Skript zur Vorlesung

Kurvenschätzung

von Prof. Dr. Michael Kohler

Sommersemester 2015

Inhaltsverzeichnis

Kapitel 1

Einführung

In dieser Vorlesung werden die folgenden drei Problemstellungen behandelt:

1. **Dichteschätzung**

   $X_1, X_2, \ldots$ seien unabhängige identisch verteilte (u.i.v.) $\mathbb{R}^d$-wertige Zufallsvariablen mit Dichte $f : \mathbb{R}^d \to \mathbb{R}$ in Bezug auf das Lebesgue-Borel Maß. Ausgehend von
   $$X_1, \ldots, X_n$$
   soll f geschätzt werden.

2. **Regressionsschätzung mit festem Design**

   Sei $m : \mathbb{R}^d \to \mathbb{R}$, seien $x_1, x_2, \cdots \in \mathbb{R}^d$, seien $\epsilon_1, \epsilon_2, \ldots$ unabhängige reelle Zufallsvariablen mit $\mathbf{E}\{\epsilon_i\} = 0$ $(i \in \mathbb{N})$ und
   $$Y_i = m(x_i) + \epsilon_i \quad (i = 1, \ldots, n).$$
   Ausgehend von
   $$(x_1, Y_1), \ldots, (x_n, Y_n)$$
   soll m geschätzt werden.

3. **Regressionsschätzung mit zufälligem Design**

   (X, Y), (X_1, Y_1), (X_2, Y_2), $\ldots$ seien u.i.v. $\mathbb{R}^d \times \mathbb{R}$–wertige Zuvallsvariablen mit $\mathbf{E}\{Y^2\} < \infty$. Sei $m : \mathbb{R}^d \to \mathbb{R}$ definiert durch $m(x) = \mathbf{E}\{Y|X = x\}$ die zugehörige Regressionsfunktion. Ausgehend von
   $$(X_1, Y_1), \ldots, (X_n, Y_n)$$
   soll m geschätzt werden

Die klassische Methode zur Lösung der obigen Schätzprobleme ist die sogenannte **parametrische Kurvenschätzung**. Hierbei wird die Bauart der zu schätzenden Funktion als bekannt vorausgesetzt und es wird angenommen, dass diese nur von endlich vielen Parametern (d.h., endlich vielen reellen Zahlen) abhängt, deren Werte unbekannt sind (z.B. bei der linearen Regression: Regressionfunktion ist eine lineare Funktion). Zur Schätzung der Funktion werden diese dann geschätzt.

Vorteil: Einfach, funktioniert auch, wenn nur wenige Daten zur Verfügung stehen, da statt einer Funktion nur endlich viele reelle Zahlen geschätzt werden müssen.

Nachteil: Eventuell großer Fehler bei der Schätzung, sofern Annahme an die Bauart falsch ist.

In dieser Vorlesung behandeln wir die sogenannte *nichtparametrische Kurvenschätzung*. Dabei ist die Bauart der zu schätzenden Funktion komplett unbekannt.

Schwerpunkte dabei sind:

1. **Universelle Konsistenz**

   Wir konstruieren Schätzverfahren so, dass sie in allen möglichen Situationen gegen die zu schätzende Funktion konvergieren, sofern der Stichprobenumfang gegen Unendlich strebt.

2. **Konvergenzgeschwindigkeit**

   Wir zeigen, dass die Konvergenz in 1. in Abhängigkeit der Situation so schnell wie möglich erfolgt (sofern der Stichprobenumfang gegen Unendlich strebt). Hierbei wird die erreichbare Geschwindigkeit der Konvergenz von der Glattheit der zu schätzenden Funktion abhängen.

3. **Adaption**

   In 2. werden Parameter des Schätzers rein datenabhängig gewählt (und zwar so, dass der gleiche Schätzer in möglichst vielen verschiedenen Situationen die jeweilige optimale Konvergenzgeschwindigkeit erreicht).

Dabei muss insbesondere geklärt werden:

1. Mit welchen Verfahren schätzen wir die Funktion ?
2. Wie wählen wir Parameter rein datenabhängig ?
3. Wie messen wir den Fehler der Schätzung ?

Literatur:

1. Devroye, Lugosi (2001). Combinatorial Methods in Density Estimation.
2. van de Geer (2001). Empirical Process in M-Estimation.
3. Györfi, Kohler, Krzyżak, Walk (2002). A distribution-free theory of nonparametric regression.

Kapitel 2

Dichteschätzung

2.1 Einführung

Sei X eine $\mathbb{R}^d$–wertige Zufallsvariable mit Dichte $f : \mathbb{R}^d \to \mathbb{R}$ (bzgl. dem LB-Maß), d.h. für die Verteilung

$$\mathbf{P}_X : \mathcal{B}_d \to \mathbb{R},\ B \mapsto \mathbf{P}_X(B) = \mathbf{P}[X \in B]$$

gilt

$$\mathbf{P}_X(B) = \int_B f(z)\,dz \quad (B \in \mathcal{B}_d) \tag{2.1}$$

(wobei $\mathcal{B}_d$ die Menge der Borleschen Mengen in $\mathbb{R}^d$ ist).

Gesucht ist eine Funktion $\bar{f} : \mathbb{R}^d \to \mathbb{R}$, mit der wir die Wahrscheinlichkeiten (2.1) approximieren können durch

$$\int_B \bar{f}(z)\,dz \tag{2.2}$$

(und damit dann Rückschlüsse auf $\mathbf{P}_X$ ziehen können).

Ist $\bar{f} : \mathbb{R}^d \to \mathbb{R}$ selbst eine Dichte, d.h. gilt

$$\bar{f}(z) \geq 0 \quad (z \in \mathbb{R}^d) \quad \text{und} \quad \int_{\mathbb{R}^d} \bar{f}(z)\,dz = 1,$$

so wissen wir nach dem **Lemma von Scheffé** (vgl. Vorlesung Mathematische Statistik, WS 14/15), dass gilt

$$\sup_{B \in \mathcal{B}_d} \left| \int_B \bar{f}(z)\,dz - \int_B f(z)\,dz \right| = \int_{\mathbb{R}^d} (f(z) - \bar{f}(z))_+\,dz = \frac{1}{2} \cdot \int_{\mathbb{R}^d} |\bar{f}(z) - f(z)|\,dz,$$

wobei $(u)_+ = \max\{u, 0\}$ $(u \in \mathbb{R})$.

Also sollte die Approximation $\bar{f}$ von f so gewählt werden, dass der sogenannte L_1–**Fehler**

$$\int_{\mathbb{R}^d} |\bar{f}(z) - f(z)|\, dz$$

möglichst klein ist.

Dies führt auf die folgende Aufgabenstellung der Dichteschätzung: Seien X, X_1, X_2, ... unabhängig identisch verteilte $\mathbb{R}^d$-wertige Zufallsvariablen mit Dichte $f : \mathbb{R}^d \to \mathbb{R}$. Ausgehend von

$$X_1, \ldots, X_n$$

soll eine Schätzung

$$f_n(\cdot) = f_n(\cdot, X_1, \ldots, X_n) : \mathbb{R}^d \to \mathbb{R}$$

so konstruiert werden, dass der L_1–Fehler der Schätzung

$$\int |f_n(x) - f(x)|\, dx = \int |f_n(x, X_1, \ldots, X_n) - f(x)|\, dx \tag{2.3}$$

möglichst klein ist.

Im Folgenden werden wir dazu für geeigenete Schätzer f_n Aussagen zum asymptotischen Verhalten von (2.3) herleiten.

Wir betrachten dazu den sogenannten **Kerndichteschätzer** von Rosenblatt und Parzen:

$$f_n(x) = \frac{1}{n \cdot h_n^d} \cdot \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) \tag{2.4}$$

(Motivation siehe Vorlesung Mathematisch Statistik, WS 14/15), wobei $h_n > 0$ die sogenannte **Bandbreite** des Schätzers und $K : \mathbb{R}^d \to \mathbb{R}$ die sogenannte **Kernfunktion** ist. An K setzen wir voraus, dass K messbar ist mit

$$\int_{\mathbb{R}^d} |K(x)|\, dx < \infty \quad \text{und} \quad \int_{\mathbb{R}^d} K(x)\, dx = 1.$$

Ist K sogar eine Dichte, so ist f_n als arithmetisches Mittel der Dichten

$$x \mapsto \frac{1}{h_n^d} \cdot K\left(\frac{x - X_i}{h_n}\right)$$

selbst als Funktion von x eine Dichte.

2.2 Konsistenz

Definition 2.1. *Eine Folge $(f_n)_n$ von Dichteschätzern heißt* **schwach** *bzw.* **stark universell konsistent**, *wenn für jede Dichte $f : \mathbb{R}^d \to \mathbb{R}$ und alle unabhängig identisch verteilten Zufallsvariablen X_1, X_2, ... mit Dichte f gilt:*

$$\int_{\mathbb{R}^d} |f_n(x, X_1, \ldots, X_n) - f(x)|\, dx \to 0 \quad \textit{nach Wk. bzw. f.s.}$$

Für den Kerndichteschätzer gilt:

Satz 2.2. *Ist $K : \mathbb{R}^d \to \mathbb{R}$ eine messbare Funktion mit*

$$\int_{\mathbb{R}^d} |K(x)|\, dx < \infty \quad \textit{und} \quad \int_{\mathbb{R}^d} K(x)\, dx = 1 \tag{2.5}$$

und ist $(h_n)_n$ eine Folge positiver Bandbreiten mit

$$h_n \to 0 \quad (n \to \infty) \quad \textit{und} \quad n \cdot h_n^d \to \infty \quad (n \to \infty), \tag{2.6}$$

so ist die Folge der Kerndichteschätzer

$$f_n(x) = \frac{1}{n \cdot h_n^d} \cdot \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)$$

stark universell konsistent, d.h. es gilt

$$\int_{\mathbb{R}^d} |f_n(x) - f(x)|\, dx \to 0 \quad f.s.$$

für alle Dichten f.

Beweis: Wurde im Spezialfall $K = 1_{[-0.5,0.5]^d}$ bereits in der Vorlesung Mathematische Statistik, WS 14/15 gezeigt. Der allgemeine Beweis erfolgt in den Übungen. □

Bemerkung: Wegen

$$\int_{\mathbb{R}^d} |f_n(x) - f(x)|\, dx \le \int_{\mathbb{R}^d} |f_n(x)|\, dx + \int_{\mathbb{R}^d} |f(x)|\, dx \le \int_{\mathbb{R}^d} |K(z)|\, dz + 1 < \infty$$

gilt in Satz 2.2 auch

$$\mathbf{E} \int_{\mathbb{R}^d} |f_n(x) - f(x)|\, dx \to 0 \quad (n \to \infty)$$

für jede Dichte f.

Bemerkung: Man kann zeigen, dass die Voraussetzung (2.6) nicht nur hinreichend, sondern auch notwendig für die Konsistenz des Kerndichteschätzers ist. Es gilt nämlich: Aus

$$\int_{\mathbb{R}^d} |f_n(x) - f(x)|\, dx \to 0 \quad \text{nach Wk.}$$

für irgendeine Dichte f folgt

$$h_n \to 0 \quad (n \to \infty) \quad \text{und} \quad n \cdot h_n^d \to \infty \quad (n \to \infty).$$

Also ist der Kerndichteschätzer genau dann stark universell konsistent, wenn er für irgendeine Dichte schwach konsistent ist.

2.3 Konvergenzgeschwindigkeit

Im Folgenden leiten wir Aussagen zur Geschwindigkeit her, mit der der erwartete L_1 Fehler

$$\mathbf{E} \int_{\mathbb{R}^d} |f_n(x) - f(x)|\, dx$$

des Kerndichteschätzers f_n im Falle einer "glatten" Dichte $f : \mathbb{R}^d \to \mathbb{R}$ gegen Null konvergiert. Unsere Glattheitsvoraussetzung an f enthält:

Definition 2.3. *Sei $p = k + r$ mit $k \in \mathbb{N}_0$ und $r \in (0, 1]$ (also $p \in (0, \infty)$) und sei $C > 0$. Eine Funktion $f : \mathbb{R}^d \to \mathbb{R}$ heißt (p, C)–***glatt***, falls für alle $k_1, \ldots, k_d \in \mathbb{N}_0$ mit $k_1 + \cdots + k_d = k$ die partielle Ableitung*

$$\frac{\partial^k f}{\partial x_1^{k_1} \ldots \partial x_d^{k_d}}$$

existiert und falls für alle $x, z \in \mathbb{R}^d$ gilt

$$\left| \frac{\partial^k f}{\partial x_1^{k_1} \ldots \partial x_d^{k_d}}(x) - \frac{\partial^k f}{\partial x_1^{k_1} \ldots \partial x_d^{k_d}}(z) \right| \leq C \cdot \|x - z\|^r.$$

Bemerkung. Im Falle von $p \leq 1$ ist eine Funktion (p, C)–glatt genau dann wenn Sie Hölder-stetig ist mit Exponent p und Hölder-Konstante C.

Das Ziel im Folgenden ist zu zeigen, dass im Falle einer (p, C)–glatten Dichte f mit kompaktem Support für den erwarteten L_1–Fehler eines geeignet definierten Kerndichteschätzers f_n gilt:

$$\mathbf{E} \int_{\mathbb{R}^d} |f_n(x) - f(x)|\, dx \leq c \cdot C^{\frac{d}{2 \cdot p + d}} \cdot n^{-\frac{p}{2 \cdot p + d}}, \tag{2.7}$$

falls die Bandbreite in Abhängigkeit von der Glattheit der Dichte gewählt wird durch

$$h_n = \bar{c} \cdot C^{\frac{-2}{2 \cdot p + d}} \cdot n^{-\frac{1}{2 \cdot p + d}}.$$

Bemerkung. Die rechte Seite von (2.7) wird umso kleiner, je

1. grösser n ist (also je mehr Daten zur Verfügung stehen),
2. kleiner d ist (also je kleiner die Dimension ist),
3. kleiner C ist und je größer p ist (also je glatter die Dichte ist).

Als erstes zeigen wir (2.7) im Falle $p \leq 1$.

Satz 2.4. *Seien X, X_1, X_2, ... unabhängig identisch verteilte Zufallsvariablen mit Dichte $f : \mathbb{R}^d \to \mathbb{R}$ bzgl. des LB–Maßes. Sei*

$$f_n(x) = \frac{1}{n \cdot h^d} \cdot \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

der Kerndichteschätzer mit Bandbreite $h > 0$ und Kern $K : \mathbb{R}^d \to \mathbb{R}$, dessen Support $supp(K)$ kompakt sei und für den gelte

$$\int_{\mathbb{R}^d} |K(x)|\, dx < \infty \quad \textit{und} \quad \int_{\mathbb{R}^d} K(x)\, dx = 1.$$

Für die Dichte f gelte:

1. *f ist (p, C)–glatt für ein $0 < p \leq 1$ und ein $C > 0$.*
2. *$supp(f)$ ist kompakt.*

Weiter sei

$$\int_{\mathbb{R}^d} \|z\|^p \cdot |K(z)|\, dz < \infty \quad \textit{und} \quad \int_{\mathbb{R}^d} |K(z)|^2 dz < \infty.$$

Dann gilt

$$\mathbf{E}\int_{\mathbb{R}^d} |f_n(x) - f(x)|\,dx \leq \frac{c_1}{\sqrt{n \cdot h^d}} + c_2 \cdot C \cdot h^p \tag{2.8}$$

für Konstanten $c_1, c_2 > 0$, die nur von $supp(f)$, d, $\int_{\mathbb{R}^d} \|z\|^p \cdot |K(z)|\,dz$ und $\int_{\mathbb{R}^d} |K(z)|^2 dz$ abhängen.

