\mathbb{P}^n_{h_o}$. □

§4.1.6 **Definition**. A sequence of statistical experiments $(\Omega_n, \mathscr{A}_n, \mathbb{P}^n_\Theta)_{n\in\mathbb{N}}$ with $\Theta \subseteq \mathbb{R}^k$ is called *locally asymptotic normal (LAN)* in $\theta_o \in \Theta$, if there is a localising rate $(\delta_n)_{n\in\mathbb{N}}$ with $\delta_n = o(1)$, a sequence of r.v.'s $(\mathcal{Z}^n_{\theta_o})_{n\in\mathbb{N}}$ and a strictly positive definite matrix $\mathcal{I}_{\theta_o}$ such that for every $h \in \mathbb{R}^k$ the following three statements hold true:

(i) $\theta_o + \delta_n h \in \Theta$ for all n sufficiently large n, i.e., $n \geqslant n_o(h)$;

(ii) $\mathcal{Z}^n_{\theta_o} \xrightarrow{d} \mathfrak{N}(0, \mathcal{I}_{\theta_o}^{-1})$ under $\mathbb{P}^n_{\theta_o}$;

(iii) $\log(\mathrm{d}\mathbb{P}^{\otimes n}_{\theta_o+\delta_n h}/\mathrm{d}\mathbb{P}^{\otimes n}_{\theta_o}) = \langle\mathcal{I}_{\theta_o}\mathcal{Z}^n_{\theta_o}, h\rangle - \frac{1}{2}\langle\mathcal{I}_{\theta_o}h, h\rangle + R_{n,h}$ where $R_{n,h} = o_{\mathbb{P}_\theta^{\otimes n}}(1)$.

The matrix $\mathcal{I}_{\theta_o}$ is called *Fisher information* at θ_o and $(\mathcal{Z}^n_{\theta_o})_{n\in\mathbb{N}}$ is called *central sequence*. □

§4.1.7 **Remark**. In a Gaussian location model $(\mathbb{R}^k, \mathscr{B}^{\otimes k}, \mathfrak{N}(\mathbb{R}^k, \mathcal{I}_{\theta_o}^{-1}))$ the log of the likelihood-ratio is given by $\log\big(\mathrm{d}\mathfrak{N}(h, \mathcal{I}_{\theta_o}^{-1})/\mathrm{d}\mathfrak{N}(0, \mathcal{I}_{\theta_o}^{-1})\big) = \langle\mathcal{I}_{\theta_o}Z, h\rangle - \frac{1}{2}\langle\mathcal{I}_{\theta_o}h, h\rangle$ where $Z \sim \mathfrak{N}(0, \mathcal{I}_{\theta_o}^{-1})$ under $\mathfrak{N}(0, \mathcal{I}_{\theta_o}^{-1})$. Consequently, if $(\Omega_n, \mathscr{A}_n, \mathbb{P}^n_\Theta)_{n\in\mathbb{N}}$ is LAN then for any finite $\mathcal{I} \subset \mathbb{R}^k$ we have $\big(\log(\mathrm{d}\mathbb{P}^{\otimes n}_{\theta_o+\delta_n h}/\mathrm{d}\mathbb{P}^{\otimes n}_{\theta_o}), h \in \mathcal{I}\big) \xrightarrow{d} \big(\log\big(\mathrm{d}\mathfrak{N}(h, \mathcal{I}_{\theta_o}^{-1})/\mathrm{d}\mathfrak{N}(0, \mathcal{I}_{\theta_o}^{-1})\big), h \in \mathcal{I}\big)$ and whence $\big(\mathrm{d}\mathbb{P}^{\otimes n}_{\theta_o+\delta_n h}/\mathrm{d}\mathbb{P}^{\otimes n}_{\theta_o}, h \in \mathcal{I}\big) \xrightarrow{d} \big(\mathrm{d}\mathfrak{N}(h, \mathcal{I}_{\theta_o}^{-1})/\mathrm{d}\mathfrak{N}(0, \mathcal{I}_{\theta_o}^{-1}), h \in \mathcal{I}\big)$ due to the continuous mapping theorem §1.1.6. In other words the sequence of statistical experiments $(\Omega_n, \mathscr{A}_n, \mathbb{P}^n_\Theta)_{n\in\mathbb{N}}$ has a Gaussian location model as limit experiment. □

§4.1.8 **Definition**. A LAN sequence of statistical experiments is called *uniformly locally asymptotic normal (ULAN)* in $\theta_o \in \Theta$, if the condition (iii) in definition §4.1.6 is replaced by

(iii') for any sequence $h_n \to h$ it holds $\log(\mathrm{d}\mathbb{P}^{\otimes n}_{\theta_o+\delta_n h_n}/\mathrm{d}\mathbb{P}^{\otimes n}_{\theta_o}) = \langle\mathcal{I}_{\theta_o}\mathcal{Z}^n_{\theta_o}, h\rangle - \frac{1}{2}\langle\mathcal{I}_{\theta_o}h, h\rangle + R_{n,h_n}$ where $R_{n,h_n} = o_{\mathbb{P}_\theta^{\otimes n}}(1)$. □

Keep in mind that a μ-dominated family $\mathbb{P}_\Theta$ with likelihood L_θ of $\mathbb{P}_\theta$ w.r.t. μ is Hellinger-differentiable in $\theta_o \in \Theta^\circ$, if there is $\dot\ell_{\theta_o} \in L^2_{\mathbb{P}_{\theta_o}}$, i.e., $\mathbb{P}_{\theta_o}\|\dot\ell_{\theta_o}\|^2 < \infty$, such that for any $h \to 0$ it holds $\|\sqrt{L_{\theta_o+h}} - \sqrt{L_{\theta_o}} - \frac{1}{2}\langle\dot\ell_{\theta_o}, h\rangle\sqrt{L_{\theta_o}}\|_{L^2_\mu} = o(\|h\|)$ (c.f. Remark §2.3.5).

§4.1.9 **Theorem**. *Let Θ be an open set in $\mathbb{R}^k$ and let $\mathbb{P}_\Theta$ be a μ-dominated family of probability measures on a measurable space $(\mathcal{X}, \mathscr{B})$ which is Hellinger-differentiable in $\theta_o \in \Theta$ with score $\dot\ell_{\theta_o}$ satisfying $\mathbb{P}_{\theta_o}\dot\ell_{\theta_o} = 0$, $\mathbb{P}_{\theta_o}\|\dot\ell_{\theta_o}\|^2 < \infty$ and strictly positive definite Fisher information matrix $\mathcal{I}_{\theta_o} = \mathbb{P}_{\theta_o}(\dot\ell_{\theta_o}\dot\ell_{\theta_o}^t)$. Then the sequence of product experiments $(\mathcal{X}^n, \mathscr{B}^{\otimes n}, \mathbb{P}_\Theta^{\otimes n})$ is ULAN in θ_o with localising rate $(\delta_n := 1/\sqrt{n})_{n\in\mathbb{N}}$ and central sequence $(\mathcal{Z}_{\theta_o}^n := \sqrt{n}\mathcal{I}_{\theta_o}^{-1}\overline{\mathbb{P}}_n\dot\ell_{\theta_o})_{n\in\mathbb{N}}$, that is, for any sequence $h_n \to h$ it holds $\log(\mathrm{d}\mathbb{P}^{\otimes n}_{\theta_o+h_n/\sqrt{n}}/\mathrm{d}\mathbb{P}^{\otimes n}_{\theta_o}) = \langle\mathcal{I}_{\theta_o}\mathcal{Z}^n_{\theta_o}, h\rangle - \frac{1}{2}\langle\mathcal{I}_{\theta_o}h, h\rangle + R_{n,h_n}$ where $R_{n,h_n} = o_{\mathbb{P}^{\otimes n}_{\theta_o}}(1)$ and $\sqrt{n}\,\overline{\mathbb{P}}_n\dot\ell_{\theta_o} \xrightarrow{d} \mathfrak{N}(0, \mathcal{I}_{\theta_o})$ under $\mathbb{P}^{\otimes n}_{\theta_o}$.*

Proof of Theorem §4.1.9 is given in the lecture. □

4.2 Hellinger-differentiability

§4.2.1 **Proposition**. *Given a statistical experiment $(\mathcal{X}, \mathscr{B}, \mathbb{P}_\Theta)$ for all $\theta \in \Theta \subset \mathbb{R}^k$ in a neighbourhood of $\theta_o \in \Theta$ let $\mathbb{P}_\theta \ll \mathbb{P}_{\theta_o}$ and let $L_{\theta,\theta_o}(x) := [\mathrm{d}\mathbb{P}_\theta/\mathrm{d}\mathbb{P}_{\theta_o}](x)$, $x \in \mathcal{X}$, be the associated likelihood function w.r.t. $\mathbb{P}_{\theta_o}$. If $\theta \mapsto L_{\theta,\theta_o}(x)$ is $L^2_{\mathbb{P}_{\theta_o}}$-differentiable in θ_o, that is, there is a map $x \mapsto \dot L_{\theta_o,\theta_o}(x)$ in $L^2_{\mathbb{P}_{\theta_o}}$ (i.e., $\mathbb{P}_{\theta_o}\|\dot L_{\theta_o,\theta_o}\|^2 < \infty$), such that*

$$\|L_{\theta,\theta_o} - L_{\theta_o,\theta_o} - \langle\dot L_{\theta_o,\theta_o}, \theta - \theta_o\rangle\|_{L^2_{\mathbb{P}_{\theta_o}}} = o(\|\theta - \theta_o\|) \quad \text{as} \quad \|\theta - \theta_o\| \to 0,$$

then $\mathbb{P}_\Theta$ is Hellinger-differentiable in θ_o with score function $\dot\ell_{\theta_o} = \dot L_{\theta_o,\theta_o}$.

Proof of Proposition §4.2.1 is given in the lecture. □

§4.2.2 **Proposition**. *Let $\mathbb{P}_\Theta$ be a μ-dominated family of probability measures on a measurable space $(\mathcal{X}, \mathscr{B})$ with open $\Theta \subset \mathbb{R}^k$ and associated likelihood function $L_\theta(x) := [\mathrm{d}\mathbb{P}_\theta/\mathrm{d}\mu](x)$, $x \in \mathcal{X}$. Suppose the following conditions hold true:*

(i) *for each $x \in \mathcal{X}$ the map $\theta \mapsto s_\theta(x) := \sqrt{L_\theta(x)}$ is continuously differentiable with derivative $\dot s_\theta(x)$,*

(ii) *$\dot s_\theta$ belongs to L^2_μ (i.e., $\mu\|\dot s_\theta\|^2 < \infty$), and hence $\mathcal{I}_\theta = \mu(\dot s_\theta \dot s_\theta^t)$ is well-defined for all $\theta \in \Theta$,*

(iii) *the map $\theta \mapsto \mathcal{I}_\theta$ is continuous.*

Then $\mathbb{P}_\Theta$ is Hellinger-differentiable with score function $\dot\ell_{\theta_o} = 2\frac{\dot s_\theta}{\sqrt{L_\theta}}\mathbb{1}_{\{L_\theta(x)>0\}}$.

Proof of Proposition §4.2.2 is given in the lecture. □

§4.2.3 **Example**. Consider a statistical location model $(\mathbb{R}, \mathbb{B}, \mathbb{P}_\mathbb{R})$ dominated by the Lebesgue measure λ with likelihood function for each $\theta \in \mathbb{R}$ given by $L_\theta(x) = g(x-\theta)$, $x \in \mathbb{R}$, where g is a strictly positive density. If g is continuously differentiable with derivative $\dot g$ satisfying $\lambda(|\dot g|^2/g) < \infty$ then due to Proposition §4.2.2 the family $\mathbb{P}_\mathbb{R}$ is Hellinger-differentiable with score function $\dot\ell_\theta = -\frac{\dot g(x-\theta)}{g(x-\theta)}$. Indeed, setting $s_\theta(x) := \sqrt{g(x-\theta)}$, we have $\dot s_\theta(x) = \frac{\partial}{\partial\theta}\sqrt{g(x-\theta)} = -\frac{1}{2}\dot g(x-\theta)/\sqrt{g(x-\theta)}$ which is continuous in θ and hence condition (i) is satisfied. Moreover conditions (ii) and (iii) hold true, since by assumption $\mathcal{I}_\theta = \lambda(\dot s_\theta)^2 = \lambda(|\dot g|^2/g) < \infty$ is constant in θ and thus continuous. Thereby, from Proposition §4.2.2 follows the claim with $\dot\ell_\theta = 2\frac{\dot s_\theta}{\sqrt{L_\theta}}\mathbb{1}_{\{L_\theta(x)>0\}} = -\dot g(x-\theta)/g(x-\theta)$. □

4.3 Limit distributions under alternatives

§4.3.1 **Theorem**. *Let $(\Omega_n, \mathscr{A}_n, \mathbb{P}^n_\Theta)_{n\in\mathbb{N}}$ be LAN in $\theta_o \in \Theta \subset \mathbb{R}^k$ with localising rate $(\delta_n)_{n\in\mathbb{N}}$, central sequence $(\mathcal{Z}^n_{\theta_o})_{n\in\mathbb{N}}$ and strictly positive definite Fisher information matrix $\mathcal{I}_{\theta_o}$. Then for any $h, h' \in \mathbb{R}^k$ the following statements hold true:*

(i) $(\mathbb{P}^n_{\theta_o+\delta_n h})_{n\in\mathbb{N}}$ *and* $(\mathbb{P}^n_{\theta_o+\delta_n h'})_{n\in\mathbb{N}}$ *are mutually contiguous, i.e.,* $\mathbb{P}^n_{\theta_o+\delta_n h} \triangleleft\triangleright \mathbb{P}^n_{\theta_o+\delta_n h'}$;

(ii) $\mathcal{Z}^n_{\theta_o} \xrightarrow{d} \mathfrak{N}(h, \mathcal{I}^{-1}_{\theta_o})$ *under* $\mathbb{P}^n_{\theta_o+\delta_n h}$.

If the sequence of statistical experiments is ULAN, then for any $h_n \to h$ and $h'_n \to h'$ in $\mathbb{R}^k$ the following statements hold true:

(i') $(\mathbb{P}^n_{\theta_o+\delta_n h_n})_{n\in\mathbb{N}}$ *and* $(\mathbb{P}^n_{\theta_o+\delta_n h'_n})_{n\in\mathbb{N}}$ *are mutually contiguous, i.e.,* $\mathbb{P}^n_{\theta_o+\delta_n h_n} \lhd\rhd \mathbb{P}^n_{\theta_o+\delta_n h'_n}$;

(ii') $\mathcal{Z}^n_{\theta_o} \xrightarrow{d} \mathfrak{N}(h, \mathcal{I}^{-1}_{\theta_o})$ *under* $\mathbb{P}^n_{\theta_o+\delta_n h_n}$.

Proof of Theorem §4.3.1 is given in the lecture. □

§4.3.2 **Corollary**. *Let $(\mathcal{X}, \mathscr{B}, \mathbb{P}_\Theta)$ be Hellinger-differentiable in θ_o with score function $\dot\ell_{\theta_o}$ such that the assumptions of Theorem §4.1.9 hold true. Given the sequence of product experiments $(\mathcal{X}^n, \mathscr{B}^{\otimes n}, \mathbb{P}^{\otimes n}_\Theta)_{n\in\mathbb{N}}$ let $(\widehat\theta_n)_{n\in\mathbb{N}}$ be a sequence of estimators of θ_o allowing an expansion $\sqrt{n}(\widehat\theta_n - \theta_o) = \sqrt{n}\,\overline{\mathbb{P}}_n\psi_{\theta_o} + o_{\mathbb{P}^{\otimes n}_{\theta_o}}(1)$ for some $\mathbb{R}^k$-valued function ψ_{θ_o} satisfying $\mathbb{P}_{\theta_o}\psi_{\theta_o} = 0$ and $\mathbb{P}_{\theta_o}\|\psi_{\theta_o}\|^2 < \infty$. For each $h \in \mathbb{R}^k$ holds $\sqrt{n}(\widehat\theta_n - \theta_o) \xrightarrow{d} \mathfrak{N}(\mathbb{P}_{\theta_o}(\psi_{\theta_o}\dot\ell^t_{\theta_o})h, \mathbb{P}_{\theta_o}(\psi_{\theta_o}\psi^t_{\theta_o}))$ under $\mathbb{P}^{\otimes n}_{\theta_o+h/\sqrt{n}}$.*

Proof of Corollary §4.3.2. By Theorem §4.1.9 $(\mathcal{X}^n, \mathscr{B}^{\otimes n}, \mathbb{P}^{\otimes n}_\Theta)_{n\in\mathbb{N}}$ is ULAN with localising rate $(\delta_n := 1/\sqrt{n})_{n\in\mathbb{N}}$ and under $\mathbb{P}^{\otimes n}_{\theta_o}$ holds $\sqrt{n}\mathbb{P}_n\dot\ell_{\theta_o} \xrightarrow{d} \mathfrak{N}(0, \mathbb{P}_{\theta_o}\dot\ell_{\theta_o}\dot\ell^t_{\theta_o})$. On the other hand side, we have $\sqrt{n}\,\overline{\mathbb{P}}_n\psi_{\theta_o} \xrightarrow{d} \mathfrak{N}(0, \mathbb{P}_{\theta_o}\psi_{\theta_o}\psi^t_{\theta_o})$. Employing Slutzky's lemma §1.1.7 under $\mathbb{P}^{\otimes n}_{\theta_o}$ it follows

$$\begin{pmatrix} \sqrt{n}(\widehat\theta_n - \theta_o) \\ \log(\mathrm{d}\mathbb{P}^{\otimes n}_{\theta_o+h/\sqrt{n}}/\mathrm{d}\mathbb{P}^{\otimes n}_{\theta_o}) \end{pmatrix} \xrightarrow{d} \mathfrak{N}\left(\begin{pmatrix} 0 \\ -\frac{1}{2}\mathbb{P}_{\theta_o}|\langle\dot\ell_{\theta_o}, h\rangle|^2 \end{pmatrix}, \begin{pmatrix} \mathbb{P}_{\theta_o}\psi_{\theta_o}\psi^t_{\theta_o} & \mathbb{P}_{\theta_o}\psi_{\theta_o}\langle\dot\ell_{\theta_o}, h\rangle \\ \mathbb{P}_{\theta_o}\langle\dot\ell_{\theta_o}, h\rangle\psi^t_{\theta_o} & \mathbb{P}_{\theta_o}|\langle\dot\ell_{\theta_o}, h\rangle|^2 \end{pmatrix} \right).$$

The assertion follows now from le Cam's third lemma as in Example §3.2.6, which completes the proof. □

§4.3.3 **Example** *(§2.3.3 continued)*. Let $\theta_o = \arg\min\{M(\theta), \theta \in \Theta\}$ with $M(\theta) := \mathbb{P}_\theta m_\theta$ for some function $m_\theta \in L_{\mathbb{P}_\theta}$. Considering an M-estimator $\widehat\theta_n := \arg\min\{M_n(\theta), \theta \in \Theta\}$ of θ_o with $M_n(\theta) := \overline{\mathbb{P}}_n m_\theta$ as in Example §2.3.3 we have $\sqrt{n}(\widehat\theta_n - \theta_o) = \sqrt{n}(\mathbb{P}_{\theta_o}\ddot m_{\theta_o})^{-1}\,\overline{\mathbb{P}}_n\dot m_{\theta_o} + o_{\mathbb{P}^{\otimes n}_{\theta_o}}(1)$, that is, $\psi_{\theta_o} := (\mathbb{P}_{\theta_o}\ddot m_{\theta_o})^{-1}\dot m_{\theta_o}$. Under $\mathbb{P}^{\otimes n}_{\theta_o+h/\sqrt{n}}$ it follows then $\sqrt{n}(\widehat\theta_n - \theta_o) \xrightarrow{d} \mathfrak{N}((\mathbb{P}_{\theta_o}\ddot m_{\theta_o})^{-1}\mathbb{P}_{\theta_o}\dot m_{\theta_o}\langle\dot\ell_{\theta_o}, h\rangle, (\mathbb{P}_{\theta_o}\ddot m_{\theta_o})^{-1}(\mathbb{P}_{\theta_o}\dot m_{\theta_o}\dot m^t_{\theta_o})(\mathbb{P}_{\theta_o}\ddot m_{\theta_o})^{-1})$. In the particular case of an MLE $\widehat\theta_n$ as in Example §2.3.4 setting $m_\theta := \ell_\theta = \log(L_\theta)$ and $\mathcal{I}_{\theta_o} := \mathbb{P}_{\theta_o}\dot\ell_{\theta_o}\dot\ell^t_{\theta_o} = -\mathbb{P}_{\theta_o}\ddot\ell_{\theta_o}$ we have $\sqrt{n}(\widehat\theta_n - \theta_o) = \mathcal{I}^{-1}_{\theta_o}\sqrt{n}\,\overline{\mathbb{P}}_n\dot\ell_{\theta_o} + o_{\mathbb{P}^{\otimes n}_{\theta_o}}(1)$ which together with $\mathcal{I}_{\theta_o}h = \mathbb{P}_{\theta_o}\langle\dot\ell_{\theta_o}, h\rangle\dot\ell_{\theta_o}$ under $\mathbb{P}^{\otimes n}_{\theta_o+h/\sqrt{n}}$ implies $\sqrt{n}(\widehat\theta_n - \theta_o) \xrightarrow{d} \mathfrak{N}(h, \mathcal{I}^{-1}_{\theta_o})$. □

§4.3.4 **Remark**. Supposing $\sqrt{n}(\widehat\theta_n - \theta_o) = \sqrt{n}\,\overline{\mathbb{P}}_n\psi_{\theta_o} + o_{\mathbb{P}^{\otimes n}_{\theta_o}}(1)$ let us further assume a transformation $A : \Theta \to \mathbb{R}^p$ that is "smooth", and hence by employing the Delta method §1.1.19,

for instance satisfies $\sqrt{n}(A(\widehat{\theta}_n)-A(\theta_o)) = \dot{A}_{\theta_o}\sqrt{n}\,\overline{\mathbb{P}}_n\psi_{\theta_o} + o_{\mathbb{P}_{\theta_o}^{\otimes n}}(1)$ under $\mathbb{P}_{\theta_o}^{\otimes n}$. Consequently, $\sqrt{n}(A(\widehat{\theta}_n)-A(\theta_o)) \xrightarrow{d} \mathfrak{N}(\dot{A}_{\theta_o}\mathbb{P}_{\theta_o}\psi_{\theta_o}\langle\dot{\ell}_{\theta_o},h\rangle, \dot{A}_{\theta_o}\mathbb{P}_{\theta_o}\psi_{\theta_o}\psi_{\theta_o}^t\dot{A}_{\theta_o}^t)$ under $\mathbb{P}_{\theta_o+h/\sqrt{n}}^{\otimes n}$ and in the special case of an MLE $\sqrt{n}(A(\widehat{\theta}_n)-A(\theta_o)) \xrightarrow{d} \mathfrak{N}(\dot{A}_{\theta_o}h, \dot{A}_{\theta_o}\mathcal{I}_{\theta_o}^{-1}\dot{A}_{\theta_o}^t)$ under $\mathbb{P}_{\theta_o+h/\sqrt{n}}^{\otimes n}$. □

§4.3.5 **Example** *(§2.4.1 continued).* Under the conditions of Corollary §4.3.2 consider the test problem $H_0 : A(\theta) = 0$ against the alternative $H_1 : A(\theta) \neq 0$ for some transformation A satisfying $\sqrt{n}(A(\widehat{\theta}_n)-A(\theta_o)) = \dot{A}_{\theta_o}\sqrt{n}\,\overline{\mathbb{P}}_n\psi_{\theta_o} + o_{\mathbb{P}_{\theta_o}^{\otimes n}}(1)$ under $\mathbb{P}_{\theta_o}^{\otimes n}$. As in section 2.4 let $\widehat{W}_n := nA(\widehat{\theta}_n)^t\widehat{\Sigma}_n^{-1}A(\widehat{\theta}_n)$ where $\widehat{\Sigma}_n$ is under $\mathbb{P}_{\theta_o}^{\otimes n}$ a consistent estimator of $\Sigma := \dot{A}_{\theta_o}\mathbb{P}_{\theta_o}\psi_{\theta_o}\psi_{\theta_o}^t\dot{A}_{\theta_o}^t$, i.e., $\widehat{\Sigma}_n = \Sigma + o_{\mathbb{P}_{\theta_o}^{\otimes n}}(1)$, then a *Wald test* is given by $\varphi_n = \mathbb{1}_{\{\widehat{W}_n > \chi^2_{p,1-\alpha}\}}$. Thereby, under H_o, that is, $A(\theta_o) = 0$, we have $\sqrt{n}A(\widehat{\theta}_n) = \dot{A}_{\theta_o}\sqrt{n}\,\overline{\mathbb{P}}_n\psi_{\theta_o} + o_{\mathbb{P}_{\theta_o}^{\otimes n}}(1)$ and $\widehat{W}_n \xrightarrow{d} W \sim \chi^2_p$ under $\mathbb{P}_{\theta_o}^{\otimes n}$ which in turn implies $\mathbb{P}_{\theta_o}^{\otimes n}(\widehat{W}_n > \chi^2_{p,1-\alpha}) \xrightarrow{n\to\infty} \mathbb{P}(W > \chi^2_{p,1-\alpha}) = \alpha$. In other words, the Wald test is asymptotically a level α test. Let us denote by $\beta_{\varphi_n}(\theta_1) = \mathbb{P}_{\theta_1}^{\otimes n}\varphi_n = \mathbb{P}_{\theta_1}^{\otimes n}(\varphi_n = 1) = \mathbb{P}_{\theta_1}^{\otimes n}(\widehat{W}_n > \chi^2_{p,1-\alpha})$ the power function of the Wald-test φ_n evaluated at θ_1 with $A(\theta_1) \neq 0$. In the sequel we consider local alternatives of the form $\theta_o + h/\sqrt{n}$ and thus we are interested in $\beta_{\varphi_n}(\theta_o + h/\sqrt{n}) = \mathbb{P}_{\theta_o+h/\sqrt{n}}^{\otimes n}(\widehat{W}_n > \chi^2_{p,1-\alpha})$. Obviously, under $\mathbb{P}_{\theta_o+h/\sqrt{n}}^{\otimes n}$ we have $\sqrt{n}A(\widehat{\theta}_n) \xrightarrow{d} \mathfrak{N}(\dot{A}_{\theta_o}\mathbb{P}_{\theta_o}\psi_{\theta_o}\langle\dot{\ell}_{\theta_o},h\rangle, \Sigma)$, or equivalently, $\Sigma^{-1/2}\sqrt{n}A(\widehat{\theta}_n) \xrightarrow{d} \mathfrak{N}(\Sigma^{-1/2}\dot{A}_{\theta_o}\mathbb{P}_{\theta_o}\psi_{\theta_o}\langle\dot{\ell}_{\theta_o},h\rangle, \mathrm{Id})$, and hence, $nA(\widehat{\theta}_n)^t\Sigma^{-1}A(\widehat{\theta}_n) \xrightarrow{d} W_h \sim \chi^2_p(\|\Sigma^{-1/2}\dot{A}_{\theta_o}\mathbb{P}_{\theta_o}\psi_{\theta_o}\langle\dot{\ell}_{\theta_o},h\rangle\|^2)$ where $\chi^2_p(a)$ denotes a non-central χ^2-distribution. Moreover, $\widehat{W}_n - nA(\widehat{\theta}_n)^t\Sigma^{-1}A(\widehat{\theta}_n) = o_{\mathbb{P}_{\theta_o}^{\otimes n}}(1)$ and thus $\widehat{W}_n - nA(\widehat{\theta}_n)^t\Sigma^{-1}A(\widehat{\theta}_n) = o_{\mathbb{P}_{\theta_o+h/\sqrt{n}}^{\otimes n}}(1)$ due to Lemma §3.2.3 by employing that $\mathbb{P}_{\theta_o}^{\otimes n}$ and $\mathbb{P}_{\theta_o+h/\sqrt{n}}^{\otimes n}$ are mutually contiguous. Consequently, $\widehat{W}_n \xrightarrow{d} W_h$ under $\mathbb{P}_{\theta_o+h/\sqrt{n}}^{\otimes n}$ and thus $\beta_{\widehat{W}_n}(\theta_o + h/\sqrt{n}) \xrightarrow{n\to\infty} \mathbb{P}(W_h \geqslant \chi^2_{p,1-\alpha})$. Note that in the particular case of an MLE we have $W_h \sim \chi^2_p(h^t\dot{A}_{\theta_o}^t(\dot{A}_{\theta_o}\mathcal{I}_{\theta_o}^{-1}\dot{A}_{\theta_o}^t)^{-1}\dot{A}_{\theta_o}h)$. □

4.4 Asymptotic power function

Let $(\Omega_n, \mathscr{A}_n, \mathbb{P}_\Theta^n)$ be LAN in $\theta_o \in \Theta \subset \mathbb{R}^p$ with localising sequence $(\delta_n)_{n\in\mathbb{N}}$, central sequence $(\mathcal{Z}_{\theta_o}^n)_{n\in\mathbb{N}}$ and strictly positive definite Fisher information matrix $\mathcal{I}_{\theta_o}$, that is, $\widetilde{\mathcal{Z}}_{\theta_o}^n := \mathcal{I}_{\theta_o}\mathcal{Z}_{\theta_o}^n \xrightarrow{d} \mathfrak{N}(0, \mathcal{I}_{\theta_o})$ under $\mathbb{P}_{\theta_o}^{\otimes n}$ and $\Lambda_n := \log(\mathrm{d}\mathbb{P}_{\theta_o+\delta_n h}^n/\mathrm{d}\mathbb{P}_{\theta_o}^n) = \langle\mathcal{I}_{\theta_o}\mathcal{Z}_{\theta_o}^n, h\rangle - \frac{1}{2}\langle\mathcal{I}_{\theta_o}h, h\rangle + o_{\mathbb{P}_{\theta_o}^{\otimes n}}(1)$. Denoting $\sigma_h^2 := \langle\mathcal{I}_{\theta_o}h, h\rangle$ from $\widetilde{\mathcal{Z}}_{\theta_o}^n \xrightarrow{d} \mathfrak{N}(0, \mathcal{I}_{\theta_o})$ under $\mathbb{P}_{\theta_o}^n$ and $\widetilde{\mathcal{Z}}_{\theta_o}^n \xrightarrow{d} \mathfrak{N}(h, \mathcal{I}_{\theta_o})$ under $\mathbb{P}_{\theta_o+\delta_n h}^n$ it follows $\Lambda_n \xrightarrow{d} \mathcal{Z}_h^o \sim \mathfrak{N}(-\frac{1}{2}\sigma_h^2, \sigma_h^2)$ under $\mathbb{P}_{\theta_o}^n$ and $\Lambda_n \xrightarrow{d} \mathcal{Z}_h^1 \sim \mathfrak{N}(\frac{1}{2}\sigma_h^2, \sigma_h^2)$ under $\mathbb{P}_{\theta_o+\delta_n h}^n$.

§4.4.1 **Example** *(Neyman-Pearson test).* Consider the elementary test problem $H_o : \mathbb{P}_{\theta_o}^n$ against $H_1 : \mathbb{P}_{\theta_1}^n$. In this situation the most powerful level-α test is of Neyman-Pearson form, i.e., $\varphi_n^\star = \mathbb{1}_{\{\Lambda_n > c_n\}}$ if $\mathbb{P}_{\theta_o}^n(\varphi_n^\star = 1) = \alpha$. Let us denote by $\beta_{\varphi_n^\star}(\theta_1) = \mathbb{P}_{\theta_1}^n\varphi_n^\star = \mathbb{P}_{\theta_1}^n(\varphi_n^\star = 1) = \mathbb{P}_{\theta_1}^n(\Lambda_n > c_n)$ the power function of $\varphi_n^\star$ evaluated at θ_1. Keep in mind that the value $\beta_{\varphi_n^\star}(\theta_1)$ equals the maximal size of the power in the class of all level-α tests, i.e., for any level-α test φ_n holds $\beta_{\varphi_n}(\theta_1) \leqslant \beta_{\varphi_n^\star}(\theta_1)$. In particular, in case of local alternatives, i.e., $H_o : \mathbb{P}_{\theta_o}^n$ against $H_1 : \mathbb{P}_{\theta_o+\delta_n h}^n$, exploiting the LAN assumption we have $\alpha = \mathbb{P}_{\theta_o}^n\varphi_n^\star = \mathbb{P}_{\theta_o}^n(\Lambda_n > c_n) \xrightarrow{n\to\infty} \mathbb{P}(\mathcal{Z}_h^o > c_{1-\alpha}) = \alpha$ which implies $c_n \xrightarrow{n\to\infty} c_{1-\alpha}$, and in addition $\beta_{\varphi_n^\star}(\theta_o + \delta_n h) = \mathbb{P}_{\theta_o+\delta_n h}^n\varphi_n^\star = \mathbb{P}_{\theta_o+\delta_n h}^n(\Lambda_n > c_n) \xrightarrow{n\to\infty} \mathbb{P}(\mathcal{Z}_h^1 > c_{1-\alpha}) =: \beta_{\varphi^\star}(h)$. □

§4.4.2 **Example**. In a Gaussian location model, i.e. $Y \bigcirc\!\!\!\!\sim \mathfrak{N}(\mathbb{R}^p, \mathcal{I}_{\theta_o}^{-1})$, consider the test problem $H_o : \mathfrak{N}(0, \mathcal{I}_{\theta_o}^{-1})$ against the alternative $H_1 : \mathfrak{N}(h, \mathcal{I}_{\theta_o}^{-1})$. It is easily seen that in this situation the log of the likelihood-ratio $\Lambda_h := \log(\mathrm{d}\mathfrak{N}(h, \mathcal{I}_{\theta_o}^{-1})/\mathrm{d}\mathfrak{N}(0, \mathcal{I}_{\theta_o}^{-1}))$ equals $\langle \mathcal{I}_{\theta_o} Y, h\rangle - \frac{1}{2}\sigma_h^2$ and thus $\Lambda_h = \mathcal{Z}_h^o \sim \mathfrak{N}(-\frac{1}{2}\sigma_h^2, \sigma_h^2)$ under the hypothesis $\mathfrak{N}(0, \mathcal{I}_{\theta_o}^{-1})$ and $\Lambda_h = \mathcal{Z}_h^1 \sim \mathfrak{N}(\frac{1}{2}\sigma_h^2, \sigma_h^2)$ under the alternative $\mathfrak{N}(h, \mathcal{I}_{\theta_o}^{-1})$. Moreover, keeping in mind that $\mathbb{P}(\mathcal{Z}_h^o > c_{1-\alpha}) = \alpha$ the most powerful level-α test $\varphi^\star$ has again Neyman-Pearson form, i.e., $\varphi^\star = \mathbb{1}_{\{\Lambda_h > c_{1-\alpha}\}}$, and its power is given by $\mathbb{P}_h(\Lambda_h > c_{1-\alpha}) = \mathbb{P}(\mathcal{Z}_h^1 > c_{1-\alpha}) = \beta_{\varphi^\star}(h)$ which again is maximal. □

§4.4.3 **Remark**. In a statistical LAN experiment the power function $\beta_{\varphi_n^\star}$ of a Neyman-Pearson test $\varphi_n^\star$ for $H_o : \mathbb{P}_{\theta_o}^n$ against $H_1 : \mathbb{P}_{\theta_1}^n$ converges point-wise as $n \to \infty$ to the power function $\beta_{\varphi^\star}$ of a Neyman-Pearson test $\varphi^\star$ for $H_o : \mathfrak{N}(0, \mathcal{I}_{\theta_o}^{-1})$ against $H_1 : \mathfrak{N}(h, \mathcal{I}_{\theta_o}^{-1})$. □

§4.4.4 **Theorem**. *Let $\Theta \subset \mathbb{R}$. Consider the one-sided test problem $H_o : \theta \leqslant \theta_o$ against $H_1 : \theta > \theta_o$. Suppose that $(\Omega_n, \mathscr{A}_n, \mathbb{P}_\Theta^n)$ is LAN in $\theta_o \in \Theta$ with localising sequence $(\delta_n)_{n\in\mathbb{N}}$, central sequence $(\mathcal{Z}_{\theta_o}^n)_{n\in\mathbb{N}}$ and strictly positive Fisher information $\mathcal{I}_{\theta_o} > 0$.*

(i) *Given any test statistic T_n satisfying $(T_n, \mathcal{I}_{\theta_o}\mathcal{Z}_{\theta_o}^n) \xrightarrow{d} \mathfrak{N}(0, \Sigma)$ with $\Sigma = \big((\sigma^2, \rho)^t, (\rho, \mathcal{I}_{\theta_o})^t\big)$ consider a randomised test $\varphi_n := \mathbb{1}_{\{T_n > c_n\}} + \gamma_n \mathbb{1}_{\{T_n = c_n\}}$ with $\gamma_n \in [0,1]$ and $c_n \in \mathbb{R}$ such that $\beta_{\varphi_n}(\theta_o) := \mathbb{P}_{\theta_o}^n \varphi_n = \mathbb{P}_{\theta_o}^n(T_n > c_n) + \gamma_n \mathbb{P}_{\theta_o}^n(T_n = c_n) = \alpha_n \overset{n\to\infty}{\longrightarrow} \alpha$. Choosing $z_{1-\alpha}$ such that $\mathbb{F}_{\mathfrak{N}(0,1)}(z_{1-\alpha}) := \mathbb{P}(Z \leqslant z_{1-\alpha}) = 1-\alpha$ with $Z \sim \mathfrak{N}(0,1)$ we have $\beta_{\varphi_n}(\theta_o + \delta_n h) = \mathbb{P}_{\theta_o+\delta_n h}^n \varphi_n \overset{n\to\infty}{\longrightarrow} \mathbb{P}(Z > z_{1-\alpha} - h\rho/\sigma) = 1 - \mathbb{F}_{\mathfrak{N}(0,1)}(z_{1-\alpha} - h\rho/\sigma) = \mathbb{F}_{\mathfrak{N}(0,1)}(-z_{1-\alpha} + h\rho/\sigma)$.*

(ii) *In the special case $T_n = \mathcal{I}_{\theta_o}\mathcal{Z}_{\theta_o}^n$ choosing $\gamma_n = 0$ and $c_n = z_{1-\alpha}\sqrt{\mathcal{I}_{\theta_o}}$, i.e., $\varphi_n^\star = \mathbb{1}_{\{\mathcal{I}_{\theta_o}\mathcal{Z}_{\theta_o}^n > z_{1-\alpha}\sqrt{\mathcal{I}_{\theta_o}}\}}$ we have $\beta_{\varphi_n^\star}(\theta_o) = \mathbb{P}_{\theta_o}^n \varphi_n^\star = \mathbb{P}_{\theta_o}^n(\sqrt{\mathcal{I}_{\theta_o}}\mathcal{Z}_{\theta_o}^n > z_{1-\alpha}) \overset{n\to\infty}{\longrightarrow} \mathbb{P}(Z > z_{1-\alpha}) = \alpha$ and $\beta_{\varphi_n^\star}(\theta_o + \delta_n h) = \mathbb{P}_{\theta_o+\delta_n h}^n \varphi_n^\star \overset{n\to\infty}{\longrightarrow} \mathbb{P}(Z > z_{1-\alpha} - h\sqrt{\mathcal{I}_{\theta_o}}) = \mathbb{F}_{\mathfrak{N}(0,1)}(-z_{1-\alpha} + h\sqrt{\mathcal{I}_{\theta_o}})$.*

Proof of Theorem §4.4.4 is given in the lecture. □

§4.4.5 **Remark**. (i) By using Theorem §3.2.5 directly it might still be possible to calculate an asymptotic power of a test if $\Lambda_n := \log(\mathrm{d}\mathbb{P}_{\theta_o+\delta_n h}^n / \mathrm{d}\mathbb{P}_{\theta_o}^n) \xrightarrow{d} Q$ under $\mathbb{P}_{\theta_o}^n$ where Q is not necessarily $\mathfrak{N}(0,1)$ distributed.

