Nichtparametrische Statistik
Gliederung zur Vorlesung im Sommersemester 2017

Martin Wahl
Humboldt-Universität zu Berlin
martin.wahl@math.hu-berlin.de

14. August 2017

Inhaltsverzeichnis

1 Dichteschätzung

In diesem Kapitel werden wir am Beispiel der Dichteschätzung einige grundlegende Themen der nichtparametrischen Statistik behandeln, wie z.B. Schätzmethoden, Konvergenzraten, untere Schranken und Wahl des Glättungsparameters. Als begleitende Literatur empfehlen wir Reiß (2012), Tsybakov (2009), Wasserman (2007) und van der Vaart (1998).

1.1 Modell

Seien $X_1, \dots, X_n$ unabhängige und identisch verteilte Zufallsvariablen auf $(\Omega, \mathscr{F}, \mathbb{P})$ mit Werten in $(\mathbb{R}, \mathscr{B}_\mathbb{R})$ und Verteilung P. Wir nehmen an, dass P eine Dichte bezüglich des Lebesgue-Maßes besitzt, das heißt es existiert ein $f \in L^1(\mathbb{R})$, $f \geq 0$, mit

$$\mathbb{P}(X_i \in A) = P(A) = \int_A f(x)\,dx \qquad \forall A \in \mathscr{B}_\mathbb{R}.$$

Die Verteilung P ist unbekannt (daher auch f). Wir kennen lediglich $X_1, \dots, X_n$, also auch die empirische Verteilung

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$$

wobei δ_x das Diracmaß in x bezeichne. Ziel ist es f zu schätzen. Dabei ist ein Schätzer von f eine Abbildung $x \mapsto \hat{f}_n(x) = \hat{f}_n(x, X_1, \dots, X_n)$ mit $\hat{f}_n : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ Borel-messbar. Um die Güte eines Schätzers zu messen, betrachten wir unter anderem folgende Risiken (sofern wohldefiniert):

Punktweises quadratisches Risiko (MSE): $\mathbb{E}(\hat{f}_n(x) - f(x))^2$ für $x \in \mathbb{R}$,

Quadratisches Risiko (MISE): $\mathbb{E}\int_\mathbb{R}(\hat{f}_n(x) - f(x))^2\,dx$,

Gleichmäßiges Risiko: $\mathbb{E}\sup_{x\in I}|\hat{f}_n(x) - f(x)|$ für ein Intervall $I \subseteq \mathbb{R}$.

1.2 Kerndichteschätzer

1.1 Definition. Eine Funktion $K \in L^1(\mathbb{R})$ mit $\int_{\mathbb{R}} K(u)\,du = 1$ heißt Kern oder Kernfunktion (kernel). Für einen Kern K und eine Bandweite $h > 0$ setzen wir

$$K_h(u) := h^{-1}K(h^{-1}u), \quad u \in \mathbb{R},$$

so dass K_h wiederum ein Kern ist.

1.2 Definition. Für einen Kern K und eine Bandweite h ist der Kerndichteschätzer definiert durch

$$\hat{f}_{n,h}(x) = \hat{f}^K_{n,h}(x) = \frac{1}{n}\sum_{i=1}^n K_h(x - X_i) = \frac{1}{nh}\sum_{i=1}^n K_h\left(\frac{x - X_i}{h}\right), \quad x \in \mathbb{R}.$$

1.3 Beispiele.

(a) $K(u) = (1/2)\mathbf{1}_{[-1,1]}(u)$ Rechteckkern.

(b) Für $K(u) = (1/2)\mathbf{1}_{[-1,1)}(u)$ gilt

$$\hat{f}_{n,h}(x) = \frac{1}{2nh}\sum_{i=1}^n \mathbf{1}_{(x-h,x+h]}(X_i) = \frac{F_n(x+h) - F_n(x-h)}{2h}$$

mit empirischer Verteilungsfunktion $F_n(y) = (1/n)\sum_{i=1}^n \mathbf{1}_{(-\infty,y]}(X_i)$.

(c) $K(u) = \frac{3}{4}(1 - u^2)\mathbf{1}_{[-1,1]}(u)$ Epanechnikov-Kern.

(d) Jede Wahrscheinlichkeitsdichte ist ein Kern, insbesondere der Gaußkern $K(x) = (2\pi)^{-1/2}e^{-x^2/2}$.

Man kann $\hat{f}_{n,h}(x)$ als Glättungsmethode interpretieren. Rund um jede Beobachtung wird auf glatte Art und Weise jeweils die Masse $1/n$ verteilt. Eine weitere Interpretation geht über die Approximation von Funktionen. Auf den ersten Blick scheint es fast unmöglich zu sein einen universalen Schätzer von f zu konstruieren, da die Menge aller Wahrscheinlichkeitsdichten sehr groß ist. Wir können jedoch f durch die Faltung (convolution) $K_h * f$ approximieren, wobei

$$K_h * f(x) := \int K_h(x - y)f(y)\,dy = \int K_h(x - y)\,P(dy)$$

(diese ist für Lebesgue-fast alle x definiert; ist f zusätzlich beschränkt, so ist $K_h * f(x)$ für alle x definiert). Gilt zum Beispiel $K = \mathbf{1}_{[-1/2,1/2]}$, so nähert sich K_h für $h \to 0$ immer mehr dem Diracmaß δ_0 an. Intuitiv erwarten wir also $K_h * f \sim \delta_0 * f = f$. Die folgene Übungsaufgabe formalisiert diese Beobachtung:

1.4 Aufgabe. *Sei K ein Kern und $f : \mathbb{R} \to \mathbb{R}$ eine meßbare Funktion.*

(a) Ist f beschränkt und stetig in x, so gilt $K_h * f(x) \to f(x)$ für $h \to 0$.

(b) Ist $f \in L^p(\mathbb{R})$, $p \geq 1$, so gilt $\|K_h * f - f\|_{L^p} \to 0$ für $h \to 0$.

Die grundlegende Beobachtung ist nun, dass wir das Integral $K_h * f(x)$ auf natürliche Art und Weise durch

$$\hat{f}_{n,h}(x) = \int K_h(x-y)\, P_n(dy) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)$$

schätzen können und zwar erwartungstreu, das heißt es gilt $\mathbb{E}\hat{f}_{n,h}(x) = K_h * f(x)$. Dies führt auf folgende Zerlegung von $\hat{f}_{n,h}(x) - f(x)$ in einen Schätz- und in einen Approximationsfehler:

$$\hat{f}_{n,h}(x) - f(x) = \frac{1}{n}\sum_{i=1}^{n}(K_h(x - X_i) - \mathbb{E}K_h(x - X_i)) + K_h * f(x) - f(x).$$

1.3 Punktweises Risiko

Ist $\hat{f}_{n,h}(x) \in L^2$, so gilt die folgende Bias-Varianz-Zerlegung für das punktweise Risiko:

$$\begin{aligned}\mathbb{E}(\hat{f}_{n,h}(x) - f(x))^2 &= \operatorname{Var}(\hat{f}_{n,h}(x)) + (\mathbb{E}\hat{f}_{n,h}(x) - f(x))^2 \\ &= \operatorname{Var}(\hat{f}_{n,h}(x)) + (K_h * f(x) - f(x))^2. \qquad (1.1)\end{aligned}$$

Wir beginnen mit dem folgendem Konsistenz-Resultat für das punktweise Risiko:

1.5 Satz. *Die Dichte f sei beschränkt und stetig in $x \in \mathbb{R}$. Außerdem sei K ein Kern mit $K \in L^2(\mathbb{R})$ und (h_n) eine Folge mit $h_n \to 0$ und $nh_n \to \infty$ für $n \to \infty$. Dann gilt*

$$\mathbb{E}(\hat{f}_{n,h_n}(x) - f(x))^2 \to 0 \qquad \text{für} \qquad n \to \infty.$$

Beweis. Wir schreiben $h = h_n$, das heißt wir unterdrücken in der Notation die Abhängigkeit von n. Für den Varianzterm gilt

$$\begin{aligned}\operatorname{Var}(\hat{f}_{n,h}(x)) &= \frac{1}{n}\operatorname{Var}(K_h(x - X_1)) \\ &\leq \frac{1}{n}\mathbb{E}K_h^2(x - X_1) \\ &= \frac{1}{nh^2}\int K^2\left(\frac{x-y}{h}\right) f(y)\, dy \leq \frac{\|f\|_\infty}{nh}\int K^2(u)\, du\end{aligned}$$

und der der letzte Ausdruck konvergiert gegen 0 für $n \to \infty$. Außerdem gilt $K_h * f(x) - f(x) \to 0$ für $n \to \infty$ nach Aufgabe 1.4 (a). Die Behauptung folgt nun aus der Bias-Varianz-Zerlegung in (1.1). □

Die Konvergenz in Satz 1.5 kann beliebig langsam sein. Unter zusätzlichen Annahmen an f kann der MSE jedoch weiter kontrolliert werden. Ein vorläufiges Resultat ist wie folgt:

1.6 Satz. *Die Dichte f sei zweimal stetig differenzierbar mit f und f'' beschränkt. Außerdem sei K ein symmetrischer Kern mit $K \in L^2(\mathbb{R})$ und $\int u^2|K(u)|\,du < \infty$. Dann gilt*

(a)

$$\mathrm{Var}(\hat{f}_{n,h}(x)) \leq \frac{\|f\|_\infty}{nh} \int K^2(u)\,du \qquad \forall x \in \mathbb{R}.$$

Falls $nh \to \infty$ für $n \to \infty$, so gilt außerdem

$$nh\,\mathrm{Var}(\hat{f}_{n,h}(x)) \to f(x) \int K^2(u)\,du \qquad \text{für} \qquad n \to \infty.$$

(b)

$$|\mathbb{E}\hat{f}_{n,h}(x) - f(x)| \leq \frac{h^2\|f''\|_\infty}{2} \int u^2|K(u)|\,du \qquad \forall x \in \mathbb{R}.$$

1.7 Korollar. *Es gelten die Voraussetzungen aus Satz 1.6 mit $\|f\|_\infty, \|f''\|_\infty \leq L$. Setze $h_n = cn^{-1/5}$ mit einer Konstante $c > 0$. Dann gilt*

$$\mathbb{E}(\hat{f}_{n,h}(x) - f(x))^2 \leq Cn^{-4/5}$$

mit einer Konstanten C die nur von c, K und L abhängt.

Beweis. Folgt aus der Bias-Varianz-Zerlegung und Satz 1.6. □

1.8 Korollar. *Es gelten die Voraussetzungen aus Satz 1.6. Sei K zusätzlich beschränkt. Es gelte $nh \to \infty$ und $n^{1/5}h \to 0$ für $n \to \infty$. Dann gilt*

$$\sqrt{nh}(\hat{f}_{n,h}(x) - f(x)) \to^d \mathcal{N}\Big(0, f(x) \int K^2(u)\,du\Big).$$

Korollar 1.8 kann verwendet werden um ein asymptotisches Konfidenzintervall $I_n(x)$ für $f(x)$ zum Niveau $\alpha > 0$, d.h. mit $\mathbb{P}(f(x) \in I_n(x)) \to 1 - \alpha$, zu konstruieren (Übung).

Beweis. Für den Bias gilt nach Satz 1.6 (c), dass

$$\sqrt{nh}|\mathbb{E}\hat{f}_{n,h}(x) - f(x)| \leq Cn^{1/2}h^{5/2} \to 0 \qquad \text{für } n \to \infty.$$

Daher reicht es zu zeigen, dass

$$\begin{aligned}&\sqrt{nh}(\hat{f}_{n,h}(x) - \mathbb{E}\hat{f}_{n,h}(x)) \\ &= \sqrt{nh}\frac{1}{n}\sum_{i=1}^{n}(K_h(x - X_i) - \mathbb{E}K_h(x - X_i)) \to^d \mathcal{N}\Big(0, f(x) \int K^2(u)\,du\Big).\end{aligned}$$

Wir wenden nun den zentralen Grenzwertsatz nach Lindeberg an, welcher folgendes besagt: Sind für alle $n \geq 1$ $Y_{n,1}, \ldots, Y_{n,n}$ unabhängige Zufallsvariablen mit (i) $\mathbb{E}Y_{n,i} = 0$, $\mathbb{E}Y_{n,i}^2 < \infty$, (ii) $\sum_{i=1}^n \mathbb{E}Y_{n,i}^2 \to \sigma^2$ und (iii) $\sum_{i=1}^n \mathbb{E}Y_{n,i}^2 \mathbf{1}(|Y_{n,i}| > \epsilon) \to 0$ für alle $\epsilon > 0$, so gilt $\sum_{i=1}^n Y_{n,i} \to^d \mathcal{N}(0, \sigma^2)$. Wir setzen

$$Y_{i,n} = \sqrt{nh}\frac{1}{n}\left(K_h(x - X_i) - \mathbb{E}K_h(x - X_i)\right)$$

an. Dann ist (i) klar, (ii) folgt aus Satz 1.6 (a) und (iii) aus der Tatsache, dass $|Y_{n,i}| \leq 2\|K\|_\infty/\sqrt{nh} \to 0$ für $n \to \infty$. □

1.9 Definition. Sei $I \subseteq \mathbb{R}$ ein Intervall und $\alpha, L > 0$. Setze $l = \lfloor \alpha \rfloor = \max\{m \in \mathbb{N} : m < \alpha\}$ und $C^l(I) = \{f : I \to \mathbb{R} : f \text{ } l\text{-mal stetig differenzierbar}\}$. Dann heißt die Menge

$$\mathcal{H}^\alpha(I; L) = \left\{ f \in C^l(I) : \sup_{x \in I} |f(x)| + \sup_{x \neq y, x, y \in I} \frac{|f^{(l)}(x) - f^{(l)}(y)|}{|x - y|^{\alpha - l}} \leq L \right\}$$

Hölder-Kugel auf I mit Parametern $\alpha, L > 0$. Gilt $I = \mathbb{R}$, so schreiben wir auch $\mathcal{H}^\alpha(L) = \mathcal{H}^\alpha(\mathbb{R}; L)$.

1.10 Definition. Ein Kern $K : \mathbb{R} \to \mathbb{R}$ ist von der Ordnung $m \in \mathbb{N}$, sofern für alle $1 \leq k \leq m$ gilt

$$\int_{\mathbb{R}} u^k K(u)\, du = 0.$$

1.11 Aufgabe. *Für jedes $m \geq 1$ existiert genau ein Polynom vom Grad $\leq m$, so dass*

$$K(u) = P(u)\mathbf{1}_{[-1,1]}(u)$$

ein Kern von der Ordnung m ist.

1.12 Proposition. *Sei $f \in \mathcal{H}^\alpha(L)$ und K ein Kern der Ordnung $\lfloor \alpha \rfloor$ mit $\int |u|^\alpha K(u)\, du < \infty$. Dann gilt*

$$|K_h * f(x) - f(x)| \leq h^\alpha \frac{L}{\lfloor \alpha \rfloor!} \int |u|^\alpha |K(u)|\, du \qquad \forall x \in \mathbb{R}.$$

Beweis. Es gilt

$$K_h * f(x) - f(x) = \int K(u)(f(x - hu) - f(x))\, du.$$

Die Taylorsche Formel besagt, dass für $l = \lfloor \alpha \rfloor$

$$f(x + y) - f(x) = \sum_{k=1}^{l-1} \frac{f^{(k)}(x)}{k!} y^k + \frac{f^{(l)}(x + \tau y)}{l!} y^l$$

mit $0 \le \tau \le 1$. Setzen wir $y = -uh$ ein und verwenden, dass K die Ordnung l besitzt, so gilt

$$\begin{aligned}
|K_h * f(x) - f(x)| &= \left| \int K(u) \frac{(-uh)^l}{l!} f^{(l)}(x - \tau hu)\, du \right| \\
&= \left| \int K(u) \frac{(-uh)^l}{l!} (f^{(l)}(x - \tau hu) - f^{(l)}(x))\, du \right| \\
&\le \int |K(u)| \frac{|uh|^l}{l!} L|\tau hu|^{\alpha - l}\, du \le h^\alpha \frac{L}{l!} \int |K(u)||u|^\alpha\, du,
\end{aligned}$$

wobei wir unterdrückt haben, dass τ von u abhängt. □

Kombinieren wir die Schranke für die Varianz aus dem Beweis von Satz 1.5 mit Proposition 1.12. so erhalten wir:

1.13 Satz. *Seien $\alpha, L > 0$ und K ein Kern der Ordnung $\lfloor \alpha \rfloor$ mit $\int |u|^\alpha K(u)\, du < \infty$ und $\int K^2(u)\, du < \infty$. Setze $h = cn^{-1/(2\alpha+1)}$. Dann gilt für alle $n \ge 1$ und $x \in \mathbb{R}$*

$$\sup_{f \in \mathcal{H}^\alpha(L), f \text{ W.-dichte}} \mathbb{E}_f (\hat{f}_{n,h}(x) - f(x))^2 \le C n^{-\frac{2\alpha}{2\alpha+1}}$$

mit einer Konstanten C die nur von c, α, L und K abhängt.

1.14 Bemerkung. In Satz 1.13 bedeutet $\mathbb{E}_f$, dass jede Beobachtung X_i die Dichte f besitzt, oder explizit:

$$\mathbb{E}_f (\hat{f}_n(x) - f(x))^2 = \int_{\mathbb{R}^n} (\hat{f}_n(x, x_1, \ldots, x_n) - f(x))^2 \prod_{i=1}^n f(x_i)\, dx_i.$$

1.15 Bemerkung (*Minimax-Ansatz*)**.** Wir suchen Schätzer für die das maximale Risiko über eine nichtparametrische Parameterklasse $\mathcal{F}$ möglichst klein ist. Satz 1.13 besagt, dass das maximale Risiko des Kerndichteschätzers über die Parametermenge $\mathcal{F} = \{f : \mathbb{R} \to [0, \infty) : \int f(x)\, dx = 1, f \in \mathcal{H}^\alpha(L)\}$ die Konvergenzrate $n^{-2\alpha/(2\alpha+1)}$ besitzt (bei geeigneter Bandbreitenwahl). Später werden wir zeigen, dass diese Rate nicht verbessert werden kann. Hierfür werden wir das maximale Risiko aus Satz 1.13 mit dem sogenannten Minimax-Risiko $\inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} \mathbb{E}_f (\hat{f}_n(x) - f(x))^2$ vergleichen.

1.4 Quadratisches Risiko

Für den Kerndichtschätzer gilt folgende Zerlegung:

1.16 Lemma. *Die Dichte erfülle $f \in L^2(\mathbb{R})$. Außerdem sei K ein Kern mit $K \in L^2(\mathbb{R})$. Dann gilt*

$$\mathbb{E} \|\hat{f}_{n,h} - f\|_{L^2}^2 = \|K_h * f - f\|_{L^2}^2 + \frac{1}{nh} \|K\|_{L^2}^2 - \frac{1}{n} \|K_h * f\|_{L^2}^2.$$

Beweis. Zunächst gilt, dass $K_h(x - X_1) \in L^2$ für Lebesgue-fast alle x, da

$$\int_{\mathbb{R}} \left(\int_{\mathbb{R}} K_h^2(x-y) f(y)\, dy \right) dx = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} K_h^2(x-y)\, dx \right) f(y)\, dy = \frac{1}{h} \|K\|_{L^2}^2 < \infty.$$

Es gilt nun

$$\begin{aligned}
\mathbb{E}\|\hat{f}_{n,h} - f\|_{L^2}^2 &= \mathbb{E} \int_{\mathbb{R}} (\hat{f}_{n,h}(x) - f(x))^2\, dx \\
&= \int_{\mathbb{R}} \mathbb{E}(\hat{f}_{n,h}(x) - f(x))^2\, dx \\
&= \int_{\mathbb{R}} \left(\operatorname{Var}(\hat{f}_{n,h}(x)) + (\mathbb{E}\hat{f}_{n,h}(x) - f(x))^2 \right) dx \\
&= \int_{\mathbb{R}} \frac{1}{n} \operatorname{Var}(K_h(x - X_1))\, dx + \|K_h * f - f\|_{L^2}^2.
\end{aligned}$$

Weiter gilt

$$\begin{aligned}
&\int_{\mathbb{R}} \frac{1}{n} \operatorname{Var}(K_h(x - X_1))\, dx \\
&= \frac{1}{n} \int_{\mathbb{R}} \int_{\mathbb{R}} K_h^2(x-y) f(y)\, dy dx - \frac{1}{n} \int_{\mathbb{R}} \left(\int_{\mathbb{R}} K_h(x-y) f(y)\, dy \right)^2 dx \\
&= \frac{1}{nh} \|K\|_{L^2}^2 - \frac{1}{n} \|K_h * f\|_{L^2}^2
\end{aligned}$$

und die Behauptung folgt. □

1.17 Aufgabe. *Die Dichte erfülle $f \in L^2(\mathbb{R})$. Sei K ein Kern mit $K \in L^2(\mathbb{R})$. Falls $h \to 0$ und $nh \to \infty$, so gilt*

$$\mathbb{E}\|\hat{f}_{n,h} - f\|_{L^2}^2 \to 0 \qquad \text{für} \qquad n \to \infty.$$

1.18 Definition. Sei $I \subseteq \mathbb{R}$ ein abgeschlossenes Intervall, $\alpha \in \mathbb{N}$ und $L > 0$. Dann heißt die Menge

$$\mathcal{S}^\alpha(I; L) = \left\{ f \in C^{\alpha-1}(I) \cap L^2(I) : f^{(\alpha-1)} \text{ absolut stetig mit } \int_I (f^{(\alpha)}(x))^2\, dx \le L^2 \right\}$$

Sobolev-Kugel auf I mit Parametern α, L Gilt $A = \mathbb{R}$, so schreiben wir auch $\mathcal{S}^\alpha(L) = \mathcal{S}^\alpha(\mathbb{R}; L)$.

1.19 Proposition. *Sei $f \in \mathcal{S}^\alpha(L)$ und K ein Kern der Ordnung $\alpha - 1$ mit $\int |u|^\alpha |K(u)|\, du < \infty$. Dann gilt*

$$\|K_h * f - f\|_{L^2} \le h^\alpha \frac{L}{\alpha!} \int |u|^\alpha |K(u)|\, du$$

Beweis. Die Taylorsche Formel besagt, dass

$$f(x+y) - f(x) = \sum_{k=1}^{\alpha-1} \frac{f^{(k)}(x)}{k!} y^k + \frac{y^\alpha}{(\alpha-1)!} \int_0^1 (1-t)^{\alpha-1} f^\alpha(x+ty)\, dt.$$

Da K die Ordnung $\alpha - 1$ besitzt folgt also

$$\begin{aligned} K_h * f(x) - f(x) &= \int K(u)(f(x-hu) - f(x))\, du \\ &= \frac{(-h)^\alpha}{(\alpha-1)!} \int \int_0^1 u^\alpha K(u)(1-t)^{\alpha-1} f^\alpha(x-thu)\, dtdu. \end{aligned}$$

Wir verwenden nun einen Spezialfall der verallgemeinerte Minkowski-Ungleichung, die besagt, dass

$$\left\| \int g(\cdot, y)\, dy \right\|_{L^2} \leq \int \|g(\cdot, y)\|_{L^2}\, dy \tag{1.2}$$

für alle Borel-messbaren Funktionen $g : \mathbb{R} \times \mathbb{R} \to [0, \infty]$. Wenden wir diese zweimal an, so folgt

$$\begin{aligned} \|K_h * f - f\|_{L^2} &\leq \frac{h^\alpha}{(\alpha-1)!} \left\| \int \int_0^1 |u^\alpha K(u)(1-t)^{\alpha-1} f^\alpha(\cdot - thu)|\, dtdu \right\|_{L^2} \\ &\leq \frac{h^\alpha}{(\alpha-1)!} \int \int_0^1 \left\| u^\alpha K(u)(1-t)^{\alpha-1} f^\alpha(\cdot - thu) \right\|_{L^2} dtdu \\ &= \frac{h^\alpha \|f^{(\alpha)}\|_{L^2}}{(\alpha-1)!} \int \int_0^1 |u|^\alpha |K(u)|(1-t)^{\alpha-1}\, dtdu \\ &\leq \frac{h^\alpha L}{\alpha!} \int |u|^\alpha |K(u)|\, du. \end{aligned}$$

Es bleibt (1.2) zu beweisen. Hierfür setzen wir $G(x) = \int g(x, y)\, dy$. Dann gilt

$$\left\| \int g(\cdot, y)\, dy \right\|_{L^2}^2 = \int G^2(x)\, dx = \sup_{\|\phi\|_{L^2} \leq 1, \phi \geq 0} \int G(x)\phi(x)\, dx.$$

Mit dem Satz von Tonelli und der Cauchy-Schwarz-Ungleichung gilt weiter

$$\int G(x)\phi(x)\, dx = \int \int g(x, y)\phi(x)\, dxdy \leq \int \|g(\cdot, y)\|_{L^2}\, dy.$$

□

Kombinieren wir Lemma 1.16 und Proposition 1.19, so erhalten wir

1.20 Satz. *Seien $\alpha \in \mathbb{N}$, $L > 0$ und K ein Kern der Ordnung $\alpha - 1$ mit $\int |u|^\alpha |K(u)|\, du < \infty$ und $\int K^2(u)\, du < \infty$. Setze $h = cn^{-1/2\alpha+1}$. Dann gilt*

$$\sup_{f \in \mathcal{S}^\alpha(L), f \text{ W.-dichte}} \mathbb{E}_f \|\hat{f}_{n,h} - f\|_{L^2}^2 \leq C L^{\frac{2}{2\alpha+1}} n^{-\frac{2\alpha}{2\alpha+1}}$$

mit einer Konstanten C die nur von c, α und K abhängt.

1.5 Projektionsschätzer

Wir nehmen im gesamten Unterkapitel an, dass f einen beschränkten Träger besitzt, welcher ohne Einschränkung gleich $[0,1]$ sei, und dass $f \in L^2([0,1])$. Es ist bekannt, dass $L^2([0,1])$ versehen mit dem Skalarprodukt $\langle g,h\rangle = \int_0^1 g(x)h(x)\,dx$ ein separabler Hilbertraum ist.

Sei nun V_d (stets) ein d-dimensionaler Vektorraum von beschränkten, rechtstetigen Funktionen mit nur endlich vielen Sprungstellen und $\phi_1,\dots,\phi_d$ eine Orthonormalbasis (ONB) von V_d bezüglich $\langle\cdot,\cdot\rangle$. Setze

$$\Pi_{V_d} : L^2[0,1] \to V_d,\ g \mapsto \sum_{j=1}^{d}\langle\phi_j, g\rangle\phi_j.$$

Dann ist Π_{V_d} die Orthogonalprojektion von $L^2([0,1])$ auf V_d und es gilt

$$\Pi_V f(x) = \sum_{j=1}^{d}\langle\phi_j, f\rangle\phi_j(x) = \int_0^1 K_{V_d}(x,y)f(y)\,dy, \qquad x\in[0,1]$$

mit

$$K_{V_d}(x,y) = \sum_{j=1}^{d}\phi_j(x)\phi_j(y)$$

unabhängig von der Wahl der ONB.

1.21 Definition. Der Projektionsschätzer ist definiert als

$$\hat f_{n,d}(x) = \int K_{V_d}(x,y)P_n(dy) = \sum_{j=1}^{d}\Big(\frac{1}{n}\sum_{i=1}^{n}\phi_j(X_i)\Big)\phi_j(x), \quad x\in[0,1].$$

1.22 Lemma. *Es gilt*

$$\mathbb{E}\|\hat f_{n,d} - f\|_{L^2}^2 = \|f-\Pi_{V_d}f\|_{L^2}^2 + \frac{1}{n}\sum_{j=1}^{d}\int_0^1\phi_j^2(x)f(x)\,dx - \frac{1}{n}\|\Pi_{V_d}f\|_{L^2}^2.$$

Beweis. Π_{V_d} erfüllt (i) $\Pi_{V_d}^2 = \Pi_{V_d}$ und (ii) $\langle\Pi_{V_d}g,h\rangle = \langle g,\Pi_{V_d}h\rangle$ für alle $g,h\in L^2[0,1]$. Dies sind gerade die definierenden Eigenschaften einer Orthogonalprojektion. Daher gilt $f-\Pi_{V_d}f \perp V_d$ und es folgt

$$\|\hat f_{n,d} - f\|_{L^2}^2 = \|\hat f_{n,d} - \Pi_{V_d}f\|_{L^2}^2 + \|f-\Pi_{V_d}f\|_{L^2}^2.$$

Weiter gilt

$$\|\hat f_{n,d} - \Pi_{V_d}f\|_{L^2}^2 = \sum_{j=1}^{d}\left(\frac{1}{n}\sum_{i=1}^{n}\phi_j(X_i) - \langle\phi_j,f\rangle\right)^2$$

und somit wegen $\langle \phi_j, f\rangle = \mathbb{E}\,\phi_j(X_i)$, dass

$$\mathbb{E}\|\hat{f}_{n,d} - \Pi_{V_d} f\|_{L^2}^2 = \frac{1}{n}\sum_{j=1}^{d} \mathrm{Var}(\phi_j(X_1)).$$

Insgesamt erhalten wir also

$$\begin{aligned}\mathbb{E}\|\hat{f}_{n,d} - f\|_{L^2}^2 &= \|f - \Pi_{V_d} f\|_{L^2}^2 + \frac{1}{n}\sum_{j=1}^{d} \mathrm{Var}(\phi_j(X_1)) \\ &= \|f - \Pi_{V_d} f\|_{L^2}^2 + \frac{1}{n}\sum_{j=1}^{d} \int_0^1 \phi_j^2(x) f(x)\,dx - \frac{1}{n}\sum_{j=1}^{d} \langle \phi_j, f\rangle^2.\end{aligned}$$

Die Behauptung folgt indem wir $\|\Pi_{V_d} f\|_2^2 = \sum_{j=1}^{d} \langle \phi_j, f\rangle^2$ einsetzen. □

Im Beweis haben wir gesehen, dass $\mathbb{E}\hat{f}_{n,d}(x) = \Pi_{V_d} f(x)$. Daher gilt für den MSE

$$\mathbb{E}(\hat{f}_{n,d}(x) - f(x))^2 = (f(x) - \Pi_{V_d} f(x))^2 + \frac{1}{n}\,\mathrm{Var}\Big(\sum_{j=1}^{d} \phi_j(x)\phi_j(X_1)\Big). \quad (1.3)$$

1.23 Beispiel (Trigonometrische Basis)**.** Die Funktionen

$$\begin{aligned}\phi_1(x) &= 1 \\ \phi_{2k}(x) &= \sqrt{2}\cos(2\pi k x) \\ \phi_{2k+1}(x) &= \sqrt{2}\sin(2\pi k x), \qquad k = 1, 2, \ldots,\end{aligned}$$

wobei $x \in [0,1]$, bilden eine ONB von $L^2([0,1])$. Sei im Folgenden $V_d = \mathrm{span}(\phi_1, \ldots, \phi_d)$ der lineare Raum aufgespannt von den ersten d Basisfunktionen. Dann gilt:

1.24 Lemma. *Sei $\alpha \in \mathbb{N}$ und $L > 0$. Dann gilt für alle $g \in \mathcal{S}^\alpha([0,1]; L)$, die zusätzlich die Bedingung $g^{(j)}(0) = g^{(j)}(1)$ für $j = 0, 1, \ldots, \alpha - 1$ erfüllen, dass*

$$\|g - \Pi_{V_d} g\|_2^2 \le \frac{L^2}{\pi^{2\alpha}} d^{-2\alpha}.$$

Beweis. Setze $\theta_k = \langle \phi_j, g\rangle$. Dann gilt $\forall j \ge 1$

$$\begin{aligned}&\theta_{2j}^2 + \theta_{2j+1}^2 \\ &= 2\left(\int_0^1 \cos(2\pi j x) g(x)\,dx\right)^2 + 2\left(\int_0^1 \sin(2\pi j x) g(x)\,dx\right)^2 \\ &= \frac{2}{(2\pi j)^2}\left(\int_0^1 \cos(2\pi j x) g'(x)\,dx\right)^2 + \frac{2}{(2\pi j)^2}\left(\int_0^1 \sin(2\pi j x) g'(x)\,dx\right)^2 \\ &= \frac{2}{(2\pi j)^{2\alpha}}\left(\int_0^1 \cos(2\pi j x) g^{(\alpha)}(x)\,dx\right)^2 + \frac{2}{(2\pi j)^{2\alpha}}\left(\int_0^1 \sin(2\pi j x) g^{(\alpha)}(x)\,dx\right)^2 \\ &= \frac{1}{(2\pi j)^{2\alpha}}\langle \phi_{2j}, g^{(\alpha)}\rangle^2 + \frac{1}{(2\pi j)^{2\alpha}}\langle \phi_{2j+1}, g^{(\alpha)}\rangle^2,\end{aligned}$$

wobei wir iterativ partielle Integration und die Tatsache, dass $g^{(j)}(0) = g^{(j)}(1)$ für $j = 0, 1, \ldots, \alpha - 1$, verwendet haben. Mit Hilfe der Parseval-Gleichung schließen wir

$$\|g - \Pi_{V_d} g\|_{L^2}^2 = \sum_{j>d} \theta_j^2 \leq \frac{\|g^{(\alpha)}\|_{L^2}^2}{(\pi d)^{2\alpha}} \leq \frac{L^2}{(\pi d)^{2\alpha}}.$$

□

1.25 Bemerkung. Eine Inspektion des Beweises zeigt, dass

$$\begin{aligned}\mathcal{W}^\alpha(L) :&= \left\{ g \in \mathcal{S}^\alpha([0,1]; L) : g^{(j)}(0) = g^{(j)}(1) \text{ für } j = 0, \ldots, \alpha - 1 \right\} \\ &\subseteq \left\{ g = \sum_{j \geq 1} \theta_j \phi_j \in L^2([0,1]) : \sum_{j \geq 1} a_j^2 \theta_j^2 \leq L^2/\pi^{2\alpha} \right\},\end{aligned}$$

mit $a_j = j^\alpha$ für j gerade und $a_j = (j-1)^\alpha$ für j ungerade. Beachte hierfür, dass $\langle \phi_1, g^{(\alpha)} \rangle = 0$ gilt. Man kann zeigen, dass sogar Gleichheit gilt (siehe z.B. Proposition 1.14 in Tsybakov (2009).

Ist nun $f \in \mathcal{W}^\alpha(L)$, wobei $\alpha \in \mathbb{N}$ und $L \geq 1$, so folgt aus Lemma 1.22 und Lemma 1.24, dass

$$\begin{aligned}\mathbb{E}\|\hat{f}_{n,d} - f\|_{L^2}^2 &\leq \frac{L^2 d^{-2\alpha}}{\pi^{2\alpha}} + \frac{1}{n} \sum_{k=1}^{d} \int_0^1 \phi_k^2(x) f(x)\, dx \\ &\leq \frac{L^2 d^{-2\alpha}}{\pi^{2\alpha}} + \frac{2d}{n} \qquad (1.4)\end{aligned}$$

wobei wir außerdem verwendet haben, dass $\|\phi_k^2\|_\infty \leq 2$. Wählen wir nun $d = \max\{m : m \leq (L^2 n)^{1/(2\alpha+1)}\}$ und benutzen die Ungleichung $d \geq (L^2 n)^{1/(2\alpha+1)}/2$, so folgt

$$\frac{L^2 d^{-2\alpha}}{\pi^{2\alpha}} + \frac{2d}{n} \leq C L^{\frac{2}{2\alpha+1}} n^{-\frac{2\alpha}{2\alpha+1}}$$

mit $C = (2/\pi)^{2\alpha} + 2$. Wir erhalten also wie schon für den Kerndichteschätzer die Konvergenzrate $n^{-2\alpha/(2\alpha+1)}$ für dass quadratische Risiko.

1.26 Beispiel. Sei $m \in \mathbb{N}$ und

$$V_m = \left\{ h : [0,1] \to \mathbb{R} : h = \sum_{j=1}^{m} c_j \mathbf{1}_{\left[\frac{j-1}{m}, \frac{j}{m}\right)}, c_j \in \mathbb{R} \right\}$$

der Raum der stückweise konstanten Funktionen. Dann ist $\sqrt{m}\mathbf{1}_{\left[\frac{j-1}{m}, \frac{j}{m}\right)}$, $j = 1, \ldots, m$, eine ONB von V_m und der zugehörige Projektionsschätzer ist gerade der Histogramm-Schätzer:

$$\hat{f}_{n,m}(x) = \sum_{j=1}^{m} m \cdot \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\left[\frac{j-1}{m}, \frac{j}{m}\right)}(X_i) \right) \mathbf{1}_{\left[\frac{j-1}{m}, \frac{j}{m}\right)}(x).$$

Erfüllt die Dichte $f \in \mathcal{H}^\alpha(L)$ mit $\alpha \le 1$, so gilt für $x \in [(j-1)/m, j/m)$

$$\begin{aligned}\mathbb{E}(\hat{f}_{n,m}(x) - f(x))^2 &\le \left(f(x) - m \int_{\frac{j-1}{m}}^{\frac{j}{m}} f(y)\,dy \right)^2 + \frac{m^2}{n} \int_{\frac{j-1}{m}}^{\frac{j}{m}} f(y)\,dy \\ &\le L^2 m^{-2\alpha} + \frac{\|f\|_\infty m}{n}.\end{aligned}$$

Wählen wir m von der Größenordnung $n^{1/(2\alpha+1)}$, so erhalten wir die Konvergenzrate $n^{-2\alpha/(2\alpha+1)}$ für den MSE. Man kann den Histogramm-Schätzer mit dem zum Rechteckkern gehörenden Kerndichteschätzer vergleichen.

1.27 Beispiel. Für glattere Funktionen mussten wir im Fall des Kerndichteschätzers einen Kern höherer Ordnung wählen. Ein Analogon im Fall des Projektionsschätzers ist wie folgt. Seien $m, r \in \mathbb{N}$ und

$$V_{m,r} = \{h : [0,1] \to \mathbb{R} : \forall j \le m \text{ ist } f|_{[\frac{j-1}{m}, \frac{j}{m})} \text{ Polynom vom Grad} \le r\}$$

der Raum der stückweisen Polynome der Ordnung r mit Knotenpunkten $0, 1/m, 2/m, \dots, 1$ (mit $[1-1/m, 1]$ für $j = m-1$). Es gilt $\dim(V_{m,r}) = (r+1)m$. Eine ONB kann mit Hilfe orthogonaler Polynome konstruiert werden.

1.28 Aufgabe. *Sei $(P_k)_{k\ge 1}$ die Folge der Legendre-Polynome definiert durch*

$$P_k(x) = \frac{1}{2^k k!} \frac{d^k}{dx^k}(x^2-1)^k, \qquad x \in [-1,1].$$

Dann gilt für alle $k, l \ge 1$

$$\int_{-1}^{1} P_k(x) P_l(x)\,dx = \frac{2}{2k+1}\delta_{kl} \qquad \textit{und} \qquad \sup_{x\in[-1,1]} |P_k(x)| \le 1.$$

Wir setzen

$$\phi_{j,k}(x) = \sqrt{2k+1}\sqrt{m} P_k(2mx - 2j + 1) \mathbf{1}_{[\frac{j-1}{m}, \frac{j}{m})}(x),$$

für $j = 1, \dots, m$ und $k = 0, \dots, r$. Dann folgt aus Aufgabe 1.28, dass diese Funktionen eine ONB von $V_{m,r}$ bilden.

1.29 Lemma. *Seien $\alpha, L > 0$. Setze $r = \lfloor \alpha \rfloor$. Dann gilt für alle $g \in \mathcal{H}^\alpha(L)$*

$$\|g - \Pi_{V_{m,r}} g\|_\infty = \sup_{x\in[0,1]} |g(x) - \Pi_{V_{m,r}} g(x)| \le C m^{-\alpha}$$

mit einer Konstanten C die nur von α und L abhängt.

Beweis. 1. Beh. Es existiert ein $h \in V_{m,r}$ mit $\|g-h\|_\infty \le Lm^{-\alpha}/r!$.
Bew. Sei hierfür O.B.d.A. $x \in [0,1/m)$ und $p(x) = \sum_{k=0}^r g^{(k)}(0)x^k/k!$ das Taylorpolynom von g vom Grad r. Dann gilt

$$\begin{aligned} |g(x)-p(x)| &= \frac{|x|^r}{r!}|g^{(r)}(\tau x) - g^{(r)}(0)| \qquad (0 \le \tau \le 1) \\ &\le \frac{|x|^r}{r!} L|\tau x|^{\alpha-r} \le L\frac{m^{-\alpha}}{r!}. \end{aligned}$$

2. Beh. Es gilt $\|\Pi_{V_{m,r}} g\|_\infty \le C\|g\|_\infty$ mit $C = \sum_{k=0}^r \sqrt{2k+1}$.
Bew. Betrachte wieder den Fall $x \in [0,1/m)$. Dann gilt

$$\begin{aligned} |\Pi_{V_{m,r}} g(x)| &= \left| \sum_{k=0}^r \left(\int_0^{\frac{1}{m}} \phi_{1,k}(y) g(y)\, dy \right) \phi_{1,k}(x) \right| \\ &\le \|g\|_\infty \sum_{k=0}^r \sqrt{2k+1}, \end{aligned}$$

wobei wir die Cauchy-Schwarz-Ungleichung und die Eigenschaft $\|\phi_{1,k}\|_\infty \le \sqrt{2k+1}\sqrt{m}$ verwendet haben.
3. Beh. Es gilt $\|g - \Pi_{V_{m,r}} g\|_\infty \le (1+C)Lm^{-\alpha}/r!$ wobei C die Konstante aus der 2. Behauptung ist.
Bew. Sei $h \in V_{m,r}$ wie in 1. Behauptung konstruiert. Dann gilt

$$\begin{aligned} \|g - \Pi_{V_{m,r}} g\|_\infty &\le \|g-h\|_\infty + \|h - \Pi_{V_{m,r}} g\|_\infty \\ &= \|g-h\|_\infty + \|\Pi_{V_{m,r}}(h-g)\|_\infty \\ &\le (1+C)\|g-h\|_\infty \\ &\le \frac{(1+C)L}{r!} m^{-\alpha}, \end{aligned}$$

wobei wir die erste und die zweite Behauptung in den letzten beiden Ungleichungen verwendet haben. □

Ist nun $f \in \mathcal{H}^\alpha(L)$ und $x \in [(j-1)/m, j/m)$ so folgt aus (1.3) und Lemma 1.29

$$\begin{aligned} \mathbb{E}(\hat{f}_{n,d}(x) - f(x))^2 &\le Cm^{-2\alpha} + \frac{1}{n} \operatorname{Var}\Big(\sum_{k=0}^r \phi_{j,k}(x)\phi_{j,k}(X_1) \Big) \\ &\le Cm^{-2\alpha} + \frac{r}{n} \sum_{k=0}^r \phi_{j,k}^2(x) \mathbb{E}\phi_{j,k}^2(X_1) \\ &\le C\left(m^{-2\alpha} + \frac{m}{n} \right), \end{aligned}$$

wobei wir außerdem die Cauchy-Schwarz-Ungleichung und die Ungleichungen $\phi_{j,k}^2(x) \le (2k+1)m$ und $\mathbb{E}\phi_{j,k}(X_1)^2 = \int_0^1 \phi_{j,k}(y) f(y)\, dy \le \|f\|_\infty$ verwendet

haben und $C = C(\alpha, L)$ eine Konstante ist, die von Zeile zu Zeile verschieden sein kann. Wählen wir nun m von der Größenordnung $n^{1/(2\alpha+1)}$, so erhalten wir wie schon für den Kerndichteschätzer die Konvergenzrate $n^{-2\alpha/(2\alpha+1)}$ für den MSE.

