Institute of Applied Mathematics
Statistics of inverse problems research group
Professor Dr. Jan JOHANNES

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Outline of the lecture course

STATISTICS 2

Summer semester 2020

Preliminary version: July 21, 2020

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 54.14.101
email: johannes@math.uni-heidelberg.de
webpage: sip.math.uni-heidelberg.de

Table of contents

Chapter 1

Preliminaries

Elements of the PROBABILITY THEORY *are recalled along the lines of the lecture Statistik 1. For a detailed exposition with many examples we refer to the text book* **Klenke** *[2008].*

§01 Fundamentals

§01.01 **Notation.** 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).

(i) We set $\mathbb{R}^+ := [0, \infty)$, $\mathbb{R}^+_{\setminus 0} := (0, \infty)$, $\mathbb{R}_{\setminus 0} := \mathbb{R} \setminus \{0\}$, $\overline{\mathbb{R}} := [-\infty, \infty]$, $\overline{\mathbb{R}}^+ := [0, \infty]$.

(ii) For $a, b \in \mathbb{R}$ with $a < b$ we write $[\![a, b]\!] := [a, b] \cap \mathbb{Z}$, $[\![a, b[\![:= [a, b) \cap \mathbb{Z}$ and $]\!]a, b]\!] := (a, b] \cap \mathbb{Z}$. Moreover, let $[\![n]\!] := [\![1, n]\!]$ and $[\![n[\![:= [\![1, n[\![$ for $n \in \mathbb{N}$.

(iii) For $a^n = (a_i)_{i \in [\![n]\!]}, b^n = (b_i)_{i \in [\![n]\!]} \in \overline{\mathbb{R}}^n$ we write $a^n < b^n$, if $a_i < b_i$ for all $i \in [\![n]\!]$. For $a^n < b^n$, define the open *rectangle* as the Cartesian product $(a^n, b^n) := \bigtimes_{i=1}^n (a_i, b_i) := (a_1, b_1) \times (a_2, b_2) \times \cdots \times (a_n, b_n)$. Analogously, we define $[a^n, b^n]$, $(a^n, b^n]$ and $[a^n, b^n)$.

(iv) We call $\overline{\mathscr{B}} := \mathscr{B}_{\overline{\mathbb{R}}}$ the Borel-σ-field over the compactified real line $\overline{\mathbb{R}}$, where the sets $\{-\infty\}$, $\{\infty\}$ and $\mathbb{R}$ are in $\overline{\mathbb{R}}$ closed and open, respectively, and hence Borel-measurable. In particular, the trace $\mathscr{B} := \mathscr{B}_{\mathbb{R}} = \overline{\mathscr{B}} \cap \mathbb{R}$ of $\overline{\mathscr{B}}$ over $\mathbb{R}$ is the Borel-σ-field over $\mathbb{R}$. Furthermore, we write $\overline{\mathscr{B}}^+ := \overline{\mathscr{B}} \cap \overline{\mathbb{R}}^+$, $\mathscr{B}^+ := \mathscr{B} \cap \mathbb{R}^+$ and $\mathscr{B}^+_{\setminus 0} := \mathscr{B} \cap \mathbb{R}^+_{\setminus 0}$.

(v) Given a measurable space $(\Omega, \mathscr{A})$ a Borel-measurable function $g : \Omega \to \mathbb{R}$ and $f : \Omega \to \overline{\mathbb{R}}$ is called *real* and *numerical*, respectively, and we write $g \in \mathscr{A}$ and $f \in \overline{\mathscr{A}}$ for short. g respectively f is called positive if $g(\Omega) \in \mathbb{R}^+$ respectively $f(\Omega) \in \overline{\mathbb{R}}^+$, then we write $g \in \mathscr{A}^+$ and $f \in \overline{\mathscr{A}}^+$- We call a Borel-measurable function $f^k = (f_i)_{i \in [\![k]\!]} : \Omega \to \overline{\mathbb{R}}^k$, that is $f_i \in \overline{\mathscr{A}}$ for each $i \in [\![k]\!]$, and $g^k = (g_i)_{i \in [\![k]\!]} : \Omega \to \mathbb{R}^k$, *numerical* and *real*, respectively and we write $f^k \in \overline{\mathscr{A}}^k$ and $g^k \in \mathscr{A}^k$ for short. □

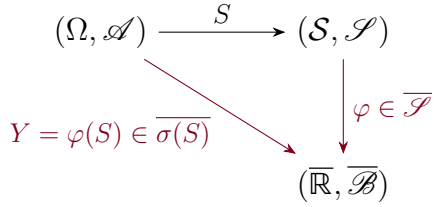
§01.02 **Property.**

(i) *For $X, Y \in \overline{\mathscr{A}}$ and $a \in \mathbb{R}$ holds: $aX \in \overline{\mathscr{A}}$ (with convention $0 \times \infty = 0$); $X \vee Y := \max(X, Y), X \wedge Y := \min(X, Y) \in \overline{\mathscr{A}}$ and particularly $X^+ := X \vee 0$, $X^- := (-X)^+ \in \overline{\mathscr{A}}^+$, $|X| \in \overline{\mathscr{A}}^+$, $\{X < Y\}, \{X \leqslant Y\}, \{X = Y\} \in \mathscr{A}$, and $\lfloor X \rfloor \in \overline{\mathscr{A}}^+$.*

(ii) *For $X^n = (X_i)_{i \in [\![n]\!]} \in \mathscr{A}^n$, i.e., $X_i \in \mathscr{A}$, $i \in [\![n]\!]$, and Borel-measurable $h : \mathbb{R}^n \to \mathbb{R}^m$ holds $h(X^n) \in \mathscr{A}^m$, and in particular $X_1 + X_2$, $X_1 - X_2$, $X_1 X_2 \in \mathscr{A}$, and $X_1 / X_2 \in \overline{\mathscr{A}}$.*

(iii) *Let $(X_n)_{n \in \mathbb{N}}$ be a sequence in $\overline{\mathscr{A}}$. Then $\sup_{n \in \mathbb{N}} X_n \in \overline{\mathscr{A}}$, $\inf_{n \in \mathbb{N}} X_n \in \overline{\mathscr{A}}$, $X_\star = \liminf_{n \to \infty} X_n \in \overline{\mathscr{A}}$ and $X^\star = \limsup_{n \to \infty} X_n \in \overline{\mathscr{A}}$. If $X := \lim_{n \to \infty} X_n$ exists, then $X \in \overline{\mathscr{A}}$.*

(iv) *Let $S : (\Omega, \mathscr{A}) \to (\mathcal{S}, \mathscr{S})$ be measurable, $\sigma(S) := S^{-1}(\mathscr{S}) \subseteq \mathscr{A}$ the sub-σ-field generated by S and $Y : \Omega \to \overline{\mathbb{R}}$. Then the following conditions are equivalent:* (a) *Y is $\sigma(S)$-measurable, symbolically $Y \in \overline{\sigma(S)}$;* (b) *There exists a measurable $\varphi : (\mathcal{S}, \mathscr{S}) \to (\overline{\mathbb{R}}, \overline{\mathscr{B}})$, in short $\varphi \in \overline{\mathscr{S}}$, with $Y = \varphi(S)$. If Y is real, bounded or positive, then φ has each of those properties too.*



The function φ is uniquely determined by Y on $S(\Omega)$, and for all $s \notin S(\Omega)$ it can be arbitrarily be extended.

(v) *For every $X \in \overline{\mathscr{A}}^+$ the sequence of simple random variables $(X_n)_{n\in\mathbb{N}}$ in $\overline{\mathscr{A}}^+$ given by $X_n := (2^{-n}\lfloor 2^n X \rfloor) \wedge n$ satisfies* (a) *$X_n \uparrow X$;* (b) *$X_n \leqslant X \wedge n$;* (c) *For each $c \in \mathbb{R}^+$ holds $\lim_{n\to\infty} X_n = X$ uniformly on $\{X \leqslant c\}$.* □

§01.03 **Notation.** For a measure μ on $(\Omega, \mathscr{A})$ we denote the integral of $f \in \overline{\mathscr{A}}$ with respect to μ by $\mu f := \int f d\mu$, if it exists. For $s \in \mathbb{R}^+_{\setminus 0}$ define $\|f\|_{\mathscr{L}_s(\mu)} := (\mu|f|^s)^{1/s}$, and $\|f\|_{\mathscr{L}_\infty(\mu)} := \inf\{c \in \mathbb{R}^+ : \mu(|f| > c) = 0\}$. For $s \in \overline{\mathbb{R}}^+_{\setminus 0} := (0, \infty]$ a function $f \in \overline{\mathscr{A}}$ is called $\mathscr{L}_s(\mu)$ $\mathscr{L}_s(\mu)$-integrable, if $\|f\|_{\mathscr{L}_s(\mu)} < \infty$. We denote the set of all $\mathscr{L}_s(\mu)$-integrable functions by $\mathscr{L}_s(\mu) := \mathscr{L}_s(\mathscr{A}, \mu) := \{f \in \overline{\mathscr{A}} : \|f\|_{\mathscr{L}_s(\mu)} < \infty\}$. Note that $\|\cdot\|_{\mathscr{L}_s(\mu)}$ is a seminorm on $\mathscr{L}_s(\mu)$ for each $s \in [1, \infty]$. Given a metric space $(\mathcal{X}, d)$ equipped with its Borel-σ-field $\mathscr{B}_\mathcal{X}$ we denote by $\mathcal{C}_b := \mathcal{C}_b(\mathcal{X})$ the set of all bounded and continuous functions mapping $\mathcal{X}$ into $\mathbb{R}$. For any finite measure μ on $(\mathcal{X}, \mathscr{B}_\mathcal{X})$ we have $\|h\|_{\mathscr{L}_\infty(\mu)} < \infty$ for all $h \in \mathcal{C}_b$ and thus $\mathcal{C}_b \subseteq \mathscr{L}_\infty(\mathscr{B}_\mathcal{X}, \mu)$ in equal. We denote by λ the Lebesgue measure on $(\mathbb{R}, \mathscr{B})$ and write shortly $\mathscr{L}_s(:) = \mathscr{L}_s(\mathscr{B}) := \mathscr{L}_s(\mathscr{B}, \lambda)$. □

§01.04 **Notation.** We understand a vector $a^k = (a_i)_{i\in[\![k]\!]}$ as a column vector, i.e., $a^k = (a_1 \cdots a_k)^t \in \overline{\mathbb{R}}^k$ and hence we identify $\overline{\mathbb{R}}^k$ and $\overline{\mathbb{R}}^{(k,1)}$. We denote by $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ the Euclidean norm and inner product on $\mathbb{R}^k$, respectively, i.e, $\|a^k\| = (\sum_{i\in[\![k]\!]} |a_i|^2)^{1/2}$ and $\langle a^k, b^k\rangle = \sum_{i\in[\![k]\!]} a_i b_i = (b^k)^t a^k$ for all $a^k, b^k \in \overline{\mathbb{R}}^k$. For $s \in \mathbb{R}^+_{\setminus 0}$ we define $\|a^k\|_s := (\sum_{i\in[\![k]\!]} |a_i|^s)^{1/s}$ and $\|a^k\|_\infty := \max_{i\in[\![k]\!]} |a_i|$. Note that $f^k \in \overline{\mathscr{A}}^k$ and $g^k \in \mathscr{A}^k$ imply $\|f^k\|_s \in \overline{\mathscr{A}}$ and $\|g^k\|_s \in \mathscr{A}$ for any $s \in \overline{\mathbb{R}}^+_{\setminus 0}$. We call $f^k = (f_i)_{i\in[\![k]\!]}$ $\mathscr{L}^k_s(\mu)$-integrable if $\|f^k\|_s \in \mathscr{L}_s(\mu)$ or equivalently $f_i \in \mathscr{L}_s(\mu)$ for each $i \in [\![k]\!]$. We define $\|f^k\|_{\mathscr{L}^k_s(\mu)} := \|\|f^k\|_p\|_{\mathscr{L}_s(\mu)}$ and $\mathscr{L}^k_s(\mu) := \mathscr{L}^k_s(\mathscr{A}, \mu) := \{f^k \in \overline{\mathscr{A}}^k : \|f^k\|_{\mathscr{L}^k_s(\mu)} < \infty\}$ with a slight abuse of notation. □

§01.05 **Notation.** Let X be a random variable, i.e. a measurable function, defined on a probability space $(\Omega, \mathscr{A}, \mathbb{P})$ with values in a measurable space $(\mathcal{X}, \mathscr{X})$. The probability measure on $(\mathcal{X}, \mathscr{X})$ induced by X is denoted by $\mathbb{P}^X := \mathbb{P} \circ X^{-1}$ and we write $X \sim \mathbb{P}^X$ for short. For $f \in \overline{\mathscr{X}}$ the expectation of f with respect to $\mathbb{P}^X$ or equivalently of $f(X)$ with respect to $\mathbb{P}$ (if it exists) is denoted by $\mathbb{E}^X f := \mathbb{P}^X f = \mathbb{P} f(X) =: \mathbb{E} f(X)$ for short. For example, when applied to the empirical measure $\widehat{\mathbb{P}}_n$ given by $\widehat{\mathbb{P}}_n(x^n) := \frac{1}{n}\sum_{i\in[\![n]\!]} \delta_{x_i}$ for $x^n = (x_i)_{i\in[\![n]\!]} \in \mathcal{X}^n$ this yields $\widehat{\mathbb{P}}_n f \in \overline{\mathscr{X}}$ with $x^n \mapsto (\widehat{\mathbb{P}}_n f)(x^n) := \frac{1}{n}\sum_{i\in[\![n]\!]} f(x_i)$. In other words, for each $x^n \in \mathcal{X}^n$, $(\widehat{\mathbb{P}}_n f)(x^n)$ is an abbreviation for the average $\frac{1}{n}\sum_{i\in[\![n]\!]} f(x_i)$. We denote by $\mathcal{W}(\mathscr{X})$ the set of

all probability measures on $(\mathcal{X}, \mathscr{X})$ and for $\mathbb{R}^n$ equipped with its Borel-σ-field $\mathscr{B}^n := \mathscr{B}_{\mathbb{R}^n}$ by $\mathcal{W}_s(\mathscr{B}^n) \subseteq \mathcal{W}(\mathscr{B}^n)$ the subset of all probability measures on $(\mathbb{R}^n, \mathscr{B}^n)$ with finite $s \in \mathbb{R}^+$ absolute mean, that is, for all $\mathbb{P} \in \mathcal{W}_s(\mathscr{B}^n)$ the identity mapping $\mathrm{id}_n : \mathbb{R}^n \to \mathbb{R}^n$ belongs to $\mathscr{L}_s^n(\mathbb{P})$. Furthermore, for $Y \sim \mathbb{P}$ we write $\mathbb{E}(Y) = \mathbb{P}(Y) := \mathbb{P}(\mathrm{id}_n) = \big(\mathbb{P}(\Pi_i)\big)_{i\in[\![n]\!]}$ using for $i \in [\![n]\!]$ the coordinate map $\Pi_i : \mathbb{R}^n \to \mathbb{R}$ with $x^n = (x_i)_{i\in[\![n]\!]} \mapsto \Pi_i(x^n) := x_i$. □

§01.06 **Property**. *Let $X \in \mathscr{L}_2^n(\mathbb{P})$, i.e. $\|X\|^2_{\mathscr{L}_2^n(\mathbb{P})} = \mathbb{P}(\|X\|^2) < \infty$. For each $b \in \mathbb{R}^n$ and $A \in \mathbb{R}^{(n,k)}$ we have $Y := AX + b \in \mathscr{L}_2^n(\mathbb{P})$. If we further denote by $\mu := \mathbb{P}X \in \mathbb{R}^k$ and $\Sigma := \mathbb{C}\mathrm{ov}(X) = \mathbb{P}(X-\mu)(X-\mu)^t = \mathbb{P}(XX^t) - \mu\mu^t \in \mathbb{R}^{(k,k)}$ expectation vector and covariance matrix of X, respectively, then $\mathbb{P}(Y) = A\mu + b \in \mathbb{R}^n$ and $\mathbb{C}\mathrm{ov}(Y) = A\Sigma A^t \in \mathbb{R}^{(n,n)}$.* □

§01.07 **Definition**. A $\mathscr{L}_2^n(\mathbb{P})$-random vector X with $\mu := \mathbb{P}(X)$ and $\Sigma := \mathbb{C}\mathrm{ov}(X)$ is *multivariate normally distributed*, $X \sim \mathrm{N}_{(\mu,\Sigma)}$ for short, if for each $c \in \mathbb{R}^k$ the real random variable $\langle X, c\rangle$ is normally distributed with mean $\langle \mu, c\rangle$ and variance $\langle \Sigma c, c\rangle$, i.e., $\langle X, c\rangle \sim \mathrm{N}_{(\langle \mu,c\rangle, \langle \Sigma c, c\rangle)}$. If Id_k denotes the k-dimensional identity matrix, then $X \sim \mathrm{N}_{(0,\mathrm{Id}_k)}$ is called a *standard normal random vector*. □

§01.08 **Property**. *A random vector $X = (X_i)_{i\in[\![k]\!]}$ is standard normal, i.e., $X \sim \mathrm{N}_{(0,\mathrm{Id}_k)}$ if and only if its components $\{X_i, i \in [\![k]\!]\}$ are independent and identically $\mathrm{N}_{(0,1)}$-distributed.* □

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

§02 Convergence of random variables

Here and subsequently, a metric space is equipped with its Borel-σ-field.

§02.01 **Definition**. Let X and X_n, $n \in \mathbb{N}$, be random variables on a probability space $(\Omega, \mathscr{A}, \mathbb{P})$ with values in a metric space $(\mathcal{X}, d)$. The sequence $(X_n)_{n\in\mathbb{N}}$ *converges to* X:

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

(b) *almost completely ($\mathbb{P}$-a.c.)*, if $\sum_{n\in\mathbb{N}} \mathbb{P}(d(X_n, X) > \varepsilon) < \infty$ for all $\varepsilon \in \mathbb{R}^+_{\setminus 0}$. We write $X_n \xrightarrow{n\to\infty} X$ $\mathbb{P}$-a.c., or briefly, $X_n \xrightarrow{\mathbb{P}\text{-a.c.}} X$.

(c) *in probability*, if $\lim_{n\to\infty} \mathbb{P}\big(d(X_n, X) \geqslant \varepsilon\big) = 0$ for all $\varepsilon \in \mathbb{R}^+_{\setminus 0}$. We write $X_n \xrightarrow{n\to\infty} X$ in $\mathbb{P}$, or briefly, $X_n \xrightarrow{\mathbb{P}} X$.

(d) *in distribution*, if $\lim_{n\to\infty} \mathbb{P}^{X_n} f = \mathbb{P}^X f$ for any $f \in \mathcal{C}_b(\mathcal{X})$. We write $X_n \xrightarrow{n\to\infty} X$ in distribution, or briefly, $X_n \xrightarrow{d} X$ and with a slight abuse of notation also $X_n \xrightarrow{d} \mathbb{P}^X$.

(e) *in $\mathscr{L}_s(\mathbb{P})$ or s-th mean*, if $\lim_{n\to\infty} \mathbb{P}(d(X_n, X)^s) = 0$. We write $X_n \xrightarrow{n\to\infty} X$ in $\mathscr{L}_s(\mathbb{P})$, or briefly, $X_n \xrightarrow{\mathscr{L}_s(\mathbb{P})} X$. □

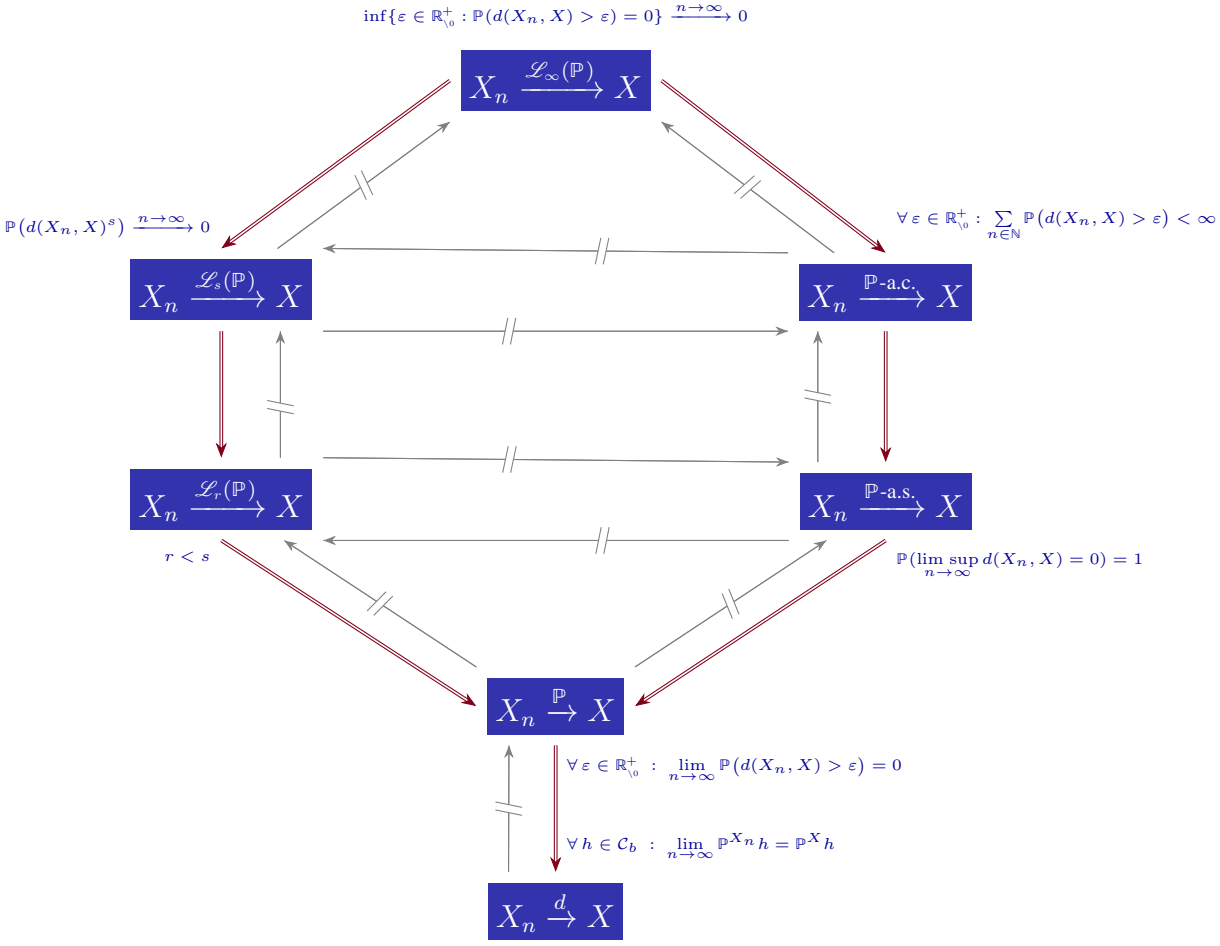
§02.02 **Remark**. Let X and X_n, $n \in \mathbb{N}$, be random vectors in $\mathbb{R}^k$, i.e., $(\mathbb{R}^k, \mathscr{B}^k)$-valued random variables, and $\|\cdot\|_s$ as in Notation §01.04. Convergence of $(X_n)_{n\in\mathbb{N}}$ to X in s-th mean, that is, $\mathbb{P}\|X_n - X\|_s^s = \|X_n - X\|^s_{\mathscr{L}_s(\mathbb{P})} \xrightarrow{n\to\infty} 0$, equals the component-wise convergence of $(X_n^i)_{n\in\mathbb{N}}$ to X^i in $\mathscr{L}_s(\mathbb{P})$, i.e., $\mathbb{P}|X_n^i - X^i|^s = \|X_n^i - X^i\|^s_{\mathscr{L}_s(\mathbb{P})} \xrightarrow{n\to\infty} 0$ for each $i \in [\![k]\!]$. □

§02.03 **Property**. *Let X and X_n, $n \in \mathbb{N}$, be random variables on a probability space $(\Omega, \mathscr{A}, \mathbb{P})$ with values in a metric space $(\mathcal{X}, d)$.*

(i) *The following statements are equivalent:* (a) $X_n \xrightarrow{\mathbb{P}\text{-a.s.}} X$; (b) $\sup_{m \geqslant n} d(X_m, X_n) \xrightarrow{\mathbb{P}} 0$; (c) $\forall \varepsilon, \delta \in \mathbb{R}_{\setminus 0}^{+} : \exists N \in \mathbb{N} : \forall n \geqslant N : \mathbb{P}\big(\bigcap_{j \geqslant n} \{d(X_j, X) \leqslant \varepsilon\}\big) \geqslant 1 - \delta$ *and* (d) $\sup_{m \geqslant n} d(X_m, X) \xrightarrow{\mathbb{P}} 0$.

(ii) *(Continuous mapping theorem) Let $g : \mathcal{X} \to \mathbb{R}$ be continuous and let $(X_n)_{n \in \mathbb{N}}$ converge to X $\mathbb{P}$-a.s. (respectively, in probability or in distribution). Then $(g(X_n))_{n \in \mathbb{N}}$ converges to $g(X)$ $\mathbb{P}$-a.s. (respectively, in probability or in distribution).*

(iii) *Counter examples show, that the converse (in gray) of the following direct implications (in red) do not hold.* □



§02.04 **Definition**. A family of $\{X_{n,j}, j \in [\![k_n]\!], n \in \mathbb{N}\}$ of real $\mathscr{L}_2$-random variables is called a standardised array, if for every $n \in \mathbb{N}$ the family $\{X_{n,j}, j \in [\![k_n]\!]\}$ is independent, centred and normed, i.e., $\mathbb{E}(X_{n,j}) = 0$, $j \in [\![k_n]\!]$ and $\sum_{j \in [\![k_n]\!]} \mathbb{V}\mathrm{ar}(X_{n,j}) = 1$. A standardised array $\{X_{n,j}, j \in [\![k_n]\!], n \in \mathbb{N}\}$ is said to satisfy

(a) the *Lindeberg condition*, if $\lim_{n \to \infty} \sum_{j \in [\![k_n]\!]} \mathbb{E}\big(X_{n,j}^2 \mathbb{1}_{\{|X_{n,j}| \geqslant \delta\}}\big) = 0$ for every $\delta \in \mathbb{R}_{\setminus 0}^{+}$;

(b) the *Lyapunov condition*, if there is $\delta \in \mathbb{R}_{\setminus 0}^{+}$ such that $\lim_{n \to \infty} \sum_{j \in [\![k_n]\!]} \mathbb{E}|X_{n,j}|^{2+\delta} = 0$. □

§02.05 **Property**. *Let $(X_n)_{n \in \mathbb{N}}$ be a sequence of independent real random variables.*

(i) *(Law of Large Numbers) Let X_n, $n \in \mathbb{N}$, be identically distributed. Then $X_1 \in \mathscr{L}_1(\mathbb{P})$ if and only if* $\lim_{n\to\infty} \frac{1}{n} \sum_{i\in[\![n]\!]} X_i = \mathbb{P}(X_1)$ $\mathbb{P}$-a.s. *(and then also in $\mathscr{L}_1(\mathbb{P})$).*

(ii) *(Lévy's equivalence theorem) For partial sums $(S_n := \sum_{i\in[\![n]\!]} X_i)_{n\in\mathbb{N}}$ $\mathbb{P}$-a.s. convergence is equivalent to convergence in probability. Otherwise, they diverge with probability one. (Kolmogorov's three-series theorem) $(S_n)_{n\in\mathbb{N}}$ converges $\mathbb{P}$-a.s. if and only if there is $\varepsilon \in \mathbb{R}^+_{\setminus 0}$ such that each of the following three conditions holds:* **(a)** $\sum_{n\in\mathbb{N}} \mathbb{P}(|X_n| > \varepsilon) < \infty$; **(b)** $\sum_{n\in\mathbb{N}} \mathbb{E}(X_n \mathbb{1}_{\{|X_n|\leqslant\varepsilon\}})$ *converges; and* **(c)** $\sum_{n\in\mathbb{N}} \mathbb{V}\mathrm{ar}(X_n \mathbb{1}_{\{|X_n|\leqslant\varepsilon\}}) < \infty$.

Let $\{X_{n,j}, j \in [\![k_n]\!], n \in \mathbb{N}\}$ be a standardised array.

(iii) *The Lyapunov condition implies the Lindeberg condition.*

(iv) *(Central Limit Theorem of Lindeberg (1922)) If the Lindeberg condition hold, then (for the row sum) $S_n^* = \sum_{j\in[\![k_n]\!]} X_{nj} \xrightarrow{d} \mathrm{N}_{(0,1)}$.* □

§02.06 **Remark** *(Law of Large Numbers).* Let X_n^k, $n \in \mathbb{N}$, be i.i.d. random vector in $\mathbb{R}^k$. Then $\|X_1^k\|_{\mathscr{L}_1(\mathbb{P})} = \mathbb{P}\|X_1^k\|_1 < \infty$ if and only if $\frac{1}{n}\sum_{i\in[\![n]\!]} X_i^k \xrightarrow{\mathbb{P}\text{-a.s.}} \mathbb{E}(X_1^k)$ (then also in $\mathscr{L}_1(\mathbb{P})$). □

§02.07 **Property** *(Portemanteau). Let X and X_n, $n \in \mathbb{N}$, be random variables on a probability space $(\Omega, \mathscr{A}, \mathbb{P})$ with values in a metric space $(\mathcal{X}, d)$. 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 $U \subseteq \mathcal{X}$;*

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

(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 $\partial B = \overline{B}\setminus B^\circ$ is the closure, interior and the boundary of B, respectively.* □

§02.08 **Property** *(Helly-Bray). Let X and X_n, $n \in \mathbb{N}$, be random vectors in $\mathbb{R}^k$ with cumulative distribution function (c.d.f.) for each $x \in \mathbb{R}^k$ given by $\mathbb{F}(x) := \mathbb{P}(X \leqslant x)$ and $\mathbb{F}_n(x) := \mathbb{P}(X_n \leqslant x)$. Then the following statements 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 $\mathbb{F}$.* □

§02.09 **Property** *(Continuous mapping theorem). Let $(\mathcal{X}_1, d_1)$ and $(\mathcal{X}_2, d_2)$ be metric spaces 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 and X_n, $n \in \mathbb{N}$, are $\mathcal{X}_1$-valued random variables with $\mathbb{P}(X \in U_\varphi) = 0$ and $X_n \xrightarrow{d} X$, then $\varphi(X_n) \xrightarrow{d} \varphi(X)$.* □

§02.10 **Property** *(Slutzky's lemma). Let X and X_n, Y_n, $n \in \mathbb{N}$, be random variables taking values in a common metric space $(\mathcal{X}, d)$ and satisfying $X_n \xrightarrow{d} X$ and $d(X_n, Y_n) \xrightarrow{\mathbb{P}} 0$. Then $Y_n \xrightarrow{d} X$.* □

§02.11 **Example**. Let X and X_n, $n \in \mathbb{N}$, be a random vector in $\mathbb{R}^k$ satisfying $X_n \xrightarrow{d} X$.

(a) If Y_n, $n \in \mathbb{N}$, are random vector in $\mathbb{R}^k$ and $c \in \mathbb{R}^k$ such that $Y_n \xrightarrow{d} c$, then $X_n + Y_n \xrightarrow{d} X + c$.

(b) If Σ_n, $n \in \mathbb{N}$ are random matrices in $\mathbb{R}^{(k,k)}$ and Σ is a matrix in $\mathbb{R}^{(k,k)}$ 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 $\Sigma_n^{-1/2} X_n \xrightarrow{d} \Sigma^{-1/2} X$. □

§02.12 **Property** *(Cramér-Wold device).* *Let X_n, $n \in \mathbb{N}$, be random vectors in $\mathbb{R}^k$. Then, the following are equivalent:* (a) *There is a random vector X with $X_n \xrightarrow{d} X$.* (b) *For any $v \in \mathbb{R}^k$, there is a real X^v with $\langle v, X_n\rangle \xrightarrow{d} X^v$. If* (a) *and* (b) *hold, then X^v and $\langle v, X\rangle$ are identically distributed (i.d.), $X^v \overset{d}{=} \langle v, X\rangle$ for short, for all $v \in \mathbb{R}^k$.* □

§02.13 **Property** *(Lindeberg-Feller CLT).* *For each $n \in \mathbb{N}$ let $\{Y_{n,j}, j \in [\![k_n]\!]\}$ be independent and centred $\mathscr{L}_2^p$-random vectors such that* (i) $\sum_{j\in[\![k_n]\!]} \mathbb{E}\|Y_{n,j}\|^2 \mathbb{1}_{\{\|Y_{n,j}\|>\varepsilon\}} \xrightarrow{n\to\infty} 0$ *for any $\varepsilon \mathbb{R}^+_{\setminus 0}$ and* (ii) $\sum_{j\in[\![k_n]\!]} \mathbb{E}(Y_{n,j}Y_{n,j}^t) \xrightarrow{n\to\infty} \Sigma$. *Then $\sum_{j\in[\![k_n]\!]} Y_{n,j} \xrightarrow{d} \mathrm{N}_{(0,\Sigma)}$.* □

§02.14 **Example**. Let X and X_n, $n \in \mathbb{N}$, be i.i.d. $\mathscr{L}_2^k(\mathbb{P})$-random vectors with $\mu = \mathbb{P}(X)$ and strictly positive definite $\Sigma = \mathbb{C}\mathrm{ov}(X)$.

(a) (CLT) $\frac{1}{\sqrt{n}} \sum_{i\in[\![n]\!]}(X_i - \mu) \xrightarrow{d} \mathrm{N}_{(0,\Sigma)}$,

(b) (LLN) $\overline{X}_n := \frac{1}{n}\sum_{i\in[\![n]\!]} X_i \xrightarrow{\mathbb{P}} \mu$,

(c) (LLN) $\frac{1}{n}\sum_{i\in[\![n]\!]} X_i X_i^t \xrightarrow{\mathbb{P}} \mathbb{E}(XX^t)$,

(d) $\widehat{\Sigma}_n := \frac{1}{n}\sum_{i\in[\![n]\!]}(X_i - \overline{X}_n)(X_i - \overline{X}_n)^t = \frac{1}{n}\sum_{i\in[\![n]\!]} X_iX_i^t - \overline{X}_n\overline{X}_n^t \xrightarrow{\mathbb{P}} \mathbb{E}(XX^t) - \mu\mu^t = \mathbb{C}\mathrm{ov}(X) = \Sigma$ (using (b) and (c) and continuous mapping theorem §02.03)

(e) $\sqrt{n}\Sigma_n^{-1/2}(\overline{X} - \mu) \xrightarrow{d} \mathrm{N}_{(0,\mathrm{Id}_k)}$ (using (a), (d) and Slutzky's lemma §02.10 as in the Example §02.11 (b)) □

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

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

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. □

§02.16 **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 and T_n, $n \in \mathbb{N}$ be random variables 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.* □

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

§02.18 **Property** *(Markov's inequality).* *If X is a $\mathscr{L}_s^k(\mathbb{P})$-random vector for some $s \geqslant 1$, then $\mathbb{P}(\|X\|_s > c) \leqslant c^{-s}\mathbb{P}(\|X\|_s^s) = c^{-s}\|X\|^s_{\mathscr{L}_s^k(\mathbb{P})}$.* □

§02.19 **Property** *(Monotone convergence).* *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of monotonically increasing real $\mathscr{L}_1(\mathbb{P})$-random variables converging $\mathbb{P}$-a.s. to a numerical random variable X, for short $X_n \uparrow X$ $\mathbb{P}$-a.s.. Then $\mathbb{P}X = \lim_{n\to\infty} \mathbb{P}X_n$.* □

§02.20 **Property** *(Dominated convergence).* *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of real $\mathscr{L}_1(\mathbb{P})$-random variables converging $\mathbb{P}$-a.s. to a numerical random variable X, i.e., $X_n \xrightarrow{\mathbb{P}\text{-a.s.}} X$. If there is a real $\mathscr{L}_1(\mathbb{P})$ random variable Y with $\sup_{n\in\mathbb{N}} |X_n| \leqslant Y$ $\mathbb{P}$-a.s. (and thus $\sup_{n\in\mathbb{N}} |X_n| \in \mathscr{L}_1(\mathbb{P})$), then $X \in \mathscr{L}_1(\mathbb{P})$ and $X_n \xrightarrow{\mathscr{L}_1(\mathbb{P})} X$.* □

§02.21 **Definition.** A sequence of random variables $(X_n)_{n\in\mathbb{N}}$ with values in a metric space $(\mathcal{X}, d)$ is called *(uniformly) tight* (straff) or *bounded in probability*, if, for any $\varepsilon \in \mathbb{R}^+_{\setminus 0}$, there exists a compact set $K_\varepsilon \subseteq \mathcal{X}$ such that $\mathbb{P}(X_n \in K_\varepsilon) \geqslant 1 - \varepsilon$ for all $n \in \mathbb{N}$. □

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

§02.23 **Example.** A sequence $(X_n)_{n\in\mathbb{N}}$ of random vectors in $\mathbb{R}^k$ 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 $n \in \mathbb{N}$. □

§02.24 **Property** *(Prohorov's theorem).* *Let X and X_n, $n \in \mathbb{N}$, be random variables with values in a Polish space.*

(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.* □

§02.25 **Landau notation.** Let X_n, $n \in \mathbb{N}$, be random variables on a probability space $(\Omega, \mathscr{A}, \mathbb{P})$ with values in a metric space $(\mathcal{X}, d)$ and let x_n, $n \in \mathbb{N}$, belong to $\mathcal{X}$.

(i) We write (a) $x_n = \mathrm{o}(1)$, if $d(x_n, 0) \xrightarrow{n\to\infty} 0$, and (b) $x_n = \mathrm{O}(1)$, if $\sup_{n\in\mathbb{N}} d(x_n, 0) < \infty$, and analogously (a) $X_n = \mathrm{o}_{\mathbb{P}}(1)$, if $X_n \xrightarrow{\mathbb{P}} 0$, and (b) $X_n = \mathrm{O}_{\mathbb{P}}(1)$, if $(X_n)_{n\in\mathbb{N}}$ is bounded in probability

(ii) Let a_n, $n \in \mathbb{N}$, be strictly positive numbers. We write (a) $x_n = \mathrm{o}(a_n)$, if $d(x_n, 0)/a_n = \mathrm{o}(1)$, and that (b) $x_n = \mathrm{O}(a_n)$, if $d(x_n, 0)/a_n = \mathrm{O}(1)$, and analogously (a) $X_n = \mathrm{o}_{\mathbb{P}}(a_n)$, if $d(X_n, 0)/a_n = \mathrm{o}_{\mathbb{P}}(1)$, and (b) $X_n = \mathrm{O}_{\mathbb{P}}(a_n)$, if $d(X_n, 0)/a_n = \mathrm{O}_{\mathbb{P}}(1)$.

(iii) Let A_n, $n \in \mathbb{N}$, be strictly positive random variables on $(\Omega, \mathscr{A}, \mathbb{P})$. We write (a) $X_n = \mathrm{o}_{\mathbb{P}}(A_n)$, if $d(X_n, 0)/A_n = \mathrm{o}_{\mathbb{P}}(1)$, and (b) $X_n = \mathrm{O}_{\mathbb{P}}(A_n)$, if $d(X_n, 0)/A_n = \mathrm{O}_{\mathbb{P}}(1)$. □

§02.26 **Property** *(Exercise).* *For real random variables the following properties hold:*

(i) $\mathrm{o}_{\mathbb{P}}(1) + \mathrm{o}_{\mathbb{P}}(1) = \mathrm{o}_{\mathbb{P}}(1)$ *meaning if* $X_n = \mathrm{o}_{\mathbb{P}}(1)$ *and* $Y_n = \mathrm{o}_{\mathbb{P}}(1)$ *then* $X_n + Y_n = \mathrm{o}_{\mathbb{P}}(1)$*;*

(ii) $\mathrm{O}_{\mathbb{P}}(1) + \mathrm{o}_{\mathbb{P}}(1) = \mathrm{O}_{\mathbb{P}}(1)$*;*

(iii) $\mathrm{O}_{\mathbb{P}}(1) \cdot \mathrm{o}_{\mathbb{P}}(1) = \mathrm{o}_{\mathbb{P}}(1)$*;*

(iv) $(1 + \mathrm{o}_{\mathbb{P}}(1))^{-1} = \mathrm{O}_{\mathbb{P}}(1)$*;*

(v) $\mathrm{o}_{\mathbb{P}}(\mathrm{O}_{\mathbb{P}}(1)) = \mathrm{o}_{\mathbb{P}}(1)$ *meaning if* $X_n = \mathrm{O}_{\mathbb{P}}(1)$ *and* $Y_n = \mathrm{o}_{\mathbb{P}}(X_n)$ *then* $Y_n = \mathrm{o}_{\mathbb{P}}(1)$. □

§03 Conditional expectation

In the reminder of this section let $(\Omega, \mathscr{A}, \mathbb{P})$ be a probability space, $\mathbb{E}$ be the expectation with respect to $\mathbb{P}$ and $\mathscr{F} \subseteq \mathscr{A}$ be a sub-σ-field of $\mathscr{A}$.

§03.01 **Notation.** We write shortly $X \in \overline{\mathscr{A}}^+$, if X is a positive numerical random variable on $(\Omega, \mathscr{A})$, i.e., $X : \Omega \to \overline{\mathbb{R}}^+$ is a $\mathscr{A}$-$\overline{\mathscr{B}}^+$-measurable function. In particular, we have $\overline{\mathscr{F}}^+ \subseteq \overline{\mathscr{A}}^+$ and for $Y \in \overline{\mathscr{F}}^+$ its expectation $\mathbb{E}(Y)$ is well-defined. □

§03.02 **Property.** *For every $X \in \overline{\mathscr{A}}^+$ exists $Y \in \overline{\mathscr{F}}^+$ with $\mathbb{E}(\mathbb{1}_F Y) = \mathbb{E}(\mathbb{1}_F X)$ for all $F \in \mathscr{F}$, where Y is unique up to $\mathbb{P}$-a.s. equality.* □

§03.03 **Definition.** A map $Y : \Omega \to \overline{\mathbb{R}}^+$ is called a (version of the) *conditional expectation* of $X \in \overline{\mathscr{A}}^+$ given $\mathscr{F}$, symbolically $\mathbb{E}\big(X\big|\mathscr{F}\big) := Y$, if

(CE1) Y is $\mathscr{F}$-$\overline{\mathscr{B}}^+$-measurable, hence $Y \in \overline{\mathscr{F}}^+$ and

(CE2) $\mathbb{E}(\mathbb{1}_F Y) = \mathbb{E}(\mathbb{1}_F X)$ for any $F \in \mathscr{F}$.

Any map $\mathbb{E}\big(\bullet\,\big|\mathscr{F}\big) : \overline{\mathscr{A}}^+ \to \overline{\mathscr{F}}^+$ with $X \mapsto \mathbb{E}\big(X\big|\mathscr{F}\big)$ is called (version of the) *conditional expectation* with respect to $\mathbb{P}$ given $\mathscr{F}$. It implies a map $\mathbb{P}\big(\bullet\,\big|\mathscr{F}\big) : \mathscr{A} \to \overline{\mathscr{F}}^+$ with $A \mapsto \mathbb{P}\big(A\big|\mathscr{F}\big) := \mathbb{E}\big(\mathbb{1}_A\big|\mathscr{F}\big)$ called (version of the) *conditional distribution* of $\mathbb{P}$ given $\mathscr{F}$. Exploiting (CE2) every version satisfies $\mathbb{E}(\mathbb{1}_F \mathbb{P}\big(A\big|\mathscr{F}\big)) = \int_F \mathbb{P}\big(A\big|\mathscr{F}\big) d\mathbb{P} = \mathbb{P}(F \cap A)$ for all $F \in \mathscr{F}$ and $A \in \mathscr{A}$. □

§03.04 **Reminder.** Let $X \in \overline{\mathscr{A}}$ be a numerical random variable. Considering the decomposition $X = X^+ - X^-$ with $X^+, X^- \in \overline{\mathscr{A}}^+$ we define for X with $\mathbb{P}(|X|) < \infty$, hence $\mathbb{E}(X^+) < \infty$ and $\mathbb{E}(X^-) < \infty$, the expectation $\mathbb{E}(X) := \mathbb{E}(X^+) - \mathbb{E}(X^-)$. Keep in mind that $\mathscr{L}_1(\mathscr{A}, \mathbb{P}) := \{X \in \overline{\mathscr{A}} : \mathbb{E}(|X|) < \infty\}$ and $\mathbb{E} : \mathscr{L}_1(\mathscr{A}, \mathbb{P}) \to \mathbb{R}$ denotes the uniquely determined expectation with respect to $\mathbb{P}$. Note that $\overline{\mathscr{F}} \subseteq \overline{\mathscr{A}}$ implies $\mathscr{L}_1(\mathscr{F}, \mathbb{P}) \subseteq \mathscr{L}_1(\mathscr{A}, \mathbb{P})$. Let $X \in \mathscr{L}_1(\mathscr{A}, \mathbb{P})$, and hence $\mathbb{E}(X^+) < \infty$ and for any version $\mathbb{E}\big(X^+\big|\mathscr{F}\big)$ holds (CE1), $\mathbb{E}\big(X^+\big|\mathscr{F}\big) \in \overline{\mathscr{F}}^+$ and (CE2), $\mathbb{E}(\mathbb{1}_F \mathbb{E}\big(X^+\big|\mathscr{F}\big)) = \mathbb{E}(\mathbb{1}_F X^+)$ for all $F \in \mathscr{F}$, in particular with $F = \Omega$ also $\mathbb{E}(\mathbb{E}\big(X^+\big|\mathscr{F}\big)) = \mathbb{E}(X^+) < \infty$. Therewith, $\mathbb{E}\big(X^+\big|\mathscr{F}\big) \in \mathscr{L}_1(\mathscr{F}, \mathbb{P})$ and analogously also for any version $\mathbb{E}\big(X^-\big|\mathscr{F}\big) \in \mathscr{L}_1(\mathscr{F}, \mathbb{P})$. Consequently, $\mathbb{E}\big(X^+\big|\mathscr{F}\big) - \mathbb{E}\big(X^-\big|\mathscr{F}\big) \in \mathscr{L}_1(\mathscr{F}, \mathbb{P})$ satisfies (CE2) too. □

§03.05 **Definition.** For $X \in \mathscr{L}_1(\mathscr{A}, \mathbb{P})$ and each version $\mathbb{E}\big(X^+\big|\mathscr{F}\big), \mathbb{E}\big(X^-\big|\mathscr{F}\big) \in \mathscr{L}_1(\mathscr{F}, \mathbb{P})$ we call $\mathbb{E}\big(X\big|\mathscr{F}\big) := \mathbb{E}\big(X^+\big|\mathscr{F}\big) - \mathbb{E}\big(X^-\big|\mathscr{F}\big) \in \mathscr{L}_1(\mathscr{F}, \mathbb{P})$ a (version of the) *conditional expectation* of X given $\mathscr{F}$. Any map

$$\mathbb{E}\big(\bullet\,\big|\mathscr{F}\big) : \mathscr{L}_1(\mathscr{A}, \mathbb{P}) \to \mathscr{L}_1(\mathscr{F}, \mathbb{P}) \text{ with } X \mapsto \mathbb{E}\big(X\big|\mathscr{F}\big) := \mathbb{E}\big(X^+\big|\mathscr{F}\big) - \mathbb{E}\big(X^-\big|\mathscr{F}\big)$$

is called a (version of the) *conditional expectation* with respect to $\mathbb{P}$ given $\mathscr{F}$. □

§03.06 **Remark.** Due to Property §03.02 versions of the conditional expectation of $X \in \overline{\mathscr{A}}^+$ or $X \in \mathscr{L}_1(\mathscr{A}, \mathbb{P})$ given $\mathscr{F}$ differ only on null sets. This property does in generally not extend to the version of the conditional expectation with respect to $\mathbb{P}$ given $\mathscr{F}$, since for each X we obtain a null set, and their union in general is not a null set. □

§03.07 **Definition.** Let $(\Omega_1, \mathscr{A}_1)$, $(\Omega_2, \mathscr{A}_2)$ be measurable spaces. A map $\kappa : \Omega_1 \times \mathscr{A}_2 \to \mathbb{R}^+$ is called *Markov kernel* (from $(\Omega_1, \mathscr{A}_1)$ to $(\Omega_2, \mathscr{A}_2)$), if

(MK1) $A_2 \mapsto \kappa(\omega_1, A_2)$ is for all $\omega_1 \in \Omega_1$ a probability measure on $(\Omega_2, \mathscr{A}_2)$, symbolically $\kappa(\omega_1, \bullet) \in \mathcal{W}(\mathscr{A}_2)$;

(MK2) $\omega_1 \mapsto \kappa(\omega_1, A_2)$ is $\mathscr{A}_1$-$\mathscr{B}$-measurable for all $A_2 \in \mathscr{A}_2$, symbolically $\kappa(\bullet, A_2) \in \mathscr{A}_1^+$. □

§03.08 **Notation.** Consider a probability space $(\Omega_1, \mathscr{A}_1, \mathbb{P})$, a measurable space $(\Omega_2, \mathscr{A}_2)$ and a Markov kernel κ (from $(\Omega_1, \mathscr{A}_1)$ to $(\Omega_2, \mathscr{A}_2)$). Then there exists an unique probability measure $\kappa \odot \mathbb{P}$ on $(\Omega_2 \times \Omega_1, \mathscr{A}_2 \otimes \mathscr{A}_1)$ determined by

$$\kappa \odot \mathbb{P}(A_2 \times A_1) = \int_{A_1} \kappa(\omega_1, A_2)\mathbb{P}(d\omega_1), \quad \text{for all } A_1 \in \mathscr{A}_1, A_2 \in \mathscr{A}_2.$$

If $f \in \overline{\mathscr{A}_2 \otimes \mathscr{A}_1}^+$ or $f \in \mathscr{L}_1(\kappa \odot \mathbb{P})$ then

$$\kappa \odot \mathbb{P} f = \int_{\Omega_2 \times \Omega_1} f(\omega_2, \omega_1)\kappa \odot \mathbb{P}(d\omega_2, d\omega_1) = \int_{\Omega_1} \int_{\Omega_2} f(\omega_2, \omega_1)\kappa(\omega_1, d\omega_2)\mathbb{P}(d\omega_1).$$

Furthermore, we denote by $\kappa\mathbb{P}$ the marginal distribution on $(\Omega_2, \mathscr{A}_2)$ induced by $\kappa \odot \mathbb{P}$, i.e. $\kappa\mathbb{P}(A_2) = \kappa \odot \mathbb{P}(A_2 \times \Omega_1) = \int_{\Omega_1} \kappa(\omega_1, A_2)\mathbb{P}(d\omega_1)$ for all $A_2 \in \mathscr{A}_2$. □

§03.09 **Definition.**

(a) $\mathbb{P}\big(\bullet \,\big|\mathscr{F}\big)$ is called *regular* (version of the) conditional distribution of $\mathbb{P}$ given $\mathscr{F}$, if $(\omega, A) \mapsto \mathbb{P}\big(A\big|\mathscr{F}\big)(\omega)$ satisfies the conditions (MK1) and (MK2), i.e. $\mathbb{P}\big(\bullet\big|\mathscr{F}\big)$ is a Markov kernel (from $(\Omega, \mathscr{F})$ to $(\Omega, \mathscr{A})$).

(b) $\mathbb{E}\big(\bullet \,\big|\mathscr{F}\big)$ is called *regular* (version of the) conditional expectation with respect to $\mathbb{P}$ given $\mathscr{F}$, if the implied conditional distribution $\mathbb{P}\big(\bullet \,\big|\mathscr{F}\big)$ of $\mathbb{P}$ given $\mathscr{F}$ is regular, and for each $\omega \in \Omega$ is $X \mapsto \mathbb{E}\big(X\big|\mathscr{F}\big)(\omega)$ the expectation with respect to $\mathbb{P}\big(\bullet \,\big|\mathscr{F}\big)(\omega)$. □

§03.10 **Property.**

(i) *Each regular conditional distribution of $\mathbb{P}$ given $\mathscr{F}$ is implied by a regular conditional expectation with respect to $\mathbb{P}$ given $\mathscr{F}$.*

(ii) *For any probability measure $\mathbb{P}$ on a polish space (Ω, d) endowed with its Borel-σ-algebra $\mathscr{B}_\Omega$ and sub-σ-field $\mathscr{F} \subseteq \mathscr{B}_\Omega$ exists a regular conditional distribution of $\mathbb{P}$ given $\mathscr{F}$.* □

§03.11 **Notation.**

(i) Let X be a random variable on $(\Omega, \mathscr{A}, \mathbb{P})$ with values in a measurable space $(\mathcal{X}, \mathscr{X})$. For $h \in \mathscr{L}_1(\mathcal{X}, \mathbb{P}^X)$ denotes $\mathbb{E}^X\big(h\big|\mathscr{F}\big) := \mathbb{E}\big(h(X)\big|\mathscr{F}\big) \in \mathscr{L}_1(\mathscr{F}, \mathbb{P})$ a conditional expectation of $h(X)$ given $\mathscr{F}$ and $\mathbb{E}^X\big(\bullet\,\big|\mathscr{F}\big) : \mathscr{L}_1(\mathcal{X}, \mathbb{P}^X) \to \mathscr{L}_1(\mathscr{F}, \mathbb{P})$ with $h \mapsto \mathbb{E}^X\big(h\big|\mathscr{F}\big)$ a *(regular)* (version of the) *conditional expectation* with respect to $\mathbb{P}^X$ given $\mathscr{F}$.

(ii) Let S be a random variable on $(\Omega, \mathscr{A}, \mathbb{P})$ with values in a measurable space $(\mathcal{S}, \mathscr{S})$. For $h \in \mathscr{L}_1(\overline{\mathscr{A}}, \mathbb{P})$ we call $\mathbb{E}\big(h\big|\sigma(S)\big) \in \mathscr{L}_1(\sigma(S), \mathbb{P})$ be a conditional expectation of h given $\mathscr{F} = \sigma(S)$. Keeping $\mathbb{E}\big(h\big|\sigma(S)\big) \in \overline{\sigma(S)}$ in mind and applying Property §01.02 (iv) there is $\varphi \in \overline{\mathscr{S}}$ with $\mathbb{E}\big(h\big|\sigma(S)\big) = \varphi(S)$, that is, $\mathbb{E}\big(h\big|\sigma(S)\big)(\omega) = \varphi(S(\omega))$, $\omega \in \Omega$. Then $\mathbb{E}\big(h\big|S\big) := \varphi \in \mathscr{L}_1(\mathscr{S}, \mathbb{P}^S)$ and $\mathbb{E}\big(h\big|S = s\big) := \varphi(s) \in \overline{\mathbb{R}}$ is called a (version of the) *conditional expectation* of h given S respectively $S = s$, and $\mathbb{E}\big(\bullet\,\big|S\big) : \mathscr{L}_1(\mathscr{A}, \mathbb{P}) \to \mathscr{L}_1(\mathscr{S}, \mathbb{P}^S)$ with $X \mapsto \mathbb{E}\big(X\big|S\big)$ a *(regular)* (version of the) *conditional expectation* with respect to $\mathbb{P}$ given S.

(iii) Let $(X, S) : (\Omega, \mathscr{A}) \to (\mathcal{X} \times \mathcal{S}, \mathscr{X} \otimes \mathscr{S})$ with joint distribution $\mathbb{P}^{(X,S)}$. We denote by $\Pi_X : \mathcal{X} \times \mathcal{S} \to \mathcal{X}$ and $\Pi_S : \mathcal{X} \times \mathcal{S} \to \mathcal{S}$ with $(x, s) \mapsto \Pi_X(x, s) := x$ and $(x, s) \mapsto \Pi_S(x, s) := s$, respectively, the corresponding coordinate maps. The marginal distribution of X respectively S is given by $\mathbb{P}^X = \mathbb{P} \circ X^{-1} = \mathbb{P} \circ \Pi_X^{-1}(X, S) = \mathbb{P}^{(X,S)} \circ \Pi_X^{-1}$ respectively

$\mathbb{P}^S = \mathbb{P}^{(X,S)} \circ \Pi_S^{-1}$. For each version $\mathbb{P}^{(X,S)}\big(\bullet\,\big|\sigma(\Pi_S)\big)$ of the conditional distribution with respect to $\mathbb{P}^{(X,S)}$ given $\sigma(\Pi_S)$, the map

$$\mathbb{P}^X\big(\bullet\,\big|S\big) : \mathscr{X} \to \overline{\mathscr{S}} \text{ with } B \mapsto \mathbb{P}^X\big(B\big|S\big) := \varphi \text{ determined by}$$
$$\mathbb{P}^X\big(B\big|\sigma(\Pi_S)\big) = \mathbb{P}^{(X,S)}\big(\Pi_X^{-1}(B)\big|\sigma(\Pi_S)\big) = \varphi(\Pi_S)$$

and analogously $\mathbb{P}^X\big(\bullet\,\big|S=s\big)$ is called (version of the) *conditional distribution* of X given S respectively $S=s$. We call a version *regular*, if $(s,B) \mapsto \mathbb{P}^X\big(B\big|S=s\big)$ is a Markov kernel (from $(\mathcal{S},\mathscr{S})$ to $(\mathcal{X},\mathscr{X})$), where due to Definition §03.03 (CE2) $\mathbb{P}^X\big(\bullet\,\big|S\big) \odot \mathbb{P}^S = \mathbb{P}^{(X,S)}$ (see Notation §03.08). Analogously, for $h \in \mathscr{L}_1(\mathscr{X},\mathbb{P}^X)$ we define a (regular) version $\mathbb{E}^X\big(h\big|S\big) \in \mathscr{L}_1(\mathscr{S},\mathbb{P}^S)$ and $\mathbb{E}^X\big(h\big|S=s\big) \in \overline{\mathbb{R}}$ of the conditional expectation of h given S respectively $S=s$. If $\mathbb{P}^X\big(\bullet\,\big|S\big)$ is a regular conditional distribution of X given S and for $s \in \mathcal{S}$ the probability measure $\mathbb{P}^X\big(\bullet\,\big|S=s\big)$ has for example a finite first absolute moment, i.e., $\mathbb{P}^X\big(\bullet\,\big|S=s\big) \in \mathcal{W}_1(\mathscr{B}^n)$ (see Notation §01.05) then $\mathbb{E}\big(X\big|S=s\big) = \mathbb{E}^X\big(\mathrm{id}_{\mathcal{X}}\big|S=s\big) = \int_{\mathcal{X}} x\mathbb{P}^X\big(dx\big|S=s\big)$.

(iv) Suppose the joint distribution $\mathbb{P}^{(X,S)}$ is dominated by a product measure $\mu\otimes\nu$ where μ and ν is a σ-finite measure on $\mathscr{X}$ and $\mathscr{S}$, respecitively, $\mu \in \mathcal{M}_\sigma(\mathscr{X})$ and $\nu \in \mathcal{M}_\sigma(\mathscr{S})$ for short. Let $\mathbb{f}^{(X,S)}$ denote a $(\mu\otimes\nu)$-density of $\mathbb{P}^{(X,S)}$. A μ- and ν-density of the marginal distribution $\mathbb{P}^X$ and $\mathbb{P}^S$ is given by $\mathbb{f}^X : x \mapsto \int_{\mathcal{S}} \mathbb{f}^{(X,S)}(x,s)\nu(ds)$ and $\mathbb{f}^S : s \mapsto \int_{\mathcal{X}} \mathbb{f}^{(X,S)}(x,s)\mu(dx)$, respectively. The $\mathbb{f}^{X|S} : \mathcal{S}\times\mathcal{X} \to \overline{\mathbb{R}}^+$ with

$$(s,x) \mapsto \mathbb{f}^{X|S=s}(x) = \frac{\mathbb{f}^{(X,S)}(x,s)}{\mathbb{f}^S(s)}\mathbb{1}_{\{\mathbb{f}^S(s)>0\}} + \mathbb{f}^X(x)\mathbb{1}_{\{\mathbb{f}^S(s)=0\}}$$

belongs to $\overline{\mathscr{S}\otimes\mathscr{X}}^+$ and it is a *μ-density of the Markov kernel* $\mathbb{P}^{X|S}$ from $(\mathcal{S},\mathscr{S})$ to $(\mathcal{X},\mathscr{X})$ defined by $(s,B) \mapsto \mathbb{P}^{X|S=s}(B) := \int_B \mathbb{f}^{X|S=s}(x)\mu(dx)$. We call $\mathbb{f}^{X|S=s}$ *conditional density* of X given $S=s$.