Insbesondere gilt für $h = c_3 \cdot C^{\frac{-2}{2\cdot p+d}} \cdot n^{-\frac{1}{2\cdot p+d}}$:

$$\mathbf{E}\int_{\mathbb{R}^d} |f_n(x) - f(x)|\,dx \leq c_4 \cdot C^{\frac{d}{2\cdot p+d}} \cdot n^{-\frac{p}{2\cdot p+d}}. \tag{2.9}$$

Bemerkung. Minimiert man für $A, B > 0$ die Funktion

$$f(u) = \frac{A}{u^{d/2}} + B \cdot u^p,$$

so gilt für die Minimalstelle

$$0 = f'(u) = -\frac{d}{2} \cdot A \cdot u^{-\frac{d}{2}-1} + B \cdot p \cdot u^{p-1},$$

was auf

$$u = \left(\frac{d}{2p} \cdot \frac{A}{B}\right)^{\frac{2}{2p+d}}$$

führt.

Mit $A = c_1/\sqrt{n}$ und $B = c_2 \cdot C$ folgt, dass die rechte Seite von (2.8) minimal wird für

$$h = c_3 \cdot C^{\frac{-2}{2\cdot p+d}} \cdot n^{-\frac{1}{2\cdot p+d}}.$$

Im Beweis verwenden wir:

Lemma 2.5. *Ist $f : \mathbb{R}^d \to \mathbb{R}$ (p, C)–glatt für ein $0 < p \leq 1$ und ein $C > 0$, und ist $K : \mathbb{R}^d \to \mathbb{R}$ messbar mit*

$$\int_{\mathbb{R}^d} |K(x)|\,dx < \infty \quad \text{und} \quad \int_{\mathbb{R}^d} K(x)\,dx = 1,$$

so gilt für jede kompakte Menge $S \subseteq \mathbb{R}$ und jedes $h > 0$:

$$\int_S \left| \int_{\mathbb{R}^d} \frac{1}{h^d} K\left(\frac{x-z}{h}\right) \cdot f(z)\,dz - f(x) \right| dx \leq \lambda(S) \cdot \int_{\mathbb{R}^d} \|u\|^p \cdot |K(u)|\,du \cdot C \cdot h^p,$$

wobei $\lambda(S)$ das LB–Maß von S ist.

Beweis. Die Voraussetzungen an K und die (p, C)–Glattheit von f implizieren

$$\begin{aligned}
&\int_S \left| \int_{\mathbb{R}^d} \frac{1}{h^d} K\left(\frac{x-z}{h}\right) \cdot f(z)\, dz - f(x) \right| dx \\
&= \int_S \left| \int_{\mathbb{R}^d} \frac{1}{h^d} K\left(\frac{x-z}{h}\right) \cdot (f(z) - f(x))\, dz \right| dx \\
&\leq \int_S \int_{\mathbb{R}^d} \frac{1}{h^d} \left| K\left(\frac{x-z}{h}\right) \right| \cdot |f(z) - f(x)|\, dz\, dx \\
&\leq \int_S \int_{\mathbb{R}^d} \frac{1}{h^d} \left| K\left(\frac{x-z}{h}\right) \right| \cdot C \cdot \|x-z\|^p\, dz\, dx \\
&= \int_S C \cdot h^p \int_{\mathbb{R}^d} |K(u)| \cdot \|u\|^p\, du\, dx = \lambda(S) \cdot C \cdot h^p \int_{\mathbb{R}^d} |K(u)| \cdot \|u\|^p\, du.
\end{aligned}$$

□

Lemma 2.6. *Ist f_n der Kerndichteschätzer aus Satz 2.4, und sind wie in Satz 2.4 die X_1, X_2, ... unabhängig identisch verteilt mit Dichte $f : \mathbb{R}^d \to \mathbb{R}$, so gilt:*

$$\int_{\mathbb{R}^d} \mathbf{Var}(f_n(x))\, dx \leq \int_{\mathbb{R}^d} K^2(u)\, du \cdot \frac{1}{n \cdot h^d}.$$

Beweis. Unter Verwendung der Unabhängigkeit und der identischen Verteiltheit der X_1, X_2, ... und des Satzes von Fubini erhalten wir

$$\begin{aligned}
\int_{\mathbb{R}^d} \mathbf{Var}(f_n(x))\, dx &= \int_{\mathbb{R}^d} \mathbf{Var}\left(\frac{1}{n \cdot h^d} \cdot \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) \right) dx \\
&= \int_{\mathbb{R}^d} \frac{1}{(n \cdot h^d)^2} \cdot n \cdot \mathbf{Var}\left(K\left(\frac{x - X_1}{h}\right) \right) dx \\
&\leq \int_{\mathbb{R}^d} \frac{1}{n \cdot h^d} \cdot \frac{1}{h^d} \cdot \mathbf{E}\left(K^2\left(\frac{x - X_1}{h}\right) \right) dx \\
&= \frac{1}{n \cdot h^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \frac{1}{h^d} \cdot K^2\left(\frac{x-z}{h}\right) \cdot f(z)\, dz\, dx \\
&= \frac{1}{n \cdot h^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \frac{1}{h^d} \cdot K^2\left(\frac{x-z}{h}\right) dx \cdot f(z)\, dz \\
&= \int_{\mathbb{R}^d} K^2(u)\, du \cdot \frac{1}{n \cdot h^d}.
\end{aligned}$$

□

Beweis von Satz 2.4. oBdA $h < 1$ (da linke Seite von (2.8) durch

$$1 + \int_{\mathbb{R}^d} |K(x)|\, dx$$

beschränkt ist).

Da K und f beide kompakten Support haben, existiert eine kompakte Menge $S \subseteq \mathbb{R}^d$ so dass mit Wahrscheinlichkeit Eins gilt:

$$f_n(x) = f(x) = 0 \quad \text{für alle } x \in \mathbb{R}^d \setminus S.$$

Damit, mit der Dreiecksungleichung, der Ungleichung von Cauchy-Schwarz und der Definition von f_n erhalten wir:

$$\begin{aligned}
&\mathbf{E} \int_{\mathbb{R}^d} |f_n(x) - f(x)|\, dx \\
&\leq \int_S \mathbf{E}|f_n(x) - \mathbf{E}(f_n(x))|\, dx + \int_S |\mathbf{E}(f_n(x)) - f(x)|\, dx \\
&\leq \int_S \sqrt{\mathbf{Var}(f_n(x))}\, dx + \int_S \left| \int_{\mathbb{R}^d} \frac{1}{h^d} K\left(\frac{x-z}{h}\right) \cdot f(z)\, dz - f(x) \right| dx \\
&\leq \sqrt{\lambda(S)} \cdot \left(\int_S \mathbf{Var}(f_n(x))\, dx \right)^{1/2} + \int_S \left| \int_{\mathbb{R}^d} \frac{1}{h^d} K\left(\frac{x-z}{h}\right) \cdot f(z)\, dz - f(x) \right| dx.
\end{aligned}$$

Mit Lemma 2.5 und Lemma 2.6 folgt (2.8), und daraus durch Einsetzen auch (2.9). □

Die Konvergenzrate in Satz 2.4 lässt sich bei stärkeren Glattheitsvoraussetzungen an f verbessern. Zur Vereinfachung der Notation zeigen wir das nur für $d = 1$.

Satz 2.7. *In Satz 2.4 sei $d = 1$ und $K : \mathbb{R} \to \mathbb{R}$ ein Kern mit kompakten Support, für den*

$$\int_{\mathbb{R}} |K(x)|\, dx < \infty, \int_{\mathbb{R}} K(x)\, dx = 1 \quad \textit{und} \quad \int_{\mathbb{R}} |K(z)|^2 dz < \infty$$

sowie

$$\int_{\mathbb{R}} u \cdot K(u)\, du = 0 \tag{2.10}$$

gelte. Für die Dichte f gelte:

1. *f ist (p, C)–glatt für ein $1 < p \leq 2$ und ein $C > 0$,*
2. *$\mathrm{supp}(f)$ ist kompakt.*

Dann gilt

$$\mathbf{E} \int_{\mathbb{R}} |f_n(x) - f(x)|\, dx \leq \frac{c_1}{\sqrt{n \cdot h}} + c_2 \cdot C \cdot h^p$$

für Konstanten $c_1, c_2 > 0$, die nur von supp(f), $\int_{\mathbb{R}} |z|^p \cdot |K(z)|\, dz$ und $\int_{\mathbb{R}} |K(z)|^2 dz$ abhängen.

Insbesondere gilt für $h = c_3 \cdot C^{\frac{-2}{2\cdot p+1}} \cdot n^{-\frac{1}{2\cdot p+1}}$:

$$\mathbf{E} \int_{\mathbb{R}} |f_n(x) - f(x)|\, dx \leq c_4 \cdot C^{\frac{1}{2\cdot p+1}} \cdot n^{-\frac{p}{2\cdot p+1}}.$$

Wie Satz 2.4 folgt auch Satz 2.7 aus Lemma 2.6 sowie der folgenden Modifikation von Lemma 2.5.

Lemma 2.8. *Ist $f : \mathbb{R} \to \mathbb{R}$ (p, C)–glatt für ein $1 < p \leq 2$ und ein $C > 0$, und ist $K : \mathbb{R} \to \mathbb{R}$ messbar mit*

$$\int_{\mathbb{R}} \max\{|x|^p, 1\} \cdot |K(x)|\, dx < \infty, \int_{\mathbb{R}} K(x)\, dx = 1 \quad \textit{und} \quad \int_{\mathbb{R}} x \cdot K(x)\, dx = 0,$$

so gilt für jede kompakte Menge $S \subseteq \mathbb{R}$ und jedes $h > 0$:

$$\int_S \left| \int_{\mathbb{R}} \frac{1}{h} K\left(\frac{x-z}{h}\right) \cdot f(z)\, dz - f(x) \right| dx \leq \lambda(S) \cdot \int_{\mathbb{R}} |u|^p \cdot |K(u)|\, du \cdot C \cdot h^p,$$

wobei $\lambda(S)$ das LB–Maß von S ist.

Beweis. Mit $\int_{\mathbb{R}} K(x)\, dx = 1$,

$$\int_{\mathbb{R}} \frac{1}{h} K\left(\frac{x-z}{h}\right) \cdot f'(x) \cdot (z-x)\, dz = f'(x) \cdot (-h) \cdot \int_{\mathbb{R}} u \cdot K(u)\, du = 0$$

und f (p, C)–glatt mit $p = 1 + r$ folgt:

$$\begin{aligned}
&\int_S \left| \int_{\mathbb{R}} \frac{1}{h} K\left(\frac{x-z}{h}\right) \cdot f(z)\, dz - f(x) \right| dx \\
&= \int_S \left| \int_{\mathbb{R}} \frac{1}{h} K\left(\frac{x-z}{h}\right) \cdot \left(f(z) - f'(x) \cdot (z-x) - f(x)\right)\, dz \right| dx \\
&= \int_S \left| \int_{\mathbb{R}} \frac{1}{h} K\left(\frac{x-z}{h}\right) \cdot \int_x^z (f'(u) - f'(x))\, du\, dz \right| dx \\
&\leq \int_S \int_{\mathbb{R}} \frac{1}{h} \left| K\left(\frac{x-z}{h}\right) \right| \cdot \int_{\min\{x,z\}}^{\max\{x,z\}} |f'(u) - f'(x)|\, du\, dz\, dx \\
&\leq \int_S \int_{\mathbb{R}} \frac{1}{h} \left| K\left(\frac{x-z}{h}\right) \right| \cdot \int_{\min\{x,z\}}^{\max\{x,z\}} C \cdot |u-x|^r\, du\, dz\, dx \\
&\leq \int_S \int_{\mathbb{R}} \frac{1}{h} \left| K\left(\frac{x-z}{h}\right) \right| \cdot C \cdot |x-z|^{1+r}\, dz\, dx
\end{aligned}$$

$$= C \cdot h^p \cdot \int_S \int_{\mathbb{R}} \frac{1}{h} \left| K\left(\frac{x-z}{h}\right)\right| \cdot \left(\frac{|x-z|}{h}\right)^p dz\, dx$$
$$= C \cdot h^p \cdot \lambda(S) \cdot \int_{\mathbb{R}} |K(u)| \cdot |u|^p du.$$

□

Bemerkung. a) Die Voraussetzung

$$\int_{\mathbb{R}} x \cdot K(x)\, dx = 0$$

ist insbesondere für alle Kerne $K : \mathbb{R} \to \mathbb{R}$ erfüllt mit

$$K(u) = K(-u) \quad (u \in \mathbb{R}).$$

b) Ist in Satz 2.7 f (p, C)–glatt für ein $p > 2$, und gilt

$$\int_{\mathbb{R}} u^l \cdot K(u)\, du = 0 \quad \text{für alle } l \in \mathbb{N}_0 \text{ mit } l < p, \tag{2.11}$$

so lässt sich in Satz 2.7 ebenfalls die Rate

$$n^{-\frac{p}{2p+1}}$$

zeigen, die für p groß dann beliebig nahe an die parametrische Rate $n^{-1/2}$ kommt. Allerdings kann (2.11) für $p > 2$ nicht gelten, sofern K nichtnegativ ist.

Im Folgenden zeigen wir, dass unter den Voraussetzungen der Sätze 2.4 und 2.7 im Allgemeinen keine besseren Konvergenzraten als dort angegeben hergeleitet werden können. Dazu betrachten wir sogenannte **Minimax-Konvergenzraten**: Für eine vorgegebene Klasse $\mathcal{F}$ von Dichten betrachten wir das Problem, einen Schätzer $f_n(\cdot) = f_n(\cdot, X_1, \ldots, X_n)$ so zu konstruieren, dass

$$\sup_{f \in \mathcal{F}} \mathbf{E} \int_{\mathbb{R}^d} |f_n(x) - f(x)|\, dx$$

"nahe am" optimalen Wert

$$\inf_{g_n} \sup_{f \in \mathcal{F}} \mathbf{E} \int_{\mathbb{R}^d} |g_n(x) - f(x)|\, dx.$$

Hierbei werden innerhalb des Erwartungswertes immer u.i.v. Zufallsvariablen X_1, $\ldots$, X_n mit Dichte f verwendet, und das Infimum oben wird über alle Schätzer $g_n(\cdot) = g_n(\cdot, X_1, \ldots, X_n)$ gebildet.

Als Menge $\mathcal{F}$ der Dichten betrachten wir die Menge aller Dichten, die (p, C)–glatt sind und einen vorgegebenen kompakten Support haben (den wir oBdA als Teilmenge von $[0,1]^d$ ansetzen).

Definition 2.9. *Seien $p, C > 0$. Dann bezeichne $\mathcal{F}^{(p,C)}$ die Menge aller Dichten $f : \mathbb{R}^d \to \mathbb{R}$ für die gilt:*

1. *f ist (p, C)–glatt.*
2. *$supp(f) \subseteq [0,1]^d$.*

Aus den Sätzen 2.4 und 2.7 folgt, dass für einen Kerndichteschätzer mit geeignet gewählter Bandbreite und geeignet gewähltem Kern gilt:

$$\sup_{f \in \mathcal{F}^{(p,C)}} \mathbf{E} \int_{\mathbb{R}^d} |f_n(x) - f(x)|\, dx \leq c \cdot n^{-\frac{p}{2p+d}}.$$

Wie unser nächster Satz zeigt, kann diese obere Schranke durch Wahl des Dichteschätzers höchstens um eine Konstante verbessert werden.

Satz 2.10. *Seien $p, C > 0$ und $\mathcal{F}^{(p,C)}$ die in Definition 2.9 eingeführte Menge von (p, C)–glatten Dichten mit kompaktem Support. Dann existiert eine nur von p, C und d abhängende Konstante $c > 0$ derart, dass für hinreichend große n und C gilt:*

$$\inf_{g_n} \sup_{f \in \mathcal{F}} \mathbf{E} \int_{\mathbb{R}^d} |g_n(x) - f(x)|\, dx \geq c \cdot n^{-\frac{p}{2p+d}}.$$

Hierbei werden innerhalb des Erwartungswertes u.i.v. Zufallsvariablen $X_1, \ldots, X_n$ mit Dichte f betrachtet, und das Infimum oben wird über alle Schätzer $g_n(\cdot) = g_n(\cdot, X_1, \ldots, X_n)$ gebildet.