1.6 Untere Schranken

Wir haben bisher Schätzer konstruiert (Kerndichteschätzer, Projektionsschätzer), die folgende Schranke erfüllen:

$$\sup_{f\in\mathcal{F}} \mathbb{E}_f d^2(\hat{f}_n, f) \le C n^{-\frac{2\alpha}{2\alpha+1}}$$

Wir haben bisher zwei Fälle betrachtet: im ersten war $d(g,h) = |g(x_0)-h(x_0)|$, $x_0 \in \mathbb{R}$, der punktweise Abstand in x_0 und $\mathcal{F} = \{f \in \mathcal{H}^\alpha(L) : f \text{ W.-dichte}\}$, im zweiten war $d(g,h) = \|g-h\|_{L^2}$ der L^2-Abstand und $\mathcal{F} = \{f \in \mathcal{S}^\alpha(L) : f \text{ W.-dichte}\}$. In diesem Kapitel wollen wir entsprechende untere Schranken beweisen. Wir betrachten für eine Menge $\mathcal{F}$ von Wahrscheinlichkeitsdichten versehen mit einer (Pseudo-)Metrik d und einen Schätzer $\hat{f}_n$ das folgende maximale quadratische Risiko

$$\sup_{f\in\mathcal{F}} \mathbb{E}_f d^2(\hat{f}_n, f) = \sup_{f\in\mathcal{F}} \int_{\mathbb{R}^n} d^2(\hat{f}(x_1,\dots,x_n), f) \prod_{i=1}^n f(x_i)\, dx_i.$$

Wir formalisieren:

1.30 Definition. Sei $\mathcal{F}$ eine nichtleere Menge. Ein Messraum $(\mathcal{X}, \mathscr{A})$ versehen mit einer Familie $(P_f)_{f\in\mathcal{F}}$ von Wahrscheinlichkeitsmaßen heißt statistisches Experiment oder statistisches Modell. Sei $\mathcal{F}$ eine Teilmenge eines (pseudo-)metrischen Raumes (S,d). Ein Schätzer $\hat{f}_n$ ist eine Borel-messbare Abbildung $\hat{f} : \mathcal{X} \to S$. Das maximale quadratische Risiko ist definiert als

$$\sup_{f\in\mathcal{F}} \int_{\mathcal{X}} d^2(\hat{f}(x), f)\, P_f(dx).$$

Das Minimax-Risiko ist definiert als

$$\inf_{\hat{f}} \sup_{f\in\mathcal{F}} \int_{\mathcal{X}} d^2(\hat{f}(x), f)\, P_f(dx),$$

wobei das Minimum über alle Schätzer $\hat{f}$ genommen wird.

1.31 Bemerkung. (a) Wie üblich schreiben wir

$$\sup_{f\in\mathcal{F}} \int_{\mathcal{X}} d^2(\hat{f}(x), f)\, P_f(dx) = \sup_{f\in\mathcal{F}} \mathbb{E}_f d^2(\hat{f}(X), f),$$

wobei X eine Beobachtung in dem statistischen Experiment ist (X hat unter $\mathbb{E}_f$ die Verteilung P_f).

(b) In den Beweisen werden wir lediglich verwenden, dass die Abbildung $x \mapsto d(\hat{f}(x), g)$ Borel-messbar ist $\forall g \in S$. Ist S separabel so ist diese Bedingung äquivalent zu $\hat{f} : \mathcal{X} \to S$ Borel-messbar.

1.32 Beispiel (Dichteschätzung)**.** Betrachte das statistische Experiment $\mathcal{E}_n = (\mathcal{X}_n, \mathscr{A}_n, (P_{f,n})_{f\in\mathcal{F}})$ mit

$$\mathcal{X}_n = \mathbb{R}^n, \qquad \mathscr{A}_n = \mathscr{B}_{\mathbb{R}^n}, \quad \text{und} \quad P_{f,n}(dx) = \prod_{i=1}^{n} f(x_i)\, dx_i$$

Im Fall des punktweise quadratischen Risikos betrachten wir $\mathcal{F} = \{f \in \mathcal{H}^\alpha(L) : f \text{ W.-dichte}\}$, $d(g,h) = |g(x_0) - h(x_0)|$ und $S = \{g : \mathbb{R} \to \mathbb{R} \text{ Borel-messbar}\}$. In diesem Fall ist eine Abbildung $\hat{f}_n : \mathbb{R}^n \to S$ Borel-messbar genau dann, wenn die Auswertung in x_0 Borel-messbar ist, d.h. wenn $x \mapsto \hat{f}_n(x)(x_0)$ Borel-messbar.

Im Fall des quadratischen Risikos betrachten wir $\mathcal{F} = \{f \in \mathcal{S}^\alpha(L) : f \text{ W.-dichte}\}$, $d(g,h) = \|g - h\|_{L^2}$ und $S = L^2(\mathbb{R})$. Da $L^2(\mathbb{R})$ separabel ist, kann man mit Hilfe des Kriteriums aus Bemerkung 1.31 (b) leicht überprüfen, ob eine Abbildung $\hat{f}_n : \mathbb{R}^n \to S$ Borel-messbar ist.

1.33 Satz (Methode von Le Cam)**.** *Sei $\mathcal{F}$ eine Teilmenge eines pseudometrischen Raumes (S,d) und $(\mathcal{X}, \mathscr{A}, (P_f)_{f\in\mathcal{F}})$ ein statistisches Experiment. Seien $f_0, f_1 \in \mathcal{F}$ zwei fixierte Elemente. Dann gilt für jeden Schätzer $\hat{f} : \mathcal{X} \to S$*

$$\sup_{f\in\mathcal{F}} \mathbb{E}_f d^2(\hat{f}(X), f) \geq \max_{j=0,1} \mathbb{E}_{f_j} d^2(\hat{f}(X), f_j) \geq \frac{d^2(f_0,f_1)}{8} \int P_{f_0} \wedge P_{f_1}.$$

1.34 Zusatz. *Es gilt außerdem*

$$\sup_{f\in\mathcal{F}} \mathbb{E}_f d^2(\hat{f}(X), f) \geq \frac{d^2(f_0,f_1)}{16} \left(\int \sqrt{P_{f_0} P_{f_1}} \right)^2.$$

1.35 Bemerkung. (a) Besitzen P_{f_0} und P_{f_1} Dichten p_0 und p_1 bezüglich eines dominierenden Maßes μ (z.B. $\mu = (P_{f_0} + P_{f_1})/2$), so definiert man

$$\int P_{f_0} \wedge P_{f_1} := \int p_0 \wedge p_1 d\mu \quad \text{und} \quad \int \sqrt{P_{f_0} P_{f_1}} := \int \sqrt{p_0 p_1} d\mu.$$

Beide Definitionen hängen nicht von der Wahl des dominierenden Maßes μ ab (Übung).

(b) Ist $P_{f_0} = P_{f_1}$, so gilt $\int \sqrt{P_{f_0} P_{f_1}} = 1$. Ist andererseits $P_{f_0} \perp P_{f_1}$, d.h. $\exists A \in \mathscr{A}$ mit $P_{f_0}(A) = 0$ und $P_{f_1}(A) = 1$, so gilt $\int \sqrt{P_{f_0} P_{f_1}} = 0$. In diesem Fall erfüllt der Schätzer $\hat{f}(x) = f_1$ falls $x \in A$ und $\hat{f} = f_0$ sonst $\max_{j=0,1} \mathbb{E}_{f_j} d^2(\hat{f}(X), f_j) = 0$.

Beweis. Die erste Ungleichung ist klar. Wir definieren nun einen Schätzer $\tilde{f}$ mit Werten in $\{f_0, f_1\}$ durch

$$d(\tilde{f}, \hat{f}) = \min_{j=0,1} d(f_j, \hat{f})$$

d.h.

$$\tilde{f}(x) = \begin{cases} f_0, & \text{falls } d(\hat{f}(x), f_0) \leq d(\hat{f}(x), f_1), \\ f_1, & \text{sonst.} \end{cases}$$

Dann ist $\tilde{f}$ ein Schätzer mit $d(\tilde{f}(x), f_j) \leq d(\tilde{f}(x), \hat{f}(x)) + d(\hat{f}(x), f_j) \leq 2d(\hat{f}(x), f_j)$. Es folgt

$$\begin{aligned} \max_{j=0,1} \mathbb{E}_{f_j} d^2(\hat{f}(X), f_j) &\geq \frac{1}{4} \max_{j=0,1} \mathbb{E}_{f_j} d^2(\tilde{f}(X), f_j) \\ &= \frac{1}{4} \max_{j=0,1} \mathbb{E}_{f_j} d^2(f_0, f_1) \mathbf{1}(\tilde{f}(X) \neq f_j) \\ &= \frac{d^2(f_0, f_1)}{4} \max_{j=0,1} \mathbb{E}_{f_j} \mathbf{1}(\tilde{f}(X) \neq f_j) \end{aligned}$$

Außerdem gilt mit ψ definiert durch $\psi(x) = \mathbf{1}(\tilde{f}(x) \neq f_0) = \mathbf{1}(\tilde{f}(x) = f_1)$, dass

$$\begin{aligned} \max_{j=0,1} \mathbb{E}_{f_j} \mathbf{1}(\tilde{f}(X) \neq f_j) &\geq \frac{1}{2} \left(\mathbb{E}_{f_0} \mathbf{1}(\tilde{f}(X) \neq f_0) + \mathbb{E}_{f_1} \mathbf{1}(\tilde{f}(X) \neq f_1) \right) \\ &= \frac{1}{2} \left(\mathbb{E}_{f_0} \psi(X) + \mathbb{E}_{f_1} (1 - \psi(X)) \right) \\ &= \frac{1}{2} \int (\psi p_0 + (1 - \psi) p_1) \, d\mu \\ &\geq \frac{1}{2} \int p_0 \wedge p_1 d\, \mu, \end{aligned}$$

wobei p_0 und p_1 Dichten von P_{f_0} und P_{f_1} bezüglich eines dominierenden Maßes μ sind und wir in der letzten Ungleichung verwendet haben, dass $p_0, p_1 \geq 0$ und $0 \leq \psi \leq 1$. Daher folgt die zweite Ungleichung. Zusatz 1.34 kann wie folgt gesehen werden. Es gilt

$$p_0 p_1 = (p_0 \vee p_1)(p_0 \wedge p_1) \leq (p_0 + p_1)(p_0 \wedge p_1).$$

Daher folgt mit Hilfe der Cauchy-Schwarz-Ungleichung

$$\begin{aligned} \left(\int \sqrt{p_0 p_1} d\mu \right)^2 &\leq \left(\int (p_0 + p_1)^{1/2} (p_0 \wedge p_1)^{1/2} d\mu \right)^2 \\ &\leq \left(\int (p_0 + p_1) d\mu \right) \left(\int p_0 \wedge p_1 d\mu \right) = 2 \int p_0 \wedge p_1 d\mu. \quad (1.5) \end{aligned}$$

Setzen wir dies in Satz 1.33 ein, so folgt die Behauptung. □

1.36 Satz. *Seien $\alpha, L > 0$ und $x_0 \in \mathbb{R}$. Dann gilt für alle $n \geq 1$ und alle Schätzer $\hat{f}_n(x_0) : \mathbb{R}^n \to \mathbb{R}$*

$$\sup_{f \in \mathcal{H}^\alpha(L), f \text{ W.-dichte}} n^{\frac{2\alpha}{2\alpha+1}} \mathbb{E}_f (\hat{f}_n(x_0, X) - f(x_0))^2 \geq c$$

mit einer Konstanten $c > 0$ die nur von α und L abhängt und $X = (X_1, \ldots, X_n)$ hat unter $\mathbb{E}_f$ unabhängige Komponenten jeweils mit Dichte f.

Beweis. Wir wenden Le Cams Methode mit dem statistischen Experiment aus Beispiel 1.32 an, d.h. mit $\mathcal{E}_n = (\mathbb{R}^n, \mathscr{B}_{\mathbb{R}^n}, (P_{f,n})_{f \in \mathcal{F}})$, wobei $P_{f,n}(dx) = \prod_{i=1}^n f(x_i)\, dx_i$ und $\mathcal{F} = \{f \in \mathcal{H}^\alpha(L) : f \text{ W.-dichte}\}$, $d(g,h) = |g(x_0) - h(x_0)|$, $S = \{g : \mathbb{R} \to \mathbb{R} \text{ Borel-messbar}\}$ und $\hat{f}_n : \mathbb{R}^n \to S$ beliebige Fortsetzung von $\hat{f}_n(x_0)$. Wir erhalten also für $f_0, f_1 \in \mathcal{F}$:

$$\begin{aligned}
&\sup_{f \in \mathcal{F}} \mathbb{E}_f (\hat{f}_n(x_0, X) - f(x))^2 \\
&\quad \geq \frac{(f_0(x_0) - f_1(x_0))^2}{16} \left(\int \sqrt{P_{f_0,n} P_{f_1,n}} \right)^2 \\
&\quad = \frac{(f_0(x_0) - f_1(x_0))^2}{16} \left(\int \sqrt{\prod_{i=1}^n f_0(x_i) f_1(x_i)}\, dx_1 \ldots dx_n \right)^2 \\
&\quad = \frac{(f_0(x_0) - f_1(x_0))^2}{16} \left(\int \sqrt{f_0 f_1}\, d\lambda \right)^{2n}. \qquad (1.6)
\end{aligned}$$

mit Lebesguemaß λ. Wir wollen nun f_0 und f_1 konstruieren. Wir behaupten, dass es Funktionen $f_0, f_1, K : \mathbb{R}^n \to \mathbb{R}$ gibt mit (Bild!)

(a) f_0 W.-dichte, $f_0 \in \mathcal{H}^\alpha(L/2)$, $f_0(y) \geq c_0 > 0$ für alle y mit $|y - x_0| \leq 1/2$,

(b) $K \in \mathcal{H}^\alpha(L/2)$ mit $\int K\, d\lambda = 0$, Träger in $[-1/2, 1/2]$, $K(0) > 0$, $\|K\|_\infty \leq c_0 \leq 1$,

(c) f_1 definiert durch $f_1(y) = f_0(y) + h^\alpha K(\frac{y - x_0}{h})$ ist W.-dichte mit $f_1 \in \mathcal{H}^\alpha(L)$ für alle $h \leq 1$.

Zunächst gilt für $f_0, K \in \mathcal{H}^\alpha(L/2)$, dass $f_1 \in \mathcal{H}^\alpha(L)$. In der Tat gilt

$$\begin{aligned}
&\sup_{u \neq v} \frac{\left| f_0^{(l)}(u) + h^{\alpha-l} K^{(l)}(\frac{u-x_0}{h}) - f_0^{(l)}(v) - h^{\alpha-l} K^{(l)}(\frac{v-x_0}{h}) \right|}{|u - v|^{\alpha-l}} \\
&\leq \sup_{u \neq v} \frac{|f_0^{(l)}(u) - f_0^{(l)}(v)|}{|u - v|^{\alpha-l}} + \sup_{u \neq v} \frac{\left| h^{\alpha-l} K^{(l)}(\frac{u-x_0}{h}) - h^{\alpha-l} K^{(l)}(\frac{v-x_0}{h}) \right|}{|u - v|^{\alpha-l}} \\
&= \sup_{u \neq v} \frac{|f_0^{(l)}(u) - f_0^{(l)}(v)|}{|u - v|^{\alpha-l}} + \sup_{u' \neq v'} \frac{|K^{(l)}(u') - K^{(l)}(v')|}{|u' - v'|^{\alpha-l}},
\end{aligned}$$

wobei wir in der letzten Gleichheit $u' = (u - x_0)/h$ und $v' = (v - x_0)/h$ gesetzt haben. Hieraus folgt, dass $f_1 \in \mathcal{H}^\alpha(L)$. Es ist nun klar, dass es viele Wahlen für f_0 und K gibt, so dass (a), (b) und (c) erfüllt sind. Es bleibt nun (1.6) zu analysieren.

1. Beh. Setze $h = n^{-\frac{1}{2\alpha+1}}$. Dann gilt

$$\left(\int \sqrt{f_0 f_1}\, d\lambda\right)^{2n} \geq 1/4.$$

Bew. Es gilt

$$\begin{aligned}\left(\int \sqrt{f_0 f_1}\, d\lambda\right)^{2n} &= \left(1 - \frac{1}{2}\int (\sqrt{f_0} - \sqrt{f_1})^2\, d\lambda\right)^{2n} \\ &= \left(1 - \frac{1}{2}\int \left(\frac{f_0 - f_1}{\sqrt{f_0} + \sqrt{f_1}}\right)^2 d\lambda\right)^{2n}.\end{aligned}$$

Verwenden wir nun die Eigenschaften in (a)-(c) und setzen $h = n^{-\frac{1}{2\alpha+1}}$ ein, so folgt

$$\begin{aligned}\left(\int \sqrt{f_0 f_1}\, d\lambda\right)^{2n} &\geq \left(1 - \frac{1}{2c_0}\int h^{2\alpha} K^2\left(\frac{y - x_0}{h}\right) dy\right)^{2n} \\ &= \left(1 - \frac{h^{2\alpha+1}\|K\|_{L^2}^2}{2c_0}\right)^{2n} \\ &\geq \left(1 - \frac{1}{2n}\right)^{2n} \geq \frac{1}{4}.\end{aligned}$$

2. Beh. Für $h = n^{-\frac{1}{2\alpha+1}}$ gilt

$$(f_0(x_0) - f_1(x_0))^2 = h^{2\alpha} K^2(0) = n^{-\frac{2\alpha}{2\alpha+1}} K^2(0).$$

Setzen wir nun die 1. und 2. Beh. in (1.6) ein, so folgt

$$n^{\frac{2\alpha}{2\alpha+1}} \sup_{f \in \mathcal{F}} \mathbb{E}_f(\hat{f}_n(x_0, X) - f(x))^2 \geq \frac{K^2(0)}{2^6}$$

und somit die Behauptung, da $K(0)$ eine Konstante ist, die nur von α und L abhängt. □

1.37 Beispiel. Man kann Satz 1.33 (bzw. Zusatz 1.34) auch auf ein parametrisches Modell anwenden. Betrachte zum Beispiel das statistische Experiment auf $(\mathbb{R}^n, \mathscr{B}_{\mathbb{R}^n})$ gegeben durch die Familie von Wahrscheinlichkeitsmaßen $(\mathcal{N}(\theta, \sigma^2)^{\otimes n})_{\theta \in \mathbb{R}}$ (wir beobachten n unabhängige normalverteilte Zufallsvariablen mit Erwartungswert θ und Varianz σ^2). Sei $\mathbb{R}$ versehen mit der

Betragsmetrik $d(\theta,\theta') = |\theta - \theta'|$. Dann gilt für alle $\theta_0, \theta_1 \in \mathbb{R}$ und alle messbaren Abbildungen $\hat{\theta}_n : \mathbb{R}^n \to \mathbb{R}$

$$\begin{aligned}
&\sup_{\theta\in\mathbb{R}} \mathbb{E}_\theta(\hat{\theta}_n(X) - \theta)^2 \\
&\geq \frac{(\theta_0-\theta_1)^2}{16}\left(\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\theta_0)^2}{4\sigma^2} - \frac{(y-\theta_1)^2}{4\sigma^2}\right)\right)^{2n} \\
&= \frac{(\theta_0-\theta_1)^2}{16} \exp\left(-\frac{n(\theta_0-\theta_1)^2}{4\sigma^2}\right).
\end{aligned}$$

Wählen wir $\theta_0 = 0$ und $\theta_1 = 2\sigma/\sqrt{n}$, so ist die rechte Seite größer gleich $\sigma^2/(4en)$. Andererseits hat der Mittelwert $(1/n)\sum_{i=1}^n X_i$ das Risiko σ^2/n und es kann gezeigt werden (siehe VL Mathematische Statistik), dass dies der Wert des Minimax-Risikos ist. Zusatz 1.34 liefert also ein bis auf die Konstante $1/(4e)$ optimales Resultat.

1.38 Definition. Sei $\mathcal{F}$ eine Teilmenge eines pseudometrischen Raumes (S,d), $(\mathcal{X}_n, \mathscr{A}_n, (P_{f,n})_{f\in\mathcal{F}})$, $n \geq 1$ eine Folge von statistischen Experimenten und

$$\mathcal{R}_n^* = \inf_{\hat{f}_n} \sup_{f\in\mathcal{F}} \mathbb{E} d^2(\hat{f}_n(X), f)$$

die zugehörigen Minimax-Risiken. Die Folge (r_n) heißt Minimax-Konvergenzrate (über $\mathcal{F}$, in d^2-Risiko) falls

$$\limsup_{n\to\infty} r_n^{-2}\mathcal{R}_n^* < \infty \quad \text{und} \quad \liminf_{n\to\infty} r_n^{-2}\mathcal{R}_n^* > 0.$$

1.39 Korollar. *Betrachte das Problem der Dichteschätzung basierend auf n unabhängigen Beobachtungen $X_1,\dots,X_n$ jeweils mit Dichte $f \in \mathcal{F} = \{f \in \mathcal{H}^\alpha(L) : f \text{ W.-dichte}\}$. Dann ist $n^{-\frac{\alpha}{2\alpha+1}}$ Minimax-Konvergenzrate über $\mathcal{F}$ im punktweise quadratischem Risiko.*

1.40 Lemma (Methode von Assouad)**.** *Sei $\mathcal{F}$ eine Teilmenge eines pseudometrischen Raumes (S,d) und $(\mathcal{X}, \mathscr{A}, (P_f)_{f\in\mathcal{F}})$ ein statistisches Experiment. Wir nehmen an, dass es Pseudometriken d_j gibt, so dass*

$$d^2(g,h) \geq \sum_{j=1}^m d_j^2(g,h).$$

Außerdem seien $\{f_\tau : \tau \in \{0,1\}^m\}$ 2^m Elemente in $\mathcal{F}$. Für $\tau \in \{0,1\}^m$ bezeichnen wir mit $\tau^j \in \{0,1\}^m$ das Element welches sich von τ nur in der $j$ten Position unterscheidet. Dann gilt für jeden Schätzer $\hat{f} : \mathcal{X} \to S$

$$\sup_{f\in\mathcal{F}} \mathbb{E}_f d^2(\hat{f}(X), f) \geq \max_{\tau\in\{0,1\}^m} \mathbb{E}_{f_\tau} d^2(\hat{f}(X), f_\tau) \geq \frac{m}{8} \min_{j,\tau} d_j^2(f_\tau, f_{\tau^j}) \int P_{f_\tau} \wedge P_{f_{\tau^j}},$$

wobei das Minimum über alle $j = 1,\dots,m$ und $\tau \in \{0,1\}^m$ genommen wird.

1.41 Zusatz. *Es gilt außerdem*

$$\sup_{f\in\mathcal{F}} \mathbb{E}_f d^2(\hat{f}(X), f) \geq \frac{m}{16} \min_{j,\tau} d_j^2(f_\tau, f_{\tau^j}) \left(\int \sqrt{P_{f_\tau} P_{f_{\tau^j}}} \right)^2 .$$

Beweis. Die erste Ungleichung ist klar. Der Zusatz folgt aus (1.5). Weiter gilt

$$\begin{aligned}
&\max_{\tau\in\{0,1\}^m} \mathbb{E}_{f_\tau} d^2(\hat{f}(X), f_\tau) \\
&\qquad \geq \frac{1}{2^m} \sum_{\tau\in\{0,1\}^m} \mathbb{E}_{f_\tau} d^2(\hat{f}(X), f_\tau) \\
&\qquad \geq \frac{1}{2^m} \sum_{j=1}^m \sum_{\tau\in\{0,1\}^m} \mathbb{E}_{f_\tau} d_j^2(\hat{f}(X), f_\tau) \\
&\qquad = \frac{1}{2^{m+1}} \sum_{j=1}^m \sum_{\tau\in\{0,1\}^m} \Big(\mathbb{E}_{f_\tau} d_j^2(\hat{f}(X), f_\tau) + \mathbb{E}_{f_{\tau^j}} d_j^2(\hat{f}(X), f_{\tau^j}) \Big) \\
&\qquad \geq \frac{1}{2^{m+3}} \sum_{j=1}^m \sum_{\tau\in\{0,1\}^m} d_j^2(f_\tau, f_{\tau^j}) \int P_{f_\tau} \wedge P_{f_{\tau^j}} ,
\end{aligned}$$

wobei die letzte Ungleichung aus dem Beweis von Satz 1.33 folgt. Die Behauptung folgt nun indem wir das Minimum aus der Summe herausziehen. □

1.42 Satz. *Seien $\alpha \in \mathbb{N}$ und $L > 0$. Dann gilt für alle $n \geq 1$ und alle Schätzer $\hat{f}_n : \mathbb{R}^n \to L^2(\mathbb{R})$*

$$\sup_{f\in\mathcal{S}^\alpha(L), f\, W\text{-dichte}} n^{\frac{2\alpha}{2\alpha+1}} \mathbb{E}_f \|\hat{f}_n(X) - f\|_{L^2}^2 \geq c$$

mit einer Konstanten $c > 0$ die nur von α und L abhängt und $X = (X_1, \ldots, X_n)$ hat unter $\mathbb{E}_f$ unabhängige Komponenten jeweils mit Dichte f.

1.7 Eine erste Maximalungleichung

In diesem Kapitel beweisen wir eine erste Maximalungleichung welche wir in den folgenden beiden Kapiteln auf das gleichmäßige Risiko und das Problem des adaptiven Schätzens anwenden werden. Startpunkt ist die Bernstein-Ungleichung:

1.43 Satz (Bernstein-Ungleichung)**.** *Seien $Y_1, \ldots, Y_n$ unabhängige Zufallsvariablen mit $\mathbb{E}Y_i = 0$ und*

$$\mathbb{E}\,|Y_i|^k \leq \frac{k!}{2} \sigma^2 b^{k-2} \qquad \forall k \geq 2.$$

Dann gilt

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n Y_i \geq y \right) \leq \exp\left(- \frac{ny^2}{2\sigma^2 + 2by} \right) \qquad \forall y \geq 0. \tag{1.7}$$

1.44 Zusatz. *Es gilt außerdem*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i \geq \sqrt{2\sigma^2 x} + bx\right) \leq \exp(-nx) \qquad \forall x \geq 0.$$

Offensichtlich impliziert die Bernstein-Ungleichung das schwache Gesetz der großen Zahlen (unter stärkeren Annahmen). Dieses wird in der Regel mit der Chebyshev-Ungleichung bewiesen, welche besagt, dass $\mathbb{P}((1/n)\sum_{i=1}^{n} Y_i \geq x) \leq \mathbb{E}\, Y_1^2/(ny^2) \leq \sigma^2/(ny^2)$. Im Vergleich hierzu liefert die Bernstein-Ungleichung eine stärkere exponentielle Ungleichung. Setzen wir zum Beispiel $y = t/\sqrt{n}$, so erhalten wir für t nicht zu groß annähernd Gaußsche Abweichungen (vergleiche mit Aufgabe 2.2 unten).

1.45 Bemerkung. Setze $S_n = \sum_{i=1}^{n} Y_i$. Aus Symmetriegründen gilt unter den Voraussetzungen von Satz 1.43 außerdem, dass

$$\mathbb{P}(|S_n/n| \geq y) \leq \mathbb{P}(S_n/n \geq y) + \mathbb{P}(-S_n/n \geq y) \leq 2\exp\left(-\frac{ny^2}{2\sigma^2 + 2by}\right).$$

Beweis. Wir betrachten zuerst die Momentenerzeugende Funktion. Für $\lambda \in (0, 1/b)$ gilt

$$\begin{aligned}
\mathbb{E}\, e^{\lambda Y_i} = \mathbb{E}\sum_{k\geq 0} \frac{\lambda^k Y_i^k}{k!} &\overset{(*)}{=} \sum_{k\geq 0} \mathbb{E}\,\frac{\lambda^k Y_i^k}{k!} \\
&\leq 1 + \frac{\sigma^2\lambda^2}{2}\sum_{k\geq 2}(\lambda b)^{k-2} \\
&= 1 + \frac{\sigma^2\lambda^2}{2(1-\lambda b)} \leq \exp\left(\frac{\sigma^2\lambda^2}{2(1-\lambda b)}\right),
\end{aligned}$$

wobei wir in den beiden Ungleichungen die Momentenungleichung, $\mathbb{E}\, Y_i = 0$ und die Ungleichung $1 + x \leq \exp x$ verwendet haben. Die Gleichheit in $(*)$ folgt aus dem Satz von Fubini indem man in einer ähnlichen Rechnung $\sum_{k\geq 0} \mathbb{E}\,\lambda^k |Y_i|^k/k! < \infty$ zeigt. Für $\lambda \in (0, 1/b)$ und $y \geq 0$ gilt nun wegen der Markov-Ungleichung

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i \geq y\right) = \mathbb{P}(S_n \geq ny) &= \mathbb{P}(e^{\lambda S_n} \geq e^{\lambda ny}) \\
&\leq \mathbb{E}\, e^{\lambda S_n} e^{-\lambda ny} \\
&\leq \exp\left(n\left(\frac{\sigma^2\lambda^2}{2(1-\lambda b)} - \lambda y\right)\right), \qquad (1.8)
\end{aligned}$$

wobei wir in der letzten Ungleichung die Unabhängigkeit der Y_i verwendet haben. Die Behauptung folgt nun indem wir $\lambda = y/(\sigma^2 + by)$ wählen, da dann

$$\frac{\sigma^2\lambda^2}{2(1-\lambda b)} - \lambda y = -\frac{y^2}{2(\sigma^2 + by)}.$$

Für den Zusatz schreiben wir (1.8) wir folgt um:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i \geq \frac{x}{\lambda} + \frac{\sigma^2\lambda}{2(1-\lambda b)}\right) \leq \exp(-nx) \qquad \forall \lambda \in (0, 1/b).$$

Es gilt nun

$$\frac{x}{\lambda} + \frac{\sigma^2\lambda}{2(1-\lambda b)} = xb + x\frac{1-\lambda b}{\lambda} + \frac{\sigma^2}{2}\frac{\lambda}{1-\lambda b},$$

und der letzte Ausdruck ist für $\lambda/(1-\lambda b) = \sqrt{2x/\sigma^2}$ gleich $xb + \sqrt{2\sigma^2 x}$. Einsetzen liefert die zweite Behauptung. □

Seien nun $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit Werten in einem messbaren Raum $(S, \mathscr{S})$ und Verteilung P. Außerdem sei $P_n = (1/n)\sum_{i=1}^{n} \delta_{X_i}$ die zugehörige empirische Verteilung. Ist $f : S \to \mathbb{R}$ eine P-integrierbare Funktion, so schreiben wir

$$P_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(X_i) \quad \text{und} \quad P(f) = \int_0^1 f(x)\, P(dx).$$

Ist $\mathscr{F}$ eine Menge von P-integrierbare Funktionen, so heißt der stochastische Prozess

$$f \mapsto \nu_n(f) := P_n(f) - P(f), \quad f \in \mathscr{F}$$

auch empirischer Prozess mit Indexmenge $\mathscr{F}$ (oft zieht man es jedoch vor den normalisierten stochastischen Prozess $f \mapsto \sqrt{n}\nu_n(f)$, $f \in \mathscr{F}$, als empirischen Prozess zu bezeichnen). Ein Beispiel welches zum Kerndichteschätzer führt ist $\mathcal{F} = \{K_h(x - \cdot) : x \in \mathbb{R}\}$. Ziel ist es obere Schranken für

$$\sup_{f \in \mathcal{F}} |P_n(f) - P(f)|$$

zu beweisen. Dabei konzentrieren wir uns in diesem Kapitel auf den Fall, dass $\mathcal{F}$ endlich ist. Verallgemeinerungen werden wir unter anderem in Kapitel 2 kennenlernen.

1.46 Korollar. *Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit Werten in einem messbaren Raum $(S, \mathscr{S})$ und $f : S \to [-b, b]$ eine messbare Funktion. Dann gilt*

$$\mathbb{P}\left(|P_n(f) - P(f)| \geq y\right) \leq 2\exp\left(-\frac{ny^2}{2\operatorname{Var}(f(X_1)) + 2by}\right).$$

Beweis. Setze $Y_i = f(X_i) - \mathbb{E}\, f(X_i)$. Dann sind die Y_i unabhängig, zentriert und es gilt $|Y_i| \leq 2b$ und

$$\mathbb{E}\,|Y_i|^k \leq (2b)^{k-2}\operatorname{Var}(Y_i) = \frac{2^{k-1}}{k!}\frac{k!}{2}b^{k-2}\operatorname{Var}(Y_i) \leq \frac{k!}{2}b^{k-2}\operatorname{Var}(Y_i)$$

für alle $k \geq 2$. Die Aussage folgt nun aus Bemerkung 1.45. □

1.47 Aufgabe. *Sei K eine beschränker Kern und f eine durch L beschränkte Dichte. Dann gilt für den Kerndichteschätzer*

$$\mathbb{P}\left(|\hat{f}_{n,h}(x) - \mathbb{E}\,\hat{f}_{n,h}(x)| \geq t/\sqrt{nh}\right) \leq 2\exp\left(-\frac{t^2}{2L\|K\|_{L^2}^2 + 2\|K\|_\infty t}\right) \qquad \forall t \geq 0.$$

Sei nun $f \in \mathcal{H}^\alpha(L)$ und K ein Kern welcher die Eigenschaften aus Proposition 1.12 erfüllt. In welchem Regime von t kann man eine analoge exponentielle Ungleichung für das Ereignis $\{|\hat{f}_{n,h}(x) - f(x)| \geq t n^{-\frac{\alpha}{2\alpha+1}}\}$ beweisen?

1.48 Korollar. *Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit Werten in einem messbaren Raum $(S, \mathscr{S})$ und $f_1, \ldots, f_M : S \to [-b, b]$ messbare Funktionen. Setze $\sigma^2 = \max_{j=1,\ldots,M} \operatorname{Var}(f_j(X_1))$. Dann gilt für alle $p \in \mathbb{N}$ die folgende Maximalungleichung*

$$\mathbb{E} \max_{j=1,\ldots,M} |P_n(f_j) - P(f_j)|^p \leq C\left(\frac{\sigma^2 \log 2M}{n}\right)^{p/2} + C\left(\frac{b \log 2M}{n}\right)^p$$

mit einer Konstanten C die nur von p abhängt.

Der Beweis kombiniert die Bernstein-Ungleichung mit dem folgenden Lemma:

1.49 Lemma. *Sei $Y \geq 0$ eine Zufallsvariable und $A \geq 1$, $B > 0$ zwei reelle Zahlen.*

(a) Es gelte

$$\mathbb{P}(Y \geq y) \leq A\exp(-y/B) \qquad \forall y \geq 0.$$

Dann gilt für alle $p \in \mathbb{N}$

$$\mathbb{E}\,Y^p \leq (2B\log A)^p + p!(2B)^p.$$

(b) Es gelte

$$\mathbb{P}(Y \geq y) \leq A\exp(-y^2/B^2) \qquad \forall y \geq 0.$$

Dann gilt für alle $p \in \mathbb{N}$

$$\mathbb{E}\,Y^p \leq (2B^2\log A)^{p/2} + (p/2)\Gamma\left(p/2\right)(2B^2)^{p/2},$$

wobei $\Gamma(\cdot)$ die Gammafunktion bezeichnet.

Ist $A \geq 2$ so gilt unter (a) $\mathbb{E}\,Y^p \leq C(B\log A)^p$ und unter (b) $\mathbb{E}\,Y^p \leq C(B^2\log A)^{p/2}$ mit einer Konstanten C die nur von p abhängt.

Beweis von Lemma 1.49. Wir verwenden die Formel $\mathbb{E}\, Y^p = \int_0^\infty p y^{p-1}\, \mathbb{P}(Y \geq y)\, dy$, die man mit Hilfe des Satzes von Tonelli erhält.
(a) Für jedes $y_0 \geq 0$ gilt wegen $\mathbb{P}(Y \geq y) \leq 1$, dass

$$\begin{aligned}\mathbb{E}\, Y^p &= \int_0^{y_0} p y^{p-1}\, \mathbb{P}(Y \geq y)\, dy + \int_{y_0}^\infty p y^{p-1}\, \mathbb{P}(Y \geq y)\, dy \\ &\leq y_0^p + \int_{y_0}^\infty p y^{p-1} A \exp(-y/B)\, dy \\ &= y_0^p + (2B)^p \int_{y_0/(2B)}^\infty p t^{p-1} A \exp(-2t)\, dt.\end{aligned}$$

Wähle nun $y_0 = 2B \log A$. Dann gilt

$$\begin{aligned}\mathbb{E}\, Y^p &\leq (2B \log A)^p + (2B)^p \int_{\log A}^\infty p t^{p-1} A \exp(-2t)\, dt \\ &\leq (2B \log A)^p + (2B)^p \int_0^\infty p t^{p-1} \exp(-t)\, dt \\ &\leq (2B \log A)^p + p!(2B)^p.\end{aligned}$$

(b) Für jedes $y_0 \geq 0$ gilt analog

$$\begin{aligned}\mathbb{E}\, Y^p &\leq y_0^p + \int_{y_0}^\infty p y^{p-1} A \exp(-y^2/B^2)\, dy \\ &= y_0^p + (2B^2)^{p/2} \int_{y_0/\sqrt{2B^2}}^\infty p t^{p-1} A \exp(-2t^2)\, dt.\end{aligned}$$

Wähle nun $y_0 = \sqrt{2B^2 \log A}$. Dann gilt

$$\begin{aligned}\mathbb{E}\, Y^p &\leq (2B^2 \log A)^{p/2} + (2B^2)^{p/2} \int_{\sqrt{\log A}}^\infty p t^{p-1} A \exp(-2t^2)\, dt \\ &\leq (2B^2 \log A)^{p/2} + (2B^2)^{p/2} \int_0^\infty p t^{p-1} \exp(-t^2)\, dt \\ &= (2B^2 \log A)^{p/2} + (p/2)\Gamma(p/2)(2B^2)^{p/2}.\end{aligned}$$

□

1.50 Aufgabe. *Sei $M \geq 2$.*

(a) Sind $Z_1, \ldots, Z_M \sim \mathrm{Exp}(\lambda)$, so gilt $\mathbb{E} \max_{j=1,\ldots,M} Z_j \leq C \log(M)/\lambda$.

(b) Sind $Z_1, \ldots, Z_M \sim \mathcal{N}(0, \sigma^2)$, so gilt $\mathbb{E} \max_{j=1,\ldots,M} |Z_j| \leq C\sigma\sqrt{\log(M)}$.

Beweis von Korollar 1.48. Wir verwenden die Schreibweise $\nu_n(f_j) = P_n(f_j) - P(f_j)$. Dann gilt nach der Bernstein-Ungleichung

$$\mathbb{P}\left(|\nu_n(f_j)| \geq y\right) \leq 2\exp\left(-\frac{ny^2}{2\sigma^2 + 2by}\right) \leq \begin{cases} 2\exp\left(-\frac{ny}{4b}\right), & \text{falls } y > \sigma^2/b, \\ 2\exp\left(-\frac{ny^2}{4\sigma^2}\right), & \text{falls } 0 \leq y \leq \sigma^2/b. \end{cases}$$

Wir zerlegen $|\nu_n(f_j)|$ in $A_j = |\nu_n(f_j)|\mathbf{1}(|\nu_n(f_j)| > \sigma^2/b)$ und $B_j = |\nu_n(f_j)|\mathbf{1}(|\nu_n(f_j)| \le \sigma^2/b)$. Dann gilt

$$\mathbb{P}(A_j \ge y) \le 2\exp\left(-\frac{ny}{4b}\right) \quad \text{und} \quad \mathbb{P}(B_j \ge y) \le 2\exp\left(-\frac{ny^2}{4\sigma^2}\right) \qquad \forall y \ge 0.$$

Es folgt, dass

$$\mathbb{P}(\max_{j=1,\dots,M} A_j \ge y) \le 2M\exp\left(-\frac{ny}{4b}\right) \qquad \forall y \ge 0$$

und

$$\mathbb{P}(\max_{j=1,\dots,M} B_j \ge y) \le 2M\exp\left(-\frac{ny^2}{4\sigma^2}\right) \qquad \forall y \ge 0.$$

Mit Hilfe von Lemma 1.49 schließen wir

$$\begin{aligned}\mathbb{E}\max_{j=1,\dots,M}|\nu_n(f_j)|^p &\le 2^p\,\mathbb{E}\max_{j=1,\dots,M} A_j^p + 2^p\,\mathbb{E}\max_{j=1,\dots,M} B_j^p \\ &\le C\left(\frac{\sigma^2\log 2M}{n}\right)^{p/2} + C\left(\frac{b\log 2M}{n}\right)^p,\end{aligned}$$

was zu zeigen war. □

1.8 Gleichmäßiges Risiko

1.51 Satz. *Die Dichte f erfülle $f \in \mathcal{H}^\alpha(L)$ mit $\alpha, L > 0$. Sei K ein Kern der Ordnung $\lfloor\alpha\rfloor$ welcher zusätzlich beschränkt und Lipschitz-stetig ist. Außerdem seien $a < b$ zwei reelle Zahlen. Es gelte $n \ge 2$ und $h \ge (\log n)/n$. Dann gilt für den Kerndichteschätzer*

$$\mathbb{E}\sup_{x\in[a,b]}|\hat f_{n,h}(x) - f(x)| \le C\Big(h^\alpha + \sqrt{\frac{\log n}{nh}}\Big)$$

mit einer Konstanten C die nur von α, L, K, a, b abhängt. Setzen wir $h = ((\log n)/n)^{\frac{1}{2\alpha+1}}$, so folgt

$$\mathbb{E}\sup_{x\in[a,b]}|\hat f_{n,h}(x) - f(x)| \le 2C\left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+1}}.$$

Beweis. Sei ohne Einschränkung $[a,b] = [0,1]$. Setze

$$\nu_n(x) = \nu_n(K_h(x-\cdot)) = \frac{1}{n}\sum_{i=1}^n K_h(x-X_i) - \mathbb{E}\,K_h(x-X_i) = \hat f_{n,h}(x) - \mathbb{E}\,\hat f_{n,h}(x).$$

Dann gilt mit Hilfe von Proposition 1.12

$$\begin{aligned}&\sup_{x\in[0,1]}|\hat f_{n,h}(x) - f(x)| \le \sup_{x\in[0,1]}|\nu_n(x)| + \sup_{x\in[0,1]}|\,\mathbb{E}\,\hat f_{n,h}(x) - f(x)| \\ &\le \sup_{x\in[0,1]}|\nu_n(x)| + Ch^\alpha.\end{aligned}$$

Die Idee ist nun das Intervall $[0,1]$ zu diskretisieren und dann die Maximalungleichung aus Korollar 1.48 anzuwenden. Wir setzen hierfür $S = \{\delta, 2\delta, \dots, 1\}$. Dann gilt

$$\begin{aligned}&\sup_{x\in[0,1]} |\nu_n(x)| \\ &\leq \sup_{x,y\in[0,1]:|x-y|\leq\delta} |\nu_n(x) - \nu_n(y)| + \max_{y\in S} |\nu_n(y)| \leq 2L_K\delta h^{-2} + \max_{y\in S} |\nu_n(y)|,\end{aligned}$$

wobei L_K die Lipschitz-Konstante von K sei. Es gilt nun weiter, dass $\|K_h(y-\cdot)\|_\infty \leq \|K\|_\infty/h$ und $\mathrm{Var}(K_h(y-X_1)) \leq L\|K\|_{L^2}^2/h$. Korollar 1.48 angwendet mit $p=1$ liefert also für $\delta < 1$

$$\begin{aligned}\mathbb{E}\sup_{x\in[0,1]} |\hat{f}_{n,h}(x) - f(x)| &\leq C(\delta h^{-2} + h^\alpha) + \mathbb{E}\max_{y\in S} |\nu_n(y)| \\ &\leq C\left(\delta h^{-2} + h^\alpha + \sqrt{\frac{\log(1/\delta)}{nh}} + \frac{\log(1/\delta)}{nh}\right).\end{aligned}$$

Damit der Diskretisierungsfehler vernachlässigbar ist setzen wir nun $\delta = n^{-5/2}$. Wir schließen

$$\mathbb{E}\sup_{x\in[0,1]} |\hat{f}_{n,h}(x) - f(x)| \leq C\left(n^{-1/2} + h^\alpha + \sqrt{\frac{\log n}{nh}}\right)$$

wobei wir die Voraussetzung $h \geq \log(n)/n$ eingesetzt haben. □

1.9 Adaptives Schätzen

Die bisherigen Schätzmethoden (Kerndichteschätzer, Projektionsschätzer) hängen von einem sogenannten Tuningparameter ab, nämlich der Bandbreite h beziehungsweise der Dimension d. Wir haben gesehen, dass diese in Abhängingkeit vom Glattheitsindex α gewählt wurden. Dieser ist in der Regel jedoch nicht bekannt. Wie sollen also h und d gewählt werden falls α unbehannt ist?