(ii) Keeping in mind that $\rho^2 = |\operatorname{Cov}(T_n, \mathcal{I}_{\theta_o}\mathcal{Z}_{\theta_o}^n)|^2 \leqslant \operatorname{Var}(T_n)\operatorname{Var}(\mathcal{I}_{\theta_o}\mathcal{Z}_{\theta_o}^n) = \sigma^2 \mathcal{I}_{\theta_o}$ the test $\varphi_n^\star$ given in Theorem §4.4.4 (ii) maximises the asymptotic power when considering only tests T_n as given in part (i) of Theorem §4.4.4. □

§4.4.6 **Theorem**. *Let the assumptions of Theorem §4.4.4 be satisfied. For any test φ_n of the one-sided test problem $H_o : \theta \leqslant \theta_o$ against $H_1 : \theta > \theta_o$ satisfying $\beta_{\varphi_n}(\theta_o) := \mathbb{P}_{\theta_o}^n \varphi_n = \alpha_n \overset{n\to\infty}{\longrightarrow} \alpha$ it holds*

(i) *for any $h > 0$ we have $\limsup_{n\to\infty} \beta_{\varphi_n}(\theta_o + \delta_n h) \leqslant \mathbb{F}_{\mathfrak{N}(0,1)}(-z_{1-\alpha} + h\sqrt{\mathcal{I}_{\theta_o}})$;*

(ii) *for any $h < 0$ we have $\liminf_{n\to\infty} \beta_{\varphi_n}(\theta_o + \delta_n h) \geqslant \mathbb{F}_{\mathfrak{N}(0,1)}(-z_{1-\alpha} + h\sqrt{\mathcal{I}_{\theta_o}})$.*

Proof of Theorem §4.4.6 is given in the lecture. □

§4.4.7 **Remark**. Keeping in mind Theorem §4.4.6 we call the test (sequence) $(\varphi_n^\star)_{n\in\mathbb{N}}$ given in Theorem §4.4.4 (ii) asymptotically uniformly most powerful level-α test (sequence) in the class

of all asymptotic level-α test (sequences). Its asymptotic power function equals $\mathbb{F}_{\mathfrak{N}(0,1)}(-z_{1-\alpha}+h\sqrt{\mathcal{I}_{\theta_o}})$ which is the power function of the uniformly most powerful test of $H_o : h \leqslant 0$ against $H_1 : h > 0$ in the limit Gaussian location experiment $(\mathbb{R}, \mathscr{B}, \mathfrak{N}(\mathbb{R}, \mathcal{I}_{\theta_o}^{-1}))$. □

4.5 Asymptotic relative efficiency

Let $(\Omega_n, \mathscr{A}_n, \mathbb{P}_\Theta^n)_{n\in\mathbb{N}}$ be LAN with localising rate $(\delta_n := 1/\sqrt{n})_{n\in\mathbb{N}}$. Consider a test φ_n^a satisfying the conditions of Theorem §4.4.4 (i) and hence, admitting an asymptotic power function such that $\beta_{\varphi_n^a}(\theta_o + h/\sqrt{n}) \overset{n\to\infty}{\longrightarrow} \mathbb{F}_{\mathfrak{N}(0,1)}(-z_{1-\alpha} + h\rho_a/\sigma_a)$. Thereby, choosing $\eta = h/\sqrt{n}$ the approximation $\beta_{\varphi_n^a}(\theta_o + \eta) \approx \mathbb{F}_{\mathfrak{N}(0,1)}(-z_{1-\alpha} + \eta\sqrt{n}\rho_a/\sigma_a)$ is reasonable. In analogy, if φ_n^b is another test satisfying the conditions of Theorem §4.4.4 (i) we have $\beta_{\varphi_n^b}(\theta_o + \eta) \approx \mathbb{F}_{\mathfrak{N}(0,1)}(-z_{1-\alpha} + \eta\sqrt{n}\rho_b/\sigma_b)$. Roughly speaking, this means, that at $\theta_o + \eta$ the power of the test $\varphi_{n_a}^a$ and $\varphi_{n_b}^b$ with sample size n_a and n_b, respectively, is approximately equal if $n_a\rho_a^2/\sigma_a^2 = n_b\rho_b^2/\sigma_b^2$. The quantity $\mathrm{are}(\varphi_{n_a}^a, \varphi_{n_b}^b) = (n_a/n_b) = (\rho_b^2\sigma_a^2)/(\rho_a^2\sigma_b^2)$ is called *asymptotic relative efficiency*. Meaning, that a sample of size $n_a = \mathrm{are}(\varphi_{n_a}^a, \varphi_{n_b}^b)\, n_b$ is needed for the test $\varphi_{n_a}^a$ to attain the same asymptotic power as the test $\varphi_{n_b}^b$ with sample size n_b. More precisely, if $n_a = \mathrm{are}(\varphi_{n_a}^a, \varphi_{n_b}^b)n_b$ and $n_b \to \infty$ then $\beta_{\varphi_{n_b}^b}(\theta_o + h/\sqrt{n_a}) \overset{n\to\infty}{\longrightarrow} \mathbb{F}_{\mathfrak{N}(0,1)}(-z_{1-\alpha} + h\rho_a/\sigma_a)$ and $\beta_{\varphi_{n_a}^a}(\theta_o + h/\sqrt{n_a}) \overset{n\to\infty}{\longrightarrow} \mathbb{F}_{\mathfrak{N}(0,1)}(-z_{1-\alpha} + h\rho_a/\sigma_a)$. Comparing with $\varphi_n^\star$ as in Theorem §4.4.4 (ii) allows to have a notion of *asymptotic absolute efficiency*.

4.6 Rank tests

Consider a sample $X_1, \ldots, X_n$ of independent and not necessarily identically distributed real-valued r.v.'s. Denote by $\mathcal{S}_n$ the set of all permutations of the set $[\![1, n]\!]$. Given a vector $(x_1, \ldots, x_n)$ let $(x_{s_1}, \ldots, x_{s_n})$ denote its arrangement according to the permutation $s \in \mathcal{S}_n$. More generally, $(X_{S_1}, \ldots, X_{S_n})$ denotes the arrangement of the r.v. $(X_1, \ldots, X_n)$ according to a $\mathcal{S}_n$-valued r.v. (random permutation) S. Precisely, given $X_1, \ldots, X_n$ and S defined on a common probability space $(\Omega, \mathscr{A}, \mathbb{P})$ for each $\omega \in \Omega$ letting $(x_1, \ldots, x_n) := (X_1(\omega), \ldots, X_n(\omega))$ and $s := S(\omega)$ we set $(X_{S_1}, \ldots, X_{S_n})(\omega) := (x_{s_1}, \ldots, x_{s_n})$.

§4.6.1 **Definition**. A $\mathcal{S}_n$-valued r.v. (random permutation) R is called a *rank vector* of a $\mathbb{R}^n$-valued r.v. $(X_1, \ldots, X_n)$, if $X_i = X_{R_i}$, $i \in [\![1, n]\!]$, and $X_{R_1} \leqslant X_{R_2} \leqslant \cdots \leqslant X_{R_n}$. For each $i \in [\![1, n]\!]$, the component R_i is called the *rank* of the i-th component X_i. □

Here and subsequently we assume that the law of each component of $(X_1, \ldots, X_n)$ has a density with respect to the Lebesgue measure, i.e., the associated cumulative distribution function (c.d.f.) is continuous, and we say the law is continuous, for short. Consequently, with probability one all components of $(X_1, \ldots, X_n)$ differ and thus $X_{R_1} < X_{R_2} < \cdots < X_{R_n}$. Thereby, the rank vector $(R_1, \ldots, R_n)$ is uniquely determined by $R_i = \sum_{j=1}^n \mathbb{1}_{\{X_j \leqslant X_i\}}$, $i \in [\![1, n]\!]$. In particular if $\overline{\mathbb{F}}_n(x) := \overline{\mathbb{P}}_n \mathbb{1}_{(-\infty,x]}$, $x \in \mathbb{R}$, denotes the empirical c.d.f., then it is easily seen that $i = n\overline{\mathbb{F}}_n(X_{R_i})$ and $R_i = n\overline{\mathbb{F}}_n(X_i)$ for each $i \in [\![1, n]\!]$.

§4.6.2 **Lemma** (Preliminary results). *Let $X = (X_1, \ldots, X_n)$ and the associated rank vector $R = (R_1, \ldots, R_n)$ be r.v.'s on a common probability space $(\Omega, \mathscr{A}, \mathbb{P})$. If $X_1, \ldots, X_n$ are i.i.d. real-valued r.v.'s with common continuous c.d.f. $\mathbb{F}$ and common density f w.r.t. the Lebesgue measure, then the following statements hold true:*

(i) *The $\mathcal{S}_n$-valued r.v. R has an uniform (Laplace) distribution, i.e., for all $s \in \mathcal{S}_n$ it holds $\mathbb{P}^R(s) = \mathbb{P}(R = s) = \frac{1}{n!}$;*

(ii) *The rank vector R and the ordered vector $(X_{R_1}, \dots, X_{R_n})$ are independent;*

(iii) *The ordered vector $(X_{R_1}, \dots, X_{R_n})$ admits a density w.r.t. the Lebesgue measure given by $n! \mathbb{1}_B(x) \prod_{i=1}^n f(x_i)$ with $B := \{(x_1, \dots, x_n) \in \mathbb{R}, x_1 < \dots < x_n\}$;*

(iv) *For each $i \in [\![1, n]\!]$ the R_i-th component X_{R_i} of the ordered vector admits a density w.r.t. the Lebesgue measure given by $i\binom{n}{i}|\mathbb{F}(x)|^{i-1}|1 - \mathbb{F}(x)|^{n-i} f(x)$;*

(v) *For each r.v. $T \in \mathcal{L}^1_{\mathbb{P}^X}$ holds $\mathbb{E}[T(X_1, \dots, X_n)|R = r] = \mathbb{E}[T(X_{r_1}, \dots, X_{r_n})]$ $\mathbb{P}$-a.s..*

Proof of Lemma §4.6.2 (i), (iv) and (v) is given in the lecture, while (ii) and (iii) is left as an exercise. □

§4.6.3 **Definition**. Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on $(\mathbb{R}, \mathscr{B})$. We say $\mathbb{P}$ is *stochastically smaller* than $\mathbb{Q}$, or $\mathbb{P} \preceq \mathbb{Q}$ for short, if $\mathbb{P}((c, \infty)) \leqslant \mathbb{Q}((c, \infty))$ for all $c \in \mathbb{R}$. If in addition $\mathbb{P} \neq \mathbb{Q}$, then we write $\mathbb{P} \prec \mathbb{Q}$. □

§4.6.4 **Remark**. Roughly speaking, $\mathbb{P} \preceq \mathbb{Q}$ says that realisations of $\mathbb{Q}$ are typically larger than realisations of $\mathbb{P}$. □

§4.6.5 **Example**. Consider on $(\mathbb{R}, \mathscr{B})$ two Gaussian distributions with common variance σ^2 and individual mean μ and μ', respectively, i.e., $\mathfrak{N}(\mu, \sigma^2)$ and $\mathfrak{N}(\mu', \sigma^2)$. Obviously, $\mathfrak{N}(\mu, \sigma^2) \prec \mathfrak{N}(\mu', \sigma^2)$ if and only if $\mu \lneqq \mu'$. More generally, given a location family $\mathbb{P}_{\mathbb{R}}$ as introduced in Example §4.2.3 with likelihood function for each $\theta \in \mathbb{R}$ given by $L_\theta(x) = g(x - \theta)$, $x \in \mathbb{R}$ for some strictly positive Lebesgue-density g on $\mathbb{R}$, then $\mathbb{P}_\theta \prec \mathbb{P}_{\theta'}$ holds if and only if $\theta \lneqq \theta'$. □

Aim: test the hypothesis $H_o : \mathbb{P} = \mathbb{Q}$ against the alternative $H_1 : \mathbb{P} \prec \mathbb{Q}$. Loosely speaking, this means, that we aim to reject the null hypothesis if realisations of $\mathbb{P}$ are *significantly* smaller than realisation of $\mathbb{Q}$. Therefore, consider a sample of $n = k + l$ independent real-valued r.v.'s $X_1, \dots, X_n$ where the first k r.v.'s have a common distribution $\mathbb{P}$ and the last l r.v.'s are distributed according to a common distribution $\mathbb{Q}$. Keep in mind that we want to reject if realisations of the common distribution $\mathbb{P}$ of the first k r.v.'s are *significantly* smaller than realisations of the common distribution of the last l r.v.'s. Given the rank vector $R = (R_1, \dots, R_n)$ associated to the pooled sample $(X_1, \dots, X_n)$ it seems thus reasonable to reject the hypothesis if the sum of ranks within the first group of k r.v.'s, i.e., $W_{\mathbb{P}} := \sum_{i=1}^k R_i$, takes *sufficiently* smaller values then the sum of ranks within the second group of l r.v.'s, i.e., $W_{\mathbb{Q}} := \sum_{i=k+1}^{k+l} R_i$ where obviously $W_{\mathbb{P}} + W_{\mathbb{Q}} = \sum_{i=1}^n R_i = \sum_{i=1}^n i = \frac{n(n+1)}{2}$.

§4.6.6 **Lemma**. *Defining $U_{kl} := \sum_{i=1}^k \sum_{j=k+1}^{k+l} \mathbb{1}_{\{X_i > X_j\}}$ it holds $W_{\mathbb{P}} = U_{kl} + \frac{k(k+1)}{2}$ and analogously $W_{\mathbb{Q}} = kl - U_{kl} + \frac{l(l+1)}{2}$.*

Proof of Lemma §4.6.6 is given in the lecture. □

Keeping the last lemma in mind, we use the test statistic $W_{\mathbb{P}}$ or equivalently U_{kl} to reject the hypothesis $H_0 : \mathbb{P} = \mathbb{Q}$ against the alternative $H_1 : \mathbb{P} \prec \mathbb{Q}$, if $U_{kl} < c$ or equivalently

$W_{\mathbb{P}} < c + \frac{k(k+1)}{2}$ for a certain threshold $0 < c \leqslant kl$. The test is called (one-sided) Mann-Whitney U-test or Wilcoxon two-sample rank sum test[1]. The critical value has to be chosen according to a pre-specified level α which under the null hypothesis necessitates the knowledge of the distribution of U_{kl} or an asymptotic approximation. Interestingly the next proposition shows that under the null hypothesis the distribution of U_{kl} is *distribution free* in the following sense: If $\mathbb{P} = \mathbb{Q}$ and $\mathbb{P}$ is continuous, then the distribution of U_{kl} is determined and it is independent of the underlying distribution $\mathbb{P}$.

§4.6.7 **Proposition**. *For every continuous $\mathbb{P}$ and $m \in [\![0, kl]\!]$ it holds $\mathbb{P}^{\otimes(k+l)}(U_{kl} = m) = N(m;k,l)/\binom{k+l}{m}$ where $N(m;k,l)$ denotes the number of all partitions $\sum_{i=1}^{k} m_i = m$ of m in k increasing ordered numbers $m_1 \leqslant m_2 \leqslant \cdots \leqslant m_k$ taking from the set $[\![0, l]\!]$. In particular, it holds $\mathbb{P}^{\otimes k+l}(U_{kl} = m) = \mathbb{P}^{\otimes(k+l)}(U_{kl} = kl - m)$.*

Proof of Proposition §4.6.7 is given in the lecture. □

For small values of m the partition number $N(m;k,l)$ can be calculated by combinatorical means and there exists tables gathering certain quantiles of the U_{kl}-distribution. However, for large values of m the exact calculation of quantiles of the U_{kl}-distribution may be avoided by using an appropriate asymptotic approximation. In the sequel we let k and l and thus $n = k + l$ tend to infinity, which formally means that we consider sequences $(k_n)_{n\in\mathbb{N}}$ and $(l_n)_{n\in\mathbb{N}}$ satisfying $k_n + l_n = n$ for any $n \in \mathbb{N}$. Here and subsequently we assume that $k_n/n \overset{n\to\infty}{\longrightarrow} \gamma \in (0,1)$ and hence $l_n/n \overset{n\to\infty}{\longrightarrow} 1-\gamma$. For sake of presentation, however, we drop the additional index n and write shortly $n = k + l$ with $k/n \overset{n\to\infty}{\longrightarrow} \gamma$ and hence $l/n \overset{n\to\infty}{\longrightarrow} 1-\gamma$.

§4.6.8 **Theorem**. *Let $X_1, X_2, \ldots$ be i.i.d. real-valued r.v.'s with common distribution $\mathbb{P}$ and continuous c.d.f. $\mathbb{F}$. Consider $U_{kl} := \sum_{i=1}^{k}\sum_{j=k+1}^{k+l} \mathbb{1}_{\{X_i > X_j\}}$ and define*

$$T_{kl} := l\sum_{i=1}^{k}\mathbb{F}(X_i) - k\sum_{i=k+1}^{k+l}\mathbb{F}(X_i) = l\sum_{i=1}^{k}(\mathbb{F}(X_i) - 1/2) - k\sum_{i=k+1}^{k+l}(\mathbb{F}(X_i) - 1/2).$$

Setting $n = k + l$, $v_{kl} := kl(n+1)/12$, $T^\star_{kl} := T_{kl}/\sqrt{v_{kl}}$ and $U^\star_{kl} := (U_{kl} - kl/2)/\sqrt{v_{kl}}$ if $k/n \to \gamma \in (0,1)$ then $U^\star_{kl} - T^\star_{kl} = o_{\mathbb{P}^{\otimes n}}(1)$, $T^\star_{kl} \xrightarrow{d} \mathfrak{N}(0,1)$ and thus $U^\star_{kl} \xrightarrow{d} \mathfrak{N}(0,1)$ as $n \to \infty$.

Proof of Theorem §4.6.8 is given in the lecture. □

Keeping in mind the last assertion given a sample of $n = k+l$ independent r.v.'s $X_1, \ldots, X_n$ with $X_1, \ldots, X_k \overset{i.i.d.}{\sim} \mathbb{P}$ and $X_{k+1}, \ldots, X_{k+l} \overset{i.i.d.}{\sim} \mathbb{Q}$, consider a test which rejects the null hypothesis $H_o : \mathbb{P} = \mathbb{Q}$ against the alternative $\mathbb{P} \prec \mathbb{Q}$, if $U_{kl} < kl/2 + z_\alpha\sqrt{v_{kl}}$ where $\mathbb{F}_{\mathfrak{N}(0,1)}(z_\alpha) = \alpha$. Note that, it is asymptotically a level-α test due to Theorem §4.6.8 since under the null hypothesis $\mathbb{P}^{\otimes n}(U_{kl} < kl/2 + z_\alpha\sqrt{v_{kl}}) \overset{n\to\infty}{\longrightarrow} \mathbb{F}_{\mathfrak{N}(0,1)}(z_\alpha) = \alpha$ for $k/n \overset{n\to\infty}{\longrightarrow} \gamma \in (0,1)$. Note that the null hypothesis $H_o : \mathbb{P} = \mathbb{Q}$ against the alternative $\mathbb{P} \succ \mathbb{Q}$ is analogously rejected if $U_{kl} > kl/2 + z_{1-\alpha}\sqrt{v_{kl}}$. Next we study the (asymptotic) size of the power of the rank test under

[1] The version based on $W_{\mathbb{P}}$ has been proposed by Wilcoxon [1945], while the U_{kl}-version has been independently be introduced by Mann and Whitney [1947].

local alternatives where we use that under the assumptions of Theorem §4.6.8 it holds

$$U^\star_{kl} = (U_{kl} - kl/2)/\sqrt{v_{kl}} = \sqrt{\tfrac{l}{n}}\tfrac{1}{\sqrt{k}} \sum_{i=1}^{k} \tfrac{\mathbb{F}(X_i)-1/2}{\sqrt{1/12}} - \sqrt{\tfrac{k}{n}}\tfrac{1}{\sqrt{l}} \sum_{i=k+1}^{k+l} \tfrac{\mathbb{F}(X_i)-1/2}{\sqrt{1/12}} + o_{\mathbb{P}^{\otimes n}}(1)$$
$$= \sqrt{1-\gamma}\,\sqrt{k}\,\overline{\mathbb{P}}_k g - \sqrt{\gamma}\,\sqrt{l}\,\overline{\mathbb{Q}}_l g + o_{\mathbb{P}^{\otimes n}}(1) \quad (4.1)$$

setting $g := \sqrt{12}(\mathbb{F} - 1/2)$, $\overline{\mathbb{P}}_k g := \frac{1}{k}\sum_{i=1}^{k} g(X_i)$ and $\overline{\mathbb{Q}}_l g := \frac{1}{l}\sum_{i=k+1}^{k+l} g(X_i)$ where $\overline{\mathbb{P}}_k g$ and $\overline{\mathbb{Q}}_l g$ are independent, $\mathbb{P}g = 0$, and $\mathbb{P}g^2 = 1$ by construction.

4.7 Asymptotic power of rank tests

Considering the test of the hypothesis $H_o : \mathbb{P} = \mathbb{Q}$ against the alternative $H_1 : \mathbb{P} \succ \mathbb{Q}$ we restrict our attention to the special case that $\mathbb{P}$ and $\mathbb{Q}$ belong to a location family $\mathbb{P}_\mathbb{R}$ as introduced in Example §4.2.3. Precisely, we assume that the family $\mathbb{P}_\mathbb{R}$ of probability measures on $(\mathbb{R}, \mathscr{B})$ is dominated by the Lebegues measure. For each $\theta \in \mathbb{R}$, $\mathbb{P}_\theta$ admits a likelihood function given by $L_\theta(x) = q(x-\theta)$, $x \in \mathbb{R}$, where q is a continuous and strictly positive density on $\mathbb{R}$. Recall that in this context $\mathbb{P}_\theta \prec \mathbb{P}_{\theta'}$ holds if and only if $\theta \lneqq \theta'$ (see Example §4.6.5). Observe further that we can assume that $\mathbb{Q} = \mathbb{P}_0$ (possibly after a reparametrisation). Supposing independent r.v.'s $X_1, \ldots, X_n$ with $(X_1, \ldots, X_k) \circledcirc \mathbb{P}_\mathbb{R}^{\otimes k}$ and $(X_{k+1}, \ldots, X_n) \sim \mathbb{P}_0^{\otimes l}$ their joint distribution belongs to the two sample location family $\mathbb{P}_\mathbb{R}^{k+l} := \{\mathbb{P}_\theta^{k+l} := \mathbb{P}_\theta^{\otimes k} \otimes \mathbb{P}_0^{\otimes l}, \theta \in \mathbb{R}\}$. Summarising, based on the statistical two sample location experiment $(\mathbb{R}^n, \mathscr{B}^{\otimes n}, \mathbb{P}_\mathbb{R}^{k+l})$ the aim is to test the hypothesis $H_0 : \theta = 0$ against the alternative $H_1 : \theta > 0$.

§4.7.1 **Regular location model**. A location family $\mathbb{P}_\mathbb{R}$ of probability measures on $(\mathbb{R}, \mathscr{B})$ as introduced in Example §4.2.3 is called *regular* if the density q is in addition continuously differentiable with derivative $\dot{q}$ satisfying $\lambda(|\dot{q}|^2/q) < \infty$. Following the Example §4.2.3 a regular location family $\mathbb{P}_\mathbb{R}$ is Hellinger-differentiable with score function $\dot{\ell}_\theta = -\dot{q}(x-\theta)/q(x-\theta)$ and Fisher information $\mathcal{I} := \lambda(|\dot{q}|^2/q)$. □

By applying Theorem §4.1.9 for a regular location model the associated product experiment $(\mathbb{R}^k, \mathscr{B}^{\otimes k}, \mathbb{P}_\mathbb{R}^{\otimes k})$ is ULAN with localising rate $(\delta_k := 1/\sqrt{k})_{k\in\mathbb{N}}$ and in $\theta_o = 0$ with central sequence $(\mathcal{Z}_0^k := -\sqrt{k}\mathcal{I}^{-1}\overline{\mathbb{P}}_k(\dot{q}/q))_{k\in\mathbb{N}}$. Precisely, for any sequence $h_k \to h$ it holds $\log(\mathrm{d}\mathbb{P}^{\otimes k}_{h_k/\sqrt{k}}/\mathrm{d}\mathbb{P}_0^{\otimes k}) = -h\sqrt{k}\,\overline{\mathbb{P}}_k(\dot{q}/q) - \frac{1}{2}h^2\mathcal{I} + o_{\mathbb{P}_0^{\otimes k}}(1)$ and $\sqrt{k}\,\overline{\mathbb{P}}_k(\dot{q}/q) \xrightarrow{d} \mathfrak{N}(0,\mathcal{I})$ under $\mathbb{P}_0^{\otimes k}$. Given a two sample location family $\mathbb{P}_\mathbb{R}^{k+l}$ for any $\theta \in \mathbb{R}$ the log of the likelihood-ratio satisfies $\log(\mathrm{d}\mathbb{P}_\theta^{k+l}/\mathrm{d}\mathbb{P}_0^{k+l}) = \log(\mathrm{d}\mathbb{P}_\theta^{\otimes k}/\mathrm{d}\mathbb{P}_0^{\otimes k})$. Thereby, if the location family is regular and $k/n \overset{n\to\infty}{\longrightarrow} \gamma \in (0,1)$, whence $h_k := h\sqrt{k/n} \overset{n\to\infty}{\longrightarrow} h\sqrt{\gamma}$, it follows

$$\Lambda_n := \log(\mathrm{d}\mathbb{P}^{k+l}_{h/\sqrt{n}}/\mathrm{d}\mathbb{P}_0^{k+l}) = \log(\mathrm{d}\mathbb{P}^{\otimes k}_{h_k/\sqrt{k}}/\mathrm{d}\mathbb{P}_0^{\otimes k})$$
$$= -h\,\sqrt{\gamma}\,\sqrt{k}\,\overline{\mathbb{P}}_k(\dot{q}/q) - \tfrac{\gamma}{2}h^2\mathcal{I} + o_{\mathbb{P}_0^n}(1) \quad (4.2)$$

§4.7.2 **Theorem**. *Assume a two sample regular location model. Consider for the test problem $H_0 : \theta = 0$ against $H_1 : \theta > 0$ the rank test $\varphi_n = \mathbb{1}_{\{U_{kl} > kl/2 + z_{1-\alpha}\sqrt{v_{kl}}\}} = \mathbb{1}_{\{U^\star_{kl} > z_{1-\alpha}\}}$ with $\mathbb{F}_{\mathfrak{N}(0,1)}(-z_{1-\alpha}) = \alpha$. If $k/n \overset{n\to\infty}{\longrightarrow} \gamma \in (0,1)$, then the following statements hold true:*

(i) *Under the null hypothesis $H_o : \theta = 0$ holds $\mathbb{P}_0^{k+l}\varphi_n = \mathbb{P}_0^{\otimes k+l}(U^\star_{kl} > z_{1-\alpha}) \overset{n\to\infty}{\longrightarrow} \alpha$, i.e., φ_n is an asymptotic level-α test;*

(ii) *Its power function* $\beta_{\varphi_n}(\theta) = \mathbb{P}_\theta^{k+l}\varphi_n$ *satisfies under local alternatives* $\beta_{\varphi_n}(h/\sqrt{n}) = \mathbb{P}_{h/\sqrt{n}}^{k+l}(U_{kl}^\star > z_{1-\alpha}) \overset{n\to\infty}{\longrightarrow} \mathbb{F}_{\mathfrak{N}(0,1)}(-z_{1-\alpha}+\rho)$ *with* $\rho = h(\lambda q^2)\sqrt{12\gamma(1-\gamma)}$.

Proof of Theorem §4.7.2 is given in the lecture. □

§4.7.3 **Remark**. Let us briefly consider the test of $H_0 : \theta = 0$ against the alternative $H_1 : \theta < 0$, where an asymptotic level-α test is given by $\varphi_n = \mathbb{1}_{\{U_{kl} < kl/2 + z_\alpha\sqrt{v_{kl}}\}} = \mathbb{1}_{\{U_{kl}^\star < z_\alpha\}}$. Its power function satisfies for local alternatives $\beta_{\varphi_n}(h/\sqrt{n}) = \mathbb{P}_{h/\sqrt{n}}^{k+l}(U_{kl}^\star < z_\alpha) \overset{n\to\infty}{\longrightarrow} \mathbb{F}_{\mathfrak{N}(0,1)}(z_\alpha - \rho)$. □

§4.7.4 **Gaussian two sample location model**. Consider a two sample Gaussian location experiment where $X_1, \ldots, X_n$ are independent r.v.'s with common variance $\sigma^2 > 0$ obeying $(X_1, \ldots, X_k) \sim \mathfrak{N}^{\otimes k}(\theta, \sigma^2)$ for some $\theta \in \mathbb{R}$ and $(X_{k+1}, \ldots, X_n) \sim \mathfrak{N}^{\otimes l}(0, \sigma^2)$. Consequently, their joint distribution belongs to the two sample Gaussian location family $\mathfrak{N}_{\mathbb{R}}^{k+l} := \{\mathfrak{N}_\theta^{k+l} := \mathfrak{N}^{\otimes k}(\theta, \sigma^2) \otimes \mathfrak{N}^{\otimes l}(0, \sigma^2), \theta \in \mathbb{R}\}$ which is obviously a regular. □

§4.7.5 **Example**. In a Gaussian two sample location model consider testing of the hypothesis $H_o : \theta = 0$ against $H_1 : \theta > 0$ (or $H_1 : \theta < 0$). Define $T_{kl} := l\sum_{i=1}^k X_i - k\sum_{i=k+1}^{k+l} X_i$ and $V_{kl} := \frac{kl(k+l)}{(k+l)-2}\{\sum_{i=1}^k (X_i - \frac{1}{k}\sum_{i=1}^k X_i)^2 + \sum_{i=k+1}^{k+l}(X_i - \frac{1}{l}\sum_{i=k+1}^{k+l} X_i)^2\}$ then under the null hypothesis, i.e., $(X_1, \ldots, X_n) \sim \mathfrak{N}^{\otimes n}(0, \sigma^2)$, the t-statistic $T_{kl}^\star := T_{kl}/\sqrt{V_{kl}}$ has a $\mathfrak{t}_{n-2}$-distribution with $n-2$ degrees of freedom, or $T_{kl}^\star \sim \mathfrak{t}_{n-2}$ for short. Let us denote by $\mathfrak{t}_{n-2,\kappa}$ its κ-quantile. Thereby, the t-test $\varphi_n^\star = \mathbb{1}_{\{T_{kl}^\star > \mathfrak{t}_{n-2,1-\alpha}\}}$ (or $\varphi_n^\star = \mathbb{1}_{\{T_{kl}^\star < \mathfrak{t}_{n-2,\alpha}\}}$) is a level-α test for $H_o : \theta = 0$ against $H_1 : \theta > 0$ (or $H_1 : \theta < 0$). Since a Gaussian location model is regular we can directly apply Theorem §4.7.2 to derive its asymptotic power function under local alternatives. However, Theorem §4.7.2 allows us to study a t-test in an arbitrary regular location model with mean location and variance. More precisely, in a Gaussian location family with common variance $\sigma^2 > 0$ introducing $g_\sigma(x) := x/\sigma$, $x \in \mathbb{R}$ the density q satisfies in addition $\lambda(g_\sigma q) = 0$ and $1 = \lambda(g_\sigma^2 q)$. □

§4.7.6 **Regular mean location and variance model**. Let $\sigma^2 > 1$ and $g_\sigma(x) := x/\sigma$, $x \in \mathbb{R}$. We call a regular location family with density q satisfying in addition $\lambda(g_\sigma q) = 0$ and $1 = \lambda(g_\sigma^2 q)$ a regular mean location and variance model. □

§4.7.7 **Theorem**. *Assume a two sample regular mean location and variance model. Consider for the test problem* $H_0 : \theta = 0$ *against* $H_1 : \theta > 0$ *the t-test* $\varphi_n^\star = \mathbb{1}_{\{T_{kl}^\star > t_{n-2,1-\alpha}\}}$ *with* $1 - \mathbb{F}_{\mathfrak{t}_{n-2}}(t_{n-2,1-\alpha}) = \alpha$. *If* $k/n \overset{n\to\infty}{\longrightarrow} \gamma \in (0,1)$, *then the following statements hold true:*

(i) *Under the null hypothesis* $H_o : \theta = 0$ *holds* $\mathbb{P}_0^{k+l}\varphi_n^\star \overset{n\to\infty}{\longrightarrow} \alpha$, *i.e.,* $\varphi_n^\star$ *is an asymptotic level-*α *test;*

(ii) *Its power function* $\beta_{\varphi_n^\star}(\theta) = \mathbb{P}_\theta^{k+l}\varphi_n^\star$ *satisfies under local alternatives* $\beta_{\varphi_n^\star}(h/\sqrt{n}) = \mathbb{P}_{h/\sqrt{n}}^{k+l}(T_{kl}^\star > \mathfrak{t}_{n-2,1-\alpha}) \overset{n\to\infty}{\longrightarrow} \mathbb{F}_{\mathfrak{N}(0,1)}(-z_{1-\alpha}+\rho)$ *with* $\rho = h\sigma^{-1}\sqrt{\gamma(1-\gamma)}$.