Der Beweis beruht auf:

Satz 2.11. *Sei $A_1, \ldots A_r$ eine Partition von $[0,1]^d$. Für $1 \leq i \leq r$ seien Funktionen $g_{i,0}, g_{i,1} : A_i \to \mathbb{R}_+$ so gegeben, dass für jedes $\theta = (\theta^{(1)}, \ldots, \theta^{(r)}) \in \{0,1\}^r$ die Funktion $f_\theta : \mathbb{R}^d \to \mathbb{R}$ definiert durch*

$$f_\theta(x) = \begin{cases} g_{i,\theta^{(i)}}(x), & \text{falls } x \in A_i, \\ 0, & \text{falls } x \notin [0,1]^d \end{cases}$$

eine Dichte ist. Für $\theta = (\theta^{(1)}, \ldots, \theta^{(r)}) \in \{0,1\}^r$ sei

$$\bar{\theta}_i = (\theta^{(1)}, \ldots, \theta^{(i-1)}, 1 - \theta^{(i)}, \theta^{(i+1)}, \ldots, \theta^{(r)}).$$

Es gelte

$$\inf_{\theta\in\{0,1\}^r}\inf_{1\le i\le r}\int_{\mathbb{R}^d}|f_\theta(x)-f_{\bar\theta_i}(x)|\,dx\ge\alpha>0$$

und

$$\inf_{\theta\in\{0,1\}^r}\inf_{1\le i\le r}\int_{\mathbb{R}^d}\sqrt{f_\theta(x)\cdot f_{\bar\theta_i}(x)}\,dx\ge\beta>0.$$

Dann gilt für $\mathcal{G}=\{f_\theta\,:\,\theta\in\{0,1\}^r\}$*:*

$$\inf_{g_n}\sup_{f\in\mathcal{G}}\mathbf{E}\int_{\mathbb{R}^d}|g_n(x)-f(x)|\,dx\ge\frac{1}{4}\cdot r\cdot\alpha\cdot\beta^{2n}.$$

Beweis. Seien $\Theta^{(1)},\ldots,\Theta^{(r)}$ unabhängige auf $\{0,1\}$ gleichverteilte Zufallsvariablen, und sei

$$\Theta=(\Theta^{(1)},\ldots,\Theta^{(r)}).$$

In den folgenden Abschätzungen werden für gegebenen Wert von θ bzw. Θ die Zufallsvariablen $X_1,\ldots,X_n$ als unabhängig identisch verteilte Zufallsvariablen mit Dichte f_θ bzw. f_Θ definiert.

Da eine zufällige Wahl von $\theta\in\{0,1\}^r$ niemals zu einem grösseren Fehler als die ungünstigste Wahl von θ führt, gilt dann für jeden beliebigen Schätzer g_n:

$$\begin{aligned}
&\sup_{f\in\mathcal{G}}\mathbf{E}\int_{\mathbb{R}^d}|g_n(x)-f(x)|\,dx\\
&=\sup_{\theta\in\{0,1\}^r}\mathbf{E}\int_{\mathbb{R}^d}|g_n(x,X_1,\ldots,X_n)-f_\theta(x)|\,dx\\
&\ge\mathbf{E}\int_{\mathbb{R}^d}|g_n(x,X_1,\ldots,X_n)-f_\Theta(x)|\,dx\\
&=\frac{1}{2^r}\cdot\sum_{\theta\in\{0,1\}^r}\int_{(\mathbb{R}^d)^n}\int_{\mathbb{R}^d}|g_n(x,x_1,\ldots,x_n)-f_\theta(x)|\,dx\cdot\prod_{j=1}^n f_\theta(x_j)\,d(x_1,\ldots,x_n).
\end{aligned}$$

Für $\theta=(\theta^{(1)},\ldots,\theta^{(r)})\in\{0,1\}^r$ setzen wir nun

$$\theta_{i,0}=(\theta^{(1)},\ldots,\theta^{(i-1)},0,\theta^{(i+1)},\ldots,\theta^{(r)})$$

und

$$\theta_{i,1}=(\theta^{(1)},\ldots,\theta^{(i-1)},1,\theta^{(i+1)},\ldots,\theta^{(r)}).$$

Dann kommt für jedes feste $i\in\{1,\ldots,r\}$ in

$$\theta_{i,0},\,\theta_{i,1}\quad(\theta\in\{0,1\}^r)$$

jedes $\theta \in \{0,1\}^r$ genau zweimal vor. Daher erhalten wir weiter:

$$
\begin{aligned}
&\frac{1}{2^r} \cdot \sum_{\theta\in\{0,1\}^r} \int_{(\mathbb{R}^d)^n} \int_{\mathbb{R}^d} |g_n(x,x_1,\ldots,x_n) - f_\theta(x)|\,dx \cdot \prod_{j=1}^n f_\theta(x_j)\,d(x_1,\ldots,x_n) \\
&= \frac{1}{2^{r+1}} \cdot \sum_{\theta\in\{0,1\}^r} \sum_{i=1}^r \Bigg(\int_{(\mathbb{R}^d)^n} \int_{A_i} |g_n(x,x_1,\ldots,x_n) - f_{\theta_{i,0}}(x)|\,dx \\
&\qquad\qquad \cdot \prod_{j=1}^n f_{\theta_{i,0}}(x_j)\,d(x_1,\ldots,x_n) \\
&\quad + \int_{(\mathbb{R}^d)^n} \int_{A_i} |g_n(x,x_1,\ldots,x_n) - f_{\theta_{i,1}}(x)|\,dx \\
&\qquad\qquad \cdot \prod_{j=1}^n f_{\theta_{i,1}}(x_j)\,d(x_1,\ldots,x_n) \Bigg) \\
&\geq \frac{1}{2^{r+1}} \cdot \sum_{\theta\in\{0,1\}^r} \int_{(\mathbb{R}^d)^n} \sum_{i=1}^r \int_{A_i} \Big(|g_n(x,x_1,\ldots,x_n) - f_{\theta_{i,0}}(x)| \\
&\qquad\qquad + |g_n(x,x_1,\ldots,x_n) - f_{\theta_{i,1}}(x)| \Big)\,dx \\
&\qquad\qquad \cdot \min\Big\{ \prod_{j=1}^n f_{\theta_{i,0}}(x_j), \prod_{j=1}^n f_{\theta_{i,1}}(x_j) \Big\} d(x_1,\ldots,x_n) \\
&\geq \frac{1}{2^{r+1}} \cdot \sum_{\theta\in\{0,1\}^r} \int_{(\mathbb{R}^d)^n} \sum_{i=1}^r \int_{A_i} |f_{\theta_{i,0}}(x) - f_{\theta_{i,1}}(x)| \Big)\,dx \\
&\qquad\qquad \cdot \min\Big\{ \prod_{j=1}^n f_{\theta_{i,0}}(x_j), \prod_{j=1}^n f_{\theta_{i,1}}(x_j) \Big\} d(x_1,\ldots,x_n) \\
&= \frac{1}{2^{r+1}} \cdot \sum_{\theta\in\{0,1\}^r} \int_{(\mathbb{R}^d)^n} \sum_{i=1}^r \int_{\mathbb{R}^d} |f_\theta(x) - f_{\bar{\theta}_i}(x)|\,dx \\
&\qquad\qquad \cdot \min\Big\{ \prod_{j=1}^n f_\theta(x_j), \prod_{j=1}^n f_{\bar{\theta}_i}(x_j) \Big\} d(x_1,\ldots,x_n) \\
&\geq \frac{r\cdot\alpha}{2^{r+1}} \cdot \sum_{\theta\in\{0,1\}^r} \inf_{1\leq i\leq r} \int_{(\mathbb{R}^d)^n} \min\Big\{ \prod_{j=1}^n f_\theta(x_j), \prod_{j=1}^n f_{\bar{\theta}_i}(x_j) \Big\} d(x_1,\ldots,x_n),
\end{aligned}
$$

da nach Voraussetzung gilt:

$$\inf_{\theta\in\{0,1\}^r} \inf_{1\leq i\leq r} \int_{\mathbb{R}^d} |f_\theta(x) - f_{\bar{\theta}_i}(x)|\,dx \geq \alpha.$$

Wir zeigen in Lemma 2.12 unten, dass für Dichten $f, g : \mathbb{R}^l \to \mathbb{R}$ gilt

$$\int_{\mathbb{R}^l} \min(f(x), g(x))\, dx \geq \frac{1}{2} \left(\int_{\mathbb{R}^l} \sqrt{f(x) \cdot g(x)}\, dx \right)^2,$$

womit mit dem Satz von Fubini und der zweiten Voraussetzung des Satzes folgt

$$\begin{aligned}
&\int_{(\mathbb{R}^d)^n} \min\Big\{ \prod_{j=1}^n f_\theta(x_j), \prod_{j=1}^n f_{\bar{\theta}_i}(x_j) \Big\}\, d(x_1, \ldots, x_n) \\
&\geq \frac{1}{2} \left(\int_{(\mathbb{R}^d)^n} \sqrt{\prod_{j=1}^n f_\theta(x_j) \cdot \prod_{j=1}^n f_{\bar{\theta}_i}(x_j)} \right) d(x_1, \ldots, x_n) \\
&= \frac{1}{2} \left(\prod_{j=1}^n \int_{\mathbb{R}^d} \sqrt{f_\theta(x) \cdot f_{\bar{\theta}_i}(x)}\, dx \right)^2 \\
&\geq \frac{1}{2} \left(\prod_{j=1}^n \beta \right)^2 = \frac{\beta^{2n}}{2}.
\end{aligned}$$

Insgesamt erhalten wir daraus

$$\sup_{f \in \mathcal{G}} \mathbf{E} \int_{\mathbb{R}^d} |g_n(x) - f(x)|\, dx \geq \frac{r \cdot \alpha}{2^{r+1}} \cdot \sum_{\theta \in \{0,1\}^r} \inf_{1 \leq i \leq r} \frac{\beta^{2n}}{2}. = \frac{1}{4} \cdot r \cdot \alpha \cdot \beta^{2n},$$

was zu zeigen war. □

Im Beweis haben wir verwendet:

Lemma 2.12. *Sind $f, g : \mathbb{R}^l \to \mathbb{R}$ Dichten, so gilt*

$$\int_{\mathbb{R}^l} \min(f(x), g(x))\, dx \geq \frac{1}{2} \left(\int_{\mathbb{R}^l} \sqrt{f(x) \cdot g(x)}\, dx \right)^2.$$

Beweis. Da mit f und g auch $(f+g)/2$ eine Dichte ist, gilt

$$\begin{aligned}
&2 - 2 \cdot \int_{\mathbb{R}^l} \min\{f(x), g(x)\}\, dx \\
&= 2 \cdot \int_{\mathbb{R}^l} \left(\frac{f(x) + g(x)}{2} - \min\{f(x), g(x)\} \right) dx \\
&= \int_{\mathbb{R}^l} \left(\max\{f(x), g(x)\} - \min\{f(x), g(x)\} \right) dx
\end{aligned}$$

$$
\begin{aligned}
&= \int_{\mathbb{R}^l} |f(x) - g(x)|\, dx \\
&= \int_{\mathbb{R}^l} |\sqrt{f(x)} - \sqrt{g(x)}| \cdot |\sqrt{f(x)} + \sqrt{g(x)}|\, dx.
\end{aligned}
$$

Mit der Ungleichung von Cauchy-Schwarz und f, g Dichten folgt daraus

$$
\begin{aligned}
&\left(2 - 2 \cdot \int_{\mathbb{R}^l} \min\{f(x), g(x)\}\, dx\right)^2 \\
&\leq \int_{\mathbb{R}^l} (\sqrt{f(x)} - \sqrt{g(x)})^2\, dx \cdot \int_{\mathbb{R}^l} (\sqrt{f(x)} + \sqrt{g(x)})^2\, dx \\
&= \left(2 - 2\int_{\mathbb{R}^l} \sqrt{f(x)} \cdot \sqrt{g(x)}\, dx\right) \cdot \left(2 + 2\int_{\mathbb{R}^l} \sqrt{f(x)} \cdot \sqrt{g(x)}\, dx\right) \\
&= 4 - 4\left(\int_{\mathbb{R}^l} \sqrt{f(x)} \cdot \sqrt{g(x)}\, dx\right)^2,
\end{aligned}
$$

also

$$
\begin{aligned}
&4 - 8 \cdot \int_{\mathbb{R}^l} \min\{f(x), g(x)\}\, dx + 4 \cdot \left(\int_{\mathbb{R}^l} \min\{f(x), g(x)\}\, dx\right)^2 \\
&\leq 4 - 4\left(\int_{\mathbb{R}^l} \sqrt{f(x)} \cdot \sqrt{g(x)}\, dx\right)^2.
\end{aligned}
$$

Letzteres ist äquivalent zu

$$
\begin{aligned}
&8 \cdot \int_{\mathbb{R}^l} \min\{f(x), g(x)\}\, dx \\
&\geq 4\left(\int_{\mathbb{R}^l} \sqrt{f(x)} \cdot \sqrt{g(x)}\, dx\right)^2 + 4 \cdot \left(\int_{\mathbb{R}^l} \min\{f(x), g(x)\}\, dx\right)^2 \\
&\geq 4\left(\int_{\mathbb{R}^l} \sqrt{f(x)} \cdot \sqrt{g(x)}\, dx\right)^2,
\end{aligned}
$$

was die Behauptung impliziert. □

Bemerkung. Aus dem Beweis von Satz 2.11 folgt, dass er auch gilt, wenn wir für festes $h : \mathbb{R}^d \to \mathbb{R}_+$ und festes $\emptyset \neq A \subset \mathbb{R}^d$ die Mengen $A_1, \ldots, A_r$ als Partition von A wählen und dann f_θ durch

$$
f_\theta(x) = \begin{cases} g_{i,\theta^{(i)}}(x), & \text{falls } x \in A_i, \\ h(x), & \text{falls } x \notin A \end{cases}
$$

definieren (wobei die Funktionen natürlich wieder so gewählt werden müssen, dass f_θ für alle $\theta \in \{0,1\}^r$ eine Dichte ist).

Beweis von Satz 2.10. Sei $p = k+s$ mit $k \in \mathbb{N}_0$ und $s \in (0,1]$. Für $M_n \in \mathbb{N}$ setze $r = M_n^d$ und "partitioniere" $[\frac{1}{4}, \frac{3}{4}]^d$ in r Würfel der Seitenlänge $1/(2 \cdot M_n)$. Wähle sodann eine nichtnegative (p, C)–glatte Funktion $g : \mathbb{R}^d \to \mathbb{R}$ mit $supp(g) \subseteq (-1/8, 1/8)^d$, $supp(g)$ Kugel in $\mathbb{R}^d$ und $\int_{\mathbb{R}^d} g(x)\, dx > 0$. Weiter wähle für $i \in \{1, \ldots, r\}$ zwei Punkte $a_{i,0} \in A_i$ und $a_{i,1} \in A_i$ so, dass

$$\left(a_{i,j}^{(1)} - \frac{1}{8M_n}, a_{i,j}^{(1)} + \frac{1}{8M_n}\right) \times \cdots \times \left(a_{i,j}^{(d)} - \frac{1}{8M_n}, a_{i,j}^{(d)} + \frac{1}{8M_n}\right) \quad (j \in \{0,1\})$$

zwei disjunkte Teilmengen von A_i sind.

Wähle $h : \mathbb{R}^d \to \mathbb{R}_+$ mit

1. h ist Dichte,
2. $h(x) = 0$ für $x \notin [0,1]^d$,
3. $h(x) = 1$ für $x \in \left[\frac{1}{4}, \frac{3}{4}\right]^d$,
4. h $(p, C/2)$-glatt.

Dies ist für hinreichend großes C möglich, da wir eine unendlich oft differenzierbare Funktion konstruieren können, die 1., 2. und 3. erfüllt.