Wir betrachten im Folgenden den Fall der Projektionsschätzer (eine analoge Betrachtung ist auch für den Kerndichteschätzer möglich). Seien also $(V_m)_{m\in\mathscr{M}}$ eine Menge von lineare Räumen wie oben beschrieben (man spricht von einer Menge von Modellen) und $\hat{f}_{n,m}$, $m \in \mathscr{M}$, die zugehörigen Projektionsschätzer. Eine optimale Wahl von m wäre zum Beispiel

$$m^* \in \operatorname*{argmin}_{m\in\mathscr{M}} \mathbb{E}\,\|\hat{f}_{n,m} - f\|_{L^2}^2.$$

Diese Wahl (man spricht vom sogenannten Orakelmodell) ist jedoch nicht möglich, da die rechte Seite selbst von f abhängt. Ein Ansatz besteht nun

darin das Risiko selbst zu schätzen und diesen Schätzer dann in m zu minimieren. Es gilt

$$\|\hat{f}_{n,m} - f\|_{L^2}^2 = \|\hat{f}_{n,m}\|_{L^2}^2 - 2\int_0^1 \hat{f}_{n,m}(x)f(x)\,dx + \|f\|_{L^2}^2.$$

Da der letzte Term nicht von m abhängt gilt also

$$m^* \in \operatorname*{argmin}_{m\in\mathscr{M}} \|\hat{f}_{n,m}\|_{L^2}^2 - 2\int_0^1 \hat{f}_{n,m}(x)f(x)\,dx.$$

Wollen wir den rechten Term schätzen, so ist ein erster Ansatz diesen durch $(1/n)\sum_{i=1}^n \hat{f}_{n,m}(X_i)$ zu ersetzen. Dies ist allerdings keine gute Idee. Sei hierfür $d_m = \dim V_m$ und $\hat{f}_{n,m} = \sum_{k=1}^{d_m} \hat{\theta}_k\phi_k$ mit $\phi_1,\dots,\phi_{d_m}$ ONB von V_m. Dann gilt

$$\|\hat{f}_{n,m}\|_{L^2}^2 - \frac{2}{n}\sum_{i=1}^n \hat{f}_{n,m}(X_i) = \sum_{k=1}^{d_m}\hat{\theta}_k^2 - 2\sum_{k=1}^{d_m}\hat{\theta}_k^2 = -\|\hat{f}_{n,m}\|_{L^2}^2$$

und die rechte Seite ist monoton fallend für $d_m \to \infty$. Man sollte also nicht mit den gleichen Daten Schätzer konstruieren und evaluieren, da sonst zu große Modelle ausgewählt werden (overfitting). Ein Ausweg ist es die Stichprobe in Tainings- und Evaluierungsmengen zu teilen. Hierfür gibt es mehrere Möglichkeiten:

1. Stichprobenaufspaltung (Hold-out). Zerlege $I = \{1,\dots,n\} = I_1 \cup I_2$ mit $I_1 \cap I_2 = \emptyset$. Sei $S_1 = \{X_i : i \in I_1\}$ Trainingsmenge und $S_2 = \{X_i : i \in I_2\}$ Validierungsmenge.

1. Verwende S_1 um Schätzer zu konstruieren, und zwar $\hat{f}_{I_1,m}$ Projektionsschätzer basierend auf Stichprobe S_1 und Modell V_m, $m \in \mathscr{M}$.

2. Verwende S_2 um den Schätzer zu evaluieren. Setze hierfür

$$R_n^{Ho}(m) = \|\hat{f}_{I_1,m}\|_{L^2}^2 - \frac{2}{|I_2|}\sum_{i\in I_2}\hat{f}_{I_1,m}(X_i)$$

   und betrachte das Minimierungsproblem

$$R_n^{Ho}(m) \to \min, d \in \mathscr{M}.$$

3. Ist $\hat{m}$ eine Lösung des Minimierungsproblems, so betrachte $\hat{f}_{I_1,\hat{m}}$ als finalen Schätzer.

2. V-fache Kreuzvalidierung (V-fold CV). Seien $I_1,\dots,I_V$ disjunkte Mengen mit $I = \bigcup_{k=1}^V I_k$. Setze

$$R_n^{V-FCV}(d) = \frac{1}{V}\sum_{k=1}^V \Big\{\|\hat{f}_{I\setminus I_k,m}\|_{L^2}^2 - \frac{2}{|I_k|}\sum_{i\in I_k}\hat{f}_{I\setminus I_k,m}(X_i)\Big\},$$

d.h. jede Menge wird sukzessive als Validierungsmenge herausgenommen, während die anderen Trainingsmenge sind. Danach wird gemittelt. Ist nun $\hat{m} \in \operatorname{argmin}_{m \in \mathscr{M}} R_n^{V-FCV}(m)$ so betrachten wir $\hat{f}_{I,\hat{m}}$ als finalen Schätzer.

3. n-fache Kreuzvalidierung (leave-one-out). Hier betrachtet man oft alternativ

$$R_n^{Loo}(d) = \|\hat{f}_{I,m}\|_{L^2}^2 - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{I\setminus\{i\},m}(X_i).$$

In allen Fällen ist es das Ziel einen (annähernd) unverzerrten Schätzer von $\mathbb{E}\,\|\hat{f}_{n,m} - f\|_{L^2}^2 - \|f\|_{L^2}^2$ zu konstruieren und diesen dann in $m \in \mathscr{M}$ zu minimieren. Man spricht auch von unbiased risk estimation.

1.52 Aufgabe. *Es gilt* $\mathbb{E}R_n^{Loo}(m) = \mathbb{E}\,\|\hat{f}_{n,m} - f\|_{L^2}^2 - \|f\|_{L^2}^2$.

1.53 Aufgabe. *Im Fall des Histogramm-Schätzers aus Beispiel 1.26 gilt*

$$R_n^{Loo}(m) = -\frac{n+1}{n-1}\|\hat{f}_{n,m}\|_{L^2}^2 + \frac{2m}{n-1}.$$

Analyse der Datenaufspaltung

Wir analysieren zuerst den zweiten Schritt der Datenaufspaltung. Bedingen wir auf die Trainingsmenge S_1, so sind die Schätzer $f_{I_1,m}$ deterministische Funktionen. Wir betrachten daher zuerst das folgende vereinfachte Problem. Seien $X_1, \ldots, X_n$ i.i.d. Zufallsvariablen mit Werten in $[0,1]$ und Dichte f und $f_1, \ldots, f_M \in L^2([0,1])$ deterministische Funktionen. Setze

$$\hat{m} \in \operatorname*{argmin}_{m=1,\ldots,M} \|f_m\|_{L^2}^2 - 2P_n(f_m) \quad \text{und} \quad \hat{f} = f_{\hat{m}}. \tag{1.9}$$

Der folgende Satz vergleicht den Abstand $\|\hat{f} - f\|_{L^2}$ mit dem kleinstmöglichen Abstand $\min_{m=1,\ldots,M} \|f_m - f\|_{L^2}$.

1.54 Satz. *Die Dichte f sei beschränkt durch L. Seien $f_1, \ldots, f_M \in L^2([0,1])$ paarweise verschiedene deterministische Funktionen und Φ_M eine reelle Zahl mit $\|f_m - f_{m'}\|_\infty^2 \le \Phi_M \|f_m - f_{m'}\|_{L^2}^2$ für alle $m \neq m'$. Dann gilt für den in* (1.9) *konstruierten Schätzer*

$$\mathbb{E}\,\|\hat{f} - f\|_{L^2}^2 \le 3 \min_{m=1,\ldots,M} \|f_m - f\|_{L^2}^2 + 8C\left(\frac{L\log M}{n} + \frac{\Phi_M(\log M)^2}{n^2}\right),$$

wobei C die Konstante aus Korollar 1.48 mit $p = 2$ ist.

1.55 Bemerkung. Die Konstante 3 kann durch $1+\epsilon$, $\epsilon > 0$ beliebig klein, ersetzt werden. Allerdings müssen wir gleichzeitig die Konstante 8 durch eine Konstante ersetzen, die von der Größenordnung $1/\epsilon$ ist.

Beweis. Wir setzen

$$m^* \in \operatorname*{argmin}_{j=1,\dots,M} \|f_m - f\|_{L^2}^2 = \operatorname*{argmin}_{m=1,\dots,M} \|f_m\|_{L^2}^2 - 2P(f_m)$$

und $f^* = f_{m^*}$. Es gilt nun nach Definition

$$\|\hat{f}\|_{L^2}^2 - 2P_n(\hat{f}) \leq \|f^*\|_{L^2}^2 - 2P_n(f^*)$$

und somit durch quadratische Ergänzung

$$\begin{aligned}&\|\hat{f} - f\|_{L^2}^2 - \|f\|_{L^2}^2 + 2\int_0^1 \hat{f}(x)f(x)\,dx - 2P_n(\hat{f})\\ &\leq \|f^* - f\|_{L^2}^2 - \|f\|_{L^2}^2 + 2\int_0^1 f^*(x)f(x)\,dx - 2P_n(f^*),\end{aligned}$$

was geschrieben werden kann als

$$\|\hat{f}-f\|_{L^2}^2-\|f\|_{L^2}^2+2P(\hat{f})-2P_n(\hat{f}) \leq \|f^*-f\|_{L^2}^2-\|f\|_{L^2}^2+2P(f^*)-2P_n(f^*).$$

Insgesamt folgt also

$$\|\hat{f} - f\|_{L^2}^2 \leq \|f^* - f\|_{L^2}^2 + 2\nu_n(\hat{f} - f^*)$$

was oft auch als Fundamentale Ungleichung bezeichnet wird. Im Prinzip könnte man an dieser Stelle die Maximalungleichung anwenden indem man zuerst wie folgt abschätzt $\nu_n(\hat{f} - f^*) \leq \max_{m=1,\dots,M} \nu_n(f_m - f^*)$. Letzterer Ausdruck ist das Maximum eines empirischen Prozesses deren Erwartungswert mit Hilfe von Korollar 1.48 abgeschätzt werden kann. Diese Ungleichung führt allerdings zu einem suboptimalen Faktor $\sqrt{(\log M)/n}$ und ist daher zu grob. Wir verwenden daher das folgende schärfere Argument:

$$\begin{aligned}&\|\hat{f} - f\|_{L^2}^2\\ &\leq \|f^* - f\|_{L^2}^2 + 2\|\hat{f} - f^*\|_{L^2} \cdot \max_{m\neq m^*} \frac{\nu_n(f_m - f^*)}{\|f_m - f^*\|_{L^2}}\\ &\leq \|f^* - f\|_{L^2}^2 + \frac{1}{4}\|\hat{f} - f^*\|_{L^2}^2 + 4\left(\max_{m\neq m^*} \frac{\nu_n(f_m - f^*)}{\|f_m - f^*\|_{L^2}}\right)^2\\ &\leq \|f^* - f\|_{L^2}^2 + \frac{1}{2}\|\hat{f} - f\|_{L^2}^2 + \frac{1}{2}\|f - f^*\|_{L^2}^2 + 4\left(\max_{m\neq m^*} \frac{\nu_n(f_m - f^*)}{\|f_m - f^*\|_{L^2}}\right)^2,\end{aligned}$$

wobei wir die Ungleichungen $2ab \leq a^2 + b^2$, $(a + b)^2 \leq 2a^2 + 2b^2$ und die Dreiecksungleichung verwendet haben. Umordnen der Terme liefert also

$$\|\hat{f} - f\|_{L^2}^2 \leq 3\|f - f^*\|_{L^2}^2 + 8\left(\max_{m\neq m^*} \frac{\nu_n(f_m - f^*)}{\|f_m - f^*\|_{L^2}}\right)^2,$$

d.h. im Vergleich zu dem direkten Abschätzen von oben haben wir ein zusätzliches Quadrat gewonnen. Der Term im Quadrat kann nun geschrieben werden als

$$\max_{m \neq m^*} \frac{1}{n} \sum_{i=1}^{n} h_m(X_i) - \mathbb{E}\, h_m(X_i)$$

mit normalisierten Funktionen $h_m = (f_m - f^*)/\|f_m - f^*\|_{L^2}$. Es gilt nun nach Voraussetzung

$$\mathrm{Var}(h_m(X_1)) \leq \mathbb{E}\, h_m^2(X_1) = \frac{\int_0^1 (f_m(x) - f^*(x))^2 f(x)\, dx}{\|f_m - f^*\|_{L^2}^2} \leq L$$

und

$$\|h_m\|_\infty = \frac{\|f_m - f^*\|_\infty}{\|f_m - f^*\|_{L^2}} \leq \sqrt{\Phi_M}.$$

Daher liefert Korollar 1.48, dass

$$\mathbb{E}\, \|\hat{f} - f\|_{L^2}^2 \leq 3\|f - f^*\|_{L^2}^2 + 8C \left(\frac{L \log M}{n} + \frac{\Phi_M (\log M)^2}{n^2} \right),$$

was zu zeigen war. □

Wir konstruieren nun unseren Schätzer basierend auf der Datenaufsplittung wie folgt. Seien $n_1 = [n/2]$, $n_2 = n - n_1$, $I_1 = \{1, \ldots, n_1\}$, $I_2 = \{n_1 + 1, \ldots, n\}$, $\mathscr{M} = \{1, \ldots, M\}$ mit $M \geq 2$ und $\{d_1, \ldots, d_m\}$ eine Menge von natürlichen Zahlen.

1. Konstruiere $\hat{f}_{I_1,1}, \ldots, \hat{f}_{I_1,M}$ basierend auf $S_1 = \{X_1, \ldots, X_{n_1}\}$, und zwar $\hat{f}_{I_1,m}$ Projektionsschätzer mit den ersten d_m Basisfunktionen der trigonometrischen Basis $(\phi_k)_{k \geq 1}$.
2. Wähle $\hat{m} \in \mathrm{argmin}_{m=1,\ldots,M} \|\hat{f}_{I_1,m}\|_{L^2}^2 - \frac{2}{n_2} \sum_{i \in I_2} \hat{f}_{I_1,m}(X_i)$.
3. Setze $\hat{f} = \hat{f}_{I_1,\hat{m}}$.

1.56 Satz. *Die Dichte f sei beschränkt durch L und die d_m aus 1. erfüllen $d_m \leq n$. Dann gilt für den in 1.-3. konstruierten Schätzer $\hat{f}$ die folgende Orakelungleichung*

$$\mathbb{E}\, \|\hat{f} - f\|_{L^2}^2 \leq 3 \min_{m=1,\ldots,M} \mathbb{E}\, \|\hat{f}_{I_1,m} - f\|_{L^2}^2 + C \frac{L \log M + (\log M)^2}{n}.$$

mit einer absoluten Konstante C.

Beweis. Es gilt $\mathbb{E}\, \|\hat{f} - f\|_{L^2}^2 = \mathbb{E}_{(1)}\, \mathbb{E}_{(2)}\, \|\hat{f} - f\|_{L^2}^2$ mit $\mathbb{E}_{(2)}$ Erwartung bezüglich $X_{n_1+1}, \ldots, X_n$ mit $X_1, \ldots, X_{n_1}$ festgehalten und $\mathbb{E}_{(1)}$ Erwartung bezüglich $X_1, \ldots, X_{n_1}$. Fixieren wir $X_1, \ldots, X_{n_1}$, so sind die Schätzer

$\hat{f}_{I_1,1}, \dots, \hat{f}_{I_1,M}$ deterministische Funktionen und wir können Satz 1.54 anwenden (mit $n = n_2$ und $f_m = \hat{f}_{I_1,m}$). Wir behaupten nun, dass die Voraussetzung aus Satz 1.54 mit $\Phi_M = 2n$ erfüllt sind. Setze hierfür $d = \max_{m=1,\dots,M} d_m$. Ist nun $h = \sum_{k=1}^{d} \theta_k \phi_k$ eine Funktion, so gilt wegen der Cauchy-Schwarz-Ungleichung

$$|h(x)|^2 \leq \Big(\sum_{k=1}^{d} \theta_k^2 \Big) \Big(\sum_{k=1}^{d} \phi_k^2(x) \Big) \leq 2d \|h\|_{L^2}^2$$

für alle $x \in [0,1]$ und somit $\|h\|_\infty^2 \leq 2d\|h\|_{L^2}^2$. Da alle $\hat{f}_{I_1,m}$ eine solche Darstellung besitzen und $d \leq n$ nach Voraussetzung gilt, folgt also $\|\hat{f}_{I_1,m} - \hat{f}_{I_1,m'}\|_\infty^2 \leq \Phi_M \|\hat{f}_{I_1,m} - \hat{f}_{I_1,m'}\|_{L^2}^2$ für alle $m \neq m'$ mit $\Phi_M = 2n$. Satz 1.54 liefert also ($n_2 \geq n/2$)

$$\mathbb{E}_{(2)} \|\hat{f} - f\|_{L^2}^2 \leq 3 \min_{j=m,\dots,M} \|\hat{f}_{I_1,m} - f\|_{L^2}^2 + C \frac{L \log M + (\log M)^2}{n}$$

Wir schließen

$$\begin{aligned}
\mathbb{E} \|\hat{f} - f\|_{L^2}^2 &= \mathbb{E}_{(1)} \mathbb{E}_{(2)} \|\hat{f} - f\|_{L^2}^2 \\
&\leq 3 \, \mathbb{E}_{(1)} \min_{m=1,\dots,M} \|\hat{f}_{I_1,m} - f\|_{L^2}^2 + C \frac{L \log M + (\log M)^2}{n} \\
&\leq 3 \min_{m=1,\dots,M} \mathbb{E}_{(1)} \|\hat{f}_{I_1,m} - f\|_{L^2}^2 + C \frac{L \log M + (\log M)^2}{n} \\
&= 3 \min_{m=1,\dots,M} \mathbb{E} \|\hat{f}_{I_1,m} - f\|_{L^2}^2 + C \frac{L \log M + (\log M)^2}{n}.
\end{aligned}$$

□

1.57 Korollar. *Setze $d_m = [n^{\frac{1}{2m+1}}]$, $m = 1, \dots, M$. Für $\alpha \in \mathbb{N}$ und $L > 0$ betrachte $\mathcal{F}_{\alpha,L} = \{f \in \mathcal{W}^\alpha(L) : f$ durch L beschränkte W.-dichte$\}$ Dann gilt für den in 1.-3. konstruierten Schätzer*

$$\forall \alpha \in \{1, \dots, M\} : \qquad \sup_{f \in \mathcal{F}_{\alpha,L}} \mathbb{E}_f \|\hat{f} - f\|_{L^2}^2 \leq C n^{-\frac{2\alpha}{2\alpha+1}}$$

mit einer Konstanten C die nur von L und M abhängt.

Beweis. Folgt aus (1.4) und Satz 1.56, da der Restterm kleiner ist als Cn^{-1} und für $f \in \mathcal{F}_{\alpha,L}$ folgendes gilt

$$\min_{m=1,\dots,M} \mathbb{E} \|\hat{f}_m - f\|_{L^2}^2 \overset{m=\alpha}{\leq} C n^{-\frac{2\alpha}{2\alpha+1}}.$$

□

1.58 Bemerkung. Wir haben also einen Schätzer konstruiert, der die optimalen Konvergenzraten besitzt und dabei ohne a priori-Kenntnis von α auskommt, man spricht von einem adaptiven Schätzer.

2 Sub-Gaußsche Prozesse

Die Theorie der empirischen und der sub-Gaußschen Prozesse hat sich als fundamentales Werkzeug in der (nichtparametrischen) Statistik etabliert. In diesem Kapitel werden wir einige fundamentale Resultate über (sub-)Gaußsche Prozesse kennenlernen, wie z.B. das Gaußsche Konzentrationsphänomen und Dudleys Entropie-Schranke. Diese und weitere Resultate können auch in Boucheron, Lugosi und Massart (2013), Massart (2007) und Giné und Nickl (2016) nachgelesen werden.

2.1 Konzentrationsungleichungen

In Kapitel 1 stand häufig eine nicht-asymptotische Analyse im Vordergrund. Ist dies der Fall, so sind Konzentrationsungleichungen oft die entscheidenden Hilfsmittel. In diesem Kapitel wollen wir die Bernstein-Ungleichung aus Kapitel 1 durch weitere Konzentrationsungleichungen ergänzen.

2.1 Definition. Eine reellwertige, zentrierte Zufallsvariable X heißt subgaußsch mit Parameter $\sigma^2 > 0$, falls

$$\mathbb{E}\, e^{\lambda X} \leq e^{\frac{\lambda^2 \sigma^2}{2}} \qquad \text{für alle } \lambda \in \mathbb{R}\,.$$

Schreibweise: $X \in \mathrm{SG}(\sigma^2)$.

2.2 Aufgabe. *Sei X eine reellwertige, zentrierte Zufallsvariable. Dann sind folgende Aussagen äquivalent, wobei sich die Parameter $\sigma_1, \sigma_2, \sigma_3, \sigma_4 > 0$ jeweils nur um eine absolute Konstante unterscheiden:*

(i) $X \in \mathrm{SG}(\sigma_1^2)$*;*

(ii) $\mathbb{P}(|X| \geq u) \leq 2e^{-u^2/(2\sigma_2^2)}$ *für alle* $u \geq 0$*;*

(iii) $(\mathbb{E}\,|X|^p)^{1/p} \leq \sigma_3 \sqrt{p}$ *für alle* $p \in \mathbb{N}$*;*

(iv) $\mathbb{E}\, e^{X^2/\sigma_4^2} \leq 2$*.*

Lösung. (i) impliziert (ii) mit $\sigma_2^2 = \sigma_1^2$: Für $u, \lambda \geq 0$ gilt mit der Markov-Ungleichung

$$\mathbb{P}(X \geq u) = \mathbb{P}(e^{\lambda X} \geq e^{\lambda u}) \leq \mathbb{E}\, e^{\lambda X} e^{-\lambda u} \leq e^{\lambda^2 \sigma_1^2/2 - \lambda u} \overset{\lambda = u/\sigma_1^2}{=} e^{-u^2/(2\sigma_1^2)}.$$

Setzt man $X' := -X \in \mathrm{SG}(\sigma_1^2)$, so erhält man analog $\mathbb{P}(X \leq -u) = \mathbb{P}(X' \geq u) \leq e^{-u^2/(2\sigma_1^2)}$. Mittels Sub-Additivität erhalten wir die Behauptung.
(ii) impliziert (iii) mit $\sigma_3 = e^{1/e}\sigma_2$: Verwenden wir (ii) so gilt

$$\begin{aligned} \mathbb{E}\,|X|^p &= \int_0^\infty p u^{p-1}\, \mathbb{P}(|X| \geq u)\, du \\ &\leq 2 \int_0^\infty p u^{p-1} e^{-u^2/(2\sigma_2^2)}\, du = \sigma_2^p 2^{p/2} p \Gamma(p/2). \end{aligned}$$

Ziehen wir die p-te Wurzel und verwenden wir die Ungleichungen $\Gamma(p/2) \leq (p/2)^{p/2}$, so folgt

$$(\mathbb{E}\,|X|^p)^{1/p} \leq p^{1/p}\sigma_2 p^{1/2} \leq e^{1/e}\sigma_2 p^{1/2}.$$

(iii) impliziert (iv) mit $\sigma_4^2 = 2e\sigma_3^2$: Es gilt

$$\mathbb{E}\,e^{X^2/\sigma_4^2} = 1 + \sum_{p=1}^{\infty} \frac{\mathbb{E}\,|X|^{2p}}{p!\sigma_4^{2p}} \leq 1 + \sum_{k=1}^{\infty} \frac{p^p\sigma_3^{2p}}{p!\sigma_4^{2p}} \leq 1 + \sum_{k=1}^{\infty} \left(\frac{e\sigma_3^2}{\sigma_4^2}\right)^p,$$

wobei wir die Ungleichung $p! \geq (p/e)^p$ verwendet haben. Die rechte Seite ist gleich 2, falls wir $\sigma_4^2 = 2e\sigma_3^2$ wählen.
(iv) impliziert (i) mit $\sigma_1^2 = 5\sigma_3^2$: Es gilt

$$\mathbb{E}\,e^{\lambda X} = 1 + \mathbb{E}\int_0^1 (1-t)\lambda^2 X^2 \exp(t\lambda X)\,dt \leq 1 + \frac{\lambda^2}{2}\,\mathbb{E}\,X^2 e^{|\lambda X|}, \qquad (2.1)$$

wobei wir in der ersten Gleichheit die Taylorsche Formel und die Tatsache, dass $\mathbb{E}\,X = 0$ verwendet haben. Weiter gilt mit Hilfe der Formeln $ab \leq a^2/2 + b^2/2$ und $ae^a \leq e^{2a}$, dass

$$X^2 e^{\lambda X} = X^2 e^{\lambda\sigma_4 \frac{X}{\sigma_4}} \leq X^2 e^{\frac{\lambda^2\sigma_4^2}{2}} e^{\frac{X^2}{2\sigma_4^2}} \leq 2\sigma_4^2 e^{\frac{\lambda^2\sigma_4^2}{2}} e^{\frac{X^2}{\sigma_4^2}}.$$

Setzen wir dies in (2.1) ein und verwenden Aussage (iv), so folgt

$$\mathbb{E}\,e^{\lambda X} \leq 1 + \lambda^2\sigma_4^2 e^{\frac{\lambda^2\sigma_4^2}{2}}\,\mathbb{E}\,e^{\frac{X^2}{\sigma_4^2}} \leq 1 + 2\lambda^2\sigma_4^2 e^{\frac{\lambda^2\sigma_4^2}{2}} \leq e^{\frac{5\lambda^2\sigma_4^2}{2}}.$$

□

Ist X normalverteilt mit Erwartungswert 0 und Varianz σ^2, so gilt $X \in \mathrm{SG}(\sigma^2)$. Des Weiteren ist jede beschränkte Zufallsvariable sub-Gaußsch (folgt aus Aufgabe 2.2 (iii) oder (iv)). Dass folgende Resultat liefert eine konkrete Konstante:

2.3 Lemma (Hoeffdings Lemma)**.** *Sei X eine zentrierte Zufallsvariable mit Werten in $[a, b]$. Dann gilt $X \in \mathrm{SG}((b-a)^2/4)$.*

Ist X zum Beispiel eine Rademacher Zufallsvariable, d.h. mit $\mathbb{P}(X = -1) = \mathbb{P}(X = 1) = 1/2$, so gilt $X \in \mathrm{SG}(1)$. Man kann zeigen (Übung), dass X nicht sub-Gaußsch mit echt kleinerem Parameter als 1 ist.

Beweis. Zunächst gilt

$$\Big|X - \frac{b+a}{2}\Big| \leq \frac{b-a}{2}$$

und somit

$$\mathrm{Var}(X) = \mathrm{Var}\Big(X - \frac{b+a}{2}\Big) \leq \frac{(b-a)^2}{4}. \qquad (2.2)$$

Wir müssen nun zeigen, dass

$$\psi(\lambda) := \log \mathbb{E}\, e^{\lambda X} \leq \frac{\lambda^2 (b-a)^2}{8} \qquad \forall \lambda \in \mathbb{R}\,.$$

Wir berechnen die ersten beiden Ableitungen:

$$\psi'(\lambda) = \frac{\mathbb{E}\, X e^{\lambda X}}{\mathbb{E}\, e^{\lambda X}}, \qquad \psi''(\lambda) = \frac{\mathbb{E}\, X^2 e^{\lambda X}}{\mathbb{E}\, e^{\lambda X}} - \left(\frac{\mathbb{E}\, X e^{\lambda X}}{\mathbb{E}\, e^{\lambda X}}\right)^2$$

Dass Differentiation und Integration vertauscht werden dürfen, kann man zum Beispiel mit dem Satz von der dominierten Konvergenz zeigen (siehe z.B. [**?**, Corollary 5.9]). Es gilt nun

$$\psi''(\lambda) = \mathbb{E}(X^2 e^{\lambda X} e^{-\psi(\lambda)}) - \left(\mathbb{E}(X e^{\lambda X} e^{-\psi(\lambda)})\right)^2 = \operatorname{Var}(Z),$$

wobei Z eine Zufallsvariable ist, welche die Dichte $e^{\lambda x} e^{-\psi(\lambda)}$ bezüglich der Verteilung P von X besitzt. Da Z Werte in $[a, b]$ annimmt, folgt aus (2.2) dass $\operatorname{Var}(Z) \leq (b-a)^2/4$. Es gilt nun $\psi(0) = \psi'(0) = 0$ und mit Hilfe der Taylorschen Formel folgt

$$\psi(\lambda) = \psi(0) + \lambda\psi'(0) + \frac{\lambda^2}{2}\psi''(\tau\lambda) \leq \frac{\lambda^2 (b-a)^2}{8}.$$

□

2.4 Lemma. *Sind $X_1, \ldots, X_n$ unabhängige Zufallsvariablen mit $X_i \in \mathrm{SG}(\sigma_i^2)$, so gilt $X_1 + \cdots + X_n \in \mathrm{SG}(\sigma_1^2 + \cdots + \sigma_n^2)$.*

Beweis. Es gilt

$$\mathbb{E}\, e^{\lambda(X_1 + \cdots + X_n)} = \mathbb{E} \prod_{i=1}^n e^{\lambda X_i} = \prod_{i=1}^n \mathbb{E}\, e^{\lambda X_i} \leq \prod_{i=1}^n e^{\lambda^2 \sigma_i^2/2} = e^{\lambda^2(\sigma_1^2 + \cdots + \sigma_n^2)/2}.$$

□

Kombinieren wir Lemma 2.3, Lemma 2.4 und Aufgabe 2.2, so erhalten wir:

2.5 Satz (Hoeffding-Ungleichung)**.** *Seien $X_1, \ldots, X_n$ unabhängige, zentrierte Zufallsvariablen wobei X_i Werte in $[a_i, b_i]$ annimmt. Dann gilt*

$$\mathbb{P}\left(|X_1 + \cdots + X_n| \geqslant u\right) \leqslant 2 \exp\left(-\frac{2u^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

2.6 Beispiel. Seien $\epsilon_1, \dots, \epsilon_n$ unabhängige Rademacher Zufallsvariablen. Dann ist der Rademacher Prozess definiert durch

$$X(t) = \sum_{i=1}^{n} t_i \epsilon_i, \qquad t \in \mathbb{R}^n .$$

Es gilt $X(t) - X(s) \in \mathrm{SG}(\|t-s\|_2^2)$ und somit insbesondere

$$\mathbb{P}(|X(t) - X(s)| \geq u) \leq \exp\left(-\frac{u^2}{2\|t-s\|_2^2}\right).$$

Wir haben bis jetzt nur Linearkombinationen von $X_1, \dots, X_n$ behandelt und wollen als nächstes allgemeinere Kombinationen $F(X_1, \dots, X_n)$ betrachten. Es stellt sich heraus, dass Lipschitz-Bedingungen ausreichen um obige Resultate zu verallgemeinern. Ein erstes Resultat:

2.7 Satz (McDiarmid-Ungleichung)**.** *Seien $X_1, \dots, X_n$ unabhängige Zufallsvariablen, wobei X_i Werte in einem messbaren Raum $(S_i, \mathscr{S}_i)$ annimmt und sei $F : S_1 \times \cdots \times S_n \to \mathbb{R}$ eine messbare Funktion. Wir nehmen an, dass $c_1, \dots, c_n > 0$ existieren, so dass*

$$|F(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) - F(x_1, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_n)| \leq c_i$$

für alle $x_1 \in S_1, \dots, x_n \in S_n$, $x_i' \in S_i$ (d.h. werden alle Koordinaten bis auf die i-te festgehalten, so fluktuiert F höchstens um c_i). Dann gilt

$$\mathbb{P}\left(|F(X_1, \dots, X_n) - \mathbb{E}\, F(X_1, \dots, X_n)| \geq u\right) \leqslant 2\exp\left(-\frac{2u^2}{\sum_{i=1}^n c_i^2}\right).$$

In dem Fall $F(x_1, \dots, x_n) = x_1 + \cdots + x_n$ und $S_i = [a_i, b_i]$ stimmt die McDiarmid-Ungleichung mit der Hoeffding-Ungleichung überein. Die McDiarmid-Ungleichung liefert allerdings auch eine gleichmäßige Version der Hoeffding-Ungleichung:

2.8 Aufgabe. *Seien $X_1, \dots, X_n$ i.i.d. mit Werten in $(S, \mathscr{S})$ und $\mathcal{F}$ eine abzählbare Menge von Funktionen $f : S \to [-b, b]$. Dann erfüllt F definiert durch*

$$F(X_1, \dots, X_n) = \sup_{f \in \mathcal{F}} |P_n(f) - P(f)|$$

die Bedingung aus Satz 2.7 mit $c_i = 2b/n$ und es gilt

$$\mathbb{P}\left(\left|\sup_{f\in\mathcal{F}} |P_n(f) - P(f)| - \mathbb{E}\sup_{f\in\mathcal{F}} |P_n(f) - P(f)|\right| \geq u\right) \leq 2\exp\left(-\frac{nu^2}{2b^2}\right)$$

für alle $u \geq 0$.

Beweis. Wir schreiben $X = (X_1, \dots, X_n)$. Indem wir eine Konstante von F abziehen, können wir annehmen, dass $\mathbb{E}\, F(X) = 0$ gilt. Wegen Aufgabe 2.2 reicht es zu zeigen, dass

$$\mathbb{E} \exp(\lambda F(X)) \leq \exp\left(\frac{\lambda^2 \sum_{i=1}^n c_i^2}{8}\right) \qquad \forall \lambda \in \mathbb{R}\,. \tag{2.3}$$

Wir berechnen die linke Seite iterativ indem wir in einem ersten Schritt auf $X_1, \dots, X_{n-1}$ bedingen:

$$\begin{aligned}
&\mathbb{E} \exp(\lambda F(X)) \\
&= \mathbb{E}\left(\mathbb{E}(\exp(\lambda F(X))|X_1, \dots, X_{n-1})\right) \\
&= \mathbb{E}\left(\mathbb{E}(\exp(\lambda \Delta_n(X))|X_1, \dots, X_{n-1}) \exp(\lambda\, \mathbb{E}(F(X)|X_1, \dots, X_{n-1}))\right),
\end{aligned}$$

wobei wir

$$\Delta_n(X) = F(X) - \mathbb{E}(F(X)|X_1, \dots, X_{n-1})$$

gesetzt haben. Fixieren wir $X_1, \dots, X_{n-1}$, so nimmt $\Delta_n(X)$ Werte in $[A_n, B_n]$ an, wobei

$$A_n = \inf_{x_n \in S_n} \Delta_n(X_1, \dots, X_{n-1}, x_n), \qquad B_n = \sup_{x_n \in S_n} \Delta_n(X_1, \dots, X_{n-1}, x_n)$$

und nach Voraussetzung gilt $B_n - A_n \leq c_n$. Verwenden wir außerdem, dass $E(\Delta_n(X)|X_1, \dots, X_{n-1}) = 0$ gilt, so folgt aus Lemma 2.3, dass

$$\mathbb{E} \exp(\lambda F(X)) \leq \exp\left(\frac{\lambda^2 c_n^2}{8}\right) \mathbb{E} \exp(\lambda\, \mathbb{E}(F(X)|X_1, \dots, X_{n-1})).$$

Es gilt nun

$$\mathbb{E}(F(X)|X_1, \dots, X_{n-1}) = F_{n-1}(X_1, \dots, X_{n-1}),$$

wobei F_{n-1} die gleichen Voraussetzungen wie F erfüllt mit n ersetzt durch $n-1$ und (2.3) folgt durch Iteration. □

Stärkere Konzentrationsungleichungen benötigen nur eine globale Lipschitzbedingung. Wir beschränken uns hier auf den Fall Gaußscher Zufallsvariablen.

2.9 Satz. *Seien $X_1, \dots, X_n$ unabhängige Zufallsvariablen mit $X_i \sim \mathcal{N}(0,1)$. und $F : \mathbb{R}^n \to \mathbb{R}$ L-Lipschitz mit $L > 0$ (d.h. $|F(x) - F(y)| \leq L\|x - y\|_2$ $\forall x, y \in \mathbb{R}^n$). Dann gilt*

$$\mathbb{P}\left(|F(X_1, \dots, X_n) - \mathbb{E}\, F(X_1, \dots, X_n)| \geq u\right) \leqslant 2 \exp\left(-\frac{u^2}{2C^2L^2}\right)$$

mit $C = \pi/2$.

2.10 Bemerkung. Obiges Resultat gilt sogar mit $C = 1$, der Erwartungswert kann durch den Median ersetzt werden. Man spricht auch von dem Gaußschen Konzentrationsphänomen.

Beweis. Ohne Einschränkung gelte $\mathbb{E}\, F(X) = 0$ und $L = 1$. Wegen Aufgabe 2.2 reicht es zu zeigen, dass

$$\mathbb{E} \exp(\lambda F(X)) \leq \exp(\lambda^2 C^2/2) \qquad \forall \lambda \in \mathbb{R}\,. \tag{2.4}$$

Weiter reicht es aus (2.4) für F stetig (partiell) differenzierbar zu zeigen, da der allgemeine Fall dann mit Hilfe eines Limesargumentes folgt (verwende: ist (F_n) eine Folge von Funktionen die punktweise gegen F konvergiert und für die (2.4) gilt, so folgt aus dem Lemma von Fatou, dass (2.4) auch für F gilt. Eine geeignete Folge (F_n) kann zum Beispiel mittels Faltungen konstruiert werden). Alternativ kann man im Folgenden auch auf die Einschränkung der stetigen Differenzierbarkeit verzichten und verwenden, dass jede Lipschitz-Funktion Lebesgue-fast überall differenzierbar ist. Es folgt nun, dass $\|\nabla F(x)\|_2 \leq \|F\|_{\mathrm{Lip}} \leq 1$ für alle $x \in \mathbb{R}^n$. Sei nun Y eine unabhängige Kopie von X. Dann gilt mit Hilfe der Jensenschen Ungleichung, dass

$$\mathbb{E} \exp(-\lambda F(Y)) \geq \exp(-\lambda\, \mathbb{E}\, F(Y)) = 1$$

und somit wegen der Unabhängigkeit von X und Y, dass

$$\mathbb{E} \exp(\lambda F(X)) \leq \mathbb{E} \exp(\lambda (F(X) - F(Y))).$$

Für $\theta \in [0, \pi/2]$ setzen wir

$$Z(\theta) = Y \cos\theta + X \sin\theta, \qquad Z'(\theta) = -Y \sin\theta + X \cos\theta$$

Dann sind $Z(\theta)$ und $Z'(\theta)$ unabhängige standardnormalverteilte Zufallsvektoren für alle $\theta \in [0, \pi/2]$ (die Rotationsinvarianz liefert, dass $(Z(\theta), Z'(\theta))$ für alle $\theta \in [0, \pi/2]$ die gleiche Verteilung besitzt). Es gilt nun

$$F(X) - F(Y) = \int_0^{\pi/2} \frac{d}{d\theta} F(Z(\theta))\, d\theta = \int_0^{\pi/2} \langle \nabla F(Z(\theta)), Z'(\theta) \rangle\, d\theta$$

und einsetzen liefert

$$\begin{aligned} \mathbb{E} \exp(\lambda F(X)) &\leq \mathbb{E} \exp\left(\lambda \int_0^{\pi/2} \langle \nabla F(Z(\theta)), Z'(\theta) \rangle\, d\theta \right) \\ &\leq \frac{2}{\pi} \int_0^{\pi/2} \mathbb{E} \exp\left(\frac{\lambda\pi}{2} \langle \nabla F(Z(\theta)), Z'(\theta) \rangle \right) d\theta, \end{aligned} \tag{2.5}$$

wobei wir in der letzten Ungleichung noch mal die Jensensche Ungleichung angewendet haben. Bedingen wir auf $Z(\theta)$, so ist $\langle \nabla F(Z(\theta)), Z'(\theta) \rangle$ normalverteilt mit Erwartungswert 0 und Varianz $\|\nabla F(x)\|_2^2 \leq 1$, d.h. es gilt

$$\mathbb{E}\left(\exp\left(\frac{\lambda\pi}{2} \langle \nabla F(Z(\theta)), Z'(\theta) \rangle \right) \middle| Z(\theta) \right) \leq \exp(\lambda^2 \pi^2/8).$$

Die Behauptung folgt nun indem wir diese Ungleichung in (2.5) einsetzen. □

2.11 Aufgabe. *Seien $\epsilon_1, \dots, \epsilon_n$ unabhängige Zufallsvariablen mit $\epsilon_i \sim \mathcal{N}(0,1)$, $T \subseteq \mathbb{R}^n$ eine beschränkte Teilmenge und $X(t) = \sum_{i=1}^n t_i \epsilon_i$, $t \in T$. Setze*

$$\sigma^2 = \sup_{t \in T} \|t\|_2^2 = \sup_{t \in T} \mathbb{E}\, X(t)^2$$

und sei Z

$$\textit{entweder} \quad \sup_{t \in T} X(t) \quad \textit{oder} \quad \sup_{t \in T} |X(t)|.$$

Dann gilt

$$\mathbb{P}(|Z - \mathbb{E}\, Z| \geq u) \leq 2 \exp\left(-\frac{u^2}{2C^2\sigma^2}\right) \qquad \forall u \geq 0$$

mit $C = \pi/2$.