Proof of Theorem §4.7.7 is given in the lecture. □

§4.7.8 **Remark**. Let us compare the asymptotic level-α rank-test $\varphi_n = \mathbb{1}_{\{U_{kl} > kl/2 + z_{1-\alpha}\sqrt{v_{kl}}\}}$ and t-test $\varphi_n^\star = \mathbb{1}_{\{T_{kl}^\star > \mathfrak{t}_{n-2,1-\alpha}\}}$. Using their asymptotic power functions the asymptotic relative efficiency between both tests equals $\mathrm{are}(\varphi_n, \varphi_n^\star) = 12\sigma^2(\lambda q^2)^2$. In the particular case of a Gaussian location model, i.e., $q(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-x^2/(2\sigma^2))$ if follows $\lambda q^2 = 1/(2\sqrt{\pi}\sigma)$ and

hence $\operatorname{are}(\varphi_n, \varphi_n^\star) = 3/\pi \approx 0.955$. On the other hand side, if we denote by $\mathcal{Q}$ the class of all Lebesgue-densities on $\mathbb{R}$ satisfying $\lambda(g_\sigma q) = 0$ and $\lambda(g_\sigma^2 q) = 1$, then Hodges and Lehmann [1956] have shown that $\inf_{q\in\mathcal{Q}} 12\sigma^2(\lambda q^2)^2 = 0.864$ and $\sup_{q\in\mathcal{Q}} 12\sigma^2(\lambda q^2)^2 = \infty$. □

Chapter 5

Non-parametric statistics: local smoothing

This chapter presents elements of the non-parametric inference for curves along the lines of the textbooks by **Tsybakov** *[2009] and* **Comte** *[2015] where far more details, examples and further discussions can be found.*

5.1 Non-parametric curve estimation

Non-parametric density estimation. Consider a family $\mathcal{P}$ of probability measures on $(\mathbb{R}, \mathscr{B})$ which contains the distribution of an observation X, i.e., $X \circlearrowright \mathcal{P}$. The class $\mathcal{P}$ captures the prior knowledge about the distribution of the observation. Considering a class containing just a singleton, i.e., $\mathcal{P} = \{\mathbb{P}\}$ for some probability measure $\mathbb{P}$, means that the data generating process is known in advance. On the contrary taking $\mathcal{P}$ equal to the set $\mathcal{P}(\mathbb{R})$ of all possible probability measures on $(\mathbb{R}, \mathscr{B})$ reflects a lack of prior knowledge. In a certain sense a parametric model $\mathcal{P} = \mathbb{P}_\Theta$ for some parameter set $\Theta \subset \mathbb{R}^k$ provides then a usual trade-off between both extremes. On the other hand side, for an arbitrary probability measure $\mathbb{P}$ with c.d.f. $\mathbb{F}$ given an i.i.d. sample $X_1, \dots, X_n$ with common distribution $\mathbb{P}$ a reasonable estimator of $\mathbb{F}$ is the empirical c.d.f. $\overline{\mathbb{F}}_n(t) = \overline{\mathbb{P}}_n \mathbb{1}_{(-\infty,t]}$, $t \in \mathbb{R}$. Obviously, for each $t \in \mathbb{R}$, $\overline{\mathbb{F}}_n(t)$ is an unbiased estimator of $\mathbb{F}_t$ with variance $\mathbb{V}\mathrm{ar}(\overline{\mathbb{F}}_n(t)) = \frac{1}{n}\mathbb{F}(t)(1 - \mathbb{F}(t))$ and hence $\overline{\mathbb{F}}_n(t)$ converges in probability to $\mathbb{F}(t)$. Moreover, by the Law of Large Numbers §1.1.10 almost sure convergence holds true point-wise and even uniformly due to Glivenko-Cantelli's Theorem, i.e., $\left\|\overline{\mathbb{F}}_n - \mathbb{F}\right\|_{L^\infty} = \sup_{t\in\mathbb{R}} |\overline{\mathbb{F}}_n(t) - \mathbb{F}(t)| \xrightarrow{a.s.} 0$. If we assume in addition that $\mathbb{P}$ admits a Lebesgue density then $\overline{\mathbb{F}}_n$ is the unbiased estimator with minimal variance employing the Theorem of Lehman-Scheffé. However, comparing different probability measures using their associated c.d.f.'s is visually difficult and hence typically other measures for dissimilarities are used. Consider, for instance, for two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on $(\mathbb{R}, \mathscr{B})$ their *total variation distance* given by $\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} := \sup\{|\mathbb{P}(B) - \mathbb{Q}(B)|, B \in \mathscr{B}\}$. Noting that for any continuous probability measure $\mathbb{P}$ on $(\mathbb{R}, \mathscr{B})$, $\left\|\mathbb{P} - \overline{\mathbb{P}}_n\right\|_{\mathrm{TV}} \geqslant 1$ a.s. for any $n \in \mathbb{N}$, the empirical probability measure $\overline{\mathbb{P}}_n$ is not a consistent estimator of $\mathbb{P}$ in terms of the total variation distance. In other words, the estimator will usually depend on the measure (metric, topology, etc.) we use to quantify its accuracy as an estimator of $\mathbb{P}$.

§5.1.1 **Proposition** (Scheffé's theorem). *Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures on $(\mathbb{R}, \mathscr{B})$ absolute continuous w.r.t. the Lebesgue measure λ with densities $\mathfrak{p}$ and $\mathfrak{q}$, respectively. Then $\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} = \lambda(\mathfrak{p} - \mathfrak{q})^+ = \frac{1}{2}\lambda|\mathfrak{p} - \mathfrak{q}| = \frac{1}{2}\|\mathfrak{p} - \mathfrak{q}\|_{L^1}$.*

Proof of Proposition §5.1.1 is given, for example, in Tsybakov [2009], Lemma 2.1, p.70. □

In the sequel $\mathbb{D}$ denotes a family of Lebesgue-densities on $(\mathbb{R}, \mathbb{B})$ and for each density $\mathfrak{p} \in \mathbb{D}$ let $\mathbb{P}$ and $\mathbb{E}_\mathfrak{p}$ be its associated probability measure and expectation, respectively. We consider

the statistical product experiment $(\mathbb{R}^n, \mathscr{B}^{\otimes n}, \mathbb{P}_{\mathbb{D}}^{\otimes n} = \{\mathbb{P}^{\otimes n}, \mathbb{p} = \mathrm{d}\mathbb{P}/\mathrm{d}\lambda \in \mathbb{D}\})$ and we write $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathbb{p} \in \mathbb{D}$ for short. Moreover, $\mathbb{E}_{\mathbb{p}}^{\otimes n}$ denotes the expectation w.r.t. $\mathbb{P}^{\otimes n}$. Typically, given an estimator $\widehat{\mathbb{p}}$ of $\mathbb{p}$ we consider for $p \geqslant 1$ either $\mathbb{E}_{\mathbb{p}}^{\otimes n}|\widehat{\mathbb{p}}(t) - \mathbb{p}(t)|^p$, for each $t \in \mathbb{R}$, and $\mathbb{E}_{\mathbb{p}}^{\otimes n}\|\widehat{\mathbb{p}} - \mathbb{p}\|_{L^p}^p = \mathbb{E}_{\mathbb{p}}^{\otimes n}(\lambda|\widehat{\mathbb{p}} - \mathbb{p}|^p)$ as, respectively, a local and global measure of its accuracy with a special focus on $p = 1$ or $p = 2$.

Non-parametric regression. Describe the dependence of the variation of a real-valued r.v. Y (response) on the variation of an explanatory real-valued random or deterministic variable Z by a functional relationship $Y = f(Z) + \varepsilon$ where f is the unknown functional parameter of interest. Typically, it is assumed that the error term ε either is centred, i.e., $\mathbb{E}\varepsilon = 0$, in case of deterministic explanatory variables Z, or satisfies $\mathbb{E}(\varepsilon|Z) = 0$, in case of random Z. For a detailed discussion of the deterministic case we refer to Tsybakov [2009]. Here and subsequently, we restrict our attention to the case that $X := (Y, Z)$ is a random vector with values in a measure space $(\mathcal{X}, \mathscr{B})$ and our aim is statistical inference on $f(Z) = \mathbb{E}(Y|Z)$. Typically, the distribution of $X = (Y, Z)$ is parametrised by the regression function f only, i.e., $X \sim \mathbb{P}_f$, and the dependence on the marginal distribution of the regressor Z and the conditional distribution of the error term ε given Z is not made explicit. For sake of simplicity, let us in addition suppose that Z takes its values in $\mathbb{R}$ and the joint distribution of $X = (Y, Z)$ admits a joint Lebesgue density $\mathbb{p}_{Y,Z}$. Denoting by $\mathbb{p}_Z$ the marginal density of Z we use the identity $\ell(z) := f(z)\mathbb{p}_Z(z) = \int y\mathbb{p}_{Y,Z}(y, z)dy$ which as usual holds a.s. only. Given an i.i.d. sample of X a widely used estimation strategy is then based on a separate estimation of the function ℓ and the marginal density $\mathbb{p}_Z$, say by $\widehat{\ell}$ and $\widehat{\mathbb{p}}_Z$, and forming a possibly regularised estimator $\widehat{f} = (\widehat{\ell}/\widehat{\mathbb{p}}_Z)\mathbb{1}_{\{\widehat{\mathbb{p}}_Z > 0\}}$ for the function of interest $f = \ell/\mathbb{p}_Z$. However, there are many different approaches including local smoothing techniques, orthogonal series estimation, penalised smoothing techniques and combinations of them, to name but a few. In the sequel $\mathbb{F}$ denotes a family of regression functions and for each $f \in \mathbb{F}$ let $\mathbb{P}_f$ and $\mathbb{E}_f$ be the associated probability measure of $X = (Y, Z)$ and its expectation, respectively. We denote by $\mathbb{P}_{\mathbb{F}}$ the family of all possible distributions of X, but keep in mind, that generally the distribution of X is not uniquely determined by $f \in \mathbb{F}$ only. However, an i.i.d. sample of $X = (Y, Z)$ obeying the regression model we denote by $X_1, \ldots, X_n \circledsim \mathbb{P}_{\mathbb{F}}^{\otimes n}$ or $(Y_1, Z_1), \ldots, (Y_n, Z_n) \circledsim \mathbb{P}_{\mathbb{F}}^{\otimes n}$ for short. Typically, given an estimator $\widehat{f}$ of f we consider again for $p \geqslant 1$ either $\mathbb{E}_f^{\otimes n}|\widehat{f}(z) - f(z)|^p$, for each $z \in \mathbb{R}$, and $\mathbb{E}_f^{\otimes n}\|\widehat{f} - f\|_{L^p}^p = \mathbb{E}_f^{\otimes n}(\lambda|\widehat{f} - f|^p)$ as, respectively, a local and global measure of its accuracy.

5.2 Kernel density estimation

Throughout this section let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathbb{P}$ be real-valued r.v.'s with c.d.f. $\mathbb{F}$ and Lebesgue-density $\mathbb{p} = \mathrm{d}\mathbb{P}/\mathrm{d}\lambda$.

§5.2.1 **Definition**. An integrable map $K : \mathbb{R} \to \mathbb{R}$, i.e., $\lambda|K| < \infty$, with $\lambda K = 1$ is called a *kernel*. Given $h > 0$, typically called *bandwidth*, the *kernel density estimator* of $\mathbb{p}(x)$ evaluated at a point $x \in \mathbb{R}$ is defined as $\widehat{\mathbb{p}}_h(x) := \frac{1}{n}\sum_{i=1}^n \frac{1}{h}K\big(\frac{X_i - x}{h}\big) = \overline{\mathbb{P}}_n K_h^x$ using the abbreviation $K_h^x(X) := \frac{1}{h}K\big(\frac{X-x}{h}\big)$ for $x, X \in \mathbb{R}$. □

§5.2.2 **Remark**. Starting with $\mathbb{F}(x + h) - \mathbb{F}(x - h) = \lambda(\mathbb{1}_{]x-h,x+h]}\mathbb{p})$ for any $h > 0$ we have

for h sufficiently small $\mathbb{F}(x+h) - \mathbb{F}(x-h) \approx \mathbb{p}(x)2h$. Replacing the unknown c.d.f. $\mathbb{F}$ by its empirical counter part $\overline{\mathbb{F}}_n$ Rosenblatt [1956] proposed for $\mathbb{p}(x)$ the following estimator

$$\widehat{\mathbb{p}}_h(x) := \frac{\overline{\mathbb{F}}_n(x+h) - \overline{\mathbb{F}}_n(x-h)}{2h} = \tfrac{1}{2h}\,\overline{\mathbb{P}}_n \mathbb{1}_{]x-h,x+h]} = \tfrac{1}{n}\sum_{i=1}^{n} \tfrac{1}{2h}\mathbb{1}_{]-1,1]}\big(\tfrac{X_i - x}{h}\big)$$
$$= \tfrac{1}{n}\sum_{i=1}^{n} \tfrac{1}{h} K\big(\tfrac{X_i - x}{h}\big) = \tfrac{1}{n}\sum_{i=1}^{n} K_h^x(X_i) = \overline{\mathbb{P}}_n K_h^x$$

setting $K(x) := \frac{1}{2}\mathbb{1}_{]-1,1]}(x)$ and $K_h^x(X) := \frac{1}{h}K\big(\frac{X-x}{h}\big)$ for $x, X \in \mathbb{R}$. Observe that K is a density, which in turn implies that $\widehat{f}_h$ is a density for each $h > 0$ as well. Parzen [1962] introduces a kernel K and a bandwidth h as in Definition §5.2.1 and studies the more general kernel density estimator $\widehat{\mathbb{p}}_h(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h}K\big(\frac{X_i - x}{h}\big) = \overline{\mathbb{P}}_n K_h^x$. Note that $\lambda\widehat{\mathbb{p}}_h = 1$ since $\lambda K = 1$. If the kernel is in addition positive, then $\widehat{\mathbb{p}}_h$ is a density. An alternative motivation for a kernel density estimator provides the following lemma. □

§5.2.3 **Proposition** (Bochner's lemma). *Let $g : \mathbb{R} \to \mathbb{R}$ be bounded, i.e., $\|g\|_{L^\infty} < \infty$, and continuous in a neighbourhood of $x \in \mathbb{R}$. If $Q : \mathbb{R} \to \mathbb{R}$ is integrable, i.e., $\lambda|Q| < \infty$, and $Q_h^x := \frac{1}{h}Q(\frac{\bullet - x}{h})$, then $\lim_{h\to 0} \lambda(Q_h^x g) = \lim_{h\to 0} \frac{1}{h}\int Q(\frac{z-x}{h})g(z)dz = g(x)\int Q(z)dz = g(x)\lambda Q$*

Proof of Proposition §5.2.3 is given in the lecture. □

§5.2.4 **Example**. Typically considered is a rectangular kernel $K(u) := \frac{1}{2}\mathbb{1}_{\{[-1,1]\}}(u)$, a triangular kernel $K(u) := (1-|u|)\mathbb{1}_{\{[-1,1]\}}(u)$, an Epanechnikov kernel $K(u) := \frac{3}{4}(1-u^2)\mathbb{1}_{\{[-1,1]\}}(u)$ or a Gaussian kernel $K(u) := \frac{1}{\sqrt{2\pi}}\exp(-u^2/2)$. □

Local measure of accuracy. For a kernel density estimator $\widehat{\mathbb{p}}_h$ we consider first its mean squared error at a point $x \in \mathbb{R}$, that is, $\mathcal{R}(\widehat{\mathbb{p}}_h(x), \mathbb{p}(x)) = \mathbb{E}_{\mathbb{p}}^{\otimes n}|\widehat{\mathbb{p}}_h(x) - \mathbb{p}(x)|^2 =: \mathrm{MSE}(x)$. Observe that $\mathrm{MSE}(x) = \mathbb{V}\mathrm{ar}_{\mathbb{p}}(\widehat{\mathbb{p}}_h(x)) + |\mathrm{bias}_{\mathbb{p}}(x)|^2$ with $\mathrm{bias}_{\mathbb{p}}(x) := \mathbb{E}_{\mathbb{p}}^{\otimes n}\widehat{\mathbb{p}}_h(x) - \mathbb{p}(x)$ where we study separately the variance and the bias term, i.e., $\mathbb{V}\mathrm{ar}_{\mathbb{p}}(\widehat{\mathbb{p}}_h(x))$ and $\mathrm{bias}_{\mathbb{p}}(x)$.

§5.2.5 **Lemma**. *If $\|\mathbb{p}\|_{L^\infty} < \infty$ and $\|K\|_{L^2}^2 = \lambda K^2 < \infty$, then for each $x \in \mathbb{R}$ it holds $\mathbb{V}\mathrm{ar}_{\mathbb{p}}(\widehat{\mathbb{p}}_h(x)) \leqslant (nh)^{-1}\|\mathbb{p}\|_{L^\infty}\|K\|_{L^2}^2$.*

Proof of Lemma §5.2.5 is given in the lecture. □

§5.2.6 **Remark**. Let $\mathbb{p}$ be bounded and continuous, and suppose that K belongs to $L^1 \cap L^2$ with $\lambda K = 1$. From Lemma §5.2.5 follows then $\mathbb{V}\mathrm{ar}_{\mathbb{p}}(\widehat{\mathbb{p}}_h(x)) \leqslant (nh)^{-1}\|\mathbb{p}\|_{L^\infty}\|K\|_{L^2}^2$. On the other hand, since $\mathrm{bias}_{\mathbb{p}}(x) = \lambda(K_h^x\mathbb{p}) - \mathbb{p}(x)$ from Bochner's lemma §5.2.3 follows $|\mathrm{bias}_{\mathbb{p}}(x)| = o(1)$ as $h \to 0$. By combining both results, we obtain for any sequence $(h_n)_{n\in\mathbb{N}}$ of bandwidths satisfying $nh_n \to \infty$ and $h_n = o(1)$ that $\mathcal{R}(\widehat{\mathbb{p}}_{h_n}(x), \mathbb{p}(x)) = o(1)$ as $n \to \infty$. Consequently, the kernel density estimator is consistent, but its rate of convergence might be arbitrarily slow. Here and subsequently the bandwidth depends on n but we drop from now on the additional index n and write shortly $nh \overset{n\to\infty}{\longrightarrow} \infty$ or $h = o(1)$ as $n \to \infty$. □.

§5.2.7 **Lemma**. *Let $\mathbb{p}$ be twice-differentiable with bounded second derivative $\ddot{\mathbb{p}}$, i.e., $\|\ddot{\mathbb{p}}\|_{L^\infty} < \infty$ and let the kernel K satisfy in addition $\lambda(\mathrm{id}\,K) = 0$ and $\lambda(\mathrm{id}^2|K|) < \infty$ with $\mathrm{id}(u) := u$, $u \in \mathbb{R}$. Then for each $x \in \mathbb{R}$, $h > 0$ and $n \in \mathbb{N}$ it holds $|\mathrm{bias}_{\mathbb{p}}(x)| \leqslant h^2\frac{1}{2}\|\ddot{\mathbb{p}}\|_{L^\infty}\lambda(\mathrm{id}^2|K|)$.*

Proof of Lemma §5.2.7 is given in the lecture. □

§5.2.8 **Remark.** Let $\mathbb{p}$ be bounded and twice-differentiable with bounded second derivative $\ddot{\mathbb{p}}$ and suppose that K belongs to $L^1 \cap L^2$ with $\lambda K = 1$, $\lambda(\mathrm{id}\, K) = 0$ and $\lambda(\mathrm{id}^2 |K|) < \infty$. By combination of Lemma §5.2.5 and §5.2.7 follows uniformly for all $x \in \mathbb{R}$

$$\mathcal{R}(\widehat{\mathbb{p}}_h(x), \mathbb{p}(x)) \leqslant (nh)^{-1} \|\mathbb{p}\|_{L^\infty} \|K\|_{L^2}^2 + h^4 \tfrac{1}{4} \|\ddot{\mathbb{p}}\|_{L^\infty}^2 (\lambda(\mathrm{id}^2 |K|))^2,$$

where the first and second right hand side term is increasing and decreasing, respectively, as h tends to zero. Therefore, let us minimise the right hand side as a function of h. Keep in mind that $M(h) := a(nh)^{-1} + bh^{2\beta}$, $h > 0$, attains its minimum $M(h_o) = b\big(\frac{a}{2\beta b}\big)^{1/(2\beta+1)} n^{-2\beta/(2\beta+1)}$ at $h_o = \big(\frac{a}{2\beta b}\big)^{1/(2\beta+1)} n^{-1/(2\beta+1)}$. Therefore, choosing $h_o = \big(\frac{\|\mathbb{p}\|_{L^\infty} \|K\|_{L^2}^2}{\|\ddot{\mathbb{p}}\|_{L^\infty}^2 (\lambda(\mathrm{id}^2 |K|))^2}\big)^{1/5} n^{-1/5}$ we obtain

$$\sup_{x \in \mathbb{R}} \mathcal{R}(\widehat{\mathbb{p}}_{h_o}(x), \mathbb{p}(x)) \leqslant \tfrac{1}{4} \big(\|\ddot{\mathbb{p}}\|_{L^\infty}^2 (\lambda(\mathrm{id}^2 |K|))^2 \big)^{4/5} \big(\|\mathbb{p}\|_{L^\infty} \|K\|_{L^2}^2 \big)^{1/5} n^{-4/5}.$$

We shall emphasise that the optimal bandwidth h_o depends not only on the Kernel but also on characteristics of the unknown density $\mathbb{p}$, and hence, is in general not feasible in practise. □

§5.2.9 **Proposition.** *Let $\mathbb{p}$ be bounded and continuous in x and let $K \in L^1 \cap L^2$ be bounded with $\lambda K = 1$. If $hn \to \infty$ and $h = o(1)$ then $\sqrt{nh}(\widehat{\mathbb{p}}_h(x) - \mathbb{E}_{\mathbb{p}} \widehat{\mathbb{p}}_h(x)) \xrightarrow{d} \mathfrak{N}(0, \mathbb{p}(x)\lambda K^2)$.*

Proof of Proposition §5.2.9 is given in the lecture. □

§5.2.10 **Remark.** Let $\mathbb{p}$ be bounded and twice-differentiable with continuous in x and bounded second derivative $\ddot{\mathbb{p}}$. If K satisfies $\lambda(\mathrm{id}\, K) = 0$ and $\lambda(\mathrm{id}^2 |K|) < \infty$ in addition to the assumptions of Proposition §5.2.9 then $h^{-2}\mathrm{bias}_{\mathbb{p}}(x) = \frac{1}{2}\ddot{\mathbb{p}}(x)\lambda(\mathrm{id}^2 K) + o(1)$ as $h \to 0$. Consequently, choosing $hn^{1/5} \to c > 0$ it follows $\sqrt{nh}\mathrm{bias}_{\mathbb{p}}(x) = \frac{c^{5/2}}{2}\ddot{\mathbb{p}}(x)\lambda(\mathrm{id}^2 K) + o(1)$ and hence $\sqrt{nh}(\widehat{\mathbb{p}}_h(x) - \mathbb{p}(x)) \xrightarrow{d} \mathfrak{N}(\frac{c^{5/2}}{2}\ddot{\mathbb{p}}(x)\lambda(\mathrm{id}^2 K), \mathbb{p}(x)\lambda K^2)$ due Proposition §5.2.9. On the other hand side, if $hn^{1/5} = o(1)$ it follows in analogy $\sqrt{nh}(\widehat{\mathbb{p}}_h(x) - \mathbb{p}(x)) \xrightarrow{d} \mathfrak{N}(0, \mathbb{p}(x)\lambda K^2)$. □

§5.2.11 **Definition.** For $l \in \mathbb{N}$ a map $K : \mathbb{R} \to \mathbb{R}$ is called a *kernel of order l* if the functions $\mathrm{id}^j K$, $j \in [\![0, l]\!]$, are integrable and satisfy $\lambda K = 1$ and $\lambda(\mathrm{id}^j K) = 0$, $j \in [\![1, l]\!]$. □

§5.2.12 **Remark.** For arbitrary $l \in \mathbb{N}$ the construction of a kernel of order l and several examples are given, for instance, in Tsybakov [2009], section 1.2.2, or Comte [2015] section 3.2.4. □

§5.2.13 **Definition.** For two positive numbers β and L the *Hölder class* $\mathcal{H}(\beta, L)$ on $\mathbb{R}$ is a set of $l = \lfloor \beta \rfloor$ times differentiable functions $f : \mathbb{R} \to \mathbb{R}$ whose derivative $f^{(l)}$ for any $x, y \in \mathbb{R}$ satisfies $|f^{(l)}(x) - f^{(l)}(y)| \leqslant L|x - y|^{\beta - l}$. □

§5.2.14 **Lemma.** *Let $\mathbb{p} \in \mathcal{H}(\beta, L)$ and let K be a kernel of order $l = \lfloor \beta \rfloor$ satisfying $\lambda(|\mathrm{id}|^\beta |K|) < \infty$. Then for each $x \in \mathbb{R}$, $h > 0$ and $n \in \mathbb{N}$ it holds $|\mathrm{bias}_{\mathbb{p}}(x)| \leqslant h^\beta \frac{L}{l!} \lambda(|\mathrm{id}|^\beta |K|)$.*

Proof of Lemma §5.2.14 is given in the lecture. □

§5.2.15 **Remark.** Let $\mathbb{p} \in \mathcal{H}(\beta, L)$ be bounded and suppose that K is a kernel of order $l = \lfloor \beta \rfloor$ satisfying $\lambda K^2 < \infty$ and $\lambda(|\mathrm{id}|^\beta |K|) < \infty$. By combination of Lemma §5.2.5 and §5.2.14 follows uniformly for all $x \in \mathbb{R}$

$$\mathcal{R}(\widehat{\mathbb{p}}_h(x), \mathbb{p}(x)) \leqslant (nh)^{-1} \|\mathbb{p}\|_{L^\infty} \|K\|_{L^2}^2 + h^{2\beta} \big(\tfrac{L}{l!} \lambda(|\mathrm{id}|^\beta |K|)\big)^2,$$

therefore minimising the right hand side as a function of h leads to an optimal bandwidth $h_o = c\, n^{-1/(2\beta+1)}$ with constant $c^{2\beta+1} = \frac{\|\mathbb{p}\|_{L^\infty} \|K\|_{L^2}^2}{2\beta(\frac{L}{l!}\lambda(|\mathrm{id}|^\beta |K|))^2}$. Consequently, by choosing the optimal bandwidth h_o we have $\sup_{x \in \mathbb{R}} \mathcal{R}(\widehat{\mathbb{p}}_{h_o}(x), \mathbb{p}(x)) = O(n^{-2\beta/(2\beta+1)})$. However, the optimal bandwidth h_o depends again on characteristics of the unknown density $\mathbb{p}$, and hence, is in general not feasible in practise. □

Global measure of accuracy. Assuming a density $\mathbb{p} \in L^2$ we consider next the integrated mean squared error (MISE) of the kernel density estimator $\widehat{\mathbb{p}}_h$, that is, $\mathrm{MISE} := \mathcal{R}(\widehat{\mathbb{p}}_h, \mathbb{p}) = \mathbb{E}_{\mathbb{p}}^{\otimes n} \|\widehat{\mathbb{p}}_h - \mathbb{p}\|_{L^2}^2 = \mathbb{E}_{\mathbb{p}}^{\otimes n} \lambda |\widehat{\mathbb{p}}_h - \mathbb{p}|^2$. Observe that $\mathrm{MISE} = \int_{\mathbb{R}} \mathbb{V}\mathrm{ar}_{\mathbb{p}}(\widehat{\mathbb{p}}_h(x)) dx + \int_{\mathbb{R}} \mathrm{bias}_{\mathbb{p}}^2(x) dx$ with $\mathrm{bias}_{\mathbb{p}}(x) = \lambda(K_h^x \mathbb{p}) - \mathbb{p}(x)$ where we study now separately the integrated variance and bias term.

§5.2.16 **Lemma.** *If $K \in L^2$, then for any density $\mathbb{p}$ holds $\int_{\mathbb{R}} \mathbb{V}\mathrm{ar}_{\mathbb{p}}(\widehat{\mathbb{p}}_h(x)) dx \leqslant (nh)^{-1} \|K\|_{L^2}^2$.*

Proof of Lemma §5.2.16 is given in the lecture. □

§5.2.17 **Definition.** For two positive numbers β and L the *Nikol'ski class* $\mathcal{N}(\beta, L)$ on $\mathbb{R}$ is a set of $l = \lfloor \beta \rfloor$ times differentiable functions $f : \mathbb{R} \to \mathbb{R}$ whose derivative $f^{(l)}$ for all $t \in \mathbb{R}$ satisfies $\big(\int |f^{(l)}(x+t) - f^{(l)}(x)|^2 dx \big)^{1/2} \leqslant L|t|^{\beta - l}$. □

§5.2.18 **Lemma.** *Let $\mathbb{p} \in \mathcal{N}(\beta, L)$ and let K be a kernel of order $l = \lfloor \beta \rfloor$ satisfying $\lambda(|\mathrm{id}|^\beta |K|) < \infty$. Then for each $x \in \mathbb{R}$, $h > 0$ and $n \in \mathbb{N}$ it holds $\int |\mathrm{bias}_{\mathbb{p}}(x)|^2 dx \leqslant h^{2\beta} \{\frac{L}{l!} \lambda(|\mathrm{id}|^\beta |K|)\}^2$.*

Proof of Lemma §5.2.18 is given in the lecture. □

§5.2.19 **Remark.** Let $\mathbb{p} \in L^2 \cap \mathcal{N}(\beta, L)$ and let K be a kernel of order $l = \lfloor \beta \rfloor$ satisfying $\lambda K^2 < \infty$ and $\lambda(|\mathrm{id}|^\beta |K|) < \infty$. By combination of Lemma §5.2.16 and §5.2.18 follows

$$\mathcal{R}(\widehat{\mathbb{p}}_h, \mathbb{p}) \leqslant (nh)^{-1} \|K\|_{L^2}^2 + h^{2\beta} \big(\tfrac{L}{l!} \lambda(|\mathrm{id}|^\beta |K|)\big)^2,$$

therefore minimising the right hand side as a function of h leads to an optimal bandwidth $h_o = c\, n^{-1/(2\beta+1)}$ with constant $c^{2\beta+1} = \frac{\lambda K^2}{2\beta(\frac{L}{l!}\lambda(|\mathrm{id}|^\beta |K|))^2}$. Consequently, by choosing the optimal bandwidth h_o we have $\mathcal{R}(\widehat{\mathbb{p}}_{h_o}, \mathbb{p}) = O(n^{-2\beta/(2\beta+1)})$. However, the optimal bandwidth h_o depends again on characteristics of the unknown density $\mathbb{p}$, and hence, is in general not feasible in practise. □

Data-driven bandwidth selection. Considering a kernel density estimator $\widehat{\mathbb{p}}_h$ the choice of the bandwidth h is crucial. An ideal value of the bandwidth is, for instance, given by $h_{id} = \arg\min\{\mathcal{R}(\widehat{\mathbb{p}}_h, \mathbb{p}), h > 0\}$. Note that for a given density $\mathbb{p}$, the estimator $\widehat{\mathbb{p}}_{h_{id}}$, if h_{id} exists, has minimal MISE within the family $\{\widehat{\mathbb{p}}_h, h > 0\}$ of all kernel density estimators

with fixed kernel and varying bandwidth. Unfortunately, the value h_{id} and hence $\widehat{\mathbb{p}}_{h_{id}}$ remains purely theoretical and thus is often called *oracle*, since $\mathcal{R}(\widehat{\mathbb{p}}_h, \mathbb{p})$ depends on unknown characteristics of the density $\mathbb{p}$. A common idea is to use unbiased estimation of the risk $\mathcal{R}(\widehat{\mathbb{p}}_h, \mathbb{p})$ and to minimise the unbiased estimator of the risk rather than the unknown risk itself. Note that $\mathcal{R}(\widehat{\mathbb{p}}_h, \mathbb{p}) = \mathbb{E}_{\mathbb{p}}^{\otimes n}\{\lambda\widehat{\mathbb{p}}_h^2 - 2\lambda(\widehat{\mathbb{p}}_h\mathbb{p})\} + \lambda\mathbb{p}^2$. Since the integral $\lambda\mathbb{p}^2$ does not depend on h the minimiser h_{id} of $\mathcal{R}(\widehat{\mathbb{p}}_h, \mathbb{p})$ also minimises the function $J(h) = \mathbb{E}_{\mathbb{p}}^{\otimes n}\{\lambda\widehat{\mathbb{p}}_h^2 - 2\lambda(\widehat{\mathbb{p}}_h\mathbb{p})\}$. We construct now an unbiased estimator of $J(h)$. For this purpose it is sufficient to find an unbiased estimator for each of the quantities $\mathbb{E}_{\mathbb{p}}^{\otimes n}\lambda\widehat{\mathbb{p}}_h^2$ and $\mathbb{E}_{\mathbb{p}}^{\otimes n}\lambda(\widehat{\mathbb{p}}_h\mathbb{p})$. A trivial unbiased of $\mathbb{E}_{\mathbb{p}}^{\otimes n}\lambda\widehat{\mathbb{p}}_h^2$ is $\lambda\widehat{\mathbb{p}}_h^2$. Define further $\widehat{\mathbb{p}}_h^{-i}(x) = \frac{1}{(n-1)}\sum_{j\neq i} K_h^x(X_j)$, then $\frac{1}{n}\sum_{i=1}^n \widehat{\mathbb{p}}_h^{-i}(X_i)$ is an unbiased estimator of $\mathbb{E}_{\mathbb{p}}^{\otimes n}\lambda(\widehat{\mathbb{p}}_h\mathbb{p})$. Consequently, $CV(h) := \lambda\widehat{\mathbb{p}}_h^2 - \frac{2}{n}\sum_{i=1}^n \widehat{\mathbb{p}}_h^{-i}(X_i)$ is an unbiased estimator of $J(h)$, where CV stands for "cross-validation". The function CV is called the leave-one-out cross-validation criterion or simply the cross-validation criterion. Keeping in mind, that the functions $h \mapsto \mathcal{R}(\widehat{\mathbb{p}}_h, \mathbb{p})$ and $h \mapsto \mathbb{E}_{\mathbb{p}}^{\otimes n}\{CV(h)\}$ have the same minimiser. In turn, the minimizers of $\mathbb{E}_{\mathbb{p}}^{\otimes n}\{CV(h)\}$ can be approximated by those of the function CV which can be computed from the sample: $h_{cv} = \arg\min\{CV(h), h > 0\}$ whenever the minimum is attained. Finally, we define the cross-validation estimator $\widehat{\mathbb{p}}_{h_{cv}}$. Note that this is a kernel estimator with random bandwidth h_{cv} depending on the sample only. It can be proved that under appropriate conditions the risk of the estimator $\widehat{\mathbb{p}}_{h_{cv}}$ is asymptotically equivalent to that of the ideal kernel pseudo-estimator (oracle) $\widehat{\mathbb{p}}_{h_{id}}$.

5.3 Non-parametric regression

Here and subsequently, consider i.i.d. r.v.'s $(Y, Z), (Y_1, Z_1), (Y_2, Z_2), \ldots$ obeying a non-parametric regression model $\mathbb{E}_f(Y|Z) = f(Z)$ for some unknown regression function $f \in \mathbb{F}$ as introduced in section 5.1, i.e., $(Y_1, Z_1), \ldots, (Y_n, Z_n) \backsim \mathbb{P}_{\mathbb{F}}^{\otimes n}$.