(v) As an example let $(X,S) \in \mathscr{B}^{k+l}$ be multivariate normally distributed with $\mathbb{C}\mathrm{ov}(X,S) = \Sigma_{XS}$ and marginal distributions $X \sim \mathrm{N}_{(\mu_X,\Sigma_X)}$ and $S \sim \mathrm{N}_{(\mu_S,\Sigma_S)}$, i.e.,

$$\begin{pmatrix} X \\ S \end{pmatrix} \sim \mathrm{N}_{(\mu,\Sigma)} \text{ with } \mu = \begin{pmatrix} \mu_X \\ \mu_S \end{pmatrix} \in \mathbb{R}^{k+l} \text{ and } \Sigma = \begin{pmatrix} \Sigma_X & \Sigma_{XS} \\ \Sigma_{XS}^t & \Sigma_S \end{pmatrix}.$$

Assuming $\Sigma > 0$ the joint distribution $\mathbb{P}^{(X,S)}$ admits a density with respect to the Lebesgue measure λ^{k+l} on $(\mathbb{R}^{k+l},\mathscr{B}^{k+l})$. For each $s \in \mathbb{R}^l$ the conditional density $\mathbb{f}^{X|S=s}$ as in (iv) is a density of the multivariate normal distribution $\mathrm{N}_{(\mu_{X|S=s},\Sigma_{X|S=s})}$-distribution with

$$\mu_{X|S=s} := \mu_X + \Sigma_{XS}\Sigma_S^{-1}(s-\mu_S) \in \mathbb{R}^k \text{ und } \Sigma_{X|S=s} := \Sigma_X - \Sigma_{XS}\Sigma_S^{-1}\Sigma_{SX} > 0$$

which is thus a regular conditional distribution of X given $S=s$. □

§03.12 **Property**. *Let $X,Y \in \mathscr{L}_1(\mathscr{A},\mathbb{P})$ and $\mathscr{F} \subseteq \mathscr{A}$ be a sub-σ-field. Any version of the conditional expectation satisfies the following properties $\mathbb{P}$-a.s.:*

(i) *For all $a,b \in \mathbb{R}$ holds $\mathbb{E}\big(aX+bY\big|\mathscr{F}\big) = a\mathbb{E}\big(X\big|\mathscr{F}\big) + b\mathbb{E}\big(Y\big|\mathscr{F}\big)$;* *(linear)*

(ii) *For $X \leqslant Y$ holds $\mathbb{E}\big(X\big|\mathscr{F}\big) \leqslant \mathbb{E}\big(Y\big|\mathscr{F}\big)$;* *(monotone)*

(iii) $\big|\mathbb{E}\big(X\big|\mathscr{F}\big)\big| \leqslant \mathbb{E}\big(|X|\big|\mathscr{F}\big)$; *(triangular inequality)*

(iv) *For $S \in \overline{\mathscr{A}}$ with $\mathbb{E}\big(|S|\big|\mathscr{F}\big) < \infty$ holds $\mathbb{P}(|S| < \infty) = 1$.* *(finite)*

(v) *For* $\phi : \mathbb{R} \to \mathbb{R}$ *convex with* $\phi(X) \in \mathscr{L}_1(\overline{\mathscr{A}}, \mathbb{P})$ *holds* $\phi\big(\mathbb{E}\big(X\big|\mathscr{F}\big)\big) \leqslant \mathbb{E}\big(\big(\phi(X)\big)\big|\mathscr{F}\big)$. *(Jensen's inequality)*

(vi) *For* $X_n \uparrow X$ $\mathbb{P}$-a.s. *holds* $\sup_{n\in\mathbb{N}} \mathbb{E}\big(X_n\big|\mathscr{F}\big) = \mathbb{E}\big(X\big|\mathscr{F}\big)$. *(monotone convergence)*

(vii) *For* $X_n \to X$ $\mathbb{P}$-a.s. *with* $|X_n| \leqslant Y$, $n \in \mathbb{N}$, *holds* $\lim_{n\to\infty} \mathbb{E}\big(X_n\big|\mathscr{F}\big) = \mathbb{E}\big(X\big|\mathscr{F}\big)$ $\mathbb{P}$-a.s. *and in* $\mathscr{L}_1(\mathscr{A}, \mathbb{P})$. *(dominated convergence)*

If the version is regular, i.e., $\mathbb{E}\big(\,\bullet\,\big|\mathscr{F}\big)(\omega)$ *is an expectation for all* $\omega \in \Omega$*, then the statements* (i)-(vii) *holds for all* $\omega \in \Omega$. □

§03.13 **Property**. *Let* $X, Y \in \mathscr{L}_1(\mathscr{A}, \mathbb{P})$ *and* $\mathscr{G} \subseteq \mathscr{F} \subseteq \mathscr{A}$ *sub-*σ*-fields. Any version of the conditional expectation satisfies the following properties* $\mathbb{P}$-a.s.*:*

(i) *For* $\mathbb{E}(|XY|) < \infty$ *and* $Y \in \mathscr{F}$ *holds*

$$\mathbb{E}\big(XY\big|\mathscr{F}\big) = Y\mathbb{E}\big(X\big|\mathscr{F}\big) \text{ and } \mathbb{E}\big(Y\big|\mathscr{F}\big) = \mathbb{E}\big(Y\big|\sigma(Y)\big) = Y;$$

(ii) $\mathbb{E}\big(\mathbb{E}\big(X\big|\mathscr{F}\big)\big|\mathscr{G}\big) = \mathbb{E}\big(\mathbb{E}\big(X\big|\mathscr{G}\big)\big|\mathscr{F}\big) = \mathbb{E}\big(X\big|\mathscr{G}\big)$; *(tower property)*

(iii) *If* $\sigma(X)$ *and* $\mathscr{F}$ *are independent, then* $\mathbb{E}\big(X\big|\mathscr{F}\big) = \mathbb{E}(X)$; *(independence)*

(iv) $\mathbb{E}(\mathbb{E}\big(X\big|\mathscr{F}\big)) = \mathbb{E}(X)$. *(total probability)*

(v) *For* $\overline{\mathscr{T}} := \big\{A \in \mathscr{A} \,\big|\, \mathbb{P}(A) \in \{0,1\}\big\}$ *holds* $\mathbb{E}\big(X\big|\overline{\mathscr{T}}\big) = \mathbb{E}(X)$. □

§03.14 **Property**. *Let* $\mathscr{F} \subseteq \mathscr{A}$ *be a sub-*σ*-field and* $\mathbb{E}\big(\,\bullet\,\big|\mathscr{F}\big)$ *be a conditional expectation.*

(i) $\mathbb{E}\big(\,\bullet\,\big|\overline{\mathscr{F}}\big) : \mathscr{L}_2(\mathscr{A}, \mathbb{P}) \to \mathscr{L}_2(\mathscr{F}, \mathbb{P})$ *is an orthogonal projection, that is, for all* $X \in \mathscr{L}_2(\mathscr{A}, \mathbb{P})$ *and* $Y \in \mathscr{L}_2(\mathscr{F}, \mathbb{P})$ *holds*

$$\|X - Y\|^2_{\mathscr{L}_2(\mathbb{P})} = \mathbb{E}(|X - Y|^2) \geqslant \mathbb{E}(|X - \mathbb{E}\big(X\big|\mathscr{F}\big)|^2) = \|X - \mathbb{E}\big(X\big|\mathscr{F}\big)\|^2_{\mathscr{L}_2(\mathbb{P})},$$

where equality holds if and only if $Y = \mathbb{E}\big(X\big|\mathscr{F}\big)$ $\mathbb{P}$-a.s..

(ii) $\mathbb{E}\big(\bullet\big|\mathscr{F}\big) : \mathscr{L}_s(\mathscr{A}, \mathbb{P}) \to \mathscr{L}_s(\mathscr{F}, \mathbb{P})$ *is a contraction for* $s \in [1, \infty]$*, i.e.,* $\|\mathbb{E}\big(X\big|\mathscr{F}\big)\|_{\mathscr{L}_s(\mathbb{P})} \leqslant \|X\|_{\mathscr{L}_s(\mathbb{P})}$*, and thus bounded and continuous. If* $(X_n)_{n\in\mathbb{N}}$ *converges in* $\mathscr{L}_s(\mathscr{A}, \mathbb{P})$*, then* $(\mathbb{E}\big(X_n\big|\mathscr{F}\big))_{n\in\mathbb{N}}$ *converges in* $\mathscr{L}_s(\mathscr{F}, \mathbb{P})$. □

Chapter 2

Asymptotic properties of M- and Z-estimators

Asymptotic properties of M- and Z-estimators are presented generalising the minimum contrast approach introduced in the lecture Statistik 1. For a more detailed exposition we refer to the text book van der Vaart [1998].

§04 Introduction / motivation / illustration

§04.01 **Example** *(Linear model).* The dependence of the variation of a real random variable Y_1 (response) on the variation of a random vector $X_1 = (X_{1j})_{j\in[\![k]\!]}$ in $\mathbb{R}^k$ (explanatory variable) is often described by a linear relationship $\mathbb{E}\big(Y_1\big|X_1\big) = \sum_{j\in[\![k]\!]}\gamma_j X_{1j} = X_1^t\gamma$ or equivalently $Y_1 = X_1^t\gamma + \varepsilon_1$ where ε_1 is a real random error satisfying $\mathbb{E}\big(\varepsilon_1\big|X_1\big) = 0$. We aim to infer on the unknown parameter of interest $\gamma \in \mathbb{R}^k$ from $n \in \mathbb{N}$ i.i.d. copies (Y_i, X_i), $i \in [\![n]\!]$. Writing $Y := (Y_i)_{i\in[\![n]\!]}$ and $X^t = (X_1 \cdots X_n)$ we have $\mathbb{E}\big(Y\big|X\big) = X\gamma$. Any (measurable) choice

$$\widehat{\gamma} \in \operatorname*{arg\,inf}_{\gamma\in\mathbb{R}^k} \widehat{\mathrm{M}}_n(\gamma) \quad \text{with } \widehat{\mathrm{M}}_n(\gamma) := \tfrac{1}{n}\sum_{i\in[\![n]\!]}(Y_i - X_i^t\gamma)^2 = \tfrac{1}{n}\|Y - X\gamma\|^2 \tag{04.1}$$

is called a Least Squares Estimator (LSE), where arg inf denotes the subset of vectors in $\mathbb{R}^k$ attaining the function's smallest value. If $X^tX = \sum_{i\in[\![n]\!]} X_iX_i^t$ is strictly positive definite (hence, invertible) then $\widehat{\gamma} = (X^tX)^{-1}X^tY = \big(\sum_{i\in[\![n]\!]} X_iX_i^t\big)^{-1}\sum_{i\in[\![n]\!]} Y_iX_i$ is the unique LSE. Under "usual" conditions (Example §02.14) holds $\frac{1}{n}\sum_{i\in[\![n]\!]} X_iX_i^t \xrightarrow{\mathbb{P}} \mathbb{E}(X_1X_1^t) =: \Omega$ (LLN). If in addition $\mathbb{E}(\varepsilon_i^2|X_i) = \sigma^2$, then $\frac{1}{\sqrt{n}}\sum_{i\in[\![n]\!]}\varepsilon_iX_i \xrightarrow{d} \mathrm{N}_{(0,\sigma^2\Omega)}$ (CLT). Applying Slutzky's lemma §02.10 and the continuous mapping theorem §02.09 holds $\sqrt{n}(\widehat{\gamma} - \gamma) \xrightarrow{d} \mathrm{N}_{(0,\sigma^2\Omega^{-1})}$ for $\Omega > 0$. Further inference on $\widehat{\gamma}$ (hypothesis testing, confidence intervals, etc.) is typically based on this asymptotic result. However, a linear relationship $\mathbb{E}\big(Y\big|X\big) = X\gamma$ is often too restrictive. □

§04.02 **Example** *(Generalised linear model).* Consider a real random variable Y_1 and a random vector X_1 in $\mathbb{R}^k$ obeying $\mathbb{E}\big(Y_1\big|X_1\big) = g(X_1^t\gamma)$ for a known link function $g : \mathbb{R} \to \mathbb{R}$. We aim to infer on the unknown parameter of interest $\gamma \in \mathbb{R}^k$ from $n \in \mathbb{N}$ i.i.d. copies (Y_i, X_i), $i \in [\![n]\!]$. As an illustration let us consider the effect of three different drugs on the behaviour of certain animals. In a trial each drug is given in different dose to certain animals and the number of effected animals is counted. The Table 1.1 summarises the results. Let Y_{jk} denote the counts of an effect among n_{jk} animals applying a log-dose X_{jk}, $j \in [\![J_k]\!]$ of the drug $k \in [\![K]\!]$. Assuming an "independent and identical" behaviour of the n_{jk} animals it seems reasonable to model Y_{jk} as Binomial-distributed random variable, $Y_{jk} \sim \mathrm{Bin}_{(n_{jk},\pi_{jk})}$ for short, with unknown percentage $\pi_{jk} \in (0,1)$. It may be reasonable to assume that $n_{jk}\pi_{jk} = \mathbb{E}\big(Y_{jk}\big|X_{jk}\big) = g(\gamma_k + \gamma_0X_{jk})$ where $(\gamma_k)_{k\in[\![K]\!]}$ is a drug specific factor and γ_0 is a common effect of the log-dose for all drugs. The model is called "probit" and "logit", respectively, if g is the standard-normal distribution

function and the logit-distribution function ($x \mapsto \frac{e^x}{1+e^x}$). As in Example §04.01 inference on $\gamma = (\gamma_k)_{k\in[\![0,K]\!]}$ is often based on a LSE, i.e., any (measurable) choice $\widehat{\gamma} \in \arg\inf_{\gamma\in\mathbb{R}^{K+1}} \widehat{\mathrm{M}}_n(\gamma)$ with $\widehat{\mathrm{M}}_n(\gamma) := \frac{1}{K}\sum_{k\in[\![K]\!]} \frac{1}{J_K}\sum_{j\in[\![J_k]\!]}(Y_{jk} - g(\gamma_k + \gamma_0 X_{jk}))^2$.

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.

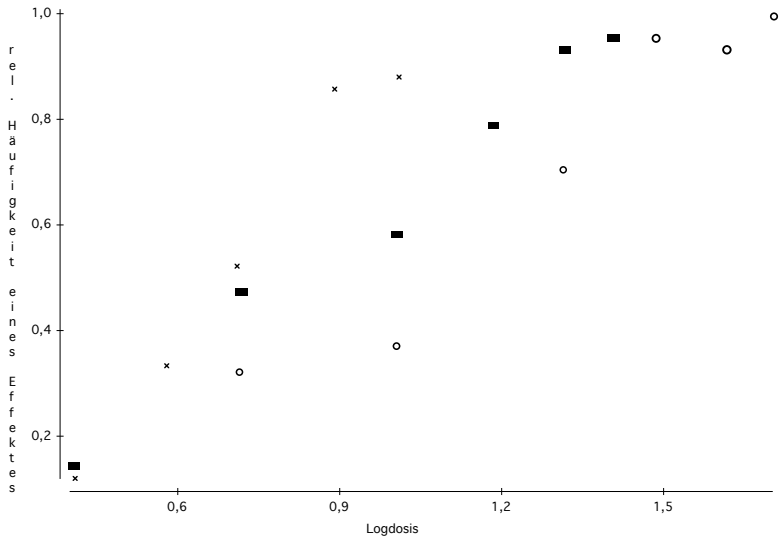
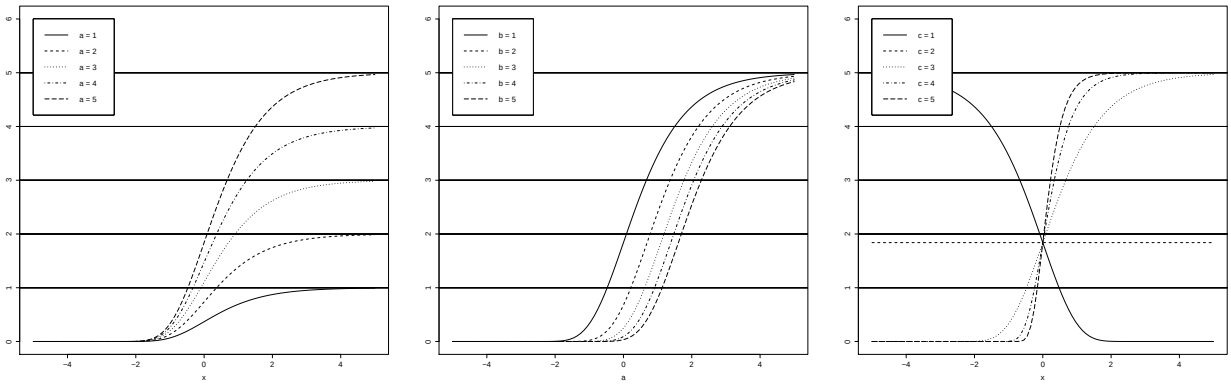


Figure 1.1: Relative frequency of the effects in dependence of the log-dose, drug 1: x; 2: ○; 3: -. □

§04.03 **Example** *(Nonlinear regression).* Consider a real random variable Y_1 and a random vector X_1 in $\mathbb{R}^k$ obeying $\mathbb{E}\big(Y_1\big|X_1\big) = g(X_1,\gamma)$ for a given link function $g : \mathbb{R}^k \times \mathbb{R}^p \to \mathbb{R}$. We aim to infer on the unknown parameter $\gamma \in \mathbb{R}^p$ from $n \in \mathbb{N}$ i.i.d. copies (Y_i, X_i), $i \in [\![n]\!]$. The next figure shows the widely used Gompertz function $g(x,(a,b,c)) = a\exp(-b\exp(x\log(c)))$.



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.

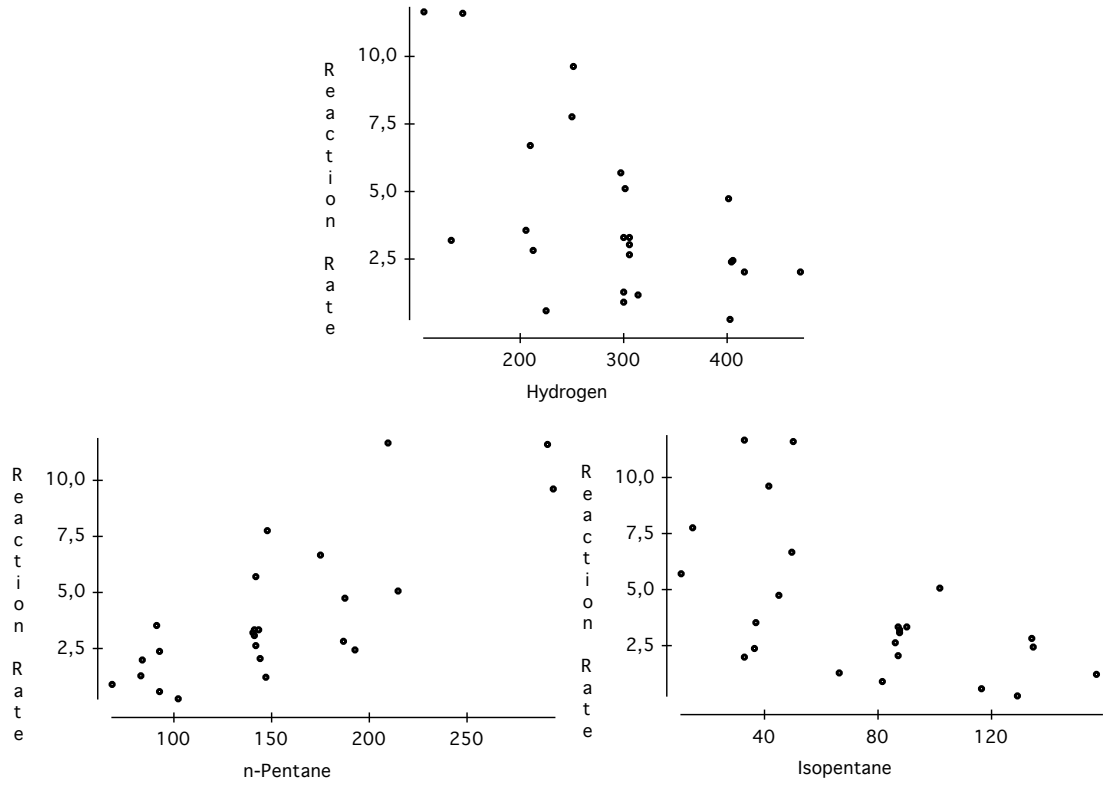


Figure 1.3: Reaction rate in dependence of the partial hydrogen, n-pentane and isopentane pressure.

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

case is given by

$$\mathbb{E}\big(Y_i\big|(X_{i1},X_{i2},X_{i3})\big)=\frac{\gamma_1\gamma_3(X_{i2}-X_{i3}/1.632)}{1+\gamma_2X_{i1}+\gamma_3X_{i2}+\gamma_4X_{i3}},\quad i\in[\![n]\!],\tag{04.2}$$

where X_{i1}, X_{i2} and X_{i3} is the partial pressure of hydrogen, isopentane and n-pentane, respectively, and $(\gamma_j)_{j\in[\![4]\!]}$ is the unknown parameter of interest. As in Example §04.01 inference on γ is often based on a LSE, i.e., any (measurable) choice $\widehat{\gamma}\in\arg\inf_{\gamma\in\mathbb{R}^4}\widehat{\mathrm{M}}_n(\gamma)$ with $\widehat{\mathrm{M}}_n(\gamma):=\frac{1}{n}\sum_{i\in[\![n]\!]}(Y_i-g(X_i,\gamma))^2$. □

§04.04 **Example** *(Quantile regression).* Consider a real random variable Y_1 and a random vector X_1 in $\mathbb{R}^k$ obeying $Y_1=X_1^t\gamma+\varepsilon_1$ with quantile condition $\mathbb{P}\big(\varepsilon_1\leqslant 0\big|X_1\big)=\alpha$ for a given probability $\alpha\in(0,1)$ or equivalently $\mathbb{P}\big(Y_1\leqslant X_1^t\gamma\big|X_1\big)=\alpha$ meaning that the conditional-α-quantile of Y_1 given X_1 equals $X_1^t\gamma$. Let q_α denote the α-quantile of $\mathbb{P}^Z\in\mathcal{W}(\mathscr{B})$, i.e., $\mathbb{P}(Z\leqslant q_\alpha)=\alpha$. Define $\tau_\alpha(z):=(1-\alpha)z^-+\alpha z^+$ 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, the α-quantile satisfies $0=\frac{\partial}{\partial q}\mathbb{E}(\tau_\alpha(Z-q))\big|_{q=q_\alpha}$, since

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

Thereby, a reasonable estimator of γ is any (measurable) choice $\widehat{\gamma}\in\arg\inf_{\gamma\in\mathbb{R}^k}\widehat{\mathrm{M}}_n(\gamma)$ with $\widehat{\mathrm{M}}_n(\gamma)=\frac{1}{n}\sum_{i\in[\![n]\!]}\tau_\alpha(Y_i-X_i^t\gamma)$. □

§04.05 **Example** *(Generalised Method of Moments).* Given a random vector Z_1 in $\mathbb{R}^p$ and a function $\mathrm{h}^J=(\mathrm{h}_j)_{j\in[\![J]\!]}:\mathbb{R}^k\times\mathbb{R}^p\to\mathbb{R}^J$ let the unknown parameter of interest $\gamma\in\mathbb{R}^k$ satisfy $\mathbb{P}^{Z_1}\mathrm{h}_j(\gamma)=\mathbb{E}\big(\mathrm{h}_j(\gamma,Z_1)\big)=0$ for all $j\in[\![J]\!]$, or $\mathbb{P}^{Z_1}\mathrm{h}^J(\gamma)=\mathbb{E}\big(\mathrm{h}^J(\gamma,Z_1)\big)=0$ for short. Supposing an i.i.d. sample $(Z_i)_{i\in[\![n]\!]}$ any (measurable) choice $\widehat{\gamma}$ satisfying $\widehat{\mathbb{P}}_n\mathrm{h}_j(\widehat{\gamma})=\frac{1}{n}\sum_{i\in[\![n]\!]}\mathrm{h}_j(\widehat{\gamma},Z_i)=0$ for all $j\in[\![J]\!]$, or $\widehat{\mathrm{H}}_n(\widehat{\gamma})=0$ with $\widehat{\mathrm{H}}_n(\gamma):=\widehat{\mathbb{P}}_n\mathrm{h}^J(\gamma)=\frac{1}{n}\sum_{i\in[\![n]\!]}\mathrm{h}^J(\gamma,Z_i)$, $\gamma\in\mathbb{R}^k$, for short, is called moment estimator. In case a moment estimator does not exist, setting $\widehat{\mathrm{M}}_n(\gamma):=(\widehat{\mathbb{P}}_n\mathrm{h}^J(\gamma))^tW_n(\widehat{\mathbb{P}}_n\mathrm{h}^J(\gamma))$ for a given weighting matrix W_n one might consider any (measurable) choice $\widehat{\gamma}\in\arg\inf_{\gamma\in\mathbb{R}^k}\widehat{\mathrm{M}}_n(\gamma)$ called a Generalised Method of Moments (GMM) estimator. □

§04.06 **Reminder**. Denote by $\mathcal{W}(\mathscr{X})$ the set of all probability measures on a measurable space $(\mathcal{X},\mathscr{X})$. For a non-empty index set Θ a family $\mathbb{P}_\Theta:=(\mathbb{P}_\theta)_{\theta\in\Theta}$ of probability measures on $\mathscr{X}$ is formally defined by the map $\Theta\to\mathcal{W}(\mathscr{X})$ with $\theta\mapsto\mathbb{P}_\theta$. Here and subsequently, for each $\theta\in\Theta$ denotes $\mathbb{E}_\theta$ the expectation with respect to $\mathbb{P}_\theta$. For a random variable X taking its values in $(\mathcal{X},\mathscr{X})$ we write shortly $X\circledcirc\mathbb{P}_\Theta$, if $X\sim\mathbb{P}_\theta$ for some $\theta\in\Theta$. If the random variables $\{X_i,i\in[\![n]\!]\}$ form an *independent and identically distributed* (i.i.d.) sample of $X\sim\mathbb{P}$ with values in $(\mathcal{X},\mathscr{X})$, then $\mathbb{P}^{\otimes n}=\otimes_{j\in[\![n]\!]}\mathbb{P}$ denotes the joint product probability measure of the family $(X_i)_{i\in[\![n]\!]}$ taking its values in the measurable product space $(\mathcal{X}^n,\mathscr{X}^{\otimes n})$. We write $\{X_i,i\in[\![n]\!]\}\overset{i.i.d.}{\sim}\mathbb{P}$ or $(X_i)_{i\in[\![n]\!]}\sim\mathbb{P}^{\otimes n}$ for short. We denote by $\mathbb{P}_\Theta^{\otimes n}:=(\mathbb{P}_\theta^{\otimes n})_{\theta\in\Theta}$ a family of product probability measures on $\mathscr{X}^{\otimes n}$. Any random variable S on $(\mathcal{X},\mathscr{X})$ taking values in a measurable space

$(\mathcal{S},\mathscr{S})$, i.e., $\mathscr{X}$-$\mathscr{S}$-measurable function $S:\mathcal{X}\to\mathcal{S}$, is called *observation* or *statistic*. We denote by $\mathbb{P}_\Theta^S := (\mathbb{P}_\theta^S)_{\theta\in\Theta}$ the family of probability measures on $(\mathcal{S},\mathscr{S})$ induced by S. A map $\gamma:\Theta\to\Gamma$ and its value $\gamma(\theta)$ for each $\theta\in\Theta$ is called *parameter* and *parameter value of interest*, respectively. A *parameter of interest* $\gamma:\Theta\to\Gamma$ is called *identifiable*, if for any $\theta_1,\theta_2\in\Theta$ from $\gamma(\theta_1)\neq\gamma(\theta_2)$ follows $\mathbb{P}_{\theta_1}\neq\mathbb{P}_{\theta_2}$. □

§04.07 **Definition**. The triple $(\mathcal{X},\mathscr{X},\mathbb{P}_\Theta)$ is called a *statistical experiment* or *statistical model*. The non-empty set Θ and $\mathcal{X}$ is called *parameter* and *sample space*, respectively. A statistical model $(\mathcal{X},\mathscr{X},\mathbb{P}_\Theta)$ is called *adequate* for a random variable X, if $X\circlearrowright\mathbb{P}_\Theta$. Given a family $\mathbb{P}_\Theta^{\otimes n}$ of product probability measures $(\mathcal{X}^n,\mathscr{X}^{\otimes n},\mathbb{P}_\Theta^{\otimes n})$ is called a *statistical product experiment*. We denote by $(\mathcal{S},\mathscr{S},\mathbb{P}_\Theta^S)$ the statistical model induced by a $(\mathcal{S},\mathscr{S})$-valued statistic S on $(\mathcal{X},\mathscr{X})$. A statistic $\widehat{\gamma}$ on $(\mathcal{X},\mathscr{X})$ with values in the measurable space $(\Gamma,\mathscr{G})$ is called *estimator* or *estimation function* for the identifiable parameter of interest γ. A statistical model $(\mathcal{X},\mathscr{X},\mathbb{P}_\Theta)$ (and the family $\mathbb{P}_\Theta$) is called *dominated*, if a σ-finite measure μ on $\mathscr{X}$ exists, $\mu\in\mathcal{M}_\sigma(\mathscr{X})$ for short, such that for each $\theta\in\Theta$ the probability measure $\mathbb{P}_\theta$ is absolutely continuous with respect to μ, i.e., $\mathbb{P}_\theta\ll\mu$. We write shortly $\mathbb{P}_\Theta\ll\mu$. Any version of the Radon-Nikodym densities

$$\mathrm{L}(\theta,x) := \frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mu}(x)\quad x\in\mathcal{X},\ \theta\in\Theta$$

considered as function of θ parametrised by x is called *likelihood* or *likelihood function* where typically it is understand as a random function $\mathrm{L}:\Theta\to\overline{\mathcal{X}}^+$ with $\theta\mapsto\mathrm{L}(\theta) := \mathrm{L}(\theta,\bullet)$. Its logarithm $\ell := \log\mathrm{L}$ (with convention $\log(0) := -\infty$) is called *log-likelihood* or *log-likelihood function*. The *likelihood* and *log-likelihood* in the corresponding dominated product experiment $(\mathcal{X}^n,\mathscr{X}^{\otimes n},\mathbb{P}_\Theta^{\otimes n})$ are $\prod_{i\in[\![n]\!]}\mathrm{L}(\theta,x_i)$ and $\sum_{i\in[\![n]\!]}\ell(\theta,x_i)$, $\theta\in\Theta$, $x^n\in\mathcal{X}^n$, respectively. □

§04.08 **Reminder**. Let $(\mathcal{X},\mathscr{X},\mathbb{P}_\Theta)$ be dominated by $\mu\in\mathcal{M}_\sigma(\mathscr{X})$. If μ is finite, then $\mu\ll\mathbb{P}_\mu := \frac{1}{\mu(\mathcal{X})}\mu\in\mathcal{W}(\mathscr{X})$ and hence $\mathbb{P}_\Theta$ is also dominated by $\mathbb{P}_\mu$. If μ is not finite, then there exists a countable and measurable partition $\{\mathcal{X}_m, m\in\mathbb{N}\}$ of $\mathcal{X}$ with $0<\mu(\mathcal{X}_m)<\infty$ for all $m\in\mathbb{N}$. For each $m\in\mathbb{N}$ define $\mathbb{P}_\mu(\bullet|\mathcal{X}_m)\in\mathcal{W}(\mathscr{X})$ with $A\mapsto\mathbb{P}_\mu(A|\mathcal{X}_m) := \frac{\mu(A\cap\mathcal{X}_m)}{\mu(\mathcal{X}_m)}$. Then holds $\mu\ll\mathbb{P}_\mu := \sum_{m\in\mathbb{N}}2^{-m}\mathbb{P}_\mu(\bullet|\mathcal{X}_m)\in\mathcal{W}(\mathscr{X})$, since $\mathbb{P}_\mu(A)=0$ implies $\mu(A\cap\mathcal{X}_m)=0$ for all $m\in\mathbb{N}$ and thus $\mu(A)=0$. Therewith, we have shown, that for each $\mu\in\mathcal{M}_\sigma(\mathscr{X})$ there is $\mathbb{P}_\mu\in\mathcal{W}(\mathscr{X})$ with $\mu\ll\mathbb{P}_\mu$ which automatically dominates $\mathbb{P}_\Theta$ too. On the other hand, there is a probability measure $\mathbb{P}_o=\sum_{i\in\mathbb{N}}c_i\mathbb{P}_{\theta_i}$ with $c_i\in\mathbb{R}^+$, $\theta_i\in\Theta$ for all $i\in\mathbb{N}$ and $\sum_{i\in\mathbb{N}}c_i=1$, and thus $\mathbb{P}_o\ll\mu$, such that $\mathbb{P}_\theta\ll\mathbb{P}_o$ for all $\theta\in\Theta$ (e.g. Statistik 1, Satz §11.04). We call any such probability measure $\mathbb{P}_o$ *privileged dominating measure*. Therefore, we eventually assume with out loss of generality that the dominating measure is indeed a probability measure. □

§04.09 **Example** *(MLE)*. Let $(\mathcal{X},\mathscr{X},\mathbb{P}_\Theta)$ be a statistical model dominated by $\mu\in\mathcal{M}_\sigma(\mathscr{X})$ with likelihood $\mathrm{L}(\theta)=\mathrm{d}\mathbb{P}_\theta/\mathrm{d}\mu$ and log-likelihood $\ell(\theta)=\log\mathrm{L}(\theta)$ for $\theta\in\Theta$ and let $(\Theta,\mathscr{T})$ be a measurable space. Any statistic $\widehat{\theta}$ on $(\mathcal{X},\mathscr{X})$ with values in $(\Theta,\mathscr{T})$ is called Maximum-Likelihood-Estimator (MLE) for θ, if $\mathrm{L}(\widehat{\theta})=\sup_{\theta\in\Theta}\mathrm{L}(\theta)$ μ-a.s. meaning $\mathrm{L}(\widehat{\theta}(x),x)=\sup_{\theta\in\Theta}\mathrm{L}(\theta,x)$ for μ-a.e. $x\in\mathcal{X}$, or equivalently $\ell(\widehat{\theta})=\sup_{\theta\in\Theta}\ell(\theta)$ μ-a.s.. Considering a statistical product experiment $(\mathcal{X}^n,\mathscr{X}^{\otimes n},\mathbb{P}_\Theta^{\otimes n})$ dominated by $\mu^{\otimes n}\in\mathcal{M}_\sigma(\mathscr{X}^{\otimes n})$ and setting $\widehat{\mathrm{M}}_n(\theta) := -\widehat{\mathbb{P}}_n\ell(\theta)$, i.e. $\widehat{\mathrm{M}}_n(\theta,x^n)=-\frac{1}{n}\sum_{i\in[\![n]\!]}\ell(\theta,x_i)$ for $x^n\in\mathcal{X}^n$, the MLE $\widehat{\theta}$ is determined by $\widehat{\theta}\in\arg\inf_{\theta\in\Theta}\widehat{\mathrm{M}}_n(\theta)$ μ-a.s.. However, in general it is not guaranteed that MLE is unique or even exits. The MLE depends on the version of the likelihood, but there exists often a canonical choice. Furthermore,

$\gamma(\widehat{\theta})$ is called MLE for a parameter of interest $\gamma : \Theta \to \Gamma$, if $\gamma(\widehat{\theta})$ is a statistic on $(\mathcal{X}^n, \mathscr{X}^{\otimes n})$ with values in $(\Gamma, \mathscr{G})$. □

§04.10 **Remark.** In all the examples the estimator $\widehat{\gamma}$ of the parameter of interest γ is determined by $\widehat{\gamma} \in \arg\inf_{\gamma\in\Gamma} \widehat{\mathrm{M}}_n(\gamma)$ for some random function $\gamma \mapsto \widehat{\mathrm{M}}_n(\gamma) \in \overline{\mathscr{X}}$ of the data. Obviously, rather than minimising (or maximising) a criterion function we might search for a zero of the associated normal or estimating equations, that is, $\widehat{\gamma}$ is determined as a zero of a random vector function $\gamma \mapsto \widehat{\mathrm{H}}_n(\gamma) \in \overline{\mathscr{X}}^k$. Note that estimator is defined $\mathbb{P}_\Theta$-a.s. only, meaning that one can change the estimator on a $\mathbb{P}_\Theta$-zero set N, i.e., $\mathbb{P}_\theta(N) = 0$ for all $\theta \in \Theta$. □

§04.11 **Definition.** Let $(\mathcal{X}_n, \mathscr{X}_n, \mathbb{P}_\Theta^n = (\mathbb{P}_\theta^n)_{\theta\in\Theta})$ for all $n \in \mathbb{N}$ be a statistical model over the same parameter space Θ and let $\gamma : \Theta \to \Gamma$ be a parameter of interest. We call a function $\mathrm{M} : \Theta\times\Gamma \to \overline{\mathbb{R}}$ and $\mathrm{H} : \Theta \times \Gamma \to \overline{\mathbb{R}}^k$ *criterion function*, if for all $\theta \in \Theta$ the function $\mathrm{M}(\theta) : \gamma \mapsto \mathrm{M}(\theta, \gamma)$, respectively $\mathrm{H}(\theta) : \gamma \mapsto \mathrm{H}(\theta, \gamma)$, has in $\gamma(\theta)$ an unique minimum, respectively an unique zero. A sequence $(\widehat{\mathrm{M}}_n)_{n\in\mathbb{N}}$ and $(\widehat{\mathrm{H}}_n)_{n\in\mathbb{N}}$ of functions $\widehat{\mathrm{M}}_n : \Gamma \times \mathcal{X}_n \to \overline{\mathbb{R}}$ and $\widehat{\mathrm{H}}_n : \Gamma \times \mathcal{X}_n \to \overline{\mathbb{R}}^k$ is called *random criterion function* or *criterion process*, if the following two conditions are satisfied:

(CP1) For all $\gamma \in \Gamma$ is $\widehat{\mathrm{M}}_n(\gamma) : x \mapsto \widehat{\mathrm{M}}_n(\gamma, x)$, respectively $\widehat{\mathrm{H}}_n(\gamma) : x \mapsto \widehat{\mathrm{H}}_n(\gamma, x)$, a statistic, that is, $\widehat{\mathrm{M}}_n(\gamma) \in \overline{\mathscr{X}_n}$, respectively $\widehat{\mathrm{H}}_n(\gamma) \in \overline{\mathscr{X}_n}^k$.

(CP2) For all $\gamma \in \Gamma$ and $\theta \in \Theta$ it holds $\widehat{\mathrm{M}}_n(\gamma) \xrightarrow{\mathbb{P}_\theta^n} \mathrm{M}(\theta, \gamma)$, respectively $\widehat{\mathrm{H}}_n(\gamma) \xrightarrow{\mathbb{P}_\theta^n} \mathrm{H}(\theta, \gamma)$.

Every (measurable) choice $\widehat{\gamma}_n : \mathcal{X}_n \to \Gamma$ (if it exists) is called a *M*-estimator, respectively a *Z*-estimator, if it satisfies

$$\widehat{\mathrm{M}}_n(\widehat{\gamma}_n) = \inf_{\gamma\in\Gamma} \widehat{\mathrm{M}}_n(\gamma) \quad \mathbb{P}_\Theta^n\text{-a.s.}, \qquad \text{respectively} \qquad \widehat{\mathrm{H}}_n(\widehat{\gamma}_n) = 0 \quad \mathbb{P}_\Theta^n\text{-a.s.},$$

or more generally, if it is, respectively, a near minimum and near zero, that is, $\widehat{\mathrm{M}}_n(\widehat{\gamma}_n) \leqslant \inf_{\gamma\in\Gamma} \widehat{\mathrm{M}}_n(\gamma) + \mathrm{o}_{\mathbb{P}_\theta^n}(1)$ and $\widehat{\mathrm{H}}_n(\widehat{\gamma}_n) = \mathrm{o}_{\mathbb{P}_\theta^n}(1)$. □

§04.12 **Remark.** There exists a measurable version of a minimum of an almost surely continuous function on a compact set (see Witting and Müller-Funk [1995], Satz 6.7). Note that in Definition §04.11 the criterion process $\widehat{\mathrm{M}}_n$ (respectively $\widehat{\mathrm{H}}_n$) is defined for each $n \in \mathbb{N}$ on a different measurable space. We write, however, shortly $\widehat{\mathrm{M}}_n(\gamma) \xrightarrow{\mathbb{P}_\theta^n} \mathrm{M}(\theta, \gamma)$, if for each $\varepsilon \in \mathbb{R}_{\setminus 0}^+$ holds $\mathbb{P}_\theta^n(|\widehat{\mathrm{M}}_n(\gamma) - \mathrm{M}(\theta, \gamma)| > \varepsilon) \xrightarrow{n\to\infty} 0$. Let us briefly consider a sample $(X_i)_{i\in[\![n]\!]} \wr\!\!\!\sim \mathbb{P}_\Theta^{\otimes n}$ of a random variable $X \wr\!\!\!\sim \mathbb{P}_\Theta$. Keeping Notation §01.05 in mind $\mathbb{P}f$ and $\widehat{\mathbb{P}}_n f$ denotes the integral of $f \in \mathscr{L}_1(\mathscr{X}, \mathbb{P})$ with respect to $\mathbb{P}$ and the empirical measure $\widehat{\mathbb{P}}_n(x) = \frac{1}{n}\sum_{i\in[\![n]\!]} \delta_{x_i}$, $x \in \mathcal{X}^n$, respectively. Revisiting each of the Examples §04.01 to §04.04 there is a function $\mathrm{m} : \Gamma\times\mathcal{X} \to \overline{\mathbb{R}}$ with $\mathrm{m}(\gamma) \in \mathscr{L}_1(\mathscr{X})$, $\gamma \in \Gamma$, such that the criterion process $\widehat{\mathrm{M}}_n$ and the associated criterion function M is for each $\gamma \in \Gamma$ given by $\widehat{\mathrm{M}}_n(\gamma) = \widehat{\mathbb{P}}_n\mathrm{m}(\gamma)$, i.e. $\widehat{\mathrm{M}}_n(\gamma, x^n) = \frac{1}{n}\sum_{i\in[\![n]\!]} \mathrm{m}(\gamma, x_i)$, $x^n \in \mathcal{X}^n$, and $\mathrm{M}(\theta, \gamma) = \mathbb{P}_\theta\mathrm{m}(\gamma) = \int_\mathcal{X} \mathrm{m}(\gamma, x)\mathbb{P}_\theta(dx)$, respectively. Analogously, a moment estimator as in Example §04.05 is a Z-estimator. By construction in each example is the condition (CP1) and with the help of the LLN (see Remark §02.06) also the condition (CP2) satisfied. Note that the GMM estimator in Example §04.05 is also a M-estimator with criterion process satisfying (CP1) and (CP2). □

§04.13 **Definition**. For two probability measure $\mathbb{P}$ and $\mathbb{Q}$ on a measurable space $(\mathcal{X}, \mathscr{X})$ is the function

$$\mathrm{KL}(\mathbb{P}|\mathbb{Q}) = \begin{cases} \mathbb{P}\big(\log \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\big) = \int \log\big(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\big) d\mathbb{P}, & \text{if } \mathbb{P} \ll \mathbb{Q}, \\ +\infty, & \text{otherwise} \end{cases}$$

called *Kullback-Leibler-divergence* of $\mathbb{P}$ with respect to $\mathbb{Q}$. □

§04.14 **Reminder**. The Kullback-Leibler-divergence satisfies $\mathrm{KL}(\mathbb{P}|\mathbb{Q}) \geqslant 0$ as well as $\mathrm{KL}(\mathbb{P}|\mathbb{Q}) = 0$ if and only if $\mathbb{P} = \mathbb{Q}$, but $\mathrm{KL}(\bullet|\bullet)$ is not symmetric. Moreover, for product measures holds $\mathrm{KL}(\mathbb{P}_1 \otimes \mathbb{P}_2|\mathbb{Q}_1 \otimes \mathbb{Q}_2) = \mathrm{KL}(\mathbb{P}_1|\mathbb{Q}_1) + \mathrm{KL}(\mathbb{P}_2|\mathbb{Q}_2)$ (e.g. Statistik 1, Lemma §20.03). □

§04.15 **Example** *(MLE, §04.09 continued.)*. Let $(\mathcal{X}^n, \mathscr{X}^{\otimes n}, \mathbb{P}_\Theta^{\otimes n})$ be a statistical product experiment dominated by a privileged measure $\mathbb{P}_o \in \mathcal{W}(\mathscr{X})$ (see Reminder §04.08) with likelihood $\mathrm{L}(\theta) = \mathrm{d}\mathbb{P}_\theta/\mathrm{d}\mathbb{P}_o$, log-likelihood $\ell = \log(\mathrm{L})$ and parameter of interest θ (i.e., $\gamma = \mathrm{id}_\Theta$). Furthermore, for all $\theta, \theta_o \in \Theta$ let $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta_o}$ be mutually dominated (i.e. $\mathbb{P}_\theta \ll \mathbb{P}_{\theta_o}$ and $\mathbb{P}_{\theta_o} \ll \mathbb{P}_\theta$, for short $\mathbb{P}_\theta \lll \mathbb{P}_{\theta_o}$), which implies $\mathbb{P}_{\theta_o} \lll \mathbb{P}_o$, and hence $-\mathrm{KL}(\mathbb{P}_{\theta_o}|\mathbb{P}_o) = \mathrm{KL}(\mathbb{P}_o|\mathbb{P}_{\theta_o})$. Then $\widehat{\mathrm{M}}_n(\theta) := -\widehat{\mathbb{P}}_n\ell(\theta) \in \overline{\mathscr{X}^{\otimes n}}$ with

$$x^n \mapsto \widehat{\mathrm{M}}_n(\theta, x^n) = -\frac{1}{n}\sum_{i\in[\![n]\!]} \ell(\theta, x_i)$$

is a criterion process associated to the criterion function $\mathrm{M}(\theta_o, \theta) := \mathrm{KL}(\mathbb{P}_{\theta_o}|\mathbb{P}_\theta) - \mathrm{KL}(\mathbb{P}_{\theta_o}|\mathbb{P}_o)$ assuming 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 $\theta \mapsto \mathrm{M}(\theta_o, \theta)$ attains its minimum $\mathrm{M}(\theta_o, \theta_o) = -\mathrm{KL}(\mathbb{P}_{\theta_o}|\mathbb{P}_o)$ uniquely at θ_o (keeping Reminder §04.14 in mind). The corresponding M-estimator is thus just a MLE. □

§05 Consistency

Here and subsequently, let (Γ, d) be a metric space endowed with its Borel-σ-algebra $\mathscr{G} := \mathscr{B}_\Gamma$, let $(\mathcal{X}_n, \mathscr{X}_n, \mathbb{P}_\Theta^n = (\mathbb{P}_\theta^n)_{\theta\in\Theta})$ for all $n \in \mathbb{N}$ be a statistical model over the parameter space Θ and let $\gamma : \Theta \to \Gamma$ be an identifiable parameter of interest.

§05.01 **Reminder**. For each $n \in \mathbb{N}$ let $\widehat{\gamma}_n$ be an estimator of γ, i.e. a statistic on $(\mathcal{X}_n, \mathscr{X}_n)$ with values in $(\Gamma, \mathscr{G})$. The sequence $(\widehat{\gamma}_n)_{n\in\mathbb{N}}$ of estimators is called *(weakly) consistent*, if for all $\varepsilon \in \mathbb{R}^+_{\setminus 0}$ holds $\mathbb{P}_\theta^n(d(\widehat{\gamma}_n, \gamma(\theta)) > \varepsilon) \xrightarrow{n\to\infty} 0$ for all $\theta \in \Theta$. Note that the estimator $\widehat{\gamma}_n$ can be defined for each $n \in \mathbb{N}$ on a different measurable space. We write, however, shortly $d(\widehat{\gamma}_n, \gamma(\theta)) \xrightarrow{\mathbb{P}_\theta^n} 0$. Moreover, saying *„$\widehat{\gamma}_n$ is consistent“* always means the sequence $(\widehat{\gamma}_n)_{n\in\mathbb{N}}$ is (weakly) consistent. □

Consider an M-estimator $\widehat{\gamma}_n$ for a random criterion function $\widehat{\mathrm{M}}_n$ with associated criterion function M, that is, $\widehat{\mathrm{M}}_n(\gamma) \xrightarrow{\mathbb{P}_\theta^n} \mathrm{M}(\theta, \gamma)$ holds point-wise for each $\gamma \in \Gamma$. For example, due to the LLN $\widehat{\mathrm{M}}_n(\gamma) = \widehat{\mathbb{P}}_n\mathrm{m}(\gamma) \xrightarrow{\mathbb{P}_\theta^{\otimes n}} \mathbb{P}_\theta\mathrm{m}(\gamma) = \mathrm{M}(\theta, \gamma)$ provided $\mathrm{m}(\gamma) \in \mathscr{L}_1(\mathscr{X}, \mathbb{P}_\theta)$. The hope is that a minimising value of $\widehat{\mathrm{M}}_n(\gamma)$ then converges to the minimising value of $\mathrm{M}(\theta, \gamma)$. However, in general point-wise convergence will not be sufficient.

§05.02 **Theorem**. *Under the assumptions and notations of Definition §04.11 any M-estimator $\widehat{\gamma}_n$ of γ, i.e., $\widehat{\mathrm{M}}_n(\widehat{\gamma}_n) \leqslant \widehat{\mathrm{M}}_n(\gamma(\theta)) + \mathrm{o}_{\mathbb{P}_\theta^n}(1)$, is consistent, i.e., $d(\widehat{\gamma}_n, \gamma(\theta)) = \mathrm{o}_{\mathbb{P}_\theta^n}(1)$, if in addition the following two conditions are satisfied:*

(CO1) $\sup_{\gamma\in\Gamma}|\widehat{\mathrm{M}}_n(\gamma) - \mathrm{M}(\theta,\gamma)| = \mathrm{o}_{\mathbb{P}_\theta^n}(1)$ *(uniform convergence in probability);*

(CO2) $\inf_{\gamma\in\Gamma: d(\gamma,\gamma(\theta))\geqslant\varepsilon} \mathrm{M}(\theta,\gamma) > \mathrm{M}(\theta,\gamma(\theta))$ *for any* $\varepsilon \in \mathbb{R}_{\setminus 0}^+$ *(identification).*

§05.03 **Proof** of Theorem §05.02. is given in the lecture. □

§05.04 **Corollary**. *Under the assumptions and notations of Definition §04.11 any Z-estimator $\widehat{\gamma}_n$ of γ, i.e., $\widehat{\mathrm{H}}_n(\widehat{\gamma}_n) = \mathrm{o}_{\mathbb{P}_\theta^n}(1)$, is consistent, i.e., $d(\widehat{\gamma}_n, \gamma(\theta)) = \mathrm{o}_{\mathbb{P}_\theta^n}(1)$, if in addition the following two conditions are satisfied:*

(CO1) $\sup_{\gamma\in\Gamma}\|\widehat{\mathrm{H}}_n(\gamma) - \mathrm{H}(\theta,\gamma)\| = \mathrm{o}_{\mathbb{P}_\theta^n}(1)$ *(uniform convergence in probability);*

(CO2) $\inf_{\gamma\in\Gamma: d(\gamma,\gamma(\theta))\geqslant\varepsilon} \|\mathrm{H}(\theta,\gamma)\| > 0 = \|\mathrm{H}(\theta,\gamma(\theta))\|$ *for any* $\varepsilon \in \mathbb{R}_{\setminus 0}^+$ *(identification).*

§05.05 **Proof** of Corollary §05.04. Setting $\widehat{\mathrm{M}}_n(\gamma) = \|\widehat{\mathrm{H}}_n(\gamma)\|$ and $\mathrm{M}(\theta,\gamma) = \|\mathrm{H}(\theta,\gamma)\|$ the claim follows directly from Theorem §05.02. □

§05.06 **Lemma**. *If* (i) Γ *is compact,* (ii) $\mathrm{M}(\theta,\gamma) > \mathrm{M}(\theta,\gamma(\theta))$ *for all* $\gamma \in \Gamma\setminus\{\gamma(\theta)\}$, *and* (iii) $\gamma \mapsto \mathrm{M}(\theta,\gamma)$ *is continuous, then* (CO2) *in Theorem §05.02 holds.*

§05.07 **Proof** of Lemma §05.06. is left as an exercise. □

§05.08 **Example** *(MLE, §04.15 continued).* Assuming in addition that the parameter space Θ is compact and that the criterion function $\theta \mapsto \mathrm{M}(\theta_o,\theta) := \mathrm{KL}(\mathbb{P}_{\theta_o}|\mathbb{P}_\theta) - \mathrm{KL}(\mathbb{P}_{\theta_o}|\mathbb{P}_{\theta_o})$ is continuous then employing Lemma §05.06 the condition (CO2) of Theorem §05.02 is satisfied. □

§05.09 **Lemma**. (CO1) *in Theorem §05.02 is satisfied, if the following conditions hold:*

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

(ii) $\gamma \mapsto \mathrm{M}(\theta,\gamma)$ *is continuous and* $\widehat{\mathrm{M}}_n(\gamma) = \mathrm{M}(\theta,\gamma) + \mathrm{o}_{\mathbb{P}_\theta^n}(1)$ *for all* $\gamma \in \Gamma$, *and*

(iii) $\lim_{\delta\downarrow 0}\limsup_{n\to\infty} \mathbb{P}_\theta^n\big(\sup_{\gamma_1,\gamma_2\in\Gamma: d(\gamma_1,\gamma_2)\leqslant\delta} |\widehat{\mathrm{M}}_n(\gamma_1) - \widehat{\mathrm{M}}_n(\gamma_2)| \geqslant \varepsilon\big) = 0$ *for all* $\varepsilon \in \mathbb{R}_{\setminus 0}^+$.

§05.10 **Proof** of Lemma §05.09. is given in the lecture. □

§05.11 **Example**. Given $(\mathcal{X}^n, \mathscr{X}^{\otimes n}, \mathbb{P}_\Theta^{\otimes n})$ and $\gamma : \Theta \to \Gamma$ for each $\gamma \in \Gamma$ let $\mathrm{m}(\gamma) \in \mathscr{X}$ be a real function $x \mapsto \mathrm{m}(\gamma, x)$ belonging to $\mathscr{L}_1(\mathscr{X}, \mathbb{P}_\theta)$. Consider $\widehat{\mathrm{M}}_n(\gamma) := \widehat{\mathbb{P}}_n\mathrm{m}(\gamma)$, i.e. $\widehat{\mathrm{M}}_n(\gamma, x^n) = \frac{1}{n}\sum_{i\in[\![n]\!]} \mathrm{m}(\gamma, x_i)$, $x^n \in \mathcal{X}^n$, and $\mathrm{M}(\theta,\gamma) := \mathbb{P}_\theta\mathrm{m}(\gamma)$ where due to the LLN §02.06 $\widehat{\mathrm{M}}_n(\gamma) = \mathrm{M}(\theta,\gamma) + \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$ for each $\gamma \in \Gamma$. Suppose in addition the following conditions:

(i) (Γ, d) is a compact metric space,

(ii) $\gamma \mapsto \mathrm{m}(\gamma, x)$ is continuous for $\mathbb{P}_\theta$-a.e. $x \in \mathcal{X}$,

(iii) there is $H \in \mathscr{L}_1(\mathscr{X}, \mathbb{P}_\theta)$ with $\sup_{\gamma\in\Gamma} |\mathrm{m}(\gamma, x)| \leqslant |H(x)|$ for $\mathbb{P}_\theta$-a.e. $x \in \mathcal{X}$, or equivalently, $\sup_{\gamma\in\Gamma} |\mathrm{m}(\gamma)|$ belongs to $\mathscr{L}_1(\mathscr{X}, \mathbb{P}_\theta)$.

Then, (I) $\gamma \mapsto \mathbb{P}_\theta \mathrm{m}(\gamma) = \mathrm{M}(\theta,\gamma)$ is continuous and (CO1) $\sup_{\gamma\in\Gamma}|\widehat{\mathrm{M}}_n(\gamma) - \mathrm{M}(\theta,\gamma)| = \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$. Indeed, by dominated convergence (see §02.20) (ii) and (iii) imply together (I). Consider (CO1). Define the random variable $\Delta_\delta^n := \sup_{\gamma_1,\gamma_2\in\Gamma:d(\gamma_1,\gamma_2)\leqslant\delta}|\widehat{\mathrm{M}}_n(\gamma_1) - \widehat{\mathrm{M}}_n(\gamma_2)| \in \overline{\mathscr{X}^{\otimes n}}$. We show below for all $\varepsilon,\eta \in \mathbb{R}_{\setminus 0}^+$ exists $\delta \in \mathbb{R}_{\setminus 0}^+$ with $\limsup_{n\to\infty}\mathbb{P}_\theta^{\otimes n}\big(\Delta_\delta^n \geqslant \varepsilon\big) \leqslant \eta$ which in turn by Lemma §05.09 implies the claim (CO1). Let $\varepsilon,\eta \in \mathbb{R}_{\setminus 0}^+$. Keeping $\Delta_\delta^1 \in \overline{\mathscr{X}}$ with $x \mapsto \Delta_\delta^1(x) = \sup_{\gamma_1,\gamma_2\in\Gamma:d(\gamma_1,\gamma_2)\leqslant\delta}|\mathrm{m}(\gamma_1,x) - \mathrm{m}(\gamma_2,x)|$ in mind and applying the elementary triangular inequality we have $\Delta_\delta^n \leqslant \widehat{\mathbb{P}}_n\Delta_\delta^1$ point-wise on $\mathcal{X}^n$. Moreover, due to (i) and (ii) for $\mathbb{P}_\theta$-a.e. $x \in \mathcal{X}$ the function $\gamma \mapsto \mathrm{m}(\gamma,x)$ is uniformly continuous on Γ, and thus $\lim_{\delta\to 0}\Delta_\delta^1(x) = 0$. Therewith, dominated convergence (see §02.20), which can be applied due to (iii), implies $\lim_{\delta\to 0}\mathbb{P}_\theta\Delta_\delta^1 = 0$. In particular there is $\delta \in \mathbb{R}_{\setminus 0}^+$ such that $\mathbb{P}_\theta\Delta_\delta^1 \leqslant \eta\varepsilon$, which in turn implies $\mathbb{P}_\theta^{\otimes n}\Delta_\delta^n \leqslant \mathbb{P}_\theta^{\otimes n}(\widehat{\mathbb{P}}_n\Delta_\delta^1) = \mathbb{P}_\theta\Delta_\delta^1 \leqslant \eta\varepsilon$. Employing Markov's inequality §02.18 the last estimate implies the claim, that is, for all $\varepsilon,\eta \in \mathbb{R}_{\setminus 0}^+$ exists $\delta \in \mathbb{R}_{\setminus 0}^+$ with $\limsup_{n\to\infty}\mathbb{P}_\theta^{\otimes n}\big(\Delta_\delta^n \geqslant \varepsilon\big) \leqslant \eta$. If in addition to (i)-(iii) and, hence (I)

(iv) there is $\gamma(\theta) \in \Gamma$ with $\mathrm{M}(\theta,\gamma) > \mathrm{M}(\theta,\gamma(\theta))$ for all $\gamma \in \Gamma\backslash\{\gamma(\theta)\}$,

then applying Lemma §05.06 it holds (CO2) $\inf_{\gamma\in\Gamma:d(\gamma,\gamma(\theta))\geqslant\varepsilon}\mathrm{M}(\theta,\gamma) > \mathrm{M}(\theta,\gamma(\theta))$. To summarise, with (CO1) and (CO2) the conditions of Theorem §05.02 are satisfied. Consequently, any *M-estimator* $\widehat{\gamma}_n$, i.e., $\widehat{\mathrm{M}}_n(\widehat{\gamma}_n) \leqslant \inf_{\gamma\in\Gamma}\widehat{\mathrm{M}}_n(\gamma) + \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$, and thus $\widehat{\mathrm{M}}_n(\widehat{\gamma}_n) \leqslant \widehat{\mathrm{M}}_n(\gamma(\theta)) + \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$, is a *consistent estimator of* γ, i.e., $d(\widehat{\gamma}_n,\gamma(\theta)) = \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$. □

§05.12 **Lemma.** (CO1) *in Corollary §05.04 is satisfied, if the following conditions hold:*

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

(ii) $\gamma \mapsto \mathrm{H}(\theta,\gamma)$ *is continuous and* $\|\widehat{\mathrm{H}}_n(\gamma) - \mathrm{H}(\theta,\gamma)\| = \mathrm{o}_{\mathbb{P}_\theta^n}(1)$ *for all* $\gamma \in \Gamma$, *and*

(iii) $\lim_{\delta\downarrow 0}\limsup_{n\to\infty}\mathbb{P}_\theta^n\big(\sup_{\gamma_1,\gamma_2\in\Gamma:d(\gamma_1,\gamma_2)\leqslant\delta}\|\widehat{\mathrm{H}}_n(\gamma_1) - \widehat{\mathrm{H}}_n(\gamma_2)\| \geqslant \varepsilon\big) = 0$ *for all* $\varepsilon \in \mathbb{R}_{\setminus 0}^+$.

§05.13 **Proof** of Lemma §05.12. is left as an exercise. □

§05.14 **Example.** Given $(\mathcal{X}^n, \mathscr{X}^{\otimes n}, \mathbb{P}_\Theta^{\otimes n})$, $\gamma : \Theta \to \Gamma$ and $(X_i)_{i\in[\![n]\!]} \sim \mathbb{P}_\theta^{\otimes n}$ for $\theta \in \Theta$, for each $\gamma \in \Gamma$ let $\mathrm{h}(\gamma) \in \overline{\mathscr{X}}^k$ be a numerical function belonging to $\mathscr{L}_1(\mathbb{P}_\theta)$ for all $\gamma \in \Gamma$. Consider $\widehat{\mathrm{H}}_n(\gamma) := \widehat{\mathbb{P}}_n\mathrm{h}(\gamma)$, i.e. $\widehat{\mathrm{H}}_n(\gamma,x^n) = \frac{1}{n}\sum_{i\in[\![n]\!]}\mathrm{h}(\gamma,x_i)$, $x^n \in \mathcal{X}^n$, and $\mathrm{H}(\theta,\gamma) := \mathbb{P}_\theta\mathrm{h}(\gamma)$ where due to the LLN §02.06 $\|\widehat{\mathrm{H}}_n(\gamma) - \mathrm{H}(\theta,\gamma)\| = \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$ for each $\gamma \in \Gamma$. Suppose in addition the following conditions:

(i) (Γ,d) is a compact metric space,

(ii) $\gamma \mapsto \mathrm{h}(\gamma,x)$ is continuous for $\mathbb{P}_\theta$-a.e. $x \in \mathcal{X}$,

(iii) $\sup_{\gamma\in\Gamma}\|\mathrm{h}(\gamma)\|$ belongs to $\mathscr{L}_1(\mathbb{P}_\theta)$.

Then, arguing line by line as in Example §05.11 (I) $\gamma \mapsto \mathbb{P}_\theta\mathrm{h}(\gamma) = \mathrm{H}(\theta,\gamma)$ is continuous and (CO1) $\sup_{\gamma\in\Gamma}\|\widehat{\mathrm{H}}_n(\gamma) - \mathrm{H}(\theta,\gamma)\| = \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$. If in addition to (i)-(iii) and hence (I)

(iv) there is $\gamma(\theta) \in \Gamma$ with $\|\mathrm{H}(\theta,\gamma)\| > 0 = \|\mathrm{H}(\theta,\gamma(\theta))\|$ for all $\gamma \in \Gamma\backslash\{\gamma(\theta)\}$,

then applying Lemma §05.06 it holds (CO2) $\inf_{\gamma\in\Gamma:d(\gamma,\gamma(\theta))\geqslant\varepsilon}\|\mathrm{H}(\theta,\gamma)\| > 0 = \|\mathrm{H}(\theta,\gamma(\theta))\|$. To summarise, with (CO1) and (CO2) the conditions of Corollary §05.04 are satisfied. Consequently, any *Z-estimator* $\widehat{\gamma}_n$, i.e., $\widehat{\mathrm{H}}_n(\widehat{\gamma}_n) = \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$ is a *consistent estimator of* γ, i.e., $d(\widehat{\gamma}_n,\gamma(\theta)) = \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$. □

§05.15 **Remark**. The conditions (CO1) and (CO2) of Corollary §05.04 (respectively, (CO1) and (CO2) of Theorem §05.02) being sufficient to ensure consistency might be weakened in specific situations as we see next. □

§05.16 **Proposition**. *Let* $\Gamma \subseteq \mathbb{R}$ *and* $\widehat{\mathrm{H}}_n(\gamma) = \mathrm{H}(\theta,\gamma)+\mathrm{o}_{\mathbb{P}_\theta^n}(1)$ *for all* $\gamma \in \Gamma$ *where* H *is a deterministic function. Assume in addition that either*

(Ia) $\gamma \mapsto \widehat{\mathrm{H}}_n(\gamma)$ *is continuous and has exactly one zero* $\widehat{\gamma}_n$*, or*

(Ib) $\gamma \mapsto \widehat{\mathrm{H}}_n(\gamma)$ *is non-decreasing with* $\widehat{\mathrm{H}}_n(\widehat{\gamma}_n) = \mathrm{o}_{\mathbb{P}_\theta^n}(1)$,

and that (II) $\mathrm{H}(\theta,\gamma(\theta)-\varepsilon) < 0 < \mathrm{H}(\theta,\gamma(\theta)+\varepsilon)$ *for every* $\varepsilon \in \mathbb{R}^+_{\setminus 0}$*. Then,* $\widehat{\gamma}_n = \gamma(\theta) + \mathrm{o}_{\mathbb{P}_\theta^n}(1)$.