Definiere dann $g_{i,0}, g_{i,1} : A_i \to \mathbb{R}$ durch

$$g_{i,0}(x) = 1 + M_n^{-p} \cdot g(M_n \cdot (x - a_{i,0})) - M_n^{-p} \cdot g(M_n \cdot (x - a_{i,1}))$$

und

$$g_{i,1}(x) = 1 - M_n^{-p} \cdot g(M_n \cdot (x - a_{i,0})) + M_n^{-p} \cdot g(M_n \cdot (x - a_{i,1})),$$

und definiere für $\theta = (\theta^{(1)}, \ldots, \theta^{(r)}) \in \{0,1\}^r$ die Funktion $f_\theta : \mathbb{R}^d \to \mathbb{R}$ durch

$$f_\theta(x) = \begin{cases} g_{i,\theta^{(i)}}(x), & \text{falls } x \in A_i, \\ h(x), & \text{falls } x \notin \left[\frac{1}{4}, \frac{3}{4}\right]^d. \end{cases}$$

Der Support von

$$x \mapsto M_n^{-p} \cdot g(M_n \cdot (x - a))$$

ist eine Teilmenge von

$$\left(a^{(1)} - \frac{1}{8M_n}, a^{(1)} + \frac{1}{8M_n}\right) \times \cdots \times \left(a^{(d)} - \frac{1}{8M_n}, a^{(d)} + \frac{1}{8M_n}\right),$$

und daher sind die Supports der Funktionen

$$x \mapsto M_n^{-p} \cdot g(M_n \cdot (x - a_{i,j})) \quad (j \in \{0,1\})$$

disjunkte Teilmengen des Inneren von A_i. Und da $\|g\|_\infty < \infty$ gilt (da g eine glatte Funktion mit kompaktem Support ist), sind alle $g_{i,j}$ und damit auch alle f_θ für genügend großes M_n nichtnegative Funktionen. Da darüberhinaus

$$\int_{\mathbb{R}^d} f_\theta(x)\,dx = \int_{[0,1]^d} h(x)\,dx = 1$$

gilt (was aus

$$\int_{A_i} f_\theta(x)\,dx = \int_{A_i} 1\,dx = \int_{A_i} h(x)\,dx$$

folgt), ist f_θ sogar eine Dichte für alle $\theta \in \{0,1\}^r$.

Wir zeigen im Folgenden:

(I) Für $\bar{C}$ geeignet gewählt ist f_θ (p, C)–glatt.

(II) Für jedes $\theta \in \{0,1\}^r$ und jedes $1 \le i \le r$ gilt:

$$\int_{\mathbb{R}^d} |f_\theta(x) - f_{\bar{\theta}_i}(x)|\,dx = 4 \cdot M_n^{-p-d} \cdot \int_{\mathbb{R}^d} g(x)\,dx =: \alpha,$$

(III) Für jedes $\theta \in \{0,1\}^r$ und jedes $1 \le i \le r$ gilt:

$$\int_{\mathbb{R}^d} \sqrt{f_\theta(x) \cdot f_{\bar{\theta}_i}(x)}\,dx \ge 1 - c \cdot M_n^{-2p-d} =: \beta.$$

Dies impliziert die Behauptung, denn aus (I) bis (III) folgt mit Satz 2.11:

$$\begin{aligned}
&\inf_{g_n} \sup_{f \in \mathcal{F}^{(p,C)}} \mathbf{E} \int_{\mathbb{R}^d} |g_n(x) - f(x)|\,dx \\
&\ge \inf_{g_n} \sup_{\theta \in \{0,1\}^r} \mathbf{E} \int_{\mathbb{R}^d} |g_n(x) - f_\theta(x)|\,dx \\
&\ge \frac{1}{4} \cdot r \cdot \alpha \cdot \beta^{2n} \\
&= \frac{1}{4} \cdot M_n^d \cdot 4 \cdot M_n^{-p-d} \cdot \int_{\mathbb{R}^d} g(x)\,dx \cdot \left(1 - c \cdot M_n^{-2p-d}\right)^{2n}.
\end{aligned}$$

Mit

$$M_n = \lceil n^{\frac{1}{2p+d}} \rceil$$

gilt

$$\left(1 - c \cdot M_n^{-2p-d}\right)^{2n} \to \exp\left(-c^2\right) \quad (n \to \infty),$$

und folglich ist

$$\inf_{g_n} \sup_{f \in \mathcal{F}} \mathbf{E} \int_{\mathbb{R}^d} |g_n(x) - f(x)|\, dx \geq \bar{c} \cdot M_n^{-p} \geq \tilde{c} \cdot n^{-\frac{p}{2p+d}}.$$

Also genügt es im Folgenden, (I) bis (III) zu zeigen.

Nachweis von (I): (I) folgt aus:

1. Sind h_1 und h_2 $(p, C/2)$–glatt, so ist $h_1 + h_2$ eine (p, C)–glatte Funktion (folgt unmittelbar aus der Definition der (p, C)–Glattheit).

2. Für jedes $a \in \mathbb{R}$ ist $h(x) = M_n^{-p} \cdot g(M_n \cdot (x - a))$ ein $(p, \bar{C})$–glatte Funktion, denn da g $(p, \bar{C})$–glatt ist, gilt für alle $k_1, \ldots, k_d \in \mathbb{N}_0$ mit $k_1 + \cdots + k_d = k$:

$$\begin{aligned}
&\left| \frac{\partial^k h}{\partial x_1^{k_1} \ldots \partial x_d^{k_d}}(x) - \frac{\partial^k h}{\partial x_1^{k_1} \ldots \partial x_d^{k_d}}(z) \right| \\
&= \left| M_n^{-p} \cdot M_n^k \cdot \left(\frac{\partial^k g}{\partial x_1^{k_1} \ldots \partial x_d^{k_d}}(M_n \cdot (x - a)) \right) \right. \\
&\qquad\qquad \left. - M_n^{-p} \cdot M_n^k \cdot \left(\frac{\partial^k g}{\partial x_1^{k_1} \ldots \partial x_d^{k_d}}(M_n \cdot (z - a)) \right) \right| \\
&\leq M_n^{-p} \cdot M_n^k \cdot \bar{C} \cdot \|M_n \cdot (x - a) - M_n \cdot (z - a)\|^s \\
&= \bar{C} \cdot \|x - z\|^s.
\end{aligned}$$

3. Sind $h_1, \ldots, h_r$ $(p, \bar{C})$–glatt und sind $supp(h_1)$, …, $supp(h_r)$ disjunkte Kugeln in $\mathbb{R}^d$, so ist

$$h = \sum_{j=1}^{r} h_j$$

$(p, 2^{1-s} \cdot \bar{C})$–glatt.

Dies gilt, da für $x \in supp(h_1)$ und $z \in supp(h_2)$ sowie $\bar{x}$, $\bar{z}$ so auf der Verbindungsstrecke von x und z gewählt, dass $\bar{x}$ bzw. $\bar{z}$ auf dem Rand von $supp(h_1)$ bzw. $supp(h_2)$ liegen, aufgrund der Konkavität der Abbildung $u \mapsto u^s$ (beachte $0 < s \leq 1$) gilt:

$$\left| \frac{\partial^k h}{\partial x_1^{k_1} \ldots \partial x_d^{k_d}}(x) - \frac{\partial^k h}{\partial x_1^{k_1} \ldots \partial x_d^{k_d}}(z) \right|$$

$$
\begin{aligned}
&\leq \left| \frac{\partial^k h_1}{\partial x_1^{k_1} \dots \partial x_d^{k_d}}(x) - \frac{\partial^k h_1}{\partial x_1^{k_1} \dots \partial x_d^{k_d}}(\bar{x}) \right| \\
&\qquad + \left| \frac{\partial^k h_2}{\partial x_1^{k_1} \dots \partial x_d^{k_d}}(z) - \frac{\partial^k h_2}{\partial x_1^{k_1} \dots \partial x_d^{k_d}}(\bar{z}) \right| \\
&\leq \bar{C} \cdot \|x - \bar{x}\|^s + \bar{C} \cdot \|z - \bar{z}\|^s \\
&= 2 \cdot \bar{C} \cdot \left(\frac{1}{2}\|x - \bar{x}\|^s + \frac{1}{2}\|z - \bar{z}\|^s \right) \\
&\leq 2 \cdot \bar{C} \cdot \left(\frac{\|x - \bar{x}\| + \|z - \bar{z}\|}{2} \right)^s \\
&\leq 2^{1-s} \cdot \bar{C} \cdot \|x - z\|^s.
\end{aligned}
$$

Nachweis von (II): Für beliebiges $\theta \in \{0,1\}^r$ und $i \in \{1, \dots, r\}$ gilt:

$$
\begin{aligned}
&\int_{\mathbb{R}^d} |f_\theta(x) - f_{\bar{\theta}_i}(x)| \, dx \\
&= \int_{A_i} |g_{i,0}(x) - g_{i,1}(x)| \, dx \\
&= \int_{A_i} \left(2 \cdot M_n^{-p} \cdot g(M_n \cdot (x - a_{i,0})) + 2 \cdot M_n^{-p} \cdot g(M_n \cdot (x - a_{i,1})) \right) dx \\
&= 4 \cdot M_n^{-p-d} \cdot \int_{\mathbb{R}^d} g(x) \, dx =: \alpha,
\end{aligned}
$$

Nachweis von (III): Für beliebiges $\theta \in \{0,1\}^r$ und $i \in \{1, \dots, r\}$ gilt wegen $f_\theta(x) = f_{\bar{\theta}_i}(x)$ für $x \notin A_i$:

$$
\begin{aligned}
&\int_{\mathbb{R}^d} \sqrt{f_\theta(x) \cdot f_{\bar{\theta}_i}(x)} \, dx \\
&= \int_{\mathbb{R}^d} f_\theta(x) \, dx - \int_{A_i} f_\theta(x) \, dx + \int_{A_i} \sqrt{g_{i,0}(x) \cdot g_{i,1}(x)} \, dx \\
&= 1 - \left(\int_{A_i} 1 \, dx - \int_{A_i} \sqrt{g_{i,0}(x) \cdot g_{i,1}(x)} \, dx \right) \\
&= 1 - \int_{A_i} \left(1 - \sqrt{g_{i,0}(x) \cdot g_{i,1}(x)} \right) dx \\
&= 1 - \int_{A_i} \frac{1 - g_{i,0}(x) \cdot g_{i,1}(x)}{1 + \sqrt{g_{i,0}(x) \cdot g_{i,1}(x)}} \, dx \\
&= 1 - \int_{A_i} \frac{\left(M_n^{-p} \cdot g(M_n \cdot (x - a_{i,0})) - M_n^{-p} \cdot g(M_n \cdot (x - a_{i,1})) \right)^2}{1 + \sqrt{g_{i,0}(x) \cdot g_{i,1}(x)}} \, dx,
\end{aligned}
$$

wobei die letzte Gleichheit aus

$$1 - (1 + a) \cdot (1 - a) = 1 - (1 - a^2) = a^2$$

folgt. Falls wir annehmen, dass $M_n^{-p} \cdot \|g\|_\infty \leq 1/2$ gilt, so folgt

$$1 + \sqrt{g_{i,0}(x) \cdot g_{i,1}(x)} \geq 1 + \frac{1}{2} = \frac{3}{2},$$

und wir erhalten

$$\begin{aligned}
&\int_{\mathbb{R}^d} \sqrt{f_\theta(x) \cdot f_{\bar{\theta}_i}(x)}\, dx \\
&\geq 1 - \frac{2}{3} \cdot \int_{A_i} \left(M_n^{-p} \cdot g(M_n \cdot (x - a_{i,0})) - M_n^{-p} \cdot g(M_n \cdot (x - a_{i,1})) \right)^2 dx \\
&= 1 - \frac{2}{3} \cdot \int_{A_i} M_n^{-2p} \cdot g^2(M_n \cdot (x - a_{i,0})) + M_n^{-2p} \cdot g^2(M_n \cdot (x - a_{i,1}))\, dx \\
&= 1 - \frac{4}{3} \cdot M_n^{-2p-d} \cdot \int_{\mathbb{R}^d} g^2(x)\, dx =: \beta.
\end{aligned}$$

□

2.4 Adaption

Wie wir in den Sätzen 2.4 und 2.7 gesehen haben, erreicht der Kerndichteschätzer die gemäß Satz 2.10 optimale Konvergenzgeschwindigkeit, sofern die Bandbreite in Abhängigkeit der Glattheit der zu schätzenden Dichte geeignet gewählt wird. In einer Anwendung ist diese aber unbekannt, so dass diese Wahl der Bandbreite nicht möglich ist.

Im Folgenden stellen wir Verfahren vor, die versuchen, die Bandbreite h des Kerndichteschätzers so zu wählen, dass gilt

$$Fehler(f_{n,\hat{h}}) \approx \min_h Fehler(f_{n,h}).$$

2.4.1 L_2–Kreuzvalidierung

Ziel der L_2–Kreuzvalidierung ist die Minimierung des erwarteten L_2–Fehlers

$$\mathbf{E} \int_{\mathbb{R}^d} |f_{n,h}(x) - f(x)|^2 dx$$

bzgl. $h > 0$.

Äquivalent dazu ist die Minimierung von

$$\begin{aligned} & T(h) \\ &= \mathbf{E} \int_{\mathbb{R}^d} |f_{n,h}(x)|^2 dx - 2 \cdot \mathbf{E} \int_{\mathbb{R}^d} f_{n,h}(x) \cdot f(x)\, dx \\ &= \mathbf{E} \int_{\mathbb{R}^d} |f_{n,h}(x)|^2 dx - 2 \cdot \int_{\mathbb{R}^d} \mathbf{E} \left\{ \frac{1}{h^d} \cdot K\left(\frac{x - X_1}{h} \right) \right\} \cdot f(x)\, dx. \end{aligned} \tag{2.12}$$

Dazu wird (2.12) geschätzt durch

$$\hat{T}_n(h) = \int_{\mathbb{R}^d} f_{n,h}(x)^2 dx - 2 \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} f_{n,h}^{(i)}(X_i), \tag{2.13}$$

wobei

$$f_{n,h}(x) = \frac{1}{n \cdot h^d} \cdot \sum_{i=1}^{n} K\left(\frac{x - X_i}{h} \right)$$

und

$$f_{n,h}^{(i)}(x) = \frac{1}{(n-1) \cdot h^d} \cdot \sum_{j=1,\ldots,n, j \neq i} K\left(\frac{x - X_j}{h} \right).$$

Lemma 2.13. *$\hat{T}_n(h)$ ist ein erwartungstreuer Schätzer von $T(h)$, d.h., es gilt*

$$\mathbf{E}\{\hat{T}_n(h)\} = T(h)$$

für alle $h > 0$.

Beweis. Es genügt zu zeigen:

$$\mathbf{E} \left\{ \frac{1}{n} \cdot \sum_{i=1}^{n} f_{n,h}^{(i)}(X_i) \right\} = \int_{\mathbb{R}^d} \mathbf{E} \left\{ \frac{1}{h^d} \cdot K\left(\frac{x - X_1}{h} \right) \right\} \cdot f(x)\, dx. \tag{2.14}$$

Mit

$$\begin{aligned} li.S. &= \mathbf{E} \left\{ f_{n,h}^{(1)}(X_1) \right\} \\ &= \mathbf{E} \left\{ \frac{1}{(n-1) \cdot h^d} \cdot \sum_{j=2,\ldots,n} K\left(\frac{X_1 - X_j}{h} \right) \right\} \\ &= \frac{1}{h^d} \cdot \mathbf{E} \left\{ K\left(\frac{X_1 - X_2}{h} \right) \right\} = \frac{1}{h^d} \cdot \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} K\left(\frac{x - z}{h} \right) \cdot f(x) \cdot f(z)\, dx\, dz \end{aligned}$$

und

$$\begin{aligned} re.S. &= \int_{\mathbb{R}^d} \frac{1}{h^d} \int_{\mathbb{R}^d} \cdot K\left(\frac{x-z}{h}\right) \cdot f(z)\, dz \cdot f(x)\, dx \\ &= \frac{1}{h^d} \cdot \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} K\left(\frac{x-z}{h}\right) \cdot f(x) \cdot f(z)\, dx\, dz \end{aligned}$$

folgt die Behauptung. □

Bei der L_2–Kreuzvalidierung wird nun

$$\hat{h} = \arg\min_{h>0} \hat{T}_n(h)$$

gesetzt (wobei das Minimum oben evt. auch nur über ein endliches Gitter von Bandbreiten gebildet wird), und dann

$$f_{n,\hat{h}}(x) = \frac{1}{n \cdot \hat{h}^d} \cdot \sum_{i=1}^{n} K\left(\frac{x - X_i}{\hat{h}}\right)$$

als Schätzer verwendet.