§5.3.1 **Assumptions and notations**. (i) The centred error term $\varepsilon := Y - f(Z)$, i.e., $\mathbb{E}_f(\varepsilon) = 0$, has a finite second moment $\sigma_\varepsilon^2 := \mathbb{E}_f(\varepsilon^2)$. (ii) ε and the real-valued explanatory variable Z are independent. (iii) The joint distribution $\mathbb{P}_f^{Y,Z}$ of (Y, Z) admits a Lebesgue density $\mathbb{p}^{Y,Z}$. The marginal Lebesgue density of Z is denoted by $\mathbb{p}^Z$. (iv) Define $\ell := f\mathbb{p}^Z = \int y\mathbb{p}^{Y,Z}(y, \bullet)dy$. □

Consider a kernel density estimator $\widehat{\mathbb{p}}_h^{Y,Z}(y, z) = \frac{1}{n}\sum_{i=1}^n K_h^y(Y_i)K_h^z(Z_i)$ of the joint density $\mathbb{p}^{Y,Z}(y, z)$ with $K_h^x(X) := \frac{1}{h}K(\frac{X-x}{h})$ for some kernel function K and bandwidth $h > 0$. Keeping in mind that $\ell = \int y\mathbb{p}^{Y,Z}(y, \bullet)dy$ and $\mathbb{p}^Z = \int \mathbb{p}^{Y,Z}(y, \bullet)dy$ their estimators are obtained by replacing the unknown density $\mathbb{p}^{Y,Z}$ by its kernel density estimator $\widehat{\mathbb{p}}_h^{Y,Z}$. If the kernel K satisfies $\lambda K = 1$ and $\lambda(\mathrm{id}\, K) = 0$, then $\widehat{\ell}_h(z) := \int y\widehat{\mathbb{p}}_h^{Y,Z}(y, z)dy = \frac{1}{n}\sum_{i=1}^n Y_i K_h^z(Z_i)$ and $\widehat{\mathbb{p}}_h^Z(z) := \int \widehat{\mathbb{p}}_h^{Y,Z}(y, z)dy = \frac{1}{n}\sum_{i=1}^n K_h^z(Z_i)$ is the usual kernel density estimator of $\mathbb{p}^Z$.

§5.3.2 **Definition**. Given a kernel K and a bandwidth h, the *Nadaraya–Watson estimator* of $f(z)$ evaluated at a point $z \in \mathbb{R}$ is defined as

$$\widehat{f}_h(z) := \frac{\widehat{\ell}_h(z)}{\widehat{\mathbb{p}}_h^Z(z)} = \frac{\frac{1}{n}\sum_{i=1}^n Y_i K_h^z(Z_i)}{\frac{1}{n}\sum_{j=1}^n K_h^z(Z_j)} = \sum_{i=1}^n Y_i \frac{K_h^z(Z_i)}{\sum_{j=1}^n K_h^z(Z_j)}, \quad \text{if } \widehat{\mathbb{p}}_h^Z(z) \neq 0$$

and $\widehat{f}_h(z) = 0$ otherwise, using the abbreviation $K_h^z(Z) := \frac{1}{h}K\big(\frac{Z-z}{h}\big)$ for $z, Z \in \mathbb{R}$. □

Local measure of accuracy. Keeping in mind that $\widehat{\mathbb{p}}_h^z$ is a kernel density estimator of $\mathbb{p}^z$ we can apply the results obtained in the last section. Therefore, it remains to consider the estimator $\widehat{\ell}_h$ of ℓ. We consider first its mean squared error at a given point $z \in \mathbb{R}$, that is, $\mathcal{R}(\widehat{\ell}_h(z), \ell(z)) = \mathbb{E}_f^{\otimes n}|\widehat{\ell}_h(z) - \ell(z)|^2 = \mathrm{MSE}(z) = \mathbb{V}\mathrm{ar}_f(\widehat{\ell}_h(z)) + |\mathrm{bias}_\ell(z)|^2$ with $\mathrm{bias}_\ell(z) := \mathbb{E}_f\widehat{\ell}_h(z) - \ell(z) = \lambda(K_h^z\ell) - \ell(z)$ where we study separately the variance and the bias term, i.e., $\mathbb{V}\mathrm{ar}_f(\widehat{\ell}_h(z))$ and $\mathrm{bias}_\ell(z)$. Obviously, as in the density estimation case replacing the density $\mathbb{p}$ by ℓ Bochner's lemma §5.2.3, Lemma §5.2.7 and §5.2.14 provide bounds for $\mathrm{bias}_\ell(z) = \lambda(K_h^z\ell) - \ell(z)$.

§5.3.3 **Lemma.** *If $\|f\|_{L^\infty} < \infty$, $\|\mathbb{p}^z\|_{L^\infty} < \infty$ and $\|K\|_{L^2}^2 = \lambda K^2 < \infty$, then for each $z \in \mathbb{R}$ it holds $\mathbb{V}\mathrm{ar}_f(\widehat{\ell}_h(z)) \leqslant (nh)^{-1}(\|f\|_{L^\infty}^2 + \sigma_\varepsilon^2)\|\mathbb{p}\|_{L^\infty}\|K\|_{L^2}^2$.*

Proof of Lemma §5.3.3 is given in the lecture. □

§5.3.4 **Remark.** Let f and $\mathbb{p}^z$, and hence, $\ell = f\mathbb{p}^z$, be bounded. Suppose that the function ℓ belongs to the Hölder class $\mathcal{H}(\beta, L)$ defined in §5.2.13 and that K is a kernel of order $l = \lfloor\beta\rfloor$ as defined in §5.2.11 satisfying $\lambda K^2 < \infty$ and $\lambda(|\mathrm{id}|^\beta|K|) < \infty$. By combination of Lemma §5.3.3 and §5.2.14 applied to ℓ rather than $\mathbb{p}^z$ follows uniformly for all $z \in \mathbb{R}$

$$\mathcal{R}(\widehat{\ell}_h(z), \ell(z)) \leqslant (nh)^{-1}(\|f\|_{L^\infty}^2 + \sigma_\varepsilon^2)\|\mathbb{p}\|_{L^\infty}\|K\|_{L^2}^2 + h^{2\beta}\big(\tfrac{L}{l!}\lambda(|\mathrm{id}|^\beta|K|)\big)^2.$$

Therefore minimising the right hand side as a function of h leads to an optimal bandwidth $h_o = c\,n^{-1/(2\beta+1)}$ with constant $c^{2\beta+1} = \frac{(\|f\|_{L^\infty}^2+\sigma_\varepsilon^2)\|\mathbb{p}\|_{L^\infty}\|K\|_{L^2}^2}{2\beta(\frac{L}{l!}\lambda(|\mathrm{id}|^\beta|K|))^2}$. Consequently, by choosing the optimal bandwidth h_o we have $\sup_{z\in\mathbb{R}}\mathcal{R}(\widehat{\ell}_{h_o}(z), \ell(z)) = O(n^{-2\beta/(2\beta+1)})$. However, the optimal bandwidth h_o depends again on characteristics of the unknown function ℓ. □

Global measure of accuracy. Assuming $\ell \in L^2$ we consider next the integrated mean squared error (MISE) of the kernel estimator $\widehat{\ell}_h$, that is, $\mathrm{MISE} := \mathcal{R}(\widehat{\ell}_h, \ell) = \mathbb{E}_f^{\otimes n}\|\widehat{\ell}_h - \ell\|_{L^2}^2 = \mathbb{E}_f\lambda|\widehat{\ell}_h - \ell|^2 = \int_{\mathbb{R}}\mathbb{V}\mathrm{ar}_f(\widehat{\ell}_h(z))dz + \int_{\mathbb{R}}\mathrm{bias}_\ell^2(z)dz$ with $\mathrm{bias}_\ell(z) = \lambda(K_h^z\ell) - \ell(z)$ where we study now separately the integrated variance and bias term. Note that, as in the density estimation case replacing the density $\mathbb{p}$ by ℓ Lemma §5.2.18 provides a bound for $\mathrm{bias}_\ell(z) = \lambda(K_h^z\ell) - \ell(z)$.

§5.3.5 **Lemma.** *If $K \in L^2$ and $\lambda(\mathbb{p}^z f^2) < \infty$, and hence $\sigma_Y^2 := \mathbb{E}_f Y^2 = \lambda(\mathbb{p}^z f^2) + \sigma_\varepsilon^2 < \infty$, then $\int_{\mathbb{R}}\mathbb{V}\mathrm{ar}_f(\widehat{\ell}_h(z))dz \leqslant (nh)^{-1}\sigma_Y^2\,\|K\|_{L^2}^2$.*

Proof of Lemma §5.3.5 is given in the lecture. □

§5.3.6 **Remark.** Let $\lambda(\mathbb{p}^z f^2) < \infty$, $\ell \in L^2 \cap \mathcal{N}(\beta, L)$ and K be a kernel of order $l = \lfloor\beta\rfloor$ satisfying $\lambda K^2 < \infty$ and $\lambda(|\mathrm{id}|^\beta|K|) < \infty$. By combination of Lemma §5.3.5 and §5.2.18 follows

$$\mathcal{R}(\widehat{\ell}_h, \ell) \leqslant (nh)^{-1}\sigma_Y^2\,\|K\|_{L^2}^2 + h^{2\beta}\big(\tfrac{L}{l!}\lambda(|\mathrm{id}|^\beta|K|)\big)^2,$$

therefore minimising the right hand side as a function of h leads to an optimal bandwidth $h_o = c\,n^{-1/(2\beta+1)}$ with constant $c^{2\beta+1} = \frac{\sigma_Y^2\lambda K^2}{2\beta(\frac{L}{l!}\lambda(|\mathrm{id}|^\beta|K|))^2}$. Consequently, by choosing the optimal bandwidth h_o we have $\mathcal{R}(\widehat{\ell}_{h_o}, \mathbb{p}) = O(n^{-2\beta/(2\beta+1)})$. However, the optimal bandwidth h_o depends again on characteristics of the unknown function ℓ. □

Under regularity conditions we have shown that the MSE of $\widehat{\ell}_h$ and $\widehat{\mathbb{p}}^z_h$ tend to zero as $n \to \infty$ provided the bandwidth is chosen appropriately. In this situation, it follows directly $\widehat{\ell}_h(z) \overset{\mathbb{P}_f^{\otimes n}}{\to} \ell(z)$ and $\widehat{\mathbb{p}}^z_h(z) \overset{\mathbb{P}_f^{\otimes n}}{\to} \mathbb{p}^z(z)$ which in turn implies $\widehat{f}_h(z) = \widehat{\ell}_h(z)/\widehat{\mathbb{p}}^z_h(z) \overset{\mathbb{P}_f^{\otimes n}}{\to} \ell(z)/\mathbb{p}^z(z) = f(z)$. Moreover, it is straightforward to show that under similar assumption as used in Proposition §5.2.9 the asymptotic normality of $\widehat{\ell}_h(z)$ holds true, which due to Slutky's lemma §1.1.7 allows then to establish the asymptotic normality of $\widehat{f}_h(z)$. In order to derive an upper bound for the MISE we use in the next assertion a regularised version of $\widehat{f}_h(z)$ which makes use of a stronger assumption, that is, $\mathbb{p}^z(z) > p_o$, $z \in \mathbb{R}$, for some known constant $p_o > 0$.

§5.3.7 **Lemma**. *Suppose that $\mathbb{p}^z(z) > p_o$, $z \in \mathbb{R}$, for some known constant $p_o > 0$. Consider the regularised Nadaraya–Watson estimator $\widehat{f}^o_h := \frac{\widehat{\ell}_h}{\widehat{\mathbb{p}}^z_h} \mathbb{1}_{\{\widehat{\mathbb{p}}^z_h > p_o/2\}}$. If $\|f\|_{L^\infty} < \infty$ then $\mathcal{R}(\widehat{f}^o_h, f) = \mathbb{E}_f^{\otimes n} \|\widehat{f}^o_h - f\|^2_{L^2} \leqslant \frac{8}{p_o^2} \mathbb{E}_f^{\otimes n} \|\widehat{\ell}_h - \ell\|^2_{L^2} + \frac{12\|f\|^2_{L^\infty}}{p_o^2} \mathbb{E}_f^{\otimes n} \|\widehat{\mathbb{p}}^z_h - \mathbb{p}^z\|^2_{L^2}$.*

Proof of Lemma §5.3.7 is given in the lecture. □

Local polynomial estimators. Let the kernel K take only non-negative values. It is easily verified, that the Nadaraya–Watson estimator $\widehat{f}_h$ satisfies

$$\widehat{f}_h(z) = \arg\min_{\theta \in \mathbb{R}} \sum_{i=1}^n (Y_i - \theta)^2 K^z_h(Z_i).$$

Therefore, $\widehat{f}_h$ is obtained by a local constant least squares approximation of the responses $\{Y_i\}$. The locality is determined by a kernel K that downweights all the Z_i that are not close to z whereas θ plays the role of a local constant to be fitted. More generally, we may define a local polynomial least squares approximation, replacing the constant θ by a polynomial of a pre-specified degree.

§5.3.8 **Definition**. For $l \in \mathbb{R}$ consider $U : \mathbb{R} \to \mathbb{R}^{l+1}$, $z \mapsto U(z) = (1, z, z^2/2!, \ldots, z^l/l!)$. Let $K : \mathbb{R} \to \mathbb{R}$ be a kernel and $h > 0$ be a bandwidth. A vector $\widehat{\theta}(z) \in \mathbb{R}^{l+1}$ satisfying

$$\widehat{\theta}(z) = \arg\min_{\theta \in \mathbb{R}^{l+1}} \sum_{i=1}^n (Y_i - \theta^t U(\tfrac{Z_i - z}{h}))^2 K^z_h(Z_i).$$

is called a *local polynomial estimator of order l* of $\theta(z) = (f(z), h\dot{f}(z), h^2\ddot{f}(z), \ldots, h^l f^{(l)}(z))$. The statistic $\widehat{f}_h(z) = U^t(0)\widehat{\theta}(z)$ is called *local polynomial estimator of order l* of $f(z)$. □

Note that $\widehat{f}_h(z)$ is simply the first coordinate of the vector $\widehat{\theta}(z)$. Obviously, the Nadaraya–Watson estimator with non-negative kernel is just a local polynomial estimator of order zero, Furthermore, properly normalised coordinates of $\widehat{\theta}(z)$ provide estimators of the derivatives $\dot{f}(z), \ddot{f}(z), \ldots, f^{(l)}(z)$. For theoretical properties of local polynomial estimators and their detailed discussion we refer to Tsybakov [2009], section 1.6.

Chapter 6

Non-parametric statistics: orthogonal series estimation

We study non-parametric estimation of a functional parameter of interest f based on a noisy version $\widehat{f} = f + n^{-1/2}\dot{W}$ of f contaminated by an additive random error $\dot{W}$ with noise level $n^{-1/2}$. The quantity $n \in \mathbb{N}$ is usually called sample size referring to statistical problems where the noisy version $\widehat{f}$ is constructed using a sample of size n. For convenience, we suppose that the function of interest f belongs to an Hilbert space and thus permits an orthogonal series expansion. We briefly recall theoretical basics and terminologies from functional analysis which allow us to formalise the statistical experiment as a sequence space model. Throughout the following chapters we illustrate the results using three particular models, namely, non-parametric regression with uniformly distributed random design, non-parametric density estimation and a Gaussian sequence space model.

6.1 Theoretical basics and terminologies

For a detailed and extensive survey on functional analysis we refer the reader, for example, to Werner [2011] or the series of textbooks by Dunford and Schwartz [1988a,b,c].

§6.1.1 **Definition**. 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 $h \in \mathbb{H}$. □

§6.1.2 **Property**.

(Cauchy-Schwarz inequality) $|\langle h_1, h_2\rangle_{\mathbb{H}}| \leqslant \|h_1\|_{\mathbb{H}} \cdot \|h_2\|_{\mathbb{H}}$ *for all* $h_1, h_2 \in \mathbb{H}$. □

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

(ii) Given $\mathcal{J} \subseteq \mathbb{Z}$, denote by $\mathbb{K}^{\mathcal{J}}$ the vector space of all $\mathbb{K}$-valued sequences over $\mathcal{J}$ where we refer to any sequence $(x_j)_{j\in\mathcal{J}} \in \mathbb{K}^{\mathcal{J}}$ as a whole by omitting its index as for example in «the sequence x» and arithmetic operations on sequences are defined element-wise, i.e., $xy := (x_j y_j)_{j\in\mathcal{J}}$. In the sequel, let $\|x\|_{\ell^p} := (\sum_{j\in\mathcal{J}} |x_j|^p)^{1/p}$, for $p \in [1, \infty)$, and $\|x\|_{\ell^\infty} := \sup_{j\in\mathcal{J}} |x_j|$. Thereby, for $p \in [1, \infty]$, consider $\ell^p(\mathcal{J}) := \{(x_j)_{j\in\mathcal{J}} \in \mathbb{K}^{\mathcal{J}}, \|x\|_{\ell^p} < \infty\}$, or ℓ^p for short, endowed with the norm $\|\cdot\|_{\ell^p}$. In particular, $\ell^2(\mathcal{J})$ is the usual *Hilbert space of square summable sequences over* $\mathcal{J}$ endowed with the inner product $\langle x, y\rangle_{\ell^2} := \sum_{j\in\mathcal{J}} x_j \overline{y}_j$

for all $x, y \in \ell^2(\mathcal{J})$.

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

(iv) Let $\mathscr{B}$ be the Borel-σ-algebra on $\mathbb{K}$. Given a measure space $(\Omega, \mathscr{A}, \mu)$ denote by $\mathbb{K}^\Omega$ the vector space of all $\mathbb{K}$-valued functions $f : \Omega \to \mathbb{K}$. Recall that $\|f\|_{L^p_\mu} = (\mu|f|^p)^{1/p}$, for $p \in [1, \infty)$, and $\|f\|_{L^\infty_\mu} := \inf\{c : \mu(|f| > c) = 0\}$, where for $p \in [1, \infty]$, we write $L^p(\Omega, \mathscr{A}, \mu) := \{f \in \mathbb{K}^\Omega, \mathscr{A}\text{-}\mathscr{B}\text{-measurable}, \|f\|_{L^p} < \infty\}$, $L^p_\mu(\Omega)$, L^p_μ or L^p for short, which is endowed with the norm $\|\cdot\|_{L^p_\mu}$ or $\|\cdot\|_{L^p}$ for short. $L^2(\Omega, \mathscr{A}, \mu)$, $L^2_\mu(\Omega)$, L^2_μ or L^2 for short, is the usual *Hilbert space of square μ-integrable, $\mathscr{A}$-$\mathscr{B}$-measurable functions on Ω* endowed with the inner product $\langle f, g\rangle_{L^2} := \mu(f\overline{g})$ for all $f, g \in L^2_\mu$.

(v) For a strictly positive function $\mathfrak{v}$ consider the *weighted norm* $\|f\|_{\mathfrak{v}}^2 := \mu(\mathfrak{v}^2 f^2)$. We define $L^2_{\mathfrak{v}}(\Omega, \mathscr{A}, \mu)$, or $L^2_{\mathfrak{v}}$ for short, as the completion of $L^2(\Omega, \mathscr{A}, \mu)$ w.r.t. $\|\cdot\|_{\mathfrak{v}}$, which is a Hilbert space endowed with $\langle f, g\rangle_{\mathfrak{v}} := \langle \mathfrak{v}f, \mathfrak{v}g\rangle_{L^2} = \mu(\mathfrak{v}^2 f\overline{g})$ for all $f, g \in L^2_{\mathfrak{v}}$.

(vi) Let X be a random variable (r.v.) on a probability space $(\Omega, \mathscr{A}, \mathbb{P})$ taking its values in a measurable space $(\mathcal{X}, \mathscr{B})$. For $p \in [1, \infty]$ we set $L^p_X := L^p(\mathcal{X}, \mathscr{B}, \mathbb{P}^X)$ where L^2_X is a Hilbert space endowed with $\langle f, g\rangle_{L^2_X} = \mathbb{P}^X(f\overline{g})$ for all $f, g \in L^2_X$. □

§6.1.4 **Definition**. 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 $\|u\|_{\mathbb{H}} = 1$, $\forall u \in \mathcal{U}$. We say $\mathcal{U}$ is an *orthonormal basis (ONB)* if $\mathcal{U} \subset \mathcal{U}'$ and $\mathcal{U}'$ is ONS, then $\mathcal{U} = \mathcal{U}'$, i.e., if it is a *complete* ONS.

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

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

is orthonormal and complete, i.e. an ONB.

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

$$e_j(t) := \exp(-\iota 2\pi j t) \text{ for } t \in [0, 1) \text{ and } j \in \mathbb{Z},$$

is orthonormal and complete, i.e. an ONB. □

§6.1.6 **Properties**.

(Pythagorean formula) *If $h_1, \ldots, h_n \in \mathbb{H}$ 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 if and only if $\|h\|_{\mathbb{H}}^2 = \sum_{u\in\mathcal{U}} |\langle h, u\rangle_{\mathbb{H}}|^2$ for all $h \in \mathbb{H}$.* □

§6.1.7 **Definition**. Let $\mathcal{U}$ be a subset of a Hilbert space $(\mathbb{H}, \langle\cdot,\cdot\rangle_{\mathbb{H}})$. Denote by $\mathbb{U} := \overline{lin}(\mathcal{U})$ 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 $\mathbb{U}^\perp := \{h \in \mathbb{H} : \langle h, u\rangle_{\mathbb{H}} = 0, \forall u \in \overline{lin}(\mathcal{U})\}$ where $\mathbb{H} = \mathbb{U} \oplus \mathbb{U}^\perp$. □

§6.1.8 **Remark**. If $\mathcal{U} \subset \mathbb{H}$ is an ONS, then there exists an ONS $\mathcal{V} \subset \mathbb{H}$ such that $\mathbb{H} = \overline{lin}(\mathcal{U}) \oplus \overline{lin}(\mathcal{V})$ 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 $h \in \mathbb{H}$. □

§6.1.9 **Definition**. Given $\mathcal{J} \subset \mathbb{Z}$, 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 $\mathcal{U} = \{u_j, j \in \mathcal{J}\}$ is a complete ONS (i.e. ONB). The Hilbert space $\mathbb{H}$ is called *separable*, if there exists a complete orthonormal sequence. □

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

§6.1.11 **Definition**. Given $\mathcal{J} \subseteq \mathbb{Z}$ we call a (possibly finite) sequence $(\mathcal{J}_m)_{m\in\mathcal{M}}$, $\mathcal{M} \subseteq \mathbb{N}$, a *nested sieve in* $\mathcal{J}$, if (i) $\mathcal{J}_k \subset \mathcal{J}_m$, for any $k \leqslant m$, $k, m \in \mathcal{M}$, (ii) $|\mathcal{J}_m| < \infty$, $m \in \mathcal{M}$, and (iii) $\cup_{m\in\mathcal{M}}\mathcal{J}_m = \mathcal{J}$. We write $\mathcal{J}_m^c := \mathcal{J}\backslash\mathcal{J}_m$, $m \in \mathcal{M}$. Denoting $[\![a, b]\!] := [a, b] \cap \mathbb{Z}$ we use typically the nested sieve $([\![1, m]\!])_{m\in\mathbb{N}}$ and $([\![-m, m]\!])_{m\in\mathbb{N}}$ in $\mathcal{J} = \mathbb{N}$ and $\mathcal{J} = \mathbb{Z}$, respectively. Analogously, given an ONS $\mathcal{U} = \{u_j, j \in \mathcal{J}\}$ and setting $\mathbb{U}_m := \overline{lin}\{u_j, j \in \mathcal{J}_m\}$, $m \in \mathcal{M}$, for a nested sieve $(\mathcal{J}_m)_{m\in\mathcal{M}}$ in $\mathcal{J}$ we call the (possibly finite) sequence $(\mathbb{U}_m)_{m\in\mathcal{M}}$ *a nested sieve in* $\mathbb{U} := \overline{lin}\{u_j, j \in \mathcal{J}\}$. We write $\mathbb{U}_m^\perp := \overline{lin}\{u_j, j \in \mathcal{J}_m^c\}$ where $\mathbb{U} = \mathbb{U}_m \oplus \mathbb{U}_m^\perp$. For convenient notations we set further $\mathbb{1}_{\mathcal{J}_m} := (\mathbb{1}_{\mathcal{J}_m}(j))_{j\in\mathcal{J}}$ with $\mathbb{1}_{\mathcal{J}_m}(j) = 1$ if $j \in \mathcal{J}_m$ and $\mathbb{1}_{\mathcal{J}_m}(j) = 0$ otherwise, and analogously $\mathbb{1}_{\mathcal{J}_m^c} := (\mathbb{1}_{\mathcal{J}_m^c}(j))_{j\in\mathcal{J}}$ □

§6.1.12 **Definition**. We call an ONS $\mathcal{U} = \{u_j, j \in \mathcal{J}\}$ in L^2_μ (respectively, in ℓ^2)

(i) *regular w.r.t. the nested sieve* $(\mathcal{J}_m)_{m\in\mathcal{M}}$ *in* $\mathcal{J}$ *and the weight sequence* $\mathfrak{v}$ if there is a finite constant $\tau_{u\mathfrak{v}} \geqslant 1$ satisfying $\|\sum_{j\in\mathcal{J}_m} \mathfrak{v}_j^2|u_j|^2\|_{L^\infty_\mu} \leqslant \tau^2_{u\mathfrak{v}} \sum_{j\in\mathcal{J}_m} \mathfrak{v}_j^2$ for all $m \in \mathcal{M}$;

(ii) *regular w.r.t. the weight sequence* $\mathfrak{a}$ if there exists a finite constant $\tau_{u\mathfrak{a}} \geqslant 1$ such that $\|\sum_{j\in\mathcal{J}} \mathfrak{a}_j^2|u_j|^2\|_{L^\infty_\mu} \leqslant \tau^2_{u\mathfrak{a}}$. □

§6.1.13 **Remark**. According to Lemma 6 of Birgé and Massart [1997] assuming in L^2 a regular ONS $\{u_j, j \in \mathbb{N}\}$ w.r.t. the nested sieve $([\![1, m]\!])_{m\in\mathbb{N}}$ and $\mathfrak{v} \equiv 1$ is exactly equivalent to following property: there exists a finite constant $\tau_u \geqslant 1$ such that for any h belonging to the subspace $\mathbb{U}_m$, spanned by the first m functions $\{u_j\}_{j=1}^m$, holds $\|h\|_{L^\infty} \leqslant \tau_u\sqrt{m}\,\|h\|_{L^2}$. Typical example are bounded basis, such as the trigonometric basis, or basis satisfying the assertion, that there exists a positive constant C_∞ such that for any $(c_1, \ldots, c_m) \in \mathbb{R}^m$, $\|\sum_{j=1}^m c_j u_j\|_{L^\infty} \leqslant C_\infty\sqrt{m}|c|_\infty$ where $|c|_\infty = \max_{1\leqslant j\leqslant m} c_j$. Birgé and Massart [1997] have shown that the last property is satisfied for piece-wise polynomials, splines and wavelets. □

§6.1.14 **Example** *(§6.1.5 (i) continued)*. Consider the *trigonometric basis* $\{\psi_j, j \in \mathbb{N}\}$ in the real Hilbert space $L^2([0, 1])$. Since $\sup_{j\in\mathbb{N}}\|\psi_j\|_{L^\infty} \leqslant \sqrt{2}$ setting $\tau^2_{\psi\mathfrak{v}} := 2$ the trigonometric basis is regular w.r.t. any nested Sieve $(\mathcal{J}_m)_{m\in\mathcal{M}}$ and sequence $\mathfrak{v}$, i.e., §6.1.12 (i) holds with $\|\sum_{j\in\mathcal{J}_m} \mathfrak{v}_j^2|\psi_j|^2\|_{L^\infty} \leqslant \tau^2_{\psi\mathfrak{v}} \sum_{j\in\mathcal{J}_m} \mathfrak{v}_j^2$. In the particular case of the nested sieve $([\![1, 1 + 2m]\!])_{m\in\mathbb{N}}$ and $\mathfrak{v} \equiv 1$, we have $\sum_{j=1}^{1+2m} |\psi_j|^2 = \mathbb{1}_{[0,1]} + \sum_{j=1}^m \{2\sin^2(2\pi j\bullet) + 2\cos^2(2\pi j\bullet)\} = 1 + 2m$ and thus, the trigonometric basis is regular with $\tau^2_\psi := 1$. Moreover, the trigonometric basis is regular w.r.t. any square-summable weight sequence $\mathfrak{a}$, i.e., $\|\mathfrak{a}\|_{\ell^2} < \infty$. Indeed, in this situation we have $\|\sum_{j\in\mathbb{N}} \mathfrak{a}_j^2|\psi_j|^2\|_{\ell^\infty} \leqslant 2\,\|\mathfrak{a}\|^2_{\ell^2}$ and hence §6.1.12 (ii) holds with $\tau^2_{\psi\mathfrak{a}} = 2\,\|\mathfrak{a}\|^2_{\ell^2}$. □

§6.1.15 **Definition**. A map $T : \mathbb{H} \to \mathbb{G}$ 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 $\mathcal{N}(T)$. □

§6.1.16 **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., there is $M > 0$ such that $\|Th\|_{\mathbb{G}} \leqslant M \|h\|_{\mathbb{H}}$ for all $h \in \mathbb{H}$.* (iii) *T is uniformly continuous.* □

§6.1.17 **Definition**. The *class of all bounded linear operators* $T : \mathbb{H} \to \mathbb{G}$ is denoted by $\mathscr{L}(\mathbb{H}, \mathbb{G})$, or $\mathscr{L}$ and in case of $\mathbb{H} = \mathbb{G}$, $\mathscr{L}(\mathbb{H})$ for short. For $T \in \mathscr{L}(\mathbb{H}, \mathbb{G})$ define its *(uniform) norm* as $\|T\|_{\mathscr{L}} := \|T\|_{\mathscr{L}(\mathbb{H},\mathbb{G})} := \sup\{\|Th\|_{\mathbb{G}} ; \ \|h\|_{\mathbb{H}} \leqslant 1, h \in \mathbb{H}\}$. □

§6.1.18 **Examples**. (i) Let M be a $(m \times k)$ matrix, then $M \in \mathscr{L}(\mathbb{R}^k, \mathbb{R}^m)$. We write $\|M\|_s := \|M\|_{\mathscr{L}(\mathbb{R}^k,\mathbb{R}^m)}$ for short. *(spectral norm)*

(ii) Let $\mathcal{U} = \{u_j, j \in \mathcal{J}\}$ be an ONS in $\mathbb{H}$ and for any $f \in \mathbb{H}$ consider its *sequence of generalised Fourier coefficients* $[f] := ([f]_j)_{j\in\mathcal{J}}$ given by $[f]_j := \langle f, u_j\rangle_{\mathbb{H}}$, $j \in \mathcal{J}$. The associated *(generalised) Fourier series transform* U defined by $f \mapsto Uf := [f]$ belongs to $\mathscr{L}(\mathbb{H}, \ell^2(\mathcal{J}))$ with $\|U\|_{\mathscr{L}} = 1$.

(iii) For a sequence $\lambda = (\lambda_j)_{j\in\mathcal{J}}$ consider the *multiplication operator* $\mathrm{M}_\lambda : \mathbb{K}^{\mathcal{J}} \to \mathbb{K}^{\mathcal{J}}$ given by $x \mapsto \mathrm{M}_\lambda x := (\lambda_j x_j)_{j\in\mathcal{J}}$. For any bounded sequence λ, i.e, $\|\lambda\|_{\ell^\infty} < \infty$, we have $\|\mathrm{M}_\lambda\|_{\mathscr{L}} \leqslant \|\lambda\|_{\ell^\infty}$ and hence, $\mathrm{M}_\lambda \in \mathscr{L}(\ell^2(\mathcal{J}))$. Analogously, given a function $\lambda : \Omega \to \mathbb{K}$ the *multiplication operator* $\mathrm{M}_\lambda : \mathbb{K}^\Omega \to \mathbb{K}^\Omega$ is defined as $f \mapsto \mathrm{M}_\lambda f := f\lambda$ where for any bounded function λ holds $\|\mathrm{M}_\lambda\|_{\mathscr{L}(L^2_\mu)} \leqslant \|\lambda\|_{L^\infty_\mu} < \infty$ and, hence $\mathrm{M}_\lambda \in \mathscr{L}(L^2_\mu)$. On the other hand side, if λ is real-valued, μ-a.s. finite and non zero, then the subset $\mathcal{D}(\mathrm{M}_\lambda) := \left\{f \in L^2_\mu : \lambda f \in L^2_\mu\right\}$ is dense in L^2_μ. In this situation the *multiplication operator* $\mathrm{M}_\lambda : L^2_\mu \supset \mathcal{D}(\mathrm{M}_\lambda) \to L^2_\mu$ is densely defined (and self-adjoint). □

§6.1.19 **Definition**. A (linear) map $\Phi : \mathbb{H} \supset \mathcal{D}(\Phi) \to \mathbb{K}$ is called *(linear) functional* and given an ONS $\{u_j, j \in \mathcal{J}\}$ 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 $\Phi \in \mathscr{L}(\mathbb{H}, \mathbb{K})$ then $\mathcal{D}(\Phi) = \mathbb{H}$. □

§6.1.20 **Property**. *Let $\Phi \in \mathscr{L}(\mathbb{H}, \mathbb{K})$.*

(Fréchet-Riesz representation) There exists a function $\phi \in \mathbb{H}$ such that $\Phi(h) = \langle\phi, h\rangle_{\mathbb{H}}$ for all $h \in \mathbb{H}$, and hence, given an ONS $\{u_j, j \in \mathcal{J}\}$ in $\mathbb{H}$ we have $[\Phi]_j = [\phi]_j$ for all $j \in \mathcal{J}$. □

§6.1.21 **Example**. Consider an ONB $\mathcal{U} = \{u_j, j \in \mathcal{J}\}$ in $L^2(\Omega)$ (or analogously in $\ell^2(\mathcal{J})$). By *evaluation at a point* $t_o \in \Omega$ we mean the linear functional Φ_{t_o} mapping $h \in L^2(\Omega)$ to $h(t_o) := \Phi_{t_o}(h) = \sum_{j\in\mathcal{J}}[h]_j u_j(t_o)$. Obviously, a point evaluation of h at t_o is well-defined, if $\sum_{j\in\mathcal{J}} |[h]_j u_j(t_o)| < \infty$. Observe that the point evaluation at t_o is generally not bounded on the subset $\{h \in L^2(\Omega) : \sum_{j\in\mathcal{J}} |[h]_j u_j(t_o)| < \infty\}$. □

§6.1.22 **Definition**. 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 $\langle Th, g\rangle_{\mathbb{G}} = \langle h, T^\star g\rangle_{\mathbb{H}}$ for all $h \in \mathbb{H}, g \in \mathbb{G}$. □

§6.1.23 **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) $\|S^\star\|_{\mathscr{L}} = \|S\|_{\mathscr{L}}$, $\|SS^\star\|_{\mathscr{L}} = \|S^\star S\|_{\mathscr{L}} = \|S\|^2_{\mathscr{L}}$.

(iii) $\mathcal{N}(S) = \mathcal{R}(S^\star)^\perp$, $\mathcal{N}(S^\star) = \mathcal{R}(S)^\perp$. □

§6.1.24 **Examples**. (i) The adjoint of a $(k \times m)$ matrix M is its $(m \times k)$ transpose matrix M^t.

(ii) Let $\mathrm{M}_\lambda \in \mathscr{L}(L^2(\Omega, \mu))$ be a *multiplication operator*, then its adjoint operator $\mathrm{M}_\lambda^\star = \mathrm{M}_{\lambda^\star}$ is a multiplication operator with $\lambda^\star(t) = \overline{\lambda(t)}$, $t \in \Omega$. □

§6.1.25 **Definition**. (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) $U \in \mathscr{L}(\mathbb{H}, \mathbb{G})$ is called *unitary*, if U is invertible with $UU^\star = \mathrm{Id}_\mathbb{G}$ and $U^\star U = \mathrm{Id}_\mathbb{H}$.

(iv) $V \in \mathscr{L}(\mathbb{H}, \mathbb{G})$ is called *partial isometry*, if $V : \mathcal{N}(V)^\perp \to \mathcal{R}(V)$ is unitary.