§05.17 **Proof** of Proposition §05.16. is given in the lecture. □

§05.18 **Example**. Consider $\mathbb{P} \in \mathcal{W}(\mathscr{B})$ and $\mathrm{h}(\gamma,t) := \operatorname{sign}(t-\gamma)$ with $\operatorname{sign}(t) := \mathbb{1}_{\{t\geqslant 0\}} - \mathbb{1}_{\{t\leqslant 0\}}$ for all $\gamma, t \in \mathbb{R}$. The sample median $\widehat{\gamma}_n$ is a (near) zero of the map $\gamma \mapsto \widehat{\mathrm{H}}_n(\gamma) := \widehat{\mathbb{P}}_n\mathrm{h}(\gamma)$, i.e. $\widehat{\mathrm{H}}_n(\gamma, x^n) = \frac{1}{n}\sum_{i\in[\![n]\!]} \mathrm{h}(\gamma, x_i)$, $x^n \in \mathbb{R}^n$. Considering $\mathrm{H}(\gamma) = \mathbb{P}\mathrm{h}(\gamma) = \mathbb{P}((\gamma,\infty)) - \mathbb{P}((-\infty,\gamma))$ we have obviously $\widehat{\mathrm{H}}_n(\gamma) = \mathrm{H}(\gamma) + \mathrm{o}_{\mathbb{P}^{\otimes n}}(1)$ for each $\gamma \in \Gamma$. Keeping in mind that $\gamma \mapsto \widehat{\mathrm{H}}_n(\gamma)$ is non-increasing from Proposition §05.16 follows consistency of the sample median $\widehat{\gamma}_n$, i.e., $\widehat{\gamma}_n = \gamma_o + \mathrm{o}_{\mathbb{P}^{\otimes n}}(1)$, if for any $\varepsilon \in \mathbb{R}^+_{\setminus 0}$ in addition $\mathrm{H}(\gamma_o-\varepsilon) > 0 > \mathrm{H}(\gamma_o+\varepsilon)$ or equivalently $\mathbb{P}((-\infty, < \gamma_o-\varepsilon)) < 1/2 < \mathbb{P}((-\infty,\gamma_o+\varepsilon))$. In other words, the sample median $\widehat{\gamma}_n$ is a consistent estimator of the population median, if it is unique. □

§06 Asymptotic normality

Here and subsequently, for $k, n \in \mathbb{N}$ let $\Gamma \subseteq \mathbb{R}^k$ be endowed with its Borel-σ-algebra $\mathscr{G} := \mathscr{B}_\Gamma$, let $(\mathcal{X}^n, \mathscr{X}^{\otimes n}, \mathbb{P}_\Theta^{\otimes n})$ be a statistical product experiment over the parameter space Θ and let $\gamma : \Theta \to \Gamma$ be an identifiable parameter of interest.

§06.01 **Heuristics**. Consider $\widehat{\mathrm{H}}_n(\gamma) = \widehat{\mathbb{P}}_n\mathrm{h}(\gamma)$, i.e. $\widehat{\mathrm{H}}_n(\gamma, x^n) = \frac{1}{n}\sum_{i\in[\![n]\!]} \mathrm{h}(\gamma, x_i)$, $x^n \in \mathcal{X}^n$, and $\mathrm{H}(\theta,\gamma) = \mathbb{P}_\theta\mathrm{h}(\gamma)$ for $\gamma \in \Gamma$ and $\theta \in \Theta$. Let $\widehat{\gamma}_n$ be a zero of $\gamma \mapsto \widehat{\mathrm{H}}_n(\gamma)$, i.e., $\widehat{\gamma}_n$ is a Z-estimator. Assume in addition that $\widehat{\gamma}_n = \gamma(\theta)+\mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$ where $\gamma(\theta)$ is a zero of $\gamma \mapsto \mathrm{H}(\theta,\gamma)$. Heuristically, consider a *Taylor expansion* of a real-valued H around $\gamma(\theta) \in \Gamma \subseteq \mathbb{R}$, that is, $0 = \widehat{\mathrm{H}}_n(\widehat{\gamma}_n) = \widehat{\mathrm{H}}_n(\gamma(\theta))+(\widehat{\gamma}_n-\gamma(\theta))\dot{\widehat{\mathrm{H}}}_n(\gamma(\theta))+\frac{1}{2}(\widehat{\gamma}_n-\gamma(\theta))^2\ddot{\widehat{\mathrm{H}}}_n(\widetilde{\gamma}_n)$ for some $\widetilde{\gamma}_n$ between $\gamma(\theta)$ and $\widehat{\gamma}_n$. Thus, rewriting the last identity $\sqrt{n}(\widehat{\gamma}_n-\gamma(\theta)) = -\sqrt{n}\widehat{\mathrm{H}}_n(\gamma(\theta))\big(\dot{\widehat{\mathrm{H}}}_n(\gamma(\theta))+\frac{1}{2}(\widehat{\gamma}_n-\gamma(\theta))\ddot{\widehat{\mathrm{H}}}_n(\widetilde{\gamma}_n)\big)^{-1}$. If $\mathrm{h}(\gamma(\theta))$ belongs to $\mathscr{L}_2(\mathbb{P}_\theta)$, then due to the CLT it holds $-\sqrt{n}(\widehat{\mathrm{H}}_n(\gamma(\theta)) - \mathrm{H}(\theta,\gamma(\theta))) = -\sqrt{n}(\widehat{\mathbb{P}}_n\mathrm{h}(\gamma(\theta)) - \mathbb{P}_\theta\mathrm{h}(\gamma(\theta))) \xrightarrow{d} \mathrm{N}_{(0,\mathbb{P}_\theta\mathrm{h}^2(\gamma(\theta)))}$. If moreover $\dot{\mathrm{h}}(\gamma(\theta)) \in \mathscr{L}_1(\mathbb{P}_\theta)$, then by the LLN $\dot{\widehat{\mathrm{H}}}_n(\gamma(\theta)) = \widehat{\mathbb{P}}_n\dot{\mathrm{h}}(\gamma(\theta)) = \mathbb{P}_\theta\dot{\mathrm{h}}(\gamma(\theta))+\mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$. If in addition $\ddot{\widehat{\mathrm{H}}}_n(\widetilde{\gamma}_n) = \mathrm{O}_{\mathbb{P}_\theta^{\otimes n}}(1)$ then employing Slutzky's lemma §02.10 it follows $\sqrt{n}(\widehat{\gamma}_n - \gamma(\theta)) \xrightarrow{d} \mathrm{N}_{(0,(\mathbb{P}_\theta\dot{\mathrm{h}}(\gamma(\theta)))^{-2}\mathbb{P}_\theta\mathrm{h}^2(\gamma(\theta)))}$. In the sequel, γ is a vector and h vector-valued. Consequently, $\dot{\mathrm{h}}(\gamma(\theta))$ is a matrix and we denote by $\|\dot{\mathrm{h}}(\gamma(\theta))\|_F$ its *Frobenius norm*, where $\|M\|_F := \big(\sum_{j\in[\![J]\!]}\sum_{k\in[\![K]\!]} M_{jk}^2\big)^{1/2}$ for any matrix $M = (M_{jk}) \in \mathbb{R}^{(J,K)}$. □

§06.02 **Theorem**. *Under the assumptions and notations of Definition §04.11 with* $\Gamma \subseteq \mathbb{R}^k$ *let* $\widehat{\gamma}_n$ *be a consistent Z-estimator of* γ*, i.e.* $\widehat{\gamma}_n = \gamma(\theta) + \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$*, with* $\widehat{\mathrm{H}}_n(\widehat{\gamma}_n) = \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(n^{-1/2})$*. Assume the*

criterion process $\widehat{\mathrm{H}}_n$ *is continuous differentiable in a neighbourhood* U *of* $\gamma(\theta) \in \mathrm{int}(\Gamma)$ *with derivative* $\dot{\widehat{\mathrm{H}}}_n := \frac{\partial}{\partial\gamma}\widehat{\mathrm{H}}_n \in \overline{\mathscr{X}}^{(k,k)}$ *and satisfies the following two conditions:*

(AN1) $\sqrt{n}\widehat{\mathrm{H}}_n(\gamma(\theta)) \xrightarrow{d} \mathrm{N}_{(0,\Omega_\theta)}$ *under* $\mathbb{P}_\theta^{\otimes n}$ *for some positive semidefinite* $\Omega_\theta \in \mathbb{R}^{(k,k)}$,

(AN2) $\sup_{\gamma\in U}\|\dot{\widehat{\mathrm{H}}}_n(\gamma) - \dot{\mathrm{H}}(\theta,\gamma)\|_F = \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$ *for some continuous matrix-valued function* $\gamma \mapsto \dot{\mathrm{H}}(\theta,\gamma)$ *with regular* $\dot{\mathrm{H}}(\theta,\gamma(\theta))$ *having* $\dot{\mathrm{H}}_\theta^{-1}$ *as inverse.*

Then $\sqrt{n}(\widehat{\gamma}_n - \gamma(\theta)) + \sqrt{n}\dot{\mathrm{H}}_\theta^{-1}\widehat{\mathrm{H}}_n(\gamma(\theta)) = \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$ *and* $\sqrt{n}(\widehat{\gamma}_n - \gamma(\theta)) \xrightarrow{d} \mathrm{N}_{(0,\dot{\mathrm{H}}_\theta^{-1}\Omega_\theta(\dot{\mathrm{H}}_\theta^{-1})^t)}$.

§06.03 **Proof** of Theorem §06.02. is given in the lecture. □

§06.04 **Corollary**. *Under the assumptions and notations of Definition §04.11 with* $\Gamma \subseteq \mathbb{R}^k$ *let* $\widehat{\gamma}_n$ *be a consistent* M*-estimator of* γ*, i.e.* $\widehat{\gamma}_n = \gamma(\theta) + \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$*, with* $\widehat{\mathrm{M}}_n(\widehat{\gamma}_n) = \inf_{\gamma\in\Gamma}\widehat{\mathrm{M}}_n(\gamma)$*. Assume the criterion process* $\widehat{\mathrm{M}}_n$ *is twice continuously differentiable in a neighbourhood* U *of* $\gamma(\theta) \in \mathrm{int}(\Gamma)$ *with derivatives* $\dot{\widehat{\mathrm{M}}}_n := \frac{\partial}{\partial\gamma}\widehat{\mathrm{M}}_n \in \overline{\mathscr{X}}^k$ *(score function) and* $\ddot{\widehat{\mathrm{M}}}_n := \frac{\partial^2}{\partial^2\gamma}\widehat{\mathrm{M}}_n \in \overline{\mathscr{X}}^{(k,k)}$ *and satisfies in addition the following two conditions:*

(AN1) $\sqrt{n}\dot{\widehat{\mathrm{M}}}_n(\gamma(\theta)) \xrightarrow{d} \mathrm{N}_{(0,\Omega_\theta)}$ *under* $\mathbb{P}_\theta^{\otimes n}$ *for some positive semidefinite* $\Omega_\theta \geqslant 0$,

(AN2) $\sup_{\gamma\in U}\|\ddot{\widehat{\mathrm{M}}}_n(\gamma) - \ddot{\mathrm{M}}(\theta,\gamma)\|_F = \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$ *for some continuous matrix-valued function* $\gamma \mapsto \ddot{\mathrm{M}}(\theta,\gamma)$ *with regular* $\ddot{\mathrm{M}}(\theta,\gamma(\theta))$ *having* $\ddot{\mathrm{M}}_\theta^{-1}$ *as inverse.*

Then $\sqrt{n}(\widehat{\gamma}_n - \gamma(\theta)) \xrightarrow{d} \mathrm{N}_{(0,\ddot{\mathrm{M}}_\theta^{-1}\Omega_\theta\ddot{\mathrm{M}}_\theta^{-1})}$.

§06.05 **Proof** of Corollary §06.04. is given in the lecture. □

§06.06 **Example** *(§05.11 continued).* Given $(\mathcal{X}^n, \mathscr{X}^{\otimes n}, \mathbb{P}_\Theta^{\otimes n})$ and $\gamma : \Theta \to \Gamma$ for each $\gamma \in \Gamma$ let $\mathrm{m}(\gamma) \in \mathscr{L}_1(\mathbb{P}_\theta)$ be a real function. Consider $\widehat{\mathrm{M}}_n(\gamma) = \widehat{\mathbb{P}}_n\mathrm{m}(\gamma)$ and $\mathrm{M}(\theta,\gamma) = \mathbb{P}_\theta\mathrm{m}(\gamma)$ where due to the LLN $\widehat{\mathrm{M}}_n(\gamma) = \mathrm{M}(\theta,\gamma) + \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$ for each $\gamma \in \Gamma$. Suppose in addition that

(i) Γ is compact,

(ii) $\gamma \mapsto \mathrm{m}(\gamma,x)$ is twice continuously differentiable in a neighbourhood U of $\gamma(\theta) \in \mathrm{int}(\Gamma)$ for $\mathbb{P}_\theta$-a.e. $x \in \mathcal{X}$ with derivatives $\dot{\mathrm{m}} := \frac{\partial}{\partial\gamma}\mathrm{m}$ and $\ddot{\mathrm{m}} := \frac{\partial^2}{\partial^2\gamma}\mathrm{m}$

(iii) $\dot{\mathrm{m}}(\gamma(\theta)) \in \mathscr{L}_2(\mathbb{P}_\theta)$ with $\mathbb{P}_\theta\dot{\mathrm{m}}(\gamma(\theta)) = 0$ and $\Omega_\theta := \mathbb{P}_\theta\dot{\mathrm{m}}(\gamma(\theta))\dot{\mathrm{m}}(\gamma(\theta))^t \geqslant 0$,

(iv) $\sup_{\gamma\in U}\|\ddot{\mathrm{m}}(\gamma)\|_F \in \mathscr{L}_1(\mathbb{P}_\theta)$ and $\ddot{\mathrm{M}}_\theta := \mathbb{P}_\theta\ddot{\mathrm{m}}(\gamma(\theta))$ is regular with inverse $\ddot{\mathrm{M}}_\theta^{-1}$.

hold true. If the M-estimator satisfies $\widehat{\gamma}_n = \gamma(\theta) + \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$ then $\sqrt{n}(\widehat{\gamma}_n - \gamma(\theta)) \xrightarrow{d} \mathrm{N}_{(0,\ddot{\mathrm{M}}_\theta^{-1}\Omega_\theta\ddot{\mathrm{M}}_\theta^{-1})}$ due to Corollary §06.04 since the conditions (AN1)-(AN2) are satisfied. Indeed, following Example §05.11, (iv) implies the condition (AN2) and due to the CLT the condition (AN1) follows from (iii). However, estimators of $\ddot{\mathrm{M}}_\theta$ and Ω_θ are necessary in order to use the asymptotic distribution to conduct inference. A typical approach to obtain these estimators is as follows. First replacing $\mathbb{P}_\theta$ by $\widehat{\mathbb{P}}_n$, the quantity $\ddot{\widehat{\mathrm{M}}}_n(\gamma) := \widehat{\mathbb{P}}_n\ddot{\mathrm{m}}(\gamma)$ and $\widehat{\Omega}_n(\gamma) = \widehat{\mathbb{P}}_n\dot{\mathrm{m}}(\gamma)\dot{\mathrm{m}}(\gamma)^t$ is just an empirical counterpart of $\ddot{\mathrm{M}}_\gamma(\gamma) = \mathbb{P}_\theta\ddot{\mathrm{m}}(\gamma)$ and $\Omega_\theta(\gamma) = \mathbb{P}_\theta\dot{\mathrm{m}}(\gamma)\dot{\mathrm{m}}(\gamma)^t$, respectively. Secondly, replace γ by its estimator $\widehat{\gamma}_n$ we obtain $\ddot{\widehat{\mathrm{M}}}_n := \ddot{\widehat{\mathrm{M}}}_n(\widehat{\gamma}_n)$ and $\widehat{\Omega}_n := \widehat{\Omega}_n(\widehat{\gamma}_n)$ as estimator of $\ddot{\mathrm{M}}_\theta = \ddot{\mathrm{M}}_\theta(\gamma(\theta))$ and $\Omega_\theta = \Omega_\theta(\gamma(\theta))$, respectively. If in addition to (i)-(iv) the following condition holds

(v) $\sup_{\gamma\in U}\|\dot{\mathrm{m}}(\gamma)\|$ belongs to $\mathscr{L}_2(\mathbb{P}_\theta)$.

Then $\sup_{\gamma\in U}\|\widehat{\ddot{\mathrm{M}}}_n(\gamma)-\ddot{\mathrm{M}}_\theta(\gamma)\|_F = \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$ and $\sup_{\gamma\in U}\|\widehat{\Omega}_n(\gamma)-\Omega_\theta(\gamma)\|_F = \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$ following line by line the arguments in Example §05.11. From these uniform convergences and $\widehat{\gamma}_n = \gamma(\theta) + \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$ follows $\widehat{\ddot{\mathrm{M}}}_n = \ddot{\mathrm{M}}_\theta + \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$ and $\widehat{\Omega}_n = \Omega_\theta + \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$ which in turn implies $\widehat{V}_n := \widehat{\ddot{\mathrm{M}}}_n^{-1}\widehat{\Omega}_n\widehat{\ddot{\mathrm{M}}}_n^{-1} = \ddot{\mathrm{M}}_\theta^{-1}\Omega_\theta\ddot{\mathrm{M}}_\theta^{-1} + \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$. Consequently, by applying Slutzky's lemma §02.10 we have $\sqrt{n}\widehat{V}_n^{-1/2}(\widehat{\gamma}_n - \gamma(\theta)) \xrightarrow{d} \mathrm{N}_{(0,\mathrm{Id}_k)}$. □

§06.07 **Example** *(MLE, §04.15 continued).* Let $(\mathcal{X}^n, \mathscr{X}^{\otimes n}, \mathbb{P}_\Theta^{\otimes n})$ with $\mathbb{P}_\theta \ll\gg \mathbb{P}_o$ for all $\theta\in\Theta$, likelihood $\mathrm{L}(\theta) = \mathrm{d}\mathbb{P}_\theta/\mathrm{d}\mathbb{P}_o$, log-likelihood $\ell = \log \mathrm{L}$ and parameter of interest θ (i.e., $\gamma = \mathrm{id}_\Theta$) as in Example §04.15. Consider the MLE $\widehat{\theta}_n$ which maximises the (joint) log-likelihood $\theta\mapsto \widehat{\mathbb{P}}_n\ell(\theta)$. Let the following conditions be satisfied:

(i) (Θ, d) is a compact metric space,

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

(iii) the map $\theta\mapsto\ell(\theta,x)$ is continuous for $\mathbb{P}_\theta$-a.e. $x\in\mathcal{X}$,

(iv) $\sup_{\theta\in\Theta}|\ell(\theta)|$ belongs to $\mathscr{L}_1(\mathbb{P}_\theta)$.

Then combining the arguments in the Examples §05.08 and §05.11 the conditions (CO1) and (CO2) of Theorem §05.02 are satisfied, which in turn implies consistency of the MLE $\widehat{\theta}_n = \theta + \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$. In addition let the following conditions be fulfilled

(v) for $\mathbb{P}_\theta$-a.e. $x\in\mathcal{X}$ the map $\theta\mapsto\ell(\theta,x)$ is twice continuously differentiable in a neighbourhood U of $\theta\in\Theta$ with derivatives $\dot{\ell} := \frac{\partial}{\partial\theta}\ell$ and $\ddot{\ell} := \frac{\partial^2}{\partial^2\theta}\ell$,

(vi) $\sup_{\theta\in U}\|\dot{\ell}(\theta)\|\in\mathscr{L}_2(\mathbb{P}_\theta)$ and $\sup_{\theta\in U}\|\ddot{\ell}(\theta)\|_F\in\mathscr{L}_1(\mathbb{P}_\theta)$,

(vii) the Fisher-information matrix $\mathcal{I}_\theta := \mathbb{P}_\theta\dot{\ell}(\theta)\dot{\ell}(\theta)^t$ is strictly positive definite.

Then the conditions (AN1) and (AN2) of Corollary §06.04, and the identity $\mathcal{I}_\theta = -\mathbb{P}_\theta\ddot{\ell}(\theta)$ are satisfied (for details see Satz §20.20 in the lecture notes Statistik 1). Therewith, the MLE satisfies $\sqrt{n}(\widehat{\theta}_n-\theta) = \sqrt{n}\mathcal{I}_\theta^{-1}\widehat{\mathbb{P}}_n\dot{\ell}(\theta) + \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$ and, consequently, $\sqrt{n}(\widehat{\theta}_n-\theta_o)\xrightarrow{d}\mathrm{N}_{(0,\mathcal{I}_\theta^{-1})}$. □

§06.08 **Remark**. The conditions (v) and (vi) in Example §06.07 can be weakened replacing differentiability by Hellinger-differentiability. Keeping the *Hellinger-distance* $H(\mathbb{P}_\theta,\mathbb{P}_{\theta_o}) = \|\mathrm{L}^{1/2}(\theta) - \mathrm{L}^{1/2}(\theta_o)\|_{\mathscr{L}_2(\mathbb{P})}$ in mind, where $\mathrm{L}^{1/2}(\theta)\in\mathscr{L}_2(\mathbb{P}_o)$ using $\|\mathrm{L}^{1/2}(\theta)\|^2_{\mathscr{L}_2(\mathbb{P})} = \mathbb{P}_o(\mathrm{L}(\theta)) = 1 < \infty$, the family $\mathbb{P}_\Theta$ is called *Hellinger-differentiable with derivative* $\dot{\ell}_{\theta_o}$ in $\theta_o\in\mathrm{int}(\Theta)\subseteq\mathbb{R}^k$, if $\dot{\ell}_{\theta_o}\in\mathscr{L}_2(\mathbb{P}_{\theta_o})$ and hence $\dot{\ell}_{\theta_o}\mathrm{L}^{1/2}(\theta_o)\in\mathscr{L}_2(\mathbb{P}_o)$ such that

$$\lim_{\theta\to\theta_o}\int_{\mathcal{X}}\left|\frac{\mathrm{L}^{1/2}(\theta,x) - \mathrm{L}^{1/2}(\theta_o,x) - \frac{1}{2}\langle\dot{\ell}_{\theta_o}(x),\theta-\theta_o\rangle\mathrm{L}^{1/2}(\theta_o,x)}{\|\theta-\theta_o\|}\right|^2\mathbb{P}_o(dx)$$
$$= \lim_{h\to0}\frac{\|\mathrm{L}^{1/2}(\theta_o+h) - \mathrm{L}^{1/2}(\theta_o) - \frac{1}{2}\langle\dot{\ell}_{\theta_o},h\rangle\mathrm{L}^{1/2}(\theta_o)\|^2_{\mathscr{L}_2(\mathbb{P})}}{\|h\|^2} = 0$$

The map $x\mapsto\dot{\ell}_{\theta_o}(x)$ is also called *score function*. Keeping $\dot{\ell}_{\theta_o}\in\mathscr{L}_2(\mathbb{P}_{\theta_o})$ in mind the *Fisher-information* matrix $\mathcal{I}_{\theta_o} = \mathbb{P}_{\theta_o}\dot{\ell}_{\theta_o}\dot{\ell}_{\theta_o}^t$ is well-defined. Note that, the score function and the Fisher-information matrix are independent of the dominating measure $\mathbb{P}_o$. □

Testing procedures

§06.09 **Heuristics.** Let $(\mathcal{X}_n, \mathscr{X}_n, \mathbb{P}_\Theta^n)$ for all $n \in \mathbb{N}$ be a statistical model over the parameter space Θ and let $\gamma : \Theta \to \Gamma$ be an identifiable parameter of interest. Given a map $A : \Gamma \to \mathbb{R}^p$ we eventually test the hypothesis $H_0 : A(\gamma) = 0$ against the alternative $H_1 : A(\gamma) \neq 0$. Typical examples include $A(\gamma) = \gamma - \gamma_o$ for a given value γ_o, or more generally, linear hypothesis $A(\gamma) = M\gamma - a_o$ for a given value a_o and matrix M. It covers in particular testing the j-th coordinate of $\gamma = (\gamma^j)_{j\in[\![k]\!]}$, i.e., $A(\gamma) = \gamma^j - \gamma_o^j$. Under regularity conditions it seems reasonable to assume an estimator $\widehat{\gamma}_n$ of γ having under $\mathbb{P}_\theta^n$ the property $\sqrt{n}(A(\widehat{\gamma}_n) - A(\gamma(\theta))) \xrightarrow{d} \mathrm{N}_{(0,\Sigma_\theta)}$ with invertible asymptotic covariance matrix Σ_θ. If we have in addition an estimator $\widehat{\Sigma}_n = \Sigma_\theta + \mathrm{o}_{\mathbb{P}_\theta^n}(1)$ at hand. Then under the hypothesis H_0, i.e., for $\mathbb{P}_\theta^n$ with $A(\gamma(\theta)) = 0$, a *Wald test* exploits the property $\widehat{W}_n := nA(\widehat{\gamma}_n)^t\widehat{\Sigma}_n^{-1}A(\widehat{\gamma}_n) \xrightarrow{d} \chi_p^2$ where χ_p^2 is a Chi-square-distribution with p degrees of freedom. Precisely, a *Wald test* rejects the hypothesis $H_0 : A(\gamma) = 0$ if $\widehat{W}_n$ exceeds the 1-α-Quantile $\chi_{p,1-\alpha}^2$ of a χ_p^2-distribution. Obviously, the Wald test does exactly meets the asymptotic level α, i.e., $\lim_{n\to\infty} \mathbb{P}_\theta^n(\widehat{W}_n \geqslant \chi_{p,1-\alpha}^2) = \mathbb{P}(W \geqslant \chi_{p,1-\alpha}^2) = \alpha$ where $W \sim \chi_p^2$. 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. □

§06.10 **Example** *(§06.06 continued).* Let $(\mathcal{X}^n, \mathscr{X}^{\otimes n}, \mathbb{P}_\Theta^{\otimes n})$, $\gamma : \Theta \to \Gamma$ be an identifiable parameter of interest and let $\mathrm{m}(\gamma) \in \mathscr{L}_1(\mathbb{P}_\theta)$ for all $\gamma \in \Gamma$. For each $\gamma \in \Gamma$ let $\widehat{\mathrm{M}}_n(\gamma) = \widehat{\mathbb{P}}_n\mathrm{m}(\gamma)$ and $\mathrm{M}(\theta,\gamma) = \mathbb{P}_\theta\mathrm{m}(\gamma)$. Under the conditions (i)-(v) in Example §06.06 an M-estimator $\widehat{\gamma}_n \in \arg\inf_{\gamma\in\Gamma} \widehat{\mathrm{M}}_n(\gamma)$ satisfies $\sqrt{n}(\widehat{\gamma}_n - \gamma(\theta)) \xrightarrow{d} \mathrm{N}_{(0,\ddot{\mathrm{M}}_\theta^{-1}\Omega_\theta\ddot{\mathrm{M}}_\theta^{-1})}$ under $\mathbb{P}_\theta^{\otimes n}$. Moreover, we have eventually access to estimators $\widehat{\ddot{\mathrm{M}}}_n = \ddot{\mathrm{M}}_\theta + \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$ and $\widehat{\Omega}_n = \Omega_\theta + \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$. Let $A : \Gamma \to \mathbb{R}^p$ be continuously differentiable in a neighbourhood of $\gamma(\theta)$ then applying the delta method §02.16 we obtain $\sqrt{n}(A(\widehat{\gamma}_n) - A(\gamma(\theta))) \xrightarrow{d} \mathrm{N}_{(0,\Sigma_\theta)}$ under $\mathbb{P}_\theta^{\otimes n}$ with $\Sigma_\theta := \dot{A}_{\gamma(\theta)}\ddot{\mathrm{M}}_\theta^{-1}\Omega_\theta\ddot{\mathrm{M}}_\theta^{-1}\dot{A}_{\gamma(\theta)}^t$. From $\dot{A}_{\widehat{\gamma}_n} = \dot{A}_{\gamma(\theta)} + \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$ follows $\widehat{\Sigma}_n := \dot{A}_{\widehat{\gamma}_n}\widehat{\ddot{\mathrm{M}}}_n^{-1}\widehat{\Omega}_n\widehat{\ddot{\mathrm{M}}}_n^{-1}\dot{A}_{\widehat{\gamma}_n}^t = \Sigma_\theta + \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$ and, thus $\sqrt{n}\widehat{\Sigma}_n^{-1/2}(A(\widehat{\gamma}_n) - A(\gamma(\theta))) \xrightarrow{d} \mathrm{N}_{(0,\mathrm{Id}_p)}$ which under H_0, i.e., for $\mathbb{P}_\theta^{(n)}$ with $A(\gamma(\theta)) = 0$, implies $\widehat{W}_n := nA(\widehat{\gamma}_n)^t\widehat{\Sigma}_n^{-1}A(\widehat{\gamma}_n) \xrightarrow{d} \chi_p^2$. □

Chapter 3

Asymptotic properties of tests

Asymptotic properties of tests under local alternatives are presented complementing the Neyman-Pearson theory introduced in the lecture Statistik 1. For a more detailed exposition we refer to the text books Witting and Müller-Funk *[1995] and* van der Vaart *[1998].*

§07 Contiguity

§07.01 Preliminaries: likelihood ratios and differentiable models

§07.01 **Motivation**. Considering a statistical model $(\mathcal{X}_n, \mathscr{X}_n, \mathbb{P}_\Theta^n)$, a parameter of interest $\gamma : \Theta \to \Gamma$, a partition $\{\mathscr{H}^0, \mathscr{H}^1\}$ of the parameter values of interests $\Gamma = \mathscr{H}^0 \biguplus \mathscr{H}^1$ (i.e. $\Gamma = \mathscr{H}^0 \cup \mathscr{H}^1$, $\emptyset = \mathscr{H}^0 \cap \mathscr{H}^1$ and $\mathscr{H}^0 \neq \emptyset \neq \mathscr{H}^1$) we are interested in a (randomised) test $\varphi_n \in \mathscr{X}_n^+$ (i.e. $\varphi_n : \mathcal{X}_n \to [0,1]$) of the hypothesis $H_0 : \mathscr{H}^0$ against the alternative $H_1 : \mathscr{H}^1$. Under regularity conditions we may have at hand an estimator $\widehat{\gamma}_n$ of γ with known asymptotic distribution. Typically the estimator $\widehat{\gamma}_n$ allows us to construct a test statistic T_n with known asymptotic distribution under H_0, i.e. under $\mathbb{P}_\theta^n$ with $\gamma(\theta) \in \mathscr{H}^0$. Exploiting the asymptotic distribution an associated test $\varphi_n = \mathbb{1}_{\{T_n \notin C_\alpha\}}$ does eventually not exceed asymptotically a given level $\alpha \in (0,1)$ under the hypothesis H_0. However, we like to investigate also its power under the alternative H_1, i.e. under a specific $\mathbb{P}_\theta^n$ with $\gamma(\theta) \in \mathscr{H}^1$. □

§07.02 **Reminder**. Let ν and μ be measures on $(\mathcal{X}, \mathscr{X})$.

(a) For any positive numerical function $\mathbb{f} \in \overline{\mathscr{X}}^+$ the map $B \mapsto \mathbb{f}\mu(B) := \mu(\mathbb{1}_B \mathbb{f}) = \int_B \mathbb{f}\, d\mu$ defines a measure $\mathbb{f}\mu$ on $(\mathcal{X}, \mathscr{X})$. Any $\mathbb{f} \in \overline{\mathscr{X}}^+$ satisfying $\nu = \mathbb{f}\mu$ is called *density* of ν with respect to μ, or *μ-density* for short.

(b) We say ν is *dominated* by μ, symbolically $\nu \ll \mu$, if for each $B \in \mathscr{X}$ with $\mu(B) = 0$ follows $\nu(B) = 0$. The measures μ and ν are called *equivalent* or *mutually dominated*, symbolically $\mu \ll\gg \nu$, if both $\nu \ll \mu$ and $\mu \ll \nu$.

(c) We say ν and μ are *orthogonal* or *singular*, symbolically $\nu \perp \mu$, if there exists $\mathcal{X} = \mathcal{X}_\mu \biguplus \mathcal{X}_\nu$ with $\mathcal{X}_\mu, \mathcal{X}_\nu \in \mathscr{X}$ and $\mu(\mathcal{X}_\nu) = 0 = \nu(\mathcal{X}_\mu)$.

We note that $g \in \mathscr{L}_1(\mathbb{f}\mu)$ if and only if $g\mathbb{f} \in \mathscr{L}_1(\mu)$. In this case holds $\mathbb{f}\mu(g) = \int g\, d(\mathbb{f}\mu) = \int (g\mathbb{f})\, d\mu = \mu(g\mathbb{f})$ (Klenke [2012], Satz 4.15, p. 93). Let additionally $\nu \in \mathcal{M}_\sigma(\mathscr{X})$ be a σ-finite measure on $(\mathcal{X}, \mathscr{X})$. If $\mathbb{f}_1\mu = \nu = \mathbb{f}_2\mu$ for $\mathbb{f}_1, \mathbb{f}_2 \in \overline{\mathscr{X}}^+$, then $\mathbb{f}_1 = \mathbb{f}_2$ μ-a.e.. In other words a density is unique up to μ-a.e. equivalence (Klenke [2012], Satz 7.29, p. 159). If in addition $\mu \in \mathcal{M}_\sigma(\mathscr{X})$, then by *Lebesgue's decomposition theorem* there exists $\nu^a, \nu^\perp \in \mathcal{M}_\sigma(\mathscr{X})$ such that $\nu = \nu^a + \nu^\perp$ with $\nu^\perp \perp \mu$ and $\nu^a = \mathbb{f}\mu$ where $\mathbb{f} \in \overline{\mathscr{X}}^+$ and $\mathbb{f} < \infty$ μ-a.e.. (Klenke [2012], Satz 7.33, p. 160) Furthermore, there is a *Radon-Nikodym-density* $\mathbb{f} \in \overline{\mathscr{X}}^+$ with $\nu = \mathbb{f}\mu$ and $\mathbb{f} < \infty$ μ-a.e. if and only if $\nu \ll \mu$ (Klenke [2012], Korollar 7.34, p. 161). If $\mathbb{f} \in \overline{\mathscr{X}}^+$ is a Radon-Nikodym-density of ν with respect to μ, i.e. $\nu = \mathbb{f}\mu$, then the positive real

function $\mathfrak{f}\mathbb{1}_{\{\mathfrak{f}\in\mathbb{R}^+\}} \in \mathscr{X}^+$ is it too. Consequently, without loss of generality we consider here and subsequently a positive real version of the Radon-Nikodym-density. □

§07.03 **Definition.** Let $\mathbb{P}_0, \mathbb{P}_1 \in \mathcal{W}(\mathscr{X})$ be probability measures on $(\mathcal{X}, \mathscr{X})$. Any positive numerical random variable $\mathrm{L} \in \overline{\mathscr{X}}^+$ satisfying

$$\mathbb{P}_0(\mathrm{L} < \infty) = 1 \text{ and } \mathbb{P}_1(B) = \mathrm{L}\mathbb{P}_0(B) + \mathbb{P}_1(B \cap \{\mathrm{L} = \infty\}) \quad \text{for all } B \in \mathscr{X} \tag{07.1}$$

is called a *likelihood ratio (LR)* of $\mathbb{P}_1$ with respect to $\mathbb{P}_0$, symbolically $\mathrm{d}\mathbb{P}_1/\mathrm{d}\mathbb{P}_0 := \mathrm{L}$. □

Here and subsequently, let $\mathbb{P}_0, \mathbb{P}_1 \in \mathcal{W}(\mathscr{X})$ and $\mathrm{L} := \mathrm{d}\mathbb{P}_1/\mathrm{d}\mathbb{P}_0$ be a likelihood ratio of $\mathbb{P}_1$ with respect to $\mathbb{P}_0$. We first note that $\mathbb{P}_0(\mathrm{L}) = \mathbb{P}_1(\mathrm{L} < \infty) \in [0,1]$ and $\mathbb{P}_0(\mathrm{L} = \infty) = 0$ by definition, and also $\mathbb{P}_1(\mathrm{L} = 0) = \mathrm{L}\mathbb{P}_0(\mathrm{L} = 0) + \mathbb{P}_1(\{\mathrm{L} = 0\} \cap \{\mathrm{L} = \infty\}) = 0$.

§07.04 **Property.**

(i) $\mathbb{P}_0 \perp \mathbb{P}_1 \Leftrightarrow \exists B \in \mathscr{X} : \mathbb{P}_0(B) = 0$ *(hence* $\mathrm{L}\mathbb{P}_0(B) = 0$*) and* $\mathbb{P}_1(B) = 1$ *(hence* $\mathbb{P}_1(B \cap \{\mathrm{L} = \infty\}) = 1$*)* $\Leftrightarrow \mathbb{P}_1(\mathrm{L} = \infty) = 1 \Leftrightarrow \mathbb{P}_0(\mathrm{L}) = 0$*;*

(ii) $\mathbb{P}_0 \not\perp \mathbb{P}_1 \Leftrightarrow \forall B \in \mathscr{X} : \mathbb{P}_0(B) = 0$ *implies* $\mathbb{P}_1(B) < 1$ *(particularly for* $B = \{\mathrm{L} = \infty\}$*)* $\Leftrightarrow \mathbb{P}_1(\mathrm{L} = \infty) < 1 \Leftrightarrow \mathbb{P}_0(\mathrm{L}) > 0$*;*

(iii) $\mathbb{P}_1 \ll \mathbb{P}_0 \Leftrightarrow \forall B \in \mathscr{X} : \mathbb{P}_0(B) = 0$ *implies* $\mathbb{P}_1(B) = 0$ *(particularly for* $B = \{\mathrm{L} = \infty\}$*)* $\Leftrightarrow \mathbb{P}_1(\mathrm{L} = \infty) = 0 \Leftrightarrow \mathbb{P}_0(\mathrm{L}) = 1$.

§07.05 **Remark.** Note that both $\mathbb{P}_0$ and $\mathbb{P}_1$ are dominated by $\mathbb{P}_\mu := \frac{1}{2}(\mathbb{P}_0 + \mathbb{P}_1) \in \mathcal{W}(\mathscr{X})$. Let $\mathfrak{f}_i \in \mathscr{X}^+$ denote a $\mathbb{P}_\mu$-density of $\mathbb{P}_i$, $i \in \{0,1\}$ (c.f. Reminder §07.02), then

$$\mathrm{L}_* = \frac{\mathfrak{f}_1}{\mathfrak{f}_0}\mathbb{1}_{\{\mathfrak{f}_0\in\mathbb{R}^+_{\setminus 0}\}} + \infty\mathbb{1}_{\{\mathfrak{f}_0=0\}\cap\{\mathfrak{f}_1\in\mathbb{R}^+_{\setminus 0}\}} \tag{07.2}$$

is a likelihood ratio of $\mathbb{P}_1$ with respect to $\mathbb{P}_0$, i.e., $\mathrm{L}_* = \mathrm{d}\mathbb{P}_1/\mathrm{d}\mathbb{P}_0$. Indeed, $\mathrm{L}_* \in \overline{\mathscr{X}}^+$ satisfies $\mathbb{P}_0(\mathrm{L}_* = \infty) \leqslant \mathbb{P}_0(\mathfrak{f}_0 = 0) = 0$ and for all $B \in \mathscr{X}$

$$\begin{aligned}\mathrm{L}_*\mathbb{P}_0(B) + \mathbb{P}_1(B \cap \{\mathrm{L}_* = \infty\}) &= \mathfrak{f}_0\mathbb{P}_\mu\left(\tfrac{\mathfrak{f}_1}{\mathfrak{f}_0}\mathbb{1}_{B\cap\{\mathfrak{f}_0\in\mathbb{R}^+_{\setminus 0}\}}\right) + \mathbb{P}_1(B \cap \{\mathfrak{f}_0 = 0\} \cap \{\mathfrak{f}_1 \in \mathbb{R}^+_{\setminus 0}\}) \\ &= \mathfrak{f}_1\mathbb{P}_\mu\left(B \cap \{\mathfrak{f}_0 \in \mathbb{R}^+_{\setminus 0}\}\right) + \mathbb{P}_1(B \cap \{\mathfrak{f}_0 = 0\}) = \mathbb{P}_1(B).\end{aligned}$$

Consequently, L_* is always a version of the likelihood ratio $\mathrm{d}\mathbb{P}_1/\mathrm{d}\mathbb{P}_0$. In general the likelihood ratio $\mathrm{d}\mathbb{P}_1/\mathrm{d}\mathbb{P}_0$ (and similar $\mathrm{d}\mathbb{P}_0/\mathrm{d}\mathbb{P}_1$) is uniquely determined by (07.1) up to $(\mathbb{P}_0 + \mathbb{P}_1)$-a.e. equivalence (Witting [1985] Satz 1.110 a), p. 112). Moreover, the positive numerical random variable $\mathrm{L}_*^{-1} = \frac{\mathfrak{f}_0}{\mathfrak{f}_1}\mathbb{1}_{\{\mathfrak{f}_1\in\mathbb{R}^+_{\setminus 0}\}} + \infty\mathbb{1}_{\{\mathfrak{f}_1=0\}\cap\{\mathfrak{f}_0\in\mathbb{R}^+_{\setminus 0}\}}$ is a version of the likelihood ratio $\mathrm{d}\mathbb{P}_0/\mathrm{d}\mathbb{P}_1$ switching the roles of $\mathbb{P}_0$ and $\mathbb{P}_1$. Consequently, (iii) can equivalently be written as $\mathbb{P}_1 \ll \mathbb{P}_0 \Leftrightarrow \mathbb{P}_1(\mathrm{d}\mathbb{P}_0/\mathrm{d}\mathbb{P}_1 = 0) = \mathbb{P}_1(\mathrm{L}_*^{-1} = 0) = \mathbb{P}_1(\mathrm{L}_* = \infty) = 0$. However, given any version $\mathrm{L} = \mathrm{d}\mathbb{P}_1/\mathrm{d}\mathbb{P}_0$ of the likelihood ratio the measure $\mathbb{P}_1$ can be written as a sum $\mathbb{P}_1 = \mathbb{P}_1^a + \mathbb{P}_1^\perp$ of two measures $\mathbb{P}_1^a, \mathbb{P}_1^\perp \in \mathcal{M}_\sigma(\mathscr{X})$ where $\mathbb{P}_1^a := \mathrm{L}\mathbb{P}_0$ and $\mathbb{P}_1^\perp := \mathbb{1}_{\{\mathrm{L}=\infty\}}\mathbb{P}_1$ with $\mathbb{P}_1^\perp(B) = \mathbb{P}_1(B \cap \{\mathrm{L} = \infty\})$, $B \in \mathscr{X}$ is, respectively, the absolute continuous and singular part of $\mathbb{P}_1$ with respect to $\mathbb{P}_0$ (Lebesgue decomposition). □

§07.06 **Property.** *The two measures* $\mathbb{P}_1^a := \mathrm{L}\mathbb{P}_0$ *and* $\mathbb{P}_1^\perp := \mathbb{1}_{\{\mathrm{L}=\infty\}}\mathbb{P}_1$ *in* $\mathcal{M}_\sigma(\mathscr{X})$ *satisfy*

(i) $\mathbb{P}_1 = \mathbb{P}_1^a + \mathbb{P}_1^\perp$, $\mathbb{P}_1^a \ll \mathbb{P}_0$*, and* $\mathbb{P}_1^\perp \perp \mathbb{P}_0$*;*

(ii) $\mathbb{P}_1(f) \geqslant \mathbb{P}_1^a(f) = \mathrm{L}\mathbb{P}_0(f) = \mathbb{P}_0(\mathrm{L}f) = \mathbb{P}_1(f\mathbb{1}_{\{\mathrm{L}<\infty\}})$ *for all* $f \in \overline{\mathscr{X}}^+$*;*

(iii) $\mathbb{P}_0 \ll \mathbb{P}_1$ *if and only if* $\mathbb{P}_0(\mathrm{L}) = 1$ *if and only if* $\mathbb{P}_1(\mathrm{d}\mathbb{P}_0/\mathrm{d}\mathbb{P}_1 = 0) = \mathbb{P}_1(\mathrm{L} = \infty) = 0$ *if and only if for all* $f \in \overline{\mathscr{X}}^+$ *holds* $\mathbb{P}_1(f) = \mathbb{P}_0(\mathrm{L}f)$. □

§07.07 **Reminder**. Consider a $\mathbb{R}^k$-valued statistic S defined on $(\mathcal{X}, \mathscr{X})$, i.e. $S \in \mathscr{X}^k$. If $\mathbb{P}_1 \ll \mathbb{P}_0$, then the probability measure $\mathbb{P}_1^S = \mathbb{P}_1 \circ S^{-1} \in \mathcal{W}(\mathscr{B}^k)$ induced by S under $\mathbb{P}_1$ can be calculated from the probability measure $\mathbb{P}_0^{(S,\mathrm{L})} = \mathbb{P}_0 \circ (S, \mathrm{L})^{-1}$ induced by the random vector (S, L) under $\mathbb{P}_0$ through the formula

$$\mathbb{P}_1(S \in B) = \mathbb{P}_1^S(\mathbb{1}_B) = \mathbb{P}_0\big(\mathbb{1}_B(S)\mathrm{L}\big) = \mathbb{P}_0^{(S,\mathrm{L})}\big(\mathbb{1}_B(\Pi_S)\Pi_\mathrm{L}\big) \qquad \text{for all } B \in \mathscr{B}^k$$

using the coordinate maps $\Pi_\mathrm{L}(S, \mathrm{L}) = \mathrm{L}$ and $\Pi_S(S, \mathrm{L}) = S$. The formula, however, is only valid under the assumption $\mathbb{P}_1 \ll \mathbb{P}_0$, since a part of $\mathbb{P}_1$ orthogonal to $\mathbb{P}_0$ can't be recovered. □

Here and subsequently, let $\mathbb{P}_\Theta = (\mathbb{P}_\theta)_{\theta\in\Theta}$ with $\Theta \subseteq \mathbb{R}^k$ be a family of probability measures on a measurable space $(\mathcal{X}, \mathscr{X})$, and for each $\theta_o, \theta \in \Theta$ let $\mathrm{L}_{\theta_o}(\theta) := \mathrm{d}\mathbb{P}_\theta/\mathrm{d}\mathbb{P}_{\theta_o}$ denote a likelihood ratio of $\mathbb{P}_\theta$ with respect to $\mathbb{P}_{\theta_o}$. Keep in mind, that $\mathrm{L}_{\theta_o}(\theta_o) = 1 (= \mathbb{1}_\mathcal{X})$.

§07.08 **Definition**. Let $s \geqslant 1$ and $\theta_o \in \mathrm{int}(\Theta)$. The statistical model $(\mathcal{X}, \mathscr{X}, \mathbb{P}_\Theta)$ (and the family $\mathbb{P}_\Theta$) is called *$\mathscr{L}_s(\theta_o)$-differentiable with derivative $\dot\ell_{\theta_o}$*, if $\dot\ell_{\theta_o} \in \mathscr{L}_s(\mathbb{P}_{\theta_o})$ and for all $\theta \to \theta_o$ hold

$$\|s(\mathrm{L}_{\theta_o}^{1/s}(\theta) - 1) - \langle\dot\ell_{\theta_o}, (\theta - \theta_o)\rangle\|_{\mathscr{L}_s(\mathbb{P}_{\theta_o})} = \mathrm{o}(\|\theta - \theta_o\|) \tag{07.3}$$

and $\mathbb{P}_\theta(\mathrm{L}_{\theta_o}(\theta) = \infty) = \mathrm{o}(\|\theta - \theta_o\|^s)$. □

§07.09 **Remark**. In case $s = 1$ the defining condition $\mathbb{P}_\theta(\mathrm{L}_{\theta_o}(\theta) = \infty) = \mathrm{o}(\|\theta - \theta_o\|)$ follows from (07.3) (Witting [1985], Hilfssatz 1.178, p164). We note that $\mathscr{L}_1(\theta_o)$-differentiability implies $\mathscr{L}_1(\mathbb{P}_{\theta_o})$-continuity of $\theta \mapsto \mathrm{L}_{\theta_o}(\theta)$ in θ_o, i.e., $\|\mathrm{L}_{\theta_o}(\theta) - \mathrm{L}_{\theta_o}(\theta_o)\|_{\mathscr{L}_1(\mathbb{P}_{\theta_o})} = \mathrm{o}(1)$ as $\theta \to \theta_o$. Since $\mathrm{L}_{\theta_o}(\theta)$ is unique up to $\mathbb{P}_\theta + \mathbb{P}_{\theta_o}$-a.e.-equivalence $\mathscr{L}_s(\theta_o)$-differentiability does not depend on the version $\mathrm{L}_{\theta_o}(\theta)$ of the likelihood ratio $\mathrm{d}\mathbb{P}_\theta/\mathrm{d}\mathbb{P}_{\theta_o}$. □

§07.10 **Lemma**. *If $\mathbb{P}_\Theta$ is $\mathscr{L}_1(\theta_o)$-differentiable with derivative $\dot\ell_{\theta_o}$, then it holds $\mathbb{P}_{\theta_o}(\dot\ell_{\theta_o}) = 0$. For any $s \geqslant r \geqslant 1$ if $\mathbb{P}_\Theta$ is $\mathscr{L}_s(\theta_o)$-differentiable with derivative $\dot\ell_{\theta_o}$, then $\mathbb{P}_\Theta$ is also $\mathscr{L}_r(\theta_o)$-differentiable with derivative $\dot\ell_{\theta_o}$.*

§07.11 **Proof** of Lemma §07.10. see Witting [1985] (Hilfssatz 1.178, p164 and Satz 1.190, p.164). □

In order to avoid additional integrability conditions in Definition §07.08 the function $\theta \mapsto s(\mathrm{L}_\mu^{1/s}(\theta) - 1)$ is considered. The next assertion formulates differentiability under additional integrability conditions.

§07.12 **Lemma**. *Let $s \geqslant 1$ and $\theta_o \in \mathrm{int}(\Theta)$. The family $\mathbb{P}_\Theta$ is $\mathscr{L}_s(\theta_o)$-differentiable with derivative $\dot\ell_{\theta_o}$, if $\dot\ell_{\theta_o} \in \mathscr{L}_s(\mathbb{P}_{\theta_o})$, $\mathrm{L}_{\theta_o}(\theta) \in \mathscr{L}_s(\mathbb{P}_{\theta_o})$ for all $\theta \in U(\theta_o)$ and for all $\theta \to \theta_o$ hold*

$$\|(\mathrm{L}_{\theta_o}(\theta) - 1) - \langle\dot\ell_{\theta_o}, \theta - \theta_o\rangle\|_{\mathscr{L}_s(\mathbb{P}_{\theta_o})} = \mathrm{o}(\|\theta - \theta_o\|)$$

and $\mathbb{P}_\theta(\mathrm{L}_{\theta_o}(\theta) = \infty) = \mathrm{o}(\|\theta - \theta_o\|^s)$.

§07.13 **Proof** of Lemma §07.12. see Witting [1985] (Satz 1.199, p.183). □

Let us assume in addition, that the family $\mathbb{P}_\Theta$ is dominated by $\mu \in \mathcal{M}_\sigma(\mathscr{X})$. For each $\theta \in \Theta$ denote by $\mathrm{L}_\mu(\theta) := \mathrm{d}\mathbb{P}_\theta/\mathrm{d}\mu \in \mathscr{X}^+$ a Radon-Nikodym density of $\mathbb{P}_\theta$ with respect to μ. Keeping Remark §07.05 in mind $\mathrm{L}_{*,\theta_o}(\theta) = \frac{\mathrm{L}_\mu(\theta)}{\mathrm{L}_\mu(\theta_o)}\mathbb{1}_{\{\mathrm{L}_\mu(\theta_o)\in\mathbb{R}^+_{\setminus 0}\}} + \infty\mathbb{1}_{\{\mathrm{L}_\mu(\theta_o)=0\}\cap\{\mathrm{L}_\mu(\theta)\in\mathbb{R}^+_{\setminus 0}\}}$ as in (07.2) is for each $\theta_o, \theta \in \Theta$ a version of the likelihood ratio $\mathrm{d}\mathbb{P}_\theta/\mathrm{d}\mathbb{P}_{\theta_o}$. We note that

$$\{\mathrm{L}_{*,\theta_o}(\theta) = \infty\} = \big\{\{\mathrm{L}_\mu(\theta_o) = 0\} \cap \{\mathrm{L}_\mu(\theta) \in \mathbb{R}^+_{\setminus 0}\}\big\} \subseteq \{\mathrm{L}_\mu(\theta_o) = 0\} =: \mathcal{N}_{\theta_o},$$

where $\mathbb{P}_{\theta_o}(\mathcal{N}_{\theta_o}) = 0$, and for all $\theta \in \Theta$ holds $\frac{\mathrm{L}_\mu(\theta)}{\mathrm{L}_\mu(\theta_o)}\mathbb{1}_{\mathcal{N}^c_{\theta_o}} = \mathrm{L}_{*,\theta_o}(\theta)\mathbb{1}_{\mathcal{N}^c_{\theta_o}} < \infty$ and $\mathbb{P}_\theta(\mathcal{N}_{\theta_o}) = \mathbb{P}_\theta(\mathcal{N}_{\theta_o} \cap \{\mathrm{L}_\mu(\theta) \in \mathbb{R}^+_{\setminus 0}\}) = \mathbb{P}_\theta(\mathrm{L}_{*,\theta_o}(\theta) = \infty) = \mathbb{P}_\theta(\mathrm{d}\mathbb{P}_\theta/\mathrm{d}\mathbb{P}_{\theta_o} = \infty)$. Decomposing the integral with respect to $\mathscr{X} = \mathcal{N}_{\theta_o} \biguplus \mathcal{N}^c_{\theta_o}$ it follows

$$\begin{aligned}
&\|2(\mathrm{L}^{1/2}_\mu(\theta) - \mathrm{L}^{1/2}_\mu(\theta_o)) - \langle\dot\ell_{\theta_o}, (\theta-\theta_o)\rangle\mathrm{L}^{1/2}_\mu(\theta_o)\|^2_{\mathscr{L}_2(\mu)}\\
&\quad= \|2(\mathrm{L}^{1/2}_{*,\theta_o}(\theta) - 1) - \langle\dot\ell_{\theta_o}, (\theta-\theta_o)\rangle\|^2_{\mathscr{L}_2(\mathbb{P}_{\theta_o})} + \|\mathbb{1}_{\mathcal{N}_{\theta_o}}2\mathrm{L}^{1/2}_\mu(\theta)\|^2_{\mathscr{L}_2(\mu)}\\
&\quad= \|2(\mathrm{L}^{1/2}_{*,\theta_o}(\theta) - 1) - \langle\dot\ell_{\theta_o}, (\theta-\theta_o)\rangle\|^2_{\mathscr{L}_2(\mathbb{P}_{\theta_o})} + 4\mathbb{P}_\theta(\mathrm{L}_{*,\theta_o}(\theta) = \infty)\\
&\quad= \|2(\mathrm{L}^{1/2}_{\theta_o}(\theta) - 1) - \langle\dot\ell_{\theta_o}, (\theta-\theta_o)\rangle\|^2_{\mathscr{L}_2(\mathbb{P}_{\theta_o})} + 4\mathbb{P}_\theta(\mathrm{L}_{\theta_o}(\theta) = \infty). \qquad (07.4)
\end{aligned}$$

Keeping Remark §06.08 in mind for $\theta_o \in \mathrm{int}(\Theta)$ the family $\mathbb{P}_\Theta$ is *Hellinger-differentiable with derivative* $\dot\ell_{\theta_o}$, if $\dot\ell_{\theta_o} \in \mathscr{L}^*_2(\mathbb{P}_{\theta_o})$, hence $\dot\ell_{\theta_o}\mathrm{L}^{1/2}_\mu(\theta_o) \in \mathscr{L}^*_2(\mu)$, and for $\theta \to \theta_o$

$$\|\mathrm{L}^{1/2}_\mu(\theta) - \mathrm{L}^{1/2}_\mu(\theta_o) - \tfrac12\langle\dot\ell_{\theta_o}, \theta-\theta_o\rangle\mathrm{L}^{1/2}_\mu(\theta_o)\|_{\mathscr{L}_2(\mu)} = \mathrm{o}(\|\theta-\theta_o\|).$$

Exploiting the identity (07.4) we obtain immediately the next property.

§07.14 **Property**. *Let $\mathbb{P}_\Theta \ll \mu \in \mathcal{M}_\sigma(\mathscr{X})$ and $\theta_o \in \mathrm{int}(\Theta)$. The family $\mathbb{P}_\Theta$ is Hellinger-differentiable with derivative $\dot\ell_{\theta_o}$ if and only if $\mathbb{P}_\Theta$ is $\mathscr{L}_2(\theta_o)$-differentiable with derivative $\dot\ell_{\theta_o}$.*

§07.15 **Proposition**. *Let $\mathbb{P}_\Theta \ll \mu \in \mathcal{M}_\sigma(\mathscr{X})$ with open $\Theta \subseteq \mathbb{R}^k$. If the likelihood $\mathrm{L}_\mu(\theta) := \mathrm{d}\mathbb{P}_\theta/\mathrm{d}\mu$, $\theta \in \Theta$, satisfies in addition the following conditions:*

(i) *for each $x \in \mathscr{X}$ the map $\theta \mapsto s(\theta, x) := \mathrm{L}^{1/2}_\mu(\theta, x)$ is continuously differentiable with derivative $\dot s := \frac{\partial}{\partial\theta}s$,*

(ii) *$\dot s(\theta) \in \mathscr{L}_2(\mu)$ for all $\theta \in \Theta$, and hence $\mathcal{I}_\theta := 4\mu(\dot s(\theta)\dot s(\theta)^t) \in \mathbb{R}^{(k,k)}_\geqslant$,*

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

Then $\mathbb{P}_\Theta$ is for all $\theta_0 \in \Theta$ Hellinger-differentiable with score function $\dot\ell_{\theta_o} = 2\frac{\dot s(\theta_o)}{s(\theta_o)}\mathbb{1}_{\{s(\theta_o)\in\mathbb{R}^+_{\setminus 0}\}}$.

§07.16 **Proof** of Proposition §07.15. is given in the lecture. □

§07.17 **Example**. Consider a statistical location model $(\mathbb{R}, \mathscr{B}, \mathbb{P}_\mathbb{R})$ dominated by the Lebesgue measure $\lambda \in \mathcal{M}_\sigma(\mathscr{B})$ with likelihood for each $\theta \in \mathbb{R}$ given by $\mathrm{L}(\theta, x) = g(x-\theta)$, $x \in \mathbb{R}$, 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 §07.15 the family $\mathbb{P}_\mathbb{R}$ is Hellinger-differentiable with score function $\dot\ell_\theta = -\dot g(x-\theta)/g(x-\theta)$. Indeed, setting $s(\theta, x) := \sqrt{g(x-\theta)}$, we have $\dot s(\theta, x) = \frac{\partial}{\partial\theta}\sqrt{g(x-\theta)} = -\frac12\dot g(x-\theta)/\sqrt{g(x-\theta)}$ which is continuous in θ and hence condition (i) is satisfied. Moreover conditions (ii) and (iii) hold true, since $\theta \mapsto \mathcal{I}_\theta = 4\lambda(\dot s(\theta))^2 = \lambda(|\dot g|^2/g) < \infty$ is constant and thus continuous. Applying Proposition §07.15 the family $\mathbb{P}_\mathbb{R}$ is Hellinger-differentiable with score function $\dot\ell_{\theta_o} = 2\frac{\dot s(\theta_o)}{s(\theta_o)}\mathbb{1}_{\{s(\theta_o)\in\mathbb{R}^+_{\setminus 0}\}} = -\dot g(x-\theta_o)/g(x-\theta_o)$. □

§07.02 Contiguity

We introduce next an asymptotic version of absolute continuity. In this section we restrict our attention to probability measures $\mathbb{P}_0^n, \mathbb{P}_1^n \in \mathcal{W}(\mathscr{X}_n)$, $n \in \mathbb{N}$, in short $(\mathbb{P}_0^n)_{n\in\mathbb{N}}, (\mathbb{P}_1^n)_{n\in\mathbb{N}} \in (\mathcal{W}(\mathscr{X}_n))_{n\in\mathbb{N}}$. We aim to obtain the limiting distribution of (test) statistics $S_n \in \mathscr{X}_n^k$, $n \in \mathbb{N}$, under $\mathbb{P}_1^n$ if its limiting distribution under $\mathbb{P}_0^n$ is known.