2.4.2 Die kombinatorische Methode zur Bandbreitenwahl

Statt des L_2–Fehlers versucht die kombinatorische Methode den L_1–Fehler zu minimieren.

Ausgangspunkt ist die Beobachtung, dass nach dem Lemma von Scheffé gilt

$$\begin{aligned} \int_{\mathbb{R}^d} |f_{n,h}(x) - f(x)|\, dx &= 2 \cdot \int_{\mathbb{R}^d} (f_{n,h}(x) - f(x))_+\, dx \\ &= 2 \cdot \int_{\mathbb{R}^d} (f(x) - f_{n,h}(x))_+\, dx, \end{aligned}$$

was impliziert

$$\int_{\mathbb{R}^d} |f_{n,h}(x) - f(x)|\, dx = 2 \cdot \max_{A \in \{[f_{n,h} > f], [f > f_{n,h}]\}} \left| \int_A f_{n,h}(x)\, dx - \int_A f(x)\, dx \right|,$$

wobei für $f, g : \mathbb{R}^d \to \mathbb{R}$

$$[f > g] := \left\{ x \in \mathbb{R}^d \,:\, f(x) > g(x) \right\}$$

gesetzt wird.

Ein erwartungstreuer Schätzer für $\int_A f(x)\,dx$ ist

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} 1_A(X_i). \tag{2.15}$$

Anstelle der in einer Anwendung unbekannten Mengen

$$[f_{n,h} > f] \quad \text{und} \quad [f > f_{n,h}]$$

werden statt dessen Mengen

$$[f_{n,h_1} > f_{n,h_2}] \quad (h_1, h_2 > 0)$$

betrachtet. Zusätzlich wird für (2.15) noch eine Unterteilung der Stichprobe durchgeführt.

Dies führt auf: Gegeben sei eine (endliche) Menge $\mathcal{P} \subseteq (0,\infty)$ von Bandbreiten und eine Zerlegung $n = n_l + n_t$ mit $n_l, n_t \in \mathbb{N}$. Für $h \in \mathcal{P}$ und $A \subseteq \mathbb{R}^d$ setzen wir

$$f_{n_l,h}(x) = \frac{1}{n_l \cdot h^d} \cdot \sum_{i=1}^{n_l} K\left(\frac{x - X_i}{h}\right)$$

und

$$\mu_{n_t}(A) = \frac{1}{n_t}\sum_{i=n_l+1}^{n} 1_A(X_i).$$

Die kombinatorische Methode wählt dann

$$\hat{h} = \arg\min_{h\in\mathcal{P}} \max_{A\in\mathcal{A}} \left| \int_A f_{n_l,h}(x)\,dx - \mu_{n_t}(A) \right|, \tag{2.16}$$

wobei

$$\mathcal{A} = \{[f_{n_l,h_1} > f_{n_l,h_2}] \,:\, h_1, h_2 \in \mathcal{P},\, h_1 \neq h_2\}$$

die Menge der sogenannten Scheffé Mengen ist, und definiert dann den Schätzer der Dichte f durch

$$\hat{f}_n(x) = f_{n_l,\hat{h}}(x). \tag{2.17}$$

Bem.: Wir nehmen hier vereinfachend an, dass $|\mathcal{P}|$ (und damit auch $\mathcal{A}$) endlich ist, so dass die obigen Minima und Maxima in der Tat existieren.

Lemma 2.14. *Ist $|\mathcal{P}| < \infty$, so gilt mit den obigen Bezeichnungen*

$$\int_{\mathbb{R}^d} |\hat{f}_n(x) - f(x)|\,dx \leq 3 \cdot \min_{h\in\mathcal{P}} \int_{\mathbb{R}^d} |f_{n_l,h} - f(x)|\,dx + 4 \cdot \Delta,$$

wobei

$$\Delta = \max_{A\in\mathcal{A}} \left| \int_A f(x)\,dx - \mu_{n_t}(A) \right|.$$

Beweis. Sei $\bar{h} \in \mathcal{P}$ mit

$$\int_{\mathbb{R}^d} |f_{n_l,\bar{h}} - f(x)|\,dx = \min_{h\in\mathcal{P}} \int_{\mathbb{R}^d} |f_{n_l,h} - f(x)|\,dx.$$

Dann gilt

$$\int_{\mathbb{R}^d} |\hat{f}_n(x) - f(x)|\,dx \leq \int_{\mathbb{R}^d} |f_{n_l,\hat{h}} - f_{n_l,\bar{h}}(x)|\,dx + \min_{h\in\mathcal{P}} \int_{\mathbb{R}^d} |f_{n_l,h} - f(x)|\,dx.$$

Mit dem Lemma von Scheffé, der Dreiecksungleichung und der Definition von $\hat{h}$ folgt:

$$\begin{aligned}
&\int_{\mathbb{R}^d} |f_{n_l,\hat{h}} - f_{n_l,\bar{h}}(x)|\,dx \\
&= 2 \cdot \max_{A\in\{[f_{n_l,\hat{h}} > f_{n_l,\bar{h}}(x)],[f_{n_l,\bar{h}}(x) > f_{n_l,\hat{h}}]\}} \left| \int_A f_{n_l,\hat{h}}\,dx - \int_A f_{n_l,\bar{h}}(x)\,dx \right| \\
&\leq 2 \cdot \max_{A\in\mathcal{A}} \left| \int_A f_{n_l,\hat{h}}\,dx - \int_A f_{n_l,\bar{h}}(x)\,dx \right| \\
&\leq 2 \cdot \max_{A\in\mathcal{A}} \left| \int_A f_{n_l,\hat{h}}\,dx - \mu_{n_t}(A) \right| \\
&\qquad +2 \cdot \max_{A\in\mathcal{A}} \left| \int_A f_{n_l,\bar{h}}(x)\,dx - \mu_{n_t}(A) \right| \\
&\leq 4 \cdot \max_{A\in\mathcal{A}} \left| \int_A f_{n_l,\bar{h}}(x)\,dx - \mu_{n_t}(A) \right| \\
&\leq 4 \cdot \max_{A\in\mathcal{A}} \left| \int_A f_{n_l,\bar{h}}(x)\,dx - \int_A f(x)\,dx \right| \\
&\qquad +4 \cdot \max_{A\in\mathcal{A}} \left| \int_A f(x)\,dx - \mu_{n_t}(A) \right| \\
&\leq 4 \cdot \sup_{A\in\mathcal{B}_d} \left| \int_A f_{n_l,\bar{h}}(x)\,dx - \int_A f(x)\,dx \right| \\
&\qquad +4 \cdot \max_{A\in\mathcal{A}} \left| \int_A f(x)\,dx - \mu_{n_t}(A) \right|
\end{aligned}$$

$$= 2 \cdot \int_{\mathbb{R}^d} |f_{n_l,\bar{h}}(x)\,dx - f(x)|\,dx + 4 \cdot \Delta,$$

was die Behauptung impliziert. □

Seien nun X, X_1, X_2, ... unabhängige identisch verteilte $\mathbb{R}^d$–wertige Zufallsvariablen mit Dichte $f : \mathbb{R}^d \to \mathbb{R}$. Dann gilt nach der Ungleichung von Hoeffding für für $A \in \mathcal{B}_d$ und $\epsilon > 0$:

$$\begin{aligned}
&\mathbf{P}\left\{\left|\int_A f(x)\,dx - \mu_{n_t}(A)\right| > \epsilon\right\} \\
&= \mathbf{P}\left\{\left|\frac{1}{n_t}\sum_{i=n_l+1}^{n_l+n_t} 1_A(X_i) - \mathbf{E}\{1_A(X)\}\right| > \epsilon\right\} \\
&\leq 2 \cdot \exp\left(-2 \cdot n_t \cdot \epsilon^2\right).
\end{aligned}$$

Für beliebiges $\delta > 0$ impliziert dies

$$\begin{aligned}
\mathbf{E}\Delta &= \int_0^\infty \mathbf{P}\left\{\Delta > t\right\}\,dt \\
&\leq \delta + \int_\delta^\infty \mathbf{E}\left\{\mathbf{P}\left\{\Delta > t\right\}\left|X_1,\ldots,X_{n_l}\right.\right\}\,dt \\
&\leq \delta + \int_\delta^\infty \mathbf{E}\left\{|\mathcal{A}|^2 \cdot \max_{A\in\mathcal{A}} \mathbf{P}\left\{\left|\int_A f(x)\,dx - \mu_{n_t}(A)\right| > t\right\}\left|X_1,\ldots,X_{n_l}\right.\right\}\,dt \\
&\leq \delta + \int_\delta^\infty |\mathcal{P}|^2 \cdot 2 \cdot \exp\left(-2 \cdot n_t \cdot t^2\right)\,dt \\
&\leq \delta + \int_\delta^\infty |\mathcal{P}|^2 \cdot 2 \cdot \exp\left(-2 \cdot n_t \cdot \delta \cdot t\right)\,dt \\
&= \delta + |\mathcal{P}|^2 \cdot \frac{1}{n_t \cdot \delta} \cdot \exp\left(-2 \cdot n_t \cdot \delta^2\right),
\end{aligned}$$

woraus für $|\mathcal{P}| \geq 3$ mit $\delta = \frac{\sqrt{\ln(|\mathcal{P}|}}{\sqrt{n_t}}$ folgt:

$$\mathbf{E}\Delta \leq \frac{\sqrt{\ln(|\mathcal{P}|)}}{\sqrt{n_t}} + \frac{1}{\sqrt{n_t} \cdot \sqrt{\ln(|P|)}} \leq 2 \cdot \frac{\sqrt{\ln(|\mathcal{P}|)}}{\sqrt{n_t}}. \tag{2.18}$$

Daher folgt aus Lemma 2.14:

Satz 2.15. *Seien X, X_1, X_2, ... unabhängige identisch verteilte $\mathbb{R}^d$–wertige Zufallsvariablen mit Dichte $f : \mathbb{R}^d \to \mathbb{R}$. Sei $n = n_l + n_t$ mit $n_l, n_t \in \mathbb{N}$ und sei $\mathcal{P} \subseteq (0,\infty)$ eine endliche Menge mit $|\mathcal{P}| \geq 3$. Für $K : \mathbb{R}^d \to \mathbb{R}$ sei*

$$f_{n_l,h}(x) = \frac{1}{n_l \cdot h^d} \cdot \sum_{i=1}^{n_l} K\left(\frac{x - X_i}{h}\right),$$

und sei die Bandbreite $\hat{h}$ durch die kombinatorische Methode wie in (2.16) gewählt. Dann gilt:

$$\mathbf{E}\int_{\mathbb{R}^d} |\hat{f}_n(x) - f(x)|\,dx \leq 3 \cdot \min_{h\in\mathcal{P}} \mathbf{E}\int_{\mathbb{R}^d} |f_{n_l,h} - f(x)|\,dx + 8 \cdot \frac{\sqrt{\ln(|\mathcal{P}|}}{\sqrt{n_t}}.$$

Beweis: Die Behauptung folgt aus Lemma 2.14 und (2.18). □

Bem. Wählt man in Satz 2.15 z.B. $n_t \approx n/2 \approx n_l$,

$$\mathcal{P} = \left\{2^k \,:\, k \in \{-n, -n+1, \ldots, n\}\right\}$$

und K geeignet, so gilt für jedes $0 < p \leq 1$ und jedes $C > 0$ für n hinreichend groß:

$$\sup_{f\in\mathcal{F}^{(p,C)}} \mathbf{E}\int_{\mathbb{R}^d} |f_n(x) - f(x)|\,dx \leq c \cdot C^{\frac{d}{2\cdot p+d}} n^{-\frac{p}{2p+d}},$$

für eine nur von d abhängende Konstante c.

Kapitel 3

Regressionsschätzung bei festem Design

3.1 Einführung

Gegeben sind Daten

$$(x_{1,n}, Y_{1,n}), \ldots, (x_{n,n}, Y_{n,n}),$$

wobei die $x_{1,n}, \ldots, x_{n,n} \in \mathbb{R}^d$ sind und

$$Y_{i,n} = m(x_{i,n}) + \epsilon_{i,n} \quad (i = 1, \ldots, n)$$

gilt für eine Funktion $m : \mathbb{R}^d \to \mathbb{R}$ und unabhängige reelle Zufallsvariablen

$$\epsilon_{1,n}, \ldots, \epsilon_{n,n}$$

mit

$$\mathbf{E}(\epsilon_{i,n}) = 0 \quad (i = 1, \ldots, n).$$

Meist ist hierbei $d = 1$, manchmal auch $d = 2$, größere Werte kommen für d eher selten vor.

Gesucht ist eine Schätzung

$$m_n(\cdot) = m_n(\cdot, (x_{1,n}, Y_{1,n}), \ldots, (x_{n,n}, Y_{n,n})) : \mathbb{R}^d \to \mathbb{R}$$

von m bzw. eventuell auch nur Schätzungen von $m(x_{1,n}), \ldots, m(x_{n,n})$.

Beispiel: Die obige Problemstellung tritt z.B. bei der Schätzung der Festigkeit von Werkstoffen auf. Diese wird üblicherweise beschrieben durch eine sogenannte Dehnungs-Woehler-Linie, die beschreibt, nach wievielen zyklischen Belastungen zu einer vorgebenen Dehnung ein Werkstoff bricht. Hierbei wäre dann $x_{i,n}$ gerade die beim i-ten Versuch eingestellte Dehnung, und $Y_{i,n}$ wäre die Anzahl der zyklischen Belastungen mit dieser Dehnung bis zum Brechen des Werkstoffes. Geplottet wird dabei üblicherweise die Abhängigkeit der Dehnung von der Anzahl der Zyklen (also gerade die Umkehrfunktion zu der obigen Funktion). Standardmäßig führt man dazu 12 bis 16 Versuche durch, was einen Zeitaufwand von etwa einem Monat verursacht.

Im Folgenden betrachten wir sogenannte *Kleinste-Quadrate-Schätzer*:

Für einen vorgegebenen Raum $\mathcal{F}_n$ von Funktionen $f : \mathbb{R}^d \to \mathbb{R}$ setzen wir

$$m_n(\cdot) = \arg\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |f(x_{i,n}) - Y_{i,n}|^2,$$

d.h. $m_n(\cdot) = m_n(\cdot, (x_{1,n}, Y_{1,n}), \ldots, (x_{n,n}, Y_{n,n}))$ erfüllt

$$m_n(\cdot) = m_n(\cdot, (x_{1,n}, Y_{1,n}), \ldots, (x_{n,n}, Y_{n,n})) \in \mathcal{F}_n$$

sowie

$$\frac{1}{n} \sum_{i=1}^{n} |m_n(x_{i,n}) - Y_{i,n}|^2 = \min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |f(x_{i,n}) - Y_{i,n}|^2.$$

Dabei setzen wir vereinfachend voraus, dass obiges Minimum existiert. Sollte es nicht eindeutig sein, wählen wir irgendeine Funktion, die das Minimum annimmt.

Ziel im Folgenden ist die Abschätzung des durchschnittlichen quadratischen Fehlers (auch empirischen L_2-Fehlers)

$$\frac{1}{n} \sum_{i=1}^{n} |m_n(x_{i,n}) - m(x_{i,n})|^2$$

von m_n in Abhängigkeit von $\mathcal{F}_n$.

Zur Vereinfachung der Schreibweise schreiben wir ab sofort x_i bzw. Y_i bzw. ϵ_i statt $x_{i,n}$ bzw. $Y_{i,n}$ bzw. $\epsilon_{i,n}$.