(v) $T \in \mathscr{L}(\mathbb{H})$ is called *self-adjoint*, if $T = T^\star$, i.e., $\langle Th, g\rangle_\mathbb{H} = \langle h, T^\star g\rangle_\mathbb{H}$ for all $h, g \in \mathbb{H}$.

(vi) $T \in \mathscr{L}(\mathbb{H})$ is called *normal*, if $TT^\star = T^\star T$, i.e., $\langle Th, Tg\rangle_\mathbb{H} = \langle T^\star h, T^\star g\rangle_\mathbb{H}$ for all $h, g \in \mathbb{H}$.

(vii) A self-adjoint $T \in \mathscr{L}(\mathbb{H})$ is called *non-negative* or $T \geqslant 0$ for short, if $\langle Th, h\rangle_\mathbb{H} \geqslant 0$ for all $h \in \mathbb{H}$ and *strictly positive* or $T > 0$ for short, if $\langle Th, h\rangle_\mathbb{H} > 0$ for all $h \in \mathbb{H}\backslash\{0\}$. □

(viii) $\Pi \in \mathscr{L}(\mathbb{H})$ is called *projection* if $\Pi^2 = \Pi$. For $\Pi \neq 0$ are equivalent: (a) Π is an orthogonal projection ($\mathbb{H} = \mathcal{R}(\Pi) \oplus \mathcal{N}(\Pi)$); (b) $\|\Pi\|_\mathscr{L} = 1$; (c) Π is non-negative.

§6.1.26 **Examples** *(§6.1.18 continued)*. (i) The *(generalised) Fourier series transform* U (see §6.1.18 (ii)) is a partial isometry with adjoint operator $U^\star x = \sum_{j\in\mathcal{J}} x_j u_j$ for $x \in \ell^2(\mathcal{J})$. Moreover, the orthogonal projection $\Pi_\mathbb{U}$ onto $\mathbb{U}$ satisfies $\Pi_\mathbb{U} f = U^\star U f = \sum_{j\in\mathcal{J}} [f]_j u_j$ for all $f \in \mathbb{H}$. If $\mathcal{U} = \{u_j, j \in \mathcal{J}\}$ is complete (i.e. ONB), then U is invertible with $UU^\star = \mathrm{Id}_{\ell^2}$ and $U^\star U = \mathrm{Id}_\mathbb{H}$ due to Parseval's formula, and hence U is unitary.

(ii) A *multiplication operator* $\mathrm{M}_\lambda \in \mathscr{L}(L^2_\mu)$ (see §6.1.18 (iii)) is normal. If λ is in addition real, it is self-adjoint and if λ is non-negative, then it is non-negative. □

6.2 Abstract smoothness condition

§6.2.1 **Notations**. Let $\mathcal{U} = \{u_j, j \in \mathcal{J}\}$ be an ONS with $\mathbb{U} = \overline{lin}\,\{u_j, j \in \mathcal{J}\} \subseteq \mathbb{H}$. For $h, g \in \mathbb{H}$ we denote by $[h] := ([h]_j)_{j\in\mathcal{J}} = Uh$ the *sequence of generalised Fourier coefficients* $[h]_j := \langle h, u_j\rangle_\mathbb{H}$ and given a strictly positive sequence of weights $\mathfrak{v} = (\mathfrak{v}_j)_{j\in\mathcal{J}}$, we define $\langle h, g\rangle_\mathfrak{v}^2 := \langle \mathfrak{v}[h], \mathfrak{v}[g]\rangle_{\ell^2} = \sum_{j\in\mathcal{J}} \mathfrak{v}_j^2 [h]_j \overline{[g]}_j$ and $\|h\|_\mathfrak{v}^2 := \sum_{j\in\mathcal{J}} \mathfrak{v}_j^2 |[h]_j|^2$. Obviously, $\langle\cdot,\cdot\rangle_\mathfrak{v}$ and $\|\cdot\|_\mathfrak{v}$ restricted on $\mathbb{U}$ defines on $\mathbb{U}$ a *weighted inner product* and it induced *weighted norm*, respectively. We denote by $\mathbb{U}_\mathfrak{v}$ the completion of $\mathbb{U}$ w.r.t. $\|\cdot\|_\mathfrak{v}$. If $(u_j)_{j\in\mathcal{J}}$ is complete in $\mathbb{H}$ then let $\mathbb{H}_\mathfrak{v}$ be the completion of $\mathbb{H}$ w.r.t. $\|\cdot\|_\mathfrak{v}$. □

§6.2.2 **Example** *(§6.1.14 continued)*. Consider the real Hilbert space $L^2([0,1])$ and the *trigonometric basis* $\{\psi_j, j \in \mathbb{N}\}$. Define further a weighted norm $\|\cdot\|_\mathfrak{v}$ w.r.t. the trigonometric basis, that is, $\|h\|_\mathfrak{v} := \sum_{j\in\mathbb{N}} \mathfrak{v}_j^2 |\langle h, \psi_j\rangle_{L^2}|^2$. Denote by $L^2_\mathfrak{v}([0,1])$ or $L^2_\mathfrak{v}$ for short, the completion of $L^2([0,1])$ w.r.t. $\|\cdot\|_\mathfrak{v}$.

(P) If we set $\mathfrak{v}_1 = 1$, $\mathfrak{v}_{2k} = \mathfrak{v}_{2k+1} = j^p$, $p \in \mathbb{N}$, $k \in \mathbb{N}$, then $L^2_{\mathfrak{v}}([0,1])$ 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 $\|h\|^2_{\mathfrak{v}}$ equals the L^2-norm of its p-th weak derivative $h^{(p)}$ (Tsybakov [2009]).

(E) If, on the contrary, $\mathfrak{v}_j = \exp(-1 + j^{2p})$, $p > 1/2$, $j \in \mathbb{N}$, then $L^2_{\mathfrak{v}}([0,1])$ is a *class of analytic functions* (Kawata [1972]).

Note that, the trigonometric basis is regular w.r.t. the weight sequence $1/\mathfrak{v} = \mathfrak{v}^{-1} = (\mathfrak{v}_j^{-1})$ as in §6.1.12 (ii), i.e., $\|1/\mathfrak{v}\|_{\ell^2} < \infty$, in case **(P)** whenever $p > 1/2$ and in case **(E)** if $p > 0$. □

§6.2.3 **Definition** *(Abstract smoothness condition).* Given a strictly positive sequence of weights $\mathfrak{a} = (\mathfrak{a}_j)_{j\in\mathcal{J}}$ and an ONS $\mathcal{U} = \{u_j, j \in \mathcal{J}\}$ in $\mathbb{H}$ consider the associated weighted norm $\|\cdot\|_{1/\mathfrak{a}}$ and the completion $\mathbb{U}_{1/\mathfrak{a}}$ of $\mathbb{U}$. Let $r > 0$ be a constant. We assume in the following that the function of interest f belongs to the ellipsoid $\mathbb{F}^r_{\mathfrak{a}} := \{h \in \mathbb{U}_{1/\mathfrak{a}} : \|h\|^2_{1/\mathfrak{a}} \leqslant r^2\}$ and hence, $\Pi_{\mathbb{U}^\perp} f = 0$. □

§6.2.4 **Lemma**. *Let $\mathbb{F}^r_{\mathfrak{a}}$ be a class of functions w.r.t. an ONS $\mathcal{U} = \{u_j, j \in \mathcal{J}\}$ in L^2_μ (or analogously in ℓ^2) as given in §6.2.3. If the ONS is regular w.r.t. the weight sequence $\mathfrak{a}$ as in §6.1.12 (ii) for some finite constant $\tau_{u\mathfrak{a}} \geqslant 1$, then for each $f \in \mathbb{F}^r_{\mathfrak{a}}$ holds $\|f\|_{L^\infty_\mu} \leqslant \tau_{u\mathfrak{a}} \|f\|_{1/\mathfrak{a}} \leqslant r\tau_{u\mathfrak{a}}$.*

Proof of Lemma §6.2.4 is given in the lecture. □

§6.2.5 **Examples** *(§6.2.2 continued).* Consider $L^2_{\mathfrak{v}}([0,1])$ w.r.t. the *trigonometric basis* $\{\psi_j, j \in \mathbb{N}\}$ and a weight sequence $\mathfrak{v}$ satisfying either §6.2.2 **(P)** with $p > 1/2$ or §6.2.2 **(E)** with $p > 0$. In both cases setting $\tau^2_{\psi\mathfrak{v}} = 2\|1/\mathfrak{v}\|^2_{\ell^2} < \infty$ the trigonometric basis is regular w.r.t. the weight sequence $1/\mathfrak{v}$. Consequently, setting $\mathfrak{a} = 1/\mathfrak{v}$ from Lemma §6.2.4 follows $\sup\{\|f\|^2_{L^\infty}, f \in L^2_{1/\mathfrak{a}}([0,1])\} \leqslant 2\|f\|^2_{1/\mathfrak{a}} \|\mathfrak{a}\|^2_{\ell^2}$. □

§6.2.6 **Definition** *(Regular linear functionals).* Consider an ONS $\mathcal{U} = \{u_j, j \in \mathcal{J}\}$ in $\mathbb{H}$ which belongs to the domain $\mathcal{D}(\Phi)$ of a linear functional Φ. In order to guarantee that $\mathbb{U}_{1/\mathfrak{a}}$ and hence the class $\mathbb{F}^r_{\mathfrak{a}}$ of functions of interest as in §6.2.3 are contained in $\mathcal{D}(\Phi)$ and that $\Phi(f) = \sum_{j\in\mathcal{J}} [\Phi]_j [f]_j$ holds for all $f \in \mathbb{F}^r_{\mathfrak{a}}$, it is sufficient that $\|[\Phi]\|^2_{\mathfrak{a}} := \sum_{j\in\mathcal{J}} |[\Phi]_j|^2 \mathfrak{a}_j^2 < \infty$. Indeed, $|\Phi(f)|^2 \leqslant \|f\|^2_{1/\mathfrak{a}} \|[\Phi]\|^2_{\mathfrak{a}}$ for any $f \in \mathbb{U}_{1/\mathfrak{a}}$ and hence $\Phi \in \mathscr{L}(\mathbb{U}_{1/\mathfrak{a}}, \mathbb{K})$ with $\|\Phi\|_{\mathscr{L}} \leqslant \|[\Phi]\|_{\mathfrak{a}}$. We denote by $\mathcal{L}_{\mathfrak{a}}$ the set of all linear functionals with $\|[\Phi]\|^2_{\mathfrak{a}} < \infty$. □

§6.2.7 **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. The assumption $\Phi \in \mathcal{L}_{\mathfrak{a}}$, however, enables us in specific cases to deal with more demanding functionals, such as in §6.2.8 below the evaluation of the solution at a given point. □

§6.2.8 **Example** *(§6.1.21 continued).* Consider an ONB $\mathcal{U} = \{u_j, j \in \mathcal{J}\}$ in $L^2(\Omega)$ and the *evaluation at a point* $t_o \in \Omega$ given by $\Phi_{t_o}(h) = \sum_{j\in\mathcal{J}} [h]_j u_j(t_o)$. Consider the completion $L^2_{1/\mathfrak{a}}(\Omega)$ of $L^2(\Omega)$ w.r.t. a weighted norm $\|\cdot\|_{1/\mathfrak{a}}$ derived from $\mathcal{U}$ and a strictly positive sequence $\mathfrak{a}$. Since $|\Phi_{t_o}(h)|^2 \leqslant \|h\|^2_{1/\mathfrak{a}} \sum_{j\in\mathcal{J}} \mathfrak{a}_j^2 |u_j(t_o)|^2$ the point evaluation in t_o is bounded on $L^2_{1/\mathfrak{a}}(\Omega)$ and, thus, belongs to $\mathscr{L}(L^2_{1/\mathfrak{a}}(\Omega), \mathbb{K})$, if $\sum_{j\in\mathcal{J}} \mathfrak{a}_j^2 |u_j(t_o)|^2 < \infty$. Consequently, if the ONS $\mathcal{U}$ is regular w.r.t. the weight sequence $\mathfrak{a}$, i.e., §6.1.12 (ii) holds for some finite constant $\tau_{u\mathfrak{a}} \geqslant 1$, then $\|\Phi_{t_o}\|_{\mathscr{L}(L^2_{1/\mathfrak{a}}(\Omega),\mathbb{K})} \leqslant \tau_{u\mathfrak{a}}$ uniformly for any $t_o \in \Omega$. Revisiting the particular situation of

Example §6.2.2 and its continuation in §6.2.5, that is, $L^2_{\mathfrak{v}}([0,1])$ w.r.t. the *trigonometric basis* $\{\psi_j, j \in \mathbb{N}\}$ and weight sequence $\mathfrak{v}$ satisfying either §6.2.2 **(P)** with $p > 1/2$ or §6.2.2 **(E)** with $p > 0$, recall that the trigonometric basis is regular w.r.t. $\mathfrak{a} = 1/\mathfrak{v}$ and hence, the point evaluation Φ_{t_o} belongs to $\mathscr{L}(L^2_{1/\mathfrak{a}}([0,1]), \mathbb{R})$, i.e., $\|\Phi_{t_o}\|_{\mathscr{L}} \leqslant \sqrt{2}\,\|\mathfrak{a}\|_{\ell^2}$ for each $t_o \in [0,1]$. □

6.3 Approximation by dimension reduction

Here and subsequently, we consider a class of functions $\mathbb{F}^r_{\mathfrak{a}}$ as given in §6.2.3 w.r.t. an ONS $\{u_j, j \in \mathcal{J}\}$ in $\mathbb{H}$ and a strictly positive sequence $\mathfrak{a} = (\mathfrak{a}_j)_{j\in\mathcal{J}}$. Moreover, we assume a nested sieve $(\mathcal{J}_m)_{m\in\mathcal{M}}$ in $\mathcal{J}$ and its associated nested sieve $(\mathbb{U}_m)_{m\in\mathcal{M}}$ in $\mathbb{U}$ (see §6.1.11). For $f \in \mathbb{U}$ we consider the orthogonal projection $f_m = \Pi_{\mathbb{U}_m} f$ of f onto $\mathbb{U}_m$. Observe, that we have $f = U^\star[f]$ while $f_m = \sum_{j\in\mathcal{J}}([f]_j \mathbb{1}_{\mathcal{J}_m}(j))u_j = U^\star([f]\mathbb{1}_{\mathcal{J}_m})$ by using the sequence of indicators $\mathbb{1}_{\mathcal{J}_m} := (\mathbb{1}_{\mathcal{J}_m}(j))_{j\in\mathcal{J}}$. We shall measure the accuracy of the approximation f_m of f by its distance $\mathfrak{d}_{\text{ist}}(f_m, f)$ where $\mathfrak{d}_{\text{ist}}(\cdot,\cdot)$ is a certain semi metric. Note that in general $\mathfrak{d}_{\text{ist}}(f_m, f)$ is not monotone in m and hence we define $\text{bias}_m(f) := \sup\{\mathfrak{d}_{\text{ist}}(f, f_k), k \geqslant m, k \in \mathcal{M}\}$ as the approximation error. We are particularly interested in the following two cases.

§6.3.1 **Definition**. For $f \in \mathbb{F}_{\mathfrak{a}}$, and hence $\Pi_{\mathbb{U}^\perp} f = 0$, let $f_m = \Pi_{\mathbb{U}_m} f \in \mathbb{U}_m$ denote its orthogonal projection onto $\mathbb{U}_m$. Keep in mind that $\mathbb{U}^\perp$ and $\mathbb{U}_m^\perp$ denotes the orthogonal complement of $\mathbb{U}$ and $\mathbb{U}_m$ in $\mathbb{H}$ and $\mathbb{U}$, respectively.

(global) Given the ONS $\{u_j, j \in \mathcal{J}\}$ and a strictly positive sequence $\mathfrak{v}$ consider the completion $\mathbb{U}_{\mathfrak{v}}$ of $\mathbb{U}$ w.r.t. a weighted norm $\|\cdot\|_{\mathfrak{v}}$. If $\mathbb{F}_{\mathfrak{a}} \subset \mathbb{U}_{\mathfrak{v}}$, then $\mathfrak{d}^{\mathfrak{v}}_{\text{ist}}(h_1, h_2) := \|h_1 - h_2\|_{\mathfrak{v}}$, $h_1, h_2 \in \mathbb{U}_{\mathfrak{v}}$ defines a *global distance* on $\mathbb{U}_{\mathfrak{v}}$ and for $f \in \mathbb{F}_{\mathfrak{a}}$ we denote by $\text{bias}^{\mathfrak{v}}_m(f) := \|\Pi_{\mathbb{U}_m} f - f\|_{\mathfrak{v}} = \|\Pi_{\mathbb{U}_m^\perp} f\|_{\mathfrak{v}} = \sup\{\mathfrak{d}^{\mathfrak{v}}_{\text{ist}}(f, f_k), k \geqslant m, k \in \mathcal{M}\}$ the *global approximation error*.

(local) Let Φ be a linear functional and $\mathbb{F}_{\mathfrak{a}} \subset \mathcal{D}(\Phi)$, then $\mathfrak{d}^{\Phi}_{\text{ist}}(h_1, h_1) := |\Phi(h_1 - h_2)|$, $h_1, h_2 \in \mathcal{D}(\Phi)$, defines a *local distance* and we denote by $\text{bias}^{\Phi}_m(f) := \sup\{|\Phi(\Pi_{\mathbb{U}_k^\perp} f)|, k \geqslant m, k \in \mathcal{M}\} = \sup\{\mathfrak{d}^{\Phi}_{\text{ist}}(f, f_k), k \geqslant m, k \in \mathcal{M}\}$ the *local approximation error*. □

§6.3.2 **Lemma**. *Consider the orthogonal projection $f_m = \Pi_{\mathbb{U}_m} f \in \mathbb{U}_m$ as theoretical approximation of $f \in \mathbb{F}^r_{\mathfrak{a}}$. For each $m \in \mathcal{M}$ let $(\mathfrak{a}\mathfrak{v})_{(m)} := \|\mathfrak{a}\mathfrak{v}\mathbb{1}_{\mathcal{J}^c_m}\|_{\ell^\infty} = \sup\{\mathfrak{a}_j\mathfrak{v}_j, j \in \mathcal{J}^c_m\}$, then $\text{bias}^{\mathfrak{v}}_m(f) \leqslant r\,(\mathfrak{a}\mathfrak{v})_{(m)}$. On the other hand if $\Phi \in \mathcal{L}_{\mathfrak{a}}$ as in §6.2.6, then for each $m \in \mathcal{M}$, $\sum_{j\in\mathcal{J}^c_m} |[\Phi]_j|^2\mathfrak{a}_j^2 = \|[\Phi]\mathbb{1}_{\mathcal{J}^c_m}\|^2_{\mathfrak{a}} \leqslant \|[\Phi]\|^2_{\mathfrak{a}} < \infty$ and $(\text{bias}^{\Phi}_m(f))^2 \leqslant r^2\,\|[\Phi]\mathbb{1}_{\mathcal{J}^c_m}\|^2_{\mathfrak{a}}$.*

Proof of Lemma §6.3.2 is given in the lecture. □

§6.3.3 **Notations**. (i) For $f \in \mathbb{H}$ considering the sequence of generalised Fourier coefficients $[f]$ as in §6.2.1 introduce its sub-vector $[f]_{\underline{m}} := ([f]_j)_{j\in\mathcal{J}_m}$, where $[\Pi_{\mathbb{U}_m} f]_{\underline{m}} = [f]_{\underline{m}}$.

(ii) For $T \in \mathscr{L}(\mathbb{H})$ denote by $[T]$ the (infinite) matrix with generic entries $[T]_{k,j} := \langle u_k, Tu_j\rangle_{\mathbb{H}}$. For $m \in \mathcal{M}$, let $[T]_{\underline{m}}$ denote the $(|\mathcal{J}_m| \times |\mathcal{J}_m|)$-sub-matrix of $[T]$ given by $[T]_{\underline{m}} := ([T]_{k,j})_{j,k\in\mathcal{J}_m}$. Note that $[T^\star]_{\underline{m}} = [T]^t_{\underline{m}}$. Clearly, if we restrict $\Pi_{\mathbb{U}_m} T \Pi_{\mathbb{U}_m}$ to an operator from $\mathbb{U}_m$ to itself, then it can be represented by the matrix $[T]_{\underline{m}}$.

(iii) Given the identity $\text{Id} \in \mathscr{L}(\mathbb{H})$ the $|\mathcal{J}_m|$-dimensional identity matrix is denoted by $[\text{Id}]_{\underline{m}}$.

(iv) Consider the generalised Fourier series transform $U \in \mathscr{L}(\mathbb{H}, \ell^2(\mathcal{J}))$ as in §6.1.26 (i). Let $\text{M}_{\mathfrak{v}} : \mathbb{K}^{\mathcal{J}} \to \mathbb{K}^{\mathcal{J}}$ denote the multiplication operator $x \mapsto \text{M}_{\mathfrak{v}} x = \mathfrak{v} \cdot x$, define $\nabla_{\mathfrak{v}} :=$

$U^\star \mathrm{M}_\mathfrak{v} U : \mathbb{H} \supset \mathcal{D}(\nabla_\mathfrak{v}) \to \mathbb{H}$ and denote by $[\nabla_\mathfrak{v}]_{\underline{m}}$ the $|\mathcal{J}_m|$-dimensional diagonal matrix with diagonal entries $(\mathfrak{v}_j)_{j\in\mathcal{J}_m}$. Note that, $[\nabla_\mathfrak{v}]_{\underline{m}}^s = [\nabla_{\mathfrak{v}^s}]_{\underline{m}}$, $s \in \mathbb{R}$.

(v) Keep in mind the Euclidean norm $\|\cdot\|$ of a vector and the weighted norm $\|\cdot\|_\mathfrak{v}$ w.r.t. an ONS $\{u_j, j \in \mathcal{J}\}$ in $\mathbb{H}$. For all $f \in \mathbb{U}_m$ we have $\|f\|_\mathfrak{v}^2 = [f]_{\underline{m}}^t [\nabla_\mathfrak{v}]_{\underline{m}}^{-1} [f]_{\underline{m}} = \|[\nabla_\mathfrak{v}]_{\underline{m}}^{-1/2} [f]_{\underline{m}}\|^2$. □

6.4 Stochastic process on Hilbert spaces

Here and subsequently, $(\mathbb{H}, \langle\cdot,\cdot\rangle_\mathbb{H})$ and $\mathcal{U}$ denotes a separable Hilbert space and a subset of $\mathbb{H}$, respectively. Considering the product spaces $\mathbb{K}^\mathbb{H} = \bigtimes_{h\in\mathbb{H}} \mathbb{K}$ and $\mathbb{K}^\mathcal{U} = \bigtimes_{u\in\mathcal{U}} \mathbb{K}$ the mapping $\Pi_\mathcal{U} : \mathbb{K}^\mathbb{H} \to \mathbb{K}^\mathcal{U}$ given by $y = (y_h, h \in \mathbb{H}) \mapsto (y_u, u \in \mathcal{U}) =: \Pi_\mathcal{U} y$ is called canonical projection and for each $h \in \mathbb{H}$ in particular $\Pi_h : \mathbb{K}^\mathbb{H} \to \mathbb{K}$ given by $y = (y_{h'}, h' \in \mathbb{H}) \mapsto y_h =: \Pi_h y$ is called coordinate map. Moreover, $\mathscr{B}$ denotes the Borel-σ-algebra on $\mathbb{K}$ and $\mathbb{K}^\mathbb{H}$ is equipped with the product Borel-σ-algebra $\mathscr{B}^{\otimes\mathbb{H}} := \bigotimes_{h\in\mathbb{H}} \mathscr{B}$. Recall that $\mathscr{B}^{\otimes\mathbb{H}}$ equals the smallest σ-algebra such that all coordinate maps Π_h, $h \in \mathbb{H}$ are measurable. i.e., $\mathscr{B}^{\otimes\mathbb{H}} = \sigma(\Pi_h, h \in \mathbb{H})$.

§6.4.1 **Definition** *(Stochastic process on $\mathbb{H}$).* Let $\{Y_h, h \in \mathbb{H}\}$ be a family of $\mathbb{K}$-valued r.v.'s on a common probability space $(\Omega, \mathscr{A}, \mathbb{P})$, that is, $Y_h : \Omega \to \mathbb{K}$ is a $\mathscr{A}$-$\mathscr{B}$-measurable mapping for each $h \in \mathbb{H}$. Consider the $\mathbb{K}^\mathbb{H}$-valued r.v. $Y := (Y_h, h \in \mathbb{H})$ where $Y : \Omega \to \mathbb{K}^\mathbb{H}$ is a $\mathscr{A}$-$\mathscr{B}^{\otimes\mathbb{H}}$-measurable mapping given by $\omega \mapsto (Y_h(\omega), h \in \mathbb{H}) =: Y(\omega)$. Y is called a *stochastic process* on $\mathbb{H}$. Its *distribution* $\mathbb{P}^Y := \mathbb{P} \circ Y^{-1}$ is the image probability measure of $\mathbb{P}$ under the map Y. Further, denote by $\mathbb{P}^{\Pi_\mathcal{U} Y}$ the distribution of the stochastic process $\Pi_\mathcal{U} Y = (Y_u, u \in \mathcal{U})$ on $\mathcal{U}$. The family $\left\{\mathbb{P}^{\Pi_\mathcal{U} Y}, \mathcal{U} \subset \mathbb{H} \text{ finite}\right\}$ is called family of the finite-dimensional distributions of Y or $\mathbb{P}^Y$. In particular, $\mathbb{P}^{Y_h} := \mathbb{P}^{\Pi_h Y}$ denotes the distribution of $Y_h = \Pi_h Y$. Furthermore, we write $\mathbb{E}(Y_h)$ and $\mathbb{C}\mathrm{ov}(Y_h, Y_{h'}) := \mathbb{E}((Y_h - \mathbb{E}(Y_h))\overline{(Y_{h'} - \mathbb{E}(Y_{h'}))})$, if it is exists, for the expectation of Y_h w.r.t. $\mathbb{P}^{Y_h}$ and the covariance of Y_h and $Y_{h'}$ w.r.t. $\mathbb{P}^{\Pi_{\{h,h'\}} Y}$, respectively. □

§6.4.2 **Definition**. Let $Y := (Y_h, h \in \mathbb{H})$ be a stochastic process on $\mathbb{H}$. If $\mathbb{E}|Y_h| < \infty$ for each $h \in \mathbb{H}$ then the functional $\mu : \mathbb{H} \to \mathbb{K}$ with $h \mapsto \mathbb{E}(Y_h) =: \mu(h)$ is called *mean function* of Y. If the mean function μ is in addition linear and bounded, that is, $\mu \in \mathscr{L}(\mathbb{H}, \mathbb{K})$, then due to the Fréchet-Riesz representation theorem §6.1.20 there exists $\mu_Y \in \mathbb{H}$ such that $\mu(h) = \langle \mu_Y, h\rangle_\mathbb{H}$ for all $h \in \mathbb{H}$. The element $\mathbb{E}(Y) := \mu_Y$ is called *mean* or *expectation* of Y or $\mathbb{P}^Y$. If $\mathbb{E}|Y_h|^2 < \infty$ for each $h \in \mathbb{H}$ then the mapping $\mathrm{cov} : \mathbb{H} \times \mathbb{H} \to \mathbb{K}$ with $(h, h') \mapsto \mathbb{C}\mathrm{ov}(Y_h, Y_{h'}) =: \mathrm{cov}(h, h')$ is called *covariance function* of Y. If the covariance function cov is in addition a bounded bilinear form, then there is $\Gamma_Y \in \mathscr{L}(\mathbb{H})$ such that $\mathrm{cov}(h, h') = \langle \Gamma_Y h, h'\rangle_\mathbb{H} = \langle h, \Gamma_Y h'\rangle_\mathbb{H}$ for all $h, h' \in \mathbb{H}$. The operator Γ_Y is called *covariance operator* of Y or $\mathbb{P}^Y$. If Y admits a mean function μ and a covariance function cov then we write shortly $Y \sim \mathfrak{L}(\mu, \mathrm{cov})$. Analogously, $Y \sim \mathfrak{L}(\mu_Y, \Gamma)$ if there is an expectation $\mu_Y \in \mathbb{H}$ and a covariance operator $\Gamma_Y \in \mathscr{L}(\mathbb{H})$. □

§6.4.3 **Property**. *A covariance operator $\Gamma_Y \in \mathscr{L}(\mathbb{H})$ associated with a stochastic process Y on $\mathbb{H}$ is self-adjoint and non-negative definite.* □

§6.4.4 **Example** *(Non-parametric density estimation).* Let X be a r.v. taking its values in the interval $[0, 1]$ with distribution $\mathbb{P}$, c.d.f. $\mathbb{F}$ and admitting a Lebesgue-density $\mathbb{p} = \mathrm{d}\mathbb{P}/\mathrm{d}\lambda$ (see section 5.2). Given $h \in L_X^1$ as introduced in §6.1.3 (vi) denote by $\mathbb{E}_\mathbb{p}(h(X)) = \mathbb{P}h = \lambda(h\mathbb{p})$ the expectation of $h(X)$ w.r.t. $\mathbb{P}$. For convenience we suppose that the density $\mathbb{p}$ is square integrable, i.e., $\mathbb{p}$ belongs to the real Hilbert space $L^2 := L^2([0, 1])$ equipped with its usual inner product $\langle\cdot,\cdot\rangle_{L^2}$ (compare §6.1.3 (iv)). Thereby, for any $h \in L^2$ we have $\langle \mathbb{p}, h\rangle_{L^2} = \lambda(\mathbb{p}h) =$

$\mathbb{P}h = \mathbb{E}_{\mathbb{P}}(h(X))$. Assuming an i.i.d. sample $X_i \sim \mathbb{P}$, $i \in [\![1,n]\!]$, let $Y = (Y_h, h \in L^2)$ be the stochastic process on L^2 defined for each $h \in L^2$ by $Y_h := \overline{\mathbb{P}}_n h = \frac{1}{n}\sum_{i=1}^n h(X_i)$. Obviously, the mean function μ of Y satisfies $\mu(h) = \mathbb{E}(Y_h) = \mathbb{P}^{\otimes n}(\overline{\mathbb{P}}_n h) = \mathbb{P}h = \langle \mathbb{P}, h\rangle_{L^2}$ and hence, $Y_h = \langle \mathbb{P}, h\rangle_{L^2} + \frac{1}{\sqrt{n}}\dot{W}_h$ with $\dot{W}_h := n^{1/2}(\overline{\mathbb{P}}_n h - \mathbb{P}h)$. Moreover, the stochastic process $\dot{W} := (\dot{W}_h, h \in L^2)$ of error terms admits a covariance function given for all $h, h' \in L^2$ by $\mathbb{C}\mathrm{ov}(\dot{W}_h, \dot{W}_{h'}) = \mathbb{P}(hh') - \mathbb{P}h\mathbb{P}h' = \mathbb{P}((h-\mathbb{P}h)(h'-\mathbb{P}h')) = \mathbb{C}\mathrm{ov}(h(X), h(X'))$. Observe that $\mathbb{P}h\mathbb{P}h' = \langle \mathrm{M}_{\mathbb{P}} h, \mathbb{1}_{[0,1]}\rangle_{L^2}\langle \mathbb{1}_{[0,1]}, \mathrm{M}_{\mathbb{P}} h'\rangle_{L^2} = \langle \Pi_{\{\mathbb{1}_{[0,1]}\}}\mathrm{M}_{\mathbb{P}} h, \mathrm{M}_{\mathbb{P}} h'\rangle_{L^2}$ and $\mathbb{P}(hh') - \mathbb{P}h\mathbb{P}h' = \langle \Gamma_{\mathbb{P}} h, h'\rangle_{L^2}$ with $\Gamma_{\mathbb{P}} = \mathrm{M}_{\mathbb{P}} - \mathrm{M}_{\mathbb{P}}\Pi_{\{\mathbb{1}_{[0,1]}\}}\mathrm{M}_{\mathbb{P}}$, and thus, $\dot{W} \sim \mathfrak{L}(0, \Gamma_{\mathbb{P}})$ and consequently, $Y = \mathbb{P} + \frac{1}{n}\dot{W} \sim \mathfrak{L}(\mathbb{P}, \frac{1}{n}\Gamma_{\mathbb{P}})$. □

§6.4.5 **Example** *(Non-parametric regression).* Let (X, Z) obey a non-parametric regression model $\mathbb{E}_f(X|Z) = f(Z)$ satisfying the Assumptions §5.3.1 (see section 5.3). For convenience, in addition the regressor Z is supposed to be uniformly distributed on the interval $[0,1]$, i.e., $Z \sim \mathfrak{U}[0,1]$, and the regression function f is assumed to be square integrable, i.e., $f \in L^2 := L^2([0,1])$. Keep in mind that by Assumption §5.3.1 (ii) the centred error term $\varepsilon = X - f(Z)$ and the explanatory variable Z are independent. Given $h \in L^2$ denote by $\mathbb{E}_f(Xh(Z)) = \mathbb{P}_f[\mathrm{id}\otimes h]$ with $[\mathrm{id}\otimes h](X,Z) = Xh(Z)$ the expectation of $Xh(Z) = \{f(Z) + \varepsilon\}h(Z)$ w.r.t. the joint distribution $\mathbb{P}_f$ of (X,Z), where $\mathbb{E}_f[\varepsilon h(Z)] = 0$ and hence, $\mathbb{E}_f[Xh(Z)] = \mathbb{E}_f[f(Z)h(Z)] = \lambda(fh) = \langle f, h\rangle_{L^2}$. Assuming an i.i.d. sample (X_i, Z_i), $i \in [\![1,n]\!]$, from $\mathbb{P}_f$, let $Y = (Y_h)_{h\in L^2}$ be the stochastic process on L^2 given for each $h \in L^2$ by $Y_h := n^{-1}\sum_{i=1}^n X_i h(Z_i) = \overline{\mathbb{P}}_n[\mathrm{id}\otimes h]$. Obviously, the mean function μ of Y satisfies $\mu(h) = \mathbb{E}(Y_h) = \mathbb{E}_f[Xh(Z)] = \langle f, h\rangle_{L^2}$ and hence, $Y_h = \langle f, h\rangle_{L^2} + \frac{1}{\sqrt{n}}\dot{W}_h$ where $\dot{W}_h := n^{1/2}(\overline{\mathbb{P}}_n[\mathrm{id}\otimes h] - \mathbb{P}_f[\mathrm{id}\otimes h])$ is centred. The stochastic process $\dot{W} := (\dot{W}_h, h \in L^2)$ of error terms admits a covariance function given for all $h, h' \in L^2$ by $\mathbb{C}\mathrm{ov}(\dot{W}_h, \dot{W}_{h'}) = \mathbb{P}_f([\mathrm{id}\otimes h][\mathrm{id}\otimes h']) - \mathbb{P}_f[\mathrm{id}\otimes h]\mathbb{P}_f[\mathrm{id}\otimes h'] = \mathbb{C}\mathrm{ov}(Xh(Z), Xh'(Z)) = \sigma_\varepsilon^2\langle h, h'\rangle_{L^2} + \langle \mathrm{M}_f h, \mathrm{M}_f h'\rangle_{L^2} - \langle \Pi_{\{\mathbb{1}_{[0,1]}\}}\mathrm{M}_f h, \mathrm{M}_f h'\rangle_{L^2} = \sigma_\varepsilon^2\langle h, h'\rangle_{L^2} + \langle \mathrm{M}_f \Pi^{\perp}_{\{\mathbb{1}_{[0,1]}\}}\mathrm{M}_f h, h'\rangle_{L^2} = \langle \Gamma_f h, h'\rangle_{L^2}$ with $\Gamma_f = \sigma_\varepsilon^2\,\mathrm{Id}_{L^2} + \mathrm{M}_f\Pi^{\perp}_{\{\mathbb{1}_{[0,1]}\}}\mathrm{M}_f$, and hence, $\dot{W} \sim \mathfrak{L}(0, \Gamma_f)$ and consequently, $Y = f + \frac{1}{n}\dot{W} \sim \mathfrak{L}(f, \frac{1}{n}\Gamma_f)$. □