§07.18 **Definition.** Let $\mathbb{P}_0^n, \mathbb{P}_1^n \in \mathcal{W}(\mathscr{X}_n)$, $n \in \mathbb{N}$. The sequence $(\mathbb{P}_1^n)_{n\in\mathbb{N}}$ is called *contiguous* with respect to $(\mathbb{P}_0^n)_{n\in\mathbb{N}}$, symbolically $\mathbb{P}_1^n \lhd \mathbb{P}_0^n$, if for any sequence $(B_n)_{n\in\mathbb{N}} \in (\mathscr{X}_n)_{n\in\mathbb{N}}$ with $\lim_{n\to\infty} \mathbb{P}_0^n(B_n) = 0$ holds $\lim_{n\to\infty} \mathbb{P}_1^n(B_n) = 0$. The sequences $(\mathbb{P}_1^n)_{n\in\mathbb{N}}$ and $(\mathbb{P}_0^n)_{n\in\mathbb{N}}$ are called *mutually contiguous*, symbolically $\mathbb{P}_0^n \lhd\rhd \mathbb{P}_1^n$, if both $\mathbb{P}_1^n \lhd \mathbb{P}_0^n$ and $\mathbb{P}_0^n \lhd \mathbb{P}_1^n$. □

§07.19 **Lemma.** *Let $\mathbb{P}_0^n, \mathbb{P}_1^n \in \mathcal{W}(\mathscr{X}_n)$, $n \in \mathbb{N}$.*

(i) $\mathbb{P}_1^n \lhd \mathbb{P}_0^n \Leftrightarrow$ *for all* $(S_n)_{n\in\mathbb{N}} \in (\mathscr{X}_n^k)_{n\in\mathbb{N}}$ *holds:* $S_n \xrightarrow{\mathbb{P}_0^n} 0 \Rightarrow S_n \xrightarrow{\mathbb{P}_1^n} 0$;

(ii) *For any statistic* $S_n : (\mathscr{X}_n, \mathscr{X}_n) \to (\mathcal{S}, \mathscr{S})$, $n \in \mathbb{N}$, *holds:* $\mathbb{P}_1^n \lhd \mathbb{P}_0^n \Rightarrow \mathbb{P}_1^n \circ S_n^{-1} \lhd \mathbb{P}_0^n \circ S_n^{-1}$;

(iii) *For any sub-sequence* $(n_k)_{k\in\mathbb{N}}$ *in* $\mathbb{N}$ *holds:* $\mathbb{P}_1^n \lhd \mathbb{P}_0^n \Rightarrow \mathbb{P}_1^{n_k} \lhd \mathbb{P}_0^{n_k}$;

(iv) $\mathbb{P}_1^n \lhd \mathbb{P}_0^n \Leftrightarrow$ *for any* $\varepsilon \in \mathbb{R}_{\setminus 0}^+$ *exists* $\delta \in \mathbb{R}_{\setminus 0}^+$ *such that for all* $(B_n)_{n\in\mathbb{N}} \in (\mathscr{X}_n)_{n\in\mathbb{N}}$ *holds:* $\limsup_{n\to\infty} \mathbb{P}_0^n(B_n) < \delta \Rightarrow \limsup_{n\to\infty} \mathbb{P}_1^n(B_n) < \varepsilon$;

(v) *Let* $(S_n)_{n\in\mathbb{N}} \in (\mathscr{X}_n^k)_{n\in\mathbb{N}}$ *and* $\mathbb{P}_1^n \lhd \mathbb{P}_0^n$, *then:*

(v-a) $\mathbb{P}_0^n \circ S_n^{-1} \xrightarrow{d} \mathbb{P}_0$ *and* $\mathbb{P}_1^n \circ S_n^{-1} \xrightarrow{d} \mathbb{P}_1 \Rightarrow \mathbb{P}_1 \ll \mathbb{P}_0$;

(v-b) $(\mathbb{P}_0^n \circ S_n^{-1})_{n\in\mathbb{N}}$ *tight* $\Rightarrow (\mathbb{P}_1^n \circ S_n^{-1})_{n\in\mathbb{N}}$ *tight.*

§07.20 **Proof** of Lemma §07.19. (i) „⇒“ and its converse follows applying the definition on $B_n = \{\|S_n\| > \varepsilon\}$ for any $\varepsilon \in \mathbb{R}_{\setminus 0}^+$ and $S_n = \mathbb{1}_{B_n}$, $n \in \mathbb{N}$, respectively. (ii) and (iii) follow immediately from the definition. For the proof of (iv) and (v-a) we refer to Witting and Müller-Funk [1995] (Hilfssatz 6.111, p.294 and Satz 6.113, p.295). The proof of (v-b) is given in the lecture. □

§07.21 **Remark.** Next we characterise contiguity in terms of the asymptotic behaviour of the likelihood ratio $\mathrm{L}_n = \mathrm{d}\mathbb{P}_1^n/\mathrm{d}\mathbb{P}_0^n \in \overline{\mathscr{X}_n}^+$, $n \in \mathbb{N}$. First recall that $\mathbb{P}_1^n(\mathrm{L}_n < \infty) = \mathbb{P}_0^n(\mathrm{L}_n) \in [0, 1]$ and $\mathbb{P}_0^n(\mathrm{L}_n = \infty) = \mathbb{P}_1^n(\mathrm{L}_n = 0) = 0$ for each $n \in \mathbb{N}$. Consequently, the probability measure $\mathbb{P}_0^n \circ \mathrm{L}_n^{-1} \in \mathcal{W}(\overline{\mathscr{B}})$ is concentrated in $\mathbb{R}^+$ meaning that $\mathbb{P}_0^n \circ \mathrm{L}_n^{-1}(\mathbb{R}^+) = \mathbb{P}_0^n(\mathrm{L}_n \in \mathbb{R}^+) = 1$ for each $n \in \mathbb{N}$. Moreover, $(\mathbb{P}_0^n \circ \mathrm{L}_n^{-1})_{n\in\mathbb{N}}$ is tight, since for any $\varepsilon \in \mathbb{R}_{\setminus 0}^+$ and $c > 1/\varepsilon$ holds $\mathbb{P}_0^n(\mathrm{L}_n > c) \leqslant \frac{1}{c}\mathbb{P}_0^n(\mathrm{L}_n) \leqslant \frac{1}{c} < \varepsilon$ by Markov's inequality. However, $\mathbb{P}_1^n \circ \mathrm{L}_n^{-1}$ is generally not concentrated in $\mathbb{R}^+$, but under $\mathbb{P}_1^n \lhd \mathbb{P}_0^n$ holds $\mathbb{P}_1^n(\mathrm{L}_n = \infty) \to 0$ since $\mathbb{P}_0^n(\mathrm{L}_n = \infty) = 0$ for all $n \in \mathbb{N}$. Thereby, the limit distribution of $\mathbb{P}_1^n \circ \mathrm{L}_n^{-1}$ (if it exists) is concentrated in $\mathbb{R}^+$. □

Formally, we write $\mathrm{L}_n = \mathrm{L}_n \mathbb{1}_{\{\mathrm{L}_n < \infty\}} + \infty \mathbb{1}_{\{\mathrm{L}_n = \infty\}}$, where the second summand is negligible in the sense of Slutzky's lemma under contiguity $\mathbb{P}_1^n \lhd \mathbb{P}_0^n$.

§07.22 **Definition.** $(T_n)_{n\in\mathbb{N}} \in (\overline{\mathscr{X}}_n)_{n\in\mathbb{N}}$ converges in distribution to $\mathbb{P}^T \in \mathcal{W}(\mathscr{B})$ under $\mathbb{P}^n$, shortly $T_n \xrightarrow{d} \mathbb{P}^T$ *under* $\mathbb{P}^n$, if

$$\mathbb{P}^n \circ T_n^{-1} \xrightarrow{d} \mathbb{P}^T \quad :\Leftrightarrow \quad \mathbb{P}^n \circ (T_n \mathbb{1}_{\{T_n \in \mathbb{R}\}})^{-1} \xrightarrow{d} \mathbb{P}^T \quad \text{and} \quad \mathbb{P}^n(T_n \notin \mathbb{R}) \to 0. \tag{07.5}$$

We note that any family of probability measures on $(\overline{\mathbb{R}}, \overline{\mathscr{B}})$ is tight, since $\overline{\mathbb{R}}$ is compact. A non trivial formulation of tightness for probability measures provides the next definition.

§07.23 **Definition**. A sequence $(\mathbb{P}^n)_{n\in\mathbb{N}} \in \mathcal{W}(\overline{\mathscr{B}})$ is called *asymptotically tight* if for all $\varepsilon \in \mathbb{R}^+_{\setminus 0}$ exists $M \in \mathbb{R}^+_{\setminus 0}$ and $n_o \in \mathbb{N}$ such that for all $n > n_o$ holds $\mathbb{P}^n([-M, M]^c) < \varepsilon$. □

§07.24 **Remark**. Asymptotic tightness of $(\mathbb{P}^n)_{n\in\mathbb{N}} \in \mathcal{W}(\overline{\mathscr{B}})$ is equivalently characterised by: for any $(M_n)_{n\in\mathbb{N}}$ in $\mathbb{R}$ with $M_n \uparrow \infty$ holds $\mathbb{P}^n([-M_n, M_n]^c) \xrightarrow{n\to\infty} 0$. In particular, we have immediately $\mathbb{P}^n(\{-\infty, \infty\}) \xrightarrow{n\to\infty} 0$. The concept of asymptotic tightness and tightness as in Definition §02.21 coincide if $\mathbb{P}^n(\mathbb{R}) = 1$ for all $n \in \mathbb{N}$. Furthermore, it can be shown that the claim of Prohorov's theorem Property §02.24 holds also for families of asymptotically tight probability measures. □

§07.25 **Theorem**. *For each $n \in \mathbb{N}$ let $\mathbb{P}^n_0, \mathbb{P}^n_1 \in \mathcal{W}(\mathscr{X}_n)$, let $\mathrm{L}_n := \mathrm{d}\mathbb{P}^n_1/\mathrm{d}\mathbb{P}^n_0 \in \overline{\mathscr{X}_n}^+$ be a likelihood ratio of $\mathbb{P}^n_1$ with respect to $\mathbb{P}^n_0$ and let $\mathbb{P}^{\mathrm{L}}_0, \mathbb{P}^{\mathrm{L}}_1 \in \mathcal{W}(\mathscr{B})$. Then the following statements are equivalent:*

(a1) $\mathbb{P}^n_1 \lhd \mathbb{P}^n_0$;

(a2) $\mathbb{P}^n_0(\mathrm{L}_n) \xrightarrow{n\to\infty} 1$ *and for any $\varepsilon \in \mathbb{R}^+_{\setminus 0}$ exists $M \in \mathbb{R}^+_{\setminus 0}$ with $\sup_{n\in\mathbb{N}} \mathbb{P}^n_0(\mathrm{L}_n \mathbb{1}_{\{\mathrm{L}_n > M\}}) < \varepsilon$, i.e. $(\mathbb{P}^n_0 \circ \mathrm{L}_n^{-1})_{n\in\mathbb{N}}$ is uniformly integrable;*

(a3) $(\mathbb{P}^n_1 \circ \mathrm{L}_n^{-1})_{n\in\mathbb{N}}$ *is asymptotically tight.*

If in addition $\mathrm{L}_n \xrightarrow{d} \mathbb{P}^{\mathrm{L}}_0$ under $\mathbb{P}^n_0$, i.e. $\mathbb{P}^n_0 \circ \mathrm{L}_n^{-1} \xrightarrow{d} \mathbb{P}^{\mathrm{L}}_0$, then the following statements are equivalent:

(b1) $\mathbb{P}^n_1 \lhd \mathbb{P}^n_0$;

(b2) $1 = \int_{\mathbb{R}} y\mathbb{P}^{\mathrm{L}}_0(dy) = \mathbb{P}^{\mathrm{L}}_0(\mathrm{id}_{\mathbb{R}}) = \mathbb{P}^{\mathrm{L}}_0(\mathrm{id}_{\mathbb{R}}\mathbb{1}_{\mathbb{R}})$;

(b3) $\mathrm{L}_n \xrightarrow{d} \mathbb{P}^{\mathrm{L}}_1$ *under $\mathbb{P}^n_1$ with $\mathbb{P}^{\mathrm{L}}_1(B) = \mathbb{P}^{\mathrm{L}}_0(\mathrm{id}_{\mathbb{R}}\mathbb{1}_B) = \int_B y\mathbb{P}^{\mathrm{L}}_0(dy)$ for all $B \in \mathscr{B}$.*

§07.26 **Proof** of Theorem §07.25. is given in the lecture. □

Since $\mathbb{P}^n_0(\mathrm{L}_n) = \mathbb{P}^n_1(\mathrm{L}_n < \infty)$ it holds $\mathbb{P}^n_0(\mathrm{L}_n) \to 1 \Leftrightarrow \mathbb{P}^n_1(\mathrm{L}_n = \infty) \to 0$. Keeping (07.1) in mind the mass of the absolute continuous part of $\mathbb{P}^n_1$ with respect to $\mathbb{P}^n_0$ converges two 1, if and only if, the singular part vanishes.

§07.27 **Corollary**. *Under the notations of Theorem §07.25 the following statements are equivalent:*

(i) $\mathbb{P}^n_1 \lhd \mathbb{P}^n_0$;

(ii) *if $\mathbb{P}^{n_k}_0 \circ \mathrm{L}_{n_k}^{-1} \xrightarrow{d} \mathbb{P}^{\mathrm{L}}_0 \in \mathcal{W}(\mathscr{B})$ along a sub-sequence $(n_k)_{k\in\mathbb{N}}$, then $\mathbb{P}^{\mathrm{L}}_0(\mathrm{id}_{\mathbb{R}}) = 1$;*

(iii) *if $\mathbb{P}^{n_k}_0 \circ \mathrm{L}_{n_k}^{-1} \xrightarrow{d} \mathbb{P}^{\mathrm{L}}_0 \in \mathcal{W}(\mathscr{B})$ along a sub-sequence $(n_k)_{k\in\mathbb{N}}$, then $\mathbb{P}^{n_k}_1 \circ \mathrm{L}_{n_k}^{-1} \xrightarrow{d} \mathbb{P}^{\mathrm{L}}_1$, with $\mathbb{P}^{\mathrm{L}}_1(B) = \mathbb{P}^{\mathrm{L}}_0(\mathrm{id}_{\mathbb{R}}\mathbb{1}_B)$ for all $B \in \mathscr{B}$.*

§07.28 **Proof** of Corollary §07.27. Since $(\mathbb{P}^n_0 \circ \mathrm{L}_n^{-1})_{n\in\mathbb{N}}$ is tight (Remark §07.21) the claim follows from Theorem §07.25 (b1)-(b3) by applying Prohorov's theorem §02.24. □

We are particularly interested in mutual contiguity $(\mathbb{P}^n_0 \lhd\rhd \mathbb{P}^n_1)$ of $(\mathbb{P}^n_0)_{n\in\mathbb{N}}$ and $(\mathbb{P}^n_1)_{n\in\mathbb{N}}$, which can be characterised by applying Theorem §07.25 and its analogous formulation switching the roles of $\mathbb{P}^n_0$ and $\mathbb{P}^n_1$. However, for $n \in \mathbb{N}$ the transformation of a likelihood ratio $\mathrm{L}_n = \mathrm{d}\mathbb{P}^n_1/\mathrm{d}\mathbb{P}^n_0$ into a *log-likelihood ratio* (LLR) $\ell_n := \log \mathrm{L}_n = \log\left(\mathrm{d}\mathbb{P}^n_1/\mathrm{d}\mathbb{P}^n_0\right) \in \overline{\mathscr{X}}$ captures equally both orthogonal events $\{\mathrm{L}_n = 0\}$ and $\{\mathrm{L}_n = \infty\}$. Generally, ℓ_n takes the value $-\infty$ and $+\infty$

with positive $\mathbb{P}_0^n$- and $\mathbb{P}_1^n$-probability, respectively. In other words $\mathbb{P}_0^n \circ \ell_n^{-1}$ and $\mathbb{P}_1^n \circ \ell_n^{-1}$ is concentrated in $[-\infty, \infty)$ and $(-\infty, \infty]$, respectively, since by Definition §07.03 of L_n it holds

$$\mathbb{P}_0^n(\ell_n = \infty) = 0 \quad \text{and} \quad \mathbb{P}_1^n(\ell_n = -\infty) = 0 \quad \text{for all } n \in \mathbb{N}. \tag{07.6}$$

Thereby, similar to Remark §07.21 under mutual contiguity $\mathbb{P}_0^n \lhd\rhd \mathbb{P}_1^n$ it follows

$$\mathbb{P}_1^n(\ell_n = \infty) \to 0 \quad \text{and} \quad \mathbb{P}_0^n(\ell_n = -\infty) \to 0 \quad \text{as } n \to \infty. \tag{07.7}$$

Consequently, the limit distribution of ℓ_n under both $\mathbb{P}_0^n$ and $\mathbb{P}_1^n$, if it exists, is concentrated in $\mathbb{R}$. Keeping Definition §07.22 in mind under mutual contiguity $\mathbb{P}_0^n \lhd\rhd \mathbb{P}_1^n$ convergence in distribution of ℓ_n under $\mathbb{P}_0^n$ and $\mathbb{P}_1^n$ to $\mathbb{P}_0^\ell, \mathbb{P}_1^\ell \in \mathcal{W}(\mathscr{B})$, respectively, is equivalently characterised by

$$\begin{aligned} \mathbb{P}_0^n \circ \ell_n^{-1} \xrightarrow{d} \mathbb{P}_0^\ell \quad &\Leftrightarrow \quad \mathbb{P}_0^n \circ (\ell_n \mathbb{1}_{\{\ell_n > -\infty\}})^{-1} \xrightarrow{d} \mathbb{P}_0^\ell \quad \text{and} \\ \mathbb{P}_1^n \circ \ell_n^{-1} \xrightarrow{d} \mathbb{P}_1^\ell \quad &\Leftrightarrow \quad \mathbb{P}_1^n \circ (\ell_n \mathbb{1}_{\{\ell_n < \infty\}})^{-1} \xrightarrow{d} \mathbb{P}_1^\ell. \end{aligned} \tag{07.8}$$

If $\mathrm{L}_n^{-1} = \mathrm{d}\mathbb{P}_0^n / \mathrm{d}\mathbb{P}_1^n$ is a likelihood ratio of $\mathbb{P}_0^n$ with respect to $\mathbb{P}_1^n$, as for example in Remark §07.05, then making use of the identity $\log \mathrm{L}_n^{-1} = -\log \mathrm{L}_n = -\ell_n$ the convergence in distribution of ℓ_n under $\mathbb{P}_0^n$ respectively $\mathbb{P}_1^n$ implies immediately the corresponding convergence of $\log \mathrm{L}_n^{-1}$. Similar to Theorem §07.25 (b1)-(b3) the next result characterises mutual contiguity in terms of the log-likelihood ratio ℓ_n.

§07.29 **Theorem**. *For each $n \in \mathbb{N}$ let $\mathbb{P}_0^n, \mathbb{P}_1^n \in \mathcal{W}(\mathscr{X}_n)$, let $\ell_n := \log \mathrm{L}_n = \log(\mathrm{d}\mathbb{P}_1^n / \mathrm{d}\mathbb{P}_0^n) \in \overline{\mathscr{X}_n}$ be a log-likelihood ratio such that also $\mathrm{L}_n^{-1} = \mathrm{d}\mathbb{P}_0^n / \mathrm{d}\mathbb{P}_1^n \in \overline{\mathscr{X}_n}^+$ and let $\mathbb{P}_0^\ell, \mathbb{P}_1^\ell \in \mathcal{W}(\mathscr{B})$. If in addition $\ell_n \xrightarrow{d} \mathbb{P}_0^\ell$ under $\mathbb{P}_0^n$, i.e. $\mathbb{P}_0^n \circ \ell_n^{-1} \xrightarrow{d} \mathbb{P}_0^\ell$, then the following statements are equivalent:*

(b'1) $\mathbb{P}_1^n \lhd\rhd \mathbb{P}_0^n$;

(b'2) $1 = \int_{\mathbb{R}} \exp(z) \mathbb{P}_0^\ell(dz) = \mathbb{P}_0^\ell(\exp) = \mathbb{P}_0^\ell(\exp \mathbb{1}_{\mathbb{R}})$

(b'3) $\ell_n \xrightarrow{d} \mathbb{P}_1^\ell$ *under* $\mathbb{P}_1^n$ *with* $\mathbb{P}_1^\ell(B) = \mathbb{P}_0^\ell(\exp \mathbb{1}_B) = \int_B \exp(z) \mathbb{P}_0^\ell(dz)$ *for all* $B \in \mathscr{B}$.

§07.30 **Proof** of Theorem §07.29. is given in the lecture. □

§07.31 **Remark**. Let $\mathfrak{f}_0^\ell$ and $\mathfrak{f}_1^\ell$ denote, respectively, a μ-density of $\mathbb{P}_0^\ell$ and $\mathbb{P}_1^\ell$ with respect to a measure $\mu \in \mathcal{M}_\sigma(\mathscr{B})$ dominating $\mathbb{P}_0^\ell$, and hence $\mathbb{P}_1^\ell$. The measure $\mathbb{P}_1^\ell$ in Theorem §07.29 (b'3) is equally defined by $\mathfrak{f}_1^\ell(z) = \exp(z) \mathfrak{f}_0^\ell(z)$ for μ-a.e. $z \in \mathbb{R}$. □

§07.32 **Corollary**. *Under the notations of Theorem §07.29 if $\mathbb{P}_0^n \circ \ell_n^{-1} \xrightarrow{d} \mathrm{N}_{(\mu,\sigma^2)}$ for $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+$ then the following statements are equivalent:*

(b"1) $\mathbb{P}_1^n \lhd\rhd \mathbb{P}_0^n$;

(b"2) $\mu = -\sigma^2/2$

(b"3) $\ell_n \xrightarrow{d} \mathrm{N}_{(\sigma^2/2, \sigma^2)}$ *under* $\mathbb{P}_1^n$.

§07.33 **Proof** of Corollary §07.32. is given in the lecture. □

§07.34 **Example** *(Le Cam's first lemma)*. For $n \in \mathbb{N}$ let $\mathbb{P}_0^n, \mathbb{P}_1^n \in \mathcal{W}(\mathscr{X}_n)$ and $\mathrm{L}_n := \mathrm{d}\mathbb{P}_1^n / \mathrm{d}\mathbb{P}_0^n \in \overline{\mathscr{X}_n}^+$. If $\ell_n := \log \mathrm{L}_n \xrightarrow{d} \mathrm{N}_{(-\sigma^2/2, \sigma^2)}$ under $\mathbb{P}_0^n$, then $\mathbb{P}_1^n \lhd\rhd \mathbb{P}_0^n$ and $\ell_n \xrightarrow{d} \mathrm{N}_{(\sigma^2/2, \sigma^2)}$ under $\mathbb{P}_1^n$

due to Corollary §07.32. For $\sigma > 0$ from $\sigma^{-1}(\ell_n + \sigma^2/2) \xrightarrow{d} \mathrm{N}_{(0,1)}$ under $\mathbb{P}_0^n$ follows thus $\sigma^{-1}(\ell_n + \sigma^2/2) \xrightarrow{d} \mathrm{N}_{(\sigma,1)}$ under $\mathbb{P}_1^n$. In other words in this situation there is asymptotically a location shift by σ. □

For each $n \in \mathbb{N}$ let $\mathbb{P}_0^n, \mathbb{P}_1^n \in \mathcal{W}(\mathscr{X}_n)$, let $\mathrm{L}_n := \mathrm{d}\mathbb{P}_1^n/\mathrm{d}\mathbb{P}_0^n \in \overline{\mathscr{X}_n}^+$ be a likelihood ratio of $\mathbb{P}_1^n$ with respect to $\mathbb{P}_0^n$, let $\ell_n := \log \mathrm{L}_n$ and let $S_n \in \mathscr{X}_n^k$ be a $\mathbb{R}^k$-valued statistic defined on $(\mathcal{X}_n, \mathscr{X}_n)$. We search conditions which allow to calculate the limiting distribution of (S_n, L_n) respectively (S_n, ℓ_n) under $\mathbb{P}_1^n$, from the limiting distribution of (S_n, L_n) respectively (S_n, ℓ_n) under $\mathbb{P}_0^n$. Keeping again (§07.22) in mind under mutual contiguity $\mathbb{P}_0^n \lhd\rhd \mathbb{P}_1^n$ the joint convergence in distribution of $(S_n, \mathrm{L}_n) \xrightarrow{d} \mathbb{P}_1^{(S,\mathrm{L})} \in \mathcal{W}(\mathscr{B}^{k+1})$ under $\mathbb{P}_1^n$, $(S_n, \ell_n) \xrightarrow{d} \mathbb{P}_0^{(S,\ell)} \in \mathcal{W}(\mathscr{B}^{k+1})$ under $\mathbb{P}_0^n$ and $(S_n, \ell_n) \xrightarrow{d} \mathbb{P}_1^{(S,\ell)} \in \mathcal{W}(\mathscr{B}^{k+1})$ under $\mathbb{P}_1^n$, respectively, is equally characterised by

$$\begin{aligned}
&\mathbb{P}_1^n \circ (S_n, \mathrm{L}_n)^{-1} \xrightarrow{d} \mathbb{P}_1^{(S,\mathrm{L})} \quad \Leftrightarrow \quad \mathbb{P}_1^n \circ (S_n, \mathrm{L}_n \mathbb{1}_{\{\mathrm{L}_n < \infty\}})^{-1} \xrightarrow{d} \mathbb{P}_1^{(S,\mathrm{L})}, \\
&\mathbb{P}_0^n \circ (S_n, \ell_n)^{-1} \xrightarrow{d} \mathbb{P}_0^{(S,\ell)} \quad \Leftrightarrow \quad \mathbb{P}_0^n \circ (S_n, \ell_n \mathbb{1}_{\{\ell_n > -\infty\}})^{-1} \xrightarrow{d} \mathbb{P}_0^{(S,\ell)} \quad \text{and} \\
&\mathbb{P}_1^n \circ (S_n, \ell_n)^{-1} \xrightarrow{d} \mathbb{P}_1^{(S,\ell)} \quad \Leftrightarrow \quad \mathbb{P}_1^n \circ (S_n, \ell_n \mathbb{1}_{\{\ell_n < \infty\}})^{-1} \xrightarrow{d} \mathbb{P}_1^{(S,\ell)}. \qquad (07.9)
\end{aligned}$$

Denote by $\Pi_{\mathrm{L}} := \Pi_{k+1} \in \mathscr{B}^{k+1}$, i.e. $y^{k+1} = (y_i)_{i \in [\![k+1]\!]} \mapsto \Pi_{\mathrm{L}}(y^{k+1}) := y_{k+1}$ (respectively $\Pi_\ell := \Pi_{k+1} \in \mathscr{B}^{k+1}$) the coordinate map which allows us to write $\int_C y \mathbb{P}_1^{(S,\mathrm{L})}(ds, dy) = \int_{\mathbb{R}^{k+1}} \mathbb{1}_C(s,y) \Pi_{\mathrm{L}}(s,y) \mathbb{P}_1^{(S,\mathrm{L})}(ds, dy) = \mathbb{P}_1^{(S,\mathrm{L})}(\mathbb{1}_C \Pi_{\mathrm{L}})$ for all $C \in \mathscr{B}^{k+1}$.

§07.35 **Theorem**. *For each $n \in \mathbb{N}$ let $\mathbb{P}_0^n, \mathbb{P}_1^n \in \mathcal{W}(\mathscr{X}_n)$, let $\ell_n = \log \mathrm{L}_n = \log(\mathrm{d}\mathbb{P}_1^n/\mathrm{d}\mathbb{P}_0^n) \in \overline{\mathscr{X}_n}$ be a log-likelihood ratio, and let $S_n \in \mathscr{X}_n^k$ be a $\mathbb{R}^k$-valued statistic. Then, we have*

(i) *If $(S_n, \mathrm{L}_n) \xrightarrow{d} \mathbb{P}_0^{(S,\mathrm{L})} \in \mathcal{W}(\mathscr{B}^{k+1})$ under $\mathbb{P}_0^n$ and $\mathbb{P}_0^{(S,\mathrm{L})}(\Pi_{\mathrm{L}} \mathbb{1}_{\mathbb{R}^{k+1}}) = \mathbb{P}_0^{(S,\mathrm{L})}(\Pi_{\mathrm{L}}) = 1$, then $(S_n, \mathrm{L}_n) \xrightarrow{d} \mathbb{P}_1^{(S,\mathrm{L})}$ under $\mathbb{P}_1^n$ with $\mathbb{P}_1^{(S,\mathrm{L})}(C) := \mathbb{P}_0^{(S,\mathrm{L})}(\Pi_{\mathrm{L}} \mathbb{1}_C)$ for all $C \in \mathscr{B}^{k+1}$.*

(ii) *If $(S_n, \ell_n) \xrightarrow{d} \mathbb{P}_0^{(S,\ell)} \in \mathcal{W}(\mathscr{B}^{k+1})$ under $\mathbb{P}_0^n$ and $\mathbb{P}_0^{(S,\ell)}(\exp(\Pi_\ell) \mathbb{1}_{\mathbb{R}^{k+1}}) = \mathbb{P}_0^{(S,\ell)}(\exp(\Pi_\ell)) = 1$, then $(S_n, \ell_n) \xrightarrow{d} \mathbb{P}_1^{(S,\ell)}$ under $\mathbb{P}_1^n$ with $\mathbb{P}_1^{(S,\ell)}(C) := \mathbb{P}_0^{(S,\ell)}(\exp(\Pi_\ell) \mathbb{1}_C)$ for all $C \in \mathscr{B}^{k+1}$.*

§07.36 **Proof** of Theorem §07.35. is given in the lecture. □

§07.37 **Example** *(Le Cam's third lemma)*. For each $n \in \mathbb{N}$ let $\mathbb{P}_0^n, \mathbb{P}_1^n \in \mathcal{W}(\mathscr{X}_n)$, let $\ell_n = \log \mathrm{L}_n = \log(\mathrm{d}\mathbb{P}_1^n/\mathrm{d}\mathbb{P}_0^n) \in \overline{\mathscr{X}_n}$ be a log-likelihood ratio, and let $S_n \in \mathscr{X}_n^k$ be a $\mathbb{R}^k$-valued statistic. Suppose that the limit distribution of (S_n, ℓ_n) under $\mathbb{P}_0^n$ is multivariate normal, that is

$$\mathbb{P}_0^n \circ (S_n, \ell_n)^{-1} \xrightarrow{d} \mathbb{P}_0^{(S,\ell)} = \mathrm{N}_{(v,M)} \quad \text{with} \quad v = \begin{pmatrix} \mu \\ -\frac{\sigma^2}{2} \end{pmatrix} \text{ and } M = \begin{pmatrix} \Sigma & \tau \\ \tau^t & \sigma^2 \end{pmatrix}. \qquad (07.10)$$

Then it holds $(S_n, \ell_n) \xrightarrow{d} \mathbb{P}_1^{(S,\ell)} = \mathrm{N}_{(v',M)}$ under $\mathbb{P}_1^n$ with $v' = (\mu + \tau, \sigma^2/2)^t$. Indeed, since $\mathbb{P}_0^{(S,\ell)}(\exp(\Pi_\ell)) = 1$ both assumptions of Theorem §07.35 (ii) are satisfied and hence it remains to calculate the limit distribution $\mathbb{P}_1^{(S,\ell)}(C) := \mathbb{P}_0^{(S,\ell)}(\exp(\Pi_\ell) \mathbb{1}_C)$ for all $C \in \mathscr{B}^{k+1}$. Suppose first $M > 0$, or equivalently $\Sigma > 0$ and $\sigma > 0$, then $\mathbb{P}_0^{(S,\ell)}$ has a density $\mathfrak{f}_0^{(S,\ell)}$ with respect to the Lebesgue-measure $\lambda^{k+1} \in \mathcal{M}_\sigma(\mathscr{B}^{k+1})$ and (see Remark §07.31) the Lebesgue-density $\mathfrak{f}_1^{(S,\ell)}$ of $\mathbb{P}_1^{(S,\ell)}$ satisfies $\mathfrak{f}_1^{(S,\ell)}(s,z) = \exp(z) \mathfrak{f}_0^{(S,\ell)}(s,z)$ for λ^{k+1}-a.e. $(s,z) \in \mathbb{R}^{k+1}$. Keeping the coordinate map Π_ℓ in mind we denote by $\mathfrak{f}_0^\ell$ and $\mathfrak{f}_1^\ell$ the marginal density of $\mathbb{P}_0^{(S,\ell)} \circ \Pi_\ell$ and

$\mathbb{P}_1^{(S,\ell)} \circ \Pi_\ell$, respectively. Denoting by $\mathfrak{f}_0^{S|\ell=z}$ and $\mathfrak{f}_1^{S|\ell=z}$, respectively, a conditional density of S given $\ell = z$ under the joint distribution $\mathbb{P}_1^{(S,\ell)}$ and $\mathbb{P}_1^{(S,\ell)}$ (see Notation §03.11 (iv)) we have $\mathfrak{f}_1^{S|\ell=z}(s)\mathfrak{f}_1^{\ell}(z) = \exp(z)\mathfrak{f}_0^{S|\ell=z}(s)\mathfrak{f}_0^{\ell}(z)$ for λ^{k+1}-a.e. $(s,z) \in \mathbb{R}^{k+1}$. Exploiting Theorem §07.29 (b'3) it holds $\mathfrak{f}_1^{\ell}(z) = \exp(z)\mathfrak{f}_0^{\ell}(z)$ for λ-a.e. $z \in \mathbb{R}$ (see Remark §07.31). Consequently, it remains to verify that $\mathrm{N}_{(v,M)}$ and $\mathrm{N}_{(v',M)}$ have the same conditional distribution given $\ell = z$. Indeed, both are again multivariate normal (see Notation §03.11 (v)) with equal covariance matrix $\Sigma - \sigma^2\tau\tau^t$ and conditional mean $\mathbb{P}_0^{S|\ell=z}(\mathrm{id}_{\mathbb{R}^k}) = \mu + \sigma^{-2}\tau(z + \sigma^2/2) = \mu + \tau + \sigma^{-2}\tau(z - \sigma^2/2) = \mathbb{P}_1^{S|\ell=z}(\mathrm{id}_{\mathbb{R}^k})$. The case of a positive semi-definite Σ and $\sigma^2 > 0$ follows by similar arguments when considering the projection onto the image of Σ. If $\sigma = 0$ the claim follows from Lemma §07.19 (i) together with Slutzky's lemma §02.10. In particular, note that $S_n \xrightarrow{d} \mathrm{N}_{(\mu,\Sigma)}$ under $\mathbb{P}_0^n$ and $S_n \xrightarrow{d} \mathrm{N}_{(\mu+\tau,\Sigma)}$ under $\mathbb{P}_1^n$ (see Reminder §07.07). □

§08 Local asymptotic normality (LAN)

§08.01 **Aim.** For each $n \in \mathbb{N}$ let $(\mathcal{X}_n, \mathscr{X}_n, \mathbb{P}_\Theta^n = (\mathbb{P}_\theta^n)_{\theta\in\Theta})$ with $\Theta \subseteq \mathbb{R}^k$ be a statistical experiment. We aim to approximate $(\mathcal{X}_n, \mathscr{X}_n, \mathbb{P}_\Theta^n)$ in a certain sense by a Gaussian location model after suitable reparametrisation.

§08.02 **Reminder.** Consider on $(\mathbb{R}^k, \mathscr{B}^k)$ the family $\mathrm{N}_{\mathbb{R}^k\times\{\Sigma\}} := (\mathrm{N}_{(h,\Sigma)})_{h\in\mathbb{R}^k}$ of multivariate normal distributions with common strictly positive definite covariance matrix $\Sigma \in \mathbb{R}_>^{(k,k)}$ and log-likelihood ratio $\log\big(\mathrm{dN}_{(h,\Sigma)}/\mathrm{dN}_{(0,\Sigma)}\big)(z) = \langle\Sigma^{-1}h, z\rangle - \frac{1}{2}\langle\Sigma^{-1}h, h\rangle$, $z \in \mathbb{R}^k$. Noting that for each $h \in \mathbb{R}^k$ the likelihood $\mathrm{L}(h) = \mathrm{dN}_{(h,\Sigma)}/\mathrm{d}\lambda^k$ of $\mathrm{N}_{(h,\Sigma)}$ with respect to the Lebesgue measure λ^k on $\mathbb{R}^k$ satisfies $\mathrm{L}(h,x) = \mathrm{L}(0, x-h)$ for all $x \in \mathbb{R}^k$ the statistical experiment $(\mathbb{R}^k, \mathscr{B}^k, \mathrm{N}_{\mathbb{R}^k\times\{\Sigma\}})$ is called a *Gaussian location model*. □

Consider a localised reparametrisation centred around a parameter value $\theta_o \in \mathrm{int}(\Theta)$ which is in the sequel regarded as fixed.

§08.03 **Definition.** Consider a sequence of statistical experiments $(\mathcal{X}_n, \mathscr{X}_n, \mathbb{P}_\Theta^n)$, $n \in \mathbb{N}$, 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* $\Theta_o^n := \{\delta_n^{-1}(\theta - \theta_o) : \theta \in \Theta\} \subseteq \mathbb{R}^k$. For each $\theta \in \Theta$ and associated *local parameter* $h = \delta_n^{-1}(\theta - \theta_o) \in \Theta_o^n$ rewriting $\mathbb{P}_\theta^n$ as $\mathbb{P}_{\theta_o+\delta_n h}^n$ we obtain a sequence of *localised statistical experiment* $(\mathcal{X}_n, \mathscr{X}_n, \mathbb{P}_{\delta_n\Theta_o^n+\theta_o}^n := (\mathbb{P}_{\theta_o+\delta_n h}^n)_{h\in\Theta_o^n})$, $n \in \mathbb{N}$. □

§08.04 **Remark.** In the sequel we eventually take the local parameter set Θ_o^n equal to $\mathbb{R}^k$ which is not correct if the parameter set Θ is a strict subset of $\mathbb{R}^k$. However, if $\theta_o \in \mathrm{int}(\Theta)$ 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_o^n converges to the whole of $\mathbb{R}^k$ as $n \to \infty$, i.e., $\cup_{n\in\mathbb{N}}\Theta_o^n = \mathbb{R}^k$. 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. □

§08.05 **Aim.** We show, for large n, that the localised statistical experiment $(\mathcal{X}_n, \mathscr{X}_n, \mathbb{P}_{\delta_n\mathbb{R}^k+\theta_o}^n)$ and a Gaussian location model $(\mathbb{R}^k, \mathscr{B}^k, \mathrm{N}_{\mathbb{R}^k\times\{\mathcal{I}_{\theta_o}^{-1}\}})$ are similar in statistical properties whenever the original experiments, i.e., $\theta \mapsto \mathbb{P}_\theta$, are "smooth".

§08.06 **Heuristics.** Consider a statistical experiment $(\mathcal{X}, \mathscr{X}, \mathbb{P}_\Theta)$ dominated by $\mu \in \mathcal{M}_\sigma(\mathscr{X})$, i.e.,

$\mathbb{P}_\Theta \ll \mu$, with $\Theta \subseteq \mathbb{R}$, positive real likelihood $\mathrm{L}(\theta) = \mathrm{d}\mathbb{P}_\theta/\mathrm{d}\mu \in \mathscr{X}^+$ and log-likelihood $\ell = \log \mathrm{L}$. Assume that for all $x \in \mathcal{X}$, the map $\theta \mapsto \ell(\theta, x)$ is twice differentiable with derivatives $\dot\ell := \frac{\partial}{\partial\theta}\ell$ and $\ddot\ell := \frac{\partial^2}{\partial^2\theta}\ell$. A *Taylor expansion* of the log-likelihood ratio leads to $\ell(\theta+h,x) - \ell(\theta,x) = h\dot\ell(\theta,x) + \frac{1}{2}h^2\ddot\ell(\theta,x) + \mathrm{o}_x(h^2)$ where the remainder term depends on x. Considering a product experiment $(\mathcal{X}^n, \mathscr{X}^{\otimes n}, \mathbb{P}_\Theta^{\otimes n})$ eventually it holds $\log(\mathrm{d}\mathbb{P}^{\otimes n}_{\theta+h/\sqrt{n}}/\mathrm{d}\mathbb{P}^{\otimes n}_\theta) = h\sqrt{n}\widehat{\mathbb{P}}_n\big(\dot\ell(\theta)\big) + \frac{1}{2}h^2\widehat{\mathbb{P}}_n\big(\ddot\ell(\theta)\big) + R_n$ where the score $\dot\ell$ has mean zero, i.e., $\mathbb{P}_\theta\big(\dot\ell(\theta)\big) = 0$, and the Fisher information $\mathcal{I}_\theta$ equals $-\mathbb{P}_\theta\big(\ddot\ell(\theta)\big) = \mathbb{P}_\theta\big(|\dot\ell(\theta)|^2\big)$. Setting $\mathcal{Z}_\theta^n := \sqrt{n}\widehat{\mathbb{P}}_n\big(\dot\ell(\theta)\big)$ from the central limit theorem §02.13 follows $\mathcal{Z}_\theta^n \xrightarrow{d} \mathrm{N}_{(0,\mathcal{I}_\theta)}$ under $\mathbb{P}_\theta^{\otimes n}$ while due to the law of large numbers §02.06 it holds $\widehat{\mathbb{P}}_n\ddot\ell(\theta) = -\mathcal{I}_\theta + \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$. If in addition the remainder term is negligible, i.e., $R_n = \mathrm{o}_{\mathbb{P}_\theta^{\otimes n}}(1)$, then the log-likelihood ratio permits an expansion

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

which in the limit equals the log-likelihood ratio in a Gaussian location model. □

§08.07 **Definition.** A sequence of statistical experiments $(\mathcal{X}_n, \mathscr{X}_n, \mathbb{P}_\Theta^n)_{n\in\mathbb{N}}$ with $\Theta \subseteq \mathbb{R}^k$ is called *local asymptotic normal (LAN)* in $\theta_o \in \mathrm{int}(\Theta)$, if there is a localising rate $(\delta_n)_{n\in\mathbb{N}}$ with $\delta_n = o(1)$, a sequence of statistics $(\mathcal{Z}^n_{\theta_o})_{n\in\mathbb{N}} \in (\mathscr{X}_n^k)_{n\in\mathbb{N}}$ and a matrix $\mathcal{I}_{\theta_o} \in \mathbb{R}^{(k,k)}$ such that for every $h \in \mathbb{R}^k$ the following three statements hold true:

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

(b) $\mathcal{Z}^n_{\theta_o} \xrightarrow{d} \mathrm{N}_{(0,\mathcal{I}_{\theta_o})}$ under $\mathbb{P}^n_{\theta_o}$, i.e., $\mathbb{P}^n_{\theta_o} \circ (\mathcal{Z}^n_{\theta_o})^{-1} \xrightarrow{d} \mathrm{N}_{(0,\mathcal{I}_{\theta_o})}$;

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

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

§08.08 **Comment.** If we assume in addition a strictly positive definite matrix $\mathcal{I}_{\theta_o} \in \mathbb{R}^{(k,k)}_>$ with inverse $\mathcal{I}_{\theta_o}^{-1}$ the sequence of statistics $(\widetilde{\mathcal{Z}}^n_{\theta_o} := \mathcal{I}_{\theta_o}^{-1}\mathcal{Z}^n_{\theta_o})_{n\in\mathbb{N}} \in (\mathscr{X}_n^k)_{n\in\mathbb{N}}$ is equally a central sequence satisfying $\widetilde{\mathcal{Z}}^n_{\theta_o} \xrightarrow{d} \mathrm{N}_{(0,\mathcal{I}^{-1}_{\theta_o})}$ under $\mathbb{P}^n_\theta$ and $\log(\mathrm{d}\mathbb{P}^n_{\theta_o+\delta_n h}/\mathrm{d}\mathbb{P}^n_{\theta_o}) = \langle \mathcal{I}_{\theta_o}h, \widetilde{\mathcal{Z}}^n_{\theta_o}\rangle - \frac{1}{2}\langle \mathcal{I}_{\theta_o}h, h\rangle + \mathrm{o}_{\mathbb{P}^n_{\theta_o}}(1)$. In other words the likelihood ratio $\mathrm{d}\mathbb{P}^n_{\theta_o+\delta_n h}/\mathrm{d}\mathbb{P}^n_{\theta_o}$ equals approximately the likelihood ratio $\mathrm{dN}_{(h,\mathcal{I}^{-1}_{\theta_o})}/\mathrm{dN}_{(0,\mathcal{I}^{-1}_{\theta_o})}$ as in the Reminder §08.02. Consequently, the localised statistical model $(\mathcal{X}_n, \mathscr{X}_n, \mathbb{P}^n_{\delta_n\Theta^n_o+\theta_o})$ is similar to a Gaussian location model $(\mathbb{R}^k, \mathscr{B}^{\otimes k}, \mathrm{N}_{\mathbb{R}^k\times\{\mathcal{I}^{-1}_{\theta_o}\}})$ in the sense of Definition §08.07. □

§08.09 **Definition.** A LAN sequence of statistical experiments is called *uniformly local asymptotic normal (ULAN)* in $\theta_o \in \Theta$, if the condition (c) in Definition §08.07 is replaced by

(c') for $h_n \to h$ it holds $\log(\mathrm{d}\mathbb{P}^n_{\theta_o+\delta_n h_n}/\mathrm{d}\mathbb{P}^n_{\theta_o}) = \langle \mathcal{Z}^n_{\theta_o}, h\rangle - \frac{1}{2}\langle \mathcal{I}_{\theta_o}h, h\rangle + \mathrm{o}_{\mathbb{P}^n_{\theta_o}}(1)$. □

§08.10 **Theorem.** *Let $(\mathcal{X}_n, \mathscr{X}_n, \mathbb{P}^n_\Theta)_{n\in\mathbb{N}}$ be LAN in $\theta_o \in \Theta \subseteq \mathbb{R}^k$ with localising rate $(\delta_n)_{n\in\mathbb{N}}$, central sequence $(\mathcal{Z}^n_{\theta_o})_{n\in\mathbb{N}}$ and Fisher information matrix $\mathcal{I}_{\theta_o} \in \mathbb{R}^{(k,k)}$. 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.,* $\mathbb{P}^n_{\theta_o+\delta_n h} \lhd\rhd \mathbb{P}^n_{\theta_o+\delta_n h'}$;

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

§08.11 **Proof** of Theorem §08.10. is given in the lecture. □

§08.12 **Theorem**. *Let* $\mathbb{P}_\Theta \ll \mu \in \mathcal{M}_\sigma(\mathcal{X})$ *with open* $\Theta \subseteq \mathbb{R}^k$ *be Hellinger-differentiable in* $\theta_o \in \Theta$ *with derivative* $\dot\ell_{\theta_o}$ *and Fisher information matrix* $\mathcal{I}_{\theta_o} = \mathbb{P}_{\theta_o}(\dot\ell_{\theta_o}\dot\ell^t_{\theta_o}) \in \mathbb{R}^{(k,k)}_{\geqslant}$. *Then the sequence of product experiments* $(\mathcal{X}^n, \mathscr{X}^{\otimes n}, \mathbb{P}^{\otimes n}_\Theta)$ *is ULAN in* θ_o *with localising rate* $\delta_n := n^{-1/2}$ *and central sequence* $\mathcal{Z}^n_{\theta_o} := \sqrt{n}\,\widehat{\mathbb{P}}_n(\dot\ell_{\theta_o})$, $n \in \mathbb{N}$, *that is,*

(i) $\sqrt{n}\,\widehat{\mathbb{P}}_n(\dot\ell_{\theta_o}) \xrightarrow{d} \mathrm{N}_{(0,\mathcal{I}_{\theta_o})}$ *under* $\mathbb{P}^{\otimes n}_{\theta_o}$ *and*

(ii) *for* $h_n \to h$ *it holds* $\log(\mathrm{d}\mathbb{P}^{\otimes n}_{\theta_o+h_n/\sqrt{n}}/\mathrm{d}\mathbb{P}^{\otimes n}_{\theta_o}) = \langle \mathcal{Z}^n_{\theta_o}, h\rangle - \frac{1}{2}\langle \mathcal{I}_{\theta_o}h, h\rangle + \mathrm{o}_{\mathbb{P}^{\otimes n}_{\theta_o}}(1)$.

§08.13 **Proof** of Theorem §08.12. is given in the lecture. □

§08.14 **Corollary**. *Under the assumptions of Theorem §08.12 consider for each* $n \in \mathbb{N}$ *a statistical product experiment* $(\mathcal{X}^n, \mathscr{X}^{\otimes n}, \mathbb{P}^{\otimes n}_\Theta)$ *and an estimator* $\widehat\gamma_n \in (\mathscr{X}^{\otimes n})^p$ *of a parameter of interest* $\gamma : \Theta \to \mathbb{R}^p$ *allowing an expansion* $\sqrt{n}(\widehat\gamma_n - \gamma(\theta_o)) = \sqrt{n}\,\widehat{\mathbb{P}}_n(\psi_{\theta_o}) + o_{\mathbb{P}^{\otimes n}_{\theta_o}}(1)$ *for some function* $\psi_{\theta_o} \in \mathscr{L}^p_2(\mathbb{P}_{\theta_o})$ *with* $\mathbb{P}_{\theta_o}(\psi_{\theta_o}) = 0$. *Then,* $\sqrt{n}(\widehat\gamma_n - \gamma(\theta_o)) \xrightarrow{d} \mathrm{N}_{(0,\Sigma_o)}$ *under* $\mathbb{P}^{\otimes n}_{\theta_o}$ *with* $\Sigma_o := \mathbb{P}_{\theta_o}(\psi_{\theta_o}\psi^t_{\theta_o})$ *and for each* $h \in \mathbb{R}^k$ *holds* $\sqrt{n}(\widehat\gamma_n - \gamma(\theta_o)) \xrightarrow{d} \mathrm{N}_{(\tau_h,\Sigma_o)}$ *under* $\mathbb{P}^{\otimes n}_{\theta_o+h/\sqrt{n}}$ *with* $\tau_h := \mathbb{P}_{\theta_o}(\psi_{\theta_o}\dot\ell^t_{\theta_o})h$.

§08.15 **Proof** of Corollary §08.14. is given in the lecture. □

§08.16 **Example** *(Example §06.06 continued).* Under the assumptions of Theorem §08.12 let $\gamma : \theta \to \mathbb{R}^p$ be a parameter of interest. Consider $\mathrm{m}(\gamma) \in \mathscr{L}_1(\mathbb{P}_\theta)$ for all $\gamma \in \mathbb{R}^p$, a criterion process $\widehat{\mathrm{M}}_n(\gamma) = \widehat{\mathbb{P}}_n\big(\mathrm{m}(\gamma)\big)$, a criterion function $\mathrm{M}(\theta,\gamma) = \mathbb{P}_\theta\big(\mathrm{m}(\gamma)\big)$ and a M-estimator $\widehat\gamma_n \in \arg\inf_{\gamma\in\Gamma}\{\widehat{\mathrm{M}}_n(\gamma)\}$ of $\{\gamma_o := \gamma(\theta_o)\} = \arg\inf_{\gamma\in\Gamma}\{\mathrm{M}(\theta_o,\gamma)\}$. Under regularity conditions as in Example §06.06 we have $\sqrt{n}(\widehat\gamma_n - \gamma_o) = \sqrt{n}\widehat{\mathbb{P}}_n\big(\psi_{\theta_o}\big) + \mathrm{o}_{\mathbb{P}^{\otimes n}_{\theta_o}}(1)$ with $\psi_{\theta_o} := -\ddot{\mathrm{M}}_o^{-1}\dot{\mathrm{m}}(\gamma_o)$ assuming a regular matrix $\ddot{\mathrm{M}}_o := \mathbb{P}_{\theta_o}\big(\ddot{\mathrm{m}}(\gamma_o)\big)$. Consequently, setting $\Sigma_o = \mathbb{P}_{\theta_o}(\psi_{\theta_o}\psi^t_{\theta_o}) = \ddot{\mathrm{M}}_o^{-1}\mathbb{P}_{\theta_o}(\dot{\mathrm{m}}(\gamma_o)\dot{\mathrm{m}}(\gamma_o)^t)\ddot{\mathrm{M}}_o^{-1}$ from Corollary §08.14 it follows

$$\sqrt{n}(\widehat\gamma_n - \gamma_o) \xrightarrow{d} \mathrm{N}_{(\tau_h,\Sigma_o)} \quad \text{under} \quad \mathbb{P}^{\otimes n}_{\theta_o+h/\sqrt{n}} \text{ with } \tau_h = -\ddot{\mathrm{M}}_o^{-1}\mathbb{P}_{\theta_o}(\dot{\mathrm{m}}(\gamma_o)\dot\ell^t_{\theta_o})h.$$

In the particular case of a MLE $\widehat\theta_n$ of θ, i.e., $(\gamma = \mathrm{id}_{\mathbb{R}^k})$, as in Example §06.07 setting $\mathrm{m} := -\log(\mathrm{d}\mathbb{P}_\theta/\mathrm{d}\mathbb{P}_o)$ we have $\dot{\mathrm{m}}(\theta_o) = -\dot\ell_{\theta_o}$, $\mathcal{I}_{\theta_o} = \mathbb{P}_{\theta_o}\big(\dot{\mathrm{m}}(\theta_o)\dot{\mathrm{m}}(\theta_o)^t\big) = \mathbb{P}_{\theta_o}\big(\ddot{\mathrm{m}}(\theta_o)\big) = \ddot{\mathrm{M}}_o$ and thus $\Sigma_o = \ddot{\mathrm{M}}_o^{-1}\mathbb{P}_{\theta_o}(\dot{\mathrm{m}}(\gamma_o)\dot{\mathrm{m}}(\gamma_o)^t)\ddot{\mathrm{M}}_o^{-1} = \mathcal{I}^{-1}_{\theta_o}$ and $\tau_h := -\ddot{\mathrm{M}}_o^{-1}\mathbb{P}_{\theta_o}(\dot{\mathrm{m}}(\theta_o)\dot\ell^t_{\theta_o})h = h$. Therewith, $\sqrt{n}(\widehat\theta_n - \theta_o) \xrightarrow{d} \mathrm{N}_{(h,\mathcal{I}^{-1}_{\theta_o})}$ under $\mathbb{P}^{\otimes n}_{\theta_o+h/\sqrt{n}}$. □

§08.17 **Remark**. Supposing $\sqrt{n}(\widehat\theta_n - \theta_o) = \sqrt{n}\,\widehat{\mathbb{P}}_n(\psi_{\theta_o}) + \mathrm{o}_{\mathbb{P}^{\otimes n}_{\theta_o}}(1)$ let us further assume a transformation $A : \Theta \to \mathbb{R}^p$ that is "smooth", and hence by employing the delta method §02.16, for instance satisfies $\sqrt{n}(A(\widehat\theta_n) - A(\theta_o)) = \dot A_{\theta_o}\sqrt{n}\,\widehat{\mathbb{P}}_n(\psi_{\theta_o}) + \mathrm{o}_{\mathbb{P}^{\otimes n}_{\theta_o}}(1)$. Consequently, it follows $\sqrt{n}(A(\widehat\theta_n) - A(\theta_o)) \xrightarrow{d} \mathrm{N}_{(\tau_h,\Sigma_o)}$ under $\mathbb{P}^{\otimes n}_{\theta_o+h/\sqrt{n}}$ with $\tau_h = \dot A_{\theta_o}\mathbb{P}_{\theta_o}(\psi_{\theta_o}\dot\ell^t_{\theta_o})h$ and $\Sigma_o = \dot A_{\theta_o}\mathbb{P}_{\theta_o}(\psi_{\theta_o}\psi^t_{\theta_o})\dot A^t_{\theta_o}$. In the special case of a MLE we have $\sqrt{n}(A(\widehat\theta_n) - A(\theta_o)) \xrightarrow{d} \mathrm{N}_{(\dot A_{\theta_o}h,\dot A_{\theta_o}\mathcal{I}^{-1}_{\theta_o}\dot A^t_{\theta_o})}$ under $\mathbb{P}^{\otimes n}_{\theta_o+h/\sqrt{n}}$. □

§09 Asymptotic relative efficiency

§09.01 **Heuristics** *(§06.09 and §06.10 continued).* Under the conditions of Corollary §08.14 consider the statistical testing task $H_0 : A(\theta_o) = 0$ against the alternative $H_1 : A(\theta_o) \neq 0$ for some transformation $A : \Theta \to \mathbb{R}^p$ satisfying $\sqrt{n}(A(\widehat{\theta}_n) - A(\theta_o)) = \dot{A}_{\theta_o}\sqrt{n}\,\widehat{\mathbb{P}}_n(\psi_{\theta_o}) + \mathrm{o}_{\mathbb{P}_{\theta_o}^{\otimes n}}(1)$. As in §06.09 let $\widehat{W}_n := nA(\widehat{\theta}_n)^t\widehat{\Sigma}_n^{-1}A(\widehat{\theta}_n)$ where $\widehat{\Sigma}_n = \Sigma + \mathrm{o}_{\mathbb{P}_{\theta_o}^{\otimes n}}(1)$ is a consistent estimator of $\Sigma = \dot{A}_{\theta_o}\mathbb{P}_{\theta_o}(\psi_{\theta_o}\psi_{\theta_o}^t)\dot{A}_{\theta_o}^t$, then a *Wald test* is given by $\varphi_n := \mathbb{1}_{\{\widehat{W}_n > \chi^2_{p,1-\alpha}\}}$. Thereby, under H_0, i.e. $A(\theta_o) = 0$, we have $\sqrt{n}A(\widehat{\theta}_n) = \dot{A}_{\theta_o}\sqrt{n}\,\widehat{\mathbb{P}}_n(\psi_{\theta_o}) + \mathrm{o}_{\mathbb{P}_{\theta_o}^{\otimes n}}(1)$ and $\widehat{W}_n \xrightarrow{d} \chi^2_p$ under $\mathbb{P}_{\theta_o}^{\otimes n}$ which in turn implies $\mathbb{P}_{\theta_o}^{\otimes n}(\varphi_n = 1) \xrightarrow{n\to\infty} \chi^2_p((\chi^2_{p,1-\alpha},\infty)) = \alpha$. In other words, the Wald test is asymptotically a level α test. For each $\theta \in \Theta$ let us denote $\beta_{\varphi_n}(\theta) := \mathbb{P}_\theta^{\otimes n}(\varphi_n) = \mathbb{P}_\theta^{\otimes n}(\varphi_n = 1) = \mathbb{P}_\theta^{\otimes n}(\widehat{W}_n > \chi^2_{p,1-\alpha})$ which equals the power of the Wald test φ_n under H_1, i.e. $\theta \in \Theta$ with $A(\theta) \neq 0$. In the sequel we consider local alternatives of the form $\theta = \theta_o + h/\sqrt{n}$ and thus we are interested in $\beta_{\varphi_n}(\theta_o + h/\sqrt{n}) = \mathbb{P}_{\theta_o+h/\sqrt{n}}^{\otimes n}(\widehat{W}_n > \chi^2_{p,1-\alpha})$. Keeping Remark §08.17 under $\mathbb{P}_{\theta_o+h/\sqrt{n}}^{\otimes n}$ we have $\sqrt{n}A(\widehat{\theta}_n) \xrightarrow{d} \mathrm{N}_{(\dot{A}_{\theta_o}\mathbb{P}_{\theta_o}(\psi_{\theta_o}\dot{\ell}_{\theta_o}^t)h,\Sigma)}$, assuming additionally $\Sigma > 0$ also $\Sigma^{-1/2}\sqrt{n}A(\widehat{\theta}_n) \xrightarrow{d} \mathrm{N}_{(a_h,\mathrm{Id}_p)}$ with $a_h := \Sigma^{-1/2}\dot{A}_{\theta_o}\mathbb{P}_{\theta_o}(\psi_{\theta_o}\dot{\ell}_{\theta_o})^t h$, and hence, $nA(\widehat{\theta}_n)^t\Sigma^{-1}A(\widehat{\theta}_n) \xrightarrow{d} \chi^2_p(\|a_h\|^2)$. Here $\chi^2_p(c)$ denotes a non-central χ^2-distribution with p degrees of freedom and non-centrality parameter $c \in \mathbb{R}^+$. Moreover, $\widehat{W}_n - nA(\widehat{\theta}_n)^t\Sigma^{-1}A(\widehat{\theta}_n) = \mathrm{o}_{\mathbb{P}_{\theta_o}^{\otimes n}}(1)$ and thus $\widehat{W}_n - nA(\widehat{\theta}_n)^t\Sigma^{-1}A(\widehat{\theta}_n) = \mathrm{o}_{\mathbb{P}_{\theta_o+h/\sqrt{n}}^{\otimes n}}(1)$ due to Lemma §07.19 (ii) by employing that $\mathbb{P}_{\theta_o}^{\otimes n} \lhd\rhd \mathbb{P}_{\theta_o+h/\sqrt{n}}^{\otimes n}$ are mutually contiguous. Consequently, $\widehat{W}_n \xrightarrow{d} \chi^2_p(\|a_h\|^2)$ under $\mathbb{P}_{\theta_o+h/\sqrt{n}}^{\otimes n}$ and thus $\beta_{\varphi_n}(\theta_o + h/\sqrt{n}) \xrightarrow{n\to\infty} \chi^2_p(\|a_h\|^2)\big((\chi^2_{p,1-\alpha},\infty)\big)$. Note that a_h simplifies to $h^t\dot{A}_{\theta_o}^t(\dot{A}_{\theta_o}\mathcal{I}_{\theta_o}^{-1}\dot{A}_{\theta_o}^t)^{-1}\dot{A}_{\theta_o}h$ in the particular case of a MLE $\widehat{\theta}_n$. □

§09.02 **Reminder** *(Gauß test).* In a Gaussian location model, i.e. $Y \bigcirc\!\!\!\!\sim \mathrm{N}_{\mathbb{R}^k\times\{\mathcal{I}_{\theta_o}^{-1}\}}$ with $\mathcal{I}_{\theta_o} \in \mathbb{R}_>^{(k,k)}$, consider the binary testing task $H_0 : \{\mathrm{N}_{(0,\mathcal{I}_{\theta_o}^{-1})}\}$ against the alternative $H_1 : \{\mathrm{N}_{(h,\mathcal{I}_{\theta_o}^{-1})}\}$ for some $h \in \mathbb{R}^k$. In this situation the log-likelihood ratio $\ell_h = \log(\mathrm{dN}_{(h,\mathcal{I}_{\theta_o}^{-1})}/\mathrm{dN}_{(0,\mathcal{I}_{\theta_o}^{-1})})$ satisfies $\ell_h(y) = \langle\mathcal{I}_{\theta_o}y, h\rangle - \frac{1}{2}\sigma_h^2$ for all $y \in \mathbb{R}^k$ with $\sigma_h^2 := \langle\mathcal{I}_{\theta_o}h, h\rangle$. Consequently, $\ell_h \sim \mathrm{N}_{(-\sigma_h^2/2,\sigma_h^2)}$ under $\mathrm{N}_{(0,\mathcal{I}_{\theta_o}^{-1})}$, i.e. under the hypothesis H_0, and $\ell_h \sim \mathrm{N}_{(\sigma_h^2/2,\sigma_h^2)}$ under $\mathrm{N}_{(h,\mathcal{I}_{\theta_o}^{-1})}$, i.e. under the alternative H_1. For $\alpha \in (0,1)$ let $c_{h,1-\alpha} \in \mathbb{R}$ satisfy $\mathrm{N}_{(-\sigma_h^2/2,\sigma_h^2)}\big((c_{h,1-\alpha},\infty)\big) = \alpha$ and thus $\mathrm{N}_{(0,\mathcal{I}_{\theta_o}^{-1})}(\ell_h > c_{h,1-\alpha}) = \mathrm{N}_{(-\sigma_h^2/2,\sigma_h^2)}\big((c_{h,1-\alpha},\infty)\big) = \alpha$. Keeping in mind that any most powerful level-α test has Neyman-Pearson form and the *Gauß test* $\varphi^\star := \mathbb{1}_{\{\ell_h > c_{h,1-\alpha}\}}$ is a Neyman-Pearson level-α test. Its power given by $\beta_{\varphi^\star}(h) := \mathrm{N}_{(h,\mathcal{I}_{\theta_o}^{-1})}(\varphi^\star = 1) = \mathrm{N}_{(h,\mathcal{I}_{\theta_o}^{-1})}(\ell_h > c_{h,1-\alpha}) = \mathrm{N}_{(\sigma_h^2/2,\sigma_h^2)}\big((c_{h,1-\alpha},\infty)\big)$ is maximal in the class of all level-α tests, i.e., for any level-α test φ holds $\beta_\varphi(h) \leqslant \beta_{\varphi^\star}(h)$. In other words, $\varphi^\star$ is a most powerful level-α test (Statistik 1, Satz §21.16, p.100). □

§09.03 **Example** *(Neyman-Pearson test).* Assume local asymptotic normality as in Definition §08.07 where $\ell_{h,n} := \log(\mathrm{d}\mathbb{P}_{\theta_o+\delta_n h}^n/\mathrm{d}\mathbb{P}_{\theta_o}^n) \xrightarrow{d} \mathrm{N}_{(-\sigma_h^2/2,\sigma_h^2)}$ under $\mathbb{P}_{\theta_o}^n$ with $\sigma_h^2 := \langle\mathcal{I}_{\theta_o}h, h\rangle$ for $h \in \mathbb{R}^k$. Hence by Le Cam's first lemma (Example §07.34)mutual contiguity $\mathbb{P}_{\theta_o+\delta_n h}^n \lhd\rhd \mathbb{P}_{\theta_o}^n$ and $\ell_{h,n} \xrightarrow{d} \mathrm{N}_{(\sigma_h^2/2,\sigma_h^2)}$ under $\mathbb{P}_{\theta_o+\delta_n h}^n$ hold. Consider the binary testing task of the hypothesis $H_0 : \{\mathbb{P}_{\theta_o}^n\}$ against a local alternative $H_1 : \{\mathbb{P}_{\theta_o+\delta_n h}^n\}$. In this situation $\varphi_n^\star = \mathbb{1}_{\{\ell_{h,n} > c_{h,n,1-\alpha}\}}$ is a *Neyman-Pearson test*, which is a most powerful level-α test, if $\mathbb{P}_{\theta_o}^n(\varphi_n^\star = 1) = \alpha$. Keeping its power

function $\beta_{\varphi_n^\star}(\theta) = \mathbb{P}_\theta^n(\varphi_n^\star) = \mathbb{P}_\theta^n(\varphi_n^\star = 1) = \mathbb{P}_\theta^n(\ell_{h,n} > c_{h,n,1-\alpha})$ evaluated at θ in mind the value $\beta_{\varphi_n^\star}(\theta_o + \delta_n h)$ equals the maximal size of the power in the class of all level-α tests. Considering $c_{h,1-\alpha} \in \mathbb{R}$ as in Reminder §09.02 under local asymptotic normality it follows $\alpha = \mathbb{P}_{\theta_o}^n(\varphi_n^\star) = \mathbb{P}_{\theta_o}^n(\ell_{h,n} > c_{h,n,1-\alpha}) \xrightarrow{n\to\infty} \mathrm{N}_{(-\sigma_h^2/2,\sigma_h^2)}\big((c_{h,1-\alpha},\infty)\big) = \alpha$ which implies $c_{h,n,1-\alpha} \xrightarrow{n\to\infty} c_{h,1-\alpha}$, and in addition $\beta_{\varphi_n^\star}(\theta_o + \delta_n h) = \mathbb{P}_{\theta_o+\delta_n h}^n(\varphi_n^\star) = \mathbb{P}_{\theta_o+\delta_n h}^n(\ell_{h,n} > c_{h,n,1-\alpha}) \xrightarrow{n\to\infty} \mathrm{N}_{(\sigma_h^2/2,\sigma_h^2)}\big((c_{h,1-\alpha},\infty)\big) = \beta_{\varphi^\star}(h)$ with Neyman-Pearson test $\varphi^\star$ in a Gaussian location model as in Reminder §09.02. □

§09.04 **Theorem**. *Let $\Theta \subseteq \mathbb{R}$. Consider a one-sided test task $H_0 : (-\infty, \theta_o]$ against $H_1 : (\theta_o, \infty)$. Suppose that $(\mathcal{X}_n, \mathscr{X}_n, \mathbb{P}_\Theta^n)$ is LAN in $\theta_o \in \Theta$ with localising sequence $(\delta_n)_{n\in\mathbb{N}}$, central sequence $(\mathcal{Z}_{\theta_o}^n)_{n\in\mathbb{N}} \in (\mathscr{X}_n)_{n\in\mathbb{N}}$ and strictly positive Fisher information $\mathcal{I}_{\theta_o} \in \mathbb{R}^+_{\setminus 0}$.*

(i) *Given a sequence $(T_n)_{n\in\mathbb{N}} \in (\mathscr{X}_n)_{n\in\mathbb{N}}$ of test statistics satisfying $(T_n, \mathcal{Z}_{\theta_o}^n) \xrightarrow{d} \mathrm{N}_{(0,M)}$ with $M = \big((\sigma^2,\rho)^t, (\rho,\mathcal{I}_{\theta_o})^t\big)$ consider the 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 $\beta_{\varphi_n}(\theta_o) = \mathbb{P}_{\theta_o}^n(\varphi_n) = \mathbb{P}_{\theta_o}^n(T_n > c_n) + \gamma_n \mathbb{P}_{\theta_o}^n(T_n = c_n) = \alpha_n \xrightarrow{n\to\infty} \alpha$. Choosing $z_{1-\alpha} \in \mathbb{R}$ with $1 - \mathbb{F}_{\mathrm{N}_{(0,1)}}(z_{1-\alpha}) := \mathrm{N}_{(0,1)}\big((z_{1-\alpha},\infty)\big) = \alpha$ we have*

$$\beta_{\varphi_n}(\theta_o + \delta_n h) = \mathbb{P}_{\theta_o+\delta_n h}^n(\varphi_n) \xrightarrow{n\to\infty} \mathbb{F}_{\mathrm{N}_{(0,1)}}(-z_{1-\alpha} + h\rho/\sigma).$$

(ii) *In case $T_n = \mathcal{Z}_{\theta_o}^n$ consider $\varphi_n^\star = \mathbb{1}_{\{\mathcal{Z}_{\theta_o}^n > z_{1-\alpha}\mathcal{I}_{\theta_o}^{1/2}\}}$, i.e. $\gamma_n = 0$ and $c_n = z_{1-\alpha}\mathcal{I}_{\theta_o}^{1/2}$. Then $\beta_{\varphi_n^\star}(\theta_o) = \mathbb{P}_{\theta_o}^n(\varphi_n^\star) = \mathbb{P}_{\theta_o}^n(\mathcal{I}_{\theta_o}^{-1/2}\mathcal{Z}_{\theta_o}^n > z_{1-\alpha}) \xrightarrow{n\to\infty} 1 - \mathbb{F}_{\mathrm{N}_{(0,1)}}(z_{1-\alpha}) = \alpha$ and*

$$\beta_{\varphi_n^\star}(\theta_o + \delta_n h) = \mathbb{P}_{\theta_o+\delta_n h}^n(\varphi_n^\star) \xrightarrow{n\to\infty} \mathbb{F}_{\mathrm{N}_{(0,1)}}(-z_{1-\alpha} + h\mathcal{I}_{\theta_o}^{1/2}).$$

§09.05 **Proof** of Theorem §09.04. is given in the lecture. □

§09.06 **Remark**.