2.2 Der Begriff der Entropie

2.12 Definition. Sei (T, d) ein (totalbeschränkter) pseudometrischer Raum. Für $t \in T$ und $\epsilon > 0$ sei $B(t, \epsilon) = \{s \in T : d(s,t) \leq \epsilon\}$ die abgeschlossene Kugel um t mit Radius ϵ. Eine ϵ-Überdeckung ist eine endliche Menge $\{t_1, \dots, t_N\}$ aus T mit $T \subseteq \bigcup_{j=1}^N B(t_j, \epsilon)$. Das kleinstmögliche solche N heißt ϵ-Überdeckungszahl (covering number) und wird mit $N(T, d, \epsilon)$ bezeichnet:

$$N(T,d,\epsilon) = \min\Big\{N : \text{ es existieren } t_1, \dots, t_N \in T \text{ mit } T \subseteq \bigcup_{j=1}^N B(t_j, \epsilon)\Big\}.$$

Der Logarithmus der Überdeckungszahl wird auch metrische Entropie genannt. Die ϵ-Packzahl $M(T, d, \epsilon)$ ist definiert durch

$$M(T,d,\epsilon) = \max\{M : \text{ es existieren } t_1, \dots, t_M \in T \text{ mit } \min_{i \neq j} d(t_i, t_j) > \epsilon\}.$$

Das folgende Resultat zeigt, dass die Überdeckungszahl und die Packzahl äquivalent sind.

2.13 Lemma. *Es gilt $N(T,d,\epsilon) \leq M(T,d,\epsilon) \leq N(T,d,\epsilon/2)$ für alle $\epsilon > 0$.*

Beweis. Für die erste Ungleichung sei $M = M(T,d,\epsilon)$ und $t_1, \dots, t_M$ mit $d(t_i, t_j) > \epsilon$ for all $i \neq j$. Ist nun $t \in T$ so gilt $d(t, t_j) \leq \epsilon$ für ein $j \leq M$ da wir sonst einen Widerspruch zur Maximalität von M erhalten. Es gilt also $T \subseteq \bigcup_{j=1}^M B(t_j, \epsilon)$ und somit $N(T,d,\epsilon) \leq M(T,d,\epsilon)$.

Für die zweite Ungleichung sei $\{t_1, \dots, t_N\}$ eine $\epsilon/2$-Überdeckung mit $N = N(T,d,\epsilon/2)$. Seien $t'_1, \dots, t'_M \in T$ Elemente mit $d(t'_i, t'_j) > \epsilon$ für alle $i' \neq j'$. Dann liegt jedes t'_j in einer Kugel $B(t_k, \epsilon/2)$ und zwei verschiedene Elemente können nicht in der gleichen Kugel liegen, wegen $d(t'_i, t'_j) > \epsilon$. Es folgt $M \leq N(T,d,\epsilon/2)$ und somit die Behauptung indem wir das Maximum nehmen. □

Wir listen einige weitere (einfache) Eigenschaften auf, die wir im Folgenden implizit verwenden werden. Ist d' eine weitere Pseudometrik mit $d' \geq d$, so gilt $N(T,d,\epsilon) \leq N(T,d',\epsilon)$ und $M(T,d,\epsilon) \leq M(T,d',\epsilon)$ für alle $\epsilon > 0$. Die Packzahl $M(T,d,\epsilon)$ hat oft den Vorteil, dass sie monoton in dem Sinn ist, dass $M(S,d,\epsilon) \leq M(T,d,\epsilon)$ für $S \subseteq T$ und alle $\epsilon > 0$. Des Weiteren, ist $S \subseteq T$ und existieren $t_1,\dots,t_N \in T$ mit $S \subseteq \bigcup_{j=1}^N B(t_j,\epsilon)$, so gilt $N(S,d,2\epsilon) \leq N$. Wird d induziert durch eine Halbnorm $\|\cdot\|$ auf dem Vektorraum T, so schreiben wir auch $N(T,\|\cdot\|,\epsilon) = N(T,d,\epsilon)$ und $M(T,\|\cdot\|,\epsilon) = M(T,d,\epsilon)$. Hier verwenden wir häufig die Formel $N(R\cdot S,\|\cdot\|,R\epsilon) = N(S,\|\cdot\|,\epsilon)$ für $S \subseteq T$ und alle $R,\epsilon > 0$.

Kugeln im $\mathbb{R}^d$

2.14 Lemma. *Sei $\|\cdot\|_2$ euklidische Norm im $\mathbb{R}^d$ und $B = B(0,1)$ abgeschlossene Einheitskugel. Dann gilt*

$$(1/\epsilon)^d \leq N(B,\|\cdot\|_2,\epsilon) \leq (3/\epsilon)^d \qquad \forall \epsilon \leq 1.$$

2.15 Bemerkung. Für $x \in \mathbb{R}^d$ und $R > 0$ folgt, dass $(R/\epsilon)^d \leq N(B(x,R),\|\cdot\|_2,\epsilon) \leq (3R/\epsilon)^d$ für alle $\epsilon \leq R$.

Beweis. Der Beweis beruht auf einem Volumenvergleichsargument welches auch in allgemeineren Situationen angewendet werden kann. Für die erste Ungleichung seien $t_1,\dots,t_N \in B$ mit $B \subseteq \bigcup_{j=1}^N B(t_j,\epsilon)$. Dann gilt

$$\operatorname{vol}(B) \leq \sum_{j=1}^N \operatorname{vol}(B(t_j,\epsilon)) = N\epsilon^d \operatorname{vol}(B)$$

und somit $N \geq (1/\epsilon)^d$. Für die zweite Ungleichung sei $M = M(B,\|\cdot\|_2,\epsilon)$ und $t_1,\dots,t_M$ mit $d(t_i,t_j) > \epsilon$ für alle $i \neq j$. Dann sind $B(t_j,\epsilon/2)$ disjunkt und es gilt $\bigcup_{j=1}^M B(t_j,\epsilon/2) \subseteq B(0,1+\epsilon/2)$. Wir erhalten also, dass

$$M(\epsilon/2)^d \operatorname{vol}(B) \leq (1+\epsilon/2)^d \operatorname{vol}(B)$$

d.h.

$$M \leq \left(\frac{1+\epsilon/2}{\epsilon/2}\right)^d = \left(\frac{2+\epsilon}{\epsilon}\right)^d \leq \left(\frac{3}{\epsilon}\right)^d \qquad \forall \epsilon \leq 1.$$

Die Behauptung folgt nun aus Lemma 2.13. □

2.16 Korollar. *Seien $(S,\mathscr{S},\mu)$ ein Maßraum, $f_1,\dots,f_d \in L^2(\mu)$ und $\mathcal{F} = \{f = \sum_{k=1}^d \theta_k f_k : \|f\|_{L^2(\mu)} \leq R\}$. Dann gilt $N(\mathcal{F},\|\cdot\|_{L^2(\mu)},\epsilon) \leq (3R/\epsilon)^d$ für alle $\epsilon \leq R$.*

Hölder-Kugeln

2.17 Lemma. *Sei* $\mathcal{F} = \{f : [0,1] \to \mathbb{R} : \|f\|_\infty, \|f'\|_\infty \leq R\}$ *mit* $R > 0$. *Dann gilt*

$$\log N(\mathcal{F}, \|\cdot\|_\infty, \epsilon) \leq C\frac{R}{\epsilon} \qquad \forall \epsilon > 0$$

mit einer absoluten Konstanten $C > 0$.

Beweis. Sei ohne Einschränkung $R = 1$. Betrachte $0 = a_0 < a_1 < \cdots < a_{N+1} = 1$ mit $a_k = k\epsilon$, $k = 0, \ldots, N$, und setze $I_1 = [a_0, a_1]$ und $I_k = (a_{k-1}, a_k]$, $k = 2, \ldots, N+1$. Für $f \in \mathcal{F}$ definiere

$$\bar{f} = \sum_{k=1}^{N+1} \epsilon[f(a_k)/\epsilon]\mathbf{1}_{I_k},$$

d.h. $\bar{f}$ ist konstant auf den Intervallen I_k und nimmt nur Vielfache von ϵ an. Es gilt $\|\bar{f} - f\|_\infty \leq 2\epsilon$ nach Konstruktion und Voraussetzung ($|\bar{f}(a_k) - f(a_k)| \leq \epsilon$ und $|f(x) - f(a_k)| \leq \epsilon$ für alle $x \in I_k$). Wir zählen nun wie viele verschiedene $\bar{f}$ es geben kann. Es gibt höchstens $[1/\epsilon]+1$ Möglichkeiten für $\bar{f}(a_1)$. Außerdem gilt

$$|\bar{f}(a_k) - \bar{f}(a_{k-1})| \leq |\bar{f}(a_k) - f(a_k)| + |f(a_k) - f(a_{k-1})| + |f(a_{k-1}) - \bar{f}(a_{k-1})| \leq 3\epsilon,$$

d.h. kennen wir $\bar{f}(a_{k-1})$, so gibt es höchstens 7 Möglichkeiten für $\bar{f}(a_k)$. Es gibt also höchstens $([1/\epsilon] + 1)7^N \leq (1/\epsilon + 1)7^{1/\epsilon}$ verschiedene $\bar{f}$. Es folgt

$$\log N(\mathcal{F}, \|\cdot\|_\infty, 4\epsilon) \leq (1/\epsilon)\log 7 + \log(1/\epsilon + 1) \leq (1 + \log 7)/\epsilon.$$

□

Mit etwas mehr Aufwand kann man sogar zeigen (siehe z.B. [12, Theorem 2.7.1])

2.18 Satz. *Für* $\alpha, R > 0$ *und* $l = \lfloor \alpha \rfloor$ *sei*

$$\mathcal{F} = \{f : [0,1] \to \mathbb{R} : \max_{k \leq l} \|f^{(k)}\|_\infty + \sup_{x \neq y} \frac{|f^{(l)}(x) - f^{(l)}(y)|}{|x - y|^{\alpha - l}} \leq R\}.$$

Dann gilt

$$\log N(\mathcal{F}, \|\cdot\|_\infty, \epsilon) \leq C\left(\frac{R}{\epsilon}\right)^{1/\alpha} \qquad \forall \epsilon > 0$$

mit einer Konstante $C > 0$ *die nur von* α *abhängt.*

Sobolev-Kugeln

2.19 Aufgabe. *Sei* $\mathcal{F} = \{f : [0,1] \to \mathbb{R} : \|f\|_{L^2}, \|f'\|_{L^2} \leq R\}$. *Dann gibt es eine absolute Konstante* $C > 0$ *mit*

$$\log N(\mathcal{F}, \|\cdot\|_{L^2}, \epsilon) \leq C\frac{R}{\epsilon}\log\left(C\frac{R}{\epsilon}\right) \qquad \forall \epsilon \leq R.$$

Es gilt sogar (siehe zum Beispiel [3, Korollar 4.3.38])

2.20 Satz. *Für* $\alpha \in \mathbb{N}$ *sei* $\mathcal{F} = \{f : [0,1] \to [0,1] : \|f^{(\alpha)}\|_{L^2} \leq R\}$. *Dann gilt*

$$\log N(\mathcal{F}, \|\cdot\|_{\infty}, \epsilon) \leq C\left(\frac{R}{\epsilon}\right)^{1/\alpha} \qquad \forall \epsilon > 0$$

mit einer Konstante $C > 0$ *die nur von* α *abhängt.*

Monotone Funktionen

2.21 Lemma. *Sei* $\mathcal{F} = \{f : \mathbb{R} \to [0,1] : f$ *monoton wachsend*$\}$. *Seien* $x_1, \ldots, x_n \in \mathbb{R}$ *und* $\|\cdot\|_{n,\infty}$ *die Halbnorm definiert durch* $\|f\|_{n,\infty} = \max_{i=1,\ldots,n} |f(x_i)|$. *Dann gilt*

$$\log N(\mathcal{F}, \|\cdot\|_{n,\infty}, \epsilon) \leq (1/\epsilon)\log(n + 1/\epsilon) \qquad \forall \epsilon > 0.$$

Beweis. Setze $\bar{f}(x_i) := \epsilon\lfloor f(x_i)/\varepsilon \rfloor$, $i = 1, \ldots, n$. Dann gilt $\|\bar{f}(x_i) - f(x_i)\|_{n,\infty} \leq \epsilon$. Wir zählen nun wie viele verschiedene $\bar{f}$ es geben kann. Sei hierfür ohne Einschränkung $x_1 \leq \cdots \leq x_n$. Es gilt

$$0 \leq \lfloor f(x_1)/\epsilon \rfloor \leq \cdots \leq \lfloor f(x_n)/\epsilon \rfloor \leq \lfloor 1/\epsilon \rfloor$$

und somit folgt, dass es

$$\binom{\lfloor 1/\epsilon \rfloor + n}{\lfloor 1/\epsilon \rfloor}$$

verschiedene $\bar{f}$ gibt. Die Aussage folgt nun aus

$$\binom{\lfloor 1/\epsilon \rfloor + n}{\lfloor 1/\epsilon \rfloor} \leqslant (\lfloor 1/\epsilon \rfloor + n)^{\lfloor 1/\epsilon \rfloor} \leqslant (1/\epsilon + n)^{1/\epsilon}.$$

□

Es gilt sogar (siehe zum Beispiel [12, Theorem 2.7.5]):

2.22 Satz. *Sei* $\mathcal{F} = \{f : \mathbb{R} \to [0,1] : f$ *monoton wachsend*$\}$, $p \in \mathbb{N}$ *und* Q *ein Wahrscheinlichkeitsmaß auf* $\mathbb{R}$. *Dann gilt*

$$\log N(\mathcal{F}, \|\cdot\|_{L^p(Q)}, \epsilon) \leq C/\epsilon \qquad \forall \epsilon > 0$$

mit einer Konstante $C > 0$ *die nur von* p *abhängt.*

2.3 Dudleys Entropieschranke

2.23 Definition. Sei (T,d) ein pseudometrischer Raum und $(X(t), t \in T)$ ein zentrierter stochastischer Prozess (eine Familie von Zufallsvariablen auf $(\Omega, \mathcal{F}, \mathbb{P})$ mit $\mathbb{E}\, X(t) = 0 \ \forall t \in T$). Dann heißt $(X(t), t \in T)$ sub-Gaußsch bezüglich d, falls die Zuwächse folgende Bedingung erfüllen:

$$\mathbb{E} \exp\left(\lambda(X(t) - X(s))\right) \leq \exp\left(\frac{\lambda^2 d(s,t)^2}{2}\right) \qquad \forall \lambda \in \mathbb{R},\ \forall s,t \in T$$

d.h. falls $X(t) - X(s) \in \mathrm{SG}(d(s,t)^2)$ für alle $s,t \in T$.

2.24 Beispiele.

(1) Sei $(X(t), t \in T)$ ein zentrierter Gaußprozess. Dann ist $(X(t), t \in T)$ sub-Gaußsch bezüglich $d(s,t) = (\mathbb{E}(X(t) - X(s))^2)^{1/2}$.

(2) Seien $\epsilon_1, \ldots, \epsilon_n$ unabhänging mit $\epsilon_i \in \mathrm{SG}(1)$, $T \subseteq \mathbb{R}^n$, $X(t) = \sum_{i=1}^n t_i \epsilon_i$, $t \in T$. Dann ist $(X(t), t \in T)$ sub-Gaußsch bezüglich des Euklidischen Abstandes, d.h. es gilt $X(t) - X(s) \in \mathrm{SG}(\|t-s\|_2^2)$ für alle $s,t \in T$.

Ziel dieses Kapitels ist es obere Schranken für $\sup_{t\in T} |X(t)|$ herzuleiten (für $(X(t), t \in T)$ sub-Gaußsch). Wir werden uns dabei auf den Erwartungswert

$$\mathbb{E} \sup_{t \in T} |X(t)| \tag{2.6}$$

konzentrieren. Allgemeinere Aussagen kann man erhalten indem man obere Schranken für den Erwartungswert mit den Konzentrationsresultaten aus Kapitel 2.1 kombiniert.

Wir betrachten zunächst den Fall, dass T endlich ist. Ein erster Versuch (2.6) abzuschätzen besteht darin wie im Beweis der Maximalungleichung in Korollar 1.48 vorzugehen: für $t_0 \in T$ beliebig gilt unter Verwendung von Aufgabe 2.2

$$\begin{aligned}
&\mathbb{P}(\max_{t\in T} |X(t) - X(t_0)| \geq u) \\
&\leq \sum_{t \in T} \mathbb{P}(|X(t) - X(t_0)| \geq u) \leq \sum_{t \in T} 2 \exp\left(-\frac{u^2}{2d(t_0,t)^2}\right) \leq 2|T| \exp\left(-\frac{u^2}{2R^2}\right)
\end{aligned}$$

mit $R = \max_{t\in T} d(t_0, t)$. Die Dreiecksungleichung und Lemma 1.49 (b) liefern nun

$$\begin{aligned}
\mathbb{E} \max_{t\in T} |X(t)| &\leq \mathbb{E}\, |X(t_0)| + \mathbb{E} \max_{t \in T} |X(t) - X(t_0)| \\
&\leq \mathbb{E}\, |X(t_0)| + CR\sqrt{\log 2|T|}
\end{aligned}$$

mit absoluter Konstante $C > 0$. Die Ungleichung gilt mit $C = \sqrt{2}$, was aus folgender Aufgabe folgt:

2.25 Aufgabe. *Seien $X_1, \dots, X_N$ Zufallsvariablen mit $X_j \in \mathrm{SG}(\sigma_j^2)$. Dann gilt*

$$\mathbb{E} \max_{j \le N} X_j \le \sqrt{2 \log N} \max_{j \le N} \sigma_j, \qquad \mathbb{E} \max_{j \le N} |X_j| \le \sqrt{2 \log 2N} \max_{j \le N} \sigma_j.$$

Die zweite Ungleichung in 2.25 liefert gute Resultate falls die Zufallsvariablen X_j annähernd unkorreliert sind (der Faktor $\sqrt{\log 2N}$ kann nicht verbessert werden falls X_j unabhängig und standardnormalverteilt sind). Allerdings wird die Ungleichung sehr schlecht falls viele Zufallsvariablen fast gleich sind.

Wir wollen nun eine obere Schranke herleiten die frei ist von der Kardinalität von T. Wir verwenden hierfür eine Methode die Chaining genannt wird. Sei zunächst $T_1 \subseteq T$ eine Teilmenge und $\pi_1 : T \to T_1$ eine Abbildung ($\pi_1(t)$ ist eine erste Approximation von t). Dann gilt

$$X(t) - X(t_0) = X(t) - X(\pi_1(t)) + X(\pi_1(t)) - X(t_0).$$

Die Idee ist nun T_1, π_1 so zu wählen, dass die Familie $(X(\pi_1(t)) - X(t_0), t \in T)$ nicht zu groß ist (Zufallsvariablen, die nah beieinander sind werden gleich gesetzt) und gleichzeitig die Zufallsvariablen $X(t) - X(\pi_1(t))$ kleiner sind als die ursprünglichen Zufallsvariablen $X(t) - X(t_0)$ (z.B. mit kleinerem sub-Gaußschen Parameter). Unsere allgemeine Strategie besteht nun darin diesen Prozess zu iterieren. Für $j = 0, \dots, J$ seien $T_j \subseteq T$ Teilmengen und $\pi_J : T \to T_j$ Abbildungen mit $T_0 = \{t_0\}$ und $T_J = T$, $\pi_J = \mathrm{id}$. Dann gilt

$$X(t) - X(t_0) = \sum_{j=1}^{J} X(\pi_j(t)) - X(\pi_{j-1}(t))$$

und somit

$$\mathbb{E} \max_{t \in T} |X(t) - X(t_0)| \le \sum_{j=1}^{J} \mathbb{E} \max_{t \in T} |X(\pi_j(t)) - X(\pi_{j-1}(t))|.$$

2.26 Satz (Dudleys Entropieschranke)**.** *Sei (T, d) ein endlicher pseudometrischer Raum und $(X(t), t \in T)$ ein zentrierter stochastischer Prozess welcher sub-Gaußsch bezüglich d ist. Dann gilt für alle $t_0 \in T$*

$$\mathbb{E} \max_{t \in T} |X(t)| \le \mathbb{E} |X(t_0)| + 12 \int_0^{R/2} \sqrt{\log 2N(T, d, \epsilon)}\, d\epsilon$$

mit $R = \max_{t \in T} d(t_0, t)$. Die Ungleichung gilt auch mit $N(T, d, \epsilon)$ ersetzt durch $M(T, d, \epsilon)$.

Beweis. Wir verwenden Chaining und benutzen bei der Konstruktion von T_j und π_j, dass obige Heuristik bzw. Approximation mit Hilfe von d umgesetzt werden kann. Sei $T_0 = \{t_0\}$, T_j δ_j-Überdeckung mit $|T_j| = N(T, d, \delta_j)$ und

$\delta_j = R2^{-j}$ und sei $\pi_j : T \to T_j$ Abbildung mit $d(t, \pi_j(t)) \leq \delta_j$ für alle $t \in T$, $j = 0, \ldots, J$. Hier wählen wir J so groß, dass $T_J = T$ gilt. Betrachte nun

$$\mathbb{E} \max_{t \in T} |X(\pi_j(t)) - X(\pi_{j-1}(t))|.$$

Dann ist $X(\pi_j(t)) - X(\pi_{j-1}(t))$ sub-Gaußsch mit Parameter

$$d(\pi_j(t), \pi_{j-1}(t))^2 \leq (d(\pi_j(t), t) + d(t, \pi_{j-1}(t)))^2 \leq (\delta_j + \delta_{j-1})^2 \leq 9\delta_j^2$$

und es gibt höchstens

$$|\{(\pi_j(t), \pi_{j-1}(t)) : t \in T\}| \leq |T_j| \cdot |T_{j-1}| \leq N(T, d, \delta_j)^2$$

verschiedene Zufallsvariablen im obigen Maximum. Aufgabe 2.25 liefert

$$\begin{aligned} \mathbb{E} \max_{t \in T} |X(\pi_j(t)) - X(\pi_{j-1}(t))| &\leq 3\delta_j \sqrt{2 \log 2N(T, d, \delta_j)^2} \\ &\leq 6\delta_j \sqrt{\log 2N(T, d, \delta_j)}. \end{aligned}$$

Wir schließen

$$\begin{aligned} \mathbb{E} \max_{t \in T} |X(t) - X(t_0)| &\leq \sum_{j=1}^{J} \mathbb{E} \max_{t \in T} |X(\pi_j(t)) - X(\pi_{j-1}(t))| \\ &\leq \sum_{j=1}^{J} 6R2^{-j} \sqrt{\log 2N(T, d, R2^{-j})} \\ &= \sum_{j=1}^{J} 12R(2^{-j} - 2^{-j-1}) \sqrt{\log 2N(T, d, R2^{-j})} \\ &\leq 12 \int_0^{R/2} \sqrt{\log 2N(T, d, \epsilon)}\, d\epsilon. \end{aligned}$$

Die Behauptung folgt nun indem wir diese Ungleichung in $\mathbb{E} \max_{t \in T} |X(t)| \leq \mathbb{E} |X(t_0)| + \mathbb{E} \max_{t \in T} |X(t) - X(t_0)|$ einsetzen. □

2.27 Bemerkung. Sind wir an oberen Schranken für $\mathbb{E} \max_{t \in T} X(t)$ interessiert, so können wir die Ungleichung $\mathbb{E} \max_{t \in T} X(t) \leq \mathbb{E} \max_{t \in T} |X(t)|$ mit Dudleyentropieschranke kombinieren. Wir können allerdings auch obigen Beweis modifizieren indem wir die Beträge weglassen und den ersten Teil von Aufgabe 2.25 anwenden. Wir erhalten dann folgende leichte Verbesserung:

$$\mathbb{E} \max_{t \in T} X(t) = \mathbb{E} \max_{t \in T} (X(t) - X(t_0)) \leq 12 \int_0^{R/2} \sqrt{\log N(T, d, \epsilon)}\, d\epsilon.$$

2.28 Korollar. *Sei (T, d) ein totalbeschränkter pseudometrischer Raum und $(X(t), t \in T)$ ein zentrierter stochastischer Prozess welcher sub-Gaußsch*

bezüglich d ist. Wir nehmen an, dass es eine dichte, abzählbare Teilmenge $T_0 \subseteq T$ gibt mit $\sup_{t\in T} |X(t)| = \sup_{t\in T_0} |X(t)|$. Dann gilt für alle $t_0 \in T$

$$\mathbb{E} \sup_{t\in T} |X(t)| \leq \mathbb{E}\,|X(t_0)| + 12 \int_0^{R/2} \sqrt{\log 2M(T,d,\epsilon)}\, d\epsilon$$

mit $R = \max_{t\in T} d(t_0, t)$.

Beweis. Aus Satz 2.26 erhalten wir für alle endlichen Teilmengen $S \subseteq T$ mit $t_0 \in S$

$$\mathbb{E} \max_{t\in S} |X(t)| \leq \mathbb{E}\,|X(t_0)| + 12 \int_0^{R/2} \sqrt{\log 2M(T,d,\epsilon)}\, d\epsilon$$

wobei wir auch die Monotonie $M(S,d,\epsilon) \leq M(T,d,\epsilon)$ verwendet haben. Die Behauptung folgt nun durch Anwendung des Satzes von der monotonen Konvergenz ($S \nearrow T_0$) und der Definition von T_0. □

2.4 Glivenko-Cantelli-Sätze

In diesem Kapitel betrachten wir eine Anwendung von Dudleys Entropieschranke auf empirische Prozesse. Seien hierfür $X_1, \ldots, X_n$ i.i.d. Zufallsvariablen mit Werten in $(S, \mathscr{S})$ und Verteilung P, $P_n = (1/n)\sum_{i=1}^n \delta_{X_i}$ die empirische Verteilung und $\mathcal{F}$ eine (abzählbare) Menge von P-integrierbaren Funktionen $f : S \to \mathbb{R}$. Für $f \in \mathcal{F}$ schreiben wir $P(f) = \int f dP$ und $P_n(f) = \int f dP_n = (1/n)\sum_{i=1}^n f(X_i)$. Gilt nun $\sup_{f\in\mathcal{F}} |P_n(f) - P(f)| \to 0$ stochastisch oder fast sicher (d.h. gilt eine gleichmäßige Version des Gesetzes der großen Zahlen), so heißt $\mathcal{F}$ P-Glivenko-Cantelli-Klasse.

2.29 Lemma (Symmetrisierungstrick)**.** *Seien $X_1, \ldots, X_n$ und $\mathcal{F}$ wie oben beschrieben und $\epsilon_1, \ldots, \epsilon_n$ unabhängige Rademacher Zufallsvariablen unabhängig von $X_1, \ldots, X_n$. Dann gilt*

$$\mathbb{E} \sup_{f\in\mathcal{F}} |P_n(f) - P(f)| \leq 2\,\mathbb{E} \sup_{f\in\mathcal{F}} \Big|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\Big|.$$

Beweis. Sei $(X_1', \ldots, X_n')$ eine unabhängige Kopie von $(X_1, \ldots, X_n)$ (d.h. $X_1, \ldots, X_n, X_1', \ldots, X_n'$ i.i.d., man spricht von einer Phantom-Stichprobe (ghost sample)). Dann sind die Zufallsvariablen $f(X_i) - f(X_i')$ unabhängig, symmetrisch und haben daher die gleiche Verteilung wie $\epsilon_i(f(X_i) - f(X_i'))$.

Es folgt

$$\begin{aligned}
\mathbb{E}\sup_{f\in\mathcal{F}}|P_n(f)-P(f)| &= \mathbb{E}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n f(X_i)-\mathbb{E}\,f(X_i')\Big| \\
&\le \mathbb{E}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n f(X_i)-f(X_i')\Big| \\
&= \mathbb{E}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n \epsilon_i(f(X_i)-f(X_i'))\Big| \\
&\le \mathbb{E}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\Big| + \mathbb{E}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i')\Big| \\
&= 2\,\mathbb{E}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\Big|,
\end{aligned}$$

wobei die erste Ungleichung mit Hilfe der Jensenschen Ungleichung folgt (alternativ kann man auch verwenden, dass für $a\in\mathbb{R}$ und $Z_f=(1/n)\sum_{i=1}^n f(X_i')$ folgendes gilt: $\sup_f|a-\mathbb{E}\,Z_f|\le\sup_f\mathbb{E}\,|a-Z_f|\le\mathbb{E}\sup_f|a-Z_f|$). □

Wollen wir Dudleys Entropieschranke anwenden, so schreiben wir

$$\mathbb{E}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\Big| = \mathbf{E}_X\mathbf{E}_\epsilon\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\Big|,$$

wobei $\mathbf{E}_\epsilon$ der Erwartungswert bezüglich $\epsilon_1,\dots,\epsilon_n$ mit $X_1,\dots,X_n$ festgehalten und $\mathbf{E}_X$ Erwartungswert bezüglich $X_1,\dots,X_n$ ist. Es gilt nun:

2.30 Satz. *Seien $\epsilon_1,\dots,\epsilon_n$ unabhängige, zentrierte Zufallsvariablen mit Werten in $[-1,1]$ und $\mathcal{F}$ eine Menge von Funktionen $f:S\to\mathbb{R}$. Außerdem seien $x_1,\dots,x_n\in S$ und $\|\cdot\|_{n,2}$ die Halbnorm definiert durch $\|f\|_{n,2}^2=(1/n)\sum_{i=1}^n f^2(x_i)$. Es gelte $R=\sup_{f\in\mathcal{F}}\|f\|_{n,2}<\infty$ und $f\equiv 0\in\mathcal{F}$. Dann gilt für alle $\delta>0$*

$$\mathbb{E}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\Big| \le 4\delta+\frac{12}{\sqrt{n}}\int_\delta^{R/2}\sqrt{\log 2N(\mathcal{F},\|\cdot\|_{n,2},u)}\,du.$$

2.31 Bemerkung. (1) Der Fall $\delta=0$ und N ersetzt durch M folgt aus Korollar 2.28 da $(1/\sqrt{n})\sum_{i=1}^n\epsilon_i(f(x_i)-g(x_i))\in\mathrm{SG}(\|f-g\|_{n,2}^2)$ für alle $f,g\in\mathcal{F}$.

(2) Die Skalierung der Euklidischen Norm führt zu $\|f\|_{n,2}\le\|f\|_\infty$. Daher können wir $\log N(\mathcal{F},\|\cdot\|_{n,2},u)\le\log N(\mathcal{F},\|\cdot\|_\infty,u)$ einsetzen und für den letzten Term die Entropieschranken aus Kapitel 2.2 verwenden.

Beweis. Wir fixieren ein $\delta>0$. Der Beweis ist analog zum Beweis von Dudleys Entropieschranke, wir wenden jedoch eine triviale Schranke auf

den letzten Term im Chaining an. Wähle $\mathcal{F}_0 = \{0\}$, $\pi_0 : \mathcal{F} \to \mathcal{F}_0$, $\mathcal{F}_j$ δ_j-Überdeckung mit $|\mathcal{F}_j| = N(\mathcal{F}, \|\cdot\|_{n,2}, \delta_j)$, $\delta_j = R2^{-j}$ und $\pi_j : \mathcal{F} \to \mathcal{F}_j$ Abbildung mit $d(f, \pi_j(f)) \leq \delta_j$ für alle $f \in \mathcal{F}$, $j = 1, \ldots, J$, wobei J später gewählt wird. Wir schreibe $\langle \epsilon, f \rangle_n = (1/n) \sum_{i=1}^n \epsilon_i f(x_i)$. Dann gilt

$$\begin{aligned}\mathbb{E} \sup_{f \in \mathcal{F}} |\langle \epsilon, f \rangle_n| &\leq \mathbb{E} \sup_{f \in \mathcal{F}} |\langle \epsilon, f - \pi_J(f) \rangle_n| + \sum_{j=1}^J \mathbb{E} \sup_{f \in \mathcal{F}} |\langle \epsilon, \pi_j(f) - \pi_{j-1}(f) \rangle_n| \\ &\leq \delta_J + \frac{6}{\sqrt{n}} \sum_{j=1}^J \delta_j \sqrt{2N(\mathcal{F}, \|\cdot\|_{n,2}, \delta_j)},\end{aligned}$$

wobei die Ungleichung $\sup_{f \in \mathcal{F}} |\langle \epsilon, f - \pi_J(f) \rangle_n| \leq \delta_J$ aus der Cauchy-Schwarz-Ungleichung folgt und wir die restlichen Terme wie im Beweis von Satz 2.26 abgeschätzt haben. Sei nun J so gewählt, dass $R2^{-J-2} \leq \delta < R2^{-J-1}$. Dann erhalten wir

$$\begin{aligned}\mathbb{E} \sup_{f \in \mathcal{F}} |\langle \epsilon, f \rangle_n| &\leq \delta_J + \frac{12}{\sqrt{n}} \sum_{j=1}^J R(2^{-j} - 2^{-j-1}) \sqrt{2N(\mathcal{F}, \|\cdot\|_{n,2}, R2^{-j})} \\ &\leq 4\delta + \frac{12}{\sqrt{n}} \int_\delta^{R/2} \sqrt{\log 2N(\mathcal{F}, \|\cdot\|_{n,2}, u)}\, du\end{aligned}$$

Die Behauptung folgt nun da δ beliebig war. □

2.32 Satz. *Sei $\mathcal{F}$ eine (abzählbare) Menge von Funktionen $f : S \to \mathbb{R}$ mit*

(a) $|f(x)| \leq b$ für alle $x \in S$ und alle $f \in \mathcal{F}$,

(b) $\frac{1}{n} \log M(\mathcal{F}, \|\cdot\|_{L^2(P_n)}, \delta) \xrightarrow{\mathbb{P}} 0$ für alle $\delta > 0$.

Dann ist $\mathcal{F}$ P-Glivenko-Cantelli, d.h. es gilt $\sup_{f \in \mathcal{F}} |P_n(f) - P(f)| \to 0$ in L^1 und stochatisch.

Die Bedingungen können noch weiter abgeschwächt werden. Ein analoger Satz gilt in dem Fall, dass $P(F) < \infty$ mit $F(x) = \sup_{f \in F} |f(x)|$, $x \in S$ (siehe zum Beispiel [10, Theorem 3.7]). Des Weiteren kann man mit Hilfe von Martingalargumenten zeigen, dass unter dieser Bedingung $\sup_{f \in \mathcal{F}} |P_n(f) - P(f)|$ immer fast sicher gegen eine Konstante konvergiert (siehe zum Beispiel [3, Proposition 3.7.8]). Daher impliziert die stochastische Konvergenz in Satz 2.32 auch die fast sichere Konvergenz gegen 0.

Beweis. $M(\mathcal{F}, \|\cdot\|_{L^2(P_n)}, \delta)$ ist in der Tat eine Zufallsvariable, da

$$M(\mathcal{F}, \|\cdot\|_{L^2(P_n)}, \delta) \geq m \;\Leftrightarrow \sup_{f_1, \ldots, f_m \in \mathcal{F}} \min_{i \neq j} P_n(f_i - f_j)^2 > \epsilon.$$

Mit Hilfe von Lemma 2.29 gilt

$$\mathbb{E} \sup_{f \in \mathcal{F}} |P_n(f) - P(f)| \leq \mathbf{E}_X \mathbf{E}_\epsilon \sup_{f \in \mathcal{F}} \Big| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) \Big|$$

Aus Satz 2.30 folgt außerdem, dass

$$\mathbf{E}_\epsilon \sup_{f\in\mathcal{F}} \Big|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\Big| \le 4\delta + \frac{12}{\sqrt{n}} \cdot \frac{b}{2} \cdot \sqrt{\log 2M(\mathcal{F}, \|\cdot\|_{n,2}, \delta)}$$

Unter Verwendung von Voraussetzung (b) folgt, dass

$$\mathbf{E}_\epsilon \sup_{f\in\mathcal{F}} \Big|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\Big| \le 4\delta \xrightarrow{\mathbb{P}} 0.$$

Außerdem ist die linke Seite beschränkt durch b. Der Satz von der dominierten Konvergenz liefert also (die Tatsche, dass stochastische Konvergenz ausreicht folgt aus einem Teilteilfolgenargument, siehe zum Beispiel [5, Lemma 3.11])

$$\mathbf{E}_X \mathbf{E}_\epsilon \sup_{f\in\mathcal{F}} \Big|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\Big| \to 0.$$

□

2.33 Beispiel. Sei $S = \mathbb{R}$ und $\mathcal{F} = \{\mathbf{1}_{(-\infty,q]} : q \in \mathbb{Q}\}$. Dann gilt $P_n(\mathbf{1}_{(-\infty,q]}) - P(\mathbf{1}_{(-\infty,q]}) = F_n(q) - F(q)$ mit den entsprechenden Verteilungsfunktionen. Wir überprüfen die Bedingungen aus Satz 2.32. Die Bedingung (a) ist klar, (b) folgt aus

$$\log M(\mathcal{F}, \|\cdot\|_{L^2(P_n)}, \delta) \le \log N(\mathcal{F}, \|\cdot\|_{L^\infty(P_n)}, \delta/2) \le \log n.$$

Wir erhalten also, dass $\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = \sup_{q\in\mathbb{Q}} |F_n(q) - F(q)| \to 0$ in L^1 und stochastisch. Analog kann man zeigen, dass alle Funtionenklassen aus Kapitel 2.2 Glivenko-Cantelli sind.

3 Nichtparametrische Regression

3.1 Modell

Regressionsmodelle sind die am häufigsten in Anwendungen vorkommenden statistischen Modellierungen. Man unterscheidet zwischen Modellen mit deterministischem und mit zufälligem Design. Wir betrachten Standardformulierungen für diese Modelle, es existieren vielfältige Verallgemeinerungen und Modifikationen.

Deterministisches Design

Wir beobachten $(x_1, Y_1), \dots, (x_n, Y_n)$ mit

$$Y_i = f_0(x_i) + \epsilon_i, \qquad i = 1, \dots, n,$$

wobei $x_1, \ldots, x_n \in S$ deterministisch (in der Regel $S \subseteq \mathbb{R}^d$), $f_0 : S \to \mathbb{R}$ eine unbekannte Funktion und $\epsilon_1, \ldots, \epsilon_n$ unabhängige Zufallsvariablen mit $\mathbb{E}\,\epsilon_i = 0$. Ziel ist es f_0 zu schätzen.

Das prototypische Regressionsmodell mit deterministischem Design ist gegeben durch äquidistante Beobachtungen auf dem Einheitsintervall und normalverteilte Fehler, d.h. $x_i = i/n$ und $\varepsilon_i \sim N(0, \sigma^2)$.

Zufälliges Design

Sei (X, Y) ein Paar von Zufallsvariablen wobei X Werte in einem messbaren Raum $(S, \mathscr{S})$ annimmt und Y reell ist mit $\mathbb{E}\,Y^2 < \infty$. Wir nehmen an, dass wir n unabhängige Kopien $(X_1, Y_1), \ldots, (X_n, Y_n)$ von (X, Y) beobachten (d.h. $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d.) und unser Ziel ist es die bedingte Erwartung

$$f_0(x) = \mathbb{E}(Y|X = x)$$

zu schätzen. Dabei interpretieren wir f_0 als beste Vorhersage von Y basierend auf X. Zur Erinnerung: Ist P^X die Verteilung von X ist, so gilt

$$\mathbb{E}(Y - f_0(X))^2 = \min_{f \in L^2(P^X)} \mathbb{E}(Y - f(X))^2.$$

Schreiben wir $\epsilon_i = Y_i - f_0(X_i)$, so kann das Modell geschrieben werden als

$$Y_i = f_0(X_i) + \epsilon_i, \qquad i = 1, \ldots, n$$

mit $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. und $E(\epsilon_i|X_i) = 0$ und $\sigma^2 = \mathrm{Var}(\epsilon_i) < \infty$. Dies führt zu einer zweiten Interpretation des Regressionsmodells mit zufälligen Design welches die Ähnlichkeit zu dem Fall des deterministischen Designs herstellt.

In beiden Fällen (deterministisch oder zufällig) heißt Y Antwortvariable (response variable), X Kovariable oder auch erklärende Variable (covariate, explanatory variable, feature), f_0 Regressionsfunktion und $\epsilon_1, \ldots, \epsilon_n$ Fehlervariablen.