§6.4.6 **Definition** *(White noise process on $\mathbb{H}$).* Let $Y := (Y_h, h \in \mathbb{H})$ be a stochastic process on $\mathbb{H}$. If $\{Y_u, u \in \mathcal{U}\}$ for an ONS $\mathcal{U}$ in $\mathbb{H}$ is a family of $\mathbb{K}$-valued, independent and identically $\mathfrak{L}(0,1)$-distributed r.v.'s, i.e., $\mathbb{P}^{\Pi_{\mathcal{U}}Y} = \otimes_{u\in\mathcal{U}}\mathbb{P}^{Y_u} = \otimes_{u\in\mathcal{U}}\mathfrak{L}(0,1) = \mathfrak{L}^{\otimes\mathcal{U}}(0,1)$, where each Y_h has zero mean and variance one, then we write shortly $\Pi_{\mathcal{U}}Y \sim \mathfrak{L}^{\otimes\mathcal{U}}(0,1)$ and call $\Pi_{\mathcal{U}}Y$ a *white noise process* on $\mathcal{U}$. If $\Pi_{\mathcal{U}}Y$ for any ONS $\mathcal{U}$ is a *white noise process* on $\mathcal{U}$ then we call Y a *white noise process* on $\mathbb{H}$. □

§6.4.7 **Remark**. Considering in example §6.4.4 or §6.4.5 the centred stochastic process $\dot{W} := (\dot{W}_h, h \in L^2)$ of error terms we note that generally there does not exists an ONB $\mathcal{U}$ in L^2 such that $\Pi_{\mathcal{U}}\dot{W}$ is a white noise process on $\mathcal{U}$. □

§6.4.8 **Property**. *Let $Y := (Y_h, h \in \mathbb{H})$ be a stochastic process on $\mathbb{H}$ admitting an expectation $\mu_Y \in \mathbb{H}$ and a covariance operator $\Gamma \in \mathscr{L}(\mathbb{H})$, i.e., $Y \sim \mathfrak{L}(\mu_Y, \Gamma)$. If there exists an ONB $\mathcal{U}$ in $\mathbb{H}$ such that $\Pi_{\mathcal{U}}Y$ is a white noise process on $\mathcal{U}$, i.e., $\Pi_{\mathcal{U}}Y \sim \mathfrak{L}^{\otimes\mathcal{U}}(0,1)$. Then we have $\mu_Y = 0 \in \mathbb{H}$ and $\Gamma = \mathrm{Id}_{\mathbb{H}}$ since $\mu_Y = \sum_{u\in\mathcal{U}}\langle \mu_Y, u\rangle_{\mathbb{H}} u = \sum_{u\in\mathcal{U}}\mathbb{E}(Y_u)u = 0$ and $\langle \Gamma\cdot, \cdot\rangle_{\mathbb{H}} = \sum_{u,u'\in\mathcal{U}}\langle u, \cdot\rangle_{\mathbb{H}}\langle \Gamma u, u'\rangle_{\mathbb{H}}\overline{\langle u', \cdot\rangle_{\mathbb{H}}} = \sum_{u,u'\in\mathcal{U}}\langle u, \cdot\rangle_{\mathbb{H}}\langle u, u'\rangle_{\mathbb{H}}\overline{\langle u', \cdot\rangle_{\mathbb{H}}} = \langle \cdot, \cdot\rangle_{\mathbb{H}}$. Consequently, for each ONB $\mathcal{V}$ in $\mathbb{H}$ the r.v.'s $\{Y_v, v \in \mathcal{V}\}$ are pairwise uncorrelated.* □

§6.4.9 **Definition** *(Gaussian process on $\mathbb{H}$).* A stochastic process $Y = (Y_h, h \in \mathbb{H})$ on $\mathbb{H}$ with mean function μ and covariance function cov is called a *Gaussian process* on $\mathbb{H}$, if the family of finite-dimensional distributions $\{\mathbb{P}^{\Pi_{\mathcal{U}}Y}, \mathcal{U} \subset \mathbb{H} \text{ finite}\}$ of Y consists of normal distributions, that is, $\Pi_{\mathcal{U}}Y = (Y_u)_{u\in\mathcal{U}}$ is normally distributed with mean vector $(\mu(u))_{u\in\mathcal{U}}$ and covariance matrix $(\mathrm{cov}(u,u'))_{u,u'\in\mathcal{U}}$. We write shortly $Y \sim \mathfrak{N}(\mu, \mathrm{cov})$ or $Y \sim \mathfrak{N}(\mu_Y, \Gamma)$, if in addition there exist an expectation $\mu_Y \in \mathbb{H}$ and a covariance operator $\Gamma \in \mathscr{L}(\mathbb{H})$ associated with Y. The Gaussian process $Y \sim \mathfrak{N}(0, \mathrm{Id}_{\mathbb{H}})$ with mean $0 \in \mathbb{H}$ and covariance operator $\mathrm{Id}_{\mathbb{H}}$ is called *iso-Gaussian process* or *Gaussian white noise process* on $\mathbb{H}$. □

§6.4.10 **Property**. *Let $Y := (Y_h, h \in \mathbb{H})$ be a Gaussian process on $\mathbb{H}$ admitting an expectation $\mu_Y \in \mathbb{H}$ and a covariance operator $\Gamma \in \mathscr{L}(\mathbb{H})$, i.e., $Y \sim \mathfrak{N}(\mu_Y, \Gamma)$. If there exists an ONB $\mathcal{U}$ in $\mathbb{H}$ such that $\Pi_{\mathcal{U}}Y$ is a Gaussian white noise process on $\mathcal{U}$, i.e., $\Pi_{\mathcal{U}}Y \sim \mathfrak{N}^{\otimes\mathcal{U}}(0,1)$, then due to §6.4.8 we have $Y \sim \mathfrak{N}(0, \mathrm{Id}_{\mathbb{H}})$ and for each ONS $\mathcal{V}$ in $\mathbb{H}$ the standard normally distributed r.v.'s $\{Y_v, v \in \mathcal{V}\}$ are pairwise uncorrelated, and hence, independent, i.e., $\Pi_{\mathcal{V}}Y \sim \mathfrak{N}^{\otimes\mathcal{V}}(0,1)$.* □

§6.4.11 **Definition** *(Random function in $\mathbb{H}$).* Let $(\mathbb{H}, \langle\cdot,\cdot\rangle_{\mathbb{H}})$ be an Hilbert space equipped with its Borel-σ-algebra $\mathscr{B}_{\mathbb{H}}$, which is induced by its topology. An $\mathscr{A}$-$\mathscr{B}_{\mathbb{H}}$-measurable map $Y : \Omega \to \mathbb{H}$ is called an $\mathbb{H}$-valued r.v. or a *random function* in $\mathbb{H}$. □

§6.4.12 **Lemma**. *Let $\mathcal{U} = \{u_j, j \in \mathbb{N}\}$ be an ONS in $\mathbb{H}$. There does not exist a random function Y in $\mathbb{H}$ such that $\Pi_{\mathcal{U}}Y$ is a Gaussian white noise process on $\mathcal{U}$.*

Proof of Lemma §6.4.12 is given in the lecture. □

6.5 Statistical experiment

Given a pre-specified ONS $\mathcal{U} = \{u_j, j \in \mathcal{J}\}$ in $\mathbb{H}$ we base our estimation procedure on the expansion of the function of interest $f \in \mathbb{U} = \overline{lin}(\mathcal{U})$. The choice of an adequate ONS is determined by the presumed information on the function of interest f formalised by the abstract smoothness conditions given in §6.2.3. However, the statistical selection of a basis from a family of bases (c.f. Birgé and Massart [1997]) is complicated, and its discussion is far beyond the scope of this lecture.

§6.5.1 **Definition** *(Sequence space model (SSM)).* Let $\dot{W} = (\dot{W}_h, h \in \mathbb{H})$ be a centred stochastic process on $\mathbb{H}$ and $n \in \mathbb{N}$ be a sample size. The stochastic process $\widehat{f} = f + \frac{1}{\sqrt{n}}\dot{W}$ on $\mathbb{H}$ is called a noisy version of $f \in \mathbb{H}$. We denote by $\mathbb{P}_f^n$ the distribution of $\widehat{f}$. If $\dot{W}$ admits a covariance operator (possibly depending on f), say Γ_f, then we eventually write $\widehat{f} \sim \mathfrak{L}(f, \frac{1}{n}\Gamma_f)$ for short. To be precise, given an ONS $\mathcal{U} = \{u_j, j \in \mathcal{J}\}$ in $\mathbb{H}$ considering the family of $\mathbb{K}$-valued r.v.'s $\{[\dot{W}]_j := \dot{W}_{u_j}, j \in \mathcal{J}\}$ the observable quantities take the form

$$[\widehat{f}]_j = \langle f, u_j\rangle_{\mathbb{H}} + \tfrac{1}{\sqrt{n}}\dot{W}_{u_j} = [f]_j + \tfrac{1}{\sqrt{n}}[\dot{W}]_j, \quad j \in \mathcal{J}. \tag{6.1}$$

We denote by $\mathbb{P}_{[f]}^n$, or $\mathfrak{L}([f], \frac{1}{n}[\Gamma_f])$, the distribution of the observable stochastic process $[\widehat{f}] = ([\widehat{f}]_j)_{j\in\mathcal{J}}$ on $\mathcal{U}$ which obviously is determined by the distribution $\mathbb{P}_f^n$, or $\mathfrak{L}(f, \frac{1}{n}\Gamma_f)$, of $\widehat{f}$. The reconstruction of the sequence $[f] = ([f]_j)_{j\in\mathcal{J}}$ and whence the function $f = U^\star[f]$ from the noisy version $\widehat{f} \sim \mathbb{P}_f^n$ is called a *(direct) sequence space model (SSM)*. Given a class $\mathbb{F}_{\mathfrak{a}}^r$ of functions of interests as in §6.2.3 define the associated family of distributions $\mathbb{P}_{\mathbb{F}_{\mathfrak{a}}^r}^n := \{\mathbb{P}_f^n, f \in \mathbb{F}_{\mathfrak{a}}^r\}$, and set $\mathbb{P}_{\mathbb{F}_{\mathfrak{a}}^r} := (\mathbb{P}_{\mathbb{F}_{\mathfrak{a}}^r}^n)_{n\in\mathbb{N}}$. □

§6.5.2 **Example** *(Gaussian sequence space model (GSSM)).* Consider a Gaussian white noise process $\dot{W} = (\dot{W}_h, h \in \mathbb{H}) \sim \mathfrak{N}(0, \mathrm{Id}_{\mathbb{H}})$ on $\mathbb{H}$ as defined in §6.4.9 and a noisy version $\widehat{f} = f + \frac{1}{\sqrt{n}}\dot{W} \sim \mathfrak{N}(f, \frac{1}{n}\mathrm{Id}_{\mathbb{H}}) = \mathbb{P}_f^n$ of a function $f \in \mathbb{H}$. Considering the projection onto an ONS $\mathcal{U} = \{u_j, j \in \mathcal{J}\}$ the observable quantities take consequently the form $[\widehat{f}]_j = [f]_j + \frac{1}{\sqrt{n}}[\dot{W}]_j$, $j \in \mathcal{J}$, where the error terms $\{[\dot{W}]_j = \dot{W}_{u_j}, j \in \mathcal{J}\}$ are independent and $\mathfrak{N}(0,1)$-distributed, i.e., $[\dot{W}] = ([\dot{W}]_j)_{j\in\mathcal{J}} \sim \mathfrak{N}^{\otimes\mathcal{J}}(0,1) = \mathfrak{N}(0, \mathrm{Id}_{\mathcal{J}})$, and thus, $[\widehat{f}] = ([\widehat{f}]_j)_{j\in\mathcal{J}}$ is a sequence of independent Gaussian random variables having mean $[f]_j$ and variance n^{-1}, i.e., $[\widehat{f}] \sim \mathbb{P}_{[f]}^n = \mathfrak{N}([f], \frac{1}{n}\mathrm{Id}_{\mathcal{J}})$. The reconstruction of the sequence $[f]$ and whence the function $f = U^\star[f]$ which we assume belongs to an ellipsoid $\mathbb{F}_{\mathfrak{a}}^r$ derived from the ONS $\mathcal{U}$ and some weight sequence $(\mathfrak{a}_j)_{j\in\mathcal{J}}$ (compare §6.2.3) from $\widehat{f} \sim \mathfrak{N}(f, \frac{1}{n}\mathrm{Id}_{\mathbb{H}})$ is called a *Gaussian (direct) sequence space model (GSSM)*. The associated family of joint distributions of sequences of Gaussian random variables is denoted by $\mathfrak{N}(\mathbb{F}_{\mathfrak{a}}^r, \frac{1}{n}\mathrm{Id}_{\mathbb{H}}) := \{\mathfrak{N}(f, \frac{1}{n}\mathrm{Id}_{\mathbb{H}}), f \in \mathbb{F}_{\mathfrak{a}}^r\}$. □

§6.5.3 **Example** *(Non-parametric density estimation §6.4.4 continued).* For $n \in \mathbb{N}$ consider an i.i.d. sample $X_i \sim \mathbb{P}$, $i \in [\![1, n]\!]$, where $\mathbb{P}$ admits a Lebesgue-density $\mathbb{p} \in L^2 = L^2([0,1])$ and $\mathbb{P}^{\otimes n}$ denotes the associated joint product distribution. Consider the centred stochastic process $\dot{W} = (\dot{W}_h, h \in L^2) \sim \mathfrak{L}(0, \Gamma_{\mathbb{p}})$ of error terms with $\Gamma_{\mathbb{p}} = \mathrm{M}_{\mathbb{p}} - \mathrm{M}_{\mathbb{p}}\Pi_{\{\mathbb{1}_{[0,1]}\}}\mathrm{M}_{\mathbb{p}}$ as introduced in §6.4.4. The non-parametric estimation of a density $\mathbb{p} \in L^2$ from an i.i.d. sample of size n may thus be based on the noisy version $\widehat{\mathbb{p}} = \mathbb{p} + \frac{1}{\sqrt{n}}\dot{W} \sim \mathfrak{L}(\mathbb{p}, \frac{1}{n}\Gamma_{\mathbb{p}})$ of the density of interest $\mathbb{p}$. In other words, given a pre-specified ONS $\{u_j, j \in \mathcal{J}\}$ the observable quantity $[\widehat{\mathbb{p}}] = ([\widehat{\mathbb{p}}]_j)_{j\in\mathcal{J}} \sim \mathbb{P}_{[\mathbb{p}]}^n$ takes for each $j \in \mathcal{J}$ with $[\dot{W}]_j := \dot{W}_{u_j}$ the form $[\widehat{\mathbb{p}}]_j = [\mathbb{p}]_j + \frac{1}{\sqrt{n}}[\dot{W}]_j = \overline{\mathbb{P}}_n u_j$. Consequently, non-parametric estimation of a density can be covered by a sequence space model, where the error process $\dot{W}$, however, is generally not a white noise process. For convenient notations let $\{\mathbb{1}_{[0,1]}\} \cup \{u_j, j \in \mathbb{N}\}$ be an ONB of L^2 for some ONS $\mathcal{U} = \{u_j, j \in \mathbb{N}\}$. Keeping in mind that $\mathbb{p}$ is a density, it admits an expansion $\mathbb{p} = \mathbb{1}_{[0,1]} + U^\star[\mathbb{p}] = \mathbb{1}_{[0,1]} + \sum_{j\in\mathbb{N}}[\mathbb{p}]_j u_j$ where $[\mathbb{p}] = U\mathbb{p} = ([\mathbb{p}]_j)_{j\in\mathbb{N}}$ with $[\mathbb{p}]_j = \mathbb{E}_{\mathbb{p}}(u_j(X))$ for $j \in \mathbb{N}$ is a sequence of unknown coefficients, and hence, $f := \Pi_{\mathbb{U}}\mathbb{p} = U^\star[\mathbb{p}]$ is the function of interest. Given the pre-specified ONS $\mathcal{U}$ the observable quantity $[\widehat{\mathbb{p}}] = ([\widehat{\mathbb{p}}]_j)_{j\in\mathbb{N}} \sim \mathbb{P}_{[\mathbb{p}]}^n$ takes for each $j \in \mathbb{N}$ the form $[\widehat{\mathbb{p}}]_j = \overline{\mathbb{P}}_n u_j$. Note that the distribution $\mathbb{P}_{[\mathbb{p}]}^n$ of the observable quantity $[\widehat{\mathbb{p}}_n]$ is determined by the distribution $\mathbb{P}^{\otimes n}$ of the sample $X_1, \ldots, X_n$. Our aim is the reconstruction of the density $\mathbb{p} = \mathbb{1}_{[0,1]} + f$ assuming that $f = \Pi_{\mathbb{U}}\mathbb{p}$ belongs to an ellipsoid $\mathbb{F}_{\mathfrak{a}}^r$ derived from the ONS $\mathcal{U} = \{u_j, j \in \mathbb{N}\}$ and some weight sequence $(\mathfrak{a}_j)_{j\in\mathbb{N}}$ (compare §6.2.3). Denoting by $\mathbb{D}$ the set of all densities on $[0,1]$ let $\mathbb{D}_{\mathfrak{a}}^r := \{\mathbb{p} \in \mathbb{D} : f = \Pi_{\mathbb{U}}\mathbb{p} \in \mathbb{F}_{\mathfrak{a}}^r\}$, and the family of probability measures associated with the observations is given by $\mathbb{P}_{\mathbb{D}_{\mathfrak{a}}^r}^{\otimes n} = \{\mathbb{P}^{\otimes n}, \mathbb{p} \in \mathbb{D}_{\mathfrak{a}}^r\}$. □

§6.5.4 **Example** *(Non-parametric regression §6.4.5 continued).* Consider $(X, Z) \sim \mathbb{P}_f$ obeying $\mathbb{E}_f(X|Z) = f(Z)$ and $Z \sim \mathfrak{U}[0,1]$ with $f \in L^2 = L^2([0,1])$. Given an i.i.d.. sample $(X_i, Z_i) \sim \mathbb{P}_f$, $i \in [\![1, n]\!]$, their joint distribution is denoted by $\mathbb{P}_f^{\otimes n}$. Consider the centred stochastic process $\dot{W} = (\dot{W}_h, h \in L^2) \sim \mathfrak{L}(0, \Gamma_f)$ of error terms as introduced in §6.4.5. The non-parametric estimation of a regression function $f \in L^2$ from an i.i.d. sample of size n may thus be based on the noisy version $\widehat{f} = f + \frac{1}{\sqrt{n}}\dot{W} \sim \mathfrak{L}(f, \frac{1}{n}\Gamma_f)$ of the regression function f. In other words, given a pre-specified ONS $\{u_j, j \in \mathcal{J}\}$ the observable quantity $[\widehat{f}] = ([\widehat{f}]_j)_{j\in\mathcal{J}} \sim \mathbb{P}_{[f]}^n$ takes for each $j \in \mathcal{J}$ the form $[\widehat{f}]_j = \overline{\mathbb{P}}_n[\mathrm{id}\otimes u_j]$. Consequently, non-parametric regression can also be covered by a sequence space model, where the error process $\dot{W}$, however, is generally not a white noise process. Our aim is the reconstruction of the regression function f assuming that it

belongs to an ellipsoid $\mathbb{F}^r_{\mathfrak{a}}$ derived from an ONB $\{u_j, j \in \mathbb{N}\}$ of L^2 and some weight sequence $(\mathfrak{a}_j)_{j\in\mathbb{N}}$ (compare §6.2.3). We denote by $\mathbb{P}^{\otimes n}_{\mathbb{F}^r_{\mathfrak{a}}} = \{\mathbb{P}_f^{\otimes n}, f \in \mathbb{F}^r_{\mathfrak{a}}\}$ the family of probability measures associated with the sample (X_i, U_i), $i \in [\![1, n]\!]$. □

6.6 Orthogonal series estimation

Here and subsequently we estimate the function of interest $f \in \mathbb{H}$ using a dimension reduction. To be more precise, let $\mathcal{U} = (u_j)_{j\in\mathcal{J}}$ be an ONS in $\mathbb{H}$ and for a nested sieve $(\mathcal{J}_m)_{m\in\mathcal{M}}$ in $\mathcal{J}$ let $(\mathbb{U}_m)_{m\in\mathcal{M}}$ be its associated nested sieve in $\mathbb{U}$. For $f = U^\star[f] \in \mathbb{U}$ we consider its orthogonal projection $f_m = \Pi_{\mathbb{U}_m} f = U^\star([f]\mathbb{1}_{\mathcal{J}_m})$ onto $\mathbb{U}_m$. We assume a noisy version $\widehat{f} \sim \mathbb{P}_f^n$ obeying an sequence space model as in §6.5.1.

§6.6.1 **Definition**. Given the orthogonal projection $f_m = U^\star([f]\mathbb{1}_{\mathcal{J}_m})$ of $f = U^\star[f]$ onto $\mathbb{U}_m$ its estimator $\widehat{f}_m = U^\star([\widehat{f}]\,\mathbb{1}_{\mathcal{J}_m})$ is called *orthogonal series estimator (OSE)* of f based on an observable quantity $[\widehat{f}]$. □

We shall measure the accuracy of the OSE $\widehat{f}_m = U^\star([\widehat{f}]\,\mathbb{1}_{\mathcal{J}_m})$ of f by its mean squared distance $\mathbb{E}_f^n|\mathfrak{d}_{\text{ist}}(\widehat{f}_m, f)|^2$ w.r.t. the distribution $\mathbb{P}_f^n$ of the noisy version $\widehat{f}$ where $\mathfrak{d}_{\text{ist}}(\cdot,\cdot)$ as in §6.3.1 is a certain semi metric, to be specified below. Moreover, we call the quantity $\mathbb{E}_f^n|\mathfrak{d}_{\text{ist}}(\widehat{f}_m, f)|^2 = \mathbb{P}_f^n|\mathfrak{d}_{\text{ist}}(\widehat{f}_m, f)|^2$ risk of the estimator $\widehat{f}_m = U^\star([\widehat{f}]\,\mathbb{1}_{\mathcal{J}_m})$.

§6.6.2 **Definition**. Given a family of OSE's $\{\widehat{f}_m, m \in \mathcal{M}\}$ of a function of interest f we call a rate $(\mathcal{R}^n_{\mathfrak{d}}(f))_{n\in\mathbb{N}}$, i.e., $\mathcal{R}^n_{\mathfrak{d}} = o(1)$, a dimension parameter $(m^n_{\mathfrak{d}})_{n\in\mathbb{N}}$ and an OSE $(\widehat{f}_{m^n_{\mathfrak{d}}})_{n\in\mathbb{N}}$, respectively, *oracle rate*, *oracle dimension* and *oracle optimal* (up to a constant $C \geqslant 1$), if

$$C^{-1}\,\mathcal{R}^n_{\mathfrak{d}}(f) \leqslant \inf_{m\in\mathcal{M}} \mathbb{E}_f^n|\mathfrak{d}_{\text{ist}}(\widehat{f}_m, f)|^2 \leqslant \mathbb{E}_f^n|\mathfrak{d}_{\text{ist}}(\widehat{f}_{m^n_{\mathfrak{d}}}, f)|^2 \leqslant C\,\mathcal{R}^n_{\mathfrak{d}}(f)$$

for all $n \in \mathbb{N}$. Consequently, up to the constant C^2 the estimator $(\widehat{f}_{m^n_{\mathfrak{d}}})_{n\in\mathbb{N}}$ attains the lower risk bound within the family of OSE's, that is, $\mathbb{E}_f^n|\mathfrak{d}_{\text{ist}}(\widehat{f}_{m^n_{\mathfrak{d}}}, f)|^2 \leqslant C^2 \inf_{m\in\mathcal{M}} \mathbb{E}_f^n|\mathfrak{d}_{\text{ist}}(\widehat{f}_m, f)|^2$. □

§6.6.3 **Remark**. Consider a family of OSE's $\{\widehat{f}_m, m \in \mathcal{M}\}$ of a function of interest f. Assume that the risk of the OSE $\widehat{f}_m$ can be decomposed as follows

$$\mathbb{E}_f^n|\mathfrak{d}_{\text{ist}}(\widehat{f}_m, f)|^2 = \mathbb{E}_f^n|\mathfrak{d}_{\text{ist}}(\widehat{f}_m, f_m)|^2 + |\mathfrak{d}_{\text{ist}}(f_m, f)|^2 \tag{6.2}$$

where $\mathbb{E}_f^n|\mathfrak{d}_{\text{ist}}(\widehat{f}_m, f_m)|^2 = o(1)$ as $n \to \infty$ for each $m \in \mathcal{M}$, and $|\mathfrak{d}_{\text{ist}}(f_m, f)|^2 = o(1)$ as $m \to \infty$. Setting $\mathcal{R}^n_{\mathfrak{d}}(m, f) := \max\left(|\mathfrak{d}_{\text{ist}}(f_m, f)|^2, \mathbb{E}_f^n|\mathfrak{d}_{\text{ist}}(\widehat{f}_m, f_m)|^2\right)$ it follows that,

$$\mathcal{R}^n_{\mathfrak{d}}(m, f) \leqslant \mathbb{E}_f^n|\mathfrak{d}_{\text{ist}}(\widehat{f}_m, f)|^2 \leqslant 2\mathcal{R}^n_{\mathfrak{d}}(m, f). \tag{6.3}$$

Let us select $m^n_{\mathfrak{d}} := \arg\min\{\mathcal{R}^n_{\mathfrak{d}}(m, f), m \in \mathcal{M}\}$ and set $\mathcal{R}^n_{\mathfrak{d}}(f) := \mathcal{R}^n_{\mathfrak{d}}(m^n_{\mathfrak{d}}, f)$. We shall emphasise that $\mathcal{R}^n_{\mathfrak{d}}(f) = \min\{\mathcal{R}^n_{\mathfrak{d}}(m, f), m \in \mathcal{M}\} = o(1)$ as $n \to \infty$. Observe that for all $\delta > 0$ there exists $m_\delta \in \mathcal{M}$ and $n_\delta \in \mathbb{N}$ such that for all $n \geqslant n_\delta$ holds $|\mathfrak{d}_{\text{ist}}(f_{m_\delta}, f)|^2 \leqslant \delta$ and $\mathbb{E}_f^n|\mathfrak{d}_{\text{ist}}(\widehat{f}_{m_\delta}, f)|^2 \leqslant \delta$, and whence $\mathcal{R}^n_{\mathfrak{d}}(f) \leqslant \mathcal{R}^n_{\mathfrak{d}}(m_\delta, f) \leqslant \delta$. However, using the dimension $m^n_{\mathfrak{d}}$ it follows immediately

$$\begin{aligned}\mathcal{R}^n_{\mathfrak{d}}(f) \leqslant \inf_{m\in\mathcal{M}} \mathbb{E}_f^n|\mathfrak{d}_{\text{ist}}(\widehat{f}_m, f)|^2 \leqslant \mathbb{E}_f^n|\mathfrak{d}_{\text{ist}}(\widehat{f}_{m^n_{\mathfrak{d}}}, f)|^2 \\ \leqslant 2\mathcal{R}^n_{\mathfrak{d}}(f) \leqslant 2\inf_{m\in\mathcal{M}} \mathbb{E}_f^n|\mathfrak{d}_{\text{ist}}(\widehat{f}_m, f)|^2\end{aligned} \tag{6.4}$$

Consequently, the rate $(\mathcal{R}^n_{\mathfrak{d}}(f))_{n\in\mathbb{N}}$, the dimension parameter $(m^n_{\mathfrak{d}})_{n\in\mathbb{N}}$ and the OSE $(\widehat{f}_{m^n_{\mathfrak{d}}})_{n\in\mathbb{N}}$, respectively, is an *oracle rate*, an *oracle dimension* and *oracle optimal* (up to the constant 2). However, the dimension parameter and thus the estimator depends on the unknown function of interest f. □

§6.6.4 **Proposition**. *Consider an ONS $\mathcal{U} = (u_j)_{j\in\mathcal{J}}$ in $\mathbb{H}$ and a nested sieve $(\mathcal{J}_m)_{m\in\mathcal{M}}$ in $\mathcal{J}$. Given for each $n \in \mathbb{N}$ a noisy version $\widehat{f} \sim \mathfrak{L}(f, \frac{1}{n}\Gamma_f)$ of $f = U^\star[f] \in \mathbb{U}$ as in §6.5.1 let the associated family of OSE's be $\{\widehat{f}_m = U^\star([\widehat{f}]\,\mathbb{1}_{\mathcal{J}_m}), m \in \mathcal{M}\}$.*

(global $\mathbb{H}_{\mathfrak{v}}$-risk) Let $f \in \mathbb{H}_{\mathfrak{v}}$, i.e., $\|\mathfrak{v}[f]\|^2_{\ell^2} < \infty$. Given the sequence of variances $\mathbb{v}^2 := (\mathbb{v}^2_j = \langle u_j, \Gamma_f u_j\rangle_{\mathbb{H}})_{j\in\mathbb{N}}$ denote for all $m \in \mathcal{M}$ and $n \in \mathbb{N}$

$$\widetilde{\mathcal{R}}^n_{\mathfrak{v}}(m, f) := \max\big(\|\mathfrak{v}[f]\mathbb{1}_{\mathcal{J}^c_m}\|^2_{\ell^2}, \tfrac{1}{n}\|\mathfrak{v}\mathbb{v}\mathbb{1}_{\mathcal{J}_m}\|^2_{\ell^2}\big),$$
$$\widetilde{m}^n_{\mathfrak{v}} := \arg\min\{\widetilde{\mathcal{R}}^n_{\mathfrak{v}}(m, f), m \in \mathcal{M}\}, \quad \text{and} \quad \widetilde{\mathcal{R}}^n_{\mathfrak{v}}(f) := \widetilde{\mathcal{R}}^n_{\mathfrak{v}}(\widetilde{m}^n_{\mathfrak{v}}, f). \tag{6.5}$$

Then, $\widetilde{\mathcal{R}}^n_{\mathfrak{v}}(f) \leqslant \inf_{m\in\mathcal{M}} \mathbb{E}^n_f\|\widehat{f}_m - f\|^2_{\mathfrak{v}} \leqslant \mathbb{E}^n_f\|\widehat{f}_{\widetilde{m}^n_{\mathfrak{v}}} - f\|^2_{\mathfrak{v}} \leqslant 2\,\widetilde{\mathcal{R}}^n_{\mathfrak{v}}(f)$ for all $n \in \mathbb{N}$, i.e., the rate $(\widetilde{\mathcal{R}}^n_{\mathfrak{v}}(f))_{n\in\mathbb{N}}$, the dimension parameter $(\widetilde{m}^n_{\mathfrak{v}})_{n\in\mathbb{N}}$ and the OSE $(\widehat{f}_{\widetilde{m}^n_{\mathfrak{v}}})_{n\in\mathbb{N}}$ is an oracle rate, an oracle dimension and oracle optimal (up to the constant 2), respectively.

(local Φ-risk) Let $\|[\Phi][f]\|_{\ell^1} < \infty$, and hence $f \in \mathcal{D}(\Phi)$, where $\Phi(f) = \sum_{j\in\mathcal{J}}[\Phi]_j[f]_j$. Given the sequence of covariance matrices $\mathbb{V} := (\mathbb{V}_m = (\langle u_j, \Gamma_f u_l\rangle_{L^2})_{j,l\in\mathcal{J}_m})_{m\in\mathcal{M}}$ denote for all $m \in \mathcal{M}$ and $n \in \mathbb{N}$

$$\widetilde{\mathcal{R}}^n_{\Phi}(m, f) := \max\big(\big\|[\Phi][f]\mathbb{1}_{\mathcal{J}^c_m}\big\|^2_{\ell^1}, \tfrac{1}{n}\big\|[\Phi]_{\underline{m}}\big\|^2_{\mathbb{V}_m}\big),$$
$$\widetilde{m}^n_{\Phi} := \arg\min\{\widetilde{\mathcal{R}}^n_{\Phi}(m, f), m \in \mathcal{M}\}, \quad \text{and} \quad \widetilde{\mathcal{R}}^n_{\Phi}(f) := \widetilde{\mathcal{R}}^n_{\Phi}(\widetilde{m}^n_{\Phi}, f). \tag{6.6}$$

Then, $\widetilde{\mathcal{R}}^n_{\Phi}(f) \leqslant \inf_{m\in\mathcal{M}} \mathbb{E}^n_f|\Phi(\widehat{f}_m - f)|^2 \leqslant \mathbb{E}^n_f|\Phi(\widehat{f}_{\widetilde{m}^n_{\Phi}} - f)|^2 \leqslant 2\,\widetilde{\mathcal{R}}^n_{\Phi}(f)$ for all $n \in \mathbb{N}$, i.e., the rate $(\widetilde{\mathcal{R}}^n_{\Phi}(f))_{n\in\mathbb{N}}$, the dimension parameter $(\widetilde{m}^n_{\Phi})_{n\in\mathbb{N}}$ and the OSE $(\widehat{f}_{\widetilde{m}^n_{\Phi}})_{n\in\mathbb{N}}$ is an oracle rate, an oracle dimension and oracle optimal (up to the constant 2), respectively.

Proof of Proposition §6.6.5 is given in the lecture. □

§6.6.5 **Corollary** (GSSM, §6.5.2 continued). *Under the assumption of Proposition §6.6.4 consider for each $n \in \mathbb{N}$ a Gaussian noisy version $\widehat{f} \sim \mathfrak{N}(f, \frac{1}{n}\,\mathrm{Id}_{\mathbb{U}})$.*

(global $\mathbb{H}_{\mathfrak{v}}$-risk) Let $\|\mathfrak{v}[f]\|^2_{\ell^2} < \infty$, i.e., $f \in \mathbb{H}_{\mathfrak{v}}$. Denote for all $m \in \mathcal{M}$ and $n \in \mathbb{N}$

$$\mathcal{R}^n_{\mathfrak{v}}(m, f) := \max\big(\|\mathfrak{v}[f]\mathbb{1}_{\mathcal{J}^c_m}\|^2_{\ell^2}, \tfrac{1}{n}\|\mathfrak{v}\mathbb{1}_{\mathcal{J}_m}\|^2_{\ell^2}\big),$$
$$m^n_{\mathfrak{v}} := \arg\min\{\mathcal{R}^n_{\mathfrak{v}}(m, f), m \in \mathcal{M}\}, \quad \text{and} \quad \mathcal{R}^n_{\mathfrak{v}}(f) := \mathcal{R}^n_{\mathfrak{v}}(m^n_{\mathfrak{v}}, f). \tag{6.7}$$

Then, $\mathcal{R}^n_{\mathfrak{v}}(f) \leqslant \inf_{m\in\mathcal{M}} \mathbb{E}^n_f\|\widehat{f}_m - f\|^2_{\mathfrak{v}} \leqslant \mathbb{E}^n_f\|\widehat{f}_{m^n_{\mathfrak{v}}} - f\|^2_{\mathfrak{v}} \leqslant 2\,\mathcal{R}^n_{\mathfrak{v}}(f)$ for all $n \in \mathbb{N}$, i.e., the rate $(\mathcal{R}^n_{\mathfrak{v}}(f))_{n\in\mathbb{N}}$, the dimension parameter $(m^n_{\mathfrak{v}})_{n\in\mathbb{N}}$ and the OSE $(\widehat{f}_{m^n_{\mathfrak{v}}})_{n\in\mathbb{N}}$ is an oracle rate, an oracle dimension and oracle optimal (up to the constant 2), respectively.