(a) By using Theorem §07.35 directly it could be possible to calculate an asymptotic power of a test if $\log(\mathrm{d}\mathbb{P}_{\theta_o+\delta_n h}^n / \mathrm{d}\mathbb{P}_{\theta_o}^n) \xrightarrow{d} \mathbb{P}$ under $\mathbb{P}_{\theta_o}^n$ where $\mathbb{P}$ equals not necessarily $\mathrm{N}_{(0,1)}$.

(b) Let $(Y_1, Y_2) \sim \mathrm{N}_{(0,M)}$ with $M = \big((\sigma^2,\rho)^t, (\rho,\mathcal{I}_{\theta_o})^t\big)$ as in Theorem §09.04 (i), then $\rho^2 = |\operatorname{Cov}(Y_1, Y_2)|^2 \leqslant \mathbb{V}\mathrm{ar}(Y_1)\,\mathbb{V}\mathrm{ar}(Y_2) = \sigma^2\mathcal{I}_{\theta_o}$. Consequently, the test $\varphi_n^\star$ given in (ii) maximises the asymptotic power when considering only a randomised test φ_n as given in part (i) of Theorem §09.04. □

§09.07 **Theorem**. *Let the assumptions of Theorem §09.04 be satisfied. Any test φ_n of the one-sided testing task $H_0 : (\infty, \theta_o]$ against $H_1 : (\theta_o, \infty)$ with $\beta_{\varphi_n}(\theta_o) := \mathbb{P}_{\theta_o}^n(\varphi_n) = \alpha_n \xrightarrow{n\to\infty} \alpha$ fulfils*

(i) $\limsup_{n\to\infty} \beta_{\varphi_n}(\theta_o + \delta_n h) \leqslant \mathbb{F}_{\mathrm{N}_{(0,1)}}(-z_{1-\alpha} + h\sqrt{\mathcal{I}_{\theta_o}})$ *for all $h \in \mathbb{R}^+_{\setminus 0}$;*

(ii) $\liminf_{n\to\infty} \beta_{\varphi_n}(\theta_o - \delta_n h) \geqslant \mathbb{F}_{\mathrm{N}_{(0,1)}}(-z_{1-\alpha} - h\sqrt{\mathcal{I}_{\theta_o}})$ *for all $h \in \mathbb{R}^+_{\setminus 0}$.*

§09.08 **Proof** of Theorem §09.07. is given in the lecture. □

§09.09 **Remark**. Keeping Theorem §09.07 in mind we call the test (sequence) $(\varphi_n^\star)_{n\in\mathbb{N}}$ given in Theorem §09.04 (ii) asymptotically uniformly most powerful level-α test (sequence) in the class of all asymptotic level-α test (sequences). Its asymptotic power function equals $\mathbb{F}_{\mathrm{N}_{(0,1)}}(-z_{1-\alpha} + h\sqrt{\mathcal{I}_{\theta_o}})$ which is the power function of the uniformly most powerful test of $H_0 : (-\infty, 0]$ against $H_1 : (0,\infty)$ in the limit Gaussian location experiment $(\mathbb{R}, \mathscr{B}, \mathrm{N}_{\mathbb{R}\times\{\mathcal{I}_{\theta_o}^{-1}\}})$. □

§09.10 **Asymptotic relative efficiency**. Let $(\mathcal{X}_n, \mathscr{X}_n, \mathbb{P}_\Theta^n)_{n\in\mathbb{N}}$ be LAN with localising rate $(\delta_n := n^{-1/2})_{n\in\mathbb{N}}$. Consider a test φ_n^a satisfying the conditions of Theorem §09.04 (i) and hence, admitting an asymptotic power function such that $\beta_{\varphi_n^a}(\theta_o+h/\sqrt{n}) \xrightarrow{n\to\infty} \mathbb{F}_{\mathbb{N}_{(0,1)}}(-z_{1-\alpha}+h\rho_a/\sigma_a)$. Thereby, choosing $\eta = h/\sqrt{n}$ the approximation $\beta_{\varphi_n^a}(\theta_o+\eta) \approx \mathbb{F}_{\mathbb{N}_{(0,1)}}(-z_{1-\alpha}+\eta\sqrt{n}\rho_a/\sigma_a)$ is reasonable. In analogy, if φ_n^b is another test satisfying the conditions of Theorem §09.04 (i) and admitting $\beta_{\varphi_n^b}(\theta_o+\eta) \approx \mathbb{F}_{\mathbb{N}_{(0,1)}}(-z_{1-\alpha}+\eta\sqrt{n}\rho_b/\sigma_b)$. 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 $\mathfrak{are}(\varphi_{n_a}^a, \varphi_{n_b}^b) = (n_a/n_b) = (\rho_b^2\sigma_a^2)/(\rho_a^2\sigma_b^2)$ is called *asymptotic relative efficiency*. Meaning, that a sample of size $n_a = \mathfrak{are}(\varphi_{n_a}^a, \varphi_{n_b}^b)\, n_b$ is needed for the test $\varphi_{n_a}^a$ to attain at $\theta_o+\eta$ approximately the same power $\mathbb{F}_{\mathbb{N}_{(0,1)}}(-z_{1-\alpha}+\eta\sqrt{n_a}\rho_a/\sigma_a) = \mathbb{F}_{\mathbb{N}_{(0,1)}}(-z_{1-\alpha}+\eta\sqrt{n_b}\rho_b/\sigma_b)$ as the test $\varphi_{n_b}^b$ with sample size n_b. A comparison with the test $\varphi_n^\star$ as in Theorem §09.04 (ii) allows analogously to introduce a notion of *asymptotic absolute efficiency*. □

§10 Rank tests

§10.01 **Reminder**. Consider on the sample space $(\mathbb{R}^n, \mathscr{B}^n)$ the statistic $T : \mathbb{R}^n \to \mathbb{R}^n$ with $x \mapsto T(x) = (T_i(x))_{i\in[\![n]\!]}$ and $T_i(x) := \min\{c \in \mathbb{R} : \sum_{j\in[\![n]\!]} \mathbb{1}_{\{x_j\leqslant c\}} \geqslant i\}$, $i \in [\![n]\!]$. Since $T_1(x) \leqslant T_2(x) \leqslant \cdots \leqslant T_n(x)$ for all $x \in \mathbb{R}^n$ the statistic T (and any other statistic with this property) is called an *order statistic*. Denote by $\mathcal{S}_n$ the symmetric group of order n, i.e. the set of all permutations of the set $[\![n]\!]$. We identify as usual a vector $s = (s_i)_{i\in[\![n]\!]} \in [\![n]\!]^n$ with the map $s : [\![n]\!] \to [\![n]\!]$, $i \mapsto s(i) := s_i$, and hence $\mathcal{S}_n \subseteq [\![n]\!]^n$. Let $s^- \in \mathcal{S}_n$ denote the inverse permutation of $s \in \mathcal{S}_n$, i.e. $\mathrm{id}_{\mathcal{S}_n} = s\circ s^- = s^-\circ s$. For a permutation $s = (s_i)_{i\in[\![n]\!]} \in \mathcal{S}_n$ and a vector $x = (x_i)_{i\in[\![n]\!]} \in \mathbb{R}^n$ we write shortly $x_s := (x_{s_i})_{i\in[\![n]\!]}$. A Borel-measurable map $S : \mathbb{R}^n \to \mathcal{S}_n$, i.e. $S^{-1}(s) \in \mathscr{B}^n$ for all $s \in \mathcal{S}_n$, is called a *random permutation* on $(\mathbb{R}^n, \mathscr{B}^n)$. The associated map $S^- : \mathbb{R}^n \to \mathcal{S}_n$ satisfying $\mathrm{id}_{\mathcal{S}_n} = S^-(x)\circ S(x) = S(x)\circ S^-(x)$ for all $x \in \mathcal{X}$ is trivially again Borel-measurable, and hence called random inverse permutation of S. Moreover the statistic $X_S : \mathbb{R}^n \to \mathbb{R}^n$ with $x \mapsto X_S(x) := (x_{S_i(x)})_{i\in[\![n]\!]} = x_{S(x)} = \sum_{s\in\mathcal{S}_n} x_s \mathbb{1}_{\{s\}}(S(x))$ (a finite sum of Borel-measurable functions $x \mapsto x_s\mathbb{1}_{S^{-1}(s)}(x)$) is called a *random arrangement*. □

§10.02 **Definition**. A random permutation $O = (O_i)_{i\in[\![n]\!]}$ on $(\mathbb{R}^n, \mathscr{B}^n)$ is called *order permutation*, if the associated random arrangement $X_O : \mathbb{R}^n \to \mathbb{R}^n$ with $x \mapsto x_{O(x)}$ is an order statistic, i.e. $x_{O_1(x)} \leqslant x_{O_2(x)} \leqslant \cdots \leqslant x_{O_n(x)}$ for all $x \in \mathbb{R}^n$. A random permutation $R = (R_i)_{i\in[\![n]\!]}$ on $(\mathbb{R}^n, \mathscr{B}^n)$ is called *rank permutation*, if its random inverse permutation $O := R^-$ is an order permutation. For $i \in [\![n]\!]$ the i-th component $R_i(x)$ of $R(x)$ is called the *rank* of the i-th component of $x \in \mathbb{R}^n$. □

§10.03 **Comment**. An order permutation O is uniquely determined on the Borel-set $\{x_i \neq x_j\} := \{(x_i)_{i\in[\![n]\!]} \in \mathbb{R}^n : x_i \neq x_j, \forall\, j \in [\![n]\!]\backslash\{i\}, \forall\, i \in [\![n]\!]\}$ only. However, for $x \in \mathbb{R}^n$, the permutation $o := O(x) \in \mathcal{S}_n$ and $i \in [\![n]\!]$ the value at the i-th position in the ordered vector x_o equals the value at the o_i-th position in the original vector x. Conversely, for the permutation $r := R(x) \in \mathcal{S}_n$ of the rank permutation $R := O^-$ the value at the r_i-th position in the ordered vector x_o equals the value at the i-th position in the original vector x. □

§10.04 **Remark**. The map $R^* = (R_i^*)_{i\in[\![n]\!]} : \mathbb{R}^n \to \mathcal{S}_n$ with $x \mapsto R_i^*(x) := \sum_{j\in[\![i]\!]} \mathbb{1}_{\{x_i=x_j\}} + \sum_{j\in[\![n]\!]} \mathbb{1}_{\{x_i>x_j\}}$ for each $i \in [\![n]\!]$ is a rank permutation. Indeed, for each $x \in \mathbb{R}^n$ we have

$r := R^*(x) \in \mathcal{S}_n$ ($r : [\![n]\!] \to [\![n]\!]$ is injective and hence bijective) and its inverse permutation $o := r^-$ satisfies $x_{o_1} \leqslant x_{o_2} \leqslant \cdots \leqslant x_{o_n}$. Furthermore, each component of R^* is $\mathscr{B}$-$2^{[\![n]\!]}$-measurable, and hence R^* is a rank permutation. On the Borel-set $\{x_i \neq x_j\}$ each rank permutation $R = (R_i)_{i\in[\![n]\!]}$ is uniquely determined by $R_i(x) = \sum_{j\in[\![n]\!]} \mathbb{1}_{\{x_j \leqslant x_i\}} = R_i^*(x)$, $i \in [\![n]\!]$. For each $y \in \mathbb{R}$ define $\widehat{\mathbb{F}}_n(y) := \widehat{\mathbb{P}}_n(\mathbb{1}_{(-\infty,y]})$ with $\widehat{\mathbb{F}}_n(y,x) := \frac{1}{n}\sum_{j\in[\![n]\!]} \mathbb{1}_{\{x_j\leqslant y\}} \in [0,1]$ for all $x \in \mathbb{R}^n$. $\widehat{\mathbb{F}}_n$ is called *empirical cumulative distribution function*. If in addition $r := R(x)$ and $o := r^-$ for $x \in \{x_i \neq x_j\}$ then $i = n\widehat{\mathbb{F}}_n(x_{o_i}, x)$ and $r_i = n\widehat{\mathbb{F}}_n(x_i, x)$ for each $i \in [\![n]\!]$. □

§10.05 **Comment**. We assume a product probability measure $\mathbb{P}^n = \bigotimes_{j\in[\![n]\!]} \mathbb{P}_j$ on the sample space $(\mathbb{R}^n, \mathscr{B}^n)$ where for each $j \in [\![n]\!]$ the marginal probability measure $\mathbb{P}_j \in \mathcal{W}(\mathscr{B})$ admits a Lebesgue density $\mathfrak{f}_j = \mathrm{d}\mathbb{P}_j/\mathrm{d}\lambda$ and hence $\mathbb{P}^n \ll \lambda^n \in \mathcal{M}_\sigma(\mathscr{B}^n)$ with Lebesgue density $\mathrm{d}\mathbb{P}^n/\mathrm{d}\lambda^n = \prod_{j\in[\![n]\!]} \mathfrak{f}_j$. Noting that the complement $\{x_i = x_j\} := \{x_i \neq x_j\}^c$ of the Borel-set $\{x_i \neq x_j\}$ is a λ^n null set, and hence it is also a $\mathbb{P}^n$ null set. Thereby, each rank permutation R on $(\mathbb{R}^n, \mathscr{B}^n)$ with corresponding order permutation $O := R^-$ satisfies $x_{O_1(x)} < x_{O_2(x)} < \cdots < x_{O_n(x)}$ for $\mathbb{P}^n$-a.e. $x \in \mathbb{R}^n$. Moreover, for $\mathbb{P}^n$-a.e. $x \in \mathbb{R}^n$ the vector of ranks $R(x)$ (and the rang permutation R) is determined by $R_i(x) = \sum_{j\in[\![n]\!]} \mathbb{1}_{\{x_j\leqslant x_i\}} = n\widehat{\mathbb{F}}_n(x_i, x)$, $i \in [\![n]\!]$. □

§10.06 **Lemma**. *Consider a product probability measure $\mathbb{P}^{\otimes n}$ on $(\mathbb{R}^n, \mathscr{B}^n)$ with identical marginal distribution $\mathbb{P} \in \mathcal{W}(\mathscr{B})$, cumulative distribution function $\mathbb{F}(y) := \mathbb{P}(\mathbb{1}_{(-\infty,y]})$, $y \in \mathbb{R}$, and Lebesgue density $\mathfrak{f} = \mathrm{d}\mathbb{P}/\mathrm{d}\lambda$. Let R and X_O with $O = R^-$ be a rang permutation on $(\mathbb{R}^n, \mathscr{B}^n)$ and the corresponding order statistic, respectively.*

(i) *R is under $\mathbb{P}^{\otimes n}$ uniformly distributed on the symmetric group $\mathcal{S}_n$, precisely, $(\mathbb{P}^{\otimes n})^R(\{s\}) = (\mathbb{P}^{\otimes n} \circ R^{-1})(\{s\}) = \mathbb{P}^{\otimes n}(R = s) = \frac{1}{n!}$, $s \in \mathcal{S}_n$, in short $R \sim (\mathbb{P}^{\otimes n})^R = \mathrm{U}_{\mathcal{S}_n}$.*

(ii) *R and X_O are independent under $\mathbb{P}^{\otimes n}$.*

(iii) *The distribution of X_O admits under $\mathbb{P}^{\otimes n}$ a Lebesgue density $\mathfrak{f}^{X_O}(x) = n!\mathbb{1}_B(x)\prod_{i\in[\![n]\!]} \mathfrak{f}(x_i)$, $x \in \mathbb{R}^n$, with $B := \{(x_i)_{i\in[\![n]\!]} \in \mathbb{R}^n, x_1 < \ldots < x_n\}$.*

(iv) *For each $i \in [\![n]\!]$ the distribution of the i-th component of X_O admits under $\mathbb{P}^{\otimes n}$ a Lebesgue density $\mathfrak{f}_i(x) = i\binom{n}{i}|\mathbb{F}(x)|^{i-1}|1-\mathbb{F}(x)|^{n-i}\mathfrak{f}(x)$, $x \in \mathbb{R}$.*

§10.07 **Proof** of Lemma §10.06. see Statistik 1 (Lemma §24.05, p.115). □

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

§10.09 **Remark**. Roughly speaking, $\mathbb{P}_o \preceq \mathbb{P}$ says that realisations of $\mathbb{P}_o$ are typically smaller than realisations of $\mathbb{P}$. □

§10.10 **Example**. For $\sigma \in \mathbb{R}^+$ consider on $(\mathbb{R}, \mathscr{B})$ a Gaussian location family $\mathrm{N}_{\mathbb{R}\times\{\sigma^2\}}$. Then for all $a, b \in \mathbb{R}$ holds $\mathrm{N}_{(a,\sigma^2)} \prec \mathrm{N}_{(b,\sigma^2)}$ if and only if $a \lneqq b$. More generally, given a location family $\mathbb{P}_{\mathbb{R}}$ on $(\mathbb{R}, \mathscr{B})$ as introduced in Example §07.17 with likelihood function $\mathrm{L}(\theta, x) = g(x-\theta)$, $\theta, x \in \mathbb{R}$, for some strictly positive Lebesgue-density g on $\mathbb{R}$. Then for all $a, b \in \mathbb{R}$ holds $\mathbb{P}_a \prec \mathbb{P}_b$ if and only if $a \lneqq b$. □

§10.11 **Heuristics**. Given a sample from each distribution $\mathbb{P}_o, \mathbb{P} \in \mathcal{W}(\mathscr{B})$ we consider the testing task $H_o : \mathbb{P} = \mathbb{P}_o$ against the alternative $H_1 : \mathbb{P}_o \prec \mathbb{P}$. Loosely speaking, this means, that we

aim to reject the null hypothesis if realisations of $\mathbb{P}_o$ are *significantly* smaller than realisation of $\mathbb{P}$. More precisely, we assume a sample of $n = m + l$ independent real random variables $\{X_i, i \in [\![n]\!]\}$ where the first m and the last l have as common marginal distribution $\mathbb{P}_o$ and $\mathbb{P}$, respectively. In other words $X = (X_i)_{i\in[\![n]\!]}$ takes its values in the pooled sample space $(\mathbb{R}^n, \mathscr{B}^n)$. Considering a rank permutation R on $(\mathbb{R}^n, \mathscr{B}^n)$ and an observation $x \in \mathbb{R}^n$ it seems reasonable to reject the hypothesis if the sum of ranks within the first group of m random variables, i.e. $W_o(x) := \sum_{i\in[\![m]\!]} R_i(x)$, takes *sufficiently* smaller values then the sum of ranks within the second group of l random variables, i.e. $W(x) := \sum_{i\in[\![l]\!]} R_{i+m}(x)$ where obviously $W_o(x) + W(x) = \sum_{i\in[\![n]\!]} R_i(x) = \sum_{i\in[\![n]\!]} i = \frac{n(n+1)}{2}$. □

§10.12 **Lemma.** *For $m, l \in \mathbb{N}$ and $n := m + l$ let $R = (R_i)_{i\in[\![n]\!]}$ be a rang permutation on $(\mathbb{R}^n, \mathscr{B}^n)$, $W_o := \sum_{i\in[\![m]\!]} R_i$, $W := \sum_{i\in[\![l]\!]} R_{i+m}$ and $U_{ml} : \mathbb{R}^n \to [\![0, ml]\!]$ with $x \mapsto U_{ml}(x) := \sum_{i\in[\![m]\!]} \sum_{j\in[\![l]\!]} \mathbb{1}_{\{x_i > x_{j+m}\}}$. Then for each $x \in \{x_i \neq x_j\}$ it holds $W_o(x) = U_{ml}(x) + \frac{m(m+1)}{2}$ and consequently $W(x) = ml - U_{ml}(x) + \frac{l(l+1)}{2}$.*

§10.13 **Proof** of Lemma §10.12. see Statistik 1 (Lemma §24.11, p.116). □

§10.14 **Comment.** Keeping Lemma §10.12 in mind, we use the test statistic W_o or equivalently U_{ml} to reject the hypothesis $H_0 : \mathbb{P} = \mathbb{P}_o$ against the alternative $H_1 : \mathbb{P}_o \prec \mathbb{P}$, if $U_{ml} < c$ or equivalently $W_o < c + \frac{m(m+1)}{2}$ for a certain threshold $c \in (0, ml]$. 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 $\alpha \in (0, 1)$ which under the null hypothesis necessitates the knowledge of the distribution of U_{ml} or an asymptotic approximation. Interestingly the next proposition shows that under the null hypothesis the distribution of U_{ml} is *distribution free* in the following sense: If $\mathbb{P}_o = \mathbb{P}$ and $\mathbb{P}$ admits a Lebesgue density, then the distribution of U_{ml} is determined and it is independent of the underlying distribution $\mathbb{P}$. □

§10.15 **Proposition.** *For $m, l \in \mathbb{N}$ and $n := m + l$ let $\mathbb{P}^{\otimes n} \in \mathcal{W}(\mathscr{B}^n)$ with identical marginal distribution $\mathbb{P} \ll \lambda$. For all $k \in [\![0, ml]\!]$ it holds $\mathbb{P}^{\otimes n}(U_{ml} = k) = N(k; m, l)/\binom{n}{k}$ where $N(k; m, l)$ denotes the number of all partitions $\sum_{i\in[\![m]\!]} k_i = k$ of k in m increasingly ordered numbers $k_1 \leqslant k_2 \leqslant \cdots \leqslant k_m$ taking from the set $[\![0, l]\!]$. In particular, it holds $\mathbb{P}^{\otimes n}(U_{ml} = k) = \mathbb{P}^{\otimes n}(U_{ml} = ml - k)$.*

§10.16 **Proof** of Proposition §10.15. see Georgii [2015] (Satz 11.26, p.342). □

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

[1] The version based on W_o has been proposed by Wilcoxon [1945], while the U_{ml}-version has been independently be introduced by Mann and Whitney [1947].

§10.18 **Theorem.** *For $m, l \in \mathbb{N}$ and $n := m+l$ let $\mathbb{P}^{\otimes n} \in \mathcal{W}(\mathscr{B}^n)$ with identical marginal distribution $\mathbb{P} \ll \lambda$, and hence continuous cumulative distribution function $\mathbb{F}$. Consider $U_{ml} : \mathbb{R}^n \to [\![0, ml]\!]$ and $T_{ml} : \mathbb{R}^n \to \mathbb{R}$ with $x \mapsto U_{ml}(x) := \sum_{i\in[\![m]\!]} \sum_{j\in[\![l]\!]} \mathbb{1}_{\{x_i > x_{j+m}\}}$ and*

$$x \mapsto T_{ml}(x) := l \sum_{i\in[\![m]\!]} \mathbb{F}(x_i) - m \sum_{i\in[\![l]\!]} \mathbb{F}(x_{i+m}) = l \sum_{i\in[\![m]\!]} (\mathbb{F}(x_i) - 1/2) - m \sum_{i\in[\![l]\!]} (\mathbb{F}(x_{i+m}) - 1/2).$$

Define further $v_{ml} := ml(n+1)/12$, $T^\star_{ml} := T_{ml}/\sqrt{v_{ml}}$ and $U^\star_{ml} := (U_{ml} - ml/2)/\sqrt{v_{ml}}$. If in addition $m/n \to \gamma \in (0,1)$ then $U^\star_{ml} - T^\star_{ml} = \mathrm{o}_{\mathbb{P}^{\otimes n}}(1)$ and $T^\star_{ml} \xrightarrow{d} \mathrm{N}_{(0,1)}$ under $\mathbb{P}^{\otimes n}$, and thus $U^\star_{ml} \xrightarrow{d} \mathrm{N}_{(0,1)}$ under $\mathbb{P}^{\otimes n}$.

§10.19 **Proof** of Theorem §10.18. see Georgii [2015] (Satz 11.29, p.344). □

§10.20 **Remark.** Considering two independent samples $(X_i)_{i\in[\![m]\!]} \sim \mathbb{P}_o^{\otimes m}$ and $(X_{i+m})_{i\in[\![l]\!]} \sim \mathbb{P}^{\otimes l}$ set $n := m+l$ and $X := (X_i)_{i\in[\![n]\!]}$. Keeping Theorem §10.18 in mind we reject the null hypothesis $H_o : \mathbb{P}_o = \mathbb{P}$ against the alternative $H_1 : \mathbb{P}_o \prec \mathbb{P}$, if $U_{ml}(X) < ml/2 + z_\alpha\sqrt{v_{ml}}$ with $\mathbb{F}_{\mathrm{N}_{(0,1)}}(z_\alpha) = \alpha \in (0,1)$. This test is asymptotically a level-α test due to Theorem §10.18 by exploiting that under the null $\mathbb{P}^{\otimes n}(U_{ml} < ml/2 + z_\alpha\sqrt{v_{ml}}) \xrightarrow{n\to\infty} \mathbb{F}_{\mathrm{N}_{(0,1)}}(z_\alpha) = \alpha$ for $m/n \xrightarrow{n\to\infty} \gamma \in (0,1)$. Note that we reject similarly the null hypothesis $H_o : \mathbb{P}_o = \mathbb{P}$ against the alternative $H_1 : \mathbb{P} \prec \mathbb{P}_o$ if $U_{ml} > ml/2 + z_{1-\alpha}\sqrt{v_{ml}}$. Next we study the (asymptotic) size of the power of the rank test under local alternatives where we use that under the assumptions of Theorem §10.18 it holds

$$\begin{aligned} U^\star_{ml} = \frac{U_{ml} - ml/2}{\sqrt{v_{ml}}} &= \sqrt{\tfrac{l}{n+1}}\tfrac{1}{\sqrt{m}} \sum_{i\in[\![m]\!]} \tfrac{\mathbb{F}(X_i)-1/2}{\sqrt{1/12}} - \sqrt{\tfrac{m}{n+1}}\tfrac{1}{\sqrt{l}} \sum_{i\in[\![l]\!]} \tfrac{\mathbb{F}(X_{i+m})-1/2}{\sqrt{1/12}} + \mathrm{o}_{\mathbb{P}^{\otimes n}}(1) \\ &= \sqrt{1-\gamma}\,\sqrt{m}\,\widehat{\mathbb{P}}_m(g) - \sqrt{\gamma}\,\sqrt{l}\,\widehat{\mathbb{P}}_l(g) + \mathrm{o}_{\mathbb{P}^{\otimes n}}(1) \quad (10.1) \end{aligned}$$

setting $g := \sqrt{12}(\mathbb{F} - 1/2)$, $\widehat{\mathbb{P}}_m(g) := \frac{1}{m}\sum_{i\in[\![m]\!]} g(X_i)$ and $\widehat{\mathbb{P}}_l(g) := \frac{1}{l}\sum_{i\in[\![l]\!]} g(X_{i+m})$ where $\widehat{\mathbb{P}}_m(g)$ and $\widehat{\mathbb{P}}_l(g)$ are independent, $\mathbb{P}(g) = 0$, and $\mathbb{P}(g^2) = 1$ by construction. □

Chapter 4

Nonparametric estimation

This chapter presents an introduction to nonparametric estimation of curves along the lines of the textbooks by **Tsybakov** *[2009] and* **Comte** *[2015] where far more details, examples and further discussions can be found.*

§12 Introduction

Nonparametric density estimation. Consider for a non-empty set of parameters Θ a family $\mathbb{P}_\Theta$ of probability measures on $(\mathbb{R}, \mathscr{B})$ which contains the distribution of an observable real random variable, $X \leadsto \mathbb{P}_\Theta$. The family $\mathbb{P}_\Theta$ captures the prior knowledge about the distribution of the observation. For example, a family given by a set of parameters Θ containing only one singleton, i.e., $\Theta = \{\theta_o\}$, and hence $X \sim \mathbb{P}_{\theta_o}$ for some probability measure $\mathbb{P}_{\theta_o} \in \mathcal{W}(\mathscr{B})$, means, the data generating process is known to us in advance. On the contrary, a parameter set $\Theta = \mathcal{W}(\mathscr{B})$ reflects a lack of prior knowledge. A parametric model $\mathbb{P}_\Theta$ for some parameter set $\Theta \subseteq \mathbb{R}^k$ provides in a certain sense a trade-off between both extremes. In this chapter our aim is to avoid an assumption of a finite dimensional set of parameters. For example, consider $\{X_i, i \in [\![n]\!]\} \overset{i.i.d.}{\sim} \mathbb{P} \in \mathcal{W}(\mathscr{B})$, that is, an independent and identically distributed sample with common probability measure $\mathbb{P} \in \mathcal{W}(\mathscr{B})$. A reasonable estimator of the associated cumulative distribution function (c.d.f.) $\mathbb{F}(t) := \mathbb{P}((-\infty, t])$, $t \in \mathbb{R}$, is the empirical cumulative distribution function (e.c.d.f.) $\widehat{\mathbb{F}}_n(t) := \widehat{\mathbb{P}}_n((-\infty, t])$, $t \in \mathbb{R}$. For each $t \in \mathbb{R}$, $\widehat{\mathbb{F}}_n(t)$ is an unbiased estimator of $\mathbb{F}(t)$ with variance $\mathbb{V}\mathrm{ar}(\widehat{\mathbb{F}}_n(t)) = \frac{1}{n}\mathbb{F}(t)(1-\mathbb{F}(t))$. Consequently, $\widehat{\mathbb{F}}_n(t)$ converges in probability to $\mathbb{F}(t)$, and thus it is a consistent estimator. Moreover, by the law of large numbers (Property §02.05 (i)) the convergence holds almost surely in any point and also uniformly, by Glivenko-Cantelli's theorem, i.e., $\|\widehat{\mathbb{F}}_n - \mathbb{F}\|_{\mathscr{L}_\infty} = \mathrm{o}(1)$ $\mathbb{P}$-a.s.. If we assume in addition that $\mathbb{P}$ admits a Lebesgue density then $\widehat{\mathbb{F}}_n$ is a unbiased estimator with minimal variance, by Lehman-Scheffé's theorem. However, comparing different probability measures using their associated c.d.f.'s is visually difficult and as a consequence, other measures for dissimilarities are typically used. Consider, for instance, for two probability measures $\mathbb{P}$ and $\mathbb{P}_o$ on $(\mathbb{R}, \mathscr{B})$ their *total variation distance* given by $\|\mathbb{P} - \mathbb{P}_o\|_{\mathrm{TV}} := \sup\{|\mathbb{P}(B) - \mathbb{P}_o(B)|, B \in \mathscr{B}\}$. We note that for any probability measure $\mathbb{P} \in \mathcal{W}(\mathscr{B})$ admitting a Lebesgue-density we have $\|\mathbb{P} - \widehat{\mathbb{P}}_n\|_{TV} = 1$ $\mathbb{P}$-a.s. for any $n \in \mathbb{N}$. As a consequence the empirical probability measure $\widehat{\mathbb{P}}_n$ is not a consistent estimator of $\mathbb{P}$ in terms of the total variation distance. In other words, dependending on the measure of accuracy (metric, topology, etc.) a different estimator of $\mathbb{P}$ might be reasonable.

§12.01 **Lemma** *(Scheffé's theorem). Let $\mathbb{P}, \mathbb{P}_o \in \mathcal{W}(\mathscr{B})$ admit a μ-density $\mathbb{p}$ and $\mathbb{p}_o$, respectively, for some $\mu \in \mathcal{M}_\sigma(\mathscr{B})$. Then $\|\mathbb{P} - \mathbb{P}_o\|_{\mathrm{TV}} = \frac{1}{2}\mu(|\mathbb{p} - \mathbb{p}_o|) = \frac{1}{2}\|\mathbb{p} - \mathbb{p}_o\|_{\mathscr{L}_1(\mu)}$.*

§12.02 **Proof** of Lemma §12.01. see Tsybakov [2009] (Lemma 2.1, p.70). □

In the sequel let $\mathcal{D}$ be the set of Lebesgue densities on $(\mathbb{R}, \mathscr{B})$, and hence $\mathcal{D} \subseteq \mathscr{L}_1 =$

$\mathscr{L}_1(\mathscr{B},\lambda)$. $\mathbb{P}_{\mathfrak{p}} = \mathfrak{p}\lambda$ and $\mathbb{E}_{\mathfrak{p}}$ denote for each density $\mathfrak{p} \in \mathcal{D}$ the associated probability measure and expectation, respectively. We consider the statistical product experiment $(\mathbb{R}^n, \mathscr{B}^n, \mathbb{P}_{\mathcal{D}}^{\otimes n} = (\mathbb{P}_{\mathfrak{p}}^{\otimes n})_{\mathfrak{p}\in\mathcal{D}})$ and $(X_i)_{i\in[\![n]\!]} \bigcirc\!\!\!\sim \mathbb{P}_{\mathcal{D}}^{\otimes n}$. Typically, for $s \geqslant 1$ we access the accuracy of an estimator $\widehat{\mathfrak{p}}$ of $\mathfrak{p}$ either by a local measure, e.g. $\mathbb{P}_{\mathfrak{p}}^{\otimes n}(|\widehat{\mathfrak{p}}(t) - \mathfrak{p}(t)|^s)$, for $t \in \mathbb{R}$, or by a global measure, e.g. $\mathbb{P}_{\mathfrak{p}}^{\otimes n}(\|\widehat{\mathfrak{p}} - \mathfrak{p}\|_{\mathscr{L}_s}^s) = \mathbb{P}_{\mathfrak{p}}^{\otimes n}(\lambda(|\widehat{\mathfrak{p}} - \mathfrak{p}|^s))$, with a focus on the special cases $s = 1$ and $s = 2$.

Nonparametric regression. We describe the dependence of the variation of a real-valued random variable Y (response) on the variation of an explanatory random variable X by a functional relationship $\mathbb{E}\big(Y\big|X = x\big) = f(x)$ where f is an unknown functional parameter of interest. For a detailed discussion of the case of a deterministic explanatory variable we refer to Tsybakov [2009]. Here and subsequently, we restrict our attention to the special case of a real-valued explanatory variable X, and hence, a random vector (Y, X) taking values in $(\mathbb{R}^2, \mathscr{B}^2)$. The joint distribution of (Y, X) is uniquely determined by the functional parameter of interest f, the conditional distribution of the error $\varepsilon := Y - f(X)$ given X and the marginal distribution of X which are generally all not known in advance. However, the joint distribution is typically parametrised by the regression function f only and we write shortly $(Y, X) \sim \mathbb{P}_f$. Thereby, the dependence on the marginal distribution $\mathbb{P}_X$ of the regressor X and the conditional distribution of the error term ε given X is usually not made explicit. For sake of simplicity, suppose in addition that the joint distribution $\mathbb{P}_f$ of (Y, X) admits a joint Lebesgue density $\mathfrak{p}$. Denoting by $\mathfrak{p}^X$ the marginal density of X we use for the conditional density $\mathfrak{p}_{Y|X}$ of Y given X the $\mathbb{P}_f$-a.s. identity $\mathfrak{p}^X\mathfrak{p}_{Y|X} = \mathfrak{p}$ (see Notation §03.11 (iii)) which allows for $\mathbb{P}_f$-a.e. $x \in \mathbb{R}$ to write

$$\begin{aligned}\mathfrak{q}(x) := f(x)\mathfrak{p}^X(x) = \mathbb{E}\big(Y\big|X = x\big)\mathfrak{p}^X(x)\\ = \int_{\mathbb{R}} y\mathfrak{p}_{Y|X=x}(y)dy\mathfrak{p}^X(x) = \int_{\mathbb{R}} y\mathfrak{p}(y,x)dy. \quad (12.1)\end{aligned}$$

Consequently, the function of interest is $\mathbb{P}_f$-a.s. given by $f = \mathfrak{q}/\mathfrak{p}^X$ which motivates the following estimation strategy. Given a sample of (Y, X) estimate separately $\mathfrak{q}$ and $\mathfrak{p}^X$, say by $\widehat{\mathfrak{q}}$ and $\widehat{\mathfrak{p}}^X$, and then form a estimator $\widehat{f} = \widehat{\mathfrak{q}}/\widehat{\mathfrak{p}}^X$ (possibly in addition to be regularised). 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 let $\mathcal{F}$ be a family of regression functions and for each $f \in \mathcal{F}$ denote by $\mathbb{P}_f$ and $\mathbb{E}_f$ the associated probability measure of (Y, X) and its expectation, respectively. We denote by $\mathbb{P}_{\mathcal{F}}$ the family of possible distributions of (Y, X), but keep in mind, that the distribution $\mathbb{P}_f$ of (Y, X) is generally not uniquely determined by $f \in \mathcal{F}$ only. If $\{(Y_i, X_i), i \in [\![n]\!]\}$ form an independent and identically distributed (i.i.d.) sample of $(Y, X) \sim \mathbb{P}_f$ then $\mathbb{P}_f^{\otimes n} = \otimes_{j\in[\![n]\!]}\mathbb{P}_f$ denotes the joint product probability measure of the family $((Y_i, X_i))_{i\in[\![n]\!]}$. We write $\{(Y_i, X_i), i \in [\![n]\!]\} \overset{i.i.d.}{\sim} \mathbb{P}_f$ or $((Y_i, X_i))_{i\in[\![n]\!]} \sim \mathbb{P}_f^{\otimes n}$ for short. We denote by $\mathbb{P}_{\mathcal{F}}^{\otimes n} := (\mathbb{P}_f^{\otimes n})_{f\in\mathcal{F}}$ the corresponding family of product probability measures. For $s \geqslant 1$ we access also the accuracy of an estimator $\widehat{f}$ of f either by a local measure, e.g. $\mathbb{P}_f^{\otimes n}(|\widehat{f}(t) - f(t)|^s)$, for $t \in \mathbb{R}$, or by a global measure, e.g. $\mathbb{P}_f^{\otimes n}(\|\widehat{f} - f\|_{\mathscr{L}_s}^s) = \mathbb{P}_f^{\otimes n}(\lambda(|\widehat{f} - f|^s))$ with a focus on the special cases $s = 1$ and $s = 2$.

§13 Kernel density estimation

Throughout this section we consider the statistical product model $(\mathbb{R}^n, \mathscr{B}^n, \mathbb{P}_{\mathcal{D}}^{\otimes n} = (\mathbb{P}_{\mathfrak{p}}^{\otimes n})_{\mathfrak{p}\in\mathcal{D}})$

and let $\{X_i, i \in [\![n]\!]\} \overset{i.i.d.}{\sim} \mathbb{P}_{\mathfrak{p}} = \mathfrak{p}\lambda \in \mathcal{W}(\mathscr{B})$ be real-valued random variables with Lebesgue density $\mathfrak{p} \in \mathcal{D} \subseteq \mathscr{L}_1 = \mathscr{L}_1(\mathscr{B}, \lambda)$ and c.d.f. $\mathbb{F}$.

§13.01 **Definition**. A function $K \in \mathscr{L}_1$ with $\lambda(K) = 1$ is called a *kernel*. Given a *bandwidth* $\mathrm{b} \in \mathbb{R}^+_{\backslash 0}$ and an evaluation point $x_o \in \mathbb{R}$ define $K_{\mathrm{b}}(x_o) \in \mathscr{L}_1$ with $x \mapsto K_{\mathrm{b}}(x_o, x) := \frac{1}{\mathrm{b}} K\big(\frac{x - x_o}{\mathrm{b}}\big)$. The statistic $\widehat{\mathfrak{p}}_{\mathrm{b}}(x_o) := \widehat{\mathbb{P}}_n K_{\mathrm{b}}(x_o) \in \mathscr{B}^n$ satisfying

$$x^n = (x_i)_{i \in [\![n]\!]} \mapsto \widehat{\mathfrak{p}}_{\mathrm{b}}(x_o, x^n) = \frac{1}{n} \sum_{i \in [\![n]\!]} K_{\mathrm{b}}(x_o, x_i) = \frac{1}{n\mathrm{b}} \sum_{i \in [\![n]\!]} K\big(\frac{x_i - x_o}{\mathrm{b}}\big)$$

is called *kernel density estimator* of $\mathfrak{p}(x_o)$. □

§13.02 **Remark**. Since $\mathbb{F}(x + \mathrm{b}) - \mathbb{F}(x - \mathrm{b}) = \mathfrak{p}\lambda(]x - \mathrm{b}, x + \mathrm{b}]) \text{ for any } \mathrm{b} \in \mathbb{R}^+_{\backslash 0}$ we have $\mathbb{F}(x + \mathrm{b}) - \mathbb{F}(x - \mathrm{b}) \approx \mathfrak{p}(x) 2\mathrm{b}$ for b sufficiently small. Replacing the unknown $\mathbb{F}$ by its empirical counter part $\widehat{\mathbb{F}}_n$ Rosenblatt [1956] proposed for $\mathfrak{p}(x)$ the estimator $\widehat{\mathfrak{p}}_{\mathrm{b}}(x) \in \mathscr{B}^n$ given by

$$\begin{aligned} x^n = (x_i)_{i \in [\![n]\!]} \mapsto \widehat{\mathfrak{p}}_{\mathrm{b}}(x, x^n) &:= \frac{\widehat{\mathbb{F}}_n(x + \mathrm{b}, x^n) - \widehat{\mathbb{F}}_n(x - \mathrm{b}, x^n)}{2\mathrm{b}} \\ &= \frac{1}{n} \sum_{i \in [\![n]\!]} \frac{1}{2\mathrm{b}} \mathbb{1}_{(-1,1]}\big(\tfrac{x_i - x}{\mathrm{b}}\big) = \frac{1}{n} \sum_{i \in [\![n]\!]} \frac{1}{\mathrm{b}} K\big(\tfrac{x_i - x}{\mathrm{b}}\big) = \frac{1}{n} \sum_{i \in [\![n]\!]} K_{\mathrm{b}}(x, x_i) \end{aligned}$$

setting $K(t) := \frac{1}{2}\mathbb{1}_{]-1,1]}(t)$ for $t \in \mathbb{R}$. Observe that K is a density, which in turn implies that $x \mapsto \widehat{\mathfrak{p}}_{\mathrm{b}}(x, x^n)$ is a density for each $h \in \mathbb{R}^+_{\backslash 0}$ and $x^n \in \mathbb{R}^n$ as well. Parzen [1962] introduces a kernel K and a bandwidth b as in Definition §13.01 and studies the more general kernel density estimator $\widehat{\mathfrak{p}}_{\mathrm{b}}(x) = \widehat{\mathbb{P}}_n K_{\mathrm{b}}(x)$, $x \in \mathbb{R}$. Note that $\lambda(\widehat{\mathfrak{p}}_{\mathrm{b}}) = 1$ since $\lambda(K) = 1$ by assumption. If the kernel K is in addition positive, i.e. $K \in \mathscr{B}^+$, then $\widehat{\mathfrak{p}}_{\mathrm{b}}$ is a density. An alternative motivation for a kernel density estimator provides the following lemma. □

§13.03 **Lemma** *(Bochner's lemma). For* $\mathrm{b} \in \mathbb{R}^+_{\backslash 0}$, $x_o \in \mathbb{R}$ *and* $Q \in \mathscr{L}_1$ *define* $Q_{\mathrm{b}}(x_o) \in \mathscr{L}_1$ *with* $Q_{\mathrm{b}}(x_o, x) := \frac{1}{\mathrm{b}} Q(\frac{x - x_o}{\mathrm{b}})$, $x \in \mathbb{R}$. *If* $g \in \mathscr{B}$ *is bounded, i.e.,* $\|g\|_{\mathscr{L}_\infty} < \infty$, *and continuous in* x_o, *then* $\lim_{\mathrm{b} \to 0} \lambda\big(g Q_{\mathrm{b}}(x_o)\big) = g(x_o)\lambda(Q)$.

§13.04 **Proof** of Lemma §13.03. is given in the lecture. □

§13.05 **Example**. Kernels typically considered are the rectangular kernel $K(u) := \frac{1}{2}\mathbb{1}_{[-1,1]}(u)$, the triangular kernel $K(u) := (1 - |u|)\mathbb{1}_{[-1,1]}(u)$, the Epanechnikov kernel $K(u) := \frac{3}{4}(1 - u^2)\mathbb{1}_{[-1,1]}(u)$ or the Gaussian kernel $K(u) := \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)$. □

Local measure of accuracy.

§13.06 **Definition**. The *mean squared error* of a kernel density estimator $\widehat{\mathfrak{p}}_h(x_o)$ satisfies

$$\mathbb{mse}(x_o) := \mathbb{E}_{\mathfrak{p}}^{\otimes n} |\widehat{\mathfrak{p}}_{\mathrm{b}}(x_o) - \mathfrak{p}(x_o)|^2 = \mathbb{var}(x_o) + |\mathbb{bias}(x_o)|^2 \quad \text{at a point } x_o \in \mathbb{R}$$

by introducing a *variance* and a *bias* term, respectively,

$$\mathbb{var}(x_o) := \mathbb{E}_{\mathfrak{p}}^{\otimes n} |\widehat{\mathfrak{p}}_{\mathrm{b}}(x_o) - \mathbb{E}_{\mathfrak{p}}^{\otimes n} \widehat{\mathfrak{p}}_{\mathrm{b}}(x_o)|^2 \quad \text{and} \quad \mathbb{bias}(x_o) := \mathbb{E}_{\mathfrak{p}}^{\otimes n} \widehat{\mathfrak{p}}_{\mathrm{b}}(x_o) - \mathfrak{p}(x_o).$$ □

In the sequel we analyse separately the variance and bias term.

§13.07 **Lemma.** *Let $\mathbb{p} \in \mathscr{L}_\infty$ and $K \in \mathscr{L}_1 \cap \mathscr{L}_2$ with $\lambda(K) = 1$. For each $x_o \in \mathbb{R}$, $h \in \mathbb{R}^+_{\setminus 0}$ and $n \in \mathbb{N}$ we have $\mathbb{var}(x_o) \leqslant (n\mathrm{b})^{-1}\|\mathbb{p}\|_{\mathscr{L}_\infty}\|K\|^2_{\mathscr{L}_2}$.*

§13.08 **Proof** of Lemma §13.07. is given in the lecture. □

§13.09 **Remark.** Let $\mathbb{p} \in \mathscr{L}_\infty$ be continuous, and suppose that $K \in \mathscr{L}_1 \cap \mathscr{L}_2$ satisfies $\lambda(K) = 1$. By Lemma §13.07 $\mathbb{var}(x_o) \leqslant (n\mathrm{b})^{-1}\|\mathbb{p}\|_{\mathscr{L}_\infty}\|K\|^2_{\mathscr{L}_2}$. On the other hand, since $\mathbb{bias}(x_o) = \lambda(\mathbb{p}K_\mathrm{b}(x_o)) - \mathbb{p}(x_o)$ from Bochner's Lemma §13.03 follows $|\mathbb{bias}(x_o)| = \mathrm{o}(1)$ as $\mathrm{b} \to 0$. By combining both results, we obtain for any sequence $(\mathrm{b}_n)_{n\in\mathbb{N}}$ of bandwidths satisfying $1 = \mathrm{o}(n\mathrm{b}_n)$, i.e. $n\mathrm{b}_n \to \infty$, and $\mathrm{b}_n = \mathrm{o}(1)$ that $\mathbb{mse}(x_o) = \mathrm{o}(1)$ as $n \to \infty$. As a consequence, 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 $1 = \mathrm{o}(n\mathrm{b})$ or $\mathrm{b} = \mathrm{o}(1)$ as $n \to \infty$. □.

§13.10 **Lemma.** *Let $\mathbb{p}$ be twice-differentiable with second derivative $\ddot{\mathbb{p}} \in \mathscr{L}_\infty$. If $K, \mathrm{id}^2_\mathbb{R}K \in \mathscr{L}_1$, $\lambda(K) = 1$ and $\lambda(\mathrm{id}_\mathbb{R}K) = 0$, then $|\mathbb{bias}(x_o)| \leqslant \mathrm{b}^2\frac{1}{2}\|\ddot{\mathbb{p}}\|_{\mathscr{L}_\infty}\lambda(\mathrm{id}^2_\mathbb{R}\,|K|)$ for all $x_o \in \mathbb{R}$, $\mathrm{b} \in \mathbb{R}^+_{\setminus 0}$.*

§13.11 **Proof** of Lemma §13.10. is given in the lecture. □

§13.12 **Remark.** Let $\mathbb{p} \in \mathscr{L}_\infty$ be twice-differentiable with second derivative $\ddot{\mathbb{p}} \in \mathscr{L}_\infty$ and suppose that $K \in \mathscr{L}_1 \cap \mathscr{L}_2$ satisfies $\mathrm{id}^2_\mathbb{R}K \in \mathscr{L}_1$, $\lambda K = 1$ and $\lambda(\mathrm{id}_\mathbb{R}K) = 0$. By combination of Lemmata §13.07 and §13.10 follows for all $\mathrm{b} \in \mathbb{R}^+_{\setminus 0}$, $n \in \mathbb{N}$ and uniformly for all $x_o \in \mathbb{R}$

$$\mathbb{mse}(x_o) \leqslant (n\mathrm{b})^{-1}\|\mathbb{p}\|_{\mathscr{L}_\infty}\|K\|^2_{\mathscr{L}_2} + \mathrm{b}^4\tfrac{1}{4}\|\ddot{\mathbb{p}}\|^2_{\mathscr{L}_\infty}(\lambda(\mathrm{id}^2_\mathbb{R}|K|))^2.$$

The first and second term on the right hand side is increasing and decreasing, respectively, as h tends to zero. Therefore, let us minimise the right hand side as a function of b. Keep in mind that $M(\mathrm{b}) := a(n\mathrm{b})^{-1} + c\mathrm{b}^{2\beta}$, $\mathrm{b} \in \mathbb{R}^+_{\setminus 0}$, attains its minimum $M(\mathrm{b}_o) = b\big(\frac{a}{2\beta c}\big)^{1/(2\beta+1)}n^{-2\beta/(2\beta+1)}$ at $\mathrm{b}_o = \big(\frac{a}{2\beta c}\big)^{1/(2\beta+1)}n^{-1/(2\beta+1)}$. Thus, choosing $\mathrm{b}_o = \big(\frac{\|\mathbb{p}\|_{\mathscr{L}_\infty}\|K\|^2_{\mathscr{L}_2}}{\|\ddot{\mathbb{p}}\|^2_{\mathscr{L}_\infty}(\lambda(\mathrm{id}^2_\mathbb{R}\,|K|))^2}\big)^{1/5}n^{-1/5}$ we obtain

$$\sup_{x_o\in\mathbb{R}} \mathbb{mse}(x_o) \leqslant \tfrac{1}{4}\big(\|\ddot{\mathbb{p}}\|^2_{\mathscr{L}_\infty}(\lambda(\mathrm{id}^2_\mathbb{R}\,|K|))^2\big)^{4/5}\big(\|\mathbb{p}\|_{\mathscr{L}_\infty}\|K\|^2_{\mathscr{L}_2}\big)^{1/5}\, n^{-4/5}.$$

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

§13.13 **Lemma.** *Let $\mathbb{p} \in \mathscr{L}_\infty$ be continuous in x_o and $K \in \mathscr{L}_1 \cap \mathscr{L}_2 \cap \mathscr{L}_\infty$ with $\lambda K = 1$. If $1 = \mathrm{o}(n\mathrm{b})$ and $\mathrm{b} = \mathrm{o}(1)$ as $n \to \infty$, then $\sqrt{n\mathrm{b}}(\widehat{\mathbb{p}}_\mathrm{b}(x_o) - \mathbb{E}^{\otimes n}_\mathbb{p}\widehat{\mathbb{p}}_\mathrm{b}(x_o)) \xrightarrow{d} \mathrm{N}_{(0,\sigma^2)}$ with $\sigma^2 := \mathbb{p}(x_o)\lambda(K^2)$.*

§13.14 **Proof** of Lemma §13.13. is given in the lecture. □

§13.15 **Remark.** In addition to the assumptions of Lemma §13.13 let $\mathbb{p}$ be twice-differentiable with second derivative $\ddot{\mathbb{p}} \in \mathscr{L}_\infty$ continuous in x_o, and let $\mathrm{id}^2_\mathbb{R}K \in \mathscr{L}_1$ with $\lambda(\mathrm{id}_\mathbb{R}K) = 0$. Then, $\mathrm{b}^{-2}\mathbb{bias}(x_o) = \frac{1}{2}\ddot{\mathbb{p}}(x_o)\lambda(\mathrm{id}^2_\mathbb{R}K) + \mathrm{o}(1)$ as $\mathrm{b} \to 0$ by Bochner's Lemma §13.03. Therefore, setting $\mu := \frac{c^{5/2}}{2}\ddot{\mathbb{p}}(x_o)\lambda(\mathrm{id}^2_\mathbb{R}K)$ we have $\sqrt{n\mathrm{b}}\mathbb{bias}(x_o) = \mu + \mathrm{o}(1)$ as $\mathrm{b}n^{1/5} \to c \in \mathbb{R}^+_{\setminus 0}$, and thus $\sqrt{n\mathrm{b}}(\widehat{\mathbb{p}}_\mathrm{b}(x_o) - \mathbb{p}(x_o)) \xrightarrow{d} \mathrm{N}_{(\mu,\sigma^2)}$ due Lemma §13.13. Moreover, we conclude similarly $\sqrt{n\mathrm{b}}(\widehat{\mathbb{p}}_\mathrm{b}(x_o) - \mathbb{p}(x_o)) \xrightarrow{d} \mathrm{N}_{(0,\sigma^2)}$, if $\mathrm{b}n^{1/5} = \mathrm{o}(1)$. □

§13.16 **Definition.** A kernel K satisfying in addition $\mathrm{id}^j_\mathbb{R}K \in \mathscr{L}_1$ and $\lambda(\mathrm{id}^j_\mathbb{R}K) = 0$ for each $j \in [\![l]\!]$ is called a *kernel of order $l \in \mathbb{N}$*. □

§13.17 **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. □

§13.18 **Notation.** We denote by $\lfloor \beta \rfloor$ the greatest integer strictly less than the real number β. □

§13.19 **Definition.** For $\beta, L \in \mathbb{R}_{\setminus 0}^{+}$ 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 $f^{(l)}$ satisfies $|f^{(l)}(x) - f^{(l)}(y)| \leqslant L|x-y|^{\beta - l}$ for all $x, y \in \mathbb{R}$. □

§13.20 **Lemma.** *Suppose that $\mathbb{p} \in \mathcal{H}(\beta, L)$ and let K be a kernel of order $l = \lfloor \beta \rfloor$ satisfying $|\mathrm{id}_{\mathbb{R}}|^{\beta} K \in \mathscr{L}_1$. Then, $|\mathbb{bias}(x_o)| \leqslant \mathrm{b}^{\beta} \frac{L}{l!} \lambda(|\mathrm{id}_{\mathbb{R}}|^{\beta} |K|)$ for each $x_o \in \mathbb{R}$, $\mathrm{b} \in \mathbb{R}_{\setminus 0}^{+}$ and $n \in \mathbb{N}$.*

§13.21 **Proof** of Lemma §13.20. is given in the lecture. □

§13.22 **Remark.** Let $\mathbb{p} \in \mathcal{H}(\beta, L)$ and suppose that $K \in \mathscr{L}_2$ is a kernel of order $l = \lfloor \beta \rfloor$ satisfying in addition $|\mathrm{id}_{\mathbb{R}}|^{\beta} K \in \mathscr{L}_1$. By combination of Lemmata §13.07 and §13.20 we conclude that uniformly for all $x_o \in \mathbb{R}$

$$\mathbb{mse}(x_o) \leqslant (n\mathrm{b})^{-1} \|\mathbb{p}\|_{\mathscr{L}_\infty} \lambda(|K|^2) + \mathrm{b}^{2\beta} \big(\tfrac{L}{l!} \lambda(\mathrm{id}_{\mathbb{R}}|^{\beta} |K|) \big)^2 .$$

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

§13.23 **Theorem.** *Suppose that $\mathbb{p} \in \mathcal{H}(\beta, L)$ and let $K \in \mathscr{L}_2$ be a kernel of order $l = \lfloor \beta \rfloor$ satisfying in addition $|\mathrm{id}_{\mathbb{R}}|^{\beta} K \in \mathscr{L}_1$. Fix $c \in \mathbb{R}_{\setminus 0}^{+}$ and set $\mathrm{b}_o := c n^{-1/(2\beta+1)}$ then for all $n \in \mathbb{N}$*

$$\sup_{x_o \in \mathbb{R}} \sup_{\mathbb{p} \in \mathcal{H}(\beta, L) \cap \mathcal{D}} \mathbb{E}_{\mathbb{p}}^{\otimes n} |\widehat{\mathbb{p}}_{\mathrm{b}_o}(x_o) - \mathbb{p}(x_o)|^2 \leqslant C n^{-2\beta/(2\beta+1)},$$

where $C \in \mathbb{R}_{\setminus 0}^{+}$ is a constant depending only on β, L, c and on the kernel K.

§13.24 **Proof** of Theorem §13.23. is given in the lecture. □

Global measure of accuracy.

§13.25 **Definition.** The *mean integrated squared error* of a kernel density estimator $\widehat{\mathbb{p}}_{\mathrm{b}} \in \mathscr{L}_2$ satisfies

$$\mathbb{mise} := \mathbb{E}_{\mathbb{p}}^{\otimes n} \|\widehat{\mathbb{p}}_{\mathrm{b}} - \mathbb{p}\|_{\mathscr{L}_2}^2 = \lambda(\mathbb{var}) + \lambda(|\mathbb{bias}|^2) \quad \text{for a density } \mathbb{p} \in \mathscr{L}_2$$

using the variance and bias term as in Definition §13.06. □

We study now separately the integrated variance and bias term.