Verlustfunktionen

Im Prinzip können wir die gleichen Verlustfunktionen wie im Fall der Dichteschätzung betrachten. Im Fall der Regressionsschätzung gibt es jedoch zwei weitere, sehr natürliche Verlustfunktionen. Im Fall des deterministischen Designs betrachten wir daher vor allem

$$\|\hat{f}_n - f_0\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} (\hat{f}_n(x_i) - f_0(x_i))^2,$$

im Fall des zufälligen Designs (Vorhersagefehler)

$$\|\hat{f}_n - f_0\|_{L^2(P^X)}^2 = \int (\hat{f}_n(x) - f_0(x))^2 P^X(dx).$$

3.2 Schätzmethoden

In diesem Kapitel geben wir einen (kurzen) Überblick über verschiedene Methoden einen Schätzer der Regressionsfunktion zu konstruieren. Wir werden sehen, dass sich einige Ideen zur Dichteschätzung auf die Schätzung der Regressionsfunktion übertragen lassen. Wir betrachten zuerst den Fall des zufälligen Designs. Wir nehmen an, dass X eine Zufallsvariable mit Werten in $\mathbb{R}$ ist und dass (X, Y) eine Dichte $f^{X,Y}$ bezüglich des Lebesgue-Maßes besitzt. Dann ist die Regressionsfunktion gegeben durch

$$f_0(x) = \mathbb{E}(Y|X = x) = \int y \frac{f^{X,Y}(x, y)}{f^X(x)}\, dy.$$

Betrachte nun für einen symmetrischen Kern K die Schätzer

$$\hat{f}_n^X(x) = \frac{1}{n} \sum_{i=1}^n K_h(x - X_i)$$

und

$$\hat{f}_n^{X,Y}(x, y) = \frac{1}{n} \sum_{i=1}^n K_h(x - X_i) K_{h'}(y - Y_i)$$

mit Bandweiten $h, h' > 0$. Dann ergibt Einsetzen (plug-in-Ansatz)

$$\hat{f}_n(x) = \int y \frac{\hat{f}_n^{X,Y}(x, y)}{\hat{f}_n^X(x)}\, dy = \frac{\frac{1}{n}\sum_{i=1}^n Y_i K_h(x - X_i)}{\frac{1}{n}\sum_{i=1}^n K_h(x - X_i)} = \sum_{i=1}^n Y_i w_i(x),$$

mit Gewichten $w_i(x) = K_h(x - X_i)/(\sum_{i=1}^n K_h(x - X_i))$, sofern $\hat{f}_n^X(x) > 0$. Dabei gilt die zweite Gleichheit wegen der Symmetrie von K:

$$\int y \hat{f}_n^{X,Y}(x, y)\, dy = \frac{1}{n} \sum_{i=1}^n K_h(x - X_i) \int y K_{h'}(y - Y_i)\, dy = \frac{1}{n} \sum_{i=1}^n K_h(x - X_i) Y_i.$$

Die Intuition hinter diesem Schätzer ist, dass er lokal diejenigen Y_i mittelt, deren Kovariablen X_i nahe bei x liegen. Welche X_i dafür in Frage kommen hängt von den Gewichten und damit vom Kern ab. Dies ergibt auch im Fall eines deterministischen Designs einen sinnvollen und sehr gebräuchlichen Kernschätzer, nämlich den **Nadaraya-Watson-Schätzer**. Dieser ist wie oben definiert durch $\hat{f}_n(x) = n^{-1} \sum_{i=1}^n Y_i w_i(x)$, aber für einen allgemeinen Kern K. Er ist wohldefiniert für alle x mit $\sum_{i=1}^n K_h(x - X_i) \neq 0$. Eine elementare Rechnung zeigt, dass der Schätzer im Fall nicht-negativer Kernfunktionen die folgende gewichtete Summe von Quadraten minimiert:

$$\hat{f}_n(x) \in \operatorname*{argmin}_{\beta_0 \in \mathbb{R}} \sum_{i=1}^n (Y_i - \beta_0)^2 K_h(x - X_i).$$

Also ist der Nadaraya-Watson-Schätzer eine lokale Approximation durch eine Konstante, die durch ein lokales Kleinste-Quadrate-Kriterium definiert wird.

Eine Verallgemeinerung dieser Konstruktion ergibt sich durch Ersetzen der lokalen Konstante durch ein lokales Polynom. Dies erlaubt einen besseren Fit bei glatten Regressionsfunktionen f_0. Betrachte hierfür das lokale Kleinste-Quadrate-Kriterium

$$\hat{\beta}(x) \in \operatorname*{argmin}_{\beta \in \mathbb{R}^{m+1}} \sum_{i=1}^{n} \Big(Y_i - \sum_{j=0}^{m} \frac{\beta_j}{j!} (X_i - x)^j \Big)^2 K_h(x - X_i).$$

Sind $\hat{\beta}_0(x), \ldots, \hat{\beta}_m(x)$ die Komponenten von $\hat{\beta}(x)$, so ist der **lokal-polynomiale Schätzer** vom Grad m definiert durch

$$\hat{f}_n(x) = \hat{\beta}_0(x).$$

Die anderen Komponenten liefern Schätzer für die entprechenden Ableitungen ($\hat{\beta}_j(x)$ ist Schätzer von $f^{(j)}(x)$). Im Fall $m = 0$ ist dies der Nadaraya-Watson-Schätzer, im Fall $m = 1$ spricht man von einem lokal-linearen Schätzer.

Die bisherigen Schätzmethoden sind lokal. Der Nadaraya-Watson-Schätzer ist eine lokale Mittelung, der lokal-polynomiale Schätzer eine lokale Modellierung. Wir wollen nun einige globale Methoden vorstellen. Da die Regressionsfunktion $f_0(x) = \mathbb{E}(Y|X = x)$ den Ausdruck $\mathbb{E}(Y - f(X))^2$ über alle messbaren Funktionen f mit $f(X) \in L^2$ minimiert, ist die Idee nun das Risiko $\mathbb{E}(Y - f(X))^2$ durch das empirische Risiko $n^{-1} \sum_{i=1}^{n} (Y_i - f(X_i))^2$ zu ersetzen und dieses dann über eine geeignete Menge von Funktionen zu minimieren. Dabei dürfen wir nicht über alle Funktionen minimieren, da die beste Lösung $f^*(X_i) := Y_i$ nur die Beobachtungen interpoliert, ohne die Struktur von f_0 aufzudecken.

Sei nun $\mathcal{F}$ eine Menge von Funktionen. Dann ist der **Kleinste-Quadrate-Schätzer** definiert durch

$$\hat{f}_n \in \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2, \tag{3.1}$$

d.h.

$$\hat{f}_n \in \mathcal{F} \quad \text{und} \quad \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{f}_n(X_i))^2 = \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2.$$

In den meisten Fällen existiert tatsächlich ein Minumum. Im Allgemeinen ist es jedoch nicht eindeutig. Ist die Existenz eines Minimums nicht gewährleistet, so kann man auch eine approximative Lösung suchen (dies kann aus numerischer Sicht sogar sinnvoll sein, wenn ein Minimum existiert). Man sucht dann beispielsweise für ein $\rho \geq 0$ (z.B. $\rho = n^{-2}$) eine Funktion $\hat{f}_n \in \mathcal{F}$ mit

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{f}_n(X_i))^2 \leq \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \rho.$$

Wir unterscheiden zwischen zwei Klassen von Kleinste-Quadrate-Schätzern. Im ersten Fall ist $\mathcal{F} = \text{span}(\phi_1, \dots, \phi_{d_n})$ ein linearer Raum, der in der Regel mit der Anzahl der Beobachtungen wächst (*method of sieves*). Das kann zum Beispiel der Raum aller trigonometrischen Polynome vom Grad kleiner gleich d_n sein, oder der Raum aller stückweisen Polynome aus Kapitel 1. Letztere Wahl führt zu einem dem lokal-polynomialen Schätzer ähnlichen Schätzer. Ist $\mathcal{F}$ ein linearer Raum, so besitzt das Minimierungsproblem (3.1) immer eine Lösung. Diese kann folgendermaßen berechnet werden. Sei $\mathbf{Y} = (Y_1, \dots, Y_n)^T$, $\mathbf{X} = (\phi_k(X_i))_{1 \le i \le n, 1 \le k \le d_n}$. Dann kann (3.1) geschrieben werden als $\hat{f}_n = \sum_{k=1}^{d_n} \hat{\theta}_k \phi_k$ mit

$$\hat{\boldsymbol{\theta}} \in \text{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^{d_n}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\theta}\|_2^2.$$

Elementare Rechnungen liefern, dass $\hat{\boldsymbol{\theta}}$ genau dann eine Lösung ist, wenn $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\theta}} = \mathbf{X}^T\mathbf{Y}$ gilt. Dass diese Gleichung eine Lösung besitzt kann zum Beispiel mit Hilfe der Singulärwertzerlegung gesehen werden. Diese besagt, dass $\mathbf{X}$ geschrieben werden kann als $\mathbf{X} = \sum_{j=1}^{r} \sigma_j u_j v_j^T$ mit $r = \text{rank}(\mathbf{X})$, $\sigma_1 \ge \cdots \ge \sigma_r > 0$ und Orthonormalsystemen $u_1, \dots, u_r$ und $v_1, \dots, v_r$ in $\mathbb{R}^n$ und $\mathbb{R}^{d_n}$. Eine Lösung ist nun gegeben durch $\hat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X})^+\mathbf{X}^T\mathbf{Y}$ mit $(\mathbf{X}^T\mathbf{X})^+ = \sum_{j=1}^{r} \sigma_j^{-2} v_j v_j^T$. Man kann zeigen, dass dies die Lösung mit minimaler Euklidischer Norm ist. Aus der Definition (3.1) ist einfach zu sehen, dass im Fall linearer Räume für $\hat{\mathbf{f}}_n = (\hat{f}_n(X_1), \dots, \hat{f}_n(X_n))$, $\hat{\mathbf{f}}_n = \boldsymbol{\Pi}\mathbf{Y}$ gilt, wobei $\boldsymbol{\Pi}$ die Orthogonalprojektion von $\mathbb{R}^n$ auf den von den Spalten von $\mathbf{X}$ aufgespannten linearen Raum ist. Daher sprechen wir in diesem Fall auch vom **Projektionsschätzer**.

Im zweiten Fall betrachtet man einen nichtlinearen Raum $\mathcal{F}$ welcher in der Regel nicht von n abhängt und für den angenommen wird, dass er f_0 enthält. Beispiele sind $\{f : \mathbb{R} \to \mathbb{R} : f \text{ monoton wachsend}\}$ oder $\{f : [0,1] \to \mathbb{R} : \int_0^1 (f^{(m)}(x))^2 \, dx \le L^2\}$ mit $m \in \mathbb{N}$. Den entsprechenden Schätzer in (3.1) werden wir in den nächsten beiden Kapiteln mit Hilfe der Entropiemethoden aus Kapitel 2 analysieren.

Wenn wir einen zusätzlichen Strafterm (*penalty*) einführen anstatt eine eingeschränkte Parametermenge zu betrachten, dann erhalten wir den **penalisierten Kleinste-Quadrate-Schätzer**

$$\hat{f}_n \in \text{argmin}_f \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda \, \text{Pen}(f),$$

wobei das Minimum über alle messbaren Funktionen genommen wird für die $\text{Pen}(f) \ge 0$ definiert ist. Eine oft benutzte Wahl (im Fall, dass $x_i \in [0,1]$) ist gegeben durch die *roughness penalty*

$$\text{Pen}(f) = \int_0^1 (f^{(m)}(x))^2 \, dx.$$

Der Parameter λ heißt dabei *Glättungsparameter*. Wenn $\lambda = 0$ ist, dann erhalten wir eine Funktion die die Daten interpoliert. Wenn $\lambda = \infty$, dann darf die

Lösung keine m-te Ableitung haben und wir erhalten einen Kleinste-Quadrate-Schätzer mit dem Raum der Polynome vom Grad kleiner gleich $m-1$. Für allgemeines λ ist der resultierende Schätzer eng verwandt mit dem Kleinste-Quadrate-Schätzer in (3.1) mit $\mathcal{F} = \{f : [0,1] \to \mathbb{R} : \int_0^1 (f^{(m)}(x))^2\,dx \le L\}$. Beide Schätzer stimmen sogar überein sofern $\lambda = \phi(L)$ geeignet gewählt wird. Sind alle Kovariablen verschieden und gilt $n \ge m$, so kann man zeigen, dass es in beiden Fällen ein eindeutiges Minimum $\hat{f}_n$ gibt welches folgende Eigenschaften besitzt: $\hat{f}_n \in C^{2m-2}([0,1])$, $\hat{f}_n$ ist ein Polynom vom Grad kleiner gleich $2m-1$ auf jedem der Intervalle (X_i, X_{i+1}) und $\hat{f}_n$ ist Polynom vom Grad $m-1$ auf $(0, X_1]$ und $(X_n, 1]$. Der Schätzer $\hat{f}_n$ wird auch **smoothing spline** genannt (vgl. Übung für den Fall $m = 2$).

Zuletzt sei noch darauf hingewiesen, dass sich alle Schätzmethoden leicht auf den Fall $S \subseteq \mathbb{R}^d$ erweitern lassen.

3.3 Konsistenz des Kleinste-Quadrate-Schätzers

In diesem Kapitel zeigen wir mit Hilfe der Entropiemethoden aus Kapitel 2, dass der Kleinste-Quadrate-Schätzer (KQS) konsistent ist. Wir nehmen an, dass die Regressionsfunktion f_0 in einer gegebenen (nichtlinearen) Funktionenklasse $\mathcal{F}$ enthalten ist und betrachten den KQS

$$\hat{f}_n \in \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2$$

(X_i ersetzt durch x_i im Fall des deterministischen Designs). Wir nehmen im Folgenden stets an, dass ein Minimum existiert, eine (einfache) Modifikation der Resultate ist möglich falls $\hat{f}_n$ eine annähernde Lösung ist.

1. Fall: Deterministisches Design

Wir schreiben $Y = (Y_1, \dots, Y_n)^T$, $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T$. Sind $u, v \in \mathbb{R}^n$ so schreiben wir $\langle u, v \rangle_n = (1/n) \sum_{i=1}^n u_i v_i$ und $\|u\|_n^2 = (1/n) \sum_{i=1}^n u_i^2$ für das durch n normalisierte Euklidische Skalarprodukt und die zugehörige Norm. Für eine Funktion $f : S \to \mathbb{R}$ schreiben wir außerdem

$$\|f\|_n^2 = \frac{1}{n} \sum_{i=1}^n f^2(x_i),$$

$$\|Y - f\|_n^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - f(x_i))^2,$$

$$\langle \epsilon, f \rangle_n = \sum_{i=1}^n \epsilon_i f(x_i),$$

d.h. betrachten wir f unter $\|\cdot\|_n$ oder $\langle \cdot, \cdot \rangle_n$, so identifizieren wir f mit dem Vektor $(f(x_1), \dots, f(x_n))$. Wir wollen den Fehler $\|\hat{f}_n - f_0\|_n^2$ analysieren. Ausgangspunkt ist:

3.1 Lemma (Fundamentale Ungleichung)**.** *Es gilt*

$$\|\hat{f}_n - f_0\|_n^2 \leq 2\langle \epsilon, \hat{f}_n - f_0 \rangle_n.$$

Beweis. Nach Definition minimiert $\hat{f}_n$ den Ausdruck $\|Y - f\|_n^2$ über $f \in \mathcal{F}$. Da $f_0 \in \mathcal{F}$ erhalten wir also

$$\|Y - \hat{f}_n\|_n^2 \leq \|Y - f_0\|_n^2.$$

Es folgt

$$\|Y - f_0\|_n^2 + 2\langle \epsilon, f_0 - \hat{f}_n \rangle_n + \|\hat{f}_n - f_0\|_n^2 \leq \|Y - f_0\|_n^2.$$

□

Wollen wir für den Fehler $\|\hat{f}_n - f_0\|_n^2$ die Asymptotik $n \to \infty$ analysieren, so betrachten wir ein Dreieckschema $Y_{i,n} = f_0(x_{i,n}) + \epsilon_{i,n}$, $i = 1, \ldots, n$, $n \geq 1$ (betrachte zum Beispiel den prototypischen Fall $x_i = i/n$). Die Abhängigkeit von n wir jedoch meist unterdrückt.

3.2 Satz. *Die Fehlerterme $\epsilon_i = \epsilon_{i,n}$ nehmen Werte in $[-b, b]$ an. Es gelte*

$$\frac{1}{n} \log N(\mathcal{F}_n(2b), \|\cdot\|_n, \delta) \to 0 \qquad \forall \delta > 0$$

mit $\mathcal{F}_n(2b) = \{f \in \mathcal{F} : \|f - f_0\|_n \leq 2b\}$. Dann gilt

$$\mathbb{E}\,\|\hat{f}_n - f_0\|_n^2 \to 0 \qquad \textit{für} \qquad n \to \infty.$$

Beweis. Mit Hilfe von Lemma 3.1 und der Cauchy-Schwarz-Ungleichung gilt

$$\|\hat{f}_n - f_0\|_n^2 \leq 2\langle \epsilon, \hat{f}_n - f_0 \rangle_n \leq 2\|\epsilon\|_n \|\hat{f}_n - f_0\|_n \leq 2b\|\hat{f}_n - f_0\|_n.$$

Daher gilt immer $\|\hat{f}_n - f_0\|_n \leq 2b$ und somit $\hat{f}_n \in \mathcal{F}_n(2b)$. Wenden wir Lemma 3.1 erneut an, so erhalten wir

$$\|\hat{f}_n - f_0\|_n^2 \leq 2\langle \epsilon, \hat{f}_n - f_0 \rangle_n \leq 2 \sup_{f \in \mathcal{F}(2b)} \langle \epsilon, f - f_0 \rangle_n$$

Wir fixieren ein $\delta > 0$. Dann folgt aus Satz 2.30, dass

$$\begin{aligned}
(1/2)\,\mathbb{E}\,\|\hat{f}_n - f_0\|_n^2 &\leq \mathbb{E} \sup_{f \in \mathcal{F}_n(2b)} \langle \epsilon, f - f_0 \rangle_n \\
&\leq \mathbb{E} \sup_{f \in \mathcal{F}_n(2b)} |\langle \epsilon, f - f_0 \rangle_n| \\
&\leq 8b\delta + \frac{12b}{\sqrt{n}} \int_\delta^b \sqrt{\log 2N(\mathcal{F}_n(2b), \|\cdot\|_n, u)}\, du \\
&\leq 8b\delta + \frac{12b^2}{\sqrt{n}} \sqrt{\log 2N(\mathcal{F}_n(2b), \|\cdot\|_n, \delta)}.
\end{aligned}$$

Nach Voraussetzung gilt somit $\limsup_{n\to\infty} \mathbb{E}\,\|\hat{f}_n - f_0\|_n^2 \leq 16b^2\delta$. Da $\delta > 0$ beliebig ist, folgt die Behauptung. □

3.3 Aufgabe. *Im Regressionsmodell mit deterministischen Design und gleichmäßig beschränkten Fehlervariablen verwende Satz 3.2 um zu zeigen, dass der Kleinste-Quadrate-Schätzer L^1-konsistent ist in den folgenden Fällen*

(a) $f_0 \in \mathcal{F}$ *mit* $\mathcal{F} = \{f : [0,1] \to \mathbb{R} : \|f'\|_{L^2} \leq 1\}$,

(b) $f_0 \in \mathcal{F}$ *mit* $\|f_0\|_\infty \leq K$ *und* $\mathcal{F} = \{f : \mathbb{R} \to \mathbb{R} : f$ *monoton wachsend*$\}$.

Mit etwas mehr Aufwand kann man sogar zeigen:

3.4 Aufgabe. *Die Fehlerterme erfüllen*

$$\lim_{b\to\infty} \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(\epsilon_i^2 \mathbf{1}(|\epsilon_i| > b)) = 0$$

und es gelte $(1/n) \log N(\mathcal{F}_n(b), \|\cdot\|_n, \delta) \to 0$ *für alle* $\delta > 0$ *und alle* $b > 0$, *wobei* $\mathcal{F}_n(b) = \{f \in \mathcal{F} : \|f - f_0\|_n \leq b\}$. *Dann gilt* $\|\hat{f}_n - f_0\|_n^2 \xrightarrow{\mathbb{P}} 0$.

2. Fall: Zufälliges Design

Wir n nun den Fall des zufälligen Designs und analysieren den Vorhersagefehler

$$\|\hat{f}_n - f_0\|^2_{L^2(P^X)} = \int (\hat{f}_n(x) - f_0(x))^2 P^X(dx),$$

wobei P^X die Verteilung von X. Ausgangspunkt ist:

3.5 Lemma. *Unter den Voraussetzung zu Beginn von Kapitel 3.3 gilt im Fall des zufälligen Designs*

$$\|\hat{f}_n - f_0\|^2_{L^2(P^X)} \leq 2 \sup_{f\in\mathcal{F}} \Big| \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 - \mathbb{E}(Y - f(X))^2 \Big|$$

Beweis. Ist $f \in \mathcal{F}$, so gilt

$$\begin{aligned}
&\mathbb{E}(Y - f(X))^2 \\
&= \mathbb{E}(Y - f_0(X))^2 + 2\,\mathbb{E}(Y - f_0(X))(f_0(X) - f(X)) + \mathbb{E}(f_0(X) - f(X))^2 \\
&= \mathbb{E}(Y - f_0(X))^2 + \mathbb{E}(f_0(X) - f(X))^2
\end{aligned}$$

und somit

$$\|f - f_0\|^2_{L^2(P^X)} = \mathbb{E}(f(X) - f_0(X))^2 = \mathbb{E}(Y - f(X))^2 - \mathbb{E}(Y - f_0(X))^2.$$

Setzen wir $f = \hat{f}_n$, so erhalten wir

$$\begin{aligned}
\|\hat{f}_n - f_0\|^2_{L^2(P^X)} &= \int (y - \hat{f}_n(x))^2 dP^{X,Y}(x,y) - \int (y - f_0(x))^2 dP^{X,Y}(x,y) \\
&\leq \int (y - \hat{f}_n(x))^2 dP^{X,Y}(x,y) - \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{f}_n(X_i))^2 \\
&\quad + \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_0(X_i))^2 - \int (y - f_0(x))^2 dP^{X,Y}(x,y),
\end{aligned}$$

wobei wir in der Ungleichung die Definition des KQS eingesetzt und die Tatsache $f_0 \in \mathcal{F}$ verwendet haben. Wir schließen

$$\|\hat{f}_n - f_0\|^2_{L^2(P^X)} \leq 2 \sup_{f \in \mathcal{F}} \Big| \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2 - \int (y - f(x))^2 dP^{X,Y}(x,y) \Big|$$

und die Behauptung folgt. □

3.6 Satz. *Betrachte das Regressionsmodell mit zufälligem Design. Es gelten die Voraussetzung zu Beginn von Kapitel 3.3. Weiter gelte* $|Y| \leq b$ *fast sicher und* $\sup_{f \in \mathcal{F}} \|f\|_\infty \leq b$. *Außerdem sei* $\mathcal{F}$ *abzählbar. Es gelte*

$$\frac{1}{n} \log M(\mathcal{F}, \|\cdot\|_{L^2(P_n^X)}, \delta) \xrightarrow{\mathbb{P}} 0 \qquad \forall \delta > 0.$$

Dann gilt

$$E\|\hat{f}_n - f_0\|^2_{L^2(P^X)} \to 0 \qquad \textit{für} \qquad n \to \infty.$$

Beweis. (Aus $|Y| \leq b$ folgt $|f_0(X)| \leq b$ f.s. Daher ist die Bedingung $\sup_{f \in \mathcal{F}} \|f\|_\infty \leq b$ kein Widerspruch zu $f_0 \in \mathcal{F}$.) Lemma 3.5 besagt, dass

$$\|\hat{f}_n - f_0\|^2_{L^2(P^X)} \leq 2 \sup_{f \in \mathcal{F}} \Big| \frac{1}{n} \sum_{i=1}^n h(Z_i) - \mathbb{E}\, h(Z_i) \Big|$$

mit $Z_i = (X_i, Y_i)$ und $\mathcal{H} = \{h : S \times [-b,b] \to \mathbb{R} : h(x,y) = (y - f(x))^2, f \in \mathcal{F}\}$. Die Behauptung folgt also falls die Voraussetzungen aus Satz 2.32 für $\mathcal{H}$ gelten. Dabei ist (i) klar. Um (ii) zu zeigen, seien $h_1, \ldots, h_M \in \mathcal{H}$ mit $M = M(\mathcal{H}, \|\cdot\|_{L^2(P_n^Z)}, \delta)$ und $\|h_j - h_k\|_{L^2(P_n^Z)} > \delta$ für alle $j \neq k$. Sind nun h_1, h_2 zwei Funktionen aus $\mathcal{H}$ mit $h_1(x,y) = (y - f_1(x))^2$, $h_2(x,y) = (y - f_2(x))^2$ und $\|h_1 - h_2\|_{L^2(P_n^Z)} > \delta$, so gilt

$$\begin{aligned}
\delta^2 &< \frac{1}{n} \sum_{i=1}^n (h_1(Z_i) - h_2(Z_i))^2 \\
&= \frac{1}{n} \sum_{i=1}^n ((Y_i - f_1(X_i))^2 - (Y_i - f_2(X_i))^2)^2 \\
&= \frac{1}{n} \sum_{i=1}^n ((f_1(X_i) - f_2(X_i))(2Y_i - f_1(X_i) - f_2(X_i)))^2 \\
&\leq \frac{(4b)^2}{n} \sum_{i=1}^n (f_1(X_i) - f_2(X_i))^2 \qquad \text{f.s.}
\end{aligned}$$

wobei wir verwendet haben, dass $|2Y_i - f_1(X_i) - f_2(X_i)| \leq 4b$ f.s. nach Voraussetzung. Daher gilt $M(\mathcal{H}, \|\cdot\|_{L^2(P_n^Z)}, \delta) \leq M(\mathcal{F}, \|\cdot\|_{L^2(P_n^X)}, \delta/(4b))$ f.s. und (ii) folgt aus der Voraussetzung. □

3.4 Konvergenzraten für den KQS

In diesem Kapitel wollen wir mit Hilfe der Entropiemethoden aus Kapitel 2 Konvergenzraten für den KQS herleiten. Dabei beschränken wir uns auf den Fall des deterministischen Designs mit normalverteilten Fehlern. Wir beobachten also $(x_1, Y_1), \dots, (x_n, Y_n)$ mit

$$Y_i = f_0(x_i) + \epsilon_i, \qquad i = 1, \dots, n,$$

wobei $\epsilon_1, \dots, \epsilon_n$ unabhängig mit $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Wir nehmen an, dass die Regressionsfunktion f_0 in einer gegebenen Funktionenklasse $\mathcal{F}$ enthalten ist und betrachten den KQS $\hat{f}_n$ definiert durch $\hat{f}_n \in \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (Y_i - f(x_i))^2$. Wir nehmen im Folgenden stets an, dass ein Minimum existiert. Wir schreiben

$$\mathcal{F}_n(\delta) = \{f \in \mathcal{F} : \|f - f_0\|_n \leq \delta\}, \qquad \delta > 0$$

und

$$\hat{\delta}_n^2 = \|\hat{f}_n - f_0\|_n^2.$$

Um mehr als nur Konsistenz zu erhalten, müssen wir die Fundamentale Ungleichung in Lemma 3.1 auf etwas subtilere Art und Weise anwenden:

$$\hat{\delta}_n^2 = \|\hat{f}_n - f_0\|_n^2 \leq 2\langle \epsilon, \hat{f}_n - f_0 \rangle_n \leq 2 \sup_{f \in \mathcal{F}_n(\hat{\delta}_n)} \langle \epsilon, f - f_0 \rangle_n \tag{3.2}$$

Unser Ziel ist es diese Ungleichung nach $\hat{\delta}_n$ zu lösen. Eine Vorbereitung ist die folgende Aufgabe:

3.7 Aufgabe. *Wir nennen eine Funktion $\psi : [0, \infty) \to [0, \infty)$ sub-linear falls (i) ψ ist monoton wachsend; (ii) Die Abbildung $\delta \mapsto \psi(\delta)/\delta$ ist monoton fallend $\forall \delta > 0$. Sei nun ψ eine sub-lineare Funktion mit $\psi \not\equiv 0$. Dann gilt*

(a) ψ ist stetig auf $(0, \infty)$,

(b) Die Gleichung $\psi(\delta) = \delta^2$ besitzt eine eindeutig bestimmte Lösung $\delta_0 > 0$ und es gilt $\psi(\delta) \leq \delta^2$ genau dann, wenn $\delta \geq \delta_0$.

Unser Hauptresultat ist wie folgt:

3.8 Satz. *Wir betrachten das Regressionsmodell mit deterministischen Design und i.i.d. Fehlervariablen mit $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Sei $\mathcal{F}$ eine Menge von Funktionen mit $f_0 \in \mathcal{F}$ und sei $\hat{f}_n$ der zu $\mathcal{F}$ gehörige KQS. Außerdem sei ψ_n eine sub-lineare Funktion mit*

$$\psi_n(\delta) \geq \mathbb{E} \sup_{f \in \mathcal{F}_n(\delta)} \langle \epsilon, f - f_0 \rangle_n \qquad \forall \delta > 0$$

und sei $\delta_n > 0$ die eindeutig bestimmte Lösung der Gleichung $16\psi_n(\delta_n) = \delta_n^2$. Dann gilt

$$\mathbb{P}(\|\hat{f}_n - f_0\|_n \geq \delta) \leq C \exp\left(-\frac{n\delta^2}{2^{10}\sigma^2}\right) \qquad \forall \delta \geq \delta_n$$

mit einer absoluten Konstanten $C > 0$. Insbesondere gilt

$$\mathbb{E}\,\|\hat{f}_n - f_0\|_n^2 \le \delta_n^2 + 1024C\frac{\sigma^2}{n}.$$

Beweis. Idee des Beweises ist es Ungleichung (3.2) mit dem Gaußschen Konzentrationsphänomen zu kombinieren. Wir fixieren ein $\delta \ge \delta_n$. Aus Aufgabe 2.11 (mit $C = 1$, siehe Bemerkung 2.10) folgt, dass

$$\mathbb{P}\Big(\sup_{f\in\mathcal{F}_n(\delta)}\langle\epsilon, f - f_0\rangle_n < \mathbb{E}\sup_{f\in\mathcal{F}_n(\delta)}\langle\epsilon, f - f_0\rangle_n + u\Big) \ge 1 - 2\exp\left(-\frac{nu^2}{2\delta^2\sigma^2}\right)$$

für alle $\delta \ge 0$, wobei wir verwendet haben, dass für $f \in \mathcal{F}_n(\delta)$ die Ungleichung $\mathbb{E}\langle\epsilon, f - f_0\rangle_n^2 \le \delta^2\sigma^2/n$ gilt. Wir bezeichnen obigen Ereignis mit $\mathcal{E}(\delta, u)$. Wir setzen $\delta_j = 2^j\delta$, $u_j = 2^{-4}\delta_j^2$ und $\mathcal{E} = \bigcap_{j\ge0}\mathcal{E}(\delta_j, u_j)$. Dann gilt

$$\begin{aligned}\mathbb{P}(\mathcal{E}^c) \le \sum_{j\ge0} 2\exp\left(-\frac{nu_j^2}{2\delta_j^2\sigma^2}\right) &= \sum_{j\ge0} 2\exp\left(-\frac{n2^{4j-8}\delta^4}{2^{2j+1}\delta^2\sigma^2}\right)\\ &= \sum_{j\ge0} 2\exp\left(-\frac{n2^{2j}\delta^2}{2^9\sigma^2}\right) \overset{(*)}{\le} C\exp\left(-\frac{n\delta^2}{2^{10}\sigma^2}\right).\end{aligned}$$

Dabei kann $(*)$ wie folgt gezeigt werden. Gilt $y = n\delta^2/\sigma^2 \ge 1$, so folgt aus

$$2^{2j}y \ge \frac{y}{2} + 2^{2j-1}y \ge \frac{y}{2} + 2^{2j-1},$$

dass

$$\mathbb{P}(\mathcal{E}^c) \le \sum_{j\ge0} 2\exp\left(-\frac{n\delta^2}{2^{10}\sigma^2}\right)\exp\left(-\frac{2^{2j}}{2^{10}}\right) = C\exp\left(-\frac{n\delta^2}{2^{10}\sigma^2}\right).$$

Andererseits gilt immer $\mathbb{P}(\mathcal{E}^c) \le 1$. Daher gilt $(*)$ auch in dem Fall, dass $y = n\delta^2/\sigma^2 < 1$ (durch Anpassen von C). Sei nun $\delta' \ge \delta$ beliebig. Dann gilt $\delta_j \le \delta' < 2\delta_j = \delta_{j+1}$ für ein $j \ge 0$ und wir erhalten auf dem Ereignis $\mathcal{E}$:

$$\begin{aligned}&2\sup_{f\in\mathcal{F}_n(\delta')}\langle\epsilon, f - f_0\rangle_n\\ &\le 2\sup_{f\in\mathcal{F}_n(\delta_{j+1})}\langle\epsilon, f - f_0\rangle_n\\ &< 2\,\mathbb{E}\sup_{f\in\mathcal{F}_n(\delta_{j+1})}\langle\epsilon, f - f_0\rangle_n + 2u_{j+1}\\ &\le 2\psi_n(\delta_{j+1}) + 2u_{j+1} \le 2\delta_{j+1}^2/16 + 2\delta_{j+1}^2/16 = \delta_j^2 \le \delta'^2,\end{aligned}$$

wobei wir unter anderem verwendet haben, dass $16\psi_n(\delta_{j+1}) \le \delta_{j+1}^2$ gilt wegen $\delta_{j+1} \ge \delta_n$. Das heißt auf dem Ereignis $\mathcal{E}$ gilt $2\sup_{f\in\mathcal{F}_n(\delta')}\langle\epsilon, f - f_0\rangle_n < \delta'^2$ für alle $\delta' \ge \delta$. Es folgt, dass $\{\hat{\delta}_n \ge \delta\} \cap \mathcal{E} = \emptyset$ da auf diesem Ereignis wegen

(3.2) $\hat{\delta}_n^2 \le 2\sup_{f\in\mathcal{F}_n(\hat{\delta}_n)}\langle\epsilon, \hat{f}_n - f_0\rangle_n < \hat{\delta}_n^2$ gelten würde, was ein Widerspruch ist. Somit gilt $\{\hat{\delta}_n \ge \delta\} \subseteq \mathcal{E}^c$ und es folgt

$$\mathbb{P}(\hat{\delta}_n \ge \delta) \le \mathbb{P}(\mathcal{E}^c) \le C\exp\left(-\frac{n\delta^2}{2^{10}\sigma^2}\right).$$

Da $\delta \ge \delta_n$ beliebig ist, folgt die erste Behauptung. Die zweite Behauptung folgt aus

$$\begin{aligned}\mathbb{E}\,\|\hat{f}_n - f_0\|_n^2 = \mathbb{E}\,\hat{\delta}_n^2 &= \int_0^\infty \mathbb{P}(\hat{\delta}_n^2 \ge u)\,du \\ &\le \delta_n^2 + C\int_{\delta_n^2}^\infty \exp\left(-\frac{nu}{2^{10}\sigma^2}\right)du \le \delta_n^2 + 2^{10}C\frac{\sigma^2}{n}.\end{aligned}$$

□

3.9 Beispiel (Lineare Regression)**.** Sei $\mathcal{F} = \{f = \sum_{k=1}^d \theta_k f_k : \theta \in \mathbb{R}^d\}$. Aus Korollar 2.16 folgt, dass $M(\mathcal{F}_n(\delta), \|\cdot\|_n, u) \le (6\delta/u)^d$ für alle $u \le \delta$. Dudleys Entropieschranke (siehe Satz 2.28) liefert nun

$$\begin{aligned}\mathbb{E}\sup_{f\in\mathcal{F}_n(\delta)}\langle\epsilon, f - f_0\rangle_n &\le \frac{12\sigma}{\sqrt{n}}\int_0^{\delta/2}\sqrt{\log 2M(\mathcal{F}_n(\delta), \|\cdot\|_n, u)}\,du \\ &\le \frac{12\sigma\delta}{\sqrt{n}}\int_0^{1/2}\sqrt{\log 2(6/v)^d}\,dv \\ &\le 12\delta\sqrt{\frac{\sigma^2 d}{n}}\int_0^{1/2}\sqrt{\log(12/v)}\,dv = C\delta\sqrt{\frac{\sigma^2 d}{n}} =: \psi_n(\delta).\end{aligned}$$

Offensichtlich ist ψ_n sub-linear und $\delta_n = C\sqrt{\sigma^2 d/n}$ ist die Lösung der Gleichung $\psi_n(\delta) = \delta^2$. Wir erhalten also unter den Voraussetzungen von Satz 3.8, dass $\mathbb{E}\,\|\hat{f}_n - f_0\|_n^2 \le C_2\sigma^2 d/n$.

3.10 Korollar. *Es gelten die Voraussetzungen aus Satz 3.8. Die Klasse $\mathcal{F}$ erfülle*

$$\log M(\mathcal{F}, \|\cdot\|_n, u) \le C_1 u^{-\frac{1}{\alpha}} \qquad \forall u > 0$$

mit $\alpha > 1/2$. Dann existiert eine Konstante C_2 die nur von C_1, α und σ^2 abhängt, so dass

$$\mathbb{E}\,\|\hat{f}_n - f_0\|_n^2 \le C_2 n^{-\frac{2\alpha}{2\alpha+1}} \qquad \forall n \ge 1.$$

Beweis. Dudleys Entropieschranke liefert (siehe Satz 2.28 und Bemerkung 2.27)

$$\mathbb{E}\sup_{f\in\mathcal{F}_n(\delta)}\langle\epsilon, f - f_0\rangle_n \le \frac{12\sigma}{\sqrt{n}}\int_0^{\delta/2}\sqrt{\log M(\mathcal{F}_n(\delta), \|\cdot\|_n, u)}\,du.$$

Verwenden wir außerdem die Voraussetzung, so folgt

$$\mathbb{E} \sup_{f \in \mathcal{F}_n(\delta)} \langle \epsilon, f - f_0 \rangle_n \leq \frac{12\sigma}{\sqrt{n}} \int_0^{\delta/2} \sqrt{C_1} u^{-1/(2\alpha)} \, du$$
$$= \frac{12\sqrt{C_1}\sigma}{\sqrt{n}} \frac{(\delta/2)^{1-1/(2\alpha)}}{1 - 1/(2\alpha)} = C \frac{\sigma \delta^{1-1/(2\alpha)}}{\sqrt{n}} =: \psi_n(\delta).$$

Offensichtlich ist ψ_n sub-linear und es gilt

$$\psi_n(\delta) = \delta^2 \Leftrightarrow C \frac{\sigma \delta^{1-1/(2\alpha)}}{\sqrt{n}} = \delta^2 \Leftrightarrow \delta = \left(\frac{C^2 \sigma^2}{n} \right)^{\frac{\alpha}{2\alpha+1}}.$$

Wir erhalten also die Lösung $\delta_n = (C^2\sigma^2/n)^{\alpha/(2\alpha+1)}$ und die Behauptung folgt aus dem zweiten Teil von Satz 3.8. □

3.11 Beispiel (Monotone Regression)**.** Sei $\mathcal{F} = \{f : \mathbb{R} \to [0,1] : f \text{ monoton wachsend}\}$. Dannn gilt (siehe Satz 2.22) $\log M(\mathcal{F}_n(\delta), \|\cdot\|_n, u) \leq C_1/u$ für alle $u > 0$ und somit ist die Bedingung aus Korollar 3.10 mit $\alpha = 1$ erfüllt. Es folgt, dass $\mathbb{E}\,\|\hat{f}_n - f_0\|_n^2 \leq C_2 n^{-2/3}$.

3.12 Beispiel (Glatte Regression)**.** Sei $\mathcal{F} = \{f : [0,1] \to [0,1] : \int_0^1 (f^{(\alpha)}(t))^2 \, dt \leq 1\}$, $\alpha \in \mathbb{N}$. Dannn gilt (siehe Satz 2.20) $\log M(\mathcal{F}_n(\delta), \|\cdot\|_n, u) \leq C_1 u^{-1/\alpha}$ für alle $u > 0$ und Korollar 3.10 liefert $\mathbb{E}\,\|\hat{f}_n - f_0\|_n^2 \leq C_2 n^{-2\alpha/(2\alpha+1)}$. Im Fall $x_i = i/n$ ist diese Rate (wie schon im Fall der Dichteschätzung) optimal (Übungsaufgabe) und erhalten die Minimax-Konvergenzrate (im Sinne von Definition 1.38) $n^{-\alpha/(2\alpha+1)}$.

Konvergenzraten für den Projektionsschätzer

Wir betrachten das Regressionsmodell mit deterministischen Design und i.i.d. Fehlervariablen mit $\sigma^2 := \mathbb{E}\,\epsilon_i^2 < \infty$. Seien $\phi_1, \ldots, \phi_d : S \to \mathbb{R}$ Funktionen und $\mathcal{F}_d$ der lineare Raum definiert durch

$$\mathcal{F}_d = \Big\{ \sum_{k=1}^{d} \theta_k \phi_k : \theta = (\theta_1, \ldots, \theta_d)^T \in \mathbb{R}^d \Big\}$$

und sei $\hat{f}_n$ der zu $\mathcal{F}_d$ gehörige KQS (wird auch Projektionsschätzer genannt). Im Gegensatz zu oben nehmen wir in diesem Fall nicht mehr an, dass f_0 in $\mathcal{F}$ enthalten ist. Dieser Fall entspricht den Projektionsschätzern im Fall der Dichteschätzung aus Kapitel 1.5. Die Räume $\mathcal{F}_d$ werden in Abhängigkeit von der Anzahl der Beobachtungen n gewählt (in der Regel wächst die Dimension mit der Anzahl der Beobachtungen (*method of sieves*)). Man kann zum Beispiel den Raum aller trigonometrischen Polynome vom Grad kleiner gleich $d = d_n$ wählen, oder einen Raum stückweiser Polynome. Es gilt nun folgende Bias-Varianz-Zerlegung für den Projektionsschätzer:

3.13 Aufgabe. *Wir betrachten das Regressionsmodell mit deterministischen Design und i.i.d. Fehlervariablen mit* $\sigma^2 := \mathbb{E}\,\epsilon_i^2 < \infty$. *Seien* $\phi_1, \dots, \phi_d : S \to \mathbb{R}$ *Funktionen und sei* $\hat{f}_n$ *der zu* $\mathcal{F}_d = \{\sum_{k=1}^d \theta_k \phi_k : \theta \in \mathbb{R}^d\}$ *gehörige KQS. Dann gilt*

$$\mathbb{E}\,\|\hat{f}_n - f_0\|_n^2 = \min_{f \in \mathcal{F}_d} \|f - f_0\|_n^2 + \frac{r\sigma^2}{n},$$

wobei r *die Dimension des von den Vektoren* $(\phi_1(x_1), \dots, \phi_1(x_n))^T, \dots, (\phi_d(x_1), \dots, \phi_d(x_n))^T$ *aufgespannten Unterraums von* $\mathbb{R}^n$ *ist.*

3.14 Aufgabe. *Wir betrachten das Regressionsmodell mit deterministischen Design* $x_i = i/n$ *und i.i.d. Fehlervariablen mit* $\sigma^2 := \mathbb{E}\,\epsilon_i^2 < \infty$. *Sei* $(\phi_k)_{k \geq 1}$ *die trigonometrische Basis und* $\hat{f}_n$ *der KQS über* $\operatorname{span}(\phi_1, \dots, \phi_d)$. *Außerdem sei* $\mathcal{W}^\alpha(L)$ *die periodische Sobolev-Kugel mit* $L > 0$ *und* $\alpha \in \mathbb{N}$ *mit* $\alpha \geq 2$. *Zeige:*

(a) Ist $g : [0,1] \to \mathbb{R}$ *eine stetig differenzierbare Funktion, so gilt*

$$|\|g\|_n^2 - \|g\|_{L^2}^2| \leq 2n^{-1}\|g\|_\infty \|g'\|_\infty.$$

(b) Sei $f_0 = \sum_{k=1}^\infty \theta_k \phi_k \in \mathcal{W}^\alpha(L)$ *und* Π_d *die Orthogonalprojektion auf* $\operatorname{span}(\phi_1, \dots, \phi_d)$. *Dann gilt* $\sum_{k=1}^\infty b_k^{2\alpha}\theta_k^2 \leq K := L^2/\pi^{2\alpha}$ *mit* $b_k = k$ *für* k *gerade und* $b_k = k - 1$ *für* k *ungerade und*

$$\|f_0 - \Pi_d f_0\|_\infty \leq \sqrt{2} \sum_{k>d} |\theta_k| \leq \sqrt{2K} \sqrt{\sum_{k>d} b_k^{-2\alpha}} \leq C d^{-\alpha+1/2}$$

mit einer Konstanten $C > 0$ *die nur von* α *und* L *abhängt.*

(c) Aus (a) und (b) folgt, dass $\|f_0 - \Pi_d f_0\|_n^2 \leq C d^{-2\alpha}(1 + d^2/n)$ *mit einer Konstanten* $C > 0$ *die nur von* α *und* L *abhängt.*

(d) Ist $d = [n^{1/(2\alpha+1)}]$, *so gilt*

$$\sup_{f_0 \in \mathcal{W}^\alpha(L)} \mathbb{E}_{f_0}\,\|\hat{f}_n - f_0\|_n^2 \leq C n^{-\frac{2\alpha}{2\alpha+1}}$$

mit einer Konstanten $C > 0$ *die nur von* α, L *und* σ^2 *abhängt.*

3.5 Modellwahl

Modellwahl ist eine klassische Methode in der Statistik. Ziel ist es aus einer Menge von Modellen das auszuwählen welches am besten zum Schätzen geeignet ist. Wir werden uns hier auf den Fall der nichtparametrischen Regression mit deterministischen Design konzentrieren, eine analoge Theorie existiert unter anderem auch im Fall der Dichteschätzung.