(local Φ-risk) Let $\|[\Phi][f]\|_{\ell^1} < \infty$, and hence $f \in \mathcal{D}(\Phi)$, where $\Phi(f) = \sum_{j\in\mathcal{J}}[\Phi]_j[f]_j$. Denote for all $m \in \mathcal{M}$ and $n \in \mathbb{N}$

$$\mathcal{R}^n_{\Phi}(m, f) := \max\big(|\langle[\Phi], [f]\mathbb{1}_{\mathcal{J}^c_m}\rangle_{\ell^2}|^2, \tfrac{1}{n}\,\|[\Phi]\,\mathbb{1}_{\mathcal{J}_m}\|^2_{\ell^2}\big),$$
$$m^n_{\Phi} := \arg\min\{\mathcal{R}^n_{\Phi}(m, f), m \in \mathcal{M}\}, \quad \text{and} \quad \mathcal{R}^n_{\Phi}(f) := \mathcal{R}^n_{\Phi}(m^n_{\Phi}, f). \tag{6.8}$$

Then, $\mathcal{R}^n_\Phi(f) \leqslant \inf_{m\in\mathcal{M}} \mathbb{E}^n_f|\Phi(\widehat{f}_m - f)|^2 \leqslant \mathbb{E}^n_f|\Phi(\widehat{f}_{m^n_\Phi} - f)|^2 \leqslant 2\,\mathcal{R}^n_\Phi(f)$ *for all* $n \in \mathbb{N}$, *i.e., the rate* $(\mathcal{R}^n_\Phi(f))_{n\in\mathbb{N}}$, *the dimension parameter* $(m^n_\Phi)_{n\in\mathbb{N}}$ *and the OSE* $(\widehat{f}_{m^n_\Phi})_{n\in\mathbb{N}}$ *is an oracle rate, an oracle dimension and oracle optimal (up to the constant 2), respectively.*

Proof of Corollary §6.6.5 is given in the lecture. □

§6.6.6 **Corollary** (Non-parametric density estimation §6.5.3 continued). *Consider an ONB* $\{\mathbb{1}_{[0,1]}\} \cup \mathcal{U}$ *in* $L^2[0,1]$ *with* $\mathcal{U} = \{u_j, j \in \mathbb{N}\}$ *and a nested sieve* $(\mathcal{J}_m)_{m\in\mathcal{M}}$ *in* $\mathbb{N}$. *Given for each* $n \in \mathbb{N}$ *a noisy version* $\widehat{\mathbb{p}} \sim \mathfrak{L}(\mathbb{p}, \frac{1}{n}\Gamma_{\mathbb{p}})$ *with* $\Gamma_{\mathbb{p}} = \mathrm{M}_{\mathbb{p}} - \mathrm{M}_{\mathbb{p}}\Pi_{\{\mathbb{1}_{[0,1]}\}}\mathrm{M}_{\mathbb{p}}$ *as in §6.4.4 based on an i.i.d. sample* $X_i \sim \mathbb{p}$, $i \in [\![1,n]\!]$, *let* $\{\widehat{\mathbb{p}}_m = \mathbb{1}_{[0,1]} + U^\star([\widehat{\mathbb{p}}]\,\mathbb{1}_{\mathcal{J}_m}), m \in \mathcal{M}\}$ *be a family of OSE's of* $\mathbb{p} = \mathbb{1}_{[0,1]} + U^\star[\mathbb{p}] \in L^2([0,1])$.

(global $L^2_{\mathfrak{v}}$*-risk) Let* $\|\mathfrak{v}[\mathbb{p}]\|^2_{\ell^2} < \infty$, *i.e.,* $U^\star[\mathbb{p}] \in L^2_{\mathfrak{v}}$. *Given the sequence of variances* $\mathbb{v}^2 := (\mathbb{v}^2_j = \langle u_j, \Gamma_{\mathbb{p}} u_j\rangle_{L^2})_{j\in\mathbb{N}}$ *for all* $m \in \mathcal{M}$ *and* $n \in \mathbb{N}$ *consider* $\widetilde{\mathcal{R}}^n_{\mathfrak{v}}(m,f)$, $\widetilde{m}^n_{\mathfrak{v}}$, *and* $\widetilde{\mathcal{R}}^n_{\mathfrak{v}}(f)$ *as in (6.5). Then,* $\widetilde{\mathcal{R}}^n_{\mathfrak{v}}(\mathbb{p}) \leqslant \inf_{m\in\mathcal{M}} \mathbb{E}^{\otimes n}_{\mathbb{p}}\|\widehat{\mathbb{p}}_m - \mathbb{p}\|^2_{\mathfrak{v}} \leqslant \mathbb{E}^{\otimes n}_{\mathbb{p}}\|\widehat{\mathbb{p}}_{\widetilde{m}^n_{\mathfrak{v}}} - \mathbb{p}\|^2_{\mathfrak{v}} \leqslant 2\,\widetilde{\mathcal{R}}^n_{\mathfrak{v}}(\mathbb{p})$ *for all* $n \in \mathbb{N}$, *i.e., the rate* $(\widetilde{\mathcal{R}}^n_{\mathfrak{v}}(\mathbb{p}))_{n\in\mathbb{N}}$, *the dimension parameter* $(\widetilde{m}^n_{\mathfrak{v}})_{n\in\mathbb{N}}$ *and the OSE* $(\widehat{\mathbb{p}}_{\widetilde{m}^n_{\mathfrak{v}}})_{n\in\mathbb{N}}$ *is an oracle rate, an oracle dimension and oracle optimal (up to the constant 2), respectively.*

(local Φ*-risk) Let* $\|[\Phi]\,[\mathbb{p}]\,\|_{\ell^1} < \infty$, *whence* $\mathbb{p} \in \mathcal{D}(\Phi)$ *with* $\Phi(\mathbb{p}) = \Phi(\mathbb{1}_{[0,1]}) + \sum_{j\in\mathcal{J}}[\Phi]_j[\mathbb{p}]_j$. *Given the sequence of covariance matrices* $\mathbb{V} := (\mathbb{V}_m = (\langle u_j, \Gamma_{\mathbb{p}} u_l\rangle_{L^2})_{j,l\in\mathcal{J}_m})_{m\in\mathcal{M}}$ *for all* $m \in \mathcal{M}$ *and* $n \in \mathbb{N}$ *consider* $\widetilde{\mathcal{R}}^n_\Phi(m,f)$, $\widetilde{m}^n_\Phi$, *and* $\widetilde{\mathcal{R}}^n_\Phi(f)$ *as in (6.6). Then,* $\widetilde{\mathcal{R}}^n_\Phi(\mathbb{p}) \leqslant \inf_{m\in\mathcal{M}} \mathbb{E}^{\otimes n}_{\mathbb{p}}|\Phi(\widehat{\mathbb{p}}_m - \mathbb{p})|^2 \leqslant \mathbb{E}^{\otimes n}_{\mathbb{p}}|\Phi(\widehat{\mathbb{p}}_{\widetilde{m}^n_\Phi} - \mathbb{p})|^2 \leqslant 2\,\widetilde{\mathcal{R}}^n_\Phi(\mathbb{p})$ *for all* $n \in \mathbb{N}$, *i.e., the rate* $(\widetilde{\mathcal{R}}^n_\Phi(\mathbb{p}))_{n\in\mathbb{N}}$, *the dimension parameter* $(\widetilde{m}^n_\Phi)_{n\in\mathbb{N}}$ *and the OSE* $(\widehat{\mathbb{p}}_{\widetilde{m}^n_\Phi})_{n\in\mathbb{N}}$ *is an oracle rate, an oracle dimension and oracle optimal (up to the constant 2), respectively.*

Proof of Corollary §6.6.6 is given in the lecture. □

§6.6.7 **Remark**. Let the assumptions of Corollary §6.6.6 be satisfied. Interestingly, in case of a local Φ-risk if the sequence $\mathbb{V} := (\mathbb{V}_m)_{m\in\mathcal{M}}$ satisfies $\sup\{\max(\|\mathbb{V}_m\|_s, \|\mathbb{V}^{-1}_m\|_s), m \in \mathcal{M}\} \leqslant C$ for some constant $C \geqslant 1$, i.e., the smallest and the largest eigenvalue of $\mathbb{V}_m$ is uniformly bounded from below by C^{-1} and above by C, respectively, then it follows immeditately that $C^{-1}\,\|[\Phi]\,\mathbb{1}_{\mathcal{J}_m}\|^2_{\ell^2} \leqslant \big\|[\Phi]_{\underline{m}}\big\|^2_{\mathbb{V}_m} \leqslant C\,\|[\Phi]\,\mathbb{1}_{\mathcal{J}_m}\|^2_{\ell^2}$. Consequently, choosing $\mathcal{R}^n_\Phi(m,\mathbb{p})$ as in (6.8), then the associated rate $(\mathcal{R}^n_\Phi(\mathbb{p}))_{n\in\mathbb{N}}$, dimension parameter $(m^n_\Phi)_{n\in\mathbb{N}}$ and OSE $(\widehat{\mathbb{p}}_{m^n_\Phi})_{n\in\mathbb{N}}$ is also, respectively, an *oracle rate*, an *oracle dimension* and *oracle optimal* (up to the constant $2C$). On the other hand side, in case of a (global $L^2_{\mathfrak{v}}$-risk) if the sequence of variances $\mathbb{v}^2 = (\mathbb{v}^2_j)_{j\in\mathbb{N}}$ satisfy $C^{-1} \leqslant \mathbb{v}^2_j \leqslant C$ for all $j \in \mathcal{J}$ and for some constant $C \geqslant 1$, i.e., the sequence is uniformly bounded from below by C^{-1} and above by C, respectively, then it follows that $C^{-1}\|\mathfrak{v}\mathbb{1}_{\mathcal{J}_m}\|^2_{\ell^2} \leqslant \|\mathfrak{v}\mathbb{v}\mathbb{1}_{\mathcal{J}_m}\|^2_{\ell^2} \leqslant C\|\mathfrak{v}\mathbb{1}_{\mathcal{J}_m}\|^2_{\ell^2}$. Consequently, choosing $\mathcal{R}^n_{\mathfrak{v}}(m,\mathbb{p})$ as in (6.7), then the associated rate $(\mathcal{R}^n_{\mathfrak{v}}(\mathbb{p}))_{n\in\mathbb{N}}$, dimension parameter $(m^n_{\mathfrak{v}})_{n\in\mathbb{N}}$ and OSE $(\widehat{\mathbb{p}}_{m^n_{\mathfrak{v}}})_{n\in\mathbb{N}}$ is also, respectively, an *oracle rate*, an *oracle dimension* and *oracle optimal* (up to the constant $2C$). □

§6.6.8 **Lemma**. *Under the assumptions of §6.6.6 let in addition* $0 < \mathbb{p}_0^{-1} \leqslant \mathbb{p} \leqslant \mathbb{p}_0 < \infty$ λ*-a.s. for some finite constant* $\mathbb{p}_0 \geqslant 1$.

(global $L^2_{\mathfrak{v}}$*-risk) Choosing* $\mathcal{R}^n_{\mathfrak{v}}(m,\mathbb{p})$ *as in (6.7), then the associated rate* $(\mathcal{R}^n_{\mathfrak{v}}(\mathbb{p}))_{n\in\mathbb{N}}$, *dimension parameter* $(m^n_{\mathfrak{v}})_{n\in\mathbb{N}}$ *and OSE* $(\widehat{\mathbb{p}}_{m^n_{\mathfrak{v}}})_{n\in\mathbb{N}}$ *is also, respectively, an oracle rate, an oracle dimension and oracle optimal (up to the constant* $2\mathbb{p}_0$*).*

(local Φ-risk) *Choosing $\mathcal{R}^n_\Phi(m,\mathbb{p})$ as in (6.8), then the associated rate $(\mathcal{R}^n_\Phi(\mathbb{p}))_{n\in\mathbb{N}}$, dimension parameter $(m^n_\Phi)_{n\in\mathbb{N}}$ and OSE $(\widehat{\mathbb{p}}_{m^n_\mathfrak{v}})_{n\in\mathbb{N}}$ is also, respectively, an oracle rate, an oracle dimension and oracle optimal (up to the constant $2\mathbb{p}_0$).*

Proof of Lemma §6.6.8 is given in the lecture. □

§6.6.9 **Corollary** (Non-parametric regression §6.5.4 continued). *Consider an ONB $\{u_j, j\in\mathbb{N}\}$ in $L^2[0,1]$ and a nested sieve $(\mathcal{J}_m)_{m\in\mathcal{M}}$ in $\mathbb{N}$. Given for each $n\in\mathbb{N}$ a noisy version $\widehat{f}\sim\mathfrak{L}(f,\frac{1}{n}\Gamma_f)$ with $\Gamma_f=\sigma_\varepsilon^2\,\mathrm{Id}_{L^2}+\mathrm{M}_f\Pi^\perp_{\{\mathbb{1}_{[0,1]}\}}\mathrm{M}_f$ as in §6.4.5 based on an i.i.d. sample $(X_i,Z_i)\sim\mathbb{P}_f$, $i\in[\![1,n]\!]$, obeying the Assumption §5.3.1 (section 5.3) let $\{\widehat{f}_m=U^\star([\widehat{f}]\,\mathbb{1}_{\mathcal{J}_m}), m\in\mathcal{M}\}$ be a family of OSE's of $f=U^\star[f]\in L^2([0,1])$.*

(global $L^2_\mathfrak{v}$-risk) *Let $\|\mathfrak{v}[f]\|^2_{\ell^2}<\infty$, i.e., $f\in L^2_\mathfrak{v}$. Given the sequence of variances $\mathbb{v}^2:=(\mathbb{v}_j^2=\langle u_j,\Gamma_f u_j\rangle_{L^2})_{j\in\mathbb{N}}$ for all $m\in\mathcal{M}$ and $n\in\mathbb{N}$ consider $\widetilde{\mathcal{R}}^n_\mathfrak{v}(m,f)$, $\widetilde{m}^n_\mathfrak{v}$, and $\widetilde{\mathcal{R}}^n_\mathfrak{v}(f)$ as in (6.5). Then, $\widetilde{\mathcal{R}}^n_\mathfrak{v}(f)\leqslant\inf_{m\in\mathcal{M}}\mathbb{E}_f^{\otimes n}\|\widehat{f}_m-f\|^2_\mathfrak{v}\leqslant\mathbb{E}_f^{\otimes n}\|\widehat{f}_{\widetilde{m}^n_\mathfrak{v}}-f\|^2_\mathfrak{v}\leqslant 2\,\widetilde{\mathcal{R}}^n_\mathfrak{v}(f)$ for all $n\in\mathbb{N}$, i.e., the rate $(\widetilde{\mathcal{R}}^n_\mathfrak{v}(f))_{n\in\mathbb{N}}$, the dimension parameter $(\widetilde{m}^n_\mathfrak{v})_{n\in\mathbb{N}}$ and the OSE $(\widehat{f}_{\widetilde{m}^n_\mathfrak{v}})_{n\in\mathbb{N}}$ is an oracle rate, an oracle dimension and oracle optimal (up to the constant 2), respectively.*

(local Φ-risk) *Let $\|[\Phi][f]\|_{\ell^1}<\infty$, whence $f\in\mathcal{D}(\Phi)$ and $\Phi(f)=\sum_{j\in\mathcal{J}}[\Phi]_j[f]_j$. Given the sequence of covariance matrices $\mathbb{V}:=(\mathbb{V}_m=(\langle u_j,\Gamma_f u_l\rangle_{L^2})_{j,l\in\mathcal{J}_m})_{m\in\mathcal{M}}$ consider for all $m\in\mathcal{M}$ and $n\in\mathbb{N}$, $\widetilde{\mathcal{R}}^n_\Phi(m,f)$, $\widetilde{m}^n_\Phi$, and $\widetilde{\mathcal{R}}^n_\Phi(f)$ as in (6.6). Then, $\widetilde{\mathcal{R}}^n_\Phi(f)\leqslant\inf_{m\in\mathcal{M}}\mathbb{E}_f^{\otimes n}|\Phi(\widehat{f}_m-f)|^2\leqslant\mathbb{E}_f^{\otimes n}|\Phi(\widehat{f}_{\widetilde{m}^n_\Phi}-f)|^2\leqslant 2\,\widetilde{\mathcal{R}}^n_\Phi(f)$ for all $n\in\mathbb{N}$, i.e., the rate $(\widetilde{\mathcal{R}}^n_\Phi(f))_{n\in\mathbb{N}}$, the dimension parameter $(\widetilde{m}^n_\Phi)_{n\in\mathbb{N}}$ and the OSE $(\widehat{f}_{\widetilde{m}^n_\Phi})_{n\in\mathbb{N}}$ is an oracle rate, an oracle dimension and oracle optimal (up to the constant 2), respectively.*

Proof of Corollary §6.6.9 is given in the lecture. □

§6.6.10 **Remark**. Comparing Proposition §6.6.6 and §6.6.9 we obtain immediatly analogous claims as in Remark §6.6.10 replacing the density $\mathbb{p}$ by the regression function f. □

§6.6.11 **Lemma**. *Under the assumptions of §6.6.9 let in addition $\|f\|^2_{L^\infty}<\infty$ and $\sigma_\varepsilon^2>0$.*

(global $L^2_\mathfrak{v}$-risk) *Choosing $\mathcal{R}^n_\mathfrak{v}(m,f)$ as in (6.7), then the associated rate $(\mathcal{R}^n_\mathfrak{v}(f))_{n\in\mathbb{N}}$, dimension parameter $(m^n_\mathfrak{v})_{n\in\mathbb{N}}$ and OSE $(\widehat{f}_{m^n_\mathfrak{v}})_{n\in\mathbb{N}}$ is also, respectively, an oracle rate, an oracle dimension and oracle optimal (up to the constant $2\max(\sigma_\varepsilon^{-2},\sigma_\varepsilon^2+\|f\|^2_{L^\infty})$).*

(local Φ-risk) *Choosing $\mathcal{R}^n_\Phi(m,f)$ as in (6.8), then the associated rate $(\mathcal{R}^n_\Phi(f))_{n\in\mathbb{N}}$, dimension parameter $(m^n_\Phi)_{n\in\mathbb{N}}$ and OSE $(\widehat{f}_{m^n_\mathfrak{v}})_{n\in\mathbb{N}}$ is also, respectively, an oracle rate, an oracle dimension and oracle optimal (up to the constant $2\max(\sigma_\varepsilon^{-2},\sigma_\varepsilon^2+\|f\|^2_{L^\infty})$).*

Proof of Lemma §6.6.8 is given in the lecture. □

Chapter 7

Minimax optimality

7.1 Minimax theory: a general approach

For each $n \in \mathbb{N}$ suppose that the observations are distributed according to a probability measure $\mathbb{P}_f^n$ which belongs to a family of probability measures $\mathbb{P}_{\mathbb{F}}^n$. Here and subsequently, we assume that the function of interest f is identifiable, i.e., $f_1 \neq f_2$ implies $\mathbb{P}_{f_1}^n \neq \mathbb{P}_{f_2}^n$. However, in general it does not hold that $f_1 = f_2$ implies $\mathbb{P}_{f_1}^n = \mathbb{P}_{f_2}^n$. Denote by $\mathbb{E}_f^n$ the expectation w.r.t. a measure $\mathbb{P}_f^n$ in $\mathbb{P}_{\mathbb{F}}^n$ and set $\mathbb{P}_{\mathbb{F}} := (\mathbb{P}_{\mathbb{F}}^n)_{n\in\mathbb{N}}$.

§7.1.1 **Example** *(Non-parametric density estimation §6.5.3 continued).* Consider the family $\mathbb{P}_{\mathbb{D}_{\mathfrak{a}}^r}^{\otimes n} = \{\mathbb{P}^{\otimes n}, \mathbb{p} \in \mathbb{D}_{\mathfrak{a}}^r\}$. The parametrisiation using the marginal density $\mathbb{p}$ is one-to-one, since $\mathbb{p} = \mathbb{q}$ holds if and only if $\mathbb{P}^{\otimes n} = \mathbb{Q}^{\otimes n}$. □

§7.1.2 **Example** *(Non-parametric regression §6.5.4 continued).* Consider the family $\mathbb{P}_{\mathbb{F}_{\mathfrak{a}}^r}^n$ of probability measures. The regression function f is identified, i.e., from $f_1 \neq f_2$ follows $\mathbb{P}_{f_1}^n \neq \mathbb{P}_{f_2}^n$, but it is not an one-to-one parametrisation. However, if Assumption §**??** holds true and in addition the error term is $\mathfrak{N}(0, \sigma_\varepsilon^2)$-distributed with an in advanced known variance σ_ε^2, then the parametrisatiuon is one-to-one. □

Assume furthermore, that given an observable quantity with distribution $\mathbb{P}_f^n \in \mathbb{P}_{\mathbb{F}}^n$ there is an estimator of f available that takes its values in $\mathbb{H}$, but it does not necessarily belong to $\mathbb{F}$. We shall measure the accuracy of any estimator $\widetilde{f}$ of f by its distance $\mathfrak{d}_{\text{ist}}(\widetilde{f}, f)$ where $\mathfrak{d}_{\text{ist}}(\cdot,\cdot)$ as in §6.3.1 is a certain semi metric, to be specified below. Moreover, we call the quantity $\mathbb{E}_f^n|\mathfrak{d}_{\text{ist}}(\widetilde{f}, f)|^2 = \mathbb{P}_f^n|\mathfrak{d}_{\text{ist}}(\widetilde{f}, f)|^2$ risk of the estimator $\widetilde{f}$ of f.

§7.1.3 **Definition**. Given an observable quantity with probability measure $\mathbb{P}_f^n \in \mathbb{P}_{\mathbb{F}}^n$ the *maximal risk* of an estimator $\widetilde{f}$ of the function of interest f over the family $\mathbb{P}_{\mathbb{F}}^n$ is defined by

$$\mathfrak{R}_{\mathfrak{d}}\big[\widetilde{f} \,|\, \mathbb{P}_{\mathbb{F}}^n\big] := \sup\{\mathbb{E}_f^n|\mathfrak{d}_{\text{ist}}(\widetilde{f}, f)|^2,\ \mathbb{P}_f^n \in \mathbb{P}_{\mathbb{F}}^n\}.$$

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

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

§7.1.4 **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. □

§7.1.5 **Definition**. Let $\mathfrak{R}_{\mathfrak{d}}\big[\,\cdot\,|\,\mathbb{P}_{\mathbb{F}}^n\big]$ be a maximal risk over a class $\mathbb{P}_{\mathbb{F}}^n$ of probability measures $\mathbb{P}_{\mathbb{F}} = (\mathbb{P}_{\mathbb{F}}^n)_{n\in\mathbb{N}}$. If there exist an estimator $\widehat{f}$ and constants $C_- := C_-(\mathbb{P}_{\mathbb{F}})$, $C_+ := C_+(\mathbb{P}_{\mathbb{F}})$ and a rate $\mathcal{R}_{\mathfrak{d}}^n := \mathcal{R}_{\mathfrak{d}}^n(\mathbb{P}_{\mathbb{F}}^n)$, $n \in \mathbb{N}$, with $\lim_{n\to\infty} \mathcal{R}_{\mathfrak{d}}^n = 0$, depending on the sequence $\mathbb{P}_{\mathbb{F}}$ such that

(lower) the rate $(\mathcal{R}_{\mathfrak{d}}^n)_{n\in\mathbb{N}}$ is a *lower bound* up to the constant C_- of the maximal risk over all possible estimators of f, that is

$$\inf_{\widetilde{f}} \mathfrak{R}_{\mathfrak{d}}\big[\widetilde{f}\,|\,\mathbb{P}_{\mathbb{F}}^n\big]\mathbb{P}_{\mathbb{F}} \geqslant C_-\,\mathcal{R}_{\mathfrak{d}}^n, \quad \text{for all } n \in \mathbb{N},$$

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

(upper) the rate $(\mathcal{R}_{\mathfrak{d}}^n)_{n\in\mathbb{N}}$ is an *upper bound* up to the constant C_+ of the maximal risk associated with an estimator $\widehat{f}$ of f, that is

$$\mathfrak{R}_{\mathfrak{d}}\big[\widehat{f}\,|\,\mathbb{P}_{\mathbb{F}}^n\big]\mathbb{P}_{\mathbb{F}} \leqslant C_+\,\mathcal{R}_{\mathfrak{d}}^n, \quad \text{for all } n \in \mathbb{N}.$$

Then we call $(\mathcal{R}_{\mathfrak{d}}^n)_{n\in\mathbb{N}}$, or $\mathcal{R}_{\mathfrak{d}}^n$ for short, *minimax-optimal rate of convergence* and the estimator $\widehat{f}$ *minimax-optimal (up to a constant)*. □

§7.1.6 **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. □

7.2 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].

§7.2.1 **Definition**. Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures on a common measurable space $(\Omega, \mathscr{A})$, which are absolutely continuous wrt. to a σ-finite measure μ, or $\mathbb{P}, \mathbb{Q} \ll \mu$ for short. We write $\mathbb{p} := \mathrm{d}\mathbb{P}/\mathrm{d}\mu$ and $\mathbb{q} := \mathrm{d}\mathbb{Q}/\mathrm{d}\mu$.

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

$$KL(\mathbb{P},\mathbb{Q}) := \begin{cases} \mathbb{P}\log(\mathbb{p}/\mathbb{q}), & \text{if } \mathbb{P} \ll \mathbb{Q}; \\ \infty, & \text{otherwise.} \end{cases}$$

(ii) The *Hellinger distance* between $\mathbb{P}$ and $\mathbb{Q}$ is defined by

$$H(\mathbb{P},\mathbb{Q}) := (\mu(\sqrt{\mathbb{p}} - \sqrt{\mathbb{q}})^2)^{1/2} = \big\|\sqrt{\mathbb{p}} - \sqrt{\mathbb{q}}\big\|_{L_\mu^2}$$

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

(iii) The *Hellinger affinity* is given by

$$\rho(\mathbb{P},\mathbb{Q}) := \mu(\sqrt{\mathbb{p}}\sqrt{\mathbb{q}}) = \langle\sqrt{\mathbb{p}}, \sqrt{\mathbb{q}}\rangle_{L_\mu^2}$$

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

Proof of Lemma §7.2.2 is given in the lecture. □

§7.2.3 **Lemma**. *Let $\widetilde{f}$ be an estimator and, let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures. For all $f_1, f_2 \in \mathbb{F}$ we have*

$$\mathbb{P}(|\mathfrak{d}_{ist}(\widetilde{f}, f_1)|^2) + \mathbb{Q}(|\mathfrak{d}_{ist}(\widetilde{f}, f_2)|^2) \geqslant \frac{1}{2}\,|\mathfrak{d}_{ist}(f_1, f_2)|^2\,\rho^2(\mathbb{P}, \mathbb{Q}). \tag{7.1}$$

Proof of Lemma §7.2.3 is given in the lecture. □

7.3 Lower bound based on two hypothesis

§7.3.1 **Lemma** (Lower bound based on two hypothesis). *Consider a family of probability measures $\mathbb{P}^n_{\mathbb{F}}$. For probability measures $\mathbb{P}^n_{f_1}$ and $\mathbb{P}^n_{f_2}$ in $\mathbb{P}^n_{\mathbb{F}}$ with Hellinger affinitiy $\rho^2(\mathbb{P}^n_{f_1}, \mathbb{P}^n_{f_2})$ holds*

$$\inf_{\widetilde{f}} \mathfrak{R}_{\mathfrak{d}}\big[\widetilde{f}\,|\,\mathbb{P}^n_{\mathbb{F}}\big] \geqslant \frac{1}{4}|\mathfrak{d}_{ist}(f_1, f_2)|^2\rho^2(\mathbb{P}^n_{f_1}, \mathbb{P}^n_{f_2}). \tag{7.2}$$

Proof of Lemma §7.3.1 is given in the lecture. □

§7.3.2 **Remark** *(Statistically indistinguishable).* On the one hand if $\mathbb{P}^n_{f_1}$ and $\mathbb{P}^n_{f_2}$ are statistically indistinguishable in the sense that $H(\mathbb{P}^n_{f_1}, \mathbb{P}^n_{f_2}) \leqslant 1$, then using the relationship §7.2.2 (b) we bound the Hellinger affinity from below by $\rho(\mathbb{P}^n_{f_1}, \mathbb{P}^n_{f_2}) \geqslant 1/2$, and whence due to Lemma §7.3.1 (7.2) we have

$$\inf_{\widetilde{f}} \mathfrak{R}_{\mathfrak{d}}\big[\widetilde{f}\,|\,\mathbb{P}^n_{\mathbb{F}}\big] \geqslant \frac{1}{16}|\mathfrak{d}_{\text{ist}}(f_1, f_2)|^2. \tag{7.3}$$

On the other hand if two poduct measures $\mathbb{P}^{\otimes n}_{f_1}$ and $\mathbb{P}^{\otimes n}_{f_2}$ are statistically indistinguishable in the sense that $H^2(\mathbb{P}_{f_1}, \mathbb{P}_{f_2}) \leqslant 2/n$, then using the independence, i.e., $\rho(\mathbb{P}^{\otimes n}_{f_1}, \mathbb{P}^{\otimes n}_{f_2}) = \rho(\mathbb{P}_{f_1}, \mathbb{P}_{f_2})^n$ togehter with the relationship §7.2.2 (b) it follows $\rho(\mathbb{P}^{\otimes n}_{f_1}, \mathbb{P}^{\otimes n}_{f_2}) \geqslant (1 - n^{-1})^n \geqslant 1/4$ for all $n \geqslant 2$, and whence

$$\inf_{\widetilde{f}} \mathfrak{R}_{\mathfrak{d}}\big[\widetilde{f}\,|\,\mathbb{P}^{\otimes n}_{\mathbb{F}}\big] \geqslant \frac{1}{64}|\mathfrak{d}_{\text{ist}}(f_1, f_2)|^2. \tag{7.4}$$

□

§7.3.3 **Remark** *(Lower bound for a maximal Φ-risk).* Let the class of functions of interest be an ellipsoid $\mathbb{F}^r_{\mathfrak{a}}$. Consider a local Φ-risk associated with a regular linear functional Φ (see §6.2.6). If we consider furthermore candidates $f_1 := f_*$ and $f_2 = -f_*$ for some $f_* \in \mathbb{F}^r_{\mathfrak{a}}$, then trivially $f_1, f_2 \in \mathbb{F}^r_{\mathfrak{a}}$ and $|\mathfrak{d}_{\text{ist}}(f_1, f_2)|^2 = |\Phi(f_1 - f_2)|^2 = 4|\Phi(f_*)|^2$. On the one hand if in addition $\mathbb{P}^n_{f_*}$ and $\mathbb{P}^n_{-f_*}$ are statistically indistinguishable in the sense that $H(\mathbb{P}^n_{f_*}, \mathbb{P}^n_{-f_*}) \leqslant 1$, then due to (7.3) in Remark §7.3.2 it holds

$$\inf_{\widetilde{f}} \mathfrak{R}_{\Phi}\big[\widetilde{f}\,|\,\mathbb{P}^n_{\mathbb{F}}\big] \geqslant \tfrac{1}{4}|\Phi(f_*)|^2 \quad \text{for all } n \geqslant 1. \tag{7.5}$$

On the other hand if two poduct measures $\mathbb{P}^{\otimes n}_{f_*}$ and $\mathbb{P}^{\otimes n}_{-f_*}$ are statistically indistinguishable in the sense that $H^2(\mathbb{P}_{f_*}, \mathbb{P}_{-f_*}) \leqslant 2/n$, then from (7.4) in Remark §7.3.2 it follows

$$\inf_{\widetilde{f}} \mathfrak{R}_{\mathfrak{d}}\big[\widetilde{f}\,|\,\mathbb{P}^{\otimes n}_{\mathbb{F}}\big] \geqslant \frac{1}{16}|\Phi(f_*)|^2 \quad \text{for all } n \geqslant 2. \tag{7.6}$$

However, often a minimax-optimal lower bound can be found by constructing a candidate $f_* \in \mathbb{F}^r_{\mathfrak{a}}$ that has the largest possible $|\Phi(f_*)|$-value but $\mathbb{P}^n_{f_*}$ and $\mathbb{P}^n_{-f_*}$ are still statistically indistinguishable. □

7.3.1 Examples - lower bound of a maximal Φ-risk

Assuming that the function of interest f with generalised Fourier coefficients $[f] = ([f]_j)_{j\in\mathcal{J}}$ belongs to the class of solutions $\mathbb{F}^r_{\mathfrak{a}}$ as in §6.2.3 we derive below a lower bound of a maximal Φ-risk considering the three examples: (i) Gaussian sequence space model (GSSM) §6.5.2, (ii) non-parametric regression §7.1.2, and (iii) density estimation §7.1.1. Let Φ be a regular linear functional belonging to the class $\mathcal{L}_{\mathfrak{a}}$ as given in §6.2.6 and define in analogy to (6.8) for all $n \in \mathbb{N}$ and $m \in \mathcal{M}$,

$$\mathcal{R}^n_\Phi(m,\mathfrak{a}) := \max\big(\|[\Phi]\,\mathfrak{a}\mathbb{1}_{\mathcal{J}^c_m}\|^2_{\ell^2}, \max(\mathfrak{a}^2_{(m)}, n^{-1})\|[\Phi]\,\mathbb{1}_{\mathcal{J}_m}\|^2_{\ell^2}\big),$$
$$m^n_\Phi := \arg\min\{\mathcal{R}^n_\Phi(m,\mathfrak{a}), m\in\mathcal{M}\} \text{ and } \mathcal{R}^n_\Phi(\mathfrak{a}) := \mathcal{R}^n_\Phi(m^n_\Phi,\mathfrak{a}). \quad (7.7)$$

Keep in mind the quantities $\mathcal{R}^n_\Phi(m,f)$ and $\mathcal{R}^n_\Phi(f)$ given in (6.8), for any $f\in\mathbb{F}^r_{\mathfrak{a}}$ by applying the Cauchy-Schwarz inequaltiy we have for all $m\in\mathcal{M}$, $|\langle[\Phi],[f]\mathbb{1}_{\mathcal{J}^c_m}\rangle_{\ell^2}|^2 \leqslant r^2\|[\Phi]\,\mathfrak{a}\mathbb{1}_{\mathcal{J}^c_m}\|^2_{\ell^2}$ and hence, $\mathcal{R}^n_\Phi(m,f)\leqslant(1\vee r^2)\mathcal{R}^n_\Phi(m,\mathfrak{a})$ for all $n\in\mathbb{N}$. Consequently, $\mathcal{R}^n_\Phi(f)\leqslant(1\vee r^2)\mathcal{R}^n_\Phi(\mathfrak{a})$ where $\mathcal{R}^n_\Phi(f)$ is eventually the oracle rate (see, for instance, Proposition §6.6.5). We show below that $\mathcal{R}^n_\Phi(\mathfrak{a})$ eventually is a minimax rate. We impose a minimal regularity of the linear functional Φ and the weight sequence $\mathfrak{a}$, which is formalised in the next assumption.

§7.3.4 **Assumption**. Consider a pre-specified ONS $\{u_j, j\in\mathcal{J}\}$ in $\mathbb{H}$, a nested sieve $(\mathcal{J}_m)_{m\in\mathcal{M}}$ in $\mathcal{J}$ and a strictly positive, monotonically non-increasing sequence $\mathfrak{a} = (\mathfrak{a}_j)_{j\in\mathcal{J}}$, that is, $\min\{\mathfrak{a}_j, j\in\mathcal{J}_m,\} \geqslant \sup\{\mathfrak{a}_j, j\in\mathcal{J}^c_m\} =: \mathfrak{a}_{(m)} > 0$ for any $m\in\mathcal{M}$. Suppose further that $\Phi\in\mathcal{L}_{\mathfrak{a}}$ such that $\eta := \inf\Big\{\min(n^{-1}\mathfrak{a}^{-2}_{(m^n_\Phi)}, n\mathfrak{a}^2_{(m^n_\Phi)}), n\in\mathbb{N}\Big\} > 0$.

In the proof of the next propositions we intend to apply the result presented in (7.5) or (7.6) in Remark §7.3.3 to two special choices of $f_\star\in\mathbb{F}^r_{\mathfrak{a}}$ which we specify next.