§13.26 **Lemma.** *Let $K \in \mathscr{L}_1 \cap \mathscr{L}_2$ with $\lambda(K) = 1$. We have $\lambda(\mathbb{var}) \leqslant (n\mathrm{b})^{-1} \|K\|_{\mathscr{L}_2}^2$ for any density $\mathbb{p}$, $\mathrm{b} \in \mathbb{R}_{\setminus 0}^{+}$ and $n \in \mathbb{N}$.*

§13.27 **Proof** of Lemma §13.26. is given in the lecture. □

§13.28 **Definition.** For $\beta, L \in \mathbb{R}^+_{\setminus 0}$ 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 $f^{(l)}$ satisfies $\|f^{(l)}(\bullet + t) - f^{(l)}\|_{\mathscr{L}_2} \leqslant L|t|^{\beta - l}$ for all $t \in \mathbb{R}$. □

§13.29 **Lemma.** *Suppose that $\mathbb{p} \in \mathscr{L}_2 \cap \mathcal{N}(\beta, L)$ and let K be a kernel of order $l = \lfloor \beta \rfloor$ satisfying $|\mathrm{id}_{\mathbb{R}}|^\beta K \in \mathscr{L}_1$. Then we have $\|\mathbb{b}\mathrm{ias}\|_{\mathscr{L}_2} \leqslant \mathrm{b}^\beta \frac{L}{l!} \lambda(|\mathrm{id}_{\mathbb{R}}|^\beta |K|)$ for each $\mathrm{b} \in \mathbb{R}^+_{\setminus 0}$ and $n \in \mathbb{N}$.*

§13.30 **Proof** of Lemma §13.29. is given in the lecture. □

§13.31 **Remark.** Let $\mathbb{p} \in \mathscr{L}_2 \cap \mathcal{N}(\beta, L)$ and let $K \in \mathscr{L}_2$ be a kernel of order $l = \lfloor \beta \rfloor$ satisfying $|\mathrm{id}_{\mathbb{R}}|^\beta K \in \mathscr{L}_1$. By combination of Lemmata §13.26 and §13.29 follows

$$\mathbb{m}\mathrm{ise} \leqslant (n\mathrm{b})^{-1} \|K\|^2_{\mathscr{L}_2} + \mathrm{b}^{2\beta} \big(\tfrac{L}{l!} \lambda(|\mathrm{id}_{\mathbb{R}}|^\beta |K|) \big)^2 .$$

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

§13.32 **Theorem.** *Suppose that $\mathbb{p} \in \mathscr{L}_2 \cap \mathcal{N}(\beta, L)$ and let $K \in \mathscr{L}_2$ be a kernel of order $l = \lfloor \beta \rfloor$ satisfying in addition $|\mathrm{id}_{\mathbb{R}}|^\beta K \in \mathscr{L}_1$. Fix $c \in \mathbb{R}^+_{\setminus 0}$ and set $\mathrm{b}_o = c n^{-1/(2\beta+1)}$ then for all $n \in \mathbb{N}$*

$$\mathbb{E}^{\otimes n}_{\mathbb{p}} \|\widehat{\mathbb{p}}_{\mathrm{b}_o} - \mathbb{p}\|^2_{\mathscr{L}_2} \leqslant C n^{-2\beta/(2\beta+1)},$$

where $C \in \mathbb{R}^+_{\setminus 0}$ is a constant depending only on β, L, c and on the kernel K.

§13.33 **Proof** of Theorem §13.32. is given in the lecture. □

Data-driven bandwidth selection.

§13.34 **Oracle choice.** Considering $\mathbb{m}\mathrm{ise}(\mathrm{b}) := \mathbb{E}^{\otimes n}_{\mathbb{p}} \|\widehat{\mathbb{p}}_{\mathrm{b}} - \mathbb{p}\|^2_{\mathscr{L}_2}$ of a kernel density estimator $\widehat{\mathbb{p}}_{\mathrm{b}}$ the choice of the bandwidth b is crucial. For instance, an ideal value b_o of the bandwidth satisfies $\mathbb{m}\mathrm{ise}(\mathrm{b}_o) = \inf\{\mathbb{m}\mathrm{ise}(\mathrm{b}) : \mathrm{b} \in \mathbb{R}^+_{\setminus 0}\}$. Note that for a given density $\mathbb{p}$, the estimator $\widehat{\mathbb{p}}_{\mathrm{b}_o}$, if b_o exists, has minimal $\mathbb{m}\mathrm{ise}(\mathrm{b}_o)$ within the family $\{\widehat{\mathbb{p}}_{\mathrm{b}} : \mathrm{b} \in \mathbb{R}^+_{\setminus 0}\}$ of all kernel density estimators with fixed kernel and varying bandwidth. Unfortunately, $\mathbb{m}\mathrm{ise}(\mathrm{b}) = \mathbb{E}^{\otimes n}_{\mathbb{p}} \|\widehat{\mathbb{p}}_{\mathrm{b}} - \mathbb{p}\|^2_{\mathscr{L}_2}$ depends on unknown characteristics of the density $\mathbb{p}$. Therefore, both the bandwidth b_o and the kernel density estimator $\widehat{\mathbb{p}}_{\mathrm{b}_o}$ remain purely theoretical and thus they are often called *oracle*. □

§13.35 **Cross validation.** A common idea is to minimise a unbiased estimator rather than $\mathbb{m}\mathrm{ise}(\mathrm{b}) = J(\mathrm{b}) + \lambda(\mathbb{p}^2)$ with $J(\mathrm{b}) := \mathbb{E}^{\otimes n}_{\mathbb{p}}\{\lambda(\widehat{\mathbb{p}}^2_{\mathrm{b}}) - 2\lambda(\mathbb{p}\widehat{\mathbb{p}}_{\mathrm{b}})\}$. We observe that $\lambda(\mathbb{p}^2)$ does not depend on the bandwidth b and hence, the oracle choice b_o, if it exists, satisfies $J(\mathrm{b}_o) = \min\{J(\mathrm{b}) : \mathrm{b} \in \mathbb{R}^+_{\setminus 0}\}$. To construct a unbiased estimator of $J(\mathrm{b})$ it is sufficient to estimate $\mathbb{E}^{\otimes n}_{\mathbb{p}} \lambda(\widehat{\mathbb{p}}^2_{\mathrm{b}})$ and $\mathbb{E}^{\otimes n}_{\mathbb{p}} \lambda(\widehat{\mathbb{p}}_{\mathrm{b}} \mathbb{p})$ without bias. Obviously, $\lambda(\widehat{\mathbb{p}}^2_{\mathrm{b}})$ is a unbiased estimator of $\mathbb{E}^{\otimes n}_{\mathbb{p}} \lambda(\widehat{\mathbb{p}}^2_{\mathrm{b}})$. For $x \in \mathbb{R}$, $i \in [\![n]\!]$ and $x^n = (x_i)_{i \in [\![n]\!]} \in \mathbb{R}^n$ we consider $\widehat{\mathbb{p}}^{-i}_{\mathrm{b}}(x, x^n) := \frac{1}{n-1} \sum_{j \in [\![n]\!] \setminus \{i\}} K_{\mathrm{b}}(x, x_j)$, and $\big(\widehat{\mathbb{P}}_n(\widehat{\mathbb{p}}^{-i}_{\mathrm{b}})\big)(x^n) = \frac{1}{n} \sum_{i \in [\![n]\!]} \widehat{\mathbb{p}}^{-i}_{\mathrm{b}}(x_i, x^n)$, hence $\widehat{\mathbb{p}}^{-i}_{\mathrm{b}}(x) \in \mathscr{B}^n$ and $\widehat{\mathbb{P}}_n(\widehat{\mathbb{p}}^{-i}_{\mathrm{b}}) \in \mathscr{B}^n$, where the latter by construction is an unbiased estimator of $\mathbb{E}^{\otimes n}_{\mathbb{p}} \lambda(\widehat{\mathbb{p}}_{\mathrm{b}} \mathbb{p})$. Note that for each $i \in [\![n]\!]$ the i-th coordinate map $x^n \mapsto \Pi_i(x^n) := x_i$ and $\widehat{\mathbb{p}}^{-i}_{\mathrm{b}}$ are independent in a statistical product experiment,

which in tun implies, that $x^n \mapsto \widehat{\mathbb{p}}_{\mathrm{b}}^{-i}(x_i, x^n)$ is a unbiased estimator of $\mathbb{E}_{\mathbb{p}}^{\otimes n}\lambda(\widehat{\mathbb{p}}_{\mathrm{b}}\mathbb{p})$. To summarise, for each $\mathrm{b} \in \mathbb{R}_{\setminus 0}^{+}$ the *(leave-one-out) cross-validation criterion* $\widehat{J}(\mathrm{b}) := \lambda(\widehat{\mathbb{p}}_{\mathrm{b}}^2) - \frac{1}{2}\widehat{\mathbb{P}}_n(\widehat{\mathbb{p}}_{\mathrm{b}}^{-i})$ is an unbiased estimator of $J(\mathrm{b})$, i.e., $J(\mathrm{b}) = \mathbb{E}_{\mathbb{p}}^{\otimes n}\widehat{J}(\mathrm{b})$. Recall, that the oracle b_o minimises both $\mathbb{mise}(\mathrm{b})$ and $\mathbb{E}_{\mathbb{p}}^{\otimes n}\{\widehat{J}(\mathrm{b})\}$ over $\mathrm{b} \in \mathbb{R}_{\setminus 0}^{+}$. Therefore, a reasonable and feasible choice $\widehat{\mathrm{b}}$ of the bandwidth, if it exists, satisfies $\widehat{J}(\widehat{\mathrm{b}}) = \min\{\widehat{J}(\mathrm{b}), \mathrm{b} \in \mathbb{R}_{\setminus 0}^{+}\}$. Finally, we define the cross-validation estimator $\widehat{\mathbb{p}}_{\widehat{\mathrm{b}}}$. Note that $\widehat{\mathbb{p}}_{\widehat{\mathrm{b}}}$ is a kernel density estimator with random bandwidth $\widehat{\mathrm{b}}$ depending on the sample only. Under appropriate conditions the mise of the estimator $\widehat{\mathbb{p}}_{\widehat{\mathrm{b}}}$ is asymptotically equivalent to that of the oracle kernel density (pseudo)-estimator $\widehat{\mathbb{p}}_{\mathrm{b}_o}$. □

§14 Nonparametric regression by local smoothing

Here and subsequently, we consider a statistical product experiment $(\mathbb{R}^{2n}, \mathscr{B}^{2n}, \mathbb{P}_{\mathcal{F}}^{\otimes n})$ as introduced in Section §12. Let $\{(Y_i, X_i), i \in [\![n]\!]\}$ be an i.i.d. sample of $(Y, X) \sim \mathbb{P}_f$. Introducing the coordinate maps $\Pi_Y, \Pi_X \in \mathscr{B}^2$ with $(y, x) \mapsto \Pi_Y(y, x) := y$ and $(y, x) \mapsto \Pi_X(y, x) := x$ we tactically identify Y and X with Π_Y and Π_X, respectively, and thus, (Y, X) with the identity $\mathrm{id}_{\mathbb{R}^2}$. We denote by $\mathbb{P}_X$ the marginal distribution of the regressor X and by $\mathbb{E}_f\big(Y\big|X\big)$ a conditional expectation of Y given X (see Section §03).

§14.01 **Assumption**. The random vector $(Y, X) \in (\mathscr{B}^2)^2$ obeys $\mathbb{P}_X$-a.e. a nonparametric regression model $\mathbb{E}_f\big(Y\big|X\big) = f$ for some unknown regression function $f \in \mathcal{F}$.

(NPR1) The error term $\varepsilon := Y - f(X)$ has a finite second moment, i.e., $\varepsilon \in \mathscr{L}_2(\mathbb{P}_f)$, and hence, $\mathbb{E}_f(\varepsilon) = 0$. We set $\sigma_\varepsilon^2 := \mathbb{E}_f(\varepsilon^2)$. The error term ε and the explanatory variable X are independent.

(NPR2) The joint distribution $\mathbb{P}_f$ of (Y, X) admits a joint Lebesgue density $\mathbb{p} \in (\mathscr{B}^2)^+$, i.e. $\mathbb{p} = \mathrm{d}\mathbb{P}_f/\mathrm{d}\lambda^2$ and $\mathbb{P}_f = \mathbb{p}\lambda^2$. Denote by $\mathbb{p}^X$ the marginal density of X. Using for the conditional density $\mathbb{p}_{Y|X}$ of Y given X the $\mathbb{P}_f$-a.s. identity $\mathbb{p}^X\mathbb{p}_{Y|X} = \mathbb{p}$ (see Notation §03.11 (iii)) define $\mathbb{q} := f\mathbb{p}^X$ as in (12.1). □

§14.02 **Heuristics**. Given a *bandwidth* $\mathrm{b} \in \mathbb{R}_{\setminus 0}^{+}$ and *evaluation points* $y_o, x_o \in \mathbb{R}$ define $K_{\mathrm{b}}(y_o)$ and $K_{\mathrm{b}}(x_o)$ as in Definition §13.01. The statistic $\widehat{\mathbb{p}}_{\mathrm{b}}(y_o, x_o) := \widehat{\mathbb{P}}_n\big(K_{\mathrm{b}}(y_o, Y)K_{\mathrm{b}}(x_o, X)\big) \in \mathscr{B}^{2n}$,

$$(y, x)^n = ((y_i, x_i))_{i\in[\![n]\!]} \mapsto \widehat{\mathbb{p}}_{\mathrm{b}}(y_o, x_o, (y, x)^n) = \frac{1}{n}\sum_{i\in[\![n]\!]} K_{\mathrm{b}}(y_o, y_i)K_{\mathrm{b}}(x_o, x_i)$$

is a kernel density estimator of the joint density $\mathbb{p} \in \mathscr{L}_1(\lambda^2)$ evaluated at (y_o, x_o). Exploiting $\lambda(K) = 1$ the marginal $\widehat{\mathbb{p}}_{\mathrm{b}}^X(x_o) := \widehat{\mathbb{P}}_n K_{\mathrm{b}}(x_o, X) := \int_{\mathbb{R}} \widehat{\mathbb{p}}_{\mathrm{b}}(y_o, x_o)dy_o \in \mathscr{B}^n$ is a kernel density estimator of the marginal density $\mathbb{p}^X \in \mathscr{L}_1$ evaluated at $x_o \in \mathbb{R}$. Keeping (12.1) in mind we estimate $\mathbb{q}(x_o)$ by replacing the unknown density $\mathbb{p}$ by its kernel estimator $\widehat{\mathbb{p}}_{\mathrm{b}}$, that is, $\widehat{\mathbb{q}}_{\mathrm{b}}(x_o) := \int_{\mathbb{R}} y_o\widehat{\mathbb{p}}_{\mathrm{b}}(y_o, x_o)dy_o \in \mathscr{B}^{2n}$. If the kernel K is in addition of order 1, i.e. $\lambda(\mathrm{id}_{\mathbb{R}}K) = 0$, then we have $\widehat{\mathbb{q}}_{\mathrm{b}}(x_o) = \widehat{\mathbb{P}}_n(YK_{\mathrm{b}}(x_o, X))$ where $\widehat{\mathbb{q}}_{\mathrm{b}}(x_o, (y, x)^n) = \frac{1}{n}\sum_{i\in[\![n]\!]} y_iK_{\mathrm{b}}(x_o, x_i)$ for all $(y, x)^n = ((y_i, x_i))_{i\in[\![n]\!]} \in \mathbb{R}^{2n}$. □

§14.03 **Definition**. Given a kernel K of order 1, a bandwidth $\mathrm{b} \in \mathbb{R}_{\setminus 0}^{+}$, and an evaluation point $x_o \in \mathbb{R}$

the statistic $\widehat{f}_{\mathrm{b}}(x_o) := \frac{\widehat{\mathfrak{q}}_{\mathrm{b}}(x_o)}{\widehat{\mathbb{p}}^X_{\mathrm{b}}(x_o)} \mathbb{1}_{\mathbb{R}^+_{\setminus 0}}(|\widehat{\mathbb{p}}^X_{\mathrm{b}}(x_o)|) \in \mathscr{B}^{2n}$ defined for all $(y,x)^n \in \mathbb{R}^{2n}$ by

$$\widehat{f}_{\mathrm{b}}(x_o,(y,x)^n) = \frac{\widehat{\mathfrak{q}}_{\mathrm{b}}(x_o,(y,x)^n)}{\widehat{\mathbb{p}}^X_{\mathrm{b}}(x_o,(y,x)^n)} = \sum_{i\in[\![n]\!]} y_i \frac{K_{\mathrm{b}}(x_o,x_i)}{\sum_{j\in[\![n]\!]} K_{\mathrm{b}}(x_o,x_j)}, \quad \text{if } |\widehat{\mathbb{p}}^X_{\mathrm{b}}(x_o,(y,x)^n)| \in \mathbb{R}^+_{\setminus 0}$$

and $\widehat{f}(x_o,(y,x)^n) = 0$ otherwise, is called *Nadaraya–Watson estimator* of $f(x_o)$. □

Local measure of accuracy.

§14.04 **Comment**. We make use of the properties of a kernel density estimator derived in Section §13 in order to analyse the estimator $\widehat{\mathbb{p}}^X_{\mathrm{b}}$ of $\mathbb{p}^X$. As a consequence, it remains to consider the estimator $\widehat{\mathfrak{q}}_{\mathrm{b}}$ of $\mathfrak{q}$. We consider first its mean squared error at a given point $x_o \in \mathbb{R}$, that is,

$$\mathbb{mse}_{\mathfrak{q}}(x_o) = \mathbb{E}_f^{\otimes n}|\widehat{\mathfrak{q}}_{\mathrm{b}}(x_o) - \mathfrak{q}(x_o)|^2 = \mathbb{var}_{\mathfrak{q}}(x_o) + |\mathbb{bias}_{\mathfrak{q}}(x_o)|^2$$

by introducing a *variance* and a *bias* term, respectively,

$$\mathbb{var}_{\mathfrak{q}}(x_o) := \mathbb{E}_f^{\otimes n}|\widehat{\mathfrak{q}}_{\mathrm{b}}(x_o) - \mathbb{E}_f^{\otimes n}\widehat{\mathfrak{q}}_{\mathrm{b}}(x_o)|^2 \quad \text{and} \quad \mathbb{bias}_{\mathfrak{q}}(x_o) := \mathbb{E}_f^{\otimes n}\widehat{\mathfrak{q}}_{\mathrm{b}}(x_o) - \mathfrak{q}(x_o).$$

In the sequel we analyse separately the variance and bias term. □

§14.05 **Lemma**. *Under Assumption §14.01 let $f, \mathbb{p}^X \in \mathscr{L}_\infty$ and $K \in \mathscr{L}_1 \cap \mathscr{L}_2$ with $\lambda(K) = 1$. For each $x_o \in \mathbb{R}$, $\mathrm{b} \in \mathbb{R}^+_{\setminus 0}$ and $n \in \mathbb{N}$ we have $\mathbb{var}_{\mathfrak{q}}(x_o) \leqslant (n\mathrm{b})^{-1}(\|f\|^2_{\mathscr{L}_\infty} + \sigma_\varepsilon^2)\,\|\mathbb{p}^X\|_{\mathscr{L}_\infty}\,\|K\|^2_{\mathscr{L}_2}$.*

§14.06 **Proof** of Lemma §14.05. is given in the lecture. □

Recall the Definitions §13.16 and §13.19 of a Hölder class and a higher order kernel.

§14.07 **Corollary**. *Under Assumption §14.01 let $\mathfrak{q} \in \mathcal{H}(\beta, L)$ and let K be a kernel of order $l = \lfloor\beta\rfloor$ with $|\mathrm{id}_{\mathbb{R}}|^\beta K \in \mathscr{L}_1$. Then, $|\mathbb{bias}_{\mathfrak{q}}(x_o)| \leqslant \mathrm{b}^\beta \frac{L}{l!}\lambda(|\mathrm{id}_{\mathbb{R}}|^\beta|K|)$ for each $x_o \in \mathbb{R}$, $\mathrm{b} \in \mathbb{R}^+_{\setminus 0}$, $n \in \mathbb{N}$.*

§14.08 **Proof** of Corollary §14.07. Due to the identity $\mathbb{bias}_{\mathfrak{q}}(x_o) = \lambda(K_{\mathrm{b}}(x_o)\mathfrak{q}) - \mathfrak{q}(x_o)$ the assertion follows immediately from Lemma §13.20 (replace the density $\mathbb{p}$ by $\mathfrak{q}$). □

§14.09 **Remark**. We note that $f, \mathbb{p}^X \in \mathscr{L}_\infty$ implies $\mathfrak{q} = f\mathbb{p}^X \in \mathscr{L}_\infty$. Suppose that $\mathfrak{q} \in \mathcal{H}(\beta, L)$ and let $K \in \mathscr{L}_2$ be a kernel of order $l = \lfloor\beta\rfloor$ with $|\mathrm{id}_{\mathbb{R}}|^\beta K \in \mathscr{L}_1$. Combining Lemma §14.05 and Corollary §14.07 we have

$$\sup_{x_o\in\mathbb{R}} \mathbb{mse}_{\mathfrak{q}}(x_o) \leqslant (n\mathrm{b})^{-1}(\|f\|^2_{\mathscr{L}_\infty} + \sigma_\varepsilon^2)\|\mathbb{p}^X\|_{\mathscr{L}_\infty}\|K\|^2_{\mathscr{L}_2} + \mathrm{b}^{2\beta}\big(\tfrac{L}{l!}\lambda(|\mathrm{id}_{\mathbb{R}}|^\beta|K|)\big)^2. \tag{14.1}$$

Suppose further that $\mathbb{p}^X \in \mathcal{H}(\beta, L)$, then combining Lemmata §13.07 and §13.20 an upper bound of $\mathbb{mse}_{\mathbb{p}^X}(x_o) := \mathbb{E}_f^{\otimes n}|\widehat{\mathbb{p}}^X_{\mathrm{b}}(x_o) - \mathbb{p}^X(x_o)|^2$ is given by (see Remark §13.22)

$$\sup_{x_o\in\mathbb{R}} \mathbb{mse}_{\mathbb{p}^X}(x_o) \leqslant (n\mathrm{b})^{-1}\|\mathbb{p}^X\|_{\mathscr{L}_\infty}\|K\|^2_{\mathscr{L}_2} + \mathrm{b}^{2\beta}\big(\tfrac{L}{l!}\lambda(|\mathrm{id}_{\mathbb{R}}|^\beta|K|)\big)^2. \tag{14.2}$$

Therefore minimising the right hand side in eqs. (14.1) and (14.2) as a function of b leads to an optimal bandwidth $\mathrm{b}_o = c\,n^{-1/(2\beta+1)}$ with constant $c \in \mathbb{R}^+_{\setminus 0}$ depending on f, $\mathbb{p}^X$ and K. □

§14.10 **Proposition**. *Under Assumption §14.01 suppose that $\mathfrak{q}, \mathbb{p}^X \in \mathcal{H}(\beta, L)$ and let $K \in \mathscr{L}_2$ be a kernel of order $l = \lfloor\beta\rfloor$ with $|\mathrm{id}_{\mathbb{R}}|^\beta K \in \mathscr{L}_1$. Fix $c \in \mathbb{R}^+_{\setminus 0}$ and set $\mathrm{b}_o = cn^{-1/(2\beta+1)}$. If $\mathbb{p}^X(x_o) > 0$ then $|\widehat{f}_{\mathrm{b}_o}(x_o) - f(x_o)|^2 = \mathrm{O}_{\mathbb{P}_f^{\otimes n}}(n^{-2\beta/(2\beta+1)})$.*

§14.11 **Proof** of Proposition §14.10. is given in the lecture. □

§14.12 **Remark**. It is straightforward to show that under similar assumption as used in Lemma §13.13 the asymptotic normality of $\widehat{\mathfrak{q}}_{\mathfrak{b}}(x_o)$ holds true, which due to Slutky's lemma §02.10 allows then to establish the asymptotic normality of $\widehat{f}_{\mathfrak{b}}(x_o)$. □.

Global measure of accuracy.

§14.13 **Comment**. We make use of the properties of a kernel density estimator derived in Section §13 in order to analyse the mean integrated squared error of the estimator $\widehat{\mathbb{p}}^X_{\mathfrak{b}}$ under the additional assumption $\mathbb{p}^X \in \mathscr{L}_2$. As a consequence, it remains to study the estimator $\widehat{\mathfrak{q}}_{\mathfrak{b}}$ of $\mathfrak{q} \in \mathscr{L}_2$, where

$$\mathfrak{mise}_{\mathfrak{q}} = \mathbb{E}_f^{\otimes n}\|\widehat{\mathfrak{q}}_{\mathfrak{b}} - \mathfrak{q}\|^2_{\mathscr{L}_2} = \lambda(\mathfrak{var}_{\mathfrak{q}}) + \lambda(|\mathbb{bias}_{\mathfrak{q}}|^2)$$

using the variance and bias term as in Comment §14.04. □

We study now separately the integrated variance and bias term.

§14.14 **Lemma**. *Under Assumption §14.01 let $f \in \mathscr{L}_2(\mathbb{P}_X)$ and $K \in \mathscr{L}_1 \cap \mathscr{L}_2$ with $\lambda(K) = 1$. For all $\mathfrak{b} \in \mathbb{R}^+_{\setminus 0}$ and $n \in \mathbb{N}$ we have $\lambda(\mathfrak{var}_{\mathfrak{q}}) \leqslant (n\mathfrak{b})^{-1}\sigma_Y^2\|K\|^2_{\mathscr{L}_2}$ with $\sigma_Y^2 := \mathbb{E}_f Y^2 = \mathbb{P}_X(f^2) + \sigma_\varepsilon^2$.*

§14.15 **Proof** of Lemma §14.14. is given in the lecture. □

Recall the Definitions §13.16 and §13.28 of a Nikol'ski class and a higher order kernel.

§14.16 **Corollary**. *Under Assumption §14.01 let $\mathfrak{q} \in \mathscr{L}_2 \cap \mathcal{N}(\beta, L)$ and let K be a kernel of order $l = \lfloor\beta\rfloor$ with $|\mathrm{id}_{\mathbb{R}}|^\beta K \in \mathscr{L}_1$. Then, $\|\mathbb{bias}_{\mathfrak{q}}\|_{\mathscr{L}_2} \leqslant \mathfrak{b}^\beta \frac{L}{l!}\lambda(|\mathrm{id}_{\mathbb{R}}|^\beta|K|)$ for all $\mathfrak{b} \in \mathbb{R}^+_{\setminus 0}$, $n \in \mathbb{N}$.*

§14.17 **Proof** of Corollary §14.16. Making use of the identity $\mathbb{bias}_{\mathfrak{q}}(x_o) = \lambda(K_{\mathfrak{b}}(x_o)\mathfrak{q}) - \mathfrak{q}(x_o)$ and replacing $\mathfrak{q}$ by the density $\mathbb{p}$ the assertion follows immediately from Lemma §13.29. □

§14.18 **Remark**. Let $f \in \mathscr{L}_2(\mathbb{P}_X)$, $\mathfrak{q} \in \mathscr{L}_2 \cap \mathcal{N}(\beta, L)$ and let $K \in \mathscr{L}_2$ be a kernel of order $l = \lfloor\beta\rfloor$ satisfying $|\mathrm{id}_{\mathbb{R}}|^\beta K \in \mathscr{L}_1$. Combining Lemma §14.14 and Corollary §14.16 we have

$$\mathfrak{mise}_{\mathfrak{q}} \leqslant (n\mathfrak{b})^{-1}\sigma_Y^2\|K\|^2_{\mathscr{L}_2} + \mathfrak{b}^{2\beta}\big(\tfrac{L}{l!}\lambda(|\mathrm{id}_{\mathbb{R}}|^\beta|K|)\big)^2. \tag{14.3}$$

Suppose further that $\mathbb{p}^X \in \mathscr{L}_2 \cap \mathcal{N}(\beta, L)$, then combining Lemmata §13.26 and §13.29 an upper bound of $\mathfrak{mise}_{\mathbb{p}^X} := \mathbb{E}_f^{\otimes n}\|\widehat{\mathbb{p}}^X_{\mathfrak{b}} - \mathbb{p}^X\|^2_{\mathscr{L}_2}$ is given by (see Remark §13.31)

$$\mathfrak{mise}_{\mathbb{p}^X} \leqslant (n\mathfrak{b})^{-1}\|K\|^2_{\mathscr{L}_2} + \mathfrak{b}^{2\beta}\big(\tfrac{L}{l!}\lambda(|\mathrm{id}_{\mathbb{R}}|^\beta|K|)\big)^2. \tag{14.4}$$

Therefore minimising the right hand side in eqs. (14.3) and (14.4) as a function of $\mathfrak{b}$ leads to an optimal bandwidth $\mathfrak{b}_o = c\,n^{-1/(2\beta+1)}$ with constant $c \in \mathbb{R}^+_{\setminus 0}$ depending on f, $\mathbb{p}^X$ and K. □

In order to derive an upper bound for the mise of $\widehat{f}_{\mathfrak{b}}$ we use in the next assertion a regularised version which makes use of a stronger assumption, that is, $\mathbb{p}^X(x) > p_o$ for all $x \in A$, for some known constant $p_o > 0$ and measurable support $A \in \mathscr{B}$.

§14.19 **Proposition**. *Under Assumption §14.01 suppose that $f \in \mathscr{L}_2(\mathbb{P}_X)$, $\mathfrak{q}, \mathbb{p}^X \in \mathscr{L}_2 \cap \mathcal{N}(\beta, L)$ and let $K \in \mathscr{L}_2$ be a kernel of order $l = \lfloor\beta\rfloor$ satisfying $|\mathrm{id}_{\mathbb{R}}|^\beta K \in \mathscr{L}_1$. Assume in addition that $\mathbb{p}^X(x) > p_o$ for all $x \in A$, for some known constant $p_o > 0$ and set $A \in \mathscr{B}$. Consider the*

regularised Nadaraya–Watson estimator $\widehat{f}^{o}_{\mathrm{b}}(x) := \frac{\widehat{\mathbb{q}}_{\mathrm{b}}(x)}{\widehat{\mathbb{p}}^{X}_{\mathrm{b}}(x)} \mathbb{1}_{\{\widehat{\mathbb{p}}^{X}_{\mathrm{b}}(x) > p_o/2\}}$ *for all* $x \in A$. *Fix* $c \in \mathbb{R}^{+}_{\setminus 0}$ *and set* $\mathrm{b}_o = cn^{-1/(2\beta+1)}$ *then for all* $n \in \mathbb{N}$

$$\mathbb{E}_f^{\otimes n} \| (\widehat{f}^{o}_{\mathrm{b}_o} - f) \mathbb{1}_A \|^2_{\mathscr{L}_2} \leqslant C n^{-2\beta/(2\beta+1)},$$

where $C \in \mathbb{R}^{+}_{\setminus 0}$ *is a constant depending only on* β, L, c, p_o *and on the kernel* K.

§14.20 **Proof** of Proposition §14.19. is given in the lecture. □

Local polynomial estimators.

§14.21 **Heuristics.** Let the kernel K take only values in $\mathbb{R}^{+}$. It is then easily verified, that the Nadaraya–Watson estimator $\widehat{f}_{\mathrm{b}} \in \mathscr{B}^{2n}$ as in Definition §14.03 satisfies

$$\widehat{f}_{\mathrm{b}}(x_o, (y,x)^n) \in \operatorname*{arg\,inf}_{\theta \in \mathbb{R}} \sum_{i \in [\![n]\!]} (y_i - \theta)^2 K_{\mathrm{b}}(x_o, x_i).$$

Therefore, $\widehat{f}_{\mathrm{b}}$ is obtained by a local constant least squares approximation of the responses $\{y_i\}$. The locality is determined by the kernel K that downweights all the x_i that are not close to x_o 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. □

§14.22 **Definition.** For $m \in \mathbb{R}$ consider $U : \mathbb{R} \to \mathbb{R}^{l+1}$, $z \mapsto U(z) = (1, z, z^2/2!, \ldots, z^m/m!)$. Let $K : \mathbb{R} \to \mathbb{R}$ be a kernel and $\mathrm{b} \in \mathbb{R}^{+}_{\setminus 0}$ be a bandwidth. A random vector $\widehat{\theta}(x_o) \in \mathscr{B}^{m+1}$ satisfying

$$\widehat{\theta}(x_o, (y,x)^n) \in \operatorname*{arg\,inf}_{\theta \in \mathbb{R}^{m+1}} \sum_{i \in [\![n]\!]} (y_i - \theta^t U(\tfrac{x_i - x_o}{\mathrm{b}}))^2 K_{\mathrm{b}}(x_o, x_i).$$

is called a *local polynomial estimator of order* m of $\theta(x_o) = (f(x_o), \mathrm{b}\dot{f}(x_o), \ldots, \mathrm{b}^m f^{(m)}(x_o))$. The statistic $\widehat{f}_{\mathrm{b}}(x_o) = U^t(0)\widehat{\theta}(x_o)$ is called *local polynomial estimator of order* m of $f(x_o)$. □

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

§15 Sequence space models

In the sequel we study nonparametric estimation of a functional parameter of interest θ based on a noisy version $\widehat{\theta} = \theta + n^{-1/2}\dot{W}$ of θ 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 $\widehat{\theta}$ is constructed using a sample of size n. For convenience, we eventually consider the measure space $([0,1], \mathscr{B}_{[0,1]}, \lambda_{[0,1]})$ where $\lambda_{[0,1]}$ denotes the restriction of the Lebesgue measure to $\mathscr{B}_{[0,1]}$. We exemplary suppose that the function of interest $\theta : [0,1] \to \mathbb{R}$ is Borel-measurable, i.e. $\theta \in \mathscr{B}_{[0,1]}$. In addition we assume that $\theta \in \mathscr{L}_2 := \mathscr{L}_2(\mathscr{B}_{[0,1]}, \lambda_{[0,1]})$ and thus θ permits an orthogonal series expansion. With a slight abuse of notations we write

shortly $\lambda := \lambda_{[0,1]}$ and $\mathscr{L}_s := \mathscr{L}_s(\mathscr{B}_{[0,1]}, \lambda)$ for $s \geqslant 1$ (see Notation §01.03). In Section §17 below 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 sections we illustrate the results using three particular models, namely, nonparametric regression with uniformly distributed random design, nonparametric density estimation and a Gaussian sequence space model.

§15.01 **Nonparametric density estimation.** Let $\mathcal{D}_2$ be a set of square-integrable Lebesgue densities on $([0,1], \mathscr{B}_{[0,1]})$, and hence $\mathcal{D}_2 \subseteq \mathscr{L}_2 = \mathscr{L}_2(\mathscr{B}_{[0,1]}, \lambda)$. $\mathbb{P}_\mathfrak{p} = \mathfrak{p}\lambda$ and $\mathbb{E}_\mathfrak{p}$ denote for each density $\mathfrak{p} \in \mathcal{D}_2$ the associated probability measure and expectation, respectively. We consider the statistical product experiment $([0,1]^n, \mathscr{B}^n_{[0,1]}, \mathbb{P}^{\otimes n}_{\mathcal{D}_2} = (\mathbb{P}^{\otimes n}_\mathfrak{p})_{\mathfrak{p}\in\mathcal{D}_2})$. Since $\mathfrak{p} \in \mathscr{L}_2$ we have $\mathbb{P}_\mathfrak{p}(|h|) = \lambda(|h|\mathfrak{p}) \leqslant \|h\|_{\mathscr{L}_2}\|\mathfrak{p}\|_{\mathscr{L}_2} < \infty$ for all $h \in \mathscr{L}_2$ and thus $\mathscr{L}_2 \subseteq \mathscr{L}_1(\mathbb{P}_\mathfrak{p})$ in equal. We write shortly $\mathfrak{p}_h := \lambda(h\mathfrak{p}) = \mathfrak{p}\lambda(h) = \mathbb{P}_\mathfrak{p}(h)$. We note that $\mathfrak{p} \in \mathscr{L}_2$ is uniquely determined by the family $(\mathfrak{p}_h)_{h\in\mathscr{L}_2}$ up to λ-a.s. equality (see Example §17.03 (d)). For each $h \in \mathscr{L}_2$ the statistic $\widehat{\mathfrak{p}}_h := \widehat{\mathbb{P}}_n h \in \mathscr{B}^n_{[0,1]}$ with $x^n = (x_i)_{i\in[\![n]\!]} \mapsto \widehat{\mathfrak{p}}_h(x^n) = \frac{1}{n}\sum_{i\in[\![n]\!]} h(x_i)$ is an unbiased estimator of $\mathfrak{p}_h$. The centred statistic $\dot{W}_h := n^{1/2}(\widehat{\mathbb{P}}_n h - \mathbb{P}_\mathfrak{p}(h)) \in \mathscr{B}^n_{[0,1]}$, i.e. $\dot{W}_h \in \mathscr{L}_1(\mathbb{P}^{\otimes n}_\mathfrak{p})$ with $\mathbb{P}^{\otimes n}_\mathfrak{p}(\dot{W}_h) = 0$, satisfies $\widehat{\mathfrak{p}}_h = \mathfrak{p}_h + n^{-1/2}\dot{W}_h$ by construction. Considering the families $\widehat{\mathfrak{p}} := (\widehat{\mathfrak{p}}_h)_{h\in\mathscr{L}_2}$ and $\dot{W} := (\dot{W}_h)_{h\in\mathscr{L}_2}$ of real-valued random variables defined on the common probability space $([0,1]^n, \mathscr{B}^n_{[0,1]}, \mathbb{P}^{\otimes n}_\mathfrak{p})$ we write shortly $\widehat{\mathfrak{p}} = \mathfrak{p} + n^{-1/2}\dot{W}$, meaning that, $\widehat{\mathfrak{p}}_h = \mathfrak{p}_h + n^{-1/2}\dot{W}_h$ for all $h \in \mathscr{L}_2$. □

§15.02 **Nonparametric regression.** Let $(Y, X) \in (\mathscr{B}^2)^2$ obey $\mathbb{P}_X$-a.e. a nonparametric regression model $\mathbb{E}_f(Y|X) = f$ satisfying the Assumption §14.01 (see section §14). For convenience, in addition the regressor X is supposed to be uniformly distributed on the interval $[0,1]$, i.e. $X \sim \mathrm{U}_{[0,1]}$. As a consequence, we have $\mathfrak{p}^X = \mathbb{1}_{[0,1]}$ and $\mathscr{L}_2(\mathbb{P}_X) = \mathscr{L}_2(\mathscr{B}_{[0,1]}, \lambda) = \mathscr{L}_2$. Let us denote in this situation by U_f the joint distribution of (Y, X), but keep in mind, that the conditional distribution of the error term given X is still no specified. The regression function $f \in \mathscr{B}_{[0,1]}$ is assumed to be square integrable, i.e., $f \in \mathscr{L}_2$. Recall that by Assumption §14.01 (NPR1) the centred error term $\varepsilon = Y - f(X)$ and the explanatory variable X are independent. Identifying again Y and X with the coordinate map Π_Y and Π_X, respectively, we have $f_h := \lambda(hf) = \mathbb{P}_X(fh) = \mathrm{U}_f(Yh(X))$. We note that $f \in \mathscr{L}_2$ is uniquely determined by the family $(f_h)_{h\in\mathscr{L}_2}$ up to λ-a.s. equality (see Example §17.03 (d)). For each $h \in \mathscr{L}_2$ the statistic $\widehat{f}_h := \widehat{\mathbb{P}}_n(Yh(X)) \in \mathscr{B}^{2n}$ with $(y,x)^n = ((y_i, x_i))_{i\in[\![n]\!]} \mapsto \widehat{f}_h((y,x)^n) = \frac{1}{n}\sum_{i\in[\![n]\!]} y_i h(x_i)$ is an unbiased estimator of f_h. The centred statistic $\dot{W}_h := n^{1/2}(\widehat{\mathbb{P}}_n(Yh(X)) - \mathrm{U}_f(Yh(X))) \in \mathscr{B}^{2n}$, i.e. $\dot{W}_h \in \mathscr{L}_1(\mathrm{U}^{\otimes n}_f)$ with $\mathrm{U}^{\otimes n}_f(\dot{W}_h) = 0$, satisfies $\widehat{f}_h = f_h + n^{-1/2}\dot{W}_h$ by construction. Considering the families $\widehat{f} := (\widehat{f}_h)_{h\in\mathscr{L}_2}$ and $\dot{W} := (\dot{W}_h)_{h\in\mathscr{L}_2}$ of real-valued random variables defined on the common probability space $(\mathbb{R}^{2n}, \mathscr{B}^{2n}, \mathrm{U}^{\otimes n}_f)$ we write shortly $\widehat{f} = f + n^{-1/2}\dot{W}$, meaning that, $\widehat{f}_h = f_h + n^{-1/2}\dot{W}_h$ for all $h \in \mathscr{L}_2$. □

Stochastic process on Hilbert spaces.

Here and subsequently, $(\mathbb{H}, \langle\cdot,\cdot\rangle_\mathbb{H})$ and $\mathcal{U}$ denotes a separable real Hilbert space and a subset of $\mathbb{H}$, respectively. Considering the product spaces $\mathbb{R}^\mathbb{H} = \bigtimes_{h\in\mathbb{H}}\mathbb{R}$ and $\mathbb{R}^\mathcal{U} = \bigtimes_{u\in\mathcal{U}}\mathbb{R}$ the mapping $\Pi_\mathcal{U} : \mathbb{R}^\mathbb{H} \to \mathbb{R}^\mathcal{U}$ given by $y = (y_h)_{h\in\mathbb{H}} \mapsto (y_u)_{u\in\mathcal{U}} =: \Pi_\mathcal{U} y$ is called canonical projection. In particular, for each $h \in \mathbb{H}$ the coordinate map $\Pi_h := \Pi_{\{h\}} : \mathbb{R}^\mathbb{H} \to \mathbb{R}$ is given by $y = (y_{h'})_{h'\in\mathbb{H}} \mapsto y_h =: \Pi_h y$. Moreover, $\mathbb{R}^\mathbb{H}$ 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 on $\mathbb{R}^\mathbb{H}$ such that all coordinate

maps Π_h, $h \in \mathbb{H}$ are measurable. i.e., $\mathscr{B}^{\otimes\mathbb{H}} = \sigma(\Pi_h, h \in \mathbb{H})$.

§15.03 **Stochastic process on $\mathbb{H}$.** Let $(Y_h)_{h\in\mathbb{H}}$ be a family of real random variables on a common probability space $(\Omega, \mathscr{A}, \mathbb{P})$, that is, $Y_h \in \mathscr{A}$ for each $h \in \mathbb{H}$. Consider the $\mathbb{R}^{\mathbb{H}}$-valued random variable $Y := (Y_h)_{h\in\mathbb{H}}$ where $Y : \Omega \to \mathbb{R}^{\mathbb{H}}$ is a $\mathscr{A}$-$\mathscr{B}^{\otimes\mathbb{H}}$-measurable map 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, i.e. $Y \sim \mathbb{P}^Y$ for short. Further, denote by $\mathbb{P}^{Y_\mathcal{U}} = \mathbb{P} \circ Y_\mathcal{U}^{-1} = \mathbb{P}^Y \circ \Pi_\mathcal{U}^{-1}$ the distribution of the stochastic process $Y_\mathcal{U} := \Pi_\mathcal{U} Y = (Y_u)_{u\in\mathcal{U}}$ on $\mathcal{U}$. The family $(\mathbb{P}^{Y_\mathcal{U}})_{\mathcal{U}\subseteq\mathbb{H} \text{ finite}}$ is called *family of finite-dimensional distributions* of Y or $\mathbb{P}^Y$. In particular, $\mathbb{P}^{Y_h} = \mathbb{P}^{\Pi_h Y} = \mathbb{P}^Y \circ \Pi_h^{-1}$ denotes the distribution of $Y_h = \Pi_h Y$. Furthermore, for $h, h_\circ \in \mathbb{H}$ we write $\mathbb{P}(Y_h) = \mathbb{P}^Y(\Pi_h)$ and $\mathbb{C}\mathrm{ov}(Y_h, Y_{h_\circ}) := \mathbb{P}^Y((\Pi_h - \mathbb{P}^Y(\Pi_h))(\Pi_{h_\circ} - \mathbb{P}^Y(\Pi_{h_\circ})))$, if it exists, for the expectation of Y_h and the covariance of Y_h and $Y_{h_\circ}$ with respect to $\mathbb{P}^Y$. □

§15.04 **Definition.** Let $Y := (Y_h)_{h\in\mathbb{H}} \sim \mathbb{P}^Y$ be a stochastic process on $\mathbb{H}$. If $\mathbb{P}(|Y_h|) < \infty$, i.e. $Y_h \in \mathscr{L}_1(\mathbb{P})$ or $\Pi_h \in \mathscr{L}_1(\mathbb{P}^Y)$ in equal, for each $h \in \mathbb{H}$, then the functional $\mathrm{m} : \mathbb{H} \to \mathbb{R}$ with $h \mapsto \mathrm{m}(h) := \mathbb{P}(Y_h)$ is called *mean function* of Y. If the mean function is in addition linear and bounded, that is, $\mathrm{m} \in \mathbb{L}(\mathbb{H}, \mathbb{R})$ (see Definition §17.22), then due to the Fréchet-Riesz representation theorem (Property §17.23) there exists $\theta \in \mathbb{H}$ such that $\mathrm{m}(h) = \langle\theta, h\rangle_\mathbb{H}$ for all $h \in \mathbb{H}$. The element $\mathbb{P}(Y) := \mathbb{P}^Y(\mathrm{id}_\mathbb{H}) := \theta$ is called $\mathbb{H}$-*mean* or *expectation* of Y (or $\mathbb{P}^Y$). If $\mathbb{P}(|Y_h|^2) < \infty$, i.e., $Y_h \in \mathscr{L}_2(\mathbb{P})$ or $\Pi_h \in \mathscr{L}_2(\mathbb{P}^Y)$ in equal, for each $h \in \mathbb{H}$, then the mapping $\mathrm{cov} : \mathbb{H}^2 \to \mathbb{R}$ with $(h, h_\circ) \mapsto \mathrm{cov}(h, h_\circ) := \mathbb{C}\mathrm{ov}(Y_h, Y_{h_\circ})$ is called *covariance function* of Y. If the covariance function is in addition a bounded bilinear form, then there is $\Gamma \in \mathbb{L}(\mathbb{H})$ such that $\mathrm{cov}(h, h_\circ) = \langle\Gamma h, h_\circ\rangle_\mathbb{H} = \langle h, \Gamma h_\circ\rangle_\mathbb{H}$ for all $h, h_\circ \in \mathbb{H}$. The operator Γ is called *covariance operator* of Y or $\mathbb{P}^Y$. If Y admits a mean function m and a covariance function cov then we write shortly $Y \sim \mathrm{P}_{(\mathrm{m},\mathrm{cov})}$. If there is a $\mathbb{H}$-mean $\theta = \mathbb{P}(Y) \in \mathbb{H}$ and a covariance operator $\Gamma \in \mathbb{L}(\mathbb{H})$ we write $Y \sim \mathrm{P}_{(\theta,\Gamma)}$, where for $h, h_\circ \in \mathbb{H}$ the covariance of Y_h and $Y_{h_\circ}$ equals $\Gamma_{h,h_\circ} := \langle h, \Gamma h_\circ\rangle_\mathbb{H}$, and $Y_h - \langle h, \theta\rangle_\mathbb{H}$ has mean zero and variance $\Gamma_{h,h}$, i.e. $Y_h - \langle h, \theta\rangle_\mathbb{H} \sim \mathrm{P}_{(0,\Gamma_{h,h})}$. □

§15.05 **Remark.** A covariance operator $\Gamma \in \mathbb{L}(\mathbb{H})$ associated with a stochastic process $Y \sim \mathbb{P}^Y$ on $\mathbb{H}$ is self-adjoint and non-negative definite, i.e. $\Gamma \in \mathbb{L}^+(\mathbb{H})$ (see Definition §17.28 (e)). □

§15.06 **Notation.** Given a measurable space $(\Omega, \mathscr{A}, \mu)$ introduce the μ-equivalence class $\{h\}_\mu := \{h_\circ \in \mathscr{A} : h = h_\circ\ \mu\text{-a.e.}\}$ for $h \in \mathscr{A}$. For $s \in \overline{\mathbb{R}}^+$ define the set of equivalence classes $\mathbb{L}_s(\mu) := \mathbb{L}_s(\mathscr{A}, \mu) := \{\{h\}_\mu : h \in \mathscr{L}_s(\mathscr{A}, \mu)\}$ and $\|\{h\}_\mu\|_{\mathbb{L}_s(\mu)} := \|h\|_{\mathscr{L}_s(\mu)}$ for $\{h\}_\mu \in \mathbb{L}_s(\mu)$. For $s \geqslant 1$, $(\mathbb{L}_s(\mu), \|\cdot\|_{\mathbb{L}_s(\mu)})$ is a normed vector space. Formally, we denote by $\{\bullet\}_\mu : \mathscr{L}_s(\mu) \to \mathbb{L}_s(\mu)$ the natural injection $h \mapsto \{h\}_\mu$. In case $s = 2$ the norm $\|\{h\}_\mu\|_{\mathbb{L}_2(\mu)} := \|h\|_{\mathscr{L}_2(\mu)} = (\mu(|h|^2)^{1/2}$ is induced by the inner product $(\{h\}_\mu, \{h_\circ\}_\mu) \mapsto \langle\{h\}_\mu, \{h_\circ\}_\mu\rangle_{\mathbb{L}_2(\mu)} := \mu(hh_\circ)$, and hence $(\mathbb{L}_2(\mu), \langle\cdot,\cdot\rangle_{\mathbb{L}_2(\mu)})$ is a Hilbert space. If $\lambda = \mu$ is the Lebesgue-measure then we write shortly $(\mathbb{L}_s, \langle\cdot,\cdot\rangle_{\mathbb{L}_s})$ and $\{\bullet\} : \mathscr{L}_s \to \mathbb{L}_s$. Similarly, given a set $\mathcal{D}_2 \subseteq \mathscr{L}_2 = \mathscr{L}_2(\mathscr{B}_{[0,1]}, \lambda_{[0,1]})$ of square-integrable Lebesgue densities on $([0,1], \mathscr{B}_{[0,1]})$, we write $\mathbb{D}_2 := \{\{\mathfrak{p}\}, \mathfrak{p} \in \mathcal{D}_2\} \subseteq \mathbb{L}_2 = \mathbb{L}_2(\mathscr{B}_{[0,1]}, \lambda_{[0,1]})$ for short.

§15.07 **Nonparametric density estimation** *(§15.01 continued).* Consider on $\mathscr{L}_2$ the stochastic process $\widehat{\mathfrak{p}} = (\widehat{\mathfrak{p}}_h)_{h\in\mathscr{L}_2}$ of real random variables defined on $([0,1]^n, \mathscr{B}^n_{[0,1]}, \mathbb{P}_\mathfrak{p}^{\otimes n})$ by $\widehat{\mathfrak{p}}_h := \widehat{\mathbb{P}}_n h \in \mathscr{B}^n_{[0,1]}$ for each $h \in \mathscr{L}_2$. We introduce a stochastic process $(\widehat{\mathfrak{p}}_{\{h\}})_{\{h\}\in\mathbb{L}_2}$ on $\mathbb{L}_2$ given by $\widehat{\mathfrak{p}}_{\{h\}} := \widehat{\mathfrak{p}}_h \in \mathscr{B}^n_{[0,1]}$ for $\{h\} \in \mathbb{L}_2$. Note that for each $h_\circ \in \{h\}$ we have $\widehat{\mathfrak{p}}_h = \widehat{\mathfrak{p}}_{h_\circ}$ λ^n-a.s. and thus also $\mathbb{P}_\mathfrak{p}^{\otimes n}$-a.s.. As usual we identify h with its equivalence class $\{h\}$ and write shortly $\widehat{\mathfrak{p}} = (\widehat{\mathfrak{p}}_h)_{h\in\mathbb{L}_2}$ with $\widehat{\mathfrak{p}}_h := \widehat{\mathbb{P}}_n h \in \mathscr{B}^n_{[0,1]}$ for each $h \in \mathbb{L}_2$. Meaning, that for each $h \in \mathbb{L}_2$ there is $h_\circ \in \{h\} \subseteq \mathscr{L}_2$

with $\widehat{\mathbb{p}}_h = \widehat{\mathbb{P}}_n h_o \in \mathscr{B}^n_{[0,1]}$, and hence $\widehat{\mathbb{p}}_h$ is unique only up to λ^n-a.s. equality. However, given the image probability measure $\mathbb{P}_{\mathbb{p}}^{\otimes n} \circ \widehat{\mathbb{p}}^{-1}$ for each $h \in \mathbb{L}_2$ we have $\widehat{\mathbb{p}}_h = \widehat{\mathbb{P}}h \in \mathscr{L}_1(\mathbb{P}_{\mathbb{p}}^{\otimes n})$ since $\mathscr{L}_2 \subseteq \mathscr{L}_1(\mathbb{P}_{\mathbb{p}})$ due to $\mathbb{p} \in \mathcal{D}_2 \subseteq \mathscr{L}_2$. As a consequence, $\widehat{\mathbb{p}} = (\widehat{\mathbb{p}}_h)_{h\in\mathbb{L}_2}$ admits a mean function $\mathrm{m}_{\mathbb{p}} : \mathbb{L}_2 \to \mathbb{R}$ satisfying $\mathrm{m}_{\mathbb{p}}(h) = \mathbb{P}_{\mathbb{p}}^{\otimes n}(\widehat{\mathbb{p}}_h) = \mathbb{P}_{\mathbb{p}}^{\otimes n}(\widehat{\mathbb{P}}h) = \mathbb{p}\lambda(h) = \langle \mathbb{p}, h\rangle_{\mathbb{L}_2} = \mathbb{p}_h$ for all $h \in \mathbb{L}_2$. Moreover, $\mathbb{p}$ (more precisely the λ-equivalence class $\{\mathbb{p}\}$) is the $\mathbb{L}_2$-*mean* of the stochastic process $\widehat{\mathbb{p}} = (\widehat{\mathbb{p}}_h)_{h\in\mathbb{L}_2}$. Introduce similarly the stochastic process $\dot{W} := (\dot{W}_h)_{h\in\mathbb{L}_2}$ on $\mathbb{L}_2$ given by $\dot{W}_h := n^{1/2}(\widehat{\mathbb{p}}_h - \langle \mathbb{p}, h\rangle_{\mathbb{L}_2}) \in \mathscr{B}^n_{[0,1]}$ for $h \in \mathbb{L}_2$, which allows us to write shortly $\widehat{\mathbb{p}} = \mathbb{p} + n^{-1/2}\dot{W}$, meaning that, $\widehat{\mathbb{p}}_h = \mathbb{p}_h + n^{-1/2}\dot{W}_h$ for all $h \in \mathbb{L}_2$. Since $\dot{W}_h \in \mathscr{L}_1(\mathbb{P}_{\mathbb{p}}^{\otimes n})$ has mean zero for each $h \in \mathbb{L}_2$, the $\mathbb{L}_2$-*mean* of $\dot{W}$equals zero. If in addition $\|\mathbb{p}\|_{\mathscr{L}_\infty} < \infty$, then we have $\mathbb{P}_{\mathbb{p}}(|h|^2) = \lambda(|h|^2\mathbb{p}) \leqslant \|\mathbb{p}\|_{\mathscr{L}_\infty}\|h\|^2_{\mathscr{L}_2} < \infty$ for all $h \in \mathscr{L}_2$ and thus $\mathscr{L}_2 \subseteq \mathscr{L}_2(\mathbb{P}_{\mathbb{p}})$ in equal. As a consequence for each $h \in \mathbb{L}_2$ we obtain $\widehat{\mathbb{p}}_h = \widehat{\mathbb{P}}h \in \mathscr{L}_2(\mathbb{P}_{\mathbb{p}}^{\otimes n})$ and, hence $\dot{W}_h \in \mathscr{L}_2(\mathbb{P}_{\mathbb{p}}^{\otimes n})$ by construction. The *covariance function* of $\dot{W} := (\dot{W}_h)_{h\in\mathbb{L}_2}$ is given by

$$\begin{aligned}(h, h_o) \mapsto \mathrm{cov}_{\mathbb{p}}(h, h_o) := \mathbb{C}\mathrm{ov}(\dot{W}_h, \dot{W}_{h_o}) &= \lambda(\mathbb{p}hh_o) - \lambda(\mathbb{p}h)\lambda(\mathbb{p}h_o)\\ &= \mathbb{P}_{\mathbb{p}}((h - \langle \mathbb{p}, h\rangle_{\mathbb{L}_2})(h_o - \langle \mathbb{p}, h_o\rangle_{\mathbb{L}_2})) = n\,\mathbb{C}\mathrm{ov}(\widehat{\mathbb{p}}_h, \widehat{\mathbb{p}}_{h_o}).\end{aligned}$$

Consequently, we have $\dot{W} \sim \mathrm{P}_{(0,\mathrm{cov}_{\mathbb{p}})}$ and $\widehat{\mathbb{p}} = \mathbb{p} + n^{-1/2}\dot{W} \sim \mathrm{P}_{(\mathrm{m}_{\mathbb{p}}, n^{-1}\,\mathrm{cov}_{\mathbb{p}})}$ Introduce the multiplication operator $\mathrm{M}_{\mathbb{p}} : \mathscr{B}_{[0,1]} \to \mathscr{B}_{[0,1]}$ given by $h \mapsto \mathrm{M}_{\mathbb{p}}(h) := h\mathbb{p}$. If $\|\mathbb{p}\|_{\mathscr{L}_\infty} < \infty$, then $\mathrm{M}_{\mathbb{p}} \in \mathbb{L}(\mathbb{L}_2)$ (see Example §17.21 (b)). This allows us to write $\lambda(\mathbb{p}hh_o) = \langle \mathrm{M}_{\mathbb{p}}h, h_o\rangle_{\mathbb{L}_2}$ for all $h, h_o \in \mathbb{L}_2$. Moreover, consider $\mathbb{1} := \mathbb{1}_{[0,1]} \in \mathscr{B}_{[0,1]}$ which trivially belongs to $\mathbb{L}_s$ for any $s \in \overline{\mathbb{R}}^+$. In particular, since $\mathbb{L}_2 \subseteq \mathbb{L}_1$ (indeed $\lambda(|h|) \leqslant \|\mathbb{1}\|_{\mathbb{L}_2}\|h\|_{\mathbb{L}_2} = \|h\|_{\mathbb{L}_2} < \infty$ for all $h \in \mathbb{L}_2$) we have $\langle h, \mathbb{1}\rangle_{\mathbb{L}_2} = \lambda(h)$ for all $h \in \mathbb{L}_2$ and $\mathbb{R}\mathbb{1} := \{a\mathbb{1}, a \in \mathbb{R}\} = \overline{\mathrm{lin}}\{\mathbb{1}\}$. Consider further the operator $\Pi_{\mathbb{R}\mathbb{1}} \in \mathbb{L}(\mathbb{L}_2)$ defined by $\Pi_{\mathbb{R}\mathbb{1}}h := \langle h, \mathbb{1}\rangle_{\mathbb{L}_2}\mathbb{1} = \lambda(h)\mathbb{1}$ for all $h \in \mathbb{L}_2$, which is an orthogonal projection (see Definition §17.28 (f) and Example §17.30 (a)). This allows us to write $\lambda(\mathbb{p}h)\lambda(\mathbb{p}h_o) = \langle \mathrm{M}_{\mathbb{p}}h, \mathbb{1}\rangle_{\mathbb{L}_2}\langle \mathbb{1}, \mathrm{M}_{\mathbb{p}}h_o\rangle_{\mathbb{L}_2} = \langle \Pi_{\mathbb{R}\mathbb{1}}\mathrm{M}_{\mathbb{p}}h, \mathrm{M}_{\mathbb{p}}h_o\rangle_{\mathbb{L}_2} = \langle \mathrm{M}_{\mathbb{p}}\Pi_{\mathbb{R}\mathbb{1}}\mathrm{M}_{\mathbb{p}}h, h_o\rangle_{\mathbb{L}_2}$. Summarising, if $\mathbb{p} \in \mathbb{D}_2 \cap \mathbb{L}_\infty$ then $\Gamma^{\mathbb{p}} := \mathrm{M}_{\mathbb{p}} - \mathrm{M}_{\mathbb{p}}\Pi_{\mathbb{R}\mathbb{1}}\mathrm{M}_{\mathbb{p}} \in \mathbb{L}^+(\mathbb{L}_2)$ is the *covariance operator* of $\dot{W}$, since $\mathrm{cov}_{\mathbb{p}}(h, h_o) = \langle \Gamma^{\mathbb{p}}h, h_o\rangle_{\mathbb{L}_2}$ for all $h, h_o \in \mathbb{L}_2$. We note that $\|\Gamma^{\mathbb{p}}\|_{\mathbb{L}(\mathbb{L}_2)} \leqslant \|\mathbb{p}\|_{\mathbb{L}_\infty}$ by using $\langle \Gamma^{\mathbb{p}}h, h\rangle_{\mathbb{L}_2} = \mathbb{p}\lambda((h - \mathbb{p}\lambda(h))^2) = \mathbb{p}\lambda(h^2) - (\mathbb{p}\lambda(h))^2 \leqslant \lambda(\mathbb{p}h^2) \leqslant \|\mathbb{p}\|_{\mathbb{L}_\infty}$ for all $h \in \mathbb{L}_2$ with $\|h\|^2_{\mathbb{L}_2} \leqslant 1$ together with Property §17.29 (i). Consequently, we have $\dot{W} \sim \mathrm{P}_{(0,\Gamma^{\mathbb{p}})}$ and $\widehat{\mathbb{p}} = \mathbb{p} + n^{-1/2}\dot{W} \sim \mathrm{P}_{(\mathbb{p}, n^{-1}\Gamma^{\mathbb{p}})}$. □

§15.08 **Nonparametric regression** *(§15.02 continued).* Consider the stochastic process $\widehat{f} = (\widehat{f}_h)_{h\in\mathbb{L}_2}$ on $\mathbb{L}_2$ of real valued random variables defined on $(\mathbb{R}^{2n}, \mathscr{B}^{2n}, \mathrm{U}_f^{\otimes n})$ by $\widehat{f}_h := \widehat{\mathbb{P}}(Yh(X)) \in \mathscr{B}^{2n}$ for each $h \in \mathbb{L}_2$. Here, we identify h again with its equivalence class $\{h\}$ as discussed in details in Example §15.07. Given the image probability measure $\mathrm{U}_f^{\otimes n} \circ \widehat{f}^{-1}$ for each $h \in \mathbb{L}_2$ we have $\widehat{f}_h = \widehat{\mathbb{P}}(Yh(X)) \in \mathscr{L}_1(\mathrm{U}_f^{\otimes n})$ by using $Yh(X) = (\varepsilon + f(X))h(X) \in \mathscr{L}_1(\mathrm{U}_f)$ under Assumption §14.01 (NPR1). Indeed, since $f \in \mathscr{L}_2 \subseteq \mathscr{L}_1$ we have $fh \in \mathbb{L}_1$ for each $h \in \mathbb{L}_2$ and $\mathrm{U}_f(|\varepsilon h(X)|) = \mathrm{U}_f(|\varepsilon|)\lambda(|h|) < \infty$ by Assumption §14.01 (NPR1). As a consequence $\widehat{f}$ admits a mean function $\mathrm{m}_f : \mathbb{L}_2 \to \mathbb{R}$ satisfying $\mathrm{m}_f(h) = \mathrm{U}_f^{\otimes n}(\widehat{f}_h) = \mathrm{U}_f^{\otimes n}(\widehat{\mathbb{P}}(Yh(X))) = \lambda(fh) = \langle f, h\rangle_{\mathbb{L}_2}$ for all $h \in \mathbb{L}_2$. Moreover, f (more precisely the λ-equivalence class $\{f\}$) is the $\mathbb{L}_2$-*mean* of the stochastic process $\widehat{f} = (\widehat{f}_h)_{h\in\mathbb{L}_2}$. Introduce similarly the stochastic process $\dot{W} := (\dot{W}_h)_{h\in\mathbb{L}_2}$ on $\mathbb{L}_2$ given by $\dot{W}_h := n^{1/2}(\widehat{f}_h - \langle f, h\rangle_{\mathbb{L}_2}) \in \mathscr{B}^{2n}$ for $h \in \mathbb{L}_2$, which allows us to write shortly $\widehat{f} = f + n^{-1/2}\dot{W}$, meaning that, $\widehat{f}_h = f_h + n^{-1/2}\dot{W}_h$ for all $h \in \mathbb{L}_2$. Since $\dot{W}_h \in \mathscr{L}_1(\mathrm{U}_f^{\otimes n})$ has mean zero for each $h \in \mathbb{L}_2$, the $\mathbb{L}_2$-*mean* of $\dot{W}$equals zero. If in addition