3.2 Lineare Kleinste-Quadrate-Schätzer

In diesem Abschnitt ist $\mathcal{F}_n$ ein von n abhängender *endlichdimensionaler* linearer Vektorraum. Der Schätzer ist definiert durch

$$m_n(\cdot) = \arg\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |f(x_i) - Y_i|^2 \tag{3.1}$$

3.2.1 Existenz und Berechnung des Schätzers

$B_1, \ldots, B_K : \mathbb{R}^d \to \mathbb{R}$ sei eine Basis von $\mathcal{F}_n$, wobei $K = K_n$ die Vektorraumdimension von $\mathcal{F}_n$ ist. Ist dann

$$f = \sum_{j=1}^{K} a_j B_j,$$

so gilt:

$$\frac{1}{n} \sum_{i=1}^{n} |f(x_i) - Y_i|^2 = \frac{1}{n} \left\| \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} - \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \right\|_2^2 = \frac{1}{n} \|\mathbf{Ba} - \mathbf{Y}\|_2^2,$$

wobei

$$\mathbf{B} = (B_j(x_i))_{i=1,\ldots,n;j=1,\ldots,K}\,,\ \mathbf{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_K \end{pmatrix}\,,\ \mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

und $\|z\|_2$ die Euklidische Norm von $z \in \mathbb{R}^n$ ist.

Also ist (3.1) äquivalent zu

$$m_n(\cdot) = \sum_{j=1}^{K} a_j^* \cdot B_j(\cdot) \tag{3.2}$$

und dem linearen Ausgleichsproblem

$$\|\mathbf{Ba}^* - \mathbf{Y}\|_2^2 = \min_{\mathbf{a} \in \mathbb{R}^K} \|\mathbf{Ba} - \mathbf{Y}\|_2^2. \tag{3.3}$$

Aus der Numerik ist bekannt, dass (3.3) wiederum äquivalent ist zu der sogenannten Normalengleichung

$$\mathbf{B}^T \mathbf{Ba}^* = \mathbf{B}^T \mathbf{Y}, \tag{3.4}$$

und dass eine Lösung des linearen Gleichungssystems (3.4) immer existiert.

Damit haben wir gezeigt: Ist $\mathcal{F}_n$ ein linearer Vektorraum, so existiert der Kleinste-Quadrate-Schätzer immer und kann durch Lösen eines linearen Gleichungssystems berechnet werden.

Bemerkungen. a) Die Lösung von (3.4) muss nicht eindeutig sein, allerdings ist nach dem Projektionssatz $\mathbf{Ba}^*$ eindeutig, was impliziert, dass

$$(m_n(x_1), \ldots, m_n(x_n))$$

eindeutig ist.

b) Durch Anwenden des Cram-Schmidtschen Orthonormalisierungsverfahrens bzgl. des (Semi-)Skalarproduktes

$$< f, g >_n = \frac{1}{n} \sum_{i=1}^{n} f(x_i) \cdot g(x_i)$$

kann man erreichen:

$$\frac{1}{n}\mathbf{B}^T\mathbf{B} = (< B_j, B_k >)_{j,k=1,\ldots,n} = \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

In diesem Fall ist dann

$$m_n(x) = \sum_{j=1}^{K} a_j^* \cdot B_j \quad \text{mit} \quad \mathbf{a}^* = \frac{1}{n}\mathbf{B}^T\mathbf{Y},$$

eine Funktion, die (3.1) erfüllt, wobei gilt:

$$a_j^* = \frac{1}{n} \sum_{i=1}^{n} Y_i \cdot B_j(x_i).$$

Dies impliziert

$$\mathbf{E}[m_n(x)] = \sum_{j=1}^{K} \mathbf{E}[a_j^*] \cdot B_j(x) = \sum_{j=1}^{K} \left(\frac{1}{n} \sum_{i=1}^{n} m(x_i) \cdot B_j(x_i) \right) \cdot B_j(x).$$

Man sieht, dass damit $\mathbf{E}[m_n(x)]$ die gleiche Bauart hat wie $m_n(x)$, wobei allerdings Y_i durch $m(x_i)$ ersetzt ist. Folglich ist $\mathbf{E}[m_n(x)]$ der Kleinste-Quadrate-Schätzer zu den Daten

$$(x_1, m(x_1)), \ldots, (x_n, m(x_n))$$

und es gilt:

$$\frac{1}{n}\sum_{i=1}^{n}|\mathbf{E}[m_n(x_i)] - m(x_i)|^2 = \min_{f\in\mathcal{F}_n}\frac{1}{n}\sum_{i=1}^{n}|f(x_i)-m(x_i)|^2.$$

(Da oben $m_n(x_1),\ldots,m_n(x_n)$ eindeutig ist, gilt diese Beziehung für jeden Kleinsten-Quadrate-Schätzer).

3.2.2 Konvergenzgeschwindigkeit

Hauptresultat dieses Abschnittes ist

Satz 3.1. *Ist $\mathcal{F}_n$ ein linearer Vektorraum der Dimension $K_n < \infty$ bestehend aus Funktionen $f : \mathbb{R}^d \to \mathbb{R}$, ist m_n der zugehörige Kleinste-Quadrate-Schätzer definiert durch (3.1), und*

$$\sigma^2 = \max_{i=1,\ldots,n} V(\epsilon_i),$$

dann gilt:

$$\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n}|m_n(x_i)-m(x_i)|^2\right] \le \sigma^2\cdot\frac{K_n}{n} + \min_{f\in\mathcal{F}_n}\frac{1}{n}\sum_{i=1}^{n}|f(x_i)-m(x_i)|^2.$$

Beweis. oBdA $\sigma^2 < \infty$.

Es gilt:

$$\begin{aligned}&\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n}|m_n(x_i)-m(x_i)|^2\right]\\ &= \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n}|m_n(x_i)-\mathbf{E}[m_n(x_i)]|^2\right] + \frac{1}{n}\sum_{i=1}^{n}|\mathbf{E}[m_n(x_i)]-m(x_i)|^2\end{aligned}$$

da

$$\begin{aligned}&\mathbf{E}\left[(m_n(x_i)-\mathbf{E}[m_n(x_i)])\cdot(\mathbf{E}[m_n(x_i)]-m(x_i))\right]\\ &= (\mathbf{E}[m_n(x_i)]-m(x_i))\cdot(\mathbf{E}\left[m_n(x_i)\right]-\mathbf{E}[m_n(x_i)])\\ &= (\mathbf{E}[m_n(x_i)]-m(x_i))\cdot 0 = 0.\end{aligned}$$

Wegen der Bemerkung oben ist

$$\frac{1}{n}\sum_{i=1}^{n}|\mathbf{E}[m_n(x_i)] - m(x_i)|^2 = \min_{f\in\mathcal{F}_n}\frac{1}{n}\sum_{i=1}^{n}|f(x_i)-m(x_i)|^2,$$

also genügt es im Folgenden zu zeigen:

$$\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n}|m_n(x_i)-\mathbf{E}[m_n(x_i)]|^2\right] \leq \sigma^2 \cdot \frac{K_n}{n}.$$

Dazu wählen wir eine Basis $B_1, \ldots, B_{K_n}$ von $\mathcal{F}_n$, wobei wir oBdA annehmen können, dass für

$$\mathbf{B} = (B_j(x_i))_{i=1,\ldots,n;j=1,\ldots,K_n}$$

gilt

$$\frac{1}{n}\mathbf{B}^T\mathbf{B} = \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

was oBdA

$$m_n(x) = \sum_{j=1}^{K}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i \cdot B_j(x_i)\right) \cdot B_j(x)$$

impliziert (vgl. Bemerkung oben, hierbei haben wir benützt, dass $m_n(x_i)$ eindeutig sind und nur von diesen die Behauptung abhängt).

Setze

$$B(x) = \begin{pmatrix} B_1(x) \\ \vdots \\ B_{K_n}(x) \end{pmatrix}.$$

Dann gilt

$$m_n(x) = \sum_{j=1}^{K}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i \cdot B_j(x_i)\right) \cdot B_j(x) = B(x)^T \cdot \frac{1}{n}\mathbf{B}^T\mathbf{Y}$$

und

$$\mathbf{E}[m_n(x)] = B(x)^T \cdot \frac{1}{n}\mathbf{B}^T \begin{pmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{pmatrix}.$$

Also folgt unter Beachtung von $z^2 = z \cdot z^T$ für $z \in \mathbb{R}$

$$\begin{aligned} &\mathbf{E}\left[|m_n(x)-\mathbf{E}[m_n(x)]|^2\right] \\ &= \mathbf{E}\left[\left|B(x)^T \cdot \frac{1}{n}\mathbf{B}^T \begin{pmatrix} Y_1 - m(x_1) \\ \vdots \\ Y_n - m(x_n) \end{pmatrix}\right|^2\right] \end{aligned}$$

$$= \mathbf{E}\left[B(x)^T \frac{1}{n}\mathbf{B}^T \begin{pmatrix} Y_1 - m(x_1) \\ \vdots \\ Y_n - m(x_n) \end{pmatrix} \cdot (Y_1 - m(x_1), \ldots, Y_n - m(x_n)) \cdot \frac{1}{n}\mathbf{B} \cdot B(x) \right]$$
$$= B(x)^T \frac{1}{n}\mathbf{B}^T \cdot (\mathbf{E}[(Y_i - m(x_i)) \cdot (Y_j - m(x_j))])_{1 \le i,j \le n} \cdot \frac{1}{n}\mathbf{B} \cdot B(x),$$

wobei die letzte Gleichheit aus der Linearität des Erwartungswertes folgt.

Wegen der Unabhängigkeit der Daten gilt

$$\mathbf{E}[(Y_i - m(x_i)) \cdot (Y_j - m(x_j))] = 0 \quad \text{für } i \neq j$$

und

$$\mathbf{E}[(Y_i - m(x_i)) \cdot (Y_j - m(x_j))] = \mathbf{E}[\epsilon_i^2] \quad \text{für } i = j.$$

Für $b = (b_1, \ldots, b_{K_n})^T \in \mathbb{R}^n$ gilt nun nach Definition von σ^2

$$b^T \left(\delta_{i,j} \cdot \mathbf{E}[\epsilon_i^2] \right)_{1 \le i,j \le n} b = \sum_{j=1}^{n} \mathbf{E}[\epsilon_j^2] \cdot b_j^2 \le \sigma^2 \cdot \sum_{j=1}^{n} b_j^2 = \sigma^2 \cdot b^T b,$$

woraus folgt:

$$\begin{aligned} \mathbf{E}\left[|m_n(x) - \mathbf{E}[m_n(x)]|^2 \right] &\le \sigma^2 \cdot B(x)^T \frac{1}{n}\mathbf{B}^T \cdot \frac{1}{n}\mathbf{B} \cdot B(x) \\ &\overset{s.o.}{=} \sigma^2 \cdot B(x)^T \cdot \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \cdot B(x) \cdot \frac{1}{n} \\ &\le \sigma^2 \cdot \frac{1}{n} \cdot B(x)^T \cdot B(x) \\ &= \sigma^2 \cdot \frac{1}{n} \sum_{j=1}^{K_n} |B_j(x)|^2. \end{aligned}$$

Damit erhalten wir unter Beachtung von

$$\frac{1}{n} \sum_{i=1}^{n} |B_j(x_i)|^2 \in \{0, 1\} \quad (j \in \{1, \ldots, K_n\}),$$

was aus

$$\frac{1}{n}\mathbf{B}^T\mathbf{B} = \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

folgt, dass gilt:

$$\mathbf{E}\left[\frac{1}{n} \sum_{i=1}^{n} |m_n(x_i) - \mathbf{E}[m_n(x_i)]|^2 \right]$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}\left[|m_n(x_i) - \mathbf{E}[m_n(x_i)]|^2\right] \\
&\overset{s.o.}{\leq} \frac{1}{n}\sum_{i=1}^{n} \sigma^2 \cdot \frac{1}{n}\sum_{j=1}^{K_n} |B_j(x_i)|^2 \\
&= \frac{\sigma^2}{n}\sum_{j=1}^{K_n} \frac{1}{n}\sum_{i=1}^{n} |B_j(x_i)|^2 \\
&\leq \frac{\sigma^2}{n}\sum_{j=1}^{K_n} 1 = \sigma^2 \cdot \frac{K_n}{n},
\end{aligned}
$$

was zu zeigen war. □

Korollar 3.2. *Sei*

$$Y_i = m(x_i) + \epsilon_i \quad (i = 1, \ldots, n)$$

mit $x_1, \ldots, x_n \in [0,1]$, $m : \mathbb{R} \to \mathbb{R}$ *p-fach stetig differenzierbar und* $\epsilon_1, \ldots, \epsilon_n$ *unabhängig mit* $\mathbf{E}\{\epsilon_i\} = 0$ *und* $V(\epsilon_i) \leq \sigma^2$ $(i = 1, \ldots, n)$. $\mathcal{F}_n$ *sei die Menge aller stückweisen Polynome vom Grad* $\max\{0, p-1\}$ *in Bezug auf eine äquidistante Partition von* $[0,1]$ *in* $\lceil n^{1/(2p+1)} \rceil$ *viele Intervalle.* m_n *sei der zugehörige Kleinste-Quadrate-Schätzer definiert durch (3.1). Dann gilt:*

$$\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n} |m_n(x_i) - m(x_i)|^2\right] = O\left(n^{-\frac{2p}{2p+1}}\right)$$

Beweis. Wegen

$$\frac{dim(\mathcal{F}_n)}{n} = \frac{(\max\{0, p-1\} + 1) \cdot \lceil n^{1/(2p+1)} \rceil}{n} = O\left(n^{-\frac{2p}{2p+1}}\right)$$

und (wie man unter Verwendung eines Taylorpolynoms auf jedem einzelnen Intervall sieht)

$$
\begin{aligned}
\min_{f \in \mathcal{F}_n} \frac{1}{n}\sum_{i=1}^{n} |f(x_i) - m(x_i)|^2 &\leq \inf_{f \in \mathcal{F}_n} \sup_{x \in [0,1]} |f(x) - m(x)|^2 \\
&\leq const(m) \cdot \left(\frac{1}{\lceil n^{1/(2p+1)} \rceil}\right)^{2p} \\
&= O\left(n^{-\frac{2p}{2p+1}}\right)
\end{aligned}
$$

folgt die Behauptung aus Satz 3.1. □

3.3 Nichtlineare Kleinste-Quadrate-Schätzer

3.3.1 Motivation

Wir betrachten den Kleinsten-Quadrate-Schätzer

$$m_n(\cdot) = \arg\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |f(x_i) - Y_i|^2,$$

wobei $\mathcal{F}_n$ eine gegebene (nichtlineare) Menge von Funktionen $f : \mathbb{R}^d \to \mathbb{R}$ ist. Dieser Kleinste-Quadrate-Schätzer (also eine Funktion oben, die das Minimum annimmt) sei als existent vorausgesetzt, ebenso existiere

$$m_n^*(\cdot) = \arg\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |f(x_i) - m(x_i)|^2.$$

Hierbei ist

$$\frac{1}{n} \sum_{i=1}^{n} |m_n^*(x_i) - m(x_i)|^2 = \min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |f(x_i) - m(x_i)|^2$$

der sogenannte (deterministische) *Approximationsfehler*, der bei Approximation von m durch Funktionen aus $\mathcal{F}_n$ immer mindestens ensteht.