3.15 Beispiel. Seien $\phi_1, \dots, \phi_p : S \to \mathbb{R}$ Funktionen mit p möglicherweise sehr groß (die Menge $\{\phi_1, \dots, \phi_p\}$ wird oft Wörterbuch (dictionary) genannt). Für eine Teilmenge $m \subseteq \{1, \dots, p\}$ setze $\mathcal{F}_m = \operatorname{span}(\phi_j : j \in m) = \{\sum_{j \in m} \beta_j \phi_j : \beta_j \in \mathbb{R}\}$ ($\mathcal{F}_m$ wird auch Modell genannt). Ist nun $\mathscr{M} \subseteq \mathscr{P}(\{1, \dots, p\})$ gegeben, so ist das Ziel ein $m \in \mathscr{M}$ zu finden, so dass der KQS (Projektionsschätzer) über $\mathcal{F}_m$ möglichst kleines Risiko besitzt. Das Problem besteht also darin aus dem Wörterbuch $\{\phi_1, \dots, \phi_p\}$ die signifikantesten Funktionen auszuwählen. Ein typisches Beispiel für ein Wörterbuch ist gegeben durch die ersten p Funktionen einer Basis (zum Beispiel der trigonometrischen Basis). Für ein weiteres Beispiel seien $z_1 < z_2 < \dots < z_p$ reelle Zahlen und $\phi_j(x) = \mathbf{1}(x \geq z_j)$. Dies führt zu sogenanntem change point detection.

3.16 Beispiel. Betrachtet man in Beispiel 3.15 die ersten p Funktionen der der trigonometrischen Basis und alle Mengen der Form $\{1, \dots, D\}$, $D \leq p$, so kann das Problem den besten Schätzer auszuwählen auch als Problem einen adaptiven Schätzer zu konstruieren, interpretiert werden. Ein analoges Problem ist wie folgt: Sei $S = [0, 1]$, $x_i = i/n$, $i = 1, \dots, n$ und $\mathcal{F}_m = \operatorname{span}(\mathbf{1}_{[0,1/m)}, \mathbf{1}_{[1/m,2/m)}, \dots, \mathbf{1}_{[1-1/m,1]})$ der Raum der stückweise konstanten Funktionen mit Knotenpunkten $0, 1/m, \dots, 1$. Ist nun $0 < \alpha \leq 1$ und $L > 0$, so erfüllt der KQS $\hat{f}_m$ über $\mathcal{F}_m$ mit $m = [n^{1/(2\alpha+1)}]$, dass

$$\sup_{f_0 \in \mathcal{H}^\alpha(S;L)} \mathbb{E}_{f_0} \|\hat{f}_m - f_0\|_n^2 \leq C n^{-\frac{2\alpha}{2\alpha+1}}$$

mit einer Konstanten C die nur von α, L und σ^2 abhängt. Der Schätzer besitzt also die Minimax-Konvergenzrate über $\mathcal{H}^\alpha(S; L)$ (Übungsaufgabe). Allerdings hängt der Schätzer vom unbekannten Glattheitsindex α ab. Ziel ist es einen adaptiven Schätzer zu konstruieren für den obige Schranke für das maximale Risiko für alle $0 < \alpha \leq 1$ erfüllt ist.

Wir betrachten nun im Regressionsmodell mit deterministischen Design und normalverteilten Fehlern $Y_i = f_0(x_i) + \epsilon_i$, $i = 1, \dots, n$, mit $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d. das folgende Problem:

1) Sei $(\mathcal{F}_m)_{m \in \mathscr{M}}$ eine (endliche) Menge von Modellen. In unserem Fall stets: $\mathcal{F}_m$ endlich-dimensionaler linearer Raum von Funktionen $f : S \to \mathbb{R}$;

2) Für $m \in \mathscr{M}$ sei $\hat{f}_m$ der KQS über $\mathcal{F}_m$.

Ziel ist den besten Schätzer auszuwählen. Dabei messen wir die Güte eines Schätzers durch das Risiko

$$R(m) = \mathbb{E} \|\hat{f}_m - f_0\|_n^2, \qquad m \in \mathscr{M}.$$

Eine optimale Wahl wäre

$$\hat{f}_{m^*} \qquad \text{mit} \qquad m^* \in \operatorname*{argmin}_{m \in \mathscr{M}} R(m).$$

Dies Wahl ist jedoch nicht möglich da $R(m)$ selbst von f_0 abhängt. Eine natürliche Idee ist nun das Risiko $R(m)$ durch einen Schätzer zu ersetzen und dann in $m \in \mathscr{M}$ zu minimieren. Wir wollen nun einen erwartungstreuen Schätzer von $R(m)$ konstruieren und verwenden dabei wieder die folgende Vektorenschreibweise/darstellung

$$Y = f_0 + \epsilon$$

mit $Y = (Y_1, \ldots, Y_n)^T$, $f_0 = (f_0(x_1), \ldots, f_0(x_n))^T$ und $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$. Setze außerdem $\hat{f}_m = (\hat{f}_m(x_1), \ldots, \hat{f}_m(x_n))^T$ (wir verzichten darauf Vektoren und Matrizen durch fettgedruckte Buchstaben darzustellen was den folgenden abuse of notation verhindern würde). Sei Π_m die Orthogonalprojektion von $\mathbb{R}^n$ auf $\{(f(x_1), \ldots, f(x_n))^T : f \in \mathcal{F}_m\}$ und sei $d_m = \dim(\{(f(x_1), \ldots, f(x_n))^T : f \in \mathcal{F}_m\})$ Dann gilt (siehe Aufgabe 3.13), dass $\hat{f}_m = \Pi_m Y$ und

$$\mathbb{E}\,\|f_0 - \hat{f}_m\|_n^2 = \|f_0 - \Pi_m f_0\|_n^2 + \frac{\sigma^2 d_m}{n}.$$

Verwendet man $Y - \hat{f}_m = (I - \Pi_m)Y = (I - \Pi_m)(f_0 + \epsilon)$, so erhält man analog

$$\begin{aligned}
&\mathbb{E}\,\|Y - \hat{f}_m\|_n^2 \\
&= \mathbb{E}(\|(I - \Pi_m)f_0\|_n^2 + 2\langle (I - \Pi_m)f_0, (I - \Pi_m)\epsilon\rangle_n + \|(I - \Pi_m)\epsilon\|_n^2) \\
&= \|f_0 - \Pi_m f_0\|_n^2 + 2\,\mathbb{E}\langle (I - \Pi_m)f_0, \epsilon\rangle_n + \mathbb{E}(\|\epsilon\|_n^2 - \|\Pi_m \epsilon\|_n^2) \\
&= \|f_0 - \Pi_m f_0\|_n^2 + \sigma^2 - \sigma^2 d_m/n.
\end{aligned}$$

Daher ist

$$R_n(m) = \|Y - \hat{f}_m\|_n^2 + \frac{2\sigma^2 d_m}{n} - \sigma^2$$

ein erwartungstreuer Schätzer von $R(m)$. Dies führt zu dem sogenannten Akaike-Informationskriterium (AIC) (Mallows C_p-Kriterium)

$$\hat{m}_{AIC} \in \operatorname*{argmin}_{m \in \mathscr{M}} \|Y - \hat{f}_m\|_n^2 + \frac{2\sigma^2 d_m}{n}.$$

Dieses Kriterium ist sehr natürlich und wird häufig angewendet. Es gibt jedoch Fälle in denen es schlechte Resultate liefern kann (das Kriterium berücksichtigt nicht die Streuung der $R_n(m)$ rund um $R(m)$. Wächst zum Beispiel die Anzahl der Modelle der Dimension d exponentiell in d, so wird sehr häufig ein zu großes Modell ausgewählt). Wir machen nun den folgenden allgemeinen Ansatz:

Für eine Funktion pen : $\mathscr{M} \to [0,\infty)$ (penalty function) betrachte

$$\hat{m} \in \operatorname*{argmin}_{m\in\mathscr{M}} \|Y - \hat{f}_m\|_n^2 + \operatorname{pen}(m)$$

und setze $\hat{f} = \hat{f}_{\hat{m}}$. Man spricht von Modellwahl via Strafterm. Der Schätzer $\hat{f}$ wird auch penalisierter KQS genannt. Ein erster Hinweis auf eine geeignete Wahl liefert das folgende Lemma (siehe auch die Ungleichung die wir im Fall der Dichteschätzung im Beweis von Satz 1.54 bewiesen haben):

3.17 Lemma. *Sei $\eta \in (0,1)$. Dann gilt für alle $m \in \mathscr{M}$*

$$\begin{aligned}&\eta^2\|f_0 - \hat{f}_{\hat{m}}\|_n^2\\ &\le (1-\eta+\eta^{-1})\|f_0 - \hat{f}_m\|_n^2 + \frac{1}{1-\eta}\left(\sup_{0\neq f\in\mathcal{F}_{\hat{m}}+\mathcal{F}_m}\frac{\langle\epsilon,f\rangle_n}{\|f\|_n}\right)^2 - \operatorname{pen}(\hat{m}) + \operatorname{pen}(m).\end{aligned}$$

Beweis. Wir fixieren ein $m \in \mathscr{M}$. Dann gilt nach Konstruktion

$$\|Y - \hat{f}_{\hat{m}}\|_n^2 + \operatorname{pen}(\hat{m}) \le \|Y - \hat{f}_m\|_n^2 + \operatorname{pen}(m)$$

Setzen wir $Y = f_0 + \epsilon$ ein, so folgt

$$\begin{aligned}&\|f_0 - \hat{f}_{\hat{m}}\|_n^2 + 2\langle\epsilon, f_0 - \hat{f}_{\hat{m}}\rangle_n + \|\epsilon\|_n^2 + \operatorname{pen}(\hat{m})\\ &\le \|f_0 - \hat{f}_m\|_n^2 + 2\langle\epsilon, f_0 - \hat{f}_m\rangle_n + \|\epsilon\|_n^2 + \operatorname{pen}(m)\end{aligned} \tag{3.3}$$

und somit

$$\|f_0 - \hat{f}_{\hat{m}}\|_n^2 \le \|f_0 - \hat{f}_m\|_n^2 + 2\langle\epsilon, \hat{f}_{\hat{m}} - \hat{f}_m\rangle_n - \operatorname{pen}(\hat{m}) + \operatorname{pen}(m).$$

Wir schätzen nun den zweiten Term der rechten Seite weiter ab. Mit Hilfe der Ungleichung $2xy \le ax^2 + (1/a)y^2$, $a > 0$, und der Dreiecksungleichung gilt

$$\begin{aligned}2\langle\epsilon, \hat{f}_{\hat{m}} - \hat{f}_m\rangle_n &\le 2\|\hat{f}_{\hat{m}} - \hat{f}_m\|_n \cdot \sup_{0\neq f\in\mathcal{F}_{\hat{m}}+\mathcal{F}_m}\frac{\langle\epsilon,f\rangle_n}{\|f\|_n}\\ &\overset{a=1-\eta}{\le} (1-\eta)\|\hat{f}_{\hat{m}} - \hat{f}_m\|_n^2 + \frac{1}{1-\eta}\left(\sup_{0\neq f\in\mathcal{F}_{\hat{m}}+\mathcal{F}_m}\frac{\langle\epsilon,f\rangle_n}{\|f\|_n}\right)^2\\ &\overset{a=\eta}{\le} (1-\eta)(1+\eta)\|f_0 - \hat{f}_{\hat{m}}\|_n^2 + (1-\eta)(1+1/\eta)\|f_0 - \hat{f}_m\|_n^2\\ &\quad + \frac{1}{1-\eta}\left(\sup_{0\neq f\in\mathcal{F}_{\hat{m}}+\mathcal{F}_m}\frac{\langle\epsilon,f\rangle_n}{\|f\|_n}\right)^2.\end{aligned}$$

Setzen wir dies in 3.3 ein und ordnen die Terme um, so folgt die Behauptung. □

Gesucht ist nun eine Strafterm der gerade so groß ist, dass $\mathrm{pen}(\hat{m})$ das Supremum in Lemma 3.17 kompensieren kann. Dabei spielt das Gaußsche Konzentrationsphänomen wieder eine große Rolle. Diese Strategie ist in dem folgenden Satz umgesetzt:

3.18 Satz. *Wir betrachten das Regressionsmodell mit deterministischen Design und i.i.d. Fehlervariablen mit $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Sei $(\mathcal{F}_m)_{m\in\mathscr{M}}$ eine (endliche) Menge von Modellen und $(\pi_m)_{m\in\mathscr{M}}$ eine Menge von positiven Zahlen mit $\pi_m \le 1$ und*

$$\sum_{m\in\mathscr{M}} \pi_m = \Sigma < \infty.$$

Für $K > 1$ setze

$$\mathrm{pen}(m) = \frac{K\sigma^2(\sqrt{d_m} + \sqrt{2\log(1/\pi_m)})^2}{n}.$$

Sei $\hat{m} \in \operatorname{argmin}_{m\in\mathscr{M}} \|Y - \hat{f}_m\|_n^2 + \mathrm{pen}(m)$ und $\hat{f} = \hat{f}_{\hat{m}}$ der zugehörige penalisierte KQS. Dann gilt

$$\mathbb{E}\,\|\hat{f}_{\hat{m}} - f_0\|_n^2 \le C_K \min_{m\in\mathscr{M}} \left(\mathbb{E}\,\|\hat{f}_m - f_0\|_n^2 + \frac{\sigma^2 \log(1/\pi_m)}{n} + \frac{\Sigma\sigma^2}{n} \right)$$

mit einer Konstanten $C_K > 0$ die nur von K abhängt.

3.19 Zusatz. *Erfüllt der Strafterm die Ungleichung*

$$\mathrm{pen}(m) \ge \frac{K\sigma^2(\sqrt{d_m} + \sqrt{2\log(1/\pi_m)})^2}{n}.$$

so gilt außerdem

$$\mathbb{E}\,\|\hat{f}_{\hat{m}} - f_0\|_n^2 \le C_K \min_{m\in\mathscr{M}} \left(\min_{f\in\mathcal{F}_m} \|f - f_0\|_n^2 + \mathrm{pen}(m) + \frac{\Sigma\sigma^2}{n} \right).$$

Gilt $\Sigma = 1$, so ist $(\pi_m)_{m\in\mathscr{M}}$ ein Wahrscheinlichkeitsvektor. In diesem kann sich a priori Wissen über das zugrundeliegende Modell widerspiegeln. In der Regel wird (π_m) allerdings so gewählt, dass die rechte Seite möglichst klein ist.

Beweis. Wir fixieren ein $m \in \mathscr{M}$. Aus Lemma 3.17 folgt, dass

$$\begin{aligned}
\eta^2\,\mathbb{E}\,\|\hat{f}_{\hat{m}} - f_0\|_n^2 &\le (1 - \eta + \eta^{-1})\,\mathbb{E}\,\|\hat{f}_m - f_0\|_n^2 + \mathrm{pen}(m) \\
&\quad + \mathbb{E}\left(\frac{1}{1-\eta} \left(\sup_{0\ne f\in\mathcal{F}_{\hat{m}}+\mathcal{F}_m} \frac{\langle \epsilon, f\rangle_n}{\|f\|_n} \right)^2 - \mathrm{pen}(\hat{m}) \right) \\
&\le (1 - \eta + \eta^{-1})\,\mathbb{E}\,\|\hat{f}_m - f_0\|_n^2 + \mathrm{pen}(m) \\
&\quad + \mathbb{E} \max_{m'\in\mathscr{M}} \left(\frac{1}{1-\eta} Z_{m'}^2 - \mathrm{pen}(m') \right) \qquad (3.4)
\end{aligned}$$

mit

$$Z_{m'} = \sup_{0 \neq f \in \mathcal{F}_{m'} + \mathcal{F}_m} \frac{\langle \epsilon, f \rangle_n}{\|f\|_n}, \qquad m' \in \mathscr{M}.$$

Aus Aufgabe 2.11 (mit $C = 1$, siehe Bemerkung 2.10) folgt, dass

$$\mathbb{P}(Z_{m'} - \mathbb{E}\, Z_{m'} \geq v) \leq 2 \exp\left(-\frac{nv^2}{2\sigma^2}\right) \qquad \forall v \geq 0, \tag{3.5}$$

wobei wir verwendet haben, dass $\mathbb{E}\langle \epsilon, f \rangle_n^2 / \|f\|_n^2 = \sigma^2/n$. Wähle nun eine ONB $\psi_1, \ldots, \psi_N$ von $\mathcal{F}_{m'} + \mathcal{F}_m$ bezüglich $\langle \cdot, \cdot \rangle_n$. Dann gilt mit Hilfe der Cauchy-Schwarz-Ungleichung

$$Z_{m'} = \sup_{0 \neq a \in R^N} \frac{1}{\|a\|} \sum_{k=1}^{N} a_k \langle \epsilon, \psi_k \rangle_n \leq \left(\sum_{k=1}^{N} \langle \epsilon, \psi_k \rangle_n^2 \right)^{1/2}.$$

und es folgt

$$\mathbb{E}\, Z_{m'} \leq (\mathbb{E}\, Z_{m'}^2)^{1/2} = \left(\sum_{k=1}^{N} \frac{\sigma^2 \|\psi_k\|_n^2}{n} \right)^{1/2} = \sqrt{\frac{\sigma^2 N}{n}} \leq \sqrt{\frac{\sigma^2 (d_{m'} + d_m)}{n}}.$$

Setzen wir dies in (3.5) ein und wählen außerdem

$$v = v_{m'} = \sqrt{\frac{2\sigma^2 u}{n} + \frac{2\sigma^2 \log(1/\pi_{m'})}{n}}$$

so folgt

$$\mathbb{P}\left(Z_{m'} \geq \sqrt{\frac{\sigma^2 (d_{m'} + d_m)}{n}} + v_{m'} \right) \leq 2\pi_{m'} \exp(-u)$$

und mittels σ-Sub-Additivität

$$\mathbb{P}\left(Z_{m'} \geq \sqrt{\frac{\sigma^2 (d_{m'} + d_m)}{n}} + v_{m'} \text{ für ein } m' \right) \leq 2\Sigma \exp(-u).$$

Komplementbildung liefert

$$\mathbb{P}\left(Z_{m'} < \sqrt{\frac{\sigma^2 (d_{m'} + d_m)}{n}} + v_{m'} \text{ für alle } m' \right) \geq 1 - 2\Sigma \exp(-u).$$

Wir bezeichnen das letzte Ereignis mit $\mathcal{E}$. Auf dem Ereignis $\mathcal{E}$ gilt nun für alle $m' \in \mathscr{M}$

$$\begin{aligned}
\frac{1}{1-\eta} Z_{m'}^2 &< \frac{1}{1-\eta} \left(\sqrt{\frac{\sigma^2 (d_{m'} + d_m)}{n}} + \sqrt{\frac{2\sigma^2 u}{n} + \frac{2\sigma^2 \log(1/\pi_{m'})}{n}} \right)^2 \\
&\leq \frac{1}{1-\eta} \left(\sqrt{\frac{\sigma^2 d_{m'}}{n}} + \sqrt{\frac{2\sigma^2 \log(1/\pi_{m'})}{n}} + \sqrt{\frac{2\sigma^2 u}{n}} + \sqrt{\frac{\sigma^2 d_m}{n}} \right)^2 \\
&\leq \frac{1+\eta}{1-\eta} \frac{1}{K} \operatorname{pen}(m') + \frac{1+\eta^{-1}}{1-\eta} \left(\sqrt{\frac{2\sigma^2 u}{n}} + \sqrt{\frac{\sigma^2 d_m}{n}} \right)^2,
\end{aligned}$$

wobei wir die Ungleichungen $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$, $x, y \ge 0$ und $(x+y)^2 \le (1+\eta)x^2 + (1+\eta^{-1})y^2$ verwendet haben. Es gilt nun

$$\frac{1+\eta}{1-\eta}\frac{1}{K} = 1 \Leftrightarrow \eta = \frac{K-1}{K+1}.$$

Für diese Wahl von η gilt auf dem Ereignis $\mathcal{E}$

$$\frac{1}{1-\eta} Z_{m'}^2 < \operatorname{pen}(m') + \frac{1+\eta^{-1}}{1-\eta}\left(\frac{4\sigma^2 u}{n} + \frac{2\sigma^2 d_m}{n}\right) \qquad \forall m' \in \mathscr{M}.$$

Setzen wir $C = (1+\eta^{-1})/(1-\eta)$, so folgt

$$\mathbb{P}\Big(\max_{m' \in \mathscr{M}} \Big(\frac{1}{1-\eta} Z_{m'}^2 - \operatorname{pen}(m')\Big) - \frac{2C\sigma^2 d_m}{n} \ge \frac{4C\sigma^2 u}{n}\Big) \le 2\Sigma \exp(-u).$$

Setzen wir weiter

$$V = \max_{m' \in \mathscr{M}} \Big(\frac{1}{1-\eta} Z_{m'}^2 - \operatorname{pen}(m')\Big) - \frac{2C\sigma^2 d_m}{n},$$

so schließen wir

$$\begin{aligned}
\mathbb{E}\, V \le \mathbb{E}\max(0, V) &= \int_0^\infty \mathbb{P}(\max(0,V) \ge x)\, dx \\
&= \int_0^\infty \mathbb{P}(V \ge x)\, dx \\
&\le \int_0^\infty 2\Sigma \exp\left(-\frac{nx}{4C\sigma^2}\right) dx = 8C\frac{\Sigma\sigma^2}{n}
\end{aligned}$$

Setzen wir dies in (3.4) ein so folgt

$$\eta^2\, \mathbb{E}\, \|\hat{f}_{\hat{m}} - f_0\|_n^2 \le (1-\eta+\eta^{-1})\, \mathbb{E}\, \|\hat{f}_m - f_0\|_n^2 + \operatorname{pen}(m) + C\left(\frac{2\sigma^2 d_m}{n} + \frac{8\Sigma\sigma^2}{n}\right).$$

Wegen

$$\frac{\sigma^2 d_m}{n} \le \mathbb{E}\, \|\hat{f}_m - f_0\|_n^2$$

und

$$\operatorname{pen}(m) \le \frac{2K\sigma^2 d_m}{n} + \frac{4K\sigma^2 \log(1/\pi_m)}{n} \le 2K\, \mathbb{E}\, \|\hat{f}_m - f_0\|_n^2 + \frac{4K\sigma^2 \log(1/\pi_m)}{n}$$

gilt also

$$\mathbb{E}\, \|\hat{f}_{\hat{m}} - f_0\|_n^2 \le C_K\left(\mathbb{E}\, \|\hat{f}_m - f_0\|_n^2 + \frac{\sigma^2 \log(1/\pi_m)}{n} + \frac{\Sigma\sigma^2}{n}\right).$$

und die Behauptung folgt da m beliebig war. Die Variante in Zusatz 3.19 folgt analog indem wir $K\sigma^2 d_m/n \le \operatorname{pen}(m)$ und

$$\mathbb{E}\, \|\hat{f}_m - f_0\|_n^2 = \min_{f \in \mathcal{F}_m} \|f - f_0\|_n^2 + \frac{\sigma^2 d_m}{n}$$

einsetzen. □

3.20 Beispiel. Sei $\{\phi_1, \dots, \phi_p\}$ ein Wörterbuch von Funktionen (z.B. die ersten p Funktionen der trigonometrischen Basis) und $\mathscr{M} \subseteq \mathscr{P}(\{1, \dots, p\})$ eine Teilmenge. Für $m \in \mathscr{M}$ setze $\mathcal{F}_m = \operatorname{span}(\phi_j : j \in m)$. Mit $|m|$ bezeichnen wir die Kardinalität von m.

1) Geordnete Variablenwahl Sei $\mathscr{M} = \{\{1, \dots, D\} : D \leq p\}$. Dann können wir

$$\operatorname{pen}(m) = \frac{K'\sigma^2|m|}{n}$$

wählen mit $K' > 1$ (insbesondere mit $K' = 2$ auch das AIC-Kriterium). Setze hierfür $\pi_m = \exp(-c|m|)$ mit $c > 0$ und

$$\operatorname{pen}(m) = \frac{K\sigma^2(\sqrt{|m|} + \sqrt{2\log(1/\pi_m)})^2}{n} = \frac{K(1+\sqrt{2c})^2\sigma^2|m|}{n}.$$

Daher gilt $K' = K(1+\sqrt{2c})^2$ sofern $K > 1$ und $c > 0$ geeignet gewählt werden. Es gilt

$$\sum_{m \in \mathscr{M}} \pi_m = \sum_{D=1}^{p} \exp(-cD) \leq \frac{1}{e^c - 1} =: \Sigma.$$

Aus Zusatz 3.19 folgt nun, dass

$$\mathbb{E}\,\|\hat{f} - f_0\|_n^2 \leq C \min_{m \in \mathscr{M}} \left(\min_{f \in \mathcal{F}_m} \|f - f_0\|_n^2 + \frac{\sigma^2|m|}{n} \right)$$

mit einer Konstanten C die nur von K und c abhängt. Diese Schranke kann noch leicht verbessert werden sofern wir $\mathscr{M} = \{\{1, \dots, D\} : D \leq p, d_m > d_{m'} \text{ für } m' = \{1, \dots, D-1\}\}$ wählen. Wenden wir nun Satz 3.18 mit $\pi_m = \exp(-d_m)$ an, so folgt

$$\mathbb{E}\,\|\hat{f} - f_0\|_n^2 \leq C_K \min_{m \in \mathscr{M}} \mathbb{E}\,\|\hat{f}_m - f_0\|_n^2 \tag{3.6}$$

wobei wir auch die Ungleichung $\sigma^2 \log(1/\pi_m)/n = \sigma^2 d_m/n \leq \mathbb{E}\,\|\hat{f}_m - f_0\|_n^2$ eingesetzt haben. Der penalisierte KQS ist also bis auf die Konstante C_K genauso gut wie der Orakel-Schätzer $\hat{f}_{m^*}$.

2) Vollständige Variablenwahl Sei $\mathscr{M} = \mathscr{P}(\{1, \dots, p\})$. Setze $\pi_m = \exp(-|m| \log p)$ und

$$\operatorname{pen}(m) = \frac{K\sigma^2(\sqrt{|m|} + \sqrt{2\log(1/\pi_m)})^2}{n} = \frac{K\sigma^2|m|(1+\sqrt{2\log p})^2}{n}$$

mit $K > 1$. Dann gilt

$$\sum_{m \in \mathscr{M}} \pi_m = \sum_{D=1}^{p} \binom{p}{D} \exp(-D \log p) \leq \sum_{D=1}^{p} \frac{p^D}{D!} p^{-D} \leq e$$

und Zusatz 3.19 liefert

$$\mathbb{E}\,\|\hat{f}_{\hat{m}} - f_0\|_n^2 \le C_K \min_{m\in\mathscr{M}} \left(\min_{f\in\mathcal{F}_m} \|f - f_0\|_n^2 + \frac{\sigma^2 |m|(1+\log p)}{n} \right).$$

Diese Schranke kann auch wie folgt formuliert werden

3.21 Korollar. *Für $\beta \in \mathbb{R}^p$ setze $|\beta|_0 = \sum_{j=1}^p \mathbf{1}(\beta_k \neq 0)$ und $f_\beta = \sum_{j=1}^p \beta_j \phi_j$. Dann gilt*

$$\mathbb{E}\,\|\hat{f} - f_0\|_n^2 \le C_K \inf_{\beta\in\mathbb{R}^p} \left(\|f_\beta - f_0\|_n^2 + \frac{\sigma^2 |\beta|_0 (1+\log p)}{n} \right)$$

Beweis. Setze $\operatorname{supp}(\beta) = \{j : \beta_j \neq 0\}$. Dann folgt die Behauptung durch einsetzen der Identität

$$\begin{aligned} &\min_{f\in\mathcal{F}_m} \|f - f_0\|_n^2 + \frac{\sigma^2 |m|(1+\log p)}{n} \\ &= \inf_{\beta:\operatorname{supp}(\beta)=m} \left(\|f_\beta - f_0\|_n^2 + \frac{\sigma^2 |\beta|_0 (1+\log p)}{n} \right). \end{aligned}$$

□

Die Schranke in Korollar 3.21 kann noch leicht verbessert werden indem man die Gewichte etwas anders wählt:

3.22 Aufgabe. *Für $1 \le D \le p$ gilt $D! \ge (D/e)^D$ und $\log \binom{p}{D} \le D(1+\log(p/D))$. Zeige, dass Zusatz 1.43 auch mit $\operatorname{pen}(m) = K\sigma^2(\sqrt{|m|} + \sqrt{2\log(1/\pi_m)})^2/n$ und $\pi_m = \exp(-|m|(1+\theta+\log(p/|m|))$, $\theta > 0$ angewendet werden kann und dass der resultierende penalisierte KQS $\hat{f}$ die folgende obere Schranke erfüllt:*

$$\mathbb{E}\,\|\hat{f} - f_0\|_n^2 \le C \inf_{0\neq\beta\in\mathbb{R}^p} \left\{ \|f_0 - f_\beta\|_n^2 + \frac{\sigma^2 |\beta|_0 \,(1+\log(p/|\beta|_0))}{n} \right\}.$$

mit einer Konstanten C die nur von K und θ abhängt.

3.23 Beispiel. Sei $S = [0,1]$, $x_i = i/n$, $i = 1,\dots,n$ und $\mathcal{F}_m = \operatorname{span}(\mathbf{1}_{[0,1/m)}, \mathbf{1}_{[1/m,2/m)}, \dots, \mathbf{1}_{[1-1/m,1]})$ der Raum der stückweise konstanten Funktionen mit Knotenpunkten $0, 1/m, \dots, 1$. Sei $\mathscr{M} = \{1,\dots,n\}$. Dann können wir

$$\operatorname{pen}(m) = \frac{K'\sigma^2 m}{n}$$

wählen mit $K' > 1$ (setze hierfür $\pi_m = \exp(-c|m|)$ mit $c > 0$ klein genug). Dann liefert Zusatz 3.19, dass

$$\mathbb{E}\,\|\hat{f} - f_0\|_n^2 \le C \min_{m=1,\dots,n} \left(\min_{f\in\mathcal{F}_m} \|f - f_0\|_n^2 + \frac{\sigma^2 m}{n} \right)$$

mit einer Konstanten C die nur von K und c abhängt. Angenommen es gilt nun $f_0 \in \mathcal{H}^\alpha(S;L)$ mit $0 < \alpha \leq 1$. Dann gilt $\min_{f\in\mathcal{F}_m} \|f - f_0\|_n^2 \leq L^2 m^{-2\alpha}$ und es folgt (wähle $m = [n^{1/(2\alpha+1)}]$)

$$\sup_{f_0\in\mathcal{H}^\alpha(S;L)} \mathbb{E}_{f_0} \|\hat{f} - f_0\|_n^2 \leq C n^{-\frac{2\alpha}{2\alpha+1}}$$

mit einer Konstanten $C > 0$ die nur von K, L und σ^2 abhängt. Diese Ungleichung gilt für alle $0 < \alpha \leq 1$. Das heißt wir haben einen adaptiven Schätzer konstruiert welcher die Minimax Konvergenzrate über alle Hölder-Kugeln $\mathcal{H}^\alpha(S;L)$, $0 < \alpha \leq 1$ besitzt. Eine analoge Betrachtung ist auch im Fall $\alpha > 1$ möglich, hier muss man jedoch auch Räume stückweiser Polynome zulassen. Des Weiteren kann man Hölder-Kugeln auch durch Sobolev-Kugeln ersetzen:

3.24 Aufgabe. *Kombiniere* (3.6) *mit Aufgabe 3.14 um einen adaptiven Schätzer zu konstruieren (in dem Sinn von Korollar 1.57).*

3.6 Der Lasso-Schätzer

In vielen Anwendungen gibt es verschiedene potentiell gut approximierende Basisfunktionen für die unbekannte Regressionsfunktion. Dies ist gerade bei höherdimensionalem Design wichtig, weil der Fluch der Dimension den Approximationsfehler betrifft (siehe Übung). Man denke an Funktionen $f : [0,1]^d \to \mathbb{R}$, die einerseits durch stückweise konstante Funktionen auf Unterwürfeln, andererseits auch durch Polynome kleinen Grades in d Variablen bzw. auch einer Linearkombination dieser beiden Funktionsklassen gut approximiert werden können. Die naheliegende Idee ist es, ein ganzes Wörterbuch $\{\phi_1, \dots, \phi_p\}$ von Funktionen $\phi_j : S \to \mathbb{R}$ in Betracht zu ziehen und die Regressionsfunktion durch eine Linearkombination einiger weniger dieser Ansatzfunktionen zu schätzen. Eine Lösung zu diesem Problem haben wir bereits im letzten Kapitel kennengelernt und zwar haben wir im Fall der vollständigen Variablenwahl den penalisierten KQS $\hat{f}_{\hat{m}}$ analysiert wobei

$$\hat{m} \in \operatorname*{argmin}_{m\subseteq\{1,\dots,p\}} \|Y - \hat{f}_m\|_n^2 + \lambda|m| \quad \text{mit} \quad \lambda = \frac{K\sigma^2(1+\sqrt{2\log p})^2}{n}.$$

Wir haben gesehen, dass $\hat{f}_{\hat{m}}$ gute theoretische Eigenschaften besitzt. Insbesondere haben wir in Korollar 3.21 gesehen, dass wir sehr gute obere Schranken erhalten, falls f_0 eine sparsame Darstellung besitzt in dem Sinn, dass $f_0 \approx f_\beta$ mit vielen $\beta_j = 0$ (d.h. $|\beta|_0$ klein). Diese Eigenschaft lässt sich auch in folgender Darstellung von $\hat{f}_{\hat{m}}$ intuitiv erfassen, die zu einer weiteren Interpretation von $\hat{f}_{\hat{m}}$ führt:

3.25 Aufgabe. *Es gilt $\hat{f}_{\hat{m}} = f_{\hat{\beta}}$ wobei $\hat{\beta}$ folgendes erfüllt:*

$$\hat{\beta} \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \|Y - f_\beta\|_n^2 + \lambda|\beta|_0.$$

Man kann allerdings zeigen, dass das (nicht-konvexe) Optimierungsproblem in Aufgabe 3.25 für große p nicht mehr implementierbar ist. Bis auf wenige Ausnahmen (siehe Übung für den Fall dass $\phi_1, \dots, \phi_p$ ein ONS bezüglich $\langle \cdot, \cdot \rangle_n$ bilden) muss man bei der Minimierung alle 2^p Teilmengen von $\{1, \dots, p\}$ durchlaufen, was schon im Fall $p \approx 100$ zu Berechenbarkeitsproblemen führt (NP-schweres Problem). Ein Ausweg ist es

$$|\beta|_0 = \sum_{j=1}^{p} \mathbf{1}(\beta_j \neq 0) \qquad \text{durch} \qquad |\beta|_1 = \sum_{j=1}^{p} |\beta_j|$$

zu ersetzen. Dabei ist die Abbildung $\beta \mapsto |\beta|_1$ konvex. Man spricht auch von konvexer Relaxation.

3.26 Definition. Betrachte das Regressionsmodell mit deterministischem Design. Sei $\{\phi_1, \dots, \phi_p\}$ eine Wörterbuch von Funktionen mit $\|\phi_j\|_n = 1$. Der zugehörige Lasso-Schätzer $\hat{f}^L = f_{\hat{\beta}^L}$ ist bestimmt durch das konvexe Minimierungsproblem

$$\hat{\beta}^L \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \|Y - f_\beta\|_n^2 + 2\lambda|\beta|_1.$$

Hierbei ist $\lambda > 0$ ein geeignet zu wählender Tuningparameter.

Man kann sehen (Bild!), dass der Lasso-Schätzer (wie auch der Schätzer aus Aufgabe 3.25) für λ groß Variablen auswählt, d.h. einige $\hat{\beta}_j^L$ gleich Null sind.

3.27 Lemma (Fundamentale Ungleichung)**.** *Für jedes $\beta \in \mathbb{R}^p$ gilt*

$$\|\hat{f}^L - f_0\|_n^2 \leq \|f_\beta - f_0\|_n^2 + 2\langle \epsilon, \hat{f}^L - f_\beta \rangle_n + 2\lambda|\beta|_1 - 2\lambda|\hat{\beta}^L|_1$$

Beweis. Nach Definition gilt $\|Y - \hat{f}^L\|_n^2 + 2\lambda|\hat{\beta}^L|_1 \leq \|Y - f_\beta\|_n^2 + 2\lambda|\beta|_1$. Die Behauptung folgt nun (wie schon im Beweis von Lemma 3.1) indem wir $Y = f_0 + \epsilon$ einsetzen und dann ausmultiplizieren. □

Für $m \subseteq \{1, \dots, p\}$ und $\nu \in \mathbb{R}^p$ sei ν_m der Vektor definiert durch

$$(\nu_m)_j = \begin{cases} \nu_j & \text{falls } j \in m, \\ 0 & \text{sonst.} \end{cases}$$

3.28 Lemma. *Sei $\beta \in \mathbb{R}^p$ und $m = \operatorname{supp}(\beta) = \{j : \beta_j \neq 0\}$. Dann gilt auf dem Ereignis $\mathcal{E} = \{\max_{j=1,\dots,p} |\langle \epsilon, \phi_j \rangle_n| \leq \lambda/2\}$:*

$$\|\hat{f}^L - f_0\|_n^2 + \lambda|\hat{\beta}^L - \beta|_1 \leq \|f_\beta - f_0\|_n^2 + 4\lambda|(\hat{\beta}^L - \beta)_m|_1.$$

Beweis. Auf dem Ereignis $\mathcal{E}$ gilt:

$$\begin{aligned}2\langle \epsilon, \hat{f}^L - f_\beta\rangle_n &= 2\sum_{j=1}^p (\hat{\beta}_j^L - \beta_j)\langle \epsilon, \phi_j\rangle_n \\ &\leq 2 \max_{j=1,\dots,p} |\langle \epsilon, \phi_j\rangle_n| \sum_{j=1}^p |\hat{\beta}_j^L - \beta_j| \leq \lambda |\hat{\beta}^L - \beta|_1.\end{aligned}$$

Setzen wir dies in Lemma 3.27 ein, so folgt

$$\|\hat{f}^L - f_0\|_n^2 \leq \|f_\beta - f_0\|_n^2 + \lambda|\hat{\beta}^L - \beta|_1 + 2\lambda|\beta|_1 - 2\lambda|\hat{\beta}^L|_1$$

Addieren wir $\lambda|\hat{\beta}^L - \beta|_1$ auf beiden Seiten, so schließen wir

$$\begin{aligned}\|\hat{f}^L - f_0\|_n^2 + \lambda|\hat{\beta}^L - \beta|_1 &\leq \|f_\beta - f_0\|_n^2 + 2\lambda\sum_{j=1}^p (|\hat{\beta}_j^L - \beta_j| + |\beta_j| - |\hat{\beta}_j^L|) \\ &= \|f_\beta - f_0\|_n^2 + 2\lambda\sum_{j\in m} (|\hat{\beta}_j^L - \beta_j| + |\beta_j| - |\hat{\beta}_j^L|) \\ &\leq \|f_\beta - f_0\|_n^2 + 4\lambda\sum_{j\in m} |\hat{\beta}_j^L - \beta_j|,\end{aligned}$$

wobei wir in der letzten Ungleichung die Dreiecksungleichung angewendet haben. □

3.29 Lemma. *Die Fehlervariablen erfüllen $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Wählen wir*

$$\lambda = 2\sigma\sqrt{\frac{2L + 2\log(2p)}{n}},$$

so gilt für das Ereignis $\mathcal{E}$ aus Lemma 3.28, dass $\mathbb{P}(\mathcal{E}^c) \leq \exp(-L)$.