$\|f\|_{\mathscr{L}_\infty} < \infty$, then we have $Yh(X) = (\varepsilon + f(X))h(X) \in \mathscr{L}_2(\mathrm{U}_f)$ under Assumption §14.01 (NPR1). Indeed, we have $\mathrm{U}_f(f^2(X)h^2(X)) = \lambda(f^2h^2) \leqslant \|f\|^2_{\mathscr{L}_\infty}\|h\|^2_{\mathbb{L}_2} < \infty$ for all $h \in \mathbb{L}_2$ and $\mathrm{U}_f(\varepsilon^2h^2(X)) = \sigma^2_\varepsilon\|h\|^2_{\mathbb{L}_2} < \infty$ by Assumption §14.01 (NPR1). As a consequence for all $h \in \mathbb{L}_2$ we have $\widehat{f}_h = \widehat{\mathbb{P}}_n(Yh(X)) \in \mathscr{L}_2(\mathrm{U}_f^{\otimes n})$ and, hence $\dot{W}_h \in \mathscr{L}_2(\mathrm{U}_f^{\otimes n})$ by construction. The *covariance function* of $\dot{W} := (\dot{W}_h)_{h\in\mathbb{L}_2}$ is under Assumption §14.01 given for all $h, h_o \in \mathbb{L}_2$ by

$$\begin{aligned}\mathrm{cov}_f(h, h_o) := \mathbb{C}\mathrm{ov}(\dot{W}_h, \dot{W}_{h_o}) &= \mathrm{U}_f(Y^2h(X)h_o(X)) - \mathrm{U}_f(Yh(X))\mathrm{U}_f(Yh_o(X)) \\ &= \sigma^2_\varepsilon\langle h, h_o\rangle^2_{\mathbb{L}_2} + \langle fh, fh_o\rangle_{\mathbb{L}_2} - \langle f, h\rangle_{\mathbb{L}_2}\langle f, h_o\rangle_{\mathbb{L}_2} = n\,\mathbb{C}\mathrm{ov}(\widehat{f}_h, \widehat{f}_{h_o}).\end{aligned}$$

Consequently, if $\|f\|_{\mathscr{L}_\infty} < \infty$, then we have $\dot{W} \sim \mathrm{P}_{(0,\mathrm{cov}_f)}$ and $\widehat{f} = f + n^{-1/2}\dot{W} \sim \mathrm{P}_{(\mathrm{m}_f, n^{-1}\,\mathrm{cov}_f)}$. Moreover, as in Example §15.07 both the multiplication operator M_f due to $\|f\|_{\mathscr{L}_\infty} < \infty$ and the projection operator $\Pi_{\mathbb{1}}$ belong to $\mathbb{L}(\mathbb{L}_2)$. Furthermore, introducing the orthogonal projection $\Pi^\perp_{\mathbb{1}} := \mathrm{id}_{\mathbb{L}_2} - \Pi_{\mathbb{1}} \in \mathbb{L}(\mathbb{L}_2)$ allows us to write

$$\begin{aligned}\mathrm{cov}_f(h, h_o) &= \sigma^2_\varepsilon\langle h, h_o\rangle^2_{\mathbb{L}_2} + \langle \mathrm{M}_f h, \mathrm{M}_f h_o\rangle_{\mathbb{L}_2} - \langle \Pi_{\mathbb{1}}\mathrm{M}_f h, \mathrm{M}_f h_o\rangle_{\mathbb{L}_2} \\ &= \langle \sigma^2_\varepsilon \mathrm{id}_{\mathbb{L}_2} h, h_o\rangle^2_{\mathbb{L}_2} + \langle \Pi^\perp_{\mathbb{1}}\mathrm{M}_f h, \mathrm{M}_f h_o\rangle_{\mathbb{L}_2} = \langle(\sigma^2_\varepsilon \mathrm{id}_{\mathbb{L}_2} + \mathrm{M}_f\Pi^\perp_{\mathbb{1}}\mathrm{M}_f)h, h_o\rangle^2_{\mathbb{L}_2}.\end{aligned}$$

Summarising, under Assumption §14.01 if $f \in \mathbb{L}_2 \cap \mathbb{L}_\infty$ then $\Gamma^f := \sigma^2_\varepsilon \mathrm{id}_{\mathbb{L}_2} + \mathrm{M}_f\Pi^\perp_{\mathbb{1}}\mathrm{M}_f \in \mathbb{L}^+(\mathbb{L}_2)$ is the *covariance operator* of $\dot{W}$, since $\mathrm{cov}_f(h, h_o) = \langle \Gamma^f h, h_o\rangle_{\mathbb{L}_2}$ for all $h, h_o \in \mathbb{L}_2$. We note that $\|\Gamma^f\|_{\mathbb{L}(\mathbb{L}_2)} \leqslant \sigma^2_\varepsilon + \|f\|^2_{\mathbb{L}_\infty}$ by using $\langle \Gamma^f h, h\rangle_{\mathbb{L}_2} = \sigma^2_\varepsilon\|h\|^2_{\mathbb{L}_2} + \|\Pi^\perp_{\mathbb{1}}\mathrm{M}_f h\|^2_{\mathbb{L}_2} \leqslant \sigma^2_\varepsilon + \|\mathrm{M}_f\|^2_{\mathbb{L}(\mathbb{L}_2)}$ for all $h \in \mathbb{L}_2$ with $\|h\|_{\mathbb{L}_2} \leqslant 1$, and $\|\mathrm{M}_f\|_{\mathbb{L}(\mathbb{L}_2)} \leqslant \|f\|_{\mathbb{L}_\infty}$ (see Example §17.21 (b)) together with Property §17.29 (i). Consequently, we have $\dot{W} \sim \mathrm{P}_{(0,\Gamma^f)}$ and $\widehat{f} = f + n^{-1/2}\dot{W} \sim \mathrm{P}_{(f, n^{-1}\Gamma^f)}$. □

§15.09 **Notation.** In the sequel we exploit the Hilbert space structure of $\mathbb{L}_2$ which guarantees the existence of an orthonormal basis (*ONB*) $\mathcal{U} \subseteq \mathbb{L}_2$. An ONB is an orthonormal system (ONS) which is complete (see Definition §17.04). Moreover, since $\mathbb{L}_2$ is separable, any ONS is countable, and thus there is an orthonormal sequence (*ONS*) $u_\bullet = (u_j)_{j\in\mathbb{N}}$ in $\mathbb{L}_2$ (see Definition §17.09). Rather than a stochastic process $Y := (Y_h)_{h\in\mathbb{L}_2}$ on $\mathbb{L}_2$ we consider its canonical projection $Y_{u_\bullet} = (Y_{u_j})_{j\in\mathbb{N}} := \Pi_{\mathcal{U}}Y$ on $\mathcal{U} = \{u_j, j \in \mathbb{N}\}$. □

§15.10 **White noise process on $\mathbb{H}$.** Let $Y := (Y_h)_{h\in\mathbb{H}}$ be a stochastic process on $\mathbb{H}$. For an ONS $u_\bullet = (u_j)_{j\in\mathbb{N}}$ in $\mathbb{H}$ we call the canonical projection $Y_{u_\bullet} = (Y_{u_j})_{j\in\mathbb{N}}$ a *white noise process*, if $\{Y_{u_j}, j \in \mathbb{N}\}$ is a family of independent and identically distributed random variables, where each Y_{u_j} has zero mean and variance one, $Y_{u_j} \sim \mathrm{P}_{(0,1)}$ and $Y_{u_\bullet} \sim \mathrm{P}^{\otimes\mathbb{N}}_{(0,1)}$ in short. We call Y a *white noise process* on $\mathbb{H}$, if $Y_{u_\bullet} \sim \mathrm{P}^{\otimes\mathbb{N}}_{(0,1)}$ yields for any ONS $u_\bullet$ in $\mathbb{H}$. □

§15.11 **Notation.** In other words, the distribution $\mathbb{P}^{Y_{u_\bullet}}$ of a white noise process $Y_{u_\bullet} = (Y_{u_j})_{j\in\mathbb{N}}$ equals the product of its marginal $\mathrm{P}_{(0,1)}$-distributions, i.e. $\mathbb{P}^{Y_{u_\bullet}} = \otimes_{j\in\mathbb{N}}\mathbb{P}^{Y_{u_j}} = \otimes_{j\in\mathbb{N}}\mathrm{P}_{(0,1)} = \mathrm{P}^{\otimes\mathbb{N}}_{(0,1)}$. □

§15.12 **Remark.** Consider the centred stochastic process $\dot{W} := (\dot{W}_h)_{h\in\mathbb{L}_2}$ of error terms in the Examples §15.07 and §15.08. In general there does not exists an ONB $u_\bullet = (u_j)_{j\in\mathbb{N}}$ in $\mathbb{L}_2$ such that the canonical projection $Y_{u_\bullet}$ is a white noise process. □

§15.13 **Property.** *Let $Y := (Y_h)_{h\in\mathbb{H}} \sim \mathbb{P}^Y$ be a stochastic process on $\mathbb{H}$ admitting a $\mathbb{H}$-mean $\theta \in \mathbb{H}$ and a covariance operator $\Gamma \in \mathbb{L}^+(\mathbb{H})$, i.e., $Y \sim \mathrm{P}_{(\theta,\Gamma)}$. If there exists an ONB $u_\bullet = (u_j)_{j\in\mathbb{N}}$ in $\mathbb{H}$ such that $Y_{u_\bullet}$ is a white noise process, i.e. $Y_{u_\bullet} \sim \mathrm{P}^{\otimes\mathbb{N}}_{(0,1)}$. Then we have $\theta = \sum_{j\in\mathbb{N}}\langle\theta, u_j\rangle_{\mathbb{H}}u_j = \sum_{j\in\mathbb{N}}\mathbb{P}(Y_{u_j})u_j = 0$ and $\langle\Gamma h, h_o\rangle_{\mathbb{H}} = \sum_{j,j_o\in\mathbb{N}}\langle u_j, h\rangle_{\mathbb{H}}\langle\Gamma u_j, u_{j_o}\rangle_{\mathbb{H}}\langle u_{j_o}, h_o\rangle_{\mathbb{H}} = \langle h, h_o\rangle_{\mathbb{H}}$, and thus*

$\theta = 0 \in \mathbb{H}$ *and* $\Gamma = \mathrm{id}_{\mathbb{H}}$. *As a consequence, for each ONS* $\mathcal{V}$ *in* $\mathbb{H}$ *the random variables* $\{Y_v, v \in \mathcal{V}\}$ *are pairwise uncorrelated.* □

§15.14 **Gaussian process on $\mathbb{H}$.** A stochastic process $Y = (Y_h)_{h\in\mathbb{H}}$ on $\mathbb{H}$ with mean function m and covariance function cov is called a *Gaussian process* on $\mathbb{H}$, if the family of finite-dimensional distributions $(\mathbb{P}^{Y_{\mathcal{U}}})_{\mathcal{U}\subseteq\mathbb{H}\text{ finite}}$ consists of normal distributions, that is, $Y_{\mathcal{U}} = (Y_u)_{u\in\mathcal{U}}$ is normally distributed with mean vector $(\mathrm{m}(u))_{u\in\mathcal{U}}$ and covariance matrix $(\mathrm{cov}(u,u'))_{u,u'\in\mathcal{U}}$. We write shortly $Y \sim \mathrm{N}_{(\mathrm{m},\mathrm{cov})}$ or $Y \sim \mathrm{N}_{(\theta,\Gamma)}$, if in addition there exist a $\mathbb{H}$-mean $\theta \in \mathbb{H}$ and a covariance operator $\Gamma \in \mathbb{L}^+(\mathbb{H})$ associated with Y. The Gaussian process $Y \sim \mathrm{N}_{(0,\mathrm{id}_{\mathbb{H}})}$ with $\mathbb{H}$-mean zero and covariance operator $\mathrm{id}_{\mathbb{H}}$ is called *iso-Gaussian process* or *Gaussian white noise process* on $\mathbb{H}$. □

§15.15 **Property.** *Let* $Y \sim \mathrm{N}_{(\theta,\Gamma)}$ *be a Gaussian process on* $\mathbb{H}$ *admitting a* $\mathbb{H}$*-mean* $\theta \in \mathbb{H}$ *and a covariance operator* $\Gamma \in \mathbb{L}^+(\mathbb{H})$. *If there exists an ONB* $u_\bullet = (u_j)_{j\in\mathbb{N}}$ *in* $\mathbb{H}$ *such that* $Y_{u_\bullet} = (Y_{u_j})_{j\in\mathbb{N}}$ *is a Gaussian white noise process, i.e.,* $Y_{u_\bullet} \sim \mathrm{N}_{(0,1)}^{\otimes\mathbb{N}}$, *then due to Property §15.13 we have* $Y \sim \mathrm{N}_{(0,\mathrm{id}_{\mathbb{H}})}$ *and for each ONS* $\mathcal{V}$ *in* $\mathbb{H}$ *the standard normally distributed random variables* $\{Y_v, v \in \mathcal{V}\}$ *are pairwise uncorrelated, and hence, independent, i.e.,* $\Pi_{\mathcal{V}} Y \sim \mathrm{N}_{(0,1)}^{\otimes\mathcal{V}}$. □

§15.16 **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 random variable or a *random function* in $\mathbb{H}$. □

§15.17 **Lemma.** *Let* $u_\bullet = (u_j)_{j\in\mathbb{N}}$ *be an ONS in* $\mathbb{H}$. *There does not exist a non-zero random function* Y *in* $\mathbb{H}$ *such that* $Y_{u_\bullet} = (Y_{u_j} = \langle Y, u_j\rangle_{\mathbb{H}})_{j\in\mathbb{N}}$ *is a Gaussian white noise process.*

§15.18 **Proof** of Lemma §15.17. For $j \in \mathbb{N}$ and $r > 0$ define $\mathcal{A}_j^r := \{h \in \mathbb{H} : |\langle h, u_j\rangle_{\mathbb{H}}| \leqslant r\}$, and $\mathcal{A}_\infty^r = \cap\{\mathcal{A}_j^r, j \in \mathbb{N}\}$. Obviously, it holds $\mathbb{H} = \lim_{r\to\infty} \mathcal{A}_\infty^r$ and hence, $1 = \mathbb{P}^Y(\mathbb{H}) = \lim_{r\to\infty} \mathbb{P}^Y(\mathcal{A}_\infty^r)$ for each random function Y in $\mathbb{H}$. Assume that there is a Gaussian white noise process $\Pi_u Y$, then for each $n \in \mathbb{N}$ it holds $\mathbb{P}^Y(\mathcal{A}_\infty^r) \leqslant \mathbb{P}^Y(\cap\{\mathcal{A}_j^r, j \in [\![1,n]\!]\}) = |\mathbb{P}^{Y_{u_1}}(\mathcal{A}_1^r)|^n = |\mathbb{P}(|Z| \leqslant r)|^n$ where $Z \sim \mathrm{N}_{(0,1)}$. Thereby, as $n \to \infty$ we get $\mathbb{P}^Y(\mathcal{A}_\infty^r) = 0$ for all $r > 0$ and hence it follows the contradiction $\mathbb{P}^Y(\mathbb{H}) = 0$, which completes the proof. □

Sequence space model.

Given a pre-specified ONS $u_\bullet = (u_j)_{j\in\mathbb{N}}$ in $\mathbb{H}$ we base our estimation procedure on the expansion of the function of interest $\theta \in \mathbb{U} = \overline{\mathrm{lin}}(u_j, j \in \mathbb{N})$. More precisely, we consider the sequence of *generalised Fourier coefficients* $U\theta = \theta_{u_\bullet} = (\theta_{u_j} = \langle\theta, u_j\rangle_{\mathbb{H}})_{j\in\mathbb{N}}$ which allow to reconstruct $\theta = U^\star\theta_{u_\bullet} = \sum_{j\in\mathbb{N}} \theta_{u_j} u_j$ (see Example §17.30 (a)). The choice of an adequate ONS $u_\bullet = (u_j)_{j\in\mathbb{N}}$ is determined by the presumed information on the function of interest θ formalised by an abstract smoothness conditions given in Definition §16.15. 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.

§15.19 **Notation.** Given a pre-specified ONS $u_\bullet = (u_j)_{j\in\mathbb{N}}$ in $\mathbb{H}$ and a stochastic process $Y \sim \mathrm{P}_{(\theta,\Gamma)}$ on $\mathbb{H}$ admitting a $\mathbb{H}$-mean $\theta \in \mathbb{H}$ and a covariance operator $\Gamma \in \mathbb{L}^+(\mathbb{H})$, the canonical projection $Y_{u_\bullet} = (Y_{u_j})_{j\in\mathbb{N}}$ admits a mean sequence $\theta_{u_\bullet} = (\theta_{u_j})_{j\in\mathbb{N}} \in \ell_2$ and a covariance operator $\Gamma_{u_\bullet,u_\bullet} \in \mathbb{L}^+(\ell_2)$ with (infinite) matrix representation $\Gamma_{u_\bullet,u_\bullet} = (\Gamma_{u_k,u_j})_{j,l\in\mathbb{N}} \in \mathbb{R}^{(\mathbb{N},\mathbb{N})}$ having generic entries $\Gamma_{u_k,u_j} := \langle u_k, \Gamma u_j\rangle_{\mathbb{H}} = \mathbb{C}\mathrm{ov}(Y_{u_k}, Y_{u_j})$ for $k, j \in \mathbb{N}$. By construction for each $a_\bullet, a_\bullet^o \in \ell_2$, and

hence $h := U^\star a_\bullet, h^o := U^\star a_\bullet^o \in \mathbb{U}$ we have $\Gamma_{u,u_\bullet} a_\bullet = (\Gamma_{u_j,h})_{j\in\mathbb{N}} = (\sum_{k\in\mathbb{N}} \Gamma_{u_j,u_k} a_k)_{j\in\mathbb{N}} \in \ell_2$ and $\langle b_\bullet, \Gamma_{u,u_\bullet} a_\bullet\rangle_{\ell_2} = \Gamma_{h,h^o}$. □

§15.20 **Sequence space model (SSM).** Let $\dot{W} = (\dot{W}_h)_{h\in\mathbb{H}}$ be a stochastic process on $\mathbb{H}$ with $\mathbb{H}$-mean zero and let $n \in \mathbb{N}$ be a sample size. The stochastic process $\widehat{\theta} = \theta + n^{-1/2}\dot{W}$ on $\mathbb{H}$ with $\mathbb{H}$-mean $\theta \in \mathbb{H}$ is called a *noisy version* of θ. We denote by $\mathbb{P}_\theta^n$ the distribution of $\widehat{\theta}$. If $\dot{W}$ admits a covariance operator (possibly depending on θ), say $\Gamma^\theta \in \mathbb{L}^+(\mathbb{H})$, then we eventually write $\widehat{\theta} \sim \mathrm{P}_{(\theta,n^{-1}\Gamma^\theta)}$ for short. Given an ONS $u_\bullet = (u_j)_{j\in\mathbb{N}}$ in $\mathbb{H}$ the canonical projection $\widehat{\theta}_{u_\bullet} = (\widehat{\theta}_{u_j})_{j\in\mathbb{N}}$, a sequence of observable quantities, takes the form of a *sequence space model (SSM)*

$$\widehat{\theta}_{u_j} = \langle \theta, u_j\rangle_{\mathbb{H}} + n^{-1/2}\dot{W}_{u_j} = \theta_{u_j} + n^{-1/2}\dot{W}_{u_j}, \quad j \in \mathbb{N}. \tag{15.1}$$

We denote by $\mathbb{P}^n_{\theta_{u_\bullet}}$, respectively $\mathrm{P}_{(\theta_{u_\bullet},n^{-1}\Gamma^\theta_{u_\bullet,u_\bullet})}$, the distribution of the stochastic process $\widehat{\theta}_{u_\bullet}$ which is determined by the distribution $\mathbb{P}_\theta^n$, respectively $\mathrm{P}_{(\theta,n^{-1}\Gamma^\theta)}$, of the noisy version $\widehat{\theta}$. □

§15.21 **Gaussian sequence space model.** Consider a separable real Hilbert space $(\mathbb{H}, \langle\cdot,\cdot\rangle_{\mathbb{H}})$. The parameter of interest $\theta \in \mathbb{H}$ is uniquely determined by the family $(\theta_h := \langle\theta, h\rangle_{\mathbb{H}})_{h\in\mathbb{H}}$. Let $\dot{W} := (\dot{W}_h)_{h\in\mathbb{H}} \sim \mathrm{N}_{(0,\mathrm{id}_{\mathbb{H}})}$ be a Gaussian white noise process on $\mathbb{H}$. The observable stochastic process $\widehat{\theta} := (\widehat{\theta}_h)_{h\in\mathbb{H}}$ on $\mathbb{H}$ satisfies $\widehat{\theta} = \theta + n^{-1/2}\dot{W}$, meaning that, $\widehat{\theta}_h = \theta_h + n^{-1/2}\dot{W}_h \sim \mathrm{N}_{(\theta_h, n^{-1}\|h\|^2_{\mathbb{H}})}$ for all $h \in \mathbb{H}$. In other words $\widehat{\theta} \sim \mathrm{N}_{(\theta,n^{-1}\mathrm{id}_{\mathbb{H}})}$ is a Gaussian process on $\mathbb{H}$ with $\mathbb{H}$-mean θ and covariance operator $n^{-1}\mathrm{id}_{\mathbb{H}}$. Given an ONS $u_\bullet = (u_j)_{j\in\mathbb{N}}$ in $\mathbb{H}$ the canonical projection $\widehat{\theta}_{u_\bullet} = (\widehat{\theta}_{u_j})_{j\in\mathbb{N}}$ takes the form of a *Gaussian sequence space model (SSM)*

$$\widehat{\theta}_{u_j} = \langle\theta, u_j\rangle_{\mathbb{H}} + n^{-1/2}\dot{W}_{u_j} = \theta_{u_j} + n^{-1/2}\dot{W}_{u_j},\ j \in \mathbb{N} \quad \text{with} \quad \{\dot{W}_{u_j}, i \in \mathbb{N}\} \overset{i.i.d.}{\sim} \mathrm{N}_{(0,1)}. \tag{15.2}$$

We denote by $\mathrm{N}_{(\theta_{u_\bullet},n^{-1}\mathrm{id}_{\ell_2})}$ the distribution of the stochastic process $\widehat{\theta}_{u_\bullet}$ which is determined by the distribution $\mathrm{N}_{(\theta,n^{-1}\mathrm{id}_{\mathbb{H}})}$ of the noisy version $\widehat{\theta}$. □

§16 Orthogonal series estimation

Here and subsequently, $u_\bullet = (u_j)_{j\in\mathbb{N}}$ denotes a pre-specified orthonormal sequence in a separable Hilbert space $(\mathbb{H}, \langle\cdot,\cdot\rangle_{\mathbb{H}})$. Given a noisy version $\widehat{\theta} = \theta + n^{-1/2}\dot{W}$ of a function of interest $\theta \in \mathbb{H}$ as in Definition §15.20 we study an estimator using a dimension reduction. To be more precise, given a nested sieve $(\llbracket m\rrbracket)_{m\in\mathcal{M}}$, $\mathcal{M} \subseteq \mathbb{N}$, as in Definition §17.11 we introduce a sequence of approximation spaces $(\mathbb{U}_m := \overline{\mathrm{lin}}(u_j, j \in \llbracket m\rrbracket))_{m\in\mathcal{M}}$ which form a nested sieve in $\mathbb{U} = \overline{\mathrm{lin}}(u_j, j \in \mathbb{N})$. If $\theta \in \mathbb{U}$, which is assumed from here on, then θ permits an expansion as *generalised Fourier series* $\theta = U^\star\theta_{u_\bullet} = \sum_{j\in\mathbb{N}} \theta_{u_j} u_j$ (see Example §17.30 (a)). For $m \in \mathcal{M}$ we approximate θ by its orthogonal projection $\theta^m := \Pi_{\mathbb{U}_m}\theta$ onto $\mathbb{U}_m$ (see Definition §17.28 (f)). Introducing the sequence of indicators $\mathbb{1}^m_\bullet := (\mathbb{1}^m_j)_{j\in\mathbb{N}}$ with $\mathbb{1}^m_j := \mathbb{1}_{\llbracket m\rrbracket}(j)$ for $j \in \mathbb{N}$ we have the identities $\theta^m = \sum_{j\in\llbracket m\rrbracket} \theta_{u_j} u_j = \sum_{j\in\mathbb{N}} \mathbb{1}^m_j \theta_{u_j} u_j = U^\star(\theta_{u_\bullet}\mathbb{1}^m_\bullet)$. Given the noisy version $\widehat{\theta}$ we replace the unknown sequence $\theta_{u_\bullet}$ of generalised Fourier coefficients by the canonical projection $\widehat{\theta}_{u_\bullet} = (\widehat{\theta}_{u_j})_{j\in\mathbb{N}}$ obeying a SSM as in Definition §15.20 (15.1).

§16.01 **Definition.** Given a SSM $\widehat{\theta}_{u_\bullet} \sim \mathbb{P}^n_{\theta_{u_\bullet}}$ as in (15.1) we call $\widehat{\theta}^m := U^\star(\widehat{\theta}_{u_\bullet}\mathbb{1}^m_\bullet)$ *orthogonal series estimator (OSE)* of θ for each $m \in \mathcal{M}$. □

Global measure of accuracy.

We shall measure the accuracy of the OSE $\widehat{\theta}^m = U^\star(\widehat{\theta}_{u_\bullet}\mathbb{1}^m_\bullet)$ of θ first by a global risk with respect to the distribution $\mathbb{P}^n_\theta$ of the noisy version $\widehat{\theta}$.

§16.02 **Definition**. Given a noisy version $\widehat{\theta} \sim \mathbb{P}^n_\theta$ of θ admitting a $\mathbb{H}$-mean $\theta \in \mathbb{H}$ the *global $\mathbb{H}$-risk* of a OSE $\widehat{\theta}^m = U^\star(\widehat{\theta}_{u_\bullet}\mathbb{1}^m_\bullet)$ satisfies

$$\mathbb{P}^n_\theta\|\widehat{\theta}^m - \theta\|^2_{\mathbb{H}} = \mathbb{var}_\theta + \mathbb{bias}^2_\theta$$

by introducing a *variance* term $\mathbb{var}_\theta := \mathbb{P}^n_\theta\|\widehat{\theta}^m - \theta^m\|^2_{\mathbb{H}}$ and a *bias* term $\mathbb{bias}_\theta := \|\theta^m - \theta\|_{\mathbb{H}}$. □

In the sequel we analyse separately the variance and bias term. We consider a noisy version $\widehat{\theta} = \theta + n^{-1/2}\dot{W} \sim \mathbb{P}^n_\theta$ of θ admitting a $\mathbb{H}$-mean $\theta \in \mathbb{H}$ and a covariance operator $n^{-1}\Gamma^\theta \in \mathbb{L}^+(\mathbb{H})$, i.e. $\widehat{\theta} \sim \mathrm{P}_{(\theta,n^{-1}\Gamma^\theta)}$, as in Definition §15.20, where the error process $\dot{W} \sim \mathrm{P}_{(0,\Gamma^\theta)}$ has $\mathbb{H}$-mean zero and covariance operator $\Gamma^\theta \in \mathbb{L}^+(\mathbb{H})$ (see Definition §15.04). As a consequence, given an ONS $u_\bullet = (u_j)_{j\in\mathbb{N}}$ in $\mathbb{H}$ for each $m \in \mathcal{M}$ the $\mathbb{R}^m$-valued random vector $(\dot{W}_{u_j})_{j\in[\![m]\!]}$ has mean zero and covariance matrix $(\Gamma^\theta_{u_k,u_j})_{k,j\in[\![m]\!]} \in \mathbb{R}^{(m,m)}$ (Notation §15.19).

§16.03 **Notation**. Let $m \in \mathbb{N}$. For $a_\bullet \in \mathbb{R}^{\mathbb{N}}$ and $T_{\bullet,\bullet} = (T_{k,j})_{k,j\in\mathbb{N}} \in \mathbb{R}^{(\mathbb{N},\mathbb{N})}$ we introduce, respectively, its sub-vector $[a_\bullet]_{\underline{m}} := (a_j)_{j\in[\![m]\!]} \in \mathbb{R}^m$ and its sub-matrix $[T_{\bullet,\bullet}]_{\underline{m}} := (T_{k,j})_{k,j\in[\![m]\!]} \in \mathbb{R}^{(m,m)}$. We denote the trace of $[T_{\bullet,\bullet}]_{\underline{m}}$ by $\mathrm{tr}([T_{\bullet,\bullet}]_{\underline{m}}) := \sum_{j\in[\![m]\!]} T_{j,j}$ and for $a_\bullet \in \mathbb{R}^{\mathbb{N}}$ with minimal value in $B \subseteq \mathbb{N}$ we define $\arg\min\{a_n, n \in B\} := \min\{m \in B : a_m \leqslant a_n,\ \forall n \in B\}$. □

§16.04 **Property**. *Let $\widehat{\theta} = \theta + n^{-1/2}\dot{W} \sim \mathbb{P}^n_\theta = \mathrm{P}_{(\theta,n^{-1}\Gamma^\theta)}$, then for all $m \in \mathcal{M}$ we have*

$$\mathbb{var}_\theta = \mathbb{P}^n_\theta\|U^\star(\mathbb{1}^m_\bullet(\widehat{\theta}_{u_\bullet} - \theta_{u_\bullet}))\|^2_{\mathbb{H}} = n^{-1}\mathbb{P}^n_\theta\|\mathbb{1}^m_\bullet\dot{W}_{u_\bullet}\|^2_{\ell_2} = n^{-1}\,\mathrm{tr}([\Gamma^\theta_{u_\bullet,u_\bullet}]_{\underline{m}}).$$ □

§16.05 **Definition**. Let $C \in \mathbb{R}^+_{\setminus 0}$ and for each $n \in \mathbb{N}$ let $\mathrm{R}^\circ_n \in \mathbb{R}^+_{\setminus 0}$ and $m^\circ_n \in \mathbb{N}$ satisfy

$$C^{-1}\,\mathrm{R}^\circ_n \leqslant \inf_{m\in\mathcal{M}} \mathbb{E}^n_\theta\|\widehat{\theta}^m - \theta\|^2_{\mathbb{H}} \leqslant \mathbb{E}^n_\theta\|\widehat{\theta}^{m^\circ_n} - \theta\|^2_{\mathbb{H}} \leqslant C\,\mathrm{R}^\circ_n.$$

Then we call R°_n *oracle bound*, m°_n *oracle dimension* and $\widehat{\theta}^{m^\circ_n}$ *oracle optimal* (up to the constant C). As a consequence, up to the constant C^2 the statistik $\widehat{\theta}^{m^\circ_n}$ attains the lower $\mathbb{H}$-risk bound within the family of OSE's, that is, $\mathbb{E}^n_\theta\|\widehat{\theta}^{m^\circ_n} - \theta\|^2_{\mathbb{H}} \leqslant C^2 \inf_{m\in\mathcal{M}} \mathbb{E}^n_\theta\|\widehat{\theta}^m - \theta\|^2_{\mathbb{H}}$. □

§16.06 **Oracle inequality**. *If $\widehat{\theta} \sim \mathbb{P}^n_\theta = \mathrm{P}_{(\theta,n^{-1}\Gamma^\theta)}$ then setting for $n \in \mathbb{N}$ and $m \in \mathcal{M}$*

$$\mathrm{R}^m_n(\theta) := [\|\theta^m - \theta\|^2_{\mathbb{H}} \vee n^{-1}\,\mathrm{tr}([\Gamma^\theta_{u_\bullet,u_\bullet}]_{\underline{m}})], \quad m^\circ_n := m^\circ_n(\theta) := \arg\min\{\mathrm{R}^m_n(\theta), m \in \mathcal{M}\}$$
$$\text{and} \quad \mathrm{R}^\circ_n(\theta) := \mathrm{R}^{m^\circ_n}_n(\theta) = \min\{\mathrm{R}^m_n(\theta), m \in \mathcal{M}\}$$

we have $\mathrm{R}^m_n(\theta) \leqslant \mathbb{P}^n_\theta\|\widehat{\theta}^m - \theta\|^2_{\mathbb{H}} = n^{-1}\,\mathrm{tr}([\Gamma^\theta_{u_\bullet,u_\bullet}]_{\underline{m}}) + \|\theta^m - \theta\|^2_{\mathbb{H}} \leqslant 2\mathrm{R}^m_n(\theta)$ for all $m \in \mathcal{M}$ and $n \in \mathbb{N}$. As a consequence we immediately obtain the following oracle inequality

$$\mathrm{R}^\circ_n(\theta) \leqslant \inf_{m\in\mathcal{M}} \mathbb{P}^n_\theta\|\widehat{\theta}^m - \theta\|^2_{\mathbb{H}} \leqslant \mathbb{P}^n_\theta\|\widehat{\theta}^{m^\circ_n} - \theta\|^2_{\mathbb{H}} \leqslant 2\mathrm{R}^\circ_n(\theta) \leqslant 2\inf_{m\in\mathcal{M}} \mathbb{P}^n_\theta\|\widehat{\theta}^m - \theta\|^2_{\mathbb{H}},$$

and hence, $\mathrm{R}^\circ_n(\theta)$, m°_n and the statistic $\widehat{\theta}^{m^\circ_n}$, respectively, is an oracle bound, an oracle dimension and oracle optimal (up to the constant 2).

§16.07 **Remark**. We shall emphasise that for each $m \in \mathcal{M}$ we have $n^{-1}\operatorname{tr}([\Gamma^\theta_{u_\bullet,u_\bullet}]_{\underline{m}}) = \mathrm{o}(1)$ as $n \to \infty$. As a consequence, if $\|\theta^m - \theta\|^2_{\mathbb{H}} = \mathrm{o}(1)$ as $m \to \infty$ then we obtain $\mathrm{R}^\circ_n(\theta) = \mathrm{o}(1)$ as $n \to \infty$, and thus, $\mathrm{R}^\circ_n(\theta)$ is also called an *oracle rate*. Indeed, for all $\delta \in \mathbb{R}^+_{\setminus 0}$ there exists $m_\delta \in \mathcal{M}$ and $n_\delta \in \mathbb{N}$ such that we have both $\|\theta^{m_\delta} - \theta\|^2_{\mathbb{H}} \leqslant \delta$ and $n^{-1}\operatorname{tr}([\Gamma^\theta_{u_\bullet,u_\bullet}]_{\underline{m}_\delta}) \leqslant \delta$ for all $n \geqslant n_\delta$, and whence $\mathrm{R}^\circ_n(\theta) \leqslant \mathrm{R}^{m_\delta}_n(\theta) \leqslant \delta$. However, note that the oracle dimension $m^\circ_n = m^\circ_n(\theta)$ as defined in Property §16.06 depends on the unknown parameter of interest θ, and thus also the oracle optimal statistic $\widehat{\theta}^{m^\circ_n}$. In other words $\widehat{\theta}^{m^\circ_n}$ is not a feasible estimator. □

§16.08 **Oracle inequality**. *If* $\widehat{\theta} \sim \mathbb{P}^n_\theta = \mathrm{P}_{(\theta, n^{-1}\Gamma^\theta)}$ *and if in addition there exists* $\mathbb{v}_\theta \in \mathbb{R}^+_{\setminus 0}$ *satisfying*

$$\forall h \in \mathbb{U}: \quad \mathbb{v}_\theta^{-1}\|h\|^2_{\mathbb{H}} \leqslant \langle h, \Gamma^\theta h\rangle_{\mathbb{H}} \leqslant \mathbb{v}_\theta\|h\|^2_{\mathbb{H}} \tag{16.1}$$

then setting for $n \in \mathbb{N}$ *and* $m \in \mathcal{M}$

$$\mathrm{R}^m_n(\theta) := [\|\theta^m - \theta\|^2_{\mathbb{H}} \vee n^{-1}m], \quad m^\circ_n := m^\circ_n(\theta) := \arg\min\{\mathrm{R}^m_n(\theta), m \in \mathcal{M}\}$$
$$\text{and} \quad \mathrm{R}^\circ_n(\theta) := \mathrm{R}^{m^\circ_n}_n(\theta) = \min\{\mathrm{R}^m_n(\theta), m \in \mathcal{M}\} \tag{16.2}$$

we have $\mathbb{v}_\theta^{-1}\mathrm{R}^m_n(\theta) \leqslant \mathbb{P}^n_\theta\|\widehat{\theta}^m - \theta\|^2_{\mathbb{H}} = n^{-1}\operatorname{tr}([\Gamma^\theta_{u_\bullet,u_\bullet}]_{\underline{m}}) + \|\theta^m - \theta\|^2_{\mathbb{H}} \leqslant (\mathbb{v}_\theta + 1)\mathrm{R}^m_n(\theta)$ *for all* $m \in \mathcal{M}$ *and* $n \in \mathbb{N}$. *As a consequence we immediately obtain the following oracle inequality*

$$\mathbb{v}_\theta^{-1}\mathrm{R}^\circ_n(\theta) \leqslant \inf_{m\in\mathcal{M}} \mathbb{P}^n_\theta\|\widehat{\theta}^m - \theta\|^2_{\mathbb{H}} \leqslant \mathbb{P}^n_\theta\|\widehat{\theta}^{m^\circ_n} - \theta\|^2_{\mathbb{H}} \leqslant 2\mathbb{v}_\theta\mathrm{R}^\circ_n(\theta) \leqslant 2\mathbb{v}^2_\theta \inf_{m\in\mathcal{M}} \mathbb{P}^n_\theta\|\widehat{\theta}^m - \theta\|^2_{\mathbb{H}},$$

and, hence $\mathrm{R}^\circ_n(\theta)$, m°_n *and the statistic* $\widehat{\theta}^{m^\circ_n}$, *respectively, is an oracle bound, an oracle dimension and oracle optimal (up to the constant* $2\mathbb{v}^2_\theta$*).* □

§16.09 **GSSM** *(§15.21 continued)*. If $\widehat{\theta} \sim \mathrm{N}_{(\theta, n^{-1}\mathrm{id}_{\mathbb{H}})}$ is a Gaussian process on $\mathbb{H}$ with $\mathbb{H}$-mean θ and covariance operator $n^{-1}\mathrm{id}_{\mathbb{H}} \in \mathbb{L}^+(\mathbb{H})$, then (16.1) is satisfied with $\mathbb{v}_\theta = 1$. Thereby, the statistic $\widehat{\theta}^{m^\circ_n} = U^\star(\widehat{\theta}_{u_\bullet}\mathbb{1}^{m^\circ_n}_\bullet)$ with oracle dimension m°_n as in (16.2) is *oracle optimal* (up to the constant 2) by Property §16.08. □

§16.10 **Nonparametric density estimation** *(§15.07 continued)*. Consider on $\mathbb{L}_2$ the stochastic process $\widehat{\mathbb{p}} = (\widehat{\mathbb{p}}_h)_{h\in\mathbb{L}_2}$ of real random variables defined on $([0,1]^n, \mathscr{B}^n_{[0,1]}, \mathbb{P}^{\otimes n}_{\mathbb{p}})$ by $\widehat{\mathbb{p}}_h := \widehat{\mathbb{P}}_n h \in \mathscr{B}^n$ for each $h \in \mathbb{L}_2$. If $\mathbb{p} \in \mathbb{D}_2 \cap \mathbb{L}_\infty$ then $\widehat{\mathbb{p}} = \mathbb{p} + n^{-1/2}\dot{W} \sim \mathrm{P}_{(\mathbb{p}, n^{-1}\Gamma^{\mathbb{p}})}$ is a stochastic process on $\mathbb{L}_2$ with $\mathbb{L}_2$-mean $\mathbb{p}$ and covariance operator $n^{-1}\Gamma^{\mathbb{p}} \in \mathbb{L}^+(\mathbb{L}_2)$ where $\Gamma^{\mathbb{p}} = \mathrm{M}_{\mathbb{p}} - \mathrm{M}_{\mathbb{p}}\Pi_{\mathbb{R}\mathbb{1}}\mathrm{M}_{\mathbb{p}}$. Let $u_\bullet = (u_j)_{j\in\mathbb{N}}$ be an ONS in $\mathbb{L}_2$ and let $\mathbb{U}^\perp$ denote the orthogonal complement of $\mathbb{U} = \overline{\operatorname{lin}}(u_j, j \in \mathbb{N})$ in $\mathbb{L}_2$ (see Definition §17.07). Assume that $\mathbb{1} := \mathbb{1}_{[0,1]} \in \mathbb{U}^\perp$, and thus $\langle \mathbb{1}, h\rangle_{\mathbb{L}_2} = 0$ for all $h \in \mathbb{U}$. Note that $\mathbb{p}_{\mathbb{1}} = 1$ (thus $\Pi_{\mathbb{R}\mathbb{1}}\mathbb{p} = \mathbb{1}$), and hence $\Gamma^{\mathbb{p}}\mathbb{1} = 0$, which can equally be deduced from $\widehat{\mathbb{p}}_{\mathbb{1}} = 1 \sim \mathrm{P}_{(\mathbb{p}_{\mathbb{1}}, 0)}$. As a consequence, we assume in the sequel an expansion $\mathbb{p} = \mathbb{1} + U^\star\mathbb{p}_{u_\bullet}$, which is trivially satisfied whenever $\{\mathbb{1}\} \cup \{u_j, j \in \mathbb{N}\}$ is an ONB. If in addition $\mathbb{p}^{-1} \in \mathbb{L}_\infty$, then (16.1) is satisfied with $\mathbb{v}_{\mathbb{p}} = \|\mathbb{p}\|_{\mathbb{L}_\infty} \vee \|\mathbb{p}^{-1}\|_{\mathbb{L}_\infty}$. Indeed, we have $\lambda(\mathbb{p}(h - \mathbb{p}\lambda(h))^2) = \langle \Gamma^{\mathbb{p}}h, h\rangle_{\mathbb{L}_2}$ (see Example §15.07), $\lambda((h - \mathbb{p}\lambda(h))^2) \leqslant \mathbb{v}_{\mathbb{p}}\lambda(\mathbb{p}(h - \mathbb{p}\lambda(h))^2)$ by definition, for each $h \in \mathbb{U}$, $\lambda((h - \mathbb{p}\lambda(h))^2) = \|h - \mathbb{p}\lambda(h)\mathbb{1}\|^2_{\mathbb{L}_2} = \|h\|^2_{\mathbb{L}_2} + \|\mathbb{p}\lambda(h)\mathbb{1}\|^2_{\mathbb{L}_2}$ since $\mathbb{1} \in \mathbb{U}^\perp$. Combining the bounds we obtain $\mathbb{v}_{\mathbb{p}}\langle \Gamma^{\mathbb{p}}h, h\rangle_{\mathbb{L}_2} \geqslant \|h\|^2_{\mathbb{L}_2}$, which shows the lower bound in (16.1). As for the upper bound we use $\|\Gamma^{\mathbb{p}}\|_{\mathbb{L}(\mathbb{H})} \leqslant \|\mathbb{p}\|_{\mathbb{L}_\infty} \leqslant \mathbb{v}_{\mathbb{p}}$ (see Example §15.07) together with Property §17.29 (i). Thereby, considering the canonical projection $\widehat{\mathbb{p}}_{u_\bullet} = (\widehat{\mathbb{p}}_{u_j})_{j\in\mathbb{N}}$ the statistic $\widehat{\mathbb{p}}^{m^\circ_n} = \mathbb{1} + U^\star(\widehat{\mathbb{p}}_{u_\bullet}\mathbb{1}^{m^\circ_n}_\bullet)$ (and hence $\widehat{\mathbb{p}}^{m^\circ_n} - \mathbb{p} = U^\star(\widehat{\mathbb{p}}_{u_\bullet}\mathbb{1}^{m^\circ_n}_\bullet) - U^\star\mathbb{p}_{u_\bullet}$) with oracle dimension m°_n as in (16.2) is *oracle optimal* (up to the constant $2(\|\mathbb{p}\|^2_{\mathbb{L}_\infty} \vee \|\mathbb{p}^{-1}\|^2_{\mathbb{L}_\infty})$) by Property §16.08. □

§16.11 **Nonparametric regression** *(§15.08 continued).* Consider the stochastic process $\widehat{f} = (\widehat{f}_h)_{h\in\mathbb{L}_2}$ on $\mathbb{L}_2$ of real valued random variables defined on $(\mathbb{R}^{2n}, \mathscr{B}^{2n}, \mathrm{U}_f^{\otimes n})$ by $\widehat{f}_h := \widehat{\mathbb{P}}_n(Yh(X)) \in \mathscr{B}^{2n}$ for each $h \in \mathbb{L}_2$. Under Assumption §14.01 if in addition $f \in \mathbb{L}_2 \cap \mathbb{L}_\infty$ then $\widehat{f} = f + n^{-1/2}\dot{W} \sim \mathrm{P}_{(f,n^{-1}\Gamma^f)}$ is a stochastic process on $\mathbb{L}_2$ with $\mathbb{L}_2$-mean f and covariance operator $n^{-1}\Gamma^f \in \mathbb{L}^+(\mathbb{L}_2)$ where $\Gamma^f = \sigma_\varepsilon^2 \mathrm{id}_{\mathbb{L}_2} + \mathrm{M}_f \Pi_{\mathbb{1}}^{\perp} \mathrm{M}_f$. Moreover, (16.1) is satisfied with $\mathbb{v}_f = (\sigma_\varepsilon^2 + \|f\|_{\mathbb{L}_\infty}^2) \vee \sigma_\varepsilon^{-2}$. Indeed, since $\langle \Gamma^f h, h\rangle_{\mathbb{L}_2} = \sigma_\varepsilon^2 \|h\|_{\mathbb{L}_2}^2 + \|\Pi_{\mathbb{1}}^{\perp} \mathrm{M}_f h\|_{\mathbb{L}_2}^2$ the lower bound follows immediatly, while for the upper bound we use $\|\Gamma^f\|_{\mathbb{L}(\mathbb{L}_2)} \leqslant \sigma_\varepsilon^2 + \|f\|_{\mathbb{L}_\infty}^2 \leqslant \mathbb{v}_f$ (see Example §15.08) together with Property §17.29 (i). Thereby, considering the canonical projection $\widehat{f}_{u_\bullet} = (\widehat{f}_{u_j})_{j\in\mathbb{N}}$ the statistic $\widehat{f}^{m_n^\circ} = U^\star(\widehat{f}_{u_\bullet} \mathbb{1}_\bullet^{m_n^\circ})$ with oracle dimension m_n° as in (16.2) is *oracle optimal* (up to the constant $2((\sigma_\varepsilon^2 + \|f\|_{\mathbb{L}_\infty}^2)^2 \vee \sigma_\varepsilon^{-4})$) by Property §16.08. □

§16.12 **Illustration.** Here and subsequently, we use for two sequences $a_\bullet, b_\bullet \in (\mathbb{R}_{\setminus 0}^+)^{\mathbb{N}}$ the notation $a_n \sim b_n$ if the sequence $a_\bullet/b_\bullet$ is bounded away both from zero and infinity. We illustrate the last results considering usual behaviour for the bias terms $(\|\theta^m - \theta\|_{\mathbb{H}}^2)_{m\in\mathcal{M}}$. We distinguish the following two cases

(p) there is $K \in \mathbb{N}$ with $\|\theta^{K-1} - \theta\|_{\mathbb{H}}^2 > 0$ and $\|\theta^K - \theta\|_{\mathbb{H}}^2 = 0$,

(np) for all $m \in \mathbb{N}$ holds $\|\theta^m - \theta\|_{\mathbb{H}}^2 > 0$.

Note that the expansion of θ is in case **(p)** *finite*, i.e., $\theta = \sum_{i\in[\![K]\!]} \theta_{u_j} u_j$ for some $K \in \mathbb{N}$ while in the opposite case **(np)**, it isn't. Interestingly, in case **(p)** the oracle bound is parametric, that is, $n\mathrm{R}_n^\circ(\theta) = \mathrm{O}(1)$, in case **(np)** the oracle bound is nonparametric, i.e. $\lim_{n\to\infty} n\mathrm{R}_n^\circ(\theta) = \infty$. In case **(np)** consider the following two specifications:

(P) If $\|\theta^m - \theta\|_{\mathbb{H}}^2 \sim m^{-2s}$, $s > 0$, then $m_n^\circ \sim n^{\frac{1}{2s+1}}$ and $\mathrm{R}_n^\circ(\theta) \sim n^{\frac{-2s}{2s+1}}$.

(E) If $\|\theta^m - \theta\|_{\mathbb{H}}^2 \sim \exp(-m^{2s})$, $s > 0$, then $m_n^\circ \sim (\log n)^{\frac{1}{2s}}$ and $\mathrm{R}_n^\circ(\theta) \sim (\log n)^{\frac{1}{2s}} n^{-1}$. □

§16.13 **Notation.** Recall that $u_\bullet = (u_j)_{j\in\mathbb{N}}$ is an ONS with $\mathbb{U} = \overline{\mathrm{lin}}\{u_j, j \in \mathbb{N}\} \subseteq \mathbb{H}$ and for $h \in \mathbb{H}$ denotes $h_{u_\bullet} := (h_{u_j})_{j\in\mathbb{N}} = Uh$ its generalised Fourier coefficients. $U \in \mathbb{L}(\mathbb{U}, \ell_2)$ is a unitary operator with inverse $U^\star$ (see Example §17.30 (a)). For a strictly positive sequence of weights $w_\bullet \in (\mathbb{R}_{\setminus 0}^+)^{\mathbb{N}}$ consider the Hilbert space $\ell_2(w_\bullet^2) := \{a_\bullet \in \mathbb{R}^{\mathbb{N}}, \|a_\bullet\|_{\ell_2(w_\bullet^2)} < \infty\}$ with inner product $\langle a_\bullet, b_\bullet\rangle_{\ell_2(w_\bullet^2)} = \sum_{j\in\mathbb{N}} w_j^2 a_j b_j$ and induced norm $\|\cdot\|_{\ell_2(w_\bullet^2)}$ (see Example §17.03 (c)). Let $\|w_\bullet^{-1}\|_{\ell_\infty} < \infty$, then $\ell_2(w_\bullet^2) \subseteq \ell_2$, and hence the image $U^\star(\ell_2(w_\bullet^2)) = \{U^\star a_\bullet : a_\bullet \in \ell_2(w_\bullet^2)\}$ of $\ell_2(w_\bullet^2)$ under U is a subset of $\mathbb{U}$. Moreover, $\mathbb{U}^{w_\bullet} := U^\star(\ell_2(w_\bullet^2))$ is a Hilbert space with inner product $\langle U^\star a_\bullet, U^\star a_\bullet^o\rangle_{u_\bullet,w_\bullet} := \langle a_\bullet, a_\bullet^o\rangle_{\ell_2(w_\bullet^2)}$ and induced norm $\|\cdot\|_{u_\bullet,w_\bullet}$. If $u_\bullet$ is complete in $\mathbb{H}$, i.e. $\mathbb{H} = \mathbb{U}$, then we eventually write $(\mathbb{H}^{w_\bullet}, \langle\cdot,\cdot\rangle_{u_\bullet,w_\bullet})$. □

§16.14 **Example.** Consider the real Hilbert space $\mathbb{L}_2 = \mathbb{L}_2(\mathscr{B}_{[0,1]}, \lambda_{[0,1]})$ and the *trigonometric basis* $\psi_\bullet = (\psi_j)_{j\in\mathbb{N}}$ (see Example §17.05). Define further the Hilbert space $(\mathbb{L}_2^{w_\bullet}, \langle\cdot,\cdot\rangle_{\psi_\bullet,w_\bullet})$ with respect to the trigonometric basis as in Notation §16.13.

(P) If we set $w_1 = 1$, $w_{2k} = w_{2k+1} = k^s$, $s \in \mathbb{N}$, $k \in \mathbb{N}$, then $\mathbb{L}_2^{w_\bullet}$ is a subset of the *Sobolev space* of s-times differentiable periodic functions. Moreover, up to a constant, for any function $h \in \mathbb{L}_2^{w_\bullet}$, the weighted norm $\|h\|_{\psi_\bullet,w_\bullet}^2$ equals the $\mathbb{L}_2$-norm of its s-th weak derivative $h^{(s)}$ (Tsybakov [2009]).

(E) If, on the contrary, $w_j = \exp(-1 + j^{2s})$, $s > 1/2$, $j \in \mathbb{N}$, then $\mathbb{L}_2^{w_\bullet}$ is a *class of analytic functions* (Kawata [1972]).

Note that, the trigonometric basis is $w_\bullet^{-1}$*-regular* as in Definition §17.12 (b) whenever $w_\bullet^{-1} \in \ell_2$

(see Example §17.14), and thus in the case **(P)** for $s > 1/2$ and in the case **(E)** for $s > 0$. □

§16.15 **Abstract smoothness condition.** Given a sequence of weights $\mathfrak{f}_\bullet = (\mathfrak{f}_j)_{j\in\mathbb{N}} \in (\mathbb{R}^+_{\setminus 0})^{\mathbb{N}}$ with $\|\mathfrak{f}_\bullet\|_{\ell_\infty} < \infty$ and an ONS $u_\bullet = (u_j)_{j\in\mathbb{N}}$ in $\mathbb{H}$ consider $(\mathbb{U}^{1/\mathfrak{f}_\cdot}, \|\cdot\|_{u_\bullet,\mathfrak{f}_\bullet^{-1}})$ as in Notation §16.13. Let $r \in \mathbb{R}^+_{\setminus 0}$ be a constant. We assume in the following that the function of interest belongs to the ellipsoid $\mathbb{F}^r_{u_\bullet,\mathfrak{f}_\bullet} := \{h \in \mathbb{U}^{\mathfrak{f}_\cdot^{-1}} : \|h\|^2_{u_\bullet,\mathfrak{f}_\bullet^{-1}} \leqslant r^2\} \subseteq \mathbb{U}$. □

§16.16 **Lemma.** *For $m \in \mathcal{M}$ consider the approximation $\theta^m = U^\star(\theta_{u_\bullet}\mathbb{1}^m_\bullet) \in \mathbb{U}_m$ of $\theta = U^\star(\theta_{u_\bullet}\mathbb{1}_\bullet) \in \mathbb{U}$ and set $\mathfrak{f}_{(m)} := \|\mathfrak{f}_\bullet(\mathbb{1}_\bullet - \mathbb{1}^m_\bullet)\|_{\ell_\infty} = \sup_{j\in[\![m]\!]^c} \mathfrak{f}_j$. If $\theta \in \mathbb{F}^r_{u_\bullet,\mathfrak{f}_\bullet}$, then $\mathbb{bias}_\theta = \|\theta^m - \theta\|_{\mathbb{H}} \leqslant r\,\mathfrak{f}_{(m)}$.*

§16.17 **Proof** of Lemma §16.16. is given in the lecture. □

§16.18 **Proposition.** *Let $\widehat{\theta} \sim \mathbb{P}^n_\theta = \mathrm{P}_{(\theta,n^{-1}\Gamma^\theta)}$. Setting for $m \in \mathcal{M}$ and $n \in \mathbb{N}$*

$$\mathrm{R}^m_n(\mathfrak{f}_\bullet) := [\mathfrak{f}^2_{(m)} \vee n^{-1}m], \quad m^\star_n := m^\star_n(\mathfrak{f}_\bullet) := \arg\min\{\mathrm{R}^m_n(\mathfrak{f}_\bullet), m \in \mathcal{M}\}$$
$$\text{and} \quad \mathrm{R}^\star_n(\mathfrak{f}_\bullet) := \mathrm{R}^{m^\star_n}_n(\mathfrak{f}_\bullet) = \min\{\mathrm{R}^m_n(\mathfrak{f}_\bullet), m \in \mathcal{M}\} \tag{16.3}$$

we have $\mathbb{P}^n_\theta\|\widehat{\theta}^{m^\star_n} - \theta\|^2_{\mathbb{H}} \leqslant (\|\Gamma^\theta\|_{\mathbb{L}(\mathbb{H})} + r^2)\ \mathrm{R}^\star_n(\mathfrak{f}_\bullet)$ for all $\theta \in \mathbb{F}^r_{u_\bullet,\mathfrak{f}_\bullet}$ and $n \in \mathbb{N}$.

§16.19 **Proof** of Proposition §16.18. is given in the lecture. □

§16.20 **Remark.** Arguing similarly as in Remark §16.07 we note that $\mathrm{R}^\star_n(\mathfrak{f}_\bullet) = \mathrm{o}(1)$ as $n \to \infty$, whenever $\mathfrak{f}_{(m)} = \mathrm{o}(1)$ as $m \to \infty$. The latter is satisfied, for example, if $\mathfrak{f}_\bullet \in \ell_2$. Note that the dimension $m^\star_n := m^\star_n(\mathfrak{f}_\bullet)$ as defined in 16.3 does not depend on the unknown parameter of interest θ but on the class $\mathbb{F}^r_{u_\bullet,\mathfrak{f}_\bullet}$ only, and thus also the statistic $\widehat{\theta}^{m^\star_n}$. In other words, if the regularity of θ known in advance, then the OSE $\widehat{\theta}^{m^\star_n}$ is a feasible estimator. □

§16.21 **GSSM** *(§16.09 continued).* If $\widehat{\theta} \sim \mathrm{N}_{(\theta,n^{-1}\mathrm{id}_{\mathbb{H}})}$, where $\|\mathrm{id}_{\mathbb{H}}\|_{\mathbb{L}(\mathbb{H})} = 1$. From Proposition §16.18 we obtain immediately, $\sup\{\mathbb{P}^n_\theta\|\widehat{\theta}^{m^\star_n} - \theta\|^2_{\mathbb{H}}, \theta \in \mathbb{F}^r_{u_\bullet,\mathfrak{f}_\bullet}\} \leqslant (1 + r^2)\ \mathrm{R}^\star_n(\mathfrak{f}_\bullet)$ for all $n \in \mathbb{N}$. In other words the global $\mathbb{H}$-risk of the OSE with optimally choosen dimension is not larger than $\mathrm{R}^\star_n(\mathfrak{f}_\bullet)$ (up to a constant) uniformely for all functions of interest belonging to $\mathbb{F}^r_{u_\bullet,\mathfrak{f}_\bullet}$. □

§16.22 **Nonparametric density estimation** *(§16.10 continued).* Consider on $\mathbb{L}_2$ the stochastic process $\widehat{\mathbb{p}} = (\widehat{\mathbb{p}}_h)_{h\in\mathbb{L}_2}$ of real random variables defined on $([0,1]^n, \mathscr{B}^n_{[0,1]}, \mathbb{P}^{\otimes n}_{\mathbb{p}})$ by $\widehat{\mathbb{p}}_h := \widehat{\mathbb{P}}^n h \in \mathscr{B}^n$ for each $h \in \mathbb{L}_2$. If $\mathbb{p} \in \mathbb{D}_2 \cap \mathbb{L}_\infty$ then $\widehat{\mathbb{p}} \sim \mathrm{P}_{(\mathbb{p},n^{-1}\Gamma^{\mathbb{p}})}$ with $\|\Gamma^{\mathbb{p}}\|_{\mathbb{L}(\mathbb{L}_2)} \leqslant \|\mathbb{p}\|_{\mathbb{L}_\infty}$ (see Example §15.07). From Proposition §16.18 we obtain immediately un upper bound for the gobal $\mathbb{L}_2$-risk (mise) which still depends on $\|\mathbb{p}\|_{\mathbb{L}_\infty}$. If we assume in addition that the ONS $u_\bullet$ is $\mathfrak{f}_\bullet$-regular as in Definition §17.12 (b), i.e. $\|\sum_{j\in\mathbb{N}} \mathfrak{f}^2_j|u_j|^2\|_{\mathbb{L}_\infty} \leqslant \tau^2_{u_\bullet,\mathfrak{f}_\bullet}$ for some $\tau_{u_\bullet,\mathfrak{f}_\bullet} \in \mathbb{R}^+_{\setminus 0}$. Then, we have $\|\mathbb{p}\|_{\mathbb{L}_\infty} \leqslant r\tau_{u_\bullet,\mathfrak{f}_\bullet}$ for all $\mathbb{p} \in \mathbb{F}^r_{u_\bullet,\mathfrak{f}_\bullet}$ by Lemma §17.15. As a consequence, considering the OSE $\widehat{\mathbb{p}}^{m^\star_n} = U^\star(\widehat{\mathbb{p}}_{u_\bullet}\mathbb{1}^{m^\star_n}_\bullet)$ with dimension $m^\star_n$ as in (16.3) from Proposition §16.18 we obtain, $\sup\{\mathbb{P}^{\otimes n}_{\mathbb{p}}\|\widehat{\mathbb{p}}^{m^\star_n} - \mathbb{p}\|^2_{\mathbb{L}_2}, \mathbb{p} \in \mathbb{D}_2 \cap \mathbb{F}^r_{u_\bullet,\mathfrak{f}_\bullet}\} \leqslant (r\tau_{u_\bullet,\mathfrak{f}_\bullet} + r^2)\ \mathrm{R}^\star_n(\mathfrak{f}_\bullet)$ for all $n \in \mathbb{N}$. Consider the *trigonometric basis* $\psi_\bullet = (\psi_j)_{j\in\mathbb{N}}$ defined in Example §17.05 and $w_\bullet \in (\mathbb{R}^+_{\setminus 0})^{\mathbb{N}}$ given either in Example §16.14 **(P)** or in **(E)**. If we set $\mathfrak{f}_\bullet := w_\bullet^{-1}$, then $\psi_\bullet$ is *$\mathfrak{f}_\bullet$-regular* (see Example §17.14) with $\tau^2_{u_\bullet,\mathfrak{f}_\bullet} = 2\|\mathfrak{f}_\bullet\|^2_{\ell_2}$ which is finite in case **(P)** for $s > 1/2$ and in case **(E)** for $s > 0$. In this situation we $\sup\{\mathbb{P}^{\otimes n}_{\mathbb{p}}\|\widehat{\mathbb{p}}^{m^\star_n} - \mathbb{p}\|^2_{\mathbb{L}_2}, \mathbb{p} \in \mathbb{D}_2 \cap \mathbb{F}^r_{\psi_\bullet,\mathfrak{f}_\bullet}\} \leqslant (\sqrt{2}r\|\mathfrak{f}_\bullet\|_{\ell_2} + r^2)\ \mathrm{R}^\star_n(\mathfrak{f}_\bullet)$ for all $n \in \mathbb{N}$, where $\mathrm{R}^\star_n(\mathfrak{f}_\bullet) = \mathrm{o}(1)$ as $n \to \infty$ (Remark §16.20). □

§16.23 **Nonparametric regression** *(§16.11 continued).* Consider the stochastic process $\widehat{f} = (\widehat{f}_h)_{h\in\mathbb{L}_2}$ on $\mathbb{L}_2$ of real valued random variables defined on $(\mathbb{R}^{2n}, \mathscr{B}^{2n}, \mathrm{U}_f^{\otimes n})$ by $\widehat{f}_h := \widehat{\mathbb{P}}_n(Yh(X)) \in \mathscr{B}^{2n}$ for each $h \in \mathbb{L}_2$. Under Assumption §14.01 if $f \in \mathbb{L}_2 \cap \mathbb{L}_\infty$ then $\widehat{f} \sim \mathrm{P}_{(f,n^{-1}\Gamma^f)}$ with $\|\Gamma^f\|_{\mathbb{L}(\mathbb{L}_2)} \leqslant \sigma_\varepsilon^2 + \|f\|_{\mathbb{L}_\infty}^2$ (see Example §15.08). From Proposition §16.18 we obtain immediately un upper bound for the gobal $\mathbb{L}_2$-risk (mise) which still depends on $\|f\|_{\mathbb{L}_\infty}$. If we assume in addition that the ONS $u_\bullet$ is $\mathfrak{f}_\bullet$-regular as in Definition §17.12 (b), then $\|f\|_{\mathbb{L}_\infty} \leqslant r\tau_{u_\bullet,\mathfrak{f}_\bullet}$ for all $f \in \mathbb{F}^r_{u_\bullet,\mathfrak{f}_\bullet}$ by Lemma §17.15. As a consequence, considering the OSE $\widehat{f}^{m_n^\star} = U^\star(\widehat{f}_{u_\bullet}\mathbb{1}^{m_n^\star}_\bullet)$ with dimension $m_n^\star$ as in (16.3) from Proposition §16.18 we obtain, $\sup\{\mathrm{U}_f^{\otimes n}\|\widehat{f}^{m_n^\star} - f\|_{\mathbb{L}_2}^2, f \in \mathbb{F}^r_{u_\bullet,\mathfrak{f}_\bullet}\} \leqslant (\sigma_\varepsilon^2 + r^2\tau^2_{u_\bullet,\mathfrak{f}_\bullet} + r^2)\ \mathrm{R}_n^\star(\mathfrak{f}_\bullet)$ for all $n \in \mathbb{N}$. Consider the *trigonometric basis* $\psi_\bullet = (\psi_j)_{j\in\mathbb{N}}$ defined in Example §17.05 and $\mathfrak{f}_\bullet^{-1} := w_\bullet \in (\mathbb{R}^+_{\setminus 0})^{\mathbb{N}}$ given either in Example §16.14 **(P)** or **(E)**. In this situation, similar to Example §16.22, we obtain $\sup\{\mathrm{U}_f^{\otimes n}\|\widehat{f}^{m_n^\star} - f\|_{\mathbb{L}_2}^2, f \in \mathbb{F}^r_{\psi_\bullet,\mathfrak{f}_\bullet}\} \leqslant (\sigma_\varepsilon^2 + 2r^2\|\mathfrak{f}_\bullet\|_{\ell_2}^2 + r^2)\ \mathrm{R}_n^\star(\mathfrak{f}_\bullet)$ for all $n \in \mathbb{N}$, where $\mathrm{R}_n^\star(\mathfrak{f}_\bullet) = \mathrm{o}(1)$ as $n \to \infty$ (Remark §16.20). □

§16.24 **Illustration**. Let us consider the following two specifications:

(P) If $\mathfrak{f}_m^2 \sim m^{-2s}$, $s > 0$, then $m_n^\star \sim n^{\frac{1}{2s+1}}$ and $\mathrm{R}_n^\star(\mathfrak{f}_\bullet) \sim n^{\frac{-2s}{2s+1}}$.