Nach Definition von m_n gilt wegen $m_n^* \in \mathcal{F}_n$

$$\frac{1}{n} \sum_{i=1}^{n} |m_n(x_i) - Y_i|^2 \leq \frac{1}{n} \sum_{i=1}^{n} |m_n^*(x_i) - Y_i|^2,$$

was impliziert

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^{n} |Y_i - m(x_i)|^2 + 2 \cdot \frac{1}{n} \sum_{i=1}^{n} (Y_i - m(x_i)) \cdot (m(x_i) - m_n(x_i)) + \frac{1}{n} \sum_{i=1}^{n} |m(x_i) - m_n(x_i)|^2 \\ \leq &\frac{1}{n} \sum_{i=1}^{n} |Y_i - m(x_i)|^2 + 2 \cdot \frac{1}{n} \sum_{i=1}^{n} (Y_i - m(x_i)) \cdot (m(x_i) - m_n^*(x_i)) + \frac{1}{n} \sum_{i=1}^{n} |m(x_i) - m_n^*(x_i)|^2 \end{aligned}$$

bzw.

$$\frac{1}{n} \sum_{i=1}^{n} |m_n(x_i) - m(x_i)|^2$$

$$\leq \frac{1}{n}\sum_{i=1}^{n}|m_n^*(x_i)-m(x_i)|^2+2\cdot\frac{1}{n}\sum_{i=1}^{n}(Y_i-m(x_i))\cdot(m_n(x_i)-m_n^*(x_i)). \tag{3.5}$$

Mit $(a+b)^2 \leq 2a^2+2b^2$ $(a,b\in\mathbb{R})$ folgt daraus

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}|m_n(x_i)-m_n^*(x_i)|^2\\
&\leq \frac{2}{n}\sum_{i=1}^{n}|m_n(x_i)-m(x_i)|^2+\frac{2}{n}\sum_{i=1}^{n}|m_n^*(x_i)-m(x_i)|^2\\
&\leq 4\cdot\frac{1}{n}\sum_{i=1}^{n}|m_n^*(x_i)-m(x_i)|^2+4\cdot\frac{1}{n}\sum_{i=1}^{n}(Y_i-m(x_i))\cdot(m_n(x_i)-m_n^*(x_i)).
\end{aligned} \tag{3.6}$$

Nun gilt:

Im Falle

$$\frac{1}{n}\sum_{i=1}^{n}|m_n^*(x_i)-m(x_i)|^2 \geq \frac{1}{n}\sum_{i=1}^{n}(Y_i-m(x_i))\cdot(m_n(x_i)-m_n^*(x_i))$$

ist wegen (3.5)

$$\frac{1}{n}\sum_{i=1}^{n}|m_n(x_i)-m(x_i)|^2 \leq 3\cdot\frac{1}{n}\sum_{i=1}^{n}|m_n^*(x_i)-m(x_i)|^2. \tag{3.7}$$

Andernfalls gilt nach (3.6)

$$\frac{1}{n}\sum_{i=1}^{n}|m_n(x_i)-m_n^*(x_i)|^2 \leq 8\cdot\frac{1}{n}\sum_{i=1}^{n}(Y_i-m(x_i))\cdot(m_n(x_i)-m_n^*(x_i)). \tag{3.8}$$

Also folgt für $\delta>0$ beliebig aus

$$\frac{1}{n}\sum_{i=1}^{n}|m_n(x_i)-m(x_i)|^2 > \delta+3\cdot\frac{1}{n}\sum_{i=1}^{n}|m_n^*(x_i)-m(x_i)|^2, \tag{3.9}$$

dass (3.8) gilt (da (3.7) nicht gelten kann). Unter Beachtung von

$$\frac{1}{n}\sum_{i=1}^{n}|m_n(x_i)-m(x_i)|^2 \leq 2\cdot\frac{1}{n}\sum_{i=1}^{n}|m_n(x_i)-m_n^*(x_i)|^2+2\cdot\frac{1}{n}\sum_{i=1}^{n}|m_n^*(x_i)-m(x_i)|^2$$

folgt aus (3.9) auch noch

$$2 \cdot \frac{1}{n} \sum_{i=1}^{n} |m_n(x_i) - m_n^*(x_i)|^2 > \delta.$$

Damit ist gezeigt:

$$\begin{aligned}
&\mathbf{P}\left[\frac{1}{n}\sum_{i=1}^{n}|m_n(x_i)-m(x_i)|^2 > \delta + 3\cdot \min_{f\in\mathcal{F}_n}\frac{1}{n}\sum_{i=1}^{n}|f(x_i)-m(x_i)|^2\right] \\
&\leq \mathbf{P}\left[\frac{\delta}{2} < \frac{1}{n}\sum_{i=1}^{n}|m_n(x_i)-m_n^*(x_i)|^2 \leq 8\cdot\frac{1}{n}\sum_{i=1}^{n}(Y_i-m(x_i))\cdot(m_n(x_i)-m_n^*(x_i))\right] \\
&\leq \mathbf{P}\left[\exists f\in\mathcal{F}_n : \frac{\delta}{2} < \frac{1}{n}\sum_{i=1}^{n}|f(x_i)-m_n^*(x_i)|^2 \leq 8\cdot\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\cdot(f(x_i)-m_n^*(x_i))\right],
\end{aligned}$$

wobei $\epsilon_i = Y_i - m(x_i)$.

Im Folgenden wird nun die obige Wahrscheinlichkeit abgeschätzt unter der Voraussetzung, dass die $\epsilon_1, \ldots, \epsilon_n$ unabhängige Zufallsvariablen mit Erwartungswert Null sind. Die Abschätzung wird dabei von der "Komplexität" des Funktionenraums abhängen, die wir mit den im nächsten Abschnitt vorgestellten Überdeckungszahlen beschreiben werden.

Im Falle, dass $\mathcal{F}_n$ eine endliche Menge von Funktionen ist, lässt sich die obige Wahrscheinlichkeit wie folgt abschätzen:

$$\begin{aligned}
&\mathbf{P}\left[\exists f\in\mathcal{F}_n : \frac{\delta}{2} < \frac{1}{n}\sum_{i=1}^{n}|f(x_i)-m_n^*(x_i)|^2 \leq 8\cdot\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\cdot(f(x_i)-m_n^*(x_i))\right] \\
&\leq |\mathcal{F}_n|\cdot\max_{f\in\mathcal{F}_n}\mathbf{P}\left[\frac{\delta}{2} < \frac{1}{n}\sum_{i=1}^{n}|f(x_i)-m_n^*(x_i)|^2 \leq 8\cdot\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\cdot(f(x_i)-m_n^*(x_i))\right] \\
&= |\mathcal{F}_n|\cdot\max_{f\in\mathcal{F}_n}\mathbf{P}\left[\frac{\tilde{\delta}}{8} < \frac{1}{n}\sum_{i=1}^{n}\epsilon_i\cdot(f(x_i)-m_n^*(x_i))\right],
\end{aligned}$$

wobei

$$\tilde{\delta} = \tilde{\delta}(f) = \max\left\{\frac{\delta}{2}, \frac{1}{n}\sum_{i=1}^{n}|f(x_i)-m_n^*(x_i)|^2\right\}.$$

Sei nun $f \in \mathcal{F}_n$ fest. Dann sind die Zufallsvariablen

$$\epsilon_i \cdot (f(x_i) - m_n^*(x_i)) \quad (i = 1, \ldots, n)$$

unabhängige reelle Zufallsvariablen mit Erwartungswert Null. Setzen wir nun voraus, dass die ϵ_i beschränkt sind, d.h., dass für ein $L > 0$ gilt

$$|\epsilon_i| \leq L \quad f.s. \quad (i = 1, \ldots, n),$$

dann gilt für diese Zufallsvariablen

$$|\epsilon_i \cdot (f(x_i) - m_n^*(x_i))| \leq L \cdot |f(x_i) - m_n^*(x_i)| \quad (i = 1, \ldots, n)$$

f.s. Mit der Ungleichung von Hoeffding folgt dann aber

$$\begin{aligned}
&\mathbf{P}\left[\frac{\tilde{\delta}}{4} < \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \cdot (f(x_i) - m_n^*(x_i))\right] \\
&\leq 2 \cdot \exp\left(-\frac{2 \cdot n \cdot \left(\frac{\tilde{\delta}}{8}\right)^2}{4L^2 \cdot \frac{1}{n}\sum_{i=1}^{n} |f(x_i) - m_n^*(x_i)|^2}\right) \\
&\leq 2 \cdot \exp\left(-\frac{n \cdot \delta}{128 L^2}\right),
\end{aligned}$$

wobei wir bei der letzten Umformung verwendet haben, dass

$$\frac{\tilde{\delta}^2}{\frac{1}{n}\sum_{i=1}^{n} |f(x_i) - m_n^*(x_i)|^2} \geq \frac{\tilde{\delta}^2}{\frac{\tilde{\delta}}{2}} = 2 \cdot \tilde{\delta} \geq \delta$$

nach Definition von $\tilde{\delta}$ gilt.

3.3.2 Überdeckungszahlen

Wir beschreiben Überdeckungszahlen zuerst allgemein in halbmetrischen Räumen.

Definition 3.3. *(X, d) sei ein halbmetrischer Raum. Für $x \in X$ und $\epsilon > 0$ sei*

$$U_\epsilon(x) = \{z \in X \,:\, d(x, z) < \epsilon\}$$

die Kugel um x mit Radius ϵ.

a) *$\{z_1, \ldots, z_N\} \subseteq X$ heißt* **ϵ-Überdeckung** *einer Menge $A \subseteq X$, falls gilt:*

$$A \subseteq \cup_{k=1}^{N} U_\epsilon(z_k).$$

b) *Ist $A \subseteq X$ und $\epsilon > 0$, so ist die sogenannte* **ϵ-Überdeckungszahl** *von A in (X, d) definiert als*

$$\mathcal{N}_{(X,d)}(\epsilon, A) = \inf\left\{|U| \quad : \quad U \subseteq X \text{ ist } \epsilon\text{-Überdeckung von } A\right\}.$$

Definition 3.4. *Sei $\mathcal{F}$ eine Menge von Funktionen $f : \mathbb{R}^d \to \mathbb{R}$, sei $\epsilon > 0$, $1 \leq p < \infty$ und seien $x_1, \ldots, x_n \in \mathbb{R}^d$ und $x_1^n = (x_1, \ldots, x_n)$. Dann ist die L_p-ϵ-Überdeckungszahl von $\mathcal{F}$ auf x_1^n definiert durch*

$$\mathcal{N}_p(\epsilon, \mathcal{F}, x_1^n) := \mathcal{N}_{(X,d)}(\epsilon, \mathcal{F}),$$

wobei der halbmetrische Raum (X, d) gegeben ist durch

- *X = Menge aller Funktionen $f : \mathbb{R}^d \to \mathbb{R}$,*
- *$d(f, g) = d_p(f, g) = (\frac{1}{n} \sum_{i=1}^n |f(x_i) - g(x_i)|^p)^{1/p}$.*

In anderen Worten: $\mathcal{N}_p(\epsilon, \mathcal{F}, x_1^n)$ ist das minimale $N \in \mathbb{N}$, so dass Funktionen $f_1, \ldots, f_N : \mathbb{R}^d \to \mathbb{R}$ existieren mit der Eigenschaft, dass für jedes $f \in \mathcal{F}$ gilt:

$$\min_{j=1,\ldots,N} (\frac{1}{n} \sum_{i=1}^n |f(x_i) - f_j(x_i)|^p)^{1/p} < \epsilon.$$

Im Folgenden verwenden wir L_2-Überdeckungszahlen. Eine Abschätzung einer solchen Überdeckungszahl enthält

Lemma 3.5. *Sei $\mathcal{F}$ die Menge aller monoton wachsender Funktionen $f : \mathbb{R} \to [0, 1]$. Dann gilt für beliebige $x_1, \ldots, x_n \in \mathbb{R}$ und beliebiges $\delta > 0$:*

$$\mathcal{N}_2(\delta, \mathcal{F}, x_1^n) \leq \left(n + \frac{1}{\delta} \right)^{1/\delta}.$$

Beweis. Für $f \in \mathcal{F}$ setze

$$M_i = \lfloor \frac{f(x_i)}{\delta} \rfloor \quad (i = 1, \ldots, n),$$

wobei $\lfloor z \rfloor$ die größte ganze Zahl kleiner oder gleich z ist. Wähle dann eine Funktion $g_f : \mathbb{R} \to \mathbb{R}$ mit

$$g_f(x_i) = \delta \cdot M_i \quad (i = 1, \ldots, n).$$

Dann gilt

$$g_f(x_i) = \delta \cdot \lfloor \frac{f(x_i)}{\delta} \rfloor \in (f(x_i) - \delta, f(x_i)] \quad (i = 1, \ldots, n),$$

woraus folgt

$$|f(x_i) - g_f(x_i)| = f(x_i) - g_f(x_i) \in [0, \delta)$$

sowie

$$d_2(f, g_f) < \delta.$$

Also ist

$$\{g_f \,:\, f \in \mathcal{F}\}$$

eine L_2-ϵ-Überdeckung von $\mathcal{F}$. Diese besteht aus so vielen Funktionen, wie es Zahlen

$$(M_1, \ldots, M_n) \in \mathbb{N}_0^n$$

mit

$$0 \leq M_1 \leq M_2 \leq \cdots \leq M_n \leq \lfloor \frac{1}{\delta} \rfloor$$

gibt (wobei die letzte Ungleichungskette aus f nichtnegativ und monoton wachsend mit Funktionswerten kleiner oder gleich Eins folgt).

Jedes dieses n-Tupel von Zahlen entspricht einer Ziehung von n Zahlen aus einer Grundmenge vom Umfang

$$\lfloor \frac{1}{\delta} \rfloor + 1$$

mit Zurücklegen und ohne Beachtung der Reihenfolge. Die zugehörige Formel aus der Kombinatorik führt auf

$$|\{g_f \,:\, f \in \mathcal{F}\}| \leq \binom{(\lfloor \frac{1}{\delta} \rfloor + 1) + n - 1}{n} = \binom{\lfloor \frac{1}{\delta} \rfloor + n}{n} = \binom{\lfloor \frac{1}{\delta} \rfloor + n}{\lfloor \frac{1}{\delta} \rfloor} \leq \left(\frac{1}{\delta} + n \right)^{\frac{1}{\delta}},$$

was zu zeigen war. □

Bemerkung. Mit einem deutlich aufwendigeren Beweis kann man sogar zeigen

$$\mathcal{N}_2(\delta, \mathcal{F}, x_1^n) \leq e^{c \cdot \frac{1}{\delta}}$$

für $c > 0$ geeignet, vgl. Birman und Solomjak (1967).

3.3.3 Eine uniforme Exponentialungleichung

Ziel dieses Abschnittes ist der Beweis von

Satz 3.6. *(van de Geer, 1990).*

Sei $n \in \mathbb{N}$, $x_1, \ldots, x_n \in \mathbb{R}^d$, $R > 0$ *und* $\mathcal{F}$ *eine Menge von Funktionen* $f : \mathbb{R}^d \to \mathbb{R}$ *mit*

$$\|f\|_n^2 := \frac{1}{n} \sum_{i=1}^{n} f(x_i)^2 \leq R^2 \quad (f \in \mathcal{F}). \tag{3.10}$$

Die rellen Zufallsvariablen $\epsilon_1, \ldots, \epsilon_n$ *seien unabhängig mit*

$$\mathbf{E}\epsilon_i = 0 \quad \textit{und} \quad |\epsilon_i| \leq L < \infty \quad f.s. \quad (i = 1, \ldots, n).$$

Dann existiert $c = c(L) > 0$ *so, dass für alle* $\delta > 0$ *mit*

$$\sqrt{n} \cdot \delta > c \cdot R \tag{3.11}$$

und

$$\sqrt{n} \cdot \delta \geq c \cdot \int_{\delta/(8 \cdot L)}^{R/2} (\log \mathcal{N}_2(u, \mathcal{F}, x_1^n))^{1/2} \, du \tag{3.12}$$

gilt:

$$\mathbf{P}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(x_i) \cdot \epsilon_i \right| \geq \delta \right] \leq c \cdot \exp\left(-\frac{n \cdot \delta^2}{c \cdot R^2} \right).$$

Bemerkung:

a) Messbarkeitsprobleme werden vernachlässigt (sonst Übergang zu äußerem Maß).

b) Für $|\mathcal{F}| = 1$ folgt die Behauptung aus der Ungleichung von Hoeffding.

c) Für $R = \tilde{c} \cdot \sqrt{\delta}$ (was in unserer Anwendung später der Fall sein wird) ist der Exponent der rechten Seite oben linear in δ. Damit können wir dann Konvergenzgeschwindigkeiten bis fast zu $1/n$ herleiten.

Beweis. oBdA $L \geq 1$.

oBdA $R > \delta/(2L)$, denn gilt diese Beziehung nicht, so folgt nach der Ungleichung von Cauchy-Schwarz, $|\epsilon_i| \leq L$ $f.s.$ und Beziehung (3.10):

$$\left| \frac{1}{n} \sum_{i=1}^{n} f(x_i) \cdot \epsilon_i \right| \leq \|f\|_n \cdot \sqrt{\frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2} \leq R \cdot L \leq \frac{\delta}{2} < \delta,$$

weswegen dann die Behauptung trivial ist.

Im Folgenden approximieren wir in

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(x_i)\cdot\epsilon_i\right|$$

jedes f durch eine Folge von f immer besser approximierender Funktionen (sog. Chaining-Technik).