Beweis. Erinnerung: Ist $X \sim \mathcal{N}(0,1)$, so gilt $\mathbb{P}(|X| \geq u) \leq 2\exp(-u^2/2)$ für alle $u \geq 0$. Es gilt also

$$\begin{aligned}\mathbb{P}(\mathcal{E}^c) &= \mathbb{P}(\exists j : |\langle \epsilon, \phi_j\rangle_n| > \lambda/2) \\ &\leq \sum_{j=1}^p \mathbb{P}(|\langle \epsilon, \phi_j\rangle_n| > \lambda/2) \\ &\leq \sum_{j=1}^p 2\exp\left(-\frac{n\lambda^2}{8\sigma^2}\right) = 2p\exp(-L - \log(2p)) = \exp(-L),\end{aligned}$$

wobei wir verwendet haben, dass $\langle \epsilon, \phi_j\rangle_n \sim \mathcal{N}(0, \sigma^2/n)$. □

Wir wollen nun sehen wie aus Lemma 3.28 und Lemma 3.29 eine obere Schranke für den Lasso-Schätzer folgt. Es bleibt, den Koeffizientenfehler

$|(\hat{\beta}^L - \beta)_m|_1$ mit dem Verlust $\|\hat{f}^L - f_\beta\|_n$ zu verknüpfen. Wir betrachten dabei zuerst den einfacheren Fall, dass $p \leq n$ und dass die Gram-Matrix

$$G_n = (\langle \phi_j \phi_k \rangle_n)_{j,k=1}^p$$

positiv definit ist (der Rang von G_n ist beschränkt durch $p \wedge n$), d.h.

$$\phi_{\min} = \min_{0 \neq \nu \in \mathbb{R}^p} \frac{\nu^T G_n \nu}{\|\nu\|^2} > 0$$

wobei $\|\cdot\|$ die Euklidische Norm bezeichne. Sei im Folgenden $0 \neq \beta \in \mathbb{R}^p$ und $m = \operatorname{supp}(\beta)$. Dann gilt

$$\hat{\beta}^L - \beta)_m|_1 \leq \sqrt{\frac{|\beta|_0}{\phi_{\min}}} \|\hat{f}^L - f_\beta\|_n. \tag{3.7}$$

Um dies zu beweisen setze $\nu = \hat{\beta}^L - \beta$. Dann gilt

$$|\nu_m|_1^2 = \Big(\sum_{j \in m} |\nu_j| \Big)^2 \leq |m| \sum_{j \in m} \nu_j^2 \leq |m| \|\nu\|^2 \leq \frac{|m|}{\phi_{\min}} \nu^T G_n \nu$$

Des Weiteren gilt $|m| = |\beta|_0$ und

$$\nu^T G_n \nu = \sum_{j,k=1}^p \nu_j \nu_k \langle \phi_j, \phi_k \rangle_n = \langle f_\nu, f_\nu \rangle_n = \|\hat{f}^L - f_\beta\|_n^2$$

Setzen wir dies in obige Ungleichung ein, so folgt (3.7).

Setzen wir die Behauptung in Lemma 3.28 ein, so erhalten wir auf dem Ereignis $\mathcal{E}$:

$$\begin{aligned} \|\hat{f}^L - f_0\|_n^2 &\leq \|f_\beta - f_0\|_n^2 + 4\lambda \sqrt{\frac{|\beta|_0}{\phi_{\min}}} \|\hat{f}^L - f_\beta\|_n \\ &\leq \|f_\beta - f_0\|_n^2 + \frac{16\lambda^2 |\beta|_0}{\phi_{\min}} + \frac{1}{2}\|\hat{f}^L - f_0\|_n^2 + \frac{1}{2}\|f_\beta - f_0\|_n^2, \end{aligned} \tag{3.8}$$

wobei wir zweimal die Ungleichung $2xy \leq ax^2 + (1/a)y^2$, $a > 0$ verwendet haben. Ordnen wir die Terme um wählen wir λ wie in Lemma 3.29 an, so schließen wir, dass der Lasso-Schätzer mit Wahrscheinlichkeit $\geq 1 - \exp(-L)$ die folgende Schranke erfüllt

$$\|\hat{f}^L - f_0\|_n^2 \leq \inf_{\beta \in \mathbb{R}^p} \left(3\|f_\beta - f_0\|_n^2 + \frac{2^8 \sigma^2 (L + \log(2p))}{\phi_{\min} n} |\beta|_0 \right).$$

Die rechte Seite kann dabei als eine verallgemeinerte Bias-Varianz-Zerlegung aufgefasst werden.

Wir wenden uns nun dem allgemeinen Fall zu. Ist das Wörterbuch übervollständig ($p > 0$) so ist G_n immer singulär und obige Herangehensweise ist nicht mehr möglich. Wir fordern daher eine schwächere Kompatibilität der Gram-Matrix bezüglich der L^1 Norm. Diese ist in Anwedungen häufig erfüllt, jedoch nur schwer überprüfbar, siehe [7] für eine ausführliche Diskussion.

3.30 Definition. Sei $0 \neq \beta \in \mathbb{R}^p$ und $m = \text{supp}(\beta)$. Setze

$$\mathcal{C}(\beta) = \{\nu \in \mathbb{R}^p : |\nu_{m^c}|_1 < 4|\nu_m|_1\}.$$

Dann ist die Kompatibilitätkonstante $\kappa(\beta)$ definiert durch

$$\kappa(\beta)^2 = \min_{0 \neq \nu \in \mathcal{C}(\beta)} \frac{|m|\nu^T G_n \nu}{|\nu_m|_1^2}.$$

3.31 Satz. *Wir betrachten das Regressionsmodell mit deterministischen Design und i.i.d. Fehlervariablen mit $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Für $L > 0$ betrachte den Lasso-Schätzer mit Tuningparameter*

$$\lambda = 2\sigma\sqrt{\frac{2L + 2\log(2p)}{n}}. \tag{3.9}$$

Dann gilt mit Wahrscheinlichkeit mindestens $1 - e^{-L}$

$$\|\hat{f}^L - f_0\|_n^2 \leq \inf_{0 \neq \beta \in \mathbb{R}^p} \left(5\|f_\beta - f_0\|_n^2 + \frac{2^8\sigma^2(L + \log(2p))}{n\kappa(\beta)^2}|\beta|_0 \right).$$

Diese obere Schranke (Orakelungleichung) für den Lasso Schätzer ist schwächer als die entsprechende obere Schranke für $\hat{f}_{\hat{m}}$ in Korollar 3.21. Der Lasso-Schätzer stellt einen Kompromiss zwischen Theorie und Praxis dar.

Beweis. Wir fixieren ein $\beta \neq 0$. Da $\mathbb{P}(\mathcal{E}) \geq 1 - e^{-L}$ gilt, genügt es zu zeigen, dass obige Ungleichung auf $\mathcal{E}$ erfüllt ist. Wir nehmen daher im Folgenden stets an, dass das Ereignis $\mathcal{E}$ gilt. Wir betrachten separat die Fälle $\lambda|(\hat{\beta}^L - \beta)_m|_1 \leq \|f_\beta - f_0\|_n^2$ und $\lambda|(\hat{\beta}^L - \beta)_m|_1 > \|f_\beta - f_0\|_n^2$. Ist $\lambda|(\hat{\beta}^L - \beta)_m|_1 \leq \|\hat{f}_\beta - f_0\|_n^2$, so folgt aus Lemma 3.28

$$\|\hat{f}^L - f_0\|_n^2 \leq \|f_\beta - f_0\|_n^2 + 4\lambda|(\hat{\beta}^L - \beta)_m|_1 \leq 5\|f_\beta - f_0\|_n^2.$$

Gilt andererseits $\lambda|(\hat{\beta}^L - \beta)_m|_1 > \|\hat{f}_\beta - f_0\|_n^2$, so folgt aus Lemma 3.28, dass

$$\lambda|\hat{\beta}^L - \beta|_1 \leq \|f_\beta - f_0\|_n^2 + 4\lambda|(\hat{\beta}^L - \beta)_m|_1 < 5\lambda|(\hat{\beta}^L - \beta)_m|_1.$$

Dies impliziert, dass $\nu = \hat{\beta}^L - \beta \in \mathcal{C}(\beta)$. Es folgt

$$|(\hat{\beta}^L - \beta)_m|_1^2 = |\nu_m|_1^2 \leq \frac{|m|\nu^T G_n \nu}{\kappa(\beta)^2} = \frac{|m|\|\hat{f}^L - f_\beta\|_n^2}{\kappa(\beta)^2}$$

Setzen wir die Behauptung in Lemma 3.28 ein, so erhalten indem wir wie in (3.8) vorgehen:

$$\begin{aligned}\|\hat{f}^L - f_0\|_n^2 &\leq \|f_\beta - f_0\|_n^2 + 4\lambda\frac{\sqrt{|m|}\|\hat{f}^L - f_\beta\|_n}{\kappa(\beta)} \\ &\leq \|f_\beta - f_0\|_n^2 + \frac{16\lambda^2|m|}{\kappa(\beta)^2} + \frac{1}{2}\|\hat{f}^L - f_0\|_n^2 + \frac{1}{2}\|f_\beta - f_0\|_n^2.\end{aligned}$$

Umordnen und einsetzen von λ und $|m| = |\beta|_0$ liefert

$$\|\hat{f}^L - f_0\|_n^2 \leq 3\|f_\beta - f_0\|_n^2 + \frac{2^8\sigma^2(L + \log(2p))}{n\kappa(\beta)^2}|\beta|_0.$$

Daher erhalten wir in beiden Fällen die Behauptung. □

3.7 Hochdimensionale Regressionsmodelle

In der klassischen Statistik liegt häufig folgende Situation vor: kleine Anzahl von Parametern p, große Anzahl von Beobachtungen n. Klassische Resultate beschreiben daher das Verhalten von Schätzern für $n \to \infty$ und p fest. Allerdings findet sich in aktuellen Anwendungen oft das umgekehrte Bild: große Anzahl an Parametern p, Anzahl von Beobachtungen n mit $n \approx p$ oder sogar $p \gg n$ (Beispiele: Genanalyse (der Mensch besitzt ≈ 25000 Gene), Bildanalyse ($10^3 - 10^6$ Pixel), Webseiten sammeln große Mengen an Daten ihrer Kunden, etc.).

Ein prototypisches hochdimensionales Regressionsmodell ist gegeben durch

$$Y_i = f_0(x^{(i)}) + \epsilon_i, \qquad i = 1, \ldots, n,$$

mit $\epsilon_1, \ldots, \epsilon_n$ i.i.d. mit $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^p$ mit $p \gg n$ und

$$f_0(x) = \langle x, \beta_0 \rangle$$

mit $s_0 := |\operatorname{supp}(\beta_0)|$ klein. Das heißt die Anzahl der Parameter ist sehr viel größer als die Anzahl der Beobachtungen, die meisten Parameter können bei der Beschreibung von Y allerdings herausgenommen werden. Wir betrachten im Folgenden den Lasso-Schätzer mit den Koordinatenfunktionen $\phi_j(x) = x_j$ als Wörterbuch. Wir können also Satz 3.31 anwenden, wollen dabei aber unsere funktionale Schreibweise durch eine Matrizen-Schreibweise ersetzen. Setzen wir

$$\mathbf{X} = \begin{pmatrix} x_1^{(1)} & \ldots & x_p^{(1)} \\ \vdots & & \vdots \\ x_1^{(n)} & \ldots & x_p^{(n)} \end{pmatrix}$$

so kann dass Modell auch geschrieben werden als

$$Y = \mathbf{X}\beta_0 + \epsilon \tag{3.10}$$

mit $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. Des Weiteren gelten die folgenden notationellen Entsprechungen:

$$(f_\beta(x^{(1)}), \cdots, f_\beta(x^{(n)}))^T = \mathbf{X}\beta, \qquad G_n = (1/n)\mathbf{X}^T\mathbf{X}$$

$$\|f_\beta\|_2^2 = \frac{1}{n}\|\mathbf{X}\beta\|^2, \qquad \|\hat{f}^L - f_0\|_2^2 = \frac{1}{n}\|\mathbf{X}(\hat{\beta}^L - \beta_0)\|^2$$

Wir erhalten aus Satz 3.31:

3.32 Satz. *Wir betrachten das Regressionsmodell* (3.10)*. Die Diagonalelemente von* $(1/n)\mathbf{X}^T\mathbf{X}$ *seien gleich* 1*. Sei* $s_0 = |\operatorname{supp}(\beta_0)|$*. Sei* $\hat{\beta}^L$ *der Lasso-Schätzer definiert durch* $\hat{\beta}^L = \operatorname{argmin}_{\beta\in\mathbb{R}^p} \|Y - \mathbf{X}\beta\|^2/n + \lambda|\beta|_1$ *mit* λ *wie in* (3.9)*. Dann gilt mit Wahrscheinlichkeit mindestens* $1 - e^{-L}$

$$\|\mathbf{X}(\hat{\beta} - \beta_0)\|^2 \leq \frac{2^8\sigma^2 s_0}{\kappa(\beta)^2}(L + \log(2p)). \tag{3.11}$$

4 Klassifikation und statistische Lerntheorie

Häufig müssen in der Statistik Entscheidungen zwischen zwei oder mehr Alternativen aufgrund komplexer Daten getroffen werden. Dies führt auf sogenannte Klassifikationsprobleme. Beispiele sind Spam-Filter (Klassifikationen zwischen 'mit Sicherheit Spam', 'mit Sicherheit kein Spam' und Klassen dazwischen), medizinische Datenanalyse (wie EKG weist auf Krankheit hin oder nicht) oder Schrifterkennung (ASCII-Code auf der Basis von Pixelmuster). Hier werden wir nur binäre Klassifikation mit Klassen (Labels) 0 und 1 betrachten.

4.1 Modell

Sei (X, Y) ein Paar von Zufallsvariablen wobei X Werte in $(S, \mathscr{S})$ und Y Werte in $\{0, 1\}$ annimmt. Wir bezeichnen mit η die Regressionsfunktion gegeben durch

$$\eta(x) = \mathbb{E}(Y|X = x) = \mathbb{P}(Y = 1|X = x).$$

4.1 Definition. Eine messbare Funktion $h : S \to \{0, 1\}$ heißt Klassifizierer. Für einen Klassifizierer h heißt

$$R(h) = \mathbb{P}(Y \neq h(X)) = \int \mathbf{1}_{y\neq h(x)}\, dP^{X,Y}(x, y)$$

Klassifizierungsfehler von h. Der Klassifizierer h^* definiert durch

$$h^*(x) = \begin{cases} 1, & \text{falls } \eta(x) > 1/2 \\ 0, & \text{falls } \eta(x) \leq 1/2 \end{cases}$$

heißt Bayes-Klassifizierer.

4.2 Satz. *Es gilt*

$$R(h^*) = \min_h R(h),$$

wobei das Minimum über alle Klassifizierer genommen wird.

Beweis. Für einen Klassifizierer gilt $Y - h(X) \in \{0, -1, 1\}$ und

$$\begin{aligned}\mathbb{P}(Y \neq h(X)) &= \mathbb{E}\,\mathbf{1}_{Y \neq h(X)} \\ &= \mathbb{E}(Y - h(X))^2 \\ &= \mathbb{E}(Y - \eta(X))^2 + \mathbb{E}(\eta(X) - h(X))^2.\end{aligned}$$

Der erste Term hängt nicht von h ab. Nach Definition von h^* gilt

$$(\eta(x) - h(x))^2 \geq (\eta(x) - h^*(x))^2 \qquad \forall x \in S.$$

Daher erfüllt der zweite Term

$$\mathbb{E}(\eta(X) - h(X))^2 \geq \mathbb{E}(\eta(X) - h^*(X))^2$$

und die Behauptung folgt. □

Man nennt

$$R^* = R(h^*) = \min_h R(h)$$

auch das Bayes-Risiko. Der Bayes-Klassifizierer besitzt also minimales Risiko. Die Verteilung von (X, Y) ist unbekannt, daher kann der Bayes-Klassifizierer nicht berechnet werden. Wie im Regressionsmodell mit zufälligem Design nehmen wir nun an, dass wir n unabhängige Kopien $(X_1, Y_1), \ldots, (X_n, Y_n)$ von (X, Y) beobachten. Unser Ziel ist es einen Klassifizierer $\hat{h}_n = \hat{h}_n(X_1, Y_1, \ldots, X_n, Y_n) : S \to \{0, 1\}$ zu konstruieren, so dass das sogenannte Exzessrisiko von $\hat{h}_n$

$$R(\hat{h}_n) - R^*$$

möglichst klein ist. Beachte dabei, dass $R(\hat{h}_n)$ eine Zufallsvariable ist, der Erwartungswert wird nur bezüglich (X, Y) genommen (betrachte R als Funktion definiert auf der Menge aller Klassifizierer):

$$R(\hat{h}_n) = \int \mathbf{1}_{y \neq \hat{h}_n(x)}\, dP^{X,Y}(x, y).$$

Dabei interpretieren wir $R(\hat{h}_n)$ als den mittleren Fehler den wir machen wenn wir $\hat{h}_n$ zum Klassifizieren einer neuen Beobachtung (X, Y) verwenden.

Parametrische Modellierung

Die Verteilung von (X, Y) ist bestimmt durch η und P^X. Angenommen P^X besitzt eine Dichte f bezüglich des Lebesguemaßes. Setze $\pi_1 = \mathbb{P}(Y = 1)$ und $\pi_0 = \mathbb{P}(Y = 0)$. Wir nehmen weiter an, dass $\pi_1 \in (0, 1)$. Dann gilt

$$\mathbb{P}(X \in A | Y = 1) = \frac{\mathbb{P}(X \in A, Y = 1)}{\mathbb{P}(Y = 1)} = \int_A \frac{\eta(x) f(x)}{\pi_1}\, dx, \qquad A \in \mathscr{B}_{\mathbb{R}}$$

und analog

$$\mathbb{P}(X \in A | Y = 0) = \int_A \frac{(1-\eta(x))f(x)}{\pi_0}\, dx, \qquad A \in \mathscr{B}_{\mathbb{R}}.$$

Setze $f_1(x) = \eta(x)f(x)/\pi_1$ und $f_0(x) = (1-\eta(x))f(x)/\pi_0$ (bedingten Dichten von X gegeben $Y = 1$ beziehungsweise $Y = 0$). Dann erhalten wir $f(x) = \pi_1 f_1(x) + \pi_0 f_0(x)$ und somit die Bayesformel

$$\eta(x) = \frac{\pi_1 f_1(x)}{\pi_1 f_1(x) + \pi_0 f_0(x)} \qquad (P^X\text{-f.ü.})$$

Setzen wir dies in die Definition des Bayes-Klassifizierers ein, so folgt

$$h^*(x) = \begin{cases} 1, & \text{falls } \pi_1 f_1(x) > \pi_0 f_0(x), \\ 0, & \text{falls } \pi_1 f_1(x) \leq \pi_0 f_0(x). \end{cases}$$

Beachte, dass die Verteilung von (X, Y) auch bestimmt ist durch π_1, f_0 und f_1. Ein beliebtes parametrisches Modell ist gegeben durch $S = \mathbb{R}^d$, f_0 Dichte einer Normalverteilung mit Erwartungswert μ_0 und Kovarianz Σ_0 und f_1 Dichte einer Normalverteilung mit Erwartungswert μ_1 und Kovarianz Σ_1. Gilt zusätzlich $\Sigma = \Sigma_0 = \Sigma_1$, so ist der Bayes-Klassifizierer gegeben durch

$$h^*(x) = \begin{cases} 1, & \text{falls } \left\langle \Sigma^{-1}(\mu_1 - \mu_0), x - \frac{\mu_1+\mu_0}{2} \right\rangle - \log(\pi_0/\pi_1) > 0, \\ 0, & \text{sonst}, \end{cases}$$

das heißt der Bayes-Klassifizierer ist ein affiner Klassifizierer. Ersetzen wir $\Sigma, \pi_1, \pi_0, \mu_1, \mu_0$ durch ihre (natürlichen) empirischen Versionen, so erhalten wir einen berechenbaren Klassifizierer.

Nichtparametrischer Plug-in Ansatz

Ist $\hat{\eta}_n$ ein Schätzer aus Kapitel 3, so können wir

$$\hat{h}_n(x) = \begin{cases} 1, & \text{falls } \hat{\eta}_n(x) > 1/2, \\ 0, & \text{falls } \hat{\eta}_n(x) \leq 1/2, \end{cases}$$

betrachten. Dann gilt (siehe [4, Theorem 1.1]):

4.3 Aufgabe. *Es gilt* $\mathbb{E}\, R(\hat{h}_n) - R^* \leq 2\big(\mathbb{E}\, \|\hat{\eta}_n - \eta\|^2_{L^2(P^X)}\big)^{1/2}$.

Allerdings verfolgt man in der statistischen Lerntheorie den Ansatz nur so viele Annahmen an die Verteilung von (X, Y) zu machen, wie für das vorliegende Problem (hier Klassifikation) notwendig sind. Insbesondere wollen wir keine parametrischen Annahmen machen. Da das Schätzen der Regressionsfunktion schwieriger zu sein scheint als das Klassifikationsproblem wollen wir im Folgenden auch keine Glattheitsannahmen an η machen. Die beiden oben beschriebenen Ansätze reichen also nicht aus.

4.2 Empirische Risikominimierung

4.4 Definition. Für einen Klassifizierer h heißt

$$R_n(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{Y_i \neq h(X_i)}$$

empirischer Klassifizierungsfehler von h. Ist $\mathcal{H}$ eine Menge von Klassifizieren, so heißt ein Klassifizierer $\hat{h}_n$ ERM-Klassifizierer, falls

$$\hat{h}_n \in \operatorname*{argmin}_{h \in \mathcal{H}} R_n(h).$$

4.5 Bemerkungen.

1) Ein Minimum existiert immer, da $R_n(h)$ nur die Werte $0, 1/n, \ldots, 1$ annehmen kann.

2) Empirische Risikominimierung ist ein grundlegendes Prinzip welches wir bereits im Fall der nichparametrischen Regression mit zufälligem Design kennengelernt haben: Setze $R(f) = \mathbb{E}(Y - f(X))^2$ für $f \in L^2(P^X)$. Dann erfüllt die Funktion $f^*(x) = \mathbb{E}(Y|X = x)$ gerade $R(f^*) = \min_{f \in L^2(P^X)} R(f)$. Des Weiteren setze $R_n(h) = (1/n)\sum_{i=1}^n (Y_i - f(X_i))^2$. Ist nun $\mathcal{F}$ eine Klasse von Funktionen so ist $\hat{f}_n \in \operatorname{argmin}_{f \in \mathcal{F}} R_n(f)$ (sofern ein Minimum existiert) gerade der KQS über $\mathcal{F}$. Für das entsprechende Exzessrisiko $R(\hat{f}_n) - R(f^*)$ von $\hat{f}_n$ haben wir im Beweis von Lemma 3.5 gezeigt, dass es mit dem Vorhersagefehler übereinstimmt, d.h. es gilt $R(\hat{f}_n) - R(f^*) = \mathbb{E}\,\|\hat{f}_n - f^*\|^2_{L^2(P^X)}$.

Wie schon im Fall der nichparametrischen Regression spielt die Wahl von $\mathcal{H}$ eine entscheidende Rolle. Das Analogon zur Bias-Varianz-Zerlegung ist:

$$R(\hat{h}_n) - R^* = \underbrace{R(\hat{h}_n) - \inf_{h \in \mathcal{H}} R(h)}_{(1)} + \underbrace{\inf_{h \in \mathcal{H}} R(h) - R^*}_{(2)}.$$

Dabei ist Term (2) ein (deterministischer) Bias-Term. Er misst wie nah die Klasse $\mathcal{H}$ an den Bayes-Klassifizierer herankommt und wird kleiner je größer wir $\mathcal{H}$ wählen. Der Term (1) heißt auch stochastischer Fehler. Er misst den Fehler, den wir machen wenn wir beim Minimieren R durch R_n ersetzen. Es gilt:

4.6 Lemma. *Sei $\mathcal{H}$ eine Menge von Klassifizieren und $\hat{h}_n$ ein zugehöriger ERM-Klassifizierer. Dann gilt*

$$R(\hat{h}_n) - \inf_{h \in \mathcal{H}} R(h) \leq 2 \sup_{h \in \mathcal{H}} |R_n(h) - R(h)|.$$

Beweis. Der Beweis ist analog zum Beweis von Lemma 3.5. Für ein $\epsilon > 0$ sei $h_\epsilon \in \mathcal{H}$ mit $R(h_\epsilon) \leq \inf_{h\in\mathcal{H}} R(h) + \epsilon$. Dann gilt

$$\begin{aligned} R(\hat{h}_n) - \inf_{h\in\mathcal{H}} R(h) &\leq R(\hat{h}_n) - R(h_\epsilon) + \epsilon \\ &= R(\hat{h}_n) - R_n(\hat{h}_n) + R_n(\hat{h}_n) - R(h_\epsilon) + \epsilon \\ &\leq R(\hat{h}_n) - R_n(\hat{h}_n) + R_n(h_\epsilon) - R(h_\epsilon) + \epsilon \\ &\leq 2 \sup_{h\in\mathcal{H}} |R_n(h) - R(h)| + \epsilon, \end{aligned}$$

wobei wir in der zweiten Ungleichung die Definition von $\hat{h}_n$ verwendet haben. Da $\epsilon > 0$ beliebig war, folgt die Behauptung. □

4.3 Vapnik-Chervonenkis-Theorie

Wir wollen im Folgenden obere Schranken für den stochastischen Fehler herleiten. Dabei beschränken wir uns auf den Erwartungswert. Diese Resultate können mittels der Konzentrationsungleichungen aus Kapitel 2 zu Schranken, die mit hoher Wahrscheinlichkeit gelten erweitert werden. Ein erstes, vorläufiges Resultat ist wie folgt:

4.7 Proposition. *Sei $\mathcal{H} = \{h_1, \dots, h_M\}$ eine endliche Menge von Klassifizierern und $\hat{h}_n$ ein zugehöriger ERM-Klassifizierer. Dann gilt die folgende Orakelungleichung*

$$\mathbb{E}\, R(\hat{h}_n) - \min_{h\in\mathcal{H}} R(h) \leq \sqrt{\frac{2\log(2M)}{n}}.$$

Beweis. Aus Lemma 4.6 folgt, dass

$$\mathbb{E}\, R(\hat{h}_n) - \min_{h\in\mathcal{H}} R(h) \leq 2\,\mathbb{E} \max_{j=1,\dots,M} |Z_j| \tag{4.1}$$

mit

$$Z_j = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{Y_i \neq h_j(X_i)} - \mathbb{E}\,\mathbf{1}_{Y_i \neq h_j(X_i)}.$$

Dabei gilt wegen Hoeffdings Lemma $\mathbf{1}_{Y_i\neq h_j(X_i)} - \mathbb{E}\,\mathbf{1}_{Y_i\neq h_j(X_i)} \in \mathrm{SG}(1/4)$, da die Zufallsvariable Werte in einem Intervall der Länge 1 annimmt. Aus Lemma 2.4 folgt, dass $Z_j \in \mathrm{SG}(1/(4n))$ und Aufgabe 2.25 liefert

$$E \max_{j=1,\dots,M} |Z_j| \leq \frac{1}{\sqrt{4n}}\sqrt{2\log(2M)}.$$

Setzen wir dies in (4.1) ein, so folgt die Behauptung. □

Eine typische Wahl für $\mathcal{H}$ im Fall $S = \mathbb{R}^d$ ist die Menge aller affinen Klassifizierer, d.h. $\mathcal{H} = \{h(x) = \mathbf{1}(\langle x, w\rangle - b > 0) : w \in \mathbb{R}^d, \|w\| = 1, b \in \mathbb{R}\}$.

Diese Menge ist allerdings unendlich und obiges Resultat nutzlos. Allerdings ist die Kardinalität von $\mathcal{H}$ nicht die richtige Größe um den stochastischen Fehler in den Griff zu bekommen. Eine (kombinatorische) Größe welche den stochastischen Fehler besser beschreibt ist die folgende:

4.8 Definition. Für eine Menge von Klassifizierern und $n \in \mathbb{N}$ heißt

$$\mathcal{S}_{\mathcal{H}}(n) = \max_{x_1,\dots,x_n \in S} |\{(h(x_1), \dots, h(x_n)) : h \in \mathcal{H}\}|$$

n-ter Shatter-Koeffizient (Zerschmetterungskoeffizient).

Es gilt immer $\mathcal{S}_{\mathcal{H}}(n) \leq 2^n$. Gilt $\mathcal{S}_{\mathcal{H}}(n) = 2^n$, so existieren $x_1, \dots, x_n \in S$ mit $|\{(h(x_1), \dots, h(x_n)) : h \in \mathcal{H}\}| = 2^n$. Man sagt auch, dass $\mathcal{H}$ die Menge $\{x_1, \dots, x_n\}$ zerschmettert. Man kann den Shatter-Koeffizienten auch wie folgt beschreiben: die Menge $\mathcal{H}$ steht in Bijektion zu $\mathcal{A} = \{A \subseteq S : h = \mathbf{1}_A \in \mathcal{H}\}$. Setzen wir nun $\mathcal{A}(x_1, \dots, x_n) = \{A \cap \{x_1, \dots, x_n\} : A \in \mathcal{A}\}$ für $x_1, \dots, x_n \in S$, so gilt $\mathcal{S}_{\mathcal{H}}(n) = \max_{x_1,\dots,x_n \in S} |\mathcal{A}(x_1, \dots, x_n)|$.

4.9 Aufgabe. *Sei $S = \mathbb{R}$ und seien $x_1, \dots, x_n \in S$ paarweise verschieden.*

(a) Ist $\mathcal{H} = \{\mathbf{1}_{(a,\infty)} : a \in \mathbb{R}\}$, so gilt $\mathcal{S}_{\mathcal{H}}(n) = n + 1$.

(b) Ist $\mathcal{H} = \{\mathbf{1}_{(a,b]} : a < b\}$, so gilt $\mathcal{S}_{\mathcal{H}}(n) = \binom{n+1}{2}$.

4.10 Satz. *Sei $\mathcal{H}$ eine Menge von Klassifizieren und $\hat{h}_n$ ein zugehöriger ERM-Klassifizierer. Dann gilt*

$$\mathbb{E}\, R(\hat{h}_n) - \inf_{h \in \mathcal{H}} R(h) \leq 4\sqrt{\frac{2\log(2\mathcal{S}_{\mathcal{H}}(n))}{n}}.$$

Beweis. Verwenden wir Lemma 4.6 und den Symmetrisierungstrick, so folgt

$$\begin{aligned}\mathbb{E}\, R(\hat{h}_n) - \inf_{h \in \mathcal{H}} R(h) &\leq 2\,\mathbb{E} \sup_{h \in \mathcal{H}} \Big| \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{Y_i \neq h(X_i)} - \mathbb{E}\, \mathbf{1}_{Y_i \neq h(X_i)} \Big| \\ &\leq 4\,\mathbb{E} \sup_{h \in \mathcal{H}} \Big| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \mathbf{1}_{Y_i \neq h(X_i)} \Big|,\end{aligned}$$

wobei $\epsilon_1, \dots, \epsilon_n$ i.i.d. Rademacher Zufallsvariablen sind. Wir schließen, dass

$$\mathbb{E}\, R(\hat{h}_n) - \inf_{h \in \mathcal{H}} R(h) \leq 4 \max_{y \in \{0,1\}^n} \max_{x \in \mathbb{R}^n} \mathbb{E} \sup_{h \in \mathcal{H}} \Big| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \mathbf{1}_{y_i \neq h(x_i)} \Big|.$$

Für $y \in \{0,1\}^n$ und $x \in \mathbb{R}^n$ setze

$$T_{\mathcal{H}}(x, y) = \{(\mathbf{1}_{y_1 \neq h(x_1)}, \dots, \mathbf{1}_{y_n \neq h(x_n)}) : h \in \mathcal{H}\}.$$

Dann gilt $|T_{\mathcal{H}}(x,y)| \leq \mathcal{S}_{\mathcal{H}}(n)$, da $T_{\mathcal{H}}(x,y)$ in Bijektion zu $\{(h(x_1),\dots,h(x_n)) : h \in \mathcal{H}\}$ steht und es folgt

$$\mathbb{E}\, R(\hat{h}_n) - \inf_{h\in\mathcal{H}} R(h) \leq 4 \max_{y\in\{0,1\}^n} \max_{x\in\mathbb{R}^n} \sqrt{\frac{1}{n}}\sqrt{2\log(2\mathcal{S}_{\mathcal{H}}(n))} = 4\sqrt{\frac{2\log(2\mathcal{S}_{\mathcal{H}}(n))}{n}},$$

wobei wir Aufgabe 2.25 und die Tatsache, dass $(1/n)\sum_{i=1}^n \epsilon_i \mathbf{1}_{y_i \neq h_j(x_i)} \in \mathrm{SG}(1/n)$, verwendet haben. □

4.11 Definition. Sei $\mathcal{H}$ eine Menge von Klassifizieren. Die Vapnik-Chervonenkis-Dimension (VC-Dimension) ist definiert durch

$$\mathcal{V}(\mathcal{H}) = \max\{n \in \mathbb{N}\cup\{\infty\} : \mathcal{S}_{\mathcal{H}}(n) = 2^n\}.$$

Ist die Menge leer, so setzen wir $\mathcal{V}(\mathcal{H}) = 0$.

4.12 Aufgabe. *Sei $S = \mathbb{R}^d$.*

(a) Ist $\mathcal{H} = \{h(x) = \mathbf{1}(\langle x, w\rangle > 0) : w \in \mathbb{R}^d\}$, so gilt $\mathcal{V} = d$.

(b) Ist $\mathcal{H} = \{h(x) = \mathbf{1}(\langle x, w\rangle - b > 0) : w \in \mathbb{R}^d, b \in \mathbb{R}\}$, so gilt $\mathcal{V} = d+1$.

4.13 Lemma (Sauers Lemma)**.** *Sei $\mathcal{H}$ eine Menge von Klassifizieren mit $\mathcal{V} = \mathcal{V}(\mathcal{H}) < \infty$. Dann gilt für alle $n \geq 1$*

$$\mathcal{S}_{\mathcal{H}}(n) \leq \sum_{j=0}^{\mathcal{V}} \binom{n}{j} \leq (n+1)^{\mathcal{V}}.$$

4.14 Bemerkung. Es gilt also entweder $\mathcal{S}_{\mathcal{H}}(n) = 2^n$ für alle $n \geq 1$ oder $\mathcal{S}_{\mathcal{H}}(n)$ wächst höchstens polynomiell in n.

4.15 Aufgabe. *Es gilt sogar*

$$\sum_{j=0}^{\mathcal{V}} \binom{n}{j} \leq \left(\frac{ne}{\mathcal{V}}\right)^{\mathcal{V}}.$$

Beweis. Die zweite Ungleichung folgt aus

$$\sum_{j=0}^{\mathcal{V}} \binom{n}{j} \leq \sum_{j=0}^{\mathcal{V}} \frac{n^j}{j!} \leq \sum_{j=0}^{\mathcal{V}} \binom{\mathcal{V}}{j} n^j = (n+1)^{\mathcal{V}}.$$

Wir zeigen nun per Induktion nach n, dass

$$\mathcal{S}_{\mathcal{H}}(n) \leq \sum_{j=0}^{V(\mathcal{H})} \binom{n}{j} \tag{4.2}$$

für alle $\mathcal{H}$ mit $\mathcal{V}(\mathcal{H}) < \infty$.

Sei zuerst $n = 1$. Ist $\mathcal{V}(\mathcal{H}) = 0$, so gilt $\mathcal{S}_{\mathcal{H}}(1) = 1 = \binom{1}{0}$. Ist andererseits $\mathcal{V}(\mathcal{H}) \geq 1$, so gilt $\mathcal{S}_{\mathcal{H}}(1) = 2^1 = \binom{1}{0} + \binom{1}{1}$.

Wir nehmen nun an, dass (4.2) für $n-1$ erfüllt ist. Sei $\mathcal{H}$ mit $\mathcal{V}(\mathcal{H}) \geq 1$ (im Fall $\mathcal{V}(\mathcal{H}) = 0$ gilt (4.2) wegen $\mathcal{S}_{\mathcal{H}}(n) = 1$) und $x_1, \ldots, x_n \in S$ beliebig. Setze $\mathcal{H}(x_1, \ldots, x_n) = \{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}\}$. Wir müssen also zeigen, dass

$$|\mathcal{H}(x_1, \ldots, x_n)| \leq \sum_{j=0}^{\mathcal{V}(\mathcal{H})} \binom{n}{j}. \tag{4.3}$$

Sei $\mathcal{F} = \{h|_{\{x_1,\ldots,x_n\}} : h \in \mathcal{H}\}$. Dann gilt $\mathcal{V}(\mathcal{F}) \leq \mathcal{V}(\mathcal{H})$ und die Behauptung folgt, falls wir (4.3) für $\mathcal{F}$ zeigen können. Wir nehmen daher im Folgenden an, dass $S = \{x_1, \ldots, x_n\}$ und $\mathcal{F} = \mathcal{H}$. Wir setzen

$$\mathcal{H}' = \{h \in \mathcal{H} : h(x_n) = 1 \text{ and } h - \mathbf{1}_{\{x_n\}} \in \mathcal{H}\}$$

Dann gilt $\mathcal{H}(x_1, \ldots, x_n) = \mathcal{H}'(x_1, \ldots, x_n) \cup (\mathcal{H} \setminus \mathcal{H}')(x_1, \ldots, x_n)$ und somit

$$|\mathcal{H}(x_1, \ldots, x_n)| \leq |\mathcal{H}'(x_1, \ldots, x_n)| + |(\mathcal{H} \setminus \mathcal{H}')(x_1, \ldots, x_n)|.$$

Wir beschränken beide Terme auf der rechten Seite separat.
1. Beh. Es gilt

$$|\mathcal{H}'(x_1, \ldots, x_n)| \leq \sum_{j=0}^{\mathcal{V}(\mathcal{H})-1} \binom{n-1}{j}.$$

Bew. Es gilt $|\mathcal{H}'(x_1, \ldots, x_n)| = |\mathcal{H}'(x_1, \ldots, x_{n-1})|$ da $h(x_n) = 1$ für alle $h \in \mathcal{H}'$. Außerdem gilt $\mathcal{V}(\mathcal{H}') \leq \mathcal{V}(\mathcal{H}) - 1$. In der Tat sei $\mathcal{V} = \mathcal{V}(\mathcal{H}')$ und $\{x_{i_1}, \ldots, x_{i_\mathcal{V}}\}$ eine Menge die von $\mathcal{H}'$ zerschmettert wird. Dann gilt $x_n \notin \{x_{i_1}, \ldots, x_{i_\mathcal{V}}\}$ da $h(x_n) = 1$ für alle $h \in \mathcal{H}'$. Nach Konstruktion von $\mathcal{H}'$ folgt, dass $\{x_{i_1}, \ldots, x_{i_\mathcal{V}}\} \cup \{x_n\}$ von $\mathcal{H}$ zerschmettert wird und es folgt $\mathcal{V} \leq \mathcal{V}(\mathcal{H}) - 1$. Nach Induktionsvoraussetzung schließen wir

$$|\mathcal{H}'(x_1, \ldots, x_n)| = |\mathcal{H}'(x_1, \ldots, x_{n-1})| \leq \sum_{j=0}^{\mathcal{V}(\mathcal{H})-1} \binom{n-1}{j}.$$

2. Beh. Es gilt

$$|(\mathcal{H} \setminus \mathcal{H}')(x_1, \ldots, x_n)| \leq \sum_{j=0}^{\mathcal{V}(\mathcal{H})} \binom{n-1}{j}.$$

Bew. Es gilt wieder $|(\mathcal{H} \setminus \mathcal{H}')(x_1, \ldots, x_n)| = |(\mathcal{H} \setminus \mathcal{H}')(x_1, \ldots, x_{n-1})|$. In der Tat, sind $h, h' \in \mathcal{H} \setminus \mathcal{H}'$ zwei Klassifizierer mit $h(x_i) = h'(x_i)$ für alle $i = 1, \ldots, n-1$ so folgt, dass auch $h(x_n) = h'(x_n)$ gelten muss, da sonst h oder h' in $\mathcal{H}'$ enthalten wäre. Die Behauptung folgt also wieder aus der

Induktionsvoraussetzung.
Kombinieren wir die 1. und 2. Behauptung, so folgt

$$|\mathcal{H}(x_1,\dots,x_n)| \le \sum_{j=0}^{\mathcal{V}(\mathcal{H})-1} \binom{n-1}{j} + \sum_{j=0}^{\mathcal{V}(\mathcal{H})} \binom{n-1}{j} = \sum_{j=0}^{\mathcal{V}(\mathcal{H})} \binom{n}{j},$$

was zu zeigen war. □

Kombinieren wir Satz 4.10 mit Sauers Lemma, so erhalten wir

4.16 Korollar. *Sei $\mathcal{H}$ eine Menge von Klassifizieren mit $\mathcal{V} = \mathcal{V}(\mathcal{H}) < \infty$. und $\hat{h}_n$ ein zugehöriger ERM-Klassifizierer. Dann gilt*

$$\mathbb{E}\, R(\hat{h}_n) - \inf_{h\in\mathcal{H}} R(h) \le 4\sqrt{\frac{2\mathcal{V}\log(2n+2)}{n}}.$$

4.4 VC-Dimension und Entropie

4.17 Satz. *Sei $\mathcal{H}$ eine Menge von Klassifizierern mit $\mathcal{V} = \mathcal{V}(\mathcal{H}) < \infty$ und Q ein Wahrscheinlichkeitsmaß auf $(S, \mathcal{S})$. Dann gilt*

$$\log M(\mathcal{H}, \|\cdot\|_{L^1(Q)}, \epsilon) \le \mathcal{V}\left(\log\left(\frac{4e}{\epsilon}\right) + \log\log\left(\frac{4e}{\epsilon}\right) + 1\right) \qquad \forall \epsilon \le 1. \quad (4.4)$$

4.18 Bemerkung. Ist $p \ge 1$ und sind $h, h' \in \mathcal{H}$, so gilt $\|h - h'\|^p_{L^p(Q)} = \|h - h'\|_{L^1(Q)}$, da $h - h' \in \{0, \pm 1\}$. Satz 4.17 liefert also auch die allgemeinere Schranke

$$\log M(\mathcal{H}, \|\cdot\|_{L^p(Q)}, \epsilon) \le \mathcal{V}\left(\log\left(\frac{4e}{\epsilon^p}\right) + \log\log\left(\frac{4e}{\epsilon^p}\right) + 1\right) \qquad \forall \epsilon \le 1. \quad (4.5)$$

Beweis. Ist $\mathcal{V} = 0$ so gilt $|\mathcal{H}| = 1$ und die Aussage ist klar. Wir nehmen also im Folgenden an, dass $\mathcal{V} \ge 1$. Sind $h, h' \in \mathcal{H}$ zwei Klassifizierer und sind $A, A' \subseteq S$ so dass $h = \mathbf{1}_A$ und $h' = \mathbf{1}_{A'}$ so gilt $\|h - h'\|_{L^1(Q)} = Q(A \triangle A')$, wobei $A \triangle A'$ die symmetrische Differenz von A und A' ist. Sei $M = M(\mathcal{H}, L^1(Q), \epsilon)$ und seien $h_1, \dots, h_M \in \mathcal{H}$ mit

$$\min_{j\ne k} Q(A_j \triangle A_k) = \min_{j\ne k} \|h_j - h_k\|_{L^1(Q)} > \epsilon.$$

Ohne Einschränkung sei im Folgenden $M \ge 2$ und $\log M \ge \mathcal{V}$, da (4.4) sonst klar ist. Wir wenden die probabilistische Methode an. Seien hierfür $X_1, \dots, X_n$ i.i.d. mit Verteilung Q, wobei n später geeignet gewählt wird.