(E) If $\mathfrak{f}_m^2 \sim \exp(-m^{2s})$, $s > 0$, then $m_n^\star \sim (\log n)^{\frac{1}{2s}}$ and $\mathrm{R}_n^\star(\mathfrak{f}_\bullet) \sim (\log n)^{\frac{1}{2s}} n^{-1}$. □

Local measure of accuracy.

Consider a linear functional $\Phi : \mathbb{H} \supset \mathcal{D}(\Phi) \to \mathbb{R}$, e.g. the point evaluation in Example §17.24. We assume from here on that the ONS $u_\bullet = (u_j)_{j\in\mathbb{N}}$ and hence for each $m \in \mathcal{M}$ also the orthogonal projection $\theta^m = U^\star(\theta_{u_\bullet}\mathbb{1}^m_\bullet) = \sum_{j\in[\![m]\!]} \theta_{u_j} u_j$ and the OSE $\widehat{\theta}^m = U^\star(\widehat{\theta}_{u_\bullet}\mathbb{1}^m_\bullet) = \sum_{j\in[\![m]\!]} \widehat{\theta}_{u_j} u_j$ belong to the domain $\mathcal{D}(\Phi)$ of Φ. As a consequence $\Phi(\theta^m)$ and $\Phi(\widehat{\theta}^m)$ are well-defined. Assuming in addition $\theta \in \mathcal{D}(\Phi)$ we measure the accuracy of $\widehat{\theta}^m$ by a local Φ-risk with respect to the distribution $\mathbb{P}_\theta^n$ of the noisy version $\widehat{\theta}$. Keep in mind, if $\widehat{\theta}$ admits an $\mathbb{H}$-mean $\theta \in \mathcal{D}(\Phi)$, then $\Phi(\widehat{\theta}^m)$ is an unbiased estimator of $\Phi(\theta^m)$, i.e. $\mathbb{P}_\theta^n\Phi(\widehat{\theta}^m) = \Phi(\theta^m)$, due to the linearity of the expectation and Φ.

§16.25 **Definition**. Given a noisy version $\widehat{\theta} \sim \mathbb{P}_\theta^n$ of θ admitting a $\mathbb{H}$-mean $\theta \in \mathcal{D}(\Phi) \subseteq \mathbb{H}$ the *local Φ-risk* of a OSE $\widehat{\theta}^m = U^\star(\widehat{\theta}_{u_\bullet}\mathbb{1}^m_\bullet)$ satisfies

$$\mathbb{P}_\theta^n\big(|\Phi(\widehat{\theta}^m) - \Phi(\theta)|^2\big) = \mathbb{var}_\theta + \mathbb{bias}_\theta^2$$

with *variance* term $\mathbb{var}_\theta := \mathbb{P}_\theta^n\big(|\Phi(\widehat{\theta}^m) - \Phi(\theta^m)|^2\big)$ and *bias* term $\mathbb{bias}_\theta := \Phi(\theta^m) - \Phi(\theta)$. □

In the sequel we analyse separately the variance and bias term. We consider a noisy version $\widehat{\theta} = \theta + n^{-1/2}\dot{W} \sim \mathbb{P}_\theta^n$ of θ admitting a $\mathbb{H}$-mean $\theta \in \mathbb{H}$ and a covariance operator $n^{-1}\Gamma^\theta \in \mathbb{L}^+(\mathbb{H})$, i.e. $\widehat{\theta} \sim \mathrm{P}_{(\theta,n^{-1}\Gamma^\theta)}$ as in Definition §15.20. Since the error process $\dot{W} \sim \mathrm{P}_{(0,\Gamma^\theta)}$ has $\mathbb{H}$-mean zero and covariance operator $\Gamma^\theta \in \mathbb{L}^+(\mathbb{H})$ (see Definition §15.04) for each $m \in \mathcal{M}$ the $\mathbb{R}^m$-valued random vector $[\dot{W}_{u_\bullet}]_{\underline{m}} = (\dot{W}_{u_j})_{j\in[\![m]\!]}$ has mean zero and covariance matrix $[\Gamma^\theta_{u_\bullet,u_\bullet}]_{\underline{m}} = (\Gamma^\theta_{u_k,u_j})_{k,j\in[\![m]\!]}$, i.e. $[\dot{W}_{u_\bullet}]_{\underline{m}} \sim \mathrm{P}_{(0,[\Gamma^\theta_{u_\bullet,u_\bullet}]_{\underline{m}})}$ (Notation §16.03).

§16.26 **Notation**. Let $m \in \mathbb{N}$. We set $\Phi_{u_\bullet} := (\Phi_{u_j})_{j\in\mathbb{N}}$ with the slight abuse of notations $\Phi_{u_j} := \Phi(u_j)$, $j \in \mathbb{N}$. If $\Phi \in \mathbb{L}(\mathbb{H},\mathbb{R})$ then $\mathcal{D}(\Phi) = \mathbb{H}$, and by Fréchet-Riesz representation theorem (Property §17.23) there is $\phi \in \mathbb{H}$ with $\Phi_{u_\bullet} = \phi_{u_\bullet}$, and thus $\Phi_{u_\bullet} \in \ell_2$. Recall that $[\Phi_{u_\bullet}]_{\underline{m}} \in \mathbb{R}^m$ and $[\phi_{u_\bullet}]_{\underline{m}} \in \mathbb{R}^m$ denotes a sub-vector of $\Phi_{u_\bullet}$ and $\phi_{u_\bullet}$, respectively. □

§16.27 **Property**. *Let* $\widehat{\theta} = \theta + n^{-1/2}\dot{W} \sim \mathbb{P}_\theta^n = \mathrm{P}_{(\theta, n^{-1}\Gamma^\theta)}$ *and* $\theta \in \mathcal{D}(\Phi)$, *then for all* $m \in \mathcal{M}$ *we have*

$$\begin{aligned}\mathbb{v}\mathrm{ar}_\theta = n^{-1}\mathbb{P}_\theta^n\big(|\langle \Phi_{u\cdot}\mathbb{1}_\bullet^m, \dot{W}_{u\cdot}\rangle_{\ell_2}|^2\big) &= n^{-1}\langle \Gamma^\theta_{u\cdot,u\cdot}(\Phi_{u\cdot}\mathbb{1}_\bullet^m), \Phi_{u\cdot}\mathbb{1}_\bullet^m\rangle_{\ell_2} \\ &= n^{-1}\|\Phi_{u\cdot}\mathbb{1}_\bullet^m\|^2_{\Gamma^\theta_{u\cdot,u\cdot}} = n^{-1}[\Phi_{u\cdot}]^t_{\underline{m}}[\Gamma^\theta_{u\cdot,u\cdot}]_{\underline{m}}[\Phi_{u\cdot}]_{\underline{m}}. \quad \square\end{aligned}$$

§16.28 **Definition**. Let $C \in \mathbb{R}^+_{\setminus 0}$ and for each $n \in \mathbb{N}$ let $\mathrm{R}_n^\circ \in \mathbb{R}^+_{\setminus 0}$ and $m_n^\circ \in \mathbb{N}$ satisfy

$$C^{-1}\,\mathrm{R}_n^\circ \leqslant \inf_{m\in\mathcal{M}} \mathbb{P}_\theta^n\big(|\Phi(\widehat{\theta}^m) - \Phi(\theta)|^2\big) \leqslant \mathbb{P}_\theta^n\big(|\Phi(\widehat{\theta}^{m_n^\circ}) - \Phi(\theta)|^2\big) \leqslant C\,\mathrm{R}_n^\circ.$$

Then we call R_n° *oracle bound*, m_n° *oracle dimension* and $\widehat{\theta}^{m_n^\circ}$ *oracle optimal* (up to the constant C). As a consequence, up to the constant C^2 the statistik $\widehat{\theta}^{m_n^\circ}$ attains the lower Φ-risk bound within the family of OSE's, i.e. $\mathbb{E}_\theta^n(|\Phi(\widehat{\theta}^{m_n^\circ}) - \Phi(\theta)|^2) \leqslant C^2 \inf_{m\in\mathcal{M}} \mathbb{P}_\theta^n(|\Phi(\widehat{\theta}^m) - \Phi(\theta)|^2)$. $\square$

§16.29 **Oracle inequality**. *If* $\widehat{\theta} \sim \mathbb{P}_\theta^n = \mathrm{P}_{(\theta, n^{-1}\Gamma^\theta)}$ *then setting for* $n, m \in \mathbb{N}$

$$\mathrm{R}_n^m(\theta) := [|\Phi(\theta^m - \theta)|^2 \vee n^{-1}\|\Phi_{u\cdot}\mathbb{1}_\bullet^m\|_{\Gamma^\theta_{u\cdot,u\cdot}}], \quad m_n^\circ := m_n^\circ(\theta) := \arg\min\{\mathrm{R}_n^m(\theta), m \in \mathcal{M}\}$$
$$\text{and} \quad \mathrm{R}_n^\circ(\theta) := \mathrm{R}_n^{m_n^\circ}(\theta) = \min\{\mathrm{R}_n^m(\theta), m \in \mathcal{M}\}$$

we have $\mathrm{R}_n^m(\theta) \leqslant \mathbb{P}_\theta^n\big(|\Phi(\widehat{\theta}^m) - \Phi(\theta)|^2\big) \leqslant 2\mathrm{R}_n^m(\theta)$ *for all* $m \in \mathcal{M}$ *and* $n \in \mathbb{N}$. *It follows*

$$\begin{aligned}\mathrm{R}_n^\circ(\theta) \leqslant \inf_{m\in\mathcal{M}} \mathbb{P}_\theta^n\big(|\Phi(\widehat{\theta}^m) - \Phi(\theta)|^2\big) &\leqslant \mathbb{P}_\theta^n\big(|\Phi(\widehat{\theta}^{m_n^\circ}) - \Phi(\theta)|^2\big) \\ &\leqslant 2\mathrm{R}_n^\circ(\theta) \leqslant 2\inf_{m\in\mathcal{M}} \mathbb{P}_\theta^n\big(|\Phi(\widehat{\theta}^m) - \Phi(\theta)|^2\big)\end{aligned}$$

As a consequence, $\mathrm{R}_n^\circ(\theta)$, m_n° *and the statistic* $\widehat{\theta}^{m_n^\circ}$, *respectively, is an oracle bound, an oracle dimension and oracle optimal (up to the constant* 2*).*

§16.30 **Remark**. Arguing similarly as in Remark §16.07 we note that $\mathrm{R}_n^\circ(\theta) = \mathrm{o}(1)$ as $n \to \infty$, whenever $|\Phi(\theta^m - \theta)| = \mathrm{o}(1)$ as $m \to \infty$. The latter is satisfied, for example, if $\Phi_{u\cdot}\theta_{u\cdot} \in \ell_1$ (see Definition §16.36). The oracle dimension $m_n^\circ = m_n^\circ(\theta)$ as defined in Property §16.29 depends again on the unknown parameter of interest θ, and thus also the oracle optimal statistic $\widehat{\theta}^{m_n^\circ}$. In other words $\widehat{\theta}^{m_n^\circ}$ is not a feasible estimator. $\square$

§16.31 **Oracle inequality**. *Let* $\widehat{\theta} \sim \mathbb{P}_\theta^n = \mathrm{P}_{(\theta, n^{-1}\Gamma^\theta)}$. *If* $\mathbb{v}_\theta \in \mathbb{R}^+_{\setminus 0}$ *satisfies* (16.1), *then setting for* $n, m \in \mathbb{N}$

$$\mathrm{R}_n^m(\theta) := [|\Phi(\theta^m - \theta)|^2 \vee n^{-1}\|\Phi_{u\cdot}\mathbb{1}_\bullet^m\|^2_{\ell_2}], \quad m_n^\circ := m_n^\circ(\theta) := \arg\min\{\mathrm{R}_n^m(\theta), m \in \mathcal{M}\}$$
$$\text{and} \quad \mathrm{R}_n^\circ(\theta) := \mathrm{R}_n^{m_n^\circ}(\theta) = \min\{\mathrm{R}_n^m(\theta), m \in \mathcal{M}\} \quad (16.4)$$

we have $\mathbb{v}_\theta^{-1}\mathrm{R}_n^m(\theta) \leqslant \mathbb{P}_\theta^n(|\Phi(\widehat{\theta}^m - \theta)|^2) \leqslant (\mathbb{v}_\theta + 1)\mathrm{R}_n^m(\theta)$ *for all* $m \in \mathcal{M}$ *and* $n \in \mathbb{N}$. *It follows*

$$\begin{aligned}\mathbb{v}_\theta^{-1}\mathrm{R}_n^\circ(\theta) \leqslant \inf_{m\in\mathcal{M}} \mathbb{P}_\theta^n(|\Phi(\widehat{\theta}^m - \theta)|^2) &\leqslant \mathbb{P}_\theta^n(|\Phi(\widehat{\theta}^{m_n^\circ} - \theta)|^2) \\ &\leqslant 2\mathbb{v}_\theta\mathrm{R}_n^\circ(\theta) \leqslant 2\mathbb{v}_\theta^2 \inf_{m\in\mathcal{M}} \mathbb{P}_\theta^n(|\Phi(\widehat{\theta}^m - \theta)|^2).\end{aligned}$$

As a consequence, $\mathrm{R}_n^\circ(\theta)$, m_n° *and the statistic* $\widehat{\theta}^{m_n^\circ}$, *respectively, is an oracle bound, an oracle dimension and oracle optimal (up to the constant* $2\mathbb{v}_\theta^2$*).* $\square$

§16.32 **GSSM** *(§16.09 continued).* If $\widehat{\theta} \sim \mathrm{N}_{(\theta, n^{-1}\mathrm{id}_{\mathbb{H}})}$, then (16.1) is satisfied with $\mathbb{v}_\theta = 1$. The statistic $\widehat{\theta}^{m^\circ_n} = U^\star(\widehat{\theta}_{u_\bullet}\mathbb{1}^{m^\circ_n}_\bullet)$ with oracle dimension m°_n as in (16.4) is *oracle optimal* (up to the constant 2) by Property §16.31. □

§16.33 **Nonparametric density estimation** *(§16.10 continued).* Consider on $\mathbb{L}_2$ the stochastic process $\widehat{\mathbb{p}} = (\widehat{\mathbb{p}}_h)_{h\in\mathbb{L}_2}$ of real random variables defined on $([0,1]^n, \mathscr{B}^n_{[0,1]}, \mathbb{P}^{\otimes n}_{\mathbb{p}})$ by $\widehat{\mathbb{p}}_h := \widehat{\mathbb{P}}_n h \in \mathscr{B}^n$ for each $h \in \mathbb{L}_2$. If $\mathbb{p} \in \mathbb{D}_2 \cap \mathbb{L}_\infty$, $\mathbb{p}^{-1} \in \mathbb{L}_\infty$ and $\mathbb{1}_{[0,1]} \in \mathbb{U}^\perp$ then we have $\widehat{\mathbb{p}} \sim \mathbb{P}^n_{\mathbb{p}} = \mathrm{P}_{(\mathbb{p}, n^{-1}\Gamma^{\mathbb{p}})}$ and $\mathbb{v}_{\mathbb{p}} = \|\mathbb{p}\|_{\mathbb{L}_\infty} \vee \|\mathbb{p}^{-1}\|_{\mathbb{L}_\infty}$ satisfies (16.1) (see Example §16.10) Thereby, considering the canonical projection $\widehat{\mathbb{p}}_{u_\bullet} = (\widehat{\mathbb{p}}_{u_j})_{j\in\mathbb{N}}$ the statistic $\widehat{\mathbb{p}}^{m^\circ_n} = \mathbb{1} + U^\star(\widehat{\mathbb{p}}_{u_\bullet}\mathbb{1}^{m^\circ_n}_\bullet)$ with oracle dimension m°_n as in (16.4) is *oracle optimal* for $\mathbb{p} = \mathbb{1} + U^\star\mathbb{p}_{u_\bullet}$ (up to the constant $2(\|\mathbb{p}\|^2_{\mathbb{L}_\infty} \vee \|\mathbb{p}^{-1}\|^2_{\mathbb{L}_\infty})$) by Property §16.31. □

§16.34 **Nonparametric regression** *(§16.11 continued).* Consider the stochastic process $\widehat{f} = (\widehat{f}_h)_{h\in\mathbb{L}_2}$ on $\mathbb{L}_2$ of real valued random variables defined on $\big(\mathbb{R}^{2n}, \mathscr{B}^{2n}, \mathrm{U}^{\otimes n}_f\big)$ by $\widehat{f}_h := \widehat{\mathbb{P}}_n(Yh(X)) \in \mathscr{B}^{2n}$ for each $h \in \mathbb{L}_2$. Under Assumption §14.01 if $f \in \mathbb{L}_2 \cap \mathbb{L}_\infty$ then we have $\widehat{f} \sim \mathbb{P}^n_f = \mathrm{P}_{(f, n^{-1}\Gamma^f)}$ and $\mathbb{v}_f = (\sigma^2_\varepsilon + \|f\|^2_{\mathbb{L}_\infty}) \vee \sigma^{-2}_\varepsilon$ satisfies (16.1) (see Example §16.11). Thereby, considering the canonical projection $\widehat{f}_{u_\bullet} = (\widehat{f}_{u_j})_{j\in\mathbb{N}}$ the statistic $\widehat{f}^{m^\circ_n} = U^\star(\widehat{f}_{u_\bullet}\mathbb{1}^{m^\circ_n}_\bullet)$ with oracle dimension m°_n as in (16.4) is *oracle optimal* (up to the constant $2((\sigma^2_\varepsilon + \|f\|^2_{\mathbb{L}_\infty})^2 \vee \sigma^{-4}_\varepsilon)$) by Property §16.31. □

§16.35 **Illustration**. We illustrate the last results considering usual behaviour for both the variance term $n^{-1}\|\Phi_{u_\bullet}\mathbb{1}^m_\bullet\|^2_{\ell_2}$ and the bias term $|\Phi(\theta^m - \theta)|^2$. Recalling the two cases **(p)** and **(np)** in Illustration §16.12 we distinguish the following two cases

(p) $\Phi_{u_\bullet} \in \ell_2$ or there is $K \in \mathbb{N}$ with $|\Phi(\theta^{K-1} - \theta)|^2 \in \mathbb{R}^+_{\setminus 0}$ and $\sup_{m\geqslant K} |\Phi(\theta^m - \theta)|^2 = 0$,

(np) $\Phi_{u_\bullet} \notin \ell_2$ and for all $m \in \mathbb{N}$ holds $|\Phi(\theta^m - \theta)|^2 \in \mathbb{R}^+_{\setminus 0}$.

In case **(p)** the oracle bound is again parametric, i.e. $n\mathrm{R}^\circ_n(\theta) = \mathrm{O}(1)$, while in case **(np)** the oracle bound is nonparametric, i.e. $\lim_{n\to\infty} n\mathrm{R}^\circ_n(\theta) = \infty$. In case **(np)** with $\Phi^2_{u_j} \sim j^{2a}$, $2a > -1$ and hence $\|\Phi_{u_\bullet}\mathbb{1}^m_\bullet\|^2_{\ell_2} \sim m^{2a+1}$ consider the following two specifications:

(P) If $\theta^2_{u_j} \sim j^{-2s-2}$, $s > 0$, and hence $|\Phi(\theta^m - \theta)|^2 \sim m^{-2(s-a)}$, then $m^\circ_n \sim n^{\frac{1}{2s+1}}$ and $\mathrm{R}^\circ_n(\theta) \sim n^{\frac{-2(s-a)}{2s+1}}$, where $\mathrm{R}^\circ_n(\theta) = \mathrm{o}(1)$ as $n \to \infty$ for $s > a$.

(E) If $\theta^2_{u_j} \sim j^{2s-2a-2}\exp(-j^{2s})$, $s > 0$, and hence $|\Phi(\theta^m - \theta)|^2 \sim \exp(-m^{2s})$, then $m^\circ_n \sim (\log n)^{\frac{1}{2s}}$ and $\mathrm{R}^\circ_n(\theta) \sim (\log n)^{\frac{2a+1}{2s}} n^{-1}$. □

§16.36 **Regular linear functional**. Consider an ONS $u_\bullet = (u_j)_{j\in\mathbb{N}}$ in $\mathbb{H}$ and an ellipsoid $\mathbb{F}^r_{u_\bullet,\mathfrak{f}_\bullet}$ as in Definition §16.15. We call a linear functional $\Phi : \mathbb{H} \supset \mathcal{D}(\Phi) \to \mathbb{R}$ *regular* if $u_\bullet$ belongs to the domain $\mathcal{D}(\Phi)$ of Φ and the sequence $\Phi_{u_\bullet} = (\Phi_{u_j})_{j\in\mathbb{N}}$ (see Notation §16.26) satisfies either $\Phi_{u_\bullet}\theta_{u_\bullet} \in \ell_1$ or $\Phi_{u_\bullet}\mathfrak{f}_\bullet \in \ell_2$. □

§16.37 **Remark**. We may emphasise that we neither impose that the sequence $\Phi_{u_\bullet} = (\Phi_{u_j})_{j\in\mathbb{N}}$ tends to zero nor that it is square summable. However, if $\Phi_{u_\bullet} \in \ell_2$ then $\Phi \in \mathbb{L}(\mathbb{U}, \mathbb{R})$ and $\Phi_{u_\bullet} = \phi_{u_\bullet}$, where $\phi_{u_\bullet}$ denotes the sequence of generalised Fourier coefficients of the representer ϕ of Φ given by Fréchet-Riesz representation theorem Property §17.23. Assuming a regular functional, however, enables us in specific cases to deal with more demanding functionals, such as in Example §17.24 the evaluation of the solution at a given point. We note that $\Phi_{u_\bullet} \in \ell_2(\mathfrak{f}^2_\bullet)$ implies $\Phi_{u_\bullet}\theta_{u_\bullet} \in \ell_1$ for all $\theta \in \mathbb{F}^r_{u_\bullet,\mathfrak{f}_\bullet}$ applying the Cauchy-Schwarz-inequality (Property §17.02).

Moreover, if $\Phi_{u_\bullet}\theta_{u_\bullet} \in \ell_1$ then $\Phi(\theta) = \sum_{j\in\mathbb{N}} \Phi_{u_j}\theta_{u_j}$ and $|\Phi(\theta^m - \theta)| \leqslant \|\Phi_{u_\bullet}\theta_{u_\bullet}(\mathbb{1}_\bullet - \mathbb{1}^m_\bullet)\|_{\ell_1} = \sum_{j\in[\![m]\!]^c} |\Phi_{u_j}\theta_{u_j}| = \mathrm{o}(1)$ as $m \to \infty$. As a consequence, for a regular linear functional the oracle bound given in Property §16.31 satisfies $\mathrm{R}^\circ_n(\theta) = \mathrm{o}(1)$ as $n \to \infty$ (Remark §16.30). □

§16.38 **Lemma**. *Let $\theta = U^\star(\theta_{u_\bullet}\mathbb{1}_\bullet) \in \mathbb{U}$. For $m \in \mathcal{M}$ set $\theta^m := U^\star(\theta_{u_\bullet}\mathbb{1}^m_\bullet) \in \mathbb{U}_m$ and $(\Phi_{u_\bullet}\mathfrak{f}_\bullet)_{(m)} := \|\Phi_{u_\bullet}\mathfrak{f}_\bullet(\mathbb{1}_\bullet - \mathbb{1}^m_\bullet)\|_{\ell_\infty} = \sup_{j\in[\![m]\!]^c}(\Phi_{u_j}\mathfrak{f}_j)$. If $\theta \in \mathbb{F}^r_{u_\bullet,\mathfrak{f}_\bullet}$, then $\mathbb{bias}_\theta = |\Phi(\theta^m - \theta)| \leqslant r\,(\Phi_{u_\bullet}\mathfrak{f}_\bullet)_{(m)}$.*

§16.39 **Proof** of Lemma §16.38. is given in the lecture. □

§16.40 **Proposition**. *Let $\widehat{\theta} \sim \mathbb{P}^n_\theta = \mathrm{P}_{(\theta,n^{-1}\Gamma^\theta)}$. Setting for $n, m \in \mathbb{N}$*

$$\mathrm{R}^m_n(\mathfrak{f}_\bullet) := [(\Phi_{u_\bullet}\mathfrak{f}_\bullet)^2_{(m)} \vee n^{-1}\|\Phi_{u_\bullet}\mathbb{1}^m_\bullet\|^2_{\ell_2}], \quad m^\star_n := m^\star_n(\mathfrak{f}_\bullet) := \arg\min\{\mathrm{R}^m_n(\mathfrak{f}_\bullet), m \in \mathcal{M}\}$$
$$\text{and} \quad \mathrm{R}^\star_n(\mathfrak{f}_\bullet) := \mathrm{R}^{m^\star_n}_n(\mathfrak{f}_\bullet) = \min\{\mathrm{R}^m_n(\mathfrak{f}_\bullet), m \in \mathcal{M}\} \quad (16.5)$$

we have $\mathbb{P}^n_\theta(|\Phi(\widehat{\theta}^{m^\star_n} - \theta)|^2) \leqslant (\|\Gamma^\theta\|_{\mathbb{L}(\mathbb{H})} + r^2)\ \mathrm{R}^\star_n(\mathfrak{f}_\bullet)$ for all $\theta \in \mathbb{F}^r_{u_\bullet,\mathfrak{f}_\bullet}$ and $n \in \mathbb{N}$.

§16.41 **Proof** of Proposition §16.40. is given in the lecture. □

§16.42 **Remark**. Arguing similarly as in Remark §16.07 we note that $\mathrm{R}^\star_n(\mathfrak{f}_\bullet) = \mathrm{o}(1)$ as $n \to \infty$, whenever $(\Phi_{u_\bullet}\mathfrak{f}_\bullet)_{(m)} = \mathrm{o}(1)$ as $m \to \infty$. The latter is satisfied, for example, if $\Phi_{u_\bullet}\mathfrak{f}_\bullet \in \ell_2$, i.e. Φ is a regular linear functional. Note that the dimension $m^\star_n := m^\star_n(\mathfrak{f}_\bullet)$ as defined in 16.5 does not depend on the unknown parameter of interest θ but on the class $\mathbb{F}^r_{u_\bullet,\mathfrak{f}_\bullet}$ only, and thus also the statistic $\widehat{\theta}^{m^\star_n}$. In other words, if the regularity of θ is known in advance, then the OSE $\widehat{\theta}^{m^\star_n}$ is a feasible estimator. □

§16.43 **GSSM** *(§16.32 continued)*. If $\widehat{\theta} \sim \mathrm{N}_{(\theta,n^{-1}\mathrm{id}_\mathbb{H})}$, where $\|\mathrm{id}_\mathbb{H}\|_{\mathbb{L}(\mathbb{H})} = 1$. From Proposition §16.40 we obtain immediately, $\sup\{\mathbb{P}^n_\theta(|\Phi(\widehat{\theta}^{m^\star_n} - \theta)|^2), \theta \in \mathbb{F}^r_{u_\bullet,\mathfrak{f}_\bullet}\} \leqslant (1 + r^2)\ \mathrm{R}^\star_n(\mathfrak{f}_\bullet)$ for all $n \in \mathbb{N}$. In other words the local Φ-risk of the OSE with optimally choosen dimension is not larger than $\mathrm{R}^\star_n(\mathfrak{f}_\bullet)$ (up to a constant) uniformely for all functions of interest belonging to $\mathbb{F}^r_{u_\bullet,\mathfrak{f}_\bullet}$. □

§16.44 **Nonparametric density estimation** *(§16.33 continued)*. Consider on $\mathbb{L}_2$ the stochastic process $\widehat{\mathbb{p}} = (\widehat{\mathbb{p}}_h)_{h\in\mathbb{L}_2}$ of real random variables defined on $([0,1]^n, \mathscr{B}^n_{[0,1]^n}, \mathbb{P}^{\otimes n}_\mathbb{p})$ by $\widehat{\mathbb{p}}_h := \widehat{\mathbb{P}}_n h \in \mathscr{B}^n$ for each $h \in \mathbb{L}_2$. If $\mathbb{p} \in \mathbb{L}_2 \cap \mathbb{L}_\infty$ and $u_\bullet$ is $\mathfrak{f}_\bullet$-regular (Definition §17.12 (b)) then $\widehat{\mathbb{p}} \sim \mathrm{P}_{(\mathbb{p},n^{-1}\Gamma^\mathbb{p})}$ with $\|\Gamma^\mathbb{p}\|_{\mathbb{L}(\mathbb{L}_2)} \leqslant \|\mathbb{p}\|_{\mathbb{L}_\infty}$ (see Example §15.07) and $\|\mathbb{p}\|_{\mathbb{L}_\infty} \leqslant r\tau_{u_\bullet,\mathfrak{f}_\bullet}$ for all $\mathbb{p} \in \mathbb{F}^r_{u_\bullet,\mathfrak{f}_\bullet}$ (see Example §16.10). As a consequence, the OSE $\widehat{\mathbb{p}}^{m^\star_n} = U^\star(\widehat{\mathbb{p}}_{u_\bullet}\mathbb{1}^{m^\star_n}_\bullet)$ with dimension $m^\star_n$ as in (16.5) satisfies $\sup\{\mathbb{P}^{\otimes n}_\mathbb{p}(|\Phi(\widehat{\mathbb{p}}^{m^\star_n} - \mathbb{p})|^2), \mathbb{p} \in \mathbb{F}^r_{u_\bullet,\mathfrak{f}_\bullet}\} \leqslant (r\tau_{u_\bullet,\mathfrak{f}_\bullet} + r^2)\ \mathrm{R}^\star_n(\mathfrak{f}_\bullet)$ for all $n \in \mathbb{N}$ by from Proposition §16.40. Consider as in Example §16.22 the *trigonometric basis* $\psi_\bullet = (\psi_j)_{j\in\mathbb{N}}$ and $\mathfrak{f}^{-1}_\bullet := w_\bullet \in (\mathbb{R}^+_{\setminus 0})^\mathbb{N}$ given either in Example §16.14 **(P)** or **(E)**. In this situation we $\sup\{\mathbb{P}^{\otimes n}_\mathbb{p}(|\Phi(\widehat{\mathbb{p}}^{m^\star_n} - \mathbb{p})|^2), \mathbb{p} \in \mathbb{D}_2 \cap \mathbb{F}^r_{\psi_\bullet,\mathfrak{f}_\bullet}\} \leqslant (\sqrt{2}r\|\mathfrak{f}_\bullet\|_{\ell_2} + r^2)\ \mathrm{R}^\star_n(\mathfrak{f}_\bullet)$ for all $n \in \mathbb{N}$, where $\mathrm{R}^\star_n(\mathfrak{f}_\bullet) = \mathrm{o}(1)$ as $n \to \infty$ if in addition $\Phi_{u_\bullet}\mathfrak{f}_\bullet \in \ell_2$ (Remark §16.42). □

§16.45 **Nonparametric regression** *(§16.34 continued)*. Consider the stochastic process $\widehat{f} = (\widehat{f}_h)_{h\in\mathbb{L}_2}$ on $\mathbb{L}_2$ of real valued random variables defined on $(\mathbb{R}^{2n}, \mathscr{B}^{2n}, \mathrm{U}^{\otimes n}_f)$ by $\widehat{f}_h := \widehat{\mathbb{P}}_n(Yh(X)) \in \mathscr{B}^{2n}$ for each $h \in \mathbb{L}_2$. Under Assumption §14.01 if $f \in \mathbb{L}_2 \cap \mathbb{L}_\infty$ then $\widehat{f} \sim \mathrm{P}_{(f,n^{-1}\Gamma^f)}$ with $\|\Gamma^f\|_{\mathbb{L}(\mathbb{L}_2)} \leqslant \sigma^2_\varepsilon + \|f\|^2_{\mathbb{L}_\infty}$ (see Example §15.08). From Proposition §16.18 we obtain immediately un upper bound for the gobal $\mathbb{L}_2$-risk (mise) which still depends on $\|f\|_{\mathbb{L}_\infty}$. If we assume in addition that the ONS $u_\bullet$ is $\mathfrak{f}_\bullet$-regular as in Definition §17.12 (b), then $\|f\|_{\mathbb{L}_\infty} \leqslant r\tau_{u_\bullet,\mathfrak{f}_\bullet}$ for

all $f \in \mathbb{F}^r_{u,\mathfrak{f}}$ by Lemma §17.15. As a consequence, considering the OSE $\widehat{f}^{m^\star_n} = U^\star(\widehat{f}_{u_\bullet} \mathbb{1}^{m^\star_n}_\bullet)$ with dimension $m^\star_n$ as in (16.3) from Proposition §16.18 we obtain, $\sup\{\mathbb{P}_f^{\otimes n} \| \widehat{f}^{m^\star_n} - \theta\|^2_{\mathbb{L}_2}, \theta \in \mathbb{F}^r_{u,\mathfrak{f}}\} \leqslant (\sigma^2_\varepsilon + r^2\tau^2_{u_\bullet,\mathfrak{f}_\bullet} + r^2)\ \mathrm{R}^\star_n(\mathfrak{f}_\bullet)$ for all $n \in \mathbb{N}$. Consider as in Example §16.23 the *trigonometric basis* $\psi_\bullet = (\psi_j)_{j\in\mathbb{N}}$ and $\mathfrak{f}^{-1}_\bullet := w_\bullet \in (\mathbb{R}^+_{\setminus 0})^{\mathbb{N}}$ given either in Example §16.14 **(P)** or **(E)**. In this situation we $\sup\{\mathbb{P}_f^{\otimes n}(|\Phi(\widehat{f}^{m^\star_n} - f)|^2), f \in \mathbb{F}^r_{\psi_\bullet,\mathfrak{f}_\bullet}\} \leqslant (\sigma^2_\varepsilon + r^2\tau^2_{u_\bullet,\mathfrak{f}_\bullet} + r^2)\ \mathrm{R}^\star_n(\mathfrak{f}_\bullet)$ for all $n \in \mathbb{N}$, where $\mathrm{R}^\star_n(\mathfrak{f}_\bullet) = \mathrm{o}(1)$ as $n \to \infty$ if in addition $\Phi_{u_\bullet}\mathfrak{f}_\bullet \in \ell_2$ (Remark §16.42). □

§16.46 **Illustration**. Let us consider $\Phi^2_{u_j} \sim j^{2a}$, $2a > -1$ and hence $\|\Phi_{u_\bullet}\mathbb{1}^m_\bullet\|^2_{\ell_2} \sim m^{2a+1}$, and the following two specifications:

(P) If $\mathfrak{f}^2_j \sim j^{-2s}$, $s > 0$, and hence $(\Phi_{u_\bullet}\mathfrak{f}_\bullet)^2_{(m)} \sim m^{-2(s-a)}$ for $s > a$, then $m^\star_n \sim n^{\frac{1}{2s+1}}$ and $\mathrm{R}^\star_n(\mathfrak{f}_\bullet) \sim n^{\frac{-2(s-a)}{2s+1}}$.

(E) If $\mathfrak{f}^2_j \sim j^{-2a}\exp(-j^{2s})$, $s > 0$, and hence $(\Phi_{u_\bullet}\mathfrak{f}_\bullet)^2_{(m)} \sim \exp(-m^{2s})$, then $m^\star_n \sim (\log n)^{\frac{1}{2s}}$ and $\mathrm{R}^\star_n(\mathfrak{f}_\bullet) \sim (\log n)^{\frac{2a+1}{2s}} n^{-1}$. □

§17 Supplementary materials

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].

§17.01 **Definition**. A normed real vector space $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ that is complete (in a Cauchy-sense) is called a real *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 $h \in \mathbb{H}$. □

§17.02 **Property**.

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

§17.03 **Example**.

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

(b) Denote by $\mathbb{R}^{\mathbb{N}}$ the vector space of all $\mathbb{R}$-valued sequences over $\mathbb{N}$ where we refer to any sequence $(a_j)_{j\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$ as a whole by $a_\bullet$ as for example in «the sequence $a_\bullet$» and arithmetic operations on sequences are defined component-wise, i.e., $a_\bullet b_\bullet = (a_j b_j)_{j\in\mathbb{N}}$, $a_\bullet \vee b_\bullet = (a_j \vee b_j := \max(a_j, b_j))_{j\in\mathbb{N}}$ $a_\bullet \wedge b_\bullet = (a_j \wedge b_j := \min(a_j, b_j))_{j\in\mathbb{N}}$ or $a_\bullet \leqslant cb_\bullet$ with $c \in \mathbb{R}^+$, if $a_j \leqslant cb_j$ for all $j \in \mathbb{N}$, for sequences $a_\bullet, b_\bullet \in \mathbb{R}^{\mathbb{N}}$. In the sequel, let $\|a_\bullet\|_{\ell_s} := (\sum_{j\in\mathbb{N}} |a_j|^s)^{1/s}$, for $s \in [1, \infty)$, and $\|a_\bullet\|_{\ell_\infty} := \sup\{|a_j|, j \in \mathbb{N}\}$. Thereby, for $s \in [1, \infty]$, consider $\ell_s(\mathbb{N}) := \{a_\bullet \in \mathbb{R}^{\mathbb{N}}, \|a_\bullet\|_{\ell_s} < \infty\}$, or ℓ_s for short, endowed with the norm $\|\cdot\|_{\ell_s}$. In particular, ℓ_2 is the usual *Hilbert space of square summable sequences* over $\mathbb{N}$ endowed with the inner product $\langle a_\bullet, b_\bullet\rangle_{\ell_2} := \sum_{j\in\mathbb{N}} a_j b_j$ for all $a_\bullet, b_\bullet \in \ell_2$.

(c) For a strictly positive sequence $w_\bullet \in (\mathbb{R}^+_{\setminus 0})^{\mathbb{N}}$ consider the *weighted norm* $\|a_\bullet\|^2_{\ell_2(w^2_\bullet)} := \sum_{j\in\mathbb{N}} w^2_j |a_j|^2$. We define $\ell_2(w^2_\bullet) := \{a_\bullet \in \mathbb{R}^{\mathbb{N}}, \|a_\bullet\|_{\ell_2(w^2_\bullet)} < \infty\}$, which is a Hilbert space

endowed with the inner product $\langle a_\bullet, b_\bullet\rangle_{\ell_2(w_\bullet^2)} := \langle w_\bullet a_\bullet, w_\bullet a_\bullet\rangle_{\ell_2} = \sum_{j\in\mathbb{N}} w_j^2 a_j b_j$ for all $a_\bullet, b_\bullet \in \ell_2(w_\bullet)$.

(d) For a measure space $(\Omega, \mathscr{A}, \mu)$ recall the set of all $\mathscr{L}_s(\mu)$-integrable functions given in Notation §01.03 for $s \in [1,\infty]$. The set of equivalence classes $\mathbb{L}_s(\mu) := \mathbb{L}_s(\mathscr{A},\mu) := \{\{h\}_\mu : h \in \mathscr{L}_s(\mathscr{A},\mu)\}$ is a vector space endowed with the norm $\|\{h\}_\mu\|_{\mathbb{L}_s(\mu)} := \|h\|_{\mathscr{L}_s(\mu)}$ for $\{h\}_\mu \in \mathbb{L}_s(\mu)$ (see Notation §15.06). As usual we identify a function $h \in \mathscr{L}_s(\mu)$ with its equivalence class $\{h\}_\mu$. For instance, $\mathbb{L}_2(\mathscr{A},\mu) = \{h \in \mathscr{A} : \|h\|_{\mathbb{L}_2(\mu)}\}$, or $\mathbb{L}_2(\mu)$ for short, denotes the usual *Hilbert space of square μ-integrable in $\mathscr{A}$* endowed with the inner product $\langle h, h_o\rangle_{\mathbb{L}_2(\mu)} := \mu(hh_o)$ for all $h, h_o \in \mathbb{L}_2(\mu)$. □

§17.04 **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 $\mathcal{U} \subseteq \mathcal{U}'$ and $\mathcal{U}'$ is ONS, then $\mathcal{U} = \mathcal{U}'$, i.e., if it is a *complete* ONS.

§17.05 **Example**. Consider the real Hilbert space $\mathbb{L}_2(\mathscr{B}_{[0,1]}, \lambda_{[0,1]})$ with respect to the restriction $\lambda_{[0,1]}$ of the Lebesgue measure to $\mathscr{B}_{[0,1]}$. With a slight abuse of notations we write shortly $\lambda := \lambda_{[0,1]}$ and $\mathbb{L}_2 := \mathbb{L}_2(\mathscr{B}_{[0,1]}, \lambda)$. The *trigonometric basis* given for $t \in [0,1]$ by

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

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

§17.06 **Property**.

(Pythagorean formula) *If $\{h_j, j \in [\![n]\!]\} \subseteq \mathbb{H}$ are orthogonal, then $\|\sum_{j\in[\![n]\!]} h_j\|_\mathbb{H}^2 = \sum_{j\in[\![n]\!]}\|h_j\|_\mathbb{H}^2$.*

(Bessel's inequality) *If $\mathcal{U} \subseteq \mathbb{H}$ 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 $\mathcal{U} \subseteq \mathbb{H}$ is complete if and only if $\|h\|_\mathbb{H}^2 = \sum_{u\in\mathcal{U}} |\langle h, u\rangle_\mathbb{H}|^2$ for all $h \in \mathbb{H}$.* □

§17.07 **Definition**. Let $\mathcal{U}$ be a subset of a Hilbert space $(\mathbb{H}, \langle\cdot,\cdot\rangle_\mathbb{H})$. Denote by $\mathbb{U} := \overline{\text{lin}}(\mathcal{U})$ the closure of the linear subspace spanned by the elements of $\mathcal{U}$. Its orthogonal complement in $(\mathbb{H}, \langle\cdot,\cdot\rangle_\mathbb{H})$ is defined by $\mathbb{U}^\perp := \{h \in \mathbb{H} : \langle h, u\rangle_\mathbb{H} = 0, \forall u \in \mathbb{U}\}$ where $\mathbb{H} = \mathbb{U} \oplus \mathbb{U}^\perp$. □

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

§17.09 **Definition**. A sequence $u_\bullet = (u_j)_{j\in\mathbb{N}}$ in $\mathbb{H}$ is said to be an *orthonormal sequence* (ONS), respectively, an orthonormal basis (ONB) if the subset $\{u_j, j \in \mathbb{N}\}$ is an ONS, respectively ONB. The Hilbert space $\mathbb{H}$ is called *separable*, if there exists a complete orthonormal sequence. □

§17.10 **Example**. The Hilbert space $(\mathbb{R}^k, \langle\cdot,\cdot\rangle_W)$, $(\ell_2(w_\bullet), \langle\cdot,\cdot\rangle_{\ell_2(w_\bullet)})$ and $(\mathbb{L}_2(\mu), \langle\cdot,\cdot\rangle_{\mathbb{L}_2(\mu)})$ with σ-finite measure μ are separable. □

§17.11 **Definition**. A family $([\![m]\!])_{m\in\mathcal{M}}$, $\mathcal{M} \subseteq \mathbb{N}$, is called a *nested sieve in $\mathbb{N}$*, if $\cup_{m\in\mathcal{M}}[\![m]\!] = \mathbb{N}$. We write $[\![m]\!]^c := \mathbb{N}\backslash[\![m]\!] = (m,\infty) \cap \mathbb{N}$ for $m \in \mathcal{M}$. Similarly, given an ONS $u_\bullet = (u_j)_{j\in\mathbb{N}}$ and setting $\mathbb{U}_m := \overline{\text{lin}}\{u_j, j \in [\![m]\!]\}$ for $m \in \mathcal{M}$ we call the family $(\mathbb{U}_m)_{m\in\mathcal{M}}$ *a nested sieve in*

$\mathbb{U} := \overline{\text{lin}}\{u_j, j \in \mathbb{N}\}$. We write $\mathbb{U}_m^\perp := \overline{\text{lin}}\{u_j, j \in [\![m]\!]^c\}$ where $\mathbb{U} = \mathbb{U}_m \oplus \mathbb{U}_m^\perp$. For convenient notations we set further $\mathbb{1}_\bullet^m := (\mathbb{1}_j^m)_{j\in\mathbb{N}}$ with $\mathbb{1}_j^m := \mathbb{1}_{[\![m]\!]}(j)$ for $j \in \mathbb{N}$. and $\mathbb{1}_\bullet := (1)_{j\in\mathbb{N}}$. □

§17.12 **Definition**. We call an ONS $u_\bullet = (u_j)_{j\in\mathbb{N}}$ in $\mathbb{L}_2(\mu)$ (respectively, in ℓ_2) *regular*

(a) with respect to a nested sieve $([\![m]\!])_{m\in\mathcal{M}}$, if there is a finite constant $\tau_u \geqslant 1$ satisfying $\|\sum_{j\in[\![m]\!]} |u_j|^2\|_{\mathbb{L}_\infty(\mu)} \leqslant \tau_u^2 m$ for all $m \in \mathcal{M}$;

(b) with respect to a weight sequence $w_\bullet$, *$w_\bullet$-regular* in short, if there exists a finite constant $\tau_{uw_\bullet} \geqslant 1$ such that $\|\sum_{j\in\mathbb{N}} w_j^2|u_j|^2\|_{\mathbb{L}_\infty\mu} \leqslant \tau_{uw_\bullet}^2$. □

§17.13 **Remark**. According to Lemma 6 of Birgé and Massart [1997] assuming in $\mathbb{L}_2$ a regular ONS $(u_j)_{j\in\mathbb{N}}$ the nested sieve $([\![m]\!])_{m\in\mathbb{N}}$ is exactly equivalent to following property: there exists a finite constant $\tau_u \geqslant 1$ such that for any $h \in \mathbb{U}_m$ we have $\|h\|_{\mathbb{L}_\infty} \leqslant \tau_u\sqrt{m}\|h\|_{\mathbb{L}_2}$. Typical example are bounded basis, such as the trigonometric basis, or basis obeying the following property: there is $C_\infty \in \mathbb{R}_{\setminus 0}^+$ such that for any $c^m \in \mathbb{R}^m$ yields $\|\sum_{j\in[\![m]\!]} c_j u_j\|_{\mathbb{L}_\infty} \leqslant C_\infty\sqrt{m}\max_{j\in[\![m]\!]} c_j$. Birgé and Massart [1997] have shown that the last property is satisfied for piece-wise polynomials, splines and wavelets. □

§17.14 **Example** *(Example §17.05 continued).* Consider the *trigonometric basis* $\psi_\bullet = (\psi_j)_{j\in\mathbb{N}}$ in the real Hilbert space $\mathbb{L}_2 = \mathbb{L}_2(\mathscr{B}_{[0,1]}, \lambda_{[0,1]})$. Since $\sup_{j\in\mathbb{N}}\|\psi_j\|_{\mathbb{L}_\infty} \leqslant \sqrt{2}$ setting $\tau_u^2 := 2$ the trigonometric basis is regular w.r.t. any nested Sieve $([\![m]\!])_{m\in\mathcal{M}}$, i.e., Definition §17.12 (a) holds with $\|\sum_{j\in[\![m]\!]} \psi_j^2\|_{\mathbb{L}_\infty} \leqslant 2m$. In the particular case of the nested sieve $([\![1+2m]\!])_{m\in\mathbb{N}}$, we have $\sum_{j\in[\![1+2m]\!]} \psi_j^2 = \mathbb{1} + \sum_{j\in[\![m]\!]}\{2\sin^2(2\pi j\bullet) + 2\cos^2(2\pi j\bullet)\} = 1 + 2m$ and thus, the trigonometric basis is regular with $\tau_u := 1$. Moreover, the trigonometric basis is regular with respect to any $w_\bullet \in \ell_2$. Indeed, in this situation we have $\|\sum_{j\in\mathbb{N}} w_j^2\psi_j^2\|_{\ell_\infty} \leqslant 2\|w_\bullet\|_{\ell_2}^2$ and hence Definition §17.12 (b) holds with $\tau_{uw_\bullet}^2 = 2\|w_\bullet\|_{\ell_2}^2$. □

§17.15 **Lemma**. *Let $\mathbb{F}_{u_\bullet,\mathfrak{f}_\bullet}^r$ be a class of functions an ONS $u_\bullet = (u_j)_{j\in\mathbb{N}}$ in $\mathbb{L}_2(\mu)$ (or analogously in ℓ_2) as given in Definition §16.15. If the ONS is regular wrt the weight sequence $\mathfrak{f}_\bullet$ as in Definition §17.12 (b) for some finite constant $\tau_{u\mathfrak{f}_\bullet} \geqslant 1$, then for each $f \in \mathbb{F}_{u_\bullet,\mathfrak{f}_\bullet}^r$ holds $\|f\|_{\mathbb{L}_\infty(\mu)} \leqslant \tau_{uw_\bullet}\|f\|_{1/\mathfrak{f}_\bullet} \leqslant r\tau_{uw_\bullet}$.*

§17.16 **Proof** of Lemma §17.15. Due to the Cauchy-Schwarz inequality (Property §17.02) for each $f \in \mathbb{F}_{u_\bullet,\mathfrak{f}_\bullet}^r$ we have $\|f\|_{\mathbb{L}_\infty(\mu)}^2 \leqslant \|f\|_{1/\mathfrak{f}_\bullet}^2\|\sum_{j\in\mathbb{N}} \mathfrak{f}_j^2 u_j^2\|_{\mathbb{L}_\infty(\mu)}$, which in turn implies the assertion by employing the Definition §17.12 (b) of $\tau_{u\mathfrak{f}_\bullet}$ and r. □

§17.17 **Example** *(Example §16.14 continued).* Consider $\mathbb{L}_2^{w_\bullet}$ with respect to the *trigonometric basis* $\psi_\bullet = (\psi_j)_{j\in\mathbb{N}}$ and a weight sequence $w_\bullet$ satisfying either Example §16.14 **(P)** with $p > 1/2$ or Example §16.14 **(E)** with $p > 0$. In both cases setting $\tau_{\psi_\bullet,w_\bullet}^2 = 2\|1/w_\bullet\|_{\ell_2}^2 < \infty$ the trigonometric basis is regular w.r.t. the weight sequence $1/w_\bullet$. Consequently, setting $\mathfrak{f}_\bullet = 1/w_\bullet$ from Lemma §17.15 follows $\sup\{\|f\|_{\mathscr{L}_\infty}^2, \theta \in \mathbb{L}_2^{1/\mathfrak{f}_\bullet}\} \leqslant 2\|\theta\|_{\psi_\bullet,1/\mathfrak{f}_\bullet}^2\|\mathfrak{f}_\bullet\|_{\ell_2}^2$. □

§17.18 **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 $h_1, h_2 \in \mathbb{H}, a, b \in \mathbb{R}$. Its *domain* will be denoted by $\mathcal{D}(T)$, its *range* by $\mathcal{R}(T)$ and its *null space* by $\mathcal{N}(T)$. □

§17.19 **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.* □

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

§17.21 **Example**.

(a) Let $u_\bullet = (u_j)_{j\in\mathbb{N}}$ be an ONS in $\mathbb{H}$, for any $h \in \mathbb{H}$ and $j \in \mathbb{N}$ we call $h_{u_j} := \langle h, u_j\rangle_{\mathbb{H}}$ *generalised Fourier coefficient*. We write $h_{u_\bullet} := (h_{u_j})_{j\in\mathbb{N}}$ for short. The associated *(generalised) Fourier series transform* $U : \mathbb{H} \to \mathbb{R}^{\mathbb{N}}$ defined by $h \mapsto Uh := h_{u_\bullet}$ belongs to $\mathscr{L}(\mathbb{H}, \ell_2)$ with $\|U\|_{\mathbb{L}(\mathbb{H},\ell_2)} = 1$.

(b) Consider a measure space $(\Omega, \mathscr{A}, \mu)$ and a function $f \in \mathscr{A}$ the map $\mathrm{M}_f : \mathscr{A} \to \mathscr{A}$ with $h \mapsto \mathrm{M}_f h := hf$ is called *multiplication operator*. If $\|f\|_{\mathscr{L}_\infty(\mu)} < \infty$ then we have $\mathrm{M}_f \in \mathbb{L}(\mathbb{L}_2(\mu))$ with $\|\mathrm{M}_f\|_{\mathbb{L}(\mathbb{L}_2(\mu))} \leqslant \|f\|_{\mathscr{L}_\infty(\mu)} < \infty$.

§17.22 **Definition**. A (linear) map $\Phi : \mathbb{H} \supset \mathcal{D}(\Phi) \to \mathbb{R}$ is called *(linear) functional* and given an ONS $u_\bullet = (u_j)_{j\in\mathbb{N}}$ in $\mathbb{H}$ which belongs to $\mathcal{D}(\Phi)$ we set $\Phi_{u_\bullet} := (\Phi_{u_j})_{j\in\mathbb{N}}$ with the slight abuse of notations $\Phi_{u_j} := \Phi(u_j)$, $j \in \mathbb{N}$. In particular, if $\Phi \in \mathbb{L}(\mathbb{H}, \mathbb{R})$ then $\mathcal{D}(\Phi) = \mathbb{H}$. □

§17.23 **Property**. *Let $\Phi \in \mathbb{L}(\mathbb{H}, \mathbb{R})$.*

(Fréchet-Riesz representation) There exists $\phi \in \mathbb{H}$ such that $\Phi(h) = \langle\phi, h\rangle_{\mathbb{H}}$ for all $h \in \mathbb{H}$, and hence, given an ONS $u_\bullet = (u_j)_{j\in\mathbb{N}}$ in $\mathbb{H}$ we have $\Phi_{u_j} = \Phi(u_j) = \langle\phi, u_j\rangle_{\mathbb{H}} = \phi_{u_j}$ for all $j \in \mathbb{N}$, or $\Phi_{u_\bullet} = \phi_{u_\bullet}$ for short. □

§17.24 **Example**. Consider an ONB $u_\bullet = (u_j)_{j\in\mathbb{N}}$ in $\mathbb{L}_2(\Omega, \mathscr{A}, \mu)$ (or analogously in $\ell_2(\mathbb{N})$). By *evaluation at a point* $t_o \in \Omega$ we mean the linear functional $\Phi_{t_o} : \mathbb{L}_2(\mu) \supset \mathcal{D}(\Phi_{t_o}) \to \mathbb{R}$ with $h \mapsto h(t_o) := \Phi_{t_o}(h) = \sum_{j\in\mathbb{N}} h_{u_j} u_j(t_o)$. Obviously, a point evaluation of h at t_o is well-defined, if $\sum_{j\in\mathbb{N}} |h_{u_j} u_j(t_o)| < \infty$. Observe that the point evaluation at t_o is generally not bounded on the subset $\{h \in \mathbb{L}_2(\mu) : \sum_{j\in\mathbb{N}} |h_{u_j} u_j(t_o)| < \infty\}$. □

§17.25 **Definition**. For each $T \in \mathbb{L}(\mathbb{H}, \mathbb{G})$ there is a uniquely determined *adjoint operator* $T^\star \in \mathbb{L}(\mathbb{G}, \mathbb{H})$ satisfying $\langle Th, g\rangle_{\mathbb{G}} = \langle h, T^\star g\rangle_{\mathbb{H}}$ for all $h \in \mathbb{H}, g \in \mathbb{G}$. □

§17.26 **Property**. *Let $S, T \in \mathbb{L}(\mathbb{H}_1, \mathbb{H}_2)$ and $R \in \mathbb{L}(\mathbb{H}_2, \mathbb{H}_3)$. Then we have*

(i) $(S + T)^\star = S^\star + T^\star$, $(RS)^\star = S^\star R^\star$.

(ii) $\|S^\star\|_{\mathbb{L}(\mathbb{H}_2,\mathbb{H}_1)} = \|S\|_{\mathbb{L}(\mathbb{H}_1,\mathbb{H}_2)}$, $\|SS^\star\|_{\mathbb{L}(\mathbb{H}_2,\mathbb{H}_2)} = \|S^\star S\|_{\mathbb{L}(\mathbb{H}_1,\mathbb{H}_1)} = \|S\|^2_{\mathbb{L}(\mathbb{H}_1,\mathbb{H}_2)}$.

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

§17.27 **Example**.

(a) The adjoint of a $(k \times m)$ matrix M is its $(m \times k)$ transpose matrix M^t.

(b) Let $\mathrm{M}_f \in \mathbb{L}(\mathbb{L}_2(\mu))$ be a *multiplication operator*, then its adjoint equals also a multiplication with f, i.e. $\mathrm{M}_f^\star = \mathrm{M}_f$. □

§17.28 **Definition**. Let $\mathbb{H}$ and $\mathbb{G}$ be Hilbert spaces.

(a) The *identity* in $\mathbb{L}(\mathbb{H})$ is denoted by $\mathrm{id}_{\mathbb{H}}$.

(b) If $T \in \mathbb{L}(\mathbb{H}, \mathbb{G})$, then $T : \mathcal{N}(T)^\perp \to \mathcal{R}(T)$ 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 $T^{-1} \in \mathbb{L}(\mathbb{G}, \mathbb{H})$ satisfies $\mathrm{id}_{\mathbb{G}} = TT^{-1}$ and $\mathrm{id}_{\mathbb{H}} = T^{-1}T$.

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

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

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

(f) $\Pi \in \mathbb{L}(\mathbb{H})$ is called *projection* if $\Pi^2 = \Pi\Pi = \Pi$. For $\Pi \neq 0$ are equivalent: (i) Π is an orthogonal projection ($\mathbb{H} = \mathcal{R}(\Pi) \oplus \mathcal{N}(\Pi)$); (ii) $\|\Pi\|_{\mathbb{L}(\mathbb{H})} = 1$; (iii) $\Pi \in \mathbb{L}^+(\mathbb{H})$. □

§17.29 **Property**.

(i) *If $T \in \mathbb{L}(\mathbb{H})$ is self-adjoint, then $\|T\|_{\mathbb{L}(\mathbb{H})} = \sup\{|\langle Th, h\rangle_{\mathbb{H}}| : \|h\|_{\mathbb{H}} \leqslant 1, h \in \mathbb{H}\}$.*

(ii) *If $T \in \mathbb{L}^+(\mathbb{H})$ then there exists $T^{1/2} \in \mathbb{L}^+(\mathbb{H})$ with $T = T^{1/2}T^{1/2}$.*

§17.30 **Example** *(Example §17.21 continued).*

(a) Let $u_\bullet = (u_j)_{j\in\mathbb{N}}$ be an ONS in $\mathbb{H}$ and set $\mathbb{U} := \overline{\mathrm{lin}}\,\{u_j, j \in \mathbb{N}\}$. The *(generalised) Fourier series transform* $U \in \mathbb{L}(\mathbb{H}, \ell_2)$ (see Example §17.21 (a)) is a partial isometry with adjoint $U^\star \in \mathbb{L}(\mathbb{H}, \ell_2)$ satisfying $U^\star a_\bullet = \sum_{j\in\mathbb{N}} a_j u_j$ for $a_\bullet \in \ell_2(\mathbb{N})$, i.e., $U : \mathbb{U} \to \ell_2(\mathbb{N})$ is unitary. Moreover, the orthogonal projection $\Pi_{\mathbb{U}} \in \mathbb{L}(\mathbb{H})$ onto $\mathbb{U}$ satisfies $\Pi_{\mathbb{U}} h = U^\star U h = \sum_{j\in\mathbb{N}} h_{u_j} u_j$ for all $h \in \mathbb{H}$. If $u_\bullet = (u_j)_{j\in\mathbb{N}}$ is complete (i.e. ONB), then U is invertible with $UU^\star = \mathrm{id}_{\ell_2(\mathbb{N})}$ and $U^\star U = \mathrm{id}_{\mathbb{H}}$ due to Parseval's formula, and hence U is unitary.

(b) A *multiplication operator* $\mathrm{M}_f \in \mathbb{L}(\mathbb{L}_2(\mu))$ (see Example §17.21 (b)) is self-adjoint and if $f \in \mathscr{A}$ is non-negative, i.e. $f \in \mathscr{A}^+$, then M_f is non-negative, i.e, $\mathrm{M}_f \in \mathbb{L}^+_{\setminus 0}(\mathbb{L}_2(\mu))$. □

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.

H.-O. Georgii. *Stochastik*. De Gruyter, 5. Auflage, 2015.

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

A. Klenke. *Probability theory. A comprehensive course.* London: Springer, 2008.

A. Klenke. *Wahrscheinlichkeitstheorie*. Springer Spektrum, 3., überarbeitete und ergänzte Auflage, 2012.

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.

E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956.

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

A. W. van der Vaart. *Asymptotic statistics.* Cambridge University Press, 1998.

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

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

H. Witting. Mathematische Statistik I: Parametrische Verfahren bei festem Stichprobenumfang. Stuttgart: B. G. Teubner, 1985.

H. Witting and U. Müller-Funk. *Mathematische Statistik II. Asymptotische Statistik: Parametrische Modelle und nichtparametrische Funktionale.* Stuttgart: B. G. Teubner, 1995.