§7.3.5 **Lemma**. *Consider η as in Assumption §7.3.4 and for $n\in\mathbb{N}$ let $m_* := m^n_\Phi$ as in (7.7). Define $K_* := \max(\mathfrak{a}^2_{(m_*)}, n^{-1})$, and $\zeta := \eta\min(r^2,c)$ for some $c>0$. Consider either the function (i) $f_* := (\zeta\alpha_*)^{1/2}\sum_{j\in\mathcal{J}_{m_*}}[\Phi]_j u_j$ with $\alpha_* := K_*\|[\Phi]\,\mathbb{1}_{\mathcal{J}_{m_*}}\|^{-2}_{\ell^2}$, or the function (ii) $f_* := (\zeta\alpha_*)^{1/2}\sum_{j\in\mathcal{J}^c_{m_*}}[\Phi]_j\mathfrak{a}^2_j u_j$ with $\alpha_* := \|[\Phi]\,\mathfrak{a}\mathbb{1}_{\mathcal{J}^c_{m_*}}\|^{-2}_{\ell^2}$. In both cases we have $\|f_*\|^2_{1/\mathfrak{a}}\leqslant\min(r^2,c)$, i.e., $f_\star\in\mathbb{F}^r_{\mathfrak{a}}$, and $n\,\|f_\star\|^2_{\mathbb{H}}\leqslant c$.*

Proof of Lemma §7.3.5 is given in the lecture. □

§7.3.6 **Corollary**. *Let the assumptions of Lemma §7.3.5 be satisfied. If the ONS $\mathcal{U}$ is in addition regular w.r.t. the weigth sequence $\mathfrak{a}$ as in §6.1.12 (ii), i.e., $\|\sum_{j\in\mathcal{J}}\mathfrak{a}^2_j|u_j|^2\|_{L^\infty}\leqslant\tau^2_{u\mathfrak{a}}$ for some $\tau_{u\mathfrak{a}}\geqslant 1$, then it holds $\|f_*\|^2_{L^\infty}\leqslant\tau^2_{u\mathfrak{a}}c$.*

Proof of Corollary §7.3.6 is given in the lecture. □

§7.3.7 **Proposition** (GSSM, §6.5.2 continued). *Consider the reconstruction of $f=\mathcal{U}^\star[f]\in\mathbb{F}^r_{\mathfrak{a}}$ given for each $n\in\mathbb{N}$ an observable quantity $\widehat{f}\wr\mathfrak{N}(\mathbb{F}^r_{\mathfrak{a}},\frac{1}{n}\mathrm{Id})$. Under Assumption §7.3.4 holds*

$$\inf_{\widetilde{f}}\mathfrak{R}_\Phi\big[\widetilde{f}\,|\,\mathfrak{N}(\mathbb{F}^r_{\mathfrak{a}},\tfrac{1}{n}\mathrm{Id})\big] \geqslant \frac{\eta}{8}\,\min(2r^2,1)\,\mathcal{R}^n_\Phi(\mathfrak{a}), \quad \textit{for all } n\in\mathbb{N}. \quad (7.8)$$

Proof of Proposition §7.3.7 is given in the lecture. □

§7.3.8 **Corollary**. *Under the assumptions of Proposition §7.3.7 for each $n \in \mathbb{N}$ consider the OSE $\widehat{f}_{m_\Phi^n} = U^\star([\widehat{f}]\,\mathbb{1}_{\mathcal{J}_{m_\Phi^n}})$ with m_Φ^n as in (7.7). Then, $\mathfrak{R}_\Phi\big[\widehat{f}_{m_\Phi^n} \,|\, \mathfrak{N}(\mathbb{F}_\mathfrak{a}^r, \frac{1}{n}\mathrm{Id}_\mathbb{H})\big] \leqslant (1+r^2)\,\mathcal{R}_\Phi^n(\mathfrak{a})$ for all $n \in \mathbb{N}$, i.e., the rate $(\mathcal{R}_\Phi^n(\mathfrak{a}))_{n\in\mathbb{N}}$ and the OSE $(\widehat{f}_{m_\Phi^n})_{n\in\mathbb{N}}$ are minimax-optimal (up to a constant).*

Proof of Corollary §7.3.8 is given in the lecture. □

§7.3.9 **Proposition** (Non-parametric density estimation §6.5.3 continued). *Let $\{\mathbb{1}_{[0,1]}\} \cup \mathcal{U}$ be an ONB in $L^2[0,1]$ with $\mathcal{U} = \{u_j, j \in \mathbb{N}\}$. Consider the reconstruction of a density $\mathbb{p} = \mathbb{1}_{[0,1]} + U^\star[f]$ in $\mathbb{D}_\mathfrak{a}^r$, i.e., $f \in \mathbb{F}_\mathfrak{a}^r$, given for each $n \in \mathbb{N}$ an i.i.d. sample $(X_1, \ldots, X_n) \backsim \mathbb{P}_{\mathbb{D}_\mathfrak{a}^r}^{\otimes n}$. Let the ONS $\mathcal{U}$ be in addition regular w.r.t. the weigth sequence $\mathfrak{a}$ as in §6.1.12 (ii), i.e., $\|\sum_{j\in\mathcal{J}} \mathfrak{a}_j^2 |u_j|^2\|_{L^\infty} \leqslant \tau_{u\mathfrak{a}}^2$ for some $\tau_{u\mathfrak{a}} \geqslant 1$. Under Assumption §7.3.4 we have*

$$\inf_{\widetilde{\mathbb{p}}} \mathfrak{R}_\Phi\big[\widetilde{\mathbb{p}} \,|\, \mathbb{P}_{\mathbb{D}_\mathfrak{a}^r}^{\otimes n}\big] \geqslant \frac{\eta}{16}\, \min(r^2, (4\tau_{u\mathfrak{a}}^2)^{-1})\, \mathcal{R}_\Phi^n(\mathfrak{a}), \quad \text{for all } n \geqslant 2. \tag{7.9}$$

Proof of Proposition §7.3.9 is given in the lecture. □

§7.3.10 **Remark**. We shall emphasise that assuming in Proposition §7.3.9 in addition a regular ONS $\mathcal{U}$ ensures that the specific choices $\mathbb{p}_* = \mathbb{1}_{[0,1]} + f_*$ with f_* as specified in Lemma §7.3.5 are indeed densities belonging to $\mathbb{D}_\mathfrak{a}^r$. Moreover, the specific choices satisfy $1/2 \leqslant \mathbb{p}_* \leqslant 1 + 1/2 \leqslant 2$, λ-a.s.. Thereby, due to Lemma §6.6.8 $\mathcal{R}_\Phi^n(\mathbb{p}_*)$ as given in (6.8) is an oracle rate, where $\mathcal{R}_\Phi^n(\mathbb{p}_*) \leqslant (1 \vee r^2)\mathcal{R}_\Phi^n(\mathfrak{a})$, and $\mathcal{R}_\Phi^n(\mathfrak{a})$ is a minimax-rate as formalised next. □

§7.3.11 **Corollary**. *Under the assumptions of Proposition §7.3.9 for each $n \in \mathbb{N}$ consider a noisy version $\widehat{\mathbb{p}} \sim \mathfrak{L}(\mathbb{p}, \frac{1}{n}\Gamma_\mathbb{p})$ with $\Gamma_\mathbb{p} = \mathrm{M}_\mathbb{p} - \mathrm{M}_\mathbb{p}\Pi_{\{\mathbb{1}_{[0,1]}\}}\mathrm{M}_\mathbb{p}$ as in §6.4.4 and an observable quantity $[\widehat{\mathbb{p}}] = (\overline{\mathbb{P}}_n u_j)_{j\in\mathbb{N}}$ using an i.i.d. sample $(X_1, \ldots, X_n) \backsim \mathbb{P}_{\mathbb{D}_\mathfrak{a}^r}^{\otimes n}$. Let $\widehat{\mathbb{p}}_{m_\Phi^n} = \mathbb{1}_{[0,1]} + U^\star([\widehat{\mathbb{p}}]\,\mathbb{1}_{\mathcal{J}_{m_\Phi^n}})$ be the OSE with m_Φ^n as in (7.7). Then, $\mathfrak{R}_\Phi\big[\widehat{\mathbb{p}}_{m_\Phi^n} \,|\, \mathbb{P}_{\mathbb{D}_\mathfrak{a}^r}^{\otimes n}\big] \leqslant (1 + r\tau_{u\mathfrak{a}} + r^2)\,\mathcal{R}_\Phi^n(\mathfrak{a})$ for all $n \in \mathbb{N}$, i.e., the rate $(\mathcal{R}_\Phi^n(\mathfrak{a}))_{n\in\mathbb{N}}$ and the OSE $(\widehat{\mathbb{p}}_{m_\Phi^n})_{n\in\mathbb{N}}$ are minimax-optimal (up to a constant).*

Proof of Corollary §7.3.11 is given in the lecture. □

§7.3.12 **Proposition** (Non-parametric regression §6.5.4 continued). *Consider the reconstruction of a regression function $f \in \mathbb{F}_\mathfrak{a}^r$ given for each $n \in \mathbb{N}$ an i.i.d. sample $(X_1, Z_1), \ldots, (X_n, Z_n)$ with joint distribution belonging to $\mathbb{P}_{\mathbb{F}_\mathfrak{a}^r}^{\otimes n}$ and obeying the Assumption §5.3.1 (see section 5.3). If the error term is $\mathfrak{N}(0, \sigma_\varepsilon^2)$-distributed, then under Assumption §7.3.4 we have*

$$\inf_{\widetilde{f}} \mathfrak{R}_\Phi\big[\widetilde{f} \,|\, \mathbb{P}_{\mathbb{F}_\mathfrak{a}^r}^{\otimes n}\big] \geqslant \frac{\eta}{8}\, \min(2r^2, \sigma_\varepsilon^2)\, \mathcal{R}_\Phi^n(\mathfrak{a}), \quad \text{for all } n \in \mathbb{N}. \tag{7.10}$$

Proof of Proposition §7.3.12 is given in the lecture. □

§7.3.13 **Remark**. We shall emphasise that assuming in Proposition §7.3.12 in addition normal-distributed error terms is only needed to simplify the calculation of the distance between distributions corresponding to different regression functions. On the other hand, below we derive an upper bound under Assumption §5.3.1 (see section 5.3) only. In this situation, Proposition §7.3.12 obviously provides a lower bound for any estimator since the family $\mathbb{P}_{\mathbb{F}_\mathfrak{a}^r}^{\otimes n}$ contains this specific Gaussian-error case. Moreover, if the ONS $\mathcal{U}$ is in addition regular w.r.t. the weight sequence $\mathfrak{a}$, then the specific choice f_* given by Lemma §7.4.4 satisfies $\|f_*\|_{L^\infty}^2 < \infty$ and due

to Lemma §6.6.11 $\mathcal{R}^n_\Phi(f_*)$ as given in (6.8) is an oracle rate, where $\mathcal{R}^n_\Phi(f_*) \leqslant (1 \vee r^2)\mathcal{R}^n_\Phi(\mathfrak{a})$, and $\mathcal{R}^n_\Phi(\mathfrak{a})$ is a minimax-rate as formalised next. □

§7.3.14 **Corollary**. *For each $n \in \mathbb{N}$ using an i.i.d. sample $(X_1, Z_1), \dots, (X_n, Z_n) \circledcirc \mathbb{P}^{\otimes n}_{\mathbb{F}^r_\mathfrak{a}}$ obeying the Assumption §5.3.1 (see section 5.3) consider a noisy version $\widehat{f} \sim \mathfrak{L}(f, \frac{1}{n}\Gamma_f)$ with $\Gamma_f = \sigma^2_\varepsilon \operatorname{Id}_\mathbb{H} + \mathrm{M}_{|f|^2}$ as in §6.5.4 and an observable quantity $[\widehat{f}] = (\overline{\mathbb{P}}_n[\mathrm{id} \otimes u_j])_{j\in\mathbb{N}}$. Let the ONS $\mathcal{U}$ be in addition regular w.r.t. the sequence $\mathfrak{a}$ as in §6.1.12 (ii), i.e., $\|\sum_{j\in\mathcal{J}} \mathfrak{a}_j^2 |u_j|^2\|_{L^\infty} \leqslant \tau^2_{u\mathfrak{a}}$. If $\widehat{f}_{m^n_\Phi} = U^\star([\widehat{f}]\,\mathbb{1}_{\mathcal{J}_{m^n_\Phi}})$ is the OSE with m^n_Φ as in (7.7), then $\mathfrak{R}_\Phi\big[\widehat{f}_{m^n_\Phi} \,|\, \mathbb{P}^{\otimes n}_{\mathbb{F}^r_\mathfrak{a}}\big] \leqslant (\sigma^2_\varepsilon + r^2\tau^2_{u\mathfrak{a}})\,\mathcal{R}^n_\Phi(\mathfrak{a})$ for all $n \in \mathbb{N}$, i.e., the rate $(\mathcal{R}^n_\Phi(\mathfrak{a}))_{n\in\mathbb{N}}$ and the OSE $(\widehat{\mathbb{P}}_{m^n_\Phi})_{n\in\mathbb{N}}$ are minimax-optimal (up to a constant).*

Proof of Corollary §7.3.14 is given in the lecture. □

7.4 Lower bound based on m hypothesis

§7.4.1 **Lemma** (Assouad's cube technique). *Consider a family $\mathbb{P}^n_\mathbb{F}$ of probability measures. For $|\mathcal{J}_m| < \infty$ let $\mathfrak{d}^{(j)}_{ist}(\cdot,\cdot)$, $j \in \mathcal{J}_m$, be distances such that $|\mathfrak{d}_{ist}(\cdot,\cdot)|^2 \geqslant \sum_{j\in\mathcal{J}_m} |\mathfrak{d}^{(j)}_{ist}(\cdot,\cdot)|^2$. Let $\theta := (\theta_j)_{j\in\mathcal{J}_m} \in \{-1,1\}^{|\mathcal{J}_m|} =: \Theta$ and for each $\theta \in \Theta$ introduce $\theta^{(j)} \in \Theta$ by $\theta^{(j)}_l = \theta_l$ for $j \neq l$ and $\theta^{(j)}_j = -\theta_j$. For each $\theta \in \Theta$ let $\mathbb{P}^n_{f_\theta}$ be a probability measure in $\mathbb{P}^n_\mathbb{F}$, then*

$$\inf_{\widetilde{f}} \mathfrak{R}_\mathfrak{d}\big[\widetilde{f} \,|\, \mathbb{P}^n_\mathbb{F}\big] \geqslant \frac{1}{2^{|\mathcal{J}_m|}} \sum_{\theta\in\Theta} \frac{1}{4} \sum_{j\in\mathcal{J}_m} |\mathfrak{d}^{(j)}_{ist}(f_\theta, f_{\theta^{(j)}})|^2 \rho^2(\mathbb{P}^n_{f_\theta}, \mathbb{P}^n_{f_{\theta^{(j)}}}).$$

Proof of Lemma §7.4.1 is given in the lecture. □

§7.4.2 **Remark** *(Lower bound for a maximal $\mathbb{H}_\mathfrak{v}$-risk).* Consider a global $\mathbb{H}_\mathfrak{v}$-risk with weighted norm $\|\cdot\|_\mathfrak{v}$ derived from an ONS $\mathcal{U}$ and some weight sequence $(\mathfrak{v}_j)_{j\in\mathcal{J}}$. In this situation the last assertion states

$$\inf_{\widetilde{f}} \mathfrak{R}_\mathfrak{v}\big[\widetilde{f} \,|\, \mathbb{P}^n_{\mathbb{F}^r_\mathfrak{a}}\big] \geqslant \frac{1}{2^{|\mathcal{J}_m|}} \sum_{\theta\in\Theta} \frac{1}{8} \sum_{j\in\mathcal{J}_m} \mathfrak{v}_j^2 |[f_\theta]_j - [f_{\theta^{(j)}}]_j|^2 \rho^2(\mathbb{P}^n_{f_\theta}, \mathbb{P}^n_{f_{\theta^{(j)}}}).$$

Let us assume that for each $\theta \in \Theta$ and $j \in \mathcal{J}_m$ the probability measures $\mathbb{P}^n_{f_\theta}$ and $\mathbb{P}^n_{f_{\theta^{(j)}}}$ are uniformely statistically indistinguishable in the sense that $\rho(\mathbb{P}^n_{f_\theta}, \mathbb{P}^n_{f_{\theta^{(j)}}}) \geqslant c$ for some $c > 0$. If we consider furthermore candidates $f_\theta := \sum_{j\in\mathcal{J}_m} \theta_j [f_*]_j u_j$, $\theta \in \Theta$, for some $f_* \in \mathbb{F}^r_\mathfrak{a}$, then it is easily verified that $\{f_\theta, \theta \in \Theta\} \subset \mathbb{F}^r_\mathfrak{a}$ and $\sum_{j\in\mathcal{J}_m} \mathfrak{v}_j^2 |[f_\theta]_j - [f_{\theta^{(j)}}]_j|^2 = 4\sum_{j\in\mathcal{J}_m} \mathfrak{v}_j^2 |[f_*]_j|^2 = 4\,\|\Pi_{\mathbb{U}_m} f_*\|^2_\mathfrak{v}$ which in turn implies

$$\inf_{\widetilde{f}} \mathfrak{R}_\mathfrak{v}\big[\widetilde{f} \,|\, \mathbb{P}^n_{\mathbb{F}^r_\mathfrak{a}}\big] \geqslant \frac{1}{2^{|\mathcal{J}_m|}} \sum_{\theta\in\Theta} c^2 \,\|\Pi_{\mathbb{U}_m} f_*\|^2_\mathfrak{v} = c^2 \,\|\Pi_{\mathbb{U}_m} f_*\|^2_\mathfrak{v}\,. \tag{7.11}$$

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 $\mathbb{P}^n_{f_\theta} \in \mathbb{P}^n_\mathbb{F}$, $\theta \in \Theta$, are still uniformely statistically indistinguishable. □

7.4.1 Examples - lower bound of a maximal $\mathbb{H}_{\mathfrak{v}}$-risk

Assuming that the function of interest f with generalised Fourier coefficients $[f] = ([f]_j)_{j\in\mathcal{J}}$ belongs to the class of solutions $\mathbb{F}^r_{\mathfrak{a}}$ as in §6.2.3 we derive below a lower bound of a maximal $\mathbb{H}_{\mathfrak{v}}$-risk considering the three examples: (i) Gaussian sequence space model (GSSM) §6.5.2, (ii) non-parametric regression §6.5.4, and (iii) density estimation §6.5.4. Define for $n \in \mathbb{N}$ and $m \in \mathcal{M}$,

$$\mathcal{R}^n_{\mathfrak{v}}(m,\mathfrak{a}) := \max\big((\mathfrak{a}\mathfrak{v})^2_{(m)}, n^{-1}\|\mathfrak{v}\mathbb{1}_{\mathcal{J}_m}\|^2_{\ell^2}\big),$$
$$m^n_{\mathfrak{v}} := \arg\min\{\mathcal{R}^n_{\mathfrak{v}}(m,\mathfrak{a}), m\in\mathcal{M}\} \text{ and } \mathcal{R}^n_{\mathfrak{v}}(\mathfrak{a}) := \mathcal{R}^n_{\mathfrak{v}}(m^n_{\mathfrak{v}},\mathfrak{a}). \quad (7.12)$$

Keep in mind the quantities $\mathcal{R}^n_{\mathfrak{v}}(m,f)$ and $\mathcal{R}^n_{\mathfrak{v}}(f)$ given in (6.7), where for any $f \in \mathbb{F}^r_{\mathfrak{a}}$ and $m \in \mathcal{M}$ we have $\big\|\mathfrak{v}[f]\mathbb{1}_{\mathcal{J}^c_m}\big\|_{\ell^2} \leqslant r(\mathfrak{v}\mathfrak{a})_{(m)}$, and hence, $\mathcal{R}^n_{\mathfrak{v}}(m,f) \leqslant (1\vee r^2)\mathcal{R}^n_{\mathfrak{v}}(m,\mathfrak{a})$ for all $n\in\mathbb{N}$. Consequently, $\mathcal{R}^n_{\mathfrak{v}}(f) \leqslant (1\vee r^2)\mathcal{R}^n_{\mathfrak{v}}(\mathfrak{a})$ where $\mathcal{R}^n_{\mathfrak{v}}(f)$ is eventually the oracle rate (see, for instance, Proposition §6.6.5). We show below that $\mathcal{R}^n_{\mathfrak{v}}(\mathfrak{a})$ eventually is a minimax rate. We impose a minimal regularity of the weight sequences $\mathfrak{a}$ and $\mathfrak{v}$, which is formalised in the next assumption.

§7.4.3 **Assumption**. Consider a pre-specified ONS $\{u_j, j\in\mathcal{J}\}$ in $\mathbb{H}$, a nested sieve $(\mathcal{J}_m)_{m\in\mathcal{M}}$ in $\mathcal{J}$ and strictly positive sequences $\mathfrak{v}$ and $\mathfrak{a}$ such that $\mathfrak{v}\mathfrak{a} = (\mathfrak{v}_j\mathfrak{a}_j)_{j\in\mathcal{J}}$ is monotonically non-increasing, i.e., $\min\{\mathfrak{v}_j\mathfrak{a}_j, j\in\mathcal{J}_m,\} \geqslant \sup\{\mathfrak{v}_j\mathfrak{a}_j, j\in\mathcal{J}^c_m\} = (\mathfrak{v}\mathfrak{a})_{(m)} > 0$ for any $m\in\mathcal{M}$. Suppose further that $\eta := \inf\Big\{|\mathcal{R}^n_{\mathfrak{v}}(\mathfrak{a})|^{-1}\min((\mathfrak{a}\mathfrak{v})^2_{(m^n_{\mathfrak{v}})}, \tfrac{1}{n}\|\mathfrak{v}\mathbb{1}_{\mathcal{J}_{m^n_{\mathfrak{v}}}}\|^2_{\ell^2}), n\in\mathbb{N}\Big\} > 0$.

Keep in mind, that under Assumption §7.4.3 we have $\mathbb{F}_{\mathfrak{a}} \subset \mathbb{H}_{\mathfrak{v}}$. In the proof of the next propositions we intend to apply the result presented in (7.11) in Remark §7.4.2 to a specific choice of $f_\star \in \mathbb{F}^r_{\mathfrak{a}}$ which we specify next.

§7.4.4 **Lemma**. *Consider η as in Assumption §7.4.3 and for $n\in\mathbb{N}$ let $m_* := m^n_{\mathfrak{v}}$ as in (7.12). Define $\alpha_* := (\|\mathfrak{v}\mathbb{1}_{\mathcal{J}_{m_*}}\|^2_{\ell^2}/n)^{-1}\mathcal{R}^n_{\mathfrak{v}}(\mathfrak{a}) \leqslant \eta^{-1}$ and $\zeta := \eta\min(r^2,c)$ for some $c>0$. Consider the function $f_* := (\zeta\alpha_*/n)^{1/2}\sum_{j\in\mathcal{J}_{m_*}} u_j$. Then we have $\|f_*\|^2_{1/\mathfrak{a}} \leqslant \min(r^2,c)$, i.e., $f_\star \in \mathbb{F}^r_{\mathfrak{a}}$, and $n\max\{|[f_\star]^2_j|, j\in\mathcal{J}_m\} \leqslant c$.*

Proof of Lemma §7.4.4 is given in the lecture. □

§7.4.5 **Corollary**. *Let the assumptions of Lemma §7.4.4 be satisfied. If the ONS $\mathcal{U}$ is in addition regular w.r.t. the weigth sequence $\mathfrak{a}$ as in §6.1.12 (ii), i.e., $\|\sum_{j\in\mathcal{J}}\mathfrak{a}^2_j|u_j|^2\|_{L^\infty} \leqslant \tau^2_{u\mathfrak{a}}$ for some $\tau_{u\mathfrak{a}} \geqslant 1$, then it holds $\|f_*\|^2_{L^\infty} \leqslant \tau^2_{u\mathfrak{a}}c$.*

Proof of Corollary §7.4.5 follows along the lines of the proof of Corollary §7.3.6. □

§7.4.6 **Proposition** (GSSM, §6.5.2 continued). *Consider the reconstruction of $f = \mathcal{U}^\star[f] \in \mathbb{F}^r_{\mathfrak{a}}$ given for each $n\in\mathbb{N}$ an observable quantity $\widehat{f} \curvearrowright \mathfrak{N}(\mathbb{F}^r_{\mathfrak{a}}, \frac{1}{n}\mathrm{Id})$. Under Assumption §7.4.3 we have*

$$\inf_{\widetilde{f}} \mathfrak{R}_{\mathfrak{v}}\big[\widetilde{f}\,|\,\mathfrak{N}(\mathbb{F}^r_{\mathfrak{a}}, \tfrac{1}{n}\mathrm{Id})\big] \geqslant \frac{\eta}{8}\ \min(2r^2,1)\ \mathcal{R}^n_{\mathfrak{v}}(\mathfrak{a}), \quad \textit{for all } n\in\mathbb{N}. \quad (7.13)$$

Proof of Proposition §7.4.6 is given in the lecture. □

§7.4.7 **Corollary**. *Under the assumptions of Proposition §7.4.6 for each $n \in \mathbb{N}$ consider the OSE $\widehat{f}_{m_{\mathfrak{v}}^n} = U^\star([\widehat{f}]\,\mathbb{1}_{\mathcal{J}_{m_{\mathfrak{v}}^n}})$ with $m_{\mathfrak{v}}^n$ as in (7.12). Then, $\mathfrak{R}_{\mathfrak{v}}\big[\widehat{f}_{m_{\mathfrak{v}}^n} \,|\, \mathfrak{N}(\mathbb{F}_{\mathfrak{a}}^r, \frac{1}{n}\mathrm{Id}_{\mathbb{H}})\big] \leqslant (1+r^2)\,\mathcal{R}_{\mathfrak{v}}^n(\mathfrak{a})$ for all $n \in \mathbb{N}$, i.e., the rate $(\mathcal{R}_{\mathfrak{v}}^n(\mathfrak{a}))_{n\in\mathbb{N}}$ and the OSE $(\widehat{f}_{m_{\mathfrak{v}}^n})_{n\in\mathbb{N}}$ are minimax-optimal (up to a constant).*

Proof of Corollary §7.4.7 is given in the lecture. □

§7.4.8 **Proposition** (Non-parametric density estimation §6.5.3 continued). *Consider the reconstruction of a density $\mathbb{p} = \mathbb{1}_{[0,1]} + f$ with $f \in \mathbb{F}_{\mathfrak{a}}^r$ given for each $n \in \mathbb{N}$ an i.i.d. sample $(X_1, \dots, X_n) \bigcirc\!\!\sim \mathbb{P}_{\mathbb{D}_{\mathfrak{a}}^r}^{\otimes n}$. Let the ONS $\mathcal{U}$ be in addition regular w.r.t. the weigth sequence $\mathfrak{a}$ as in §6.1.12 (ii), i.e., $\|\sum_{j\in\mathcal{J}} \mathfrak{a}_j^2 |u_j|^2\|_{L^\infty} \leqslant \tau_{u\mathfrak{a}}^2$ for some $\tau_{u\mathfrak{a}} \geqslant 1$. Under Assumption §7.4.3 we have*

$$\inf_{\widetilde{f}} \mathfrak{R}_{\mathfrak{v}}\big[\widetilde{f} \,|\, \mathbb{P}_{\mathbb{D}_{\mathfrak{a}}^r}^{\otimes n}\big] \geqslant \frac{\eta}{16}\, \min(r^2, (4\tau_{u\mathfrak{a}}^2)^{-1})\, \mathcal{R}_{\mathfrak{v}}^n(\mathfrak{a}), \quad \text{for all } n \geqslant 2. \tag{7.14}$$

Proof of Proposition §7.4.8 is given in the lecture. □

§7.4.9 **Remark**. We shall emphasise that assuming in Proposition §7.4.8 in addition a regular ONS $\mathcal{U}$ ensures that the specific choice $\mathbb{p}_* = \mathbb{1}_{[0,1]} + f_*$ with f_* as specified in Lemma §7.4.4 is indeed a density belonging to $\mathbb{D}_{\mathfrak{a}}^r$. Moreover, the specific choice satisfy $1/2 \leqslant \mathbb{p}_* \leqslant 1 + 1/2 \leqslant 2$, λ-a.s.. Thereby, due to Lemma §6.6.8 $\mathcal{R}_{\mathfrak{v}}^n(\mathbb{p}_*)$ as given in (6.7) is an oracle rate, where $\mathcal{R}_{\mathfrak{v}}^n(\mathbb{p}_*) \leqslant (1 \vee r^2)\mathcal{R}_{\mathfrak{v}}^n(\mathfrak{a})$, and $\mathcal{R}_{\mathfrak{v}}^n(\mathfrak{a})$ is a minimax-rate. □

§7.4.10 **Corollary**. *Under the assumptions of Proposition §7.4.8 for each $n \in \mathbb{N}$ consider a noisy version $\widehat{\mathbb{p}} \sim \mathfrak{L}(\mathbb{p}, \frac{1}{n}\Gamma_{\mathbb{p}})$ with $\Gamma_{\mathbb{p}} = \mathrm{M}_{\mathbb{p}} - \mathrm{M}_{\mathbb{p}}\Pi_{\{\mathbb{1}_{[0,1]}\}}\mathrm{M}_{\mathbb{p}}$ as in §6.4.4 and an observable quantity $[\widehat{\mathbb{p}}] = (\overline{\mathbb{P}}_n u_j)_{j\in\mathbb{N}}$ using an i.i.d. sample $(X_1, \dots, X_n) \bigcirc\!\!\sim \mathbb{P}_{\mathbb{D}_{\mathfrak{a}}^r}^{\otimes n}$. Let $\widehat{\mathbb{p}}_{m_{\mathfrak{v}}^n} = \mathbb{1}_{[0,1]} + U^\star([\widehat{\mathbb{p}}]\,\mathbb{1}_{\mathcal{J}_{m_{\mathfrak{v}}^n}})$ be the OSE with $m_{\mathfrak{v}}^n$ as in (7.12). Then, $\mathfrak{R}_{\mathfrak{v}}\big[\widehat{\mathbb{p}}_{m_{\mathfrak{v}}^n} \,|\, \mathbb{P}_{\mathbb{D}_{\mathfrak{a}}^r}^{\otimes n}\big] \leqslant (1 + r\tau_{u\mathfrak{a}} + r^2)\,\mathcal{R}_{\mathfrak{v}}^n(\mathfrak{a})$ for all $n \in \mathbb{N}$, i.e., the rate $(\mathcal{R}_{\mathfrak{v}}^n(\mathfrak{a}))_{n\in\mathbb{N}}$ and the OSE $(\widehat{\mathbb{p}}_{m_{\mathfrak{v}}^n})_{n\in\mathbb{N}}$ are minimax-optimal (up to a constant).*

Proof of Corollary §7.4.10 is given in the lecture. □

§7.4.11 **Proposition** (Non-parametric regression §6.5.4 continued). *Consider the reconstruction of a regression function $f \in \mathbb{F}_{\mathfrak{a}}^r$ given for each $n \in \mathbb{N}$ an i.i.d. sample $(X_1, Z_1), \dots, (X_n, Z_n)$ with joint distribution belonging to $\mathbb{P}_{\mathbb{F}_{\mathfrak{a}}^r}^{\otimes n}$ and obeying the Assumption §5.3.1 (see section 5.3). If the error term is $\mathfrak{N}(0, \sigma_\varepsilon^2)$-distributed, then under Assumption §7.4.3 we have*

$$\inf_{\widetilde{f}} \mathfrak{R}_{\mathfrak{v}}\big[\widetilde{f} \,|\, \mathbb{P}_{\mathbb{F}_{\mathfrak{a}}^r}^{\otimes n}\big] \geqslant \frac{\eta}{8}\, \min(2r^2, \sigma_\varepsilon^2)\, \mathcal{R}_{\mathfrak{v}}^n(\mathfrak{a}), \quad \text{for all } n \in \mathbb{N}. \tag{7.15}$$

Proof of Proposition §7.4.11 is given in the lecture. □

§7.4.12 **Remark**. We shall emphasise that assuming in Proposition §7.4.11 in addition normal-distributed error terms is only needed to simplify the calculation of the distance between distributions corresponding to different regression functions. On the other hand, below we derive an upper bound under Assumption §5.3.1 (see section 5.3) only. In this situation, Proposition §7.4.11 obviously provides a lower bound for any estimator since the family $\mathbb{P}_{\mathbb{F}_{\mathfrak{a}}^r}^{\otimes n}$ contains this specific Gaussian-error case. Moreover, if the ONS $\mathcal{U}$ is in addition regular w.r.t. the weight sequence $\mathfrak{a}$, then the specific choice f_* given by Lemma §7.4.4 satisfies $\|f_*\|_{L^\infty}^2 < \infty$ and due to Lemma §6.6.11 $\mathcal{R}_{\mathfrak{v}}^n(f_*)$ as given in (6.8) is an oracle rate, where $\mathcal{R}_{\mathfrak{v}}^n(f_*) \leqslant (1 \vee r^2)\mathcal{R}_{\mathfrak{v}}^n(\mathfrak{a})$, and $\mathcal{R}_{\mathfrak{v}}^n(\mathfrak{a})$ is a minimax-rate. □

§7.4.13 **Corollary**. *For each $n \in \mathbb{N}$ using an i.i.d. sample $(X_1, Z_1), \dots, (X_n, Z_n) \circledcirc \mathbb{P}_{\mathbb{F}_{\mathfrak{a}}^r}^{\otimes n}$ obeying the Assumption §5.3.1 (see section 5.3) consider a noisy version $\widehat{f} \sim \mathfrak{L}(f, \frac{1}{n}\Gamma_f)$ with $\Gamma_f = \sigma_\varepsilon^2 \operatorname{Id}_{\mathbb{H}} + \mathrm{M}_{|f|^2}$ as in §6.5.4 and an observable quantity $[\widehat{f}] = (\overline{\mathbb{P}}_n[\mathrm{id} \otimes u_j])_{j \in \mathbb{N}}$. Let the ONS $\mathcal{U}$ be in addition regular w.r.t. the sequence $\mathfrak{a}$ as in §6.1.12 (ii), i.e., $\|\sum_{j \in \mathcal{J}} \mathfrak{a}_j^2 |u_j|^2\|_{L^\infty} \leqslant \tau_{u\mathfrak{a}}^2$. If $\widehat{f}_{m_{\mathfrak{v}}^n} = U^\star([\widehat{f}]\, \mathbb{1}_{\mathcal{J}_{m_{\mathfrak{v}}^n}})$ is the OSE with $m_{\mathfrak{v}}^n$ as in (7.12), then $\mathfrak{R}_{\mathfrak{v}}\big[\widehat{f}_{m_{\mathfrak{v}}^n} \,|\, \mathbb{P}_{\mathbb{F}_{\mathfrak{a}}^r}^{\otimes n}\big] \leqslant (\sigma_\varepsilon^2 + r^2 \tau_{u\mathfrak{a}}^2)\, \mathcal{R}_{\mathfrak{v}}^n(\mathfrak{a})$ for all $n \in \mathbb{N}$, i.e., the rate $(\mathcal{R}_{\mathfrak{v}}^n(\mathfrak{a}))_{n \in \mathbb{N}}$ and the OSE $(\widehat{\mathbb{P}}_{m_{\mathfrak{v}}^n})_{n \in \mathbb{N}}$ are minimax-optimal (up to a constant).*

Proof of Corollary §7.4.13 is given in the lecture. □

Bibliography

L. Birgé and P. Massart. From model selection to adaptive estimation. Pollard, David (ed.) et al., Festschrift for Lucien Le Cam: research papers in probability and statistics. New York, NY: Springer. 55-87, 1997.

N. L. Carr. Kinetics of catalytic isomerization of n-pentane. *Industrial and Engineering Chemistry*, 52:391–396, 1960.

F. Comte. *Estimation non-paramétrique*. Spartacus-idh, Paris, 2015.

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.

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

H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18(1):50–60, 1947.

A. B. Tsybakov. *Introduction to nonparametric estimation.* Springer Series in Statistics. Springer, New York, 2009.

D. Werner. *Funktionalanalysis*. Springer-Lehrbuch, 2011.

F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

Index

Nested sieve
  $(\mathbb{U}_m)_{m\in\mathcal{M}}$ in $\mathbb{U}$, 45
  $(\mathcal{J}_m)_{m\in\mathcal{M}}$ in $\mathcal{J}$, 45
Non-parametric
  density estimation, 35, 36, 50, 51, 53, 56, 59, 63, 66
  regression, 36, 40, 51, 53, 57, 59, 63, 64, 66, 67
Norm
  spectral, $\|\cdot\|_s$, 46
  uniform operator, $\|\cdot\|_{\mathscr{L}}$, 46

Operator
  covariance, Γ, 50
  Fourier series transform, U, 46, 47
  linear functional, Φ, 46
  multiplication, M_λ, 46, 47
Operator classes
  bounded linear, $\mathscr{L}(\mathbb{H},\mathbb{G})$, 46
  linear functionals, $\mathcal{L}_{1/\mathfrak{a}}$, 48
Orthonormal system (ONS), 44
  regular, 45, 48

Sequence space model, 52, 55
  Gaussian (GSSM), 53, 55, 62, 63, 66