Dann gilt

$$\begin{aligned}
&\mathbb{P}\left(|\{(h_j(X_1),\ldots,h_j(X_n)) : j \le M\}| < M\right)\\
&= \mathbb{P}\left(\exists j,k \le M : (h_j(X_1),\ldots,h_j(X_n)) = (h_k(X_1),\ldots,h_k(X_n))\right)\\
&\le \binom{M}{2}\mathbb{P}\left((h_j(X_1),\ldots,h_j(X_n)) = (h_k(X_1),\ldots,h_k(X_n))\right)\\
&= \binom{M}{2}\mathbb{P}(\forall i \le n : X_i \notin A_j \triangle A_k)\\
&= \binom{M}{2}(1-\mathbb{P}(X_1 \in A_j \triangle A_k))^n\\
&= \binom{M}{2}(1-Q(A_j \triangle A_k))^n\\
&- \binom{M}{2}(1-\epsilon)^n < M^2\exp(-n\epsilon).
\end{aligned}$$

Wir wählen nun

$$n = \left\lceil \frac{2\log M}{\epsilon} \right\rceil.$$

Dann folgt

$$\mathbb{P}(|\{(h_j(X_1),\ldots,h_j(X_n)) : j \le M\}| < M) < \exp(2\log M - \epsilon n) \le 1,$$

d.h.

$$\mathbb{P}(|\{(h_j(X_1),\ldots,h_j(X_n)) : j \le M\}| = M) > 0.$$

Wir erhalten also

$$M \le \mathcal{S}_n(\mathcal{H}) \le \left(\frac{ne}{\mathcal{V}}\right)^{\mathcal{V}}, \tag{4.6}$$

wobei die zweiten Ungleichung aus Aufgabe 4.15 folgt. Mit $y := \log(M)/\mathcal{V}$ und $x := \log(4e/\epsilon)$ gilt also

$$y = \frac{\log M}{\mathcal{V}} \le \log\left(\frac{ne}{\mathcal{V}}\right) \le \log\left(\frac{4\log M}{\epsilon}\frac{e}{\mathcal{V}}\right) = x + \log(y)$$

Aus $y - \log y \le x$, $x, y \ge 1$, folgt $y \le x + \log x + 1$ da $y \mapsto y - \log y$ monoton wachsend ist auf $[1,\infty)$ und außerdem

$$x+\log x+1-\log(x+\log x+1) = x+1-\log(1+(\log(x)+1)/x) \ge x+1-\log(2) \ge x$$

gilt. Einsetzen liefert

$$\log M \le \mathcal{V}(\log(4e/\epsilon) + \log\log(4e/\epsilon) + 1)$$

und die Behauptung folgt. □

Kombinieren wir Bemerkung 4.18 (angewendet auf das empirischem Maß und $p = 2$) mit Dudleys Entropieschranke, so folgt für alle $y \in \{0,1\}^n$ und $x \in S^n$, dass

$$\mathbb{E} \sup_{h \in \mathcal{H}} \Big| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \mathbf{1}_{y_i \neq h_j(x_i)} \Big| \leq \frac{1}{\sqrt{n}} + \frac{12}{\sqrt{n}} \int_0^{1/2} \sqrt{\log 2M(\mathcal{H}, \|\cdot\|_n, \epsilon)}\, d\epsilon \leq C \sqrt{\frac{\mathcal{V}}{n}}.$$

Setzen wir dies in den Beweis von Satz 4.10 ein, so erhalten wir

$$\mathbb{E}\, R(\hat{h}_n) - \inf_{h \in \mathcal{H}} R(h) \leq C \sqrt{\frac{\mathcal{V}}{n}},$$

d.h. der $\log n$ Faktor aus Korollar 4.16 kann vermieden werden, indem wir Chaining anwenden.

4.5 Modellwahl

Für $m = 1, \ldots, M$ sei $\mathcal{H}_m$ eine Menge von Klassifizierern und $\hat{h}_m$ ein zu $\mathcal{H}_m$ gehöriger ERM-Klassifizierer. Ziel ist es einen Klassifizierer mit möglichst kleinem Klassifizierungsfehler auszuwählen.

4.19 Satz. *Sei*

$$\hat{m} \in \operatorname*{argmin}_{m=1,\ldots,M} R_n(\hat{h}_m) + \operatorname{pen}(m) \quad \textit{mit} \quad \operatorname{pen}(m) \geq 2\sqrt{\frac{2\log(2S_{\mathcal{H}}(n))}{n}}.$$

Dann gilt

$$\mathbb{E}\, R(\hat{h}_{\hat{m}}) \leq \min_{m=1,\ldots,M} \left(\inf_{h \in \mathcal{H}_m} R(h) + 2\operatorname{pen}(m) \right) + C\sqrt{\frac{\log(2M)}{n}}$$

mit einer absoluten Konstanten $C > 0$.

Beweis. Wir setzen

$$\Delta_n(m) = \sup_{h \in \mathcal{H}_m} |R_n(h) - R(h)|, \qquad m = 1, \ldots, M. \tag{4.7}$$

Wir fixieren nun ein $m \in \{1, \ldots, M\}$. Dann gilt

$$\begin{aligned} R(\hat{h}_{\hat{m}}) &= R(\hat{h}_{\hat{m}}) - R_n(\hat{h}_{\hat{m}}) - \operatorname{pen}(\hat{m}) + R_n(\hat{h}_{\hat{m}}) + \operatorname{pen}(\hat{m}) \\ &\leq \max_{m'=1,\ldots,M} \Big(R(\hat{h}_{m'}) - R_n(\hat{h}_{m'}) - \operatorname{pen}(m') \Big) + R_n(\hat{h}_{\hat{m}}) + \operatorname{pen}(\hat{m}) \\ &\leq \max_{m'=1,\ldots,M} \left(\Delta_n(m') - \operatorname{pen}(m') \right) + R_n(\hat{h}_{\hat{m}}) + \operatorname{pen}(\hat{m}). \end{aligned}$$

Außerdem gilt

$$R_n(\hat{h}_{\hat{m}}) + \operatorname{pen}(\hat{m}) \leq R_n(\hat{h}_m) + \operatorname{pen}(m) \leq \Delta_n(m) + \inf_{h \in \mathcal{H}_m} R(h) + \operatorname{pen}(m),$$

wobei die erste Ungleichung aus der Definition von $\hat{m}$ folgt und die zweite Ungleichung wie folgt gesehen werden kann: Für alle $h \in \mathcal{H}_m$ gilt

$$R_n(\hat{h}_m) \leq R_n(h) = R_n(h) - R(h) + R(h) \leq \Delta_n(m) + R(h)$$

und durch Infimumsbildung folgt

$$R_n(\hat{h}_m) \leq \Delta_n(m) + \inf_{h \in \mathcal{H}_m} R(h).$$

Insgesamt erhalten wir also

$$R(\hat{h}_{\hat{m}}) \leq \max_{m'=1,\dots,M} (\Delta_n(m') - \operatorname{pen}(m')) + \Delta_n(m) + \inf_{h \in \mathcal{H}_m} R(h) + \operatorname{pen}(m).$$

Nehmen wir den Erwartungswert und setzen Aufgabe 4.20 (b) und (c) ein, so folgt

$$\mathbb{E}\, R(\hat{h}_{\hat{m}}) \leq \inf_{h \in \mathcal{H}_m} R(h) + 2 \operatorname{pen}(m) + C\sqrt{\frac{\log(2M)}{n}}.$$

Da m beliebig war, folgt die Behauptung. □

4.20 Aufgabe. *Seien* $\operatorname{pen}(m)$ *und* $\Delta_n(m)$ *aus Theorem 4.19 bzw.* (4.7). *Zeige:*

(a) Es gilt

$$\mathbb{P}(|\Delta_n(m) - \mathbb{E}\,\Delta_n(m)| \geq u) \leq 2\exp(-2nu^2) \qquad \forall u \geq 0.$$

(b) Es gilt $\mathbb{E}\,\Delta_n(m) \leq \operatorname{pen}(m)$ *und*

$$\mathbb{P}\Big(\max_{m=1,\dots,M} (\Delta_n(m) - \operatorname{pen}(m)) \geq u\Big) \leq 2M\exp(-2nu^2) \quad u \geq 0.$$

(c) Aus (a) und (b) folgt, dass

$$\mathbb{E} \max_{m=1,\dots,M} (\Delta_n(m) - \operatorname{pen}(m)) \leq C\sqrt{\frac{\log(2M)}{n}}$$

mit einer absoluten Konstanten $C > 0$.

4.6 Der SVM-Klassifizierer

In den letzten Kapiteln haben wir gesehen, dass der ERM-Klassifizierer

$$\hat{h}_n^{ERM} \in \operatorname*{argmin}_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{Y_i \neq h(X_i)}$$

gute statistische Eigenschaften besitzt. Allerdings kann er in der Praxis oft nicht verwendet werden, da er nur schwer zu berechnen ist (sowohl $\mathcal{H}$ als auch

die Indikatorfunktion sind nicht konvex). Ähnlich wie im Fall der Modellwahl (Lasso-Schätzer) ist es unser Ziel eine konvexe Relaxation zu finden.
1. Schritt. Wir nehmen an, dass Y (und somit auch die Klassifizierer h) Werte in $\{\pm 1\}$ annimmt (dies wird durch die Transformation $Y \mapsto 2Y - 1$ erreicht). Dann gilt

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{Y_i \neq h(X_i)} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(-Y_i h(X_i) > 0)}.$$

2. Schritt. Wir ersetzen $\mathcal{H}$ durch eine konvexe Menge von Funktionen $f : S \to \mathbb{R}$. Ein f kann zum Klassifizeren verwendet werden indem wir

$$\operatorname{sign}(f) = \begin{cases} +1, & \text{falls } f(x) > 0, \\ -1, & \text{falls } f(x) \leq 0. \end{cases}$$

betrachten.
3. Schritt. Wir ersetzen $\mathbf{1}_{(z>0)}$ durch eine konvexe Funktion $\varphi(z)$ mit $\varphi(z) \geq \mathbf{1}_{(z>0)}$ für alle $z \in \mathbb{R}$.
Wir betrachten also Klassifizierer, die folgendes lösen

$$\hat{h}_n = \operatorname{sign}(\hat{f}_n) \quad \text{mit} \quad \hat{f}_n \in \frac{1}{n}\sum_{i=1}^{n} \varphi(-Y_i f(X_i)).$$

Dieses Minimierungsproblem kann immer noch als ERM-Problem interpretiert werden indem wir

$$R_\varphi(f) = \mathbb{E}\,\varphi(-Yf(X)) \quad \text{und} \quad R_{n,\varphi} = \frac{1}{n}\sum_{i=1}^{n} \varphi(-Y_i f(X_i))$$

setzen. Wir konzentrieren uns im Folgenden auf die folgenden beliebten Wahlen für $\mathcal{F}$ und φ:

1. $\varphi(z) = (1+z)_+$ hinge loss,
2. $\mathcal{F} = \{f \in \mathcal{F}_k : \|f\|_{\mathcal{F}_k} \leq R\}$ Kugel in einem RKHS $\mathcal{F}_k$ vom Radius R.

Hinge loss

In Analogie zu Satz 4.2 gilt

4.21 Lemma. *Im Fall $\varphi(z) = (1+z)_+$ gilt*

$$\min_{f: S \to \mathbb{R} \text{ messbar}} R_\varphi(f) = R_\varphi(h^*) = 2R^*$$

mit Bayes-Klassifizierer h^* (d.h. $h^*(x) = 1$ falls $\eta(x) = \mathbb{P}(Y = 1 | X = x) > 1/2$ und $h^*(x) = -1$ sonst).

Beweis. Durch Bedingen erhalten wir

$$\begin{aligned}\mathbb{E}(\varphi(-Yf(X))|X=x) &= \varphi(-f(x))\,\mathbb{P}(Y=1|X=x) + \varphi(f(x))\,\mathbb{P}(Y=-1|X=x)\\ &= (1-f(x))_+\eta(x) + (1+f(x))_+(1-\eta(x))\\ &\geq 2(\eta(x)\wedge(1-\eta(x))),\end{aligned}$$

wobei wir verwendet haben, dass $(1+a)_+ + (1-a)_- \geq 1+a+1-a = 2$. Gleichheit gilt, falls $f = h^*$ und die Behauptung folgt, indem wir den Erwartungswert nehmen. Die zweite Gleichheit folgt analog. □

RKHS

Ein Hilbertraum H von Funktionen $f : S \to \mathbb{R}$ heißt RKHS, falls die Punktevaluationen stetige Funktionale sind. Mit Hilfe des Rieszschen Darstellungssatzes existiert dann für alle $x \in S$ ein $k_x \in H$ mit $f(x) = \langle f, k_x\rangle$ $\forall f \in H$. Die k_x können zu einer Funktion $k : S \times S \to \mathbb{R}$ zusammen genommen werden, welche dann gewisse Eigenschaften erfüllt. Wir wählen hier jedoch einen umgekehrten Zugang zu der Theorie der RKHS und starten mit positiv definiten Kernen:

4.22 Definition. Eine Funktion $k : S \times S \to \mathbb{R}$ heißt positiv definiter Kern, falls

(a) für alle $x, y \in S$ gilt $k(x,y) = k(y,x)$,

(b) für alle $m \geq 1$, $x_1, \ldots, x_m \in S$ und $a_1, \ldots, a_m \in \mathbb{R}$ gilt

$$\sum_{j=1}^{m}\sum_{k=1}^{m} a_j a_k k(x_j, x_k) \geq 0.$$

4.23 Beispiele. Sei $S = \mathbb{R}^d$

1. $k(x,y) = \langle x, y\rangle$ linearer Kern,
2. $k(x,y) = x \wedge y$ Histogramm-Kern ($d = 1$),
1. $k(x,y) = \exp(-\|x-y\|^2/(2\sigma^2))$ Gaußscher Kern ($\sigma > 0$).

Wir überlassen es dem Leser zu überprüfen, dass die Eigenschaften aus Definition 4.22 tatsächlich erfüllt sind. Für den Gaußschen Kern kann dabei die folgende Aufgabe verwendet werden:

4.24 Aufgabe. *(a) Sind $k_1, k_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ zwei positiv definite Kerne und $a > 0$, so sind auch ak_1, $k_1 + k_2$ und $k_1 \cdot k_2$ positiv definite Kerne.*

(b) Ist $k_n : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $n \geq 1$, eine Folge von positiv definiten Kernen die punktweise gegen eine Funktion $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ konvergiert, so ist auch k ein positiv definiter Kern.

(c) *Ist $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ ein positiv definiter Kern, so ist auch $k' : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ definiert durch $k'(x, y) = k(x, y)/\sqrt{k(x, x)k(y, y)}$, falls der Zähler ungleich Null ist, und $k'(x, y) = 0$, sonst, ein positiv definiter Kern.*

Es gilt nun:

4.25 Satz. *Sei $k : S \times S \to \mathbb{R}$ ein positiv definiter Kern. Dann existiert genau ein Hilbertraum $\mathcal{F}_k$ von Funktionen $f : S \to \mathbb{R}$ mit*

(a) *$k(x, \cdot) \in \mathcal{F}_k$ für alle $x \in S$,*

(b) *$f(x) = \langle f, k(x, \cdot)\rangle_{\mathcal{F}_k}$ für alle $x \in S$ und alle $f \in \mathcal{F}_k$.*

$\mathcal{F}_k$ heißt auch der zu k gehörige RKHS (reproducing kernel Hilbert space).

Ein Beweis des Satzes ist in Appendix A.1 gegeben.

4.26 Beispiel. Sei $S = \mathbb{R}^d$ und $k(x, y) = \langle x, y\rangle$ der lineare Kern. Dann ist $\mathcal{F}_k$ der Raum aller Linearformen mit $\langle f, g\rangle_{\mathcal{F}_k} = \langle w, v\rangle$ für $f(x) = \langle w, x\rangle$ und $g(x) = \langle v, x\rangle$.

Die Norm $\|f\|_{\mathcal{F}_k}$ ist stark mit Glattheitseigenschaften von f verbunden. Dies wird in den folgenden Beispielen und Aufgaben herausgearbeitet.

4.27 Beispiel. Sei $\phi_1, \ldots, \phi_p$ ein Orthonormalsystem in $L^2([0, 1])$ und

$$k(x, y) = \sum_{k=1}^{p} \mu_k \phi_k(x)\phi_k(y) \text{ für beliebige } \mu_k > 0.$$

Dann gilt $\mathcal{F}_k = \operatorname{span}(\phi_1, \ldots, \phi_p)$ mit

$$\langle f, g\rangle_{\mathcal{F}_k} := \sum_{k=1}^{p} \mu_k^{-1} \langle f, \phi_k\rangle_{L^2} \langle g, \phi_k\rangle_{L^2}.$$

Offensichtlich ist $(\mathcal{F}_k, \langle\cdot, \cdot\rangle_{\mathcal{F}_k})$ als endlichdimensionaler Innenproduktraum ein Hilbertraum. Es bleibt also zu zeigen, dass (a) und (b) gelten. Dabei ist (a) klar und (b) folgt aus

$$\begin{aligned}\langle f, k(x, \cdot)\rangle_{\mathcal{F}_k} &= \sum_{k=1}^{p} \mu_k^{-1} \langle f, \phi_k\rangle_{L^2} \sum_{l=1}^{p} a_l \langle \phi_l, \phi_k\rangle_{L^2} \phi_l(x) \\ &= \sum_{k=1}^{K} \langle f, \phi_k\rangle_{L^2} \phi_k(x) = f(x).\end{aligned}$$

Dieses Beispiel lässt sich auf unendliche Reihenentwicklungen ($p = \infty$) verallgemeinern, sofern $k(x, y)$ wohldefiniert ist. Ein wichtiges Beispiel dafür ist wie folgt (vergleiche mit den periodischen Sobolev-Kugeln aus Bemerkung 1.25):

4.28 Aufgabe. *Sei $S = [0,1]$, $(\phi_k)_{k\geq 1}$ die trigonometrische Basis, $\mu_k = (1+k^2)^{-\alpha}$ für $\alpha > 1/2$ und $k(x,y) = \sum_{k=1}^{\infty} \mu_k \phi_k(x)\phi_k(y)$. Dann gilt*

$$\mathcal{F}_k = \left\{ f \in L^2([0,1]) : \sum_{k=1}^{\infty} \mu_k^{-1} \langle f, \phi_k \rangle_{L^2}^2 < \infty \right\}$$

mit

$$\langle f, g \rangle_{\mathcal{F}_k} = \sum_{k=1}^{\infty} \mu_k^{-1} \langle f, \phi_k \rangle_{L^2} \langle g, \phi_k \rangle_{L^2}, \qquad \forall f, g \in \mathcal{F}_k.$$

4.29 Aufgabe. *Sei $S = [0,1]$ und $k(x,y) = x \wedge y$. Dann gilt*

$$\mathcal{F}_k = \{ f : [0,1] \to \mathbb{R} : f(0) = 0, f \text{ absolut stetig}, f' \in L^2([0,1]) \}$$

mit

$$\langle f, g \rangle_{\mathcal{F}_k} = \int_0^1 f'(x) g'(x)\, dx \qquad \forall f, g \in \mathcal{F}_k.$$

(Hinweis: Verwende den Hauptsatz der Integralrechnung für absolut-stetige Funktionen.)

4.30 Aufgabe. *Sei $S = \mathbb{R}^d$ und $k(x,y) = \exp(-\|x-y\|^2/(2\sigma^2))$, $\sigma > 0$. Dann gilt*

$$\mathcal{F}_k = \{ f \in C_0(\mathbb{R}^d) \cap L^1(\mathbb{R}^d) : \int_{\mathbb{R}^d} |\mathscr{F} f(u)|^2 e^{\sigma^2 \|u\|^2/2}\, du < \infty \}$$

mit

$$\langle f, g \rangle_{\mathcal{F}_k} = \int_{\mathbb{R}^d} \overline{\mathscr{F} f(u)} \mathscr{F} g(u) e^{\sigma^2 \|u\|^2/2}\, du, \qquad \forall f, g \in \mathcal{F}_k.$$

Dabei ist $\mathscr{F} f$ die Fouriertransformierte von f definiert durch

$$\mathscr{F} f(u) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} f(y) e^{-i\langle u, y \rangle}\, dy, \qquad u \in \mathbb{R}^d.$$

(Hinweis: Verwende die Fourierinversionsformel $\mathscr{F}^2 f(x) = f(-x)$ und die Formel $\mathscr{F}\phi = \phi$ für $\phi(x) = \exp(-\|x\|^2/2)$.)

4.31 Definition. Sei $k : S \times S \to \mathbb{R}$ ein positiv definiter Kern, $\mathcal{F}_k$ der zugehörige RKHS und $R > 0$. Dann ist der SVM-Klassifizierer $\hat{h}_n^{SVM}$ definiert durch

$$\hat{h}_n^{SVM} = \operatorname{sign}(\hat{f}_n^{SVM}) \quad \text{mit} \quad \hat{f}_n^{SVM} \in \operatorname*{argmin}_{f \in \mathcal{F}_k : \|f\|_{\mathcal{F}_k} \leq R} \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i f(X_i))_+.$$

4.32 Bemerkung. Alternativ kann man auch

$$\hat{f}_n \in \operatorname*{argmin}_{f \in \mathcal{F}_k} \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i f(X_i))_+ + \lambda \|f\|_{\mathcal{F}_k}^2 \tag{4.8}$$

betrachten.

4.33 Satz (Darsteller-Formel)**.** *Sei $k : S \times S \to \mathbb{R}$ ein positiv definiter Kern, $\mathcal{F}_k$ der zugehörige RKHS und $\lambda > 0$. Dann gilt für $\hat{f}_n$ aus* (4.8)*, dass $\hat{f}_n = \sum_{j=1}^n \hat{\beta}_j k(X_j, \cdot)$ mit*

$$\hat{\beta} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^n} \left(\frac{1}{n} \sum_{i=1}^n \Big(1 - \sum_{j=1}^n \beta_j Y_i k(X_j, X_i)\Big)_+ + \sum_{i,j=1}^n \beta_i \beta_j k(X_i, X_j) \right).$$

Beweis. Sei $V = \operatorname{span}(k(X_1, \cdot), \ldots, k(X_n, \cdot))$ und $V^\perp$ das orthogonale Komplement in $\mathcal{F}_k$. Ist $f \in \mathcal{F}_k$, so schreibe $f = f_V + f_{V^\perp}$ mit $f_V \in V$ und $f_{V^\perp} \in V^\perp$. Dann gilt $\|f\|_{\mathcal{F}_k}^2 = \|f_V\|_{\mathcal{F}_k}^2 + \|f_{V^\perp}\|_{\mathcal{F}_k}^2$ und

$$f(X_i) = \langle f, k(X_i, \cdot)\rangle_{\mathcal{F}_k} = \langle f_V, k(X_i, \cdot)\rangle_{\mathcal{F}_k} = f_V(X_i).$$

Es folgt, dass

$$R_{n,\varphi}(f) + \lambda \|f\|_{\mathcal{F}_k}^2 = R_{n,\varphi}(f_V) + \|f_V\|_{\mathcal{F}_k}^2 + \|f_{V^\perp}\|_{\mathcal{F}_k}^2$$

und somit erfüllt ein Minimierer $\hat{f}_n$, dass $(\hat{f}_n)_{V^\perp} = 0$. Daher gilt

$$\hat{f}_n = \sum_{j=1}^n \hat{\beta}_j k(X_j, \cdot).$$

Dies zweite Behauptung folgt nun indem wir

$$\Big\| \sum_{j=1}^n \beta_j k(X_j, \cdot) \Big\|_{\mathcal{F}_k}^2 = \sum_{i,j=1}^n \beta_i \beta_j \langle k(X_i, \cdot), k(X_j, \cdot)\rangle_{\mathcal{F}_k} = \sum_{i,j=1}^n \beta_i \beta_j k(X_i, X_j)$$

einsetzen. □

4.34 Satz. *Sei $k : S \times S \to \mathbb{R}$ ein positiv definiter Kern, $\mathcal{F}_k$ der zugehörige RKHS, $\varphi(z) = (1 + z)_+$ das hinge loss und $R > 0$. Dann gilt*

$$\mathbb{E}\, R(\hat{h}_n^{SVM}) \le \inf_{\|f\|_{\mathcal{F}_k} \le R} R_\varphi(f) + 8R \sqrt{\frac{\mathbb{E}\, k(X, X)}{n}}$$

Beweis. Es gilt

$$\begin{aligned} R(\hat{h}_n^{SVM}) &= \int \mathbf{1}(-y \hat{h}_n^{SVM}(x) > 0)\, dP^{X,Y}(x, y) \\ &\le \int \mathbf{1}(-y \hat{f}_n^{SVM}(x) \ge 0)\, dP^{X,Y}(x, y) \\ &\le \int (1 - y \hat{f}_n^{SVM}(x))_+\, dP^{X,Y}(x, y) = R_\varphi(\hat{f}_n^{SVM}). \end{aligned}$$

Also

$$\begin{aligned} R(\hat{h}_n^{SVM}) &\le R_\varphi(\hat{f}_n^{SVM}) - \inf_{\|f\|_{\mathcal{F}_k} \le R} R_\varphi(f) + \inf_{\|f\|_{\mathcal{F}_k} \le R} R_\varphi(f) \\ &\le 2 \sup_{\|f\|_{\mathcal{F}_k} \le R} |R_{n,\varphi}(f) - R_\varphi(f)| + \inf_{\|f\|_{\mathcal{F}_k} \le R} R_\varphi(f), \end{aligned}$$

wobei die zweite Ungleichung wie im Beweis von Lemma 4.6 folgt. Es bleibt zu zeigen, dass

$$\mathbb{E} \sup_{\|f\|_{\mathcal{F}_k} \leq R} |R_{n,\varphi}(f) - R_\varphi(f)| \leq 4R\sqrt{\frac{\mathbb{E}\, k(X,X)}{n}}.$$

Mit Hilfe des Symmetrisierungstrick folgt, dass

$$\begin{aligned}
&\mathbb{E} \sup_{\|f\|_{\mathcal{F}_k} \leq R} |R_{n,\varphi}(f) - R_\varphi(f)| \\
&= \mathbb{E} \sup_{\|f\|_{\mathcal{F}_k} \leq R} \Big| \frac{1}{n} \sum_{i=1}^n (1 - Y_i f(X_i))_+ - 1 + 1 - \mathbb{E}(1 - Y_i f(X_i))_+ \Big| \\
&\leq 2\, \mathbb{E} \sup_{\|f\|_{\mathcal{F}_k} \leq R} \Big| \frac{1}{n} \sum_{i=1}^n \epsilon_i \big((1 - Y_i f(X_i))_+ - 1\big) \Big| \\
&= 2\mathbf{E}_{X,Y} \mathbf{E}_\epsilon \sup_{\|f\|_{\mathcal{F}_k} \leq R} \Big| \frac{1}{n} \sum_{i=1}^n \epsilon_i \Psi(-Y_i f(X_i)) \Big|,
\end{aligned}$$

mit $\epsilon_1, \ldots, \epsilon_n$ i.i.d. Rademacher Zufallsvariablen. Die Funktion Ψ ist 1-Lipschitz stetig und erfüllt $\Psi(0) = 0$. Das Konstraktionsprinzip (siehe [2, Theorem 11.6] oder [3, Theorem 3.2.1]) liefert nun

$$\begin{aligned}
&2\mathbf{E}_\epsilon \sup_{\|f\|_{\mathcal{F}_k} \leq R} \Big| \frac{1}{n} \sum_{i=1}^n \epsilon_i \Psi(-Y_i f(X_i)) \Big| \\
&\leq 4\mathbf{E}_\epsilon \sup_{\|f\|_{\mathcal{F}_k} \leq R} \Big| \frac{1}{n} \sum_{i=1}^n -\epsilon_i Y_i f(X_i) \Big| \\
&= 4\mathbf{E}_\epsilon \sup_{\|f\|_{\mathcal{F}_k} \leq R} \Big| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) \Big|,
\end{aligned}$$

wobei wir in der letzten Gleichheit verwendet haben, dass die $-\epsilon_i Y_i$ die gleiche (gemeinsame) Verteilung haben wie die ϵ_i. Wir verwenden nun die Hilbertraumstruktur von $\mathcal{F}_k$. Es gilt

$$\Big| \sum_{i=1}^n \epsilon_i f(X_i) \Big| = \Big| \langle f, \sum_{i=1}^n \epsilon_i k(X_i, \cdot) \rangle_{\mathcal{F}_k} \Big| \leq \|f\|_{\mathcal{F}_k} \Big\| \sum_{i=1}^n \epsilon_i k(X_i, \cdot) \Big\|_{\mathcal{F}_k},$$

wobei wir die Cauchy-Schwarz-Ungleichung angewendet haben, und

$$\Big\| \sum_{i=1}^n \epsilon_i k(X_i, \cdot) \Big\|_{\mathcal{F}_k}^2 = \sum_{i,j=1}^n \epsilon_i \epsilon_j k(X_i, X_j).$$

Wir schließen

$$\begin{aligned}\mathbb{E}\sup_{\|f\|_{\mathcal{F}_k}\le R}|R_{n,\varphi}(f)-R_\varphi(f)| &\le 4\,\mathbb{E}\sup_{\|f\|_{\mathcal{F}_k}\le R}\Big|\frac{1}{n}\sum_{i=1}^n\epsilon_i f(X_i)\Big|\\ &\le \frac{4R}{n}\,\mathbb{E}\Big\|\sum_{i=1}^n\epsilon_i k(X_i,\cdot)\Big\|_{\mathcal{F}_k}\\ &\le \frac{4R}{n}\sqrt{\mathbb{E}\Big\|\sum_{i=1}^n\epsilon_i k(X_i,\cdot)\Big\|^2_{\mathcal{F}_k}}\\ &= \frac{4R}{n}\sqrt{\mathbb{E}\sum_{i,j=1}^n\epsilon_i\epsilon_j k(X_i,X_j)} = 4R\sqrt{\frac{\mathbb{E}\,k(X,X)}{n}}.\end{aligned}$$

□

A Zusätzliche Beweise

A.1 Beweis von Satz 4.25

Wir setzen

$$\mathcal{F}_0=\Big\{f:S\to\mathbb{R}:f=\sum_{i=1}^n a_i k(x_i,\cdot),n\in\mathbb{N},x_i\in S,a_i\in\mathbb{R},i=1,\dots,n\Big\}.$$

Sind $f=\sum_{i=1}^n a_i k(x_i,\cdot)$ und $g=\sum_{j=1}^m b_j k(y_j,\cdot)$ aus $\mathcal{F}_0$, so setzen wir

$$\langle f,g\rangle_{\mathcal{F}_0}=\sum_{i=1}^n\sum_{j=1}^m a_i b_j k(x_i,y_j).$$

A.1 Lemma. *$(\mathcal{F}_0,\langle\cdot,\cdot\rangle_{\mathcal{F}_0})$ ist ein Innenproduktraum mit $f(x)=\langle f,k(x,\cdot)\rangle_{\mathcal{F}_0}$ für alle $x\in S$ und $f\in\mathcal{F}_0$.*

Beweis. Es ist klar, dass $\mathcal{F}_0$ ein Vektorraum ist. Desweiteren ist $\langle\cdot,\cdot\rangle_{\mathcal{F}_0}$ wohldefiniert, da

$$\langle f,g,\rangle\,\mathcal{F}_0=\sum_{i=1}^n\sum_{j=1}^m a_i b_j k(x_i,y_j)=\sum_{i=1}^n a_i g(x_i)=\sum_{j=1}^m b_j f(y_j). \tag{A.1}$$

Hieraus folgt, dass

$$\langle f,k(x,\cdot)\rangle_{\mathcal{F}_0}=1\cdot f(x)=f(x)$$

für alle $x\in S$ und $f\in\mathcal{F}_0$. Es ist klar, dass $\langle\cdot,\cdot\rangle_{\mathcal{F}_0}$ bilinear und symmetrisch ist und dass $\|f\|_{\mathcal{F}_0}=0$ falls $f=0$. Insbesondere gilt die Cauchy-Schwarz-Ungleichung: $|\langle f,g\rangle_{\mathcal{F}_0}|\le\|f\|_{\mathcal{F}_0}\|g\|_{\mathcal{F}_0}$ für alle $f,g\in\mathcal{F}_0$. Setzen wir $g=k(x,\cdot)$, so folgt

$$|f(x)|=|\langle f,k(x,\cdot)\rangle_{\mathcal{F}_0}|\le\|f\|_{\mathcal{F}_0}\|k(x,\cdot)\|_{\mathcal{F}_0}=\sqrt{k(x,x)}\|f\|_{\mathcal{F}_0}. \tag{A.2}$$

Somit impliziert $\|f\|_{\mathcal{F}_0}=0$, dass $f=0$. □

A.2 Lemma. *Sei (f_n) eine Cauchyfolge aus $\mathcal{F}_0$. Dann konvergiert die Folge $(f_n(x))$ für alle $x \in S$.*

Beweis. Aus (A.2) folgt, dass für alle $n, m \in \mathbb{N}$

$$|f_n(x) - f_m(x)| \leq \sqrt{k(x,x)}\|f_n - f_m\|_{\mathcal{F}_0}.$$

Daher ist für beliebiges x die Zahlenfolge $(f_n(x))$ eine Cauchyfolge und die Behauptung folgt aus der Vollständigkeit von $\mathbb{R}$. □

A.3 Lemma. *Ist (f_n) eine Cauchyfolge in $\mathcal{F}_0$ die punktweise gegen 0 konvergiert, so gilt $\|f_n\|_{\mathcal{F}_0} \to 0$ für $n \to \infty$.*

Beweis. Da Cauchyfolgen beschränkt sind, existiert ein $M > 0$ mit $\|f_n\|_{\mathcal{F}_0} \leq M$ für alle $n \geq 1$. Wähle nun N so groß, dass $\|f_n - f_N\|_{\mathcal{F}_0} \leq \epsilon/M$ für alle $n \geq N$. Da $f_N \in \mathcal{F}_0$, gibt es $m \in \mathbb{N}, x_i \in S, a_i \in \mathbb{R}, i = 1, \ldots, m$, mit $f_N = \sum_{i=1}^m k(x_i, \cdot)$. Nun gilt

$$\begin{aligned}\|f_n\|^2_{\mathcal{F}_0} = \langle f_n, f_n\rangle_{\mathcal{F}_0} &= \langle f_n - f_N, f_n\rangle_{\mathcal{F}_0} + \langle f_N, f_n\rangle_{\mathcal{F}_0} \\ &\overset{(A.1)}{=} \langle f_n - f_N, f_n\rangle_{\mathcal{F}_0} + \sum_{i=1}^m a_i f_n(x_i) \\ &\leq \epsilon + \sum_{i=1}^m a_i f_n(x_i).\end{aligned}$$

Daher gilt $\limsup_{n\to\infty} \|f_n\|^2_{\mathcal{F}_0} \leq \epsilon$. Da $\epsilon > 0$ beliebig war, folgt die Behauptung. □

Sei nun $\mathcal{F}_k$ der Raum aller Funktionen $f : S \to \mathbb{R}$ die punktweise Limiten von Cauchyfolgen (f_n) aus $\mathcal{F}_0$ sind. Für $f, g \in \mathcal{F}_k$ setze

$$\langle f, g\rangle_{\mathcal{F}_k} = \lim_{n\to\infty} \langle f_n, g_n\rangle_{\mathcal{F}_0},$$

wobei (f_n) und (g_n) Cauchyfolgen sind die punktweise gegen f und g konvergieren.

A.4 Lemma. *$(\mathcal{F}_k, \langle\cdot,\cdot\rangle_{\mathcal{F}_k})$ ist ein Hilbertraum.*

Beweis. Sind (f_n) und (g_n) zwei Cauchyfolgen in $\mathcal{F}_0$, so ist $(\langle f_n, g_n\rangle)_{\mathcal{F}_0}$ eine Cauchyfolge in $\mathbb{R}$. Diese konvergiert da $\mathbb{R}$ vollständig ist. Seien nun (f'_n) und (g'_n) zwei weitere Cauchyfolgen welche punktweise gegen f und g konvergieren. Dann sind $(f_n - f'_n)$ und $(g_n - g'_n)$ Cauchyfolgen die punktweise gegen 0 konvergieren und Lemma A.3 impliziert $\|f_n - f'_n\|_{\mathcal{F}_0}, \|f_n - f'_n\|_{\mathcal{F}_0} \to 0$ für $n \to 0$. Es folgt, dass

$$|\langle f_n, g_n\rangle_{\mathcal{F}_0} - \langle f'_n, g'_n\rangle_{\mathcal{F}_0}| \leq \|f_n - f'_n\|_{\mathcal{F}_0}\|g_n\|_{\mathcal{F}_0} + \|f'_n\|_{\mathcal{F}_0}\|g_n - g'_n\|_{\mathcal{F}_0} \to 0$$

für $n \to 0$. Daher ist $\langle \cdot, \cdot \rangle_{\mathcal{F}_k}$ wohldefiniert. Es ist nun einfach zu sehen, dass $\langle \cdot, \cdot \rangle_{\mathcal{F}_k}$ ein Skalarprodukt auf $\mathcal{F}_k$ ist welches $\langle \cdot, \cdot \rangle_{\mathcal{F}_0}$ erweitert. Gilt z.B. $\|f\|_{\mathcal{F}_k} = 0$, d.h. $\lim_{n\to 0} \|f_n\|_{\mathcal{F}_0} = 0$ mit (f_n) Cauchyfolge in $\mathcal{F}_0$ die punktweise gegen f konvergiert, so folgt $f(x) = \lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \langle f_n, k(x, \cdot)\rangle \leq \lim_{n\to\infty} \|f_n\|_{\mathcal{F}_0} \|\sqrt{k(x,x)}\| = 0$ und somit $f = 0$. Ist nun $f \in \mathcal{F}_k$ und (f_n) Cauchyfolge in $\mathcal{F}_0$ die punktweise gegen f konvergiert, so folgt aus der Konstruktion des Skalarproduktes, dass $\lim_{n\to\infty} \|f - f_n\|_{\mathcal{F}_k} = 0$. Somit liegt $\mathcal{F}_0$ dicht in $\mathcal{F}_k$. Wir zeigen nun, dass $\mathcal{F}_k$ vollständig ist. Sei hierfür (f_n) eine Cauchyfolge in $\mathcal{F}_k$. Aus der Dichtheit von $\mathcal{F}_0$ folgt, dass es für alle $n \geq 1$ Funktionen $f_n' \in \mathcal{F}_0$ gibt mit $\|f_n - f_n'\| < 1/n$. Dann ist (f_n') eine Cauchyfolge und es folgt aus Lemma A.2, dass diese punktweise gegen ein $f \in v$ konvergiert. Es folgt $\lim_{n\to 0} \|f - f_n'\|_{\mathcal{F}_k} = 0$ und somit $\lim_{n\to 0} \|f - f_n\|_{\mathcal{F}_k} = 0$. □

Es bleibt zu zeigen, dass der so konstruierte Hilbertraum $\mathcal{F}_k$ die Eigenschaften (a) und (b) erfüllt. Eigenschaft (a) ist klar. Um (b) zu zeigen seien $f \in \mathcal{F}_k$, $x \in S$ und (f_n) eine Cauchyfolge in $\mathcal{F}_0$ die punktweise gegen f konvergiert. Dann konvergiert f_n auch in Norm gegen f und es folgt aus der Stetigkeit des Skalarprodukts, dass

$$f(x) = \lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \langle f_n, k(x, \cdot)\rangle = \langle f, k(x, \cdot)\rangle.$$

□

A.5 Bemerkung. Der in Satz 4.25 konstruierte RKHS hat folgende zusätzliche Eigenschaften.

(a) Ist $k(x, \cdot)$ messbar für alle $x \in S$, so ist $\mathcal{F}_k$ ein Raum messbarer Funktionen. (Beweis hierfür: Nach Konstruktion besteht $\mathcal{F}_k$ aus Linearkombinationen von $k(x, \cdot)$ und (gewissen) punktweise Limiten von diesen.)

(b) Ist (S, d) ein metrischer Raum, $k(x, \cdot)$ stetig für alle $x \in S$ und k beschränkt, so ist $\mathcal{F}_k$ ein Raum stetiger Funktionen. (Beweis hierfür: Sei $f \in \mathcal{F}_k$, $x \in S$ und $\epsilon > 0$. Sei (f_n) eine Cauchyfolge in $\mathcal{F}_0$ die punktweise gegen f konvergiert. Dann gibt es ein $m \geq 1$ mit $\|f - f_m\| < \epsilon/(3M)$ (siehe Beweis von Lemma A.4) und $|f(x) - f_m(x)| < \epsilon/3$, wobei $M = \sup_{x\in S} \sqrt{k(x,x)} < \infty$. Da f_m stetig ist, existiert ein $\delta > 0$, so dass $|f_m(x) - f_m(y)| < \epsilon/3$ für alle y mit $d(x, y) < \delta$. Für alle y mit $d(x, y) < \epsilon$ folgt somit

$$\begin{aligned} |f(x) - f(y)| &\leq |f(x) - f_m(x)| + |f_m(x) - f_m(y)| + |f_m(y) - f(y)| \\ &< \epsilon/3 + \epsilon/3 + \sqrt{k(x,x)}\|f - f_m\| < \epsilon \end{aligned}$$

und die Behauptung folgt.)

Ist (S, d) zusätzlich separabel, so ist $\mathcal{F}_k$ separabel. (Beweis hierfür: Sei $S_0 \subseteq S$ dicht und abzählbar. Ist $f \perp \overline{\text{span}}(k(x, \cdot) : x \in S_0)$, so gilt

$f(x) = \langle f, k(x, \cdot)\rangle = 0$ für alle $x \in S_0$. Da f stetig ist folgt $f(x) = 0$ für alle $x \in S$.)

Literatur

[1] R. G. Bartle. *The elements of integration and Lebesgue measure.* John Wiley & Sons, Inc., New York, 1995.

[2] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities. A nonasymptotic theory of independence.* Oxford University Press, Oxford, 2013.

[3] E. Giné and R. Nickl. *Mathematical foundations of infinite-dimensional statistical models.* Cambridge University Press, New York, 2016.

[4] L. Györfi, M. Kohler, A. Krzyźak, and H. Walk. *A distribution-free theory of nonparametric regression.* Springer-Verlag, New York, 2002.

[5] O. Kallenberg. *Foundations of modern probability. Probability and its Applications.* Springer-Verlag, New York, 1997.

[6] P. Massart. *Concentration inequalities and model selection.* Springer, Berlin, 2007.

[7] P-Bühlmann and S. A. van de Geer. *Statistics for high-dimensional data. Methods, theory and applications.* Springer, Heidelberg, 2011.

[8] M. Reiß. *Skript zur VL Nichtparametrische Statistik im WS 12/13.*

[9] A. Tsybakov. *Introduction to nonparametric estimation.* Springer, 2009.

[10] S. A. van de Geer. *Applications of empirical process theory.* Cambridge University Press, Cambridge, 2000.

[11] A. van der Vaart. *Asymptotic statistics.* Cambridge University Press, 1998.

[12] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes. With applications to statistics.* Springer-Verlag, New York, 1996.

[13] L. Wasserman. *All of nonparametric statistics.* Springer, 2006.