Dazu:

Für $s\in\mathbb{N}_0$ sei $\{f_1^s,\ldots,f_{N_s}^s\}$ eine $R/2^s$-Überdeckung von $\mathcal{F}$ (bzgl. $\|\cdot\|_n$) minimaler Größe

$$N_s=\mathcal{N}_2\left(\frac{R}{2^s},\mathcal{F},x_1^n\right).$$

Wegen (3.10) gilt dabei oBdA $f_1^0=0$ und $N_0=1$.

Für $f\in\mathcal{F}$ sei

$$f^s\in\left\{f_1^s,\ldots,f_{N_s}^s\right\}$$

eine Funktion mit

$$\|f-f^s\|_n\le\frac{R}{2^s}.$$

Setze

$$S=\min\left\{s\ge 1\,:\,\frac{R}{2^s}\le\frac{\delta}{2\cdot L}\right\}.$$

Dann gilt wegen $f^0=f_1^0=0$ (s.o.) für beliebiges $f\in\mathcal{F}$:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} f(x_i)\cdot\epsilon_i &= \frac{1}{n}\sum_{i=1}^{n}\left(f(x_i)-f^0(x_i)\right)\cdot\epsilon_i\\ &= \frac{1}{n}\sum_{i=1}^{n}\left(f(x_i)-f^S(x_i)\right)\cdot\epsilon_i+\sum_{s=1}^{S}\frac{1}{n}\sum_{i=1}^{n}\left(f^s(x_i)-f^{s-1}(x_i)\right)\cdot\epsilon_i.\end{aligned}$$

Beachtet man, dass wegen der Ungleichung von Cauchy-Schwarz, $|\epsilon_i|\le L$ $f.s.$, der Beziehung (3.10) und der Definition von S gilt

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i)-f^S(x_i)\right)\cdot\epsilon_i\right|\le\|f-f^S\|_n\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2}\le\frac{R}{2^S}\cdot L\le\frac{\delta}{2\cdot L}\cdot L=\frac{\delta}{2},$$

so sieht man, dass für beliebige $\eta_1,\ldots,\eta_S\ge 0$ mit $\eta_1+\cdots+\eta_S\le 1$ gilt:

$$\mathbf{P}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(x_i)\cdot\epsilon_i\right|>\delta\right]$$

$$
\begin{aligned}
&\leq \mathbf{P}\left[\exists f \in \mathcal{F} : \sum_{s=1}^{S} \left| \frac{1}{n} \sum_{i=1}^{n} \left(f^s(x_i) - f^{s-1}(x_i)\right) \cdot \epsilon_i \right| > \frac{\delta}{2}\right] \\
&\leq \mathbf{P}\left[\exists f \in \mathcal{F} : \sum_{s=1}^{S} \left| \frac{1}{n} \sum_{i=1}^{n} \left(f^s(x_i) - f^{s-1}(x_i)\right) \cdot \epsilon_i \right| \geq \sum_{s=1}^{S} \eta_s \cdot \frac{\delta}{2}\right] \\
&\leq \sum_{s=1}^{S} \mathbf{P}\left[\exists f \in \mathcal{F} : \left| \frac{1}{n} \sum_{i=1}^{n} \left(f^s(x_i) - f^{s-1}(x_i)\right) \cdot \epsilon_i \right| \geq \eta_s \cdot \frac{\delta}{2}\right].
\end{aligned}
$$

Mit

$$
\|f^s - f^{s-1}\|_n^2 \leq \left(\|f^s - f\|_n + \|f - f^{s-1}\|_n\right)^2 \leq \left(\frac{R}{2^s} + \frac{R}{2^{s-1}}\right)^2 = 9 \cdot \frac{R^2}{2^{2\cdot s}}
$$

und der Ungleichung von Hoeffding folgt unter Beachtung der Tatsache, dass

$$
\left\{(f^s, f^{s-1}) \,:\, f \in \mathcal{F}\right\}
$$

aus

$$
N_s \cdot N_{s-1} \leq N_s^2 = \left(\mathcal{N}_2\left(\frac{R}{2}, \mathcal{F}, x_1^n\right)\right)^2
$$

vielen Funktionen besteht, dass gilt:

$$
\begin{aligned}
&\mathbf{P}\left[\sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(x_i) \cdot \epsilon_i \right| > \delta\right] \\
&\leq \sum_{s=1}^{S} \left(\mathcal{N}_2\left(\frac{R}{2^s}, \mathcal{F}, x_1^n\right)\right)^2 \cdot 2 \cdot \exp\left(-\frac{2 \cdot n \cdot \left(\frac{\eta_s \cdot \delta}{2}\right)^2}{4 \cdot L^2 \cdot 9 \cdot \frac{R^2}{2^{2\cdot s}}}\right) \\
&= \sum_{s=1}^{S} 2 \cdot \exp\left(2 \cdot \log \mathcal{N}_2\left(\frac{R}{2^s}, \mathcal{F}, x_1^n\right) - \frac{n \cdot \delta^2 \cdot 2^{2\cdot s}}{72 \cdot L^2 \cdot R^2} \cdot \eta_s^2\right) \\
&\leq \sum_{s=1}^{S} 2 \cdot \exp\left(-\frac{n \cdot \delta^2 \cdot 2^{2\cdot s}}{144 \cdot L^2 \cdot R^2} \cdot \eta_s^2\right),
\end{aligned}
$$

falls gilt

$$
2 \cdot \log \mathcal{N}_2\left(\frac{R}{2^s}, \mathcal{F}, x_1^n\right) - \frac{1}{2} \cdot \frac{n \cdot \delta^2 \cdot 2^{2\cdot s}}{72 \cdot L^2 \cdot R^2} \cdot \eta_s^2 \leq 0,
$$

d.h. falls gilt:

$$
\eta_s \geq \bar{\eta}_s = \frac{12 \cdot \sqrt{2} \cdot L \cdot R}{\sqrt{n} \cdot \delta \cdot 2^s} \cdot \left(\log \mathcal{N}_2\left(\frac{R}{2^s}, \mathcal{F}, x_1^n\right)\right)^{1/2}.
$$

Setze nun

$$\eta_s = \max\left\{\bar{\eta}_s, 2^{-s}\cdot\sqrt{s}\cdot\frac{1}{5}\right\}.$$

Beachtet man, dass wegen $R/2^{S-1} \geq \delta/(2L)$ (nach Definition von S) und (3.12) für $c \geq 48\sqrt{2}\cdot L$ einerseits gilt

$$\begin{aligned}
\sum_{s=1}^{S}\bar{\eta}_s &= \sum_{s=1}^{S}\frac{12\cdot\sqrt{2}\cdot L\cdot R}{\sqrt{n}\cdot\delta\cdot 2^s}\cdot\left(\log\mathcal{N}_2\left(\frac{R}{2^s},\mathcal{F},x_1^n\right)\right)^{1/2}\\
&\leq \sum_{s=1}^{S}\frac{24\cdot\sqrt{2}\cdot L}{\sqrt{n}\cdot\delta}\cdot\int_{R/2^{s+1}}^{R/2^s}\left(\log\mathcal{N}_2(u,\mathcal{F},x_1^n)\right)^{1/2}du\\
&\leq \frac{24\cdot\sqrt{2}\cdot L}{\sqrt{n}\cdot\delta}\cdot\int_{\delta/(8L)}^{R/2}\left(\log\mathcal{N}_2(u,\mathcal{F},x_1^n)\right)^{1/2}du\\
&\leq \frac{24\cdot\sqrt{2}\cdot L}{\sqrt{n}\cdot\delta}\cdot\frac{\sqrt{n}\cdot\delta}{c} = \frac{24\cdot\sqrt{2}\cdot L}{c}\leq\frac{1}{2},
\end{aligned}$$

sowie andererseits ebenfalls gilt

$$\sum_{s=1}^{S}\frac{2^{-s}\cdot\sqrt{s}}{5}\leq\frac{1}{5}\cdot\sum_{s=1}^{\infty}s\cdot\left(\frac{1}{2}\right)^s = \frac{1}{10}\cdot\frac{d}{dx}\left(\sum_{s=0}^{\infty}x^s\right)\Big|_{x=1/2} = \frac{1}{10}\cdot\frac{1}{(1-1/2)^2} = \frac{4}{10}\leq\frac{1}{2},$$

so sieht man:

$$\sum_{s=1}^{S}\eta_s\leq\sum_{s=1}^{S}\bar{\eta}_s+\sum_{s=1}^{S}\frac{2^{-s}\cdot\sqrt{s}}{5}\leq\frac{1}{2}+\frac{1}{2}=1.$$

Mit dieser Wahl von η_s erhalten wir:

$$\begin{aligned}
&\mathbf{P}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}f(x_i)\cdot\epsilon_i\right|>\delta\right]\\
&\leq\sum_{s=1}^{S}2\cdot\exp\left(-\frac{n\cdot\delta^2\cdot 2^{2\cdot s}}{144\cdot L^2\cdot R^2}\cdot\eta_s^2\right)\\
&\leq 2\cdot\sum_{s=1}^{\infty}\exp\left(-\frac{n\cdot\delta^2\cdot 2^{2\cdot s}}{144\cdot L^2\cdot R^2}\cdot 2^{-2\cdot s}\cdot s\cdot\frac{1}{25}\right)\\
&= 2\cdot\sum_{s=1}^{\infty}\exp\left(-\frac{n\cdot\delta^2}{25\cdot 144\cdot L^2\cdot R^2}\cdot s\right)\\
&= \frac{2}{1-\exp\left(-\frac{n\cdot\delta^2}{25\cdot 144\cdot L^2\cdot R^2}\right)}\cdot\exp\left(-\frac{n\cdot\delta^2}{25\cdot 144\cdot L^2\cdot R^2}\right)
\end{aligned}$$

$$\leq c \cdot \exp\left(-\frac{n \cdot \delta^2}{c \cdot R^2}\right)$$

für $c = c(L)$ genügend groß, wobei wir benutzt haben, dass nach (3.11) für $c = c(L)$ genügend groß gilt:

$$\frac{n \cdot \delta^2}{25 \cdot 144 \cdot L^2 \cdot R^2} \geq \frac{c^2}{25 \cdot 144 \cdot L^2} \geq const > 0.$$

□

3.3.4 Konvergenzgeschwindigkeit nichtlinearer Kleinste-Quadrate-Schätzer

Sei

$$Y_i = m(x_i) + \epsilon_i \quad (i = 1, \ldots, n),$$

wobei $x_1, \ldots, x_n \in \mathbb{R}^d$, $m : \mathbb{R}^d \to \mathbb{R}$ und $\epsilon_1, \ldots, \epsilon_n$ unabhängige reelle Zufallsvariablen sind mit

$$\mathbf{E}\epsilon_i = 0 \quad \text{und} \quad |\epsilon_i| \leq L \quad f.s. \quad (i = 1, \ldots, n).$$

Wir betrachten den Kleinsten-Quadrate-Schätzer

$$m_n(\cdot) = \arg\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |f(x_i) - Y_i|^2,$$

wobei $\mathcal{F}_n$ eine gegebene (nichtlineare) Menge von Funktionen $f : \mathbb{R}^d \to \mathbb{R}$ ist. Dann gilt:

Satz 3.7. *Sei*

$$\|f\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} |f(x_i)|^2 \quad \textit{und} \quad m_n^* = \arg\min_{f \in \mathcal{F}_n} \|f - m\|_n^2.$$

Gilt dann für $\delta_n > 0$

$$n \cdot \delta_n \to \infty \quad (n \to \infty) \tag{3.13}$$

und für ein $c_1 > 0$ *und alle* $\delta \geq \delta_n$

$$\sqrt{n} \cdot \delta \geq c_1 \cdot \int_{\delta/(128 \cdot L)}^{\sqrt{\delta}} \left(\log \mathcal{N}_2\left(u, \left\{f - m_n^* \,:\, f \in \mathcal{F}_n, \|f - m_n^*\|_n^2 \leq \delta\right\}, x_1^n\right)\right)^{1/2} du, \tag{3.14}$$

so folgt

$$\mathbf{P}\left[\frac{1}{n}\sum_{i=1}^{n}|m_n(x_i)-m(x_i)|^2 > c_2\cdot\left(\delta_n+\min_{f\in\mathcal{F}_n}\frac{1}{n}\sum_{i=1}^{n}|f(x_i)-m(x_i)|^2\right)\right]\to 0$$

$(n\to\infty)$ *für ein* $c_2>0$.

Beweis. Gemäß Abschnitt 3.3.1 gilt

$$\begin{aligned}
&\mathbf{P}\left[\frac{1}{n}\sum_{i=1}^{n}|m_n(x_i)-m(x_i)|^2 > \delta_n+3\cdot\min_{f\in\mathcal{F}_n}\frac{1}{n}\sum_{i=1}^{n}|f(x_i)-m(x_i)|^2\right]\\
&\le \mathbf{P}\left[\exists f\in\mathcal{F}_n:\frac{\delta_n}{2}<\frac{1}{n}\sum_{i=1}^{n}|f(x_i)-m_n^*(x_i)|^2\le 8\cdot\frac{1}{n}\sum_{i=1}^{n}(f(x_i)-m_n^*(x_i))\cdot\epsilon_i\right]\\
&\le \sum_{k=0}^{\infty}\mathbf{P}\left[\exists f\in\mathcal{F}_n:2^{k-1}\cdot\delta_n<\frac{1}{n}\sum_{i=1}^{n}|f(x_i)-m_n^*(x_i)|^2\le 2^k\cdot\delta_n\right.\\
&\qquad\qquad\left.\frac{\delta_n}{2}<\frac{1}{n}\sum_{i=1}^{n}|f(x_i)-m_n^*(x_i)|^2\le 8\cdot\frac{1}{n}\sum_{i=1}^{n}(f(x_i)-m_n^*(x_i))\cdot\epsilon_i\right]\\
&\le \sum_{k=0}^{\infty}\mathbf{P}\left[\exists f\in\mathcal{F}_n:\frac{1}{n}\sum_{i=1}^{n}|f(x_i)-m_n^*(x_i)|^2\le 2^k\cdot\delta_n,\right.\\
&\qquad\qquad\left.\frac{1}{n}\sum_{i=1}^{n}(f(x_i)-m_n^*(x_i))\cdot\epsilon_i>\frac{2^k\cdot\delta_n}{16}\right].
\end{aligned}$$

Wir wenden nun (für $k\in\mathbb{N}_0$ fest) Satz 3.6 an mit

$$\mathcal{F}=\left\{f-m_n^*\,:\,f\in\mathcal{F}_n,\|f-m_n^*\|_n^2\le 2^k\cdot\delta_n\right\}$$

(also mit $R=\sqrt{2^k\delta_n}$) und $\delta=2^k\delta_n/16$. Dann gilt für n genügend groß (3.11), also

$$\sqrt{n}\cdot\frac{2^k\delta_n}{16}>c\cdot\sqrt{2^k\delta_n},$$

wegen (3.13), und wegen

$$\sqrt{n}\cdot\frac{2^k\delta_n}{16}\ge c\cdot\int_{2^k\delta_n/(8\cdot 16\cdot L)}^{\sqrt{2^k\delta_n}/2}(\log\mathcal{N}_2(u,\mathcal{F},x_1^n))^{1/2}\,du$$

(was durch (3.14) impliziert wird, wenn man darin $\delta=2^k\delta_n$ und $c_1=16\cdot c$ setzt) gilt auch (3.12) für $R=\sqrt{2^k\delta_n}$ und $\delta=2^k\delta_n/16$.

Damit erhält man

$$\begin{aligned}
&\mathbf{P}\left[\frac{1}{n}\sum_{i=1}^{n}|m_n(x_i)-m(x_i)|^2 > \delta_n + 3\cdot\min_{f\in\mathcal{F}_n}\frac{1}{n}\sum_{i=1}^{n}|f(x_i)-m(x_i)|^2\right]\\
&\leq \sum_{k=0}^{\infty} c\cdot\exp\left(-\frac{n\cdot\left(\frac{2^k\delta_n}{16}\right)^2}{c\cdot 2^k\delta_n}\right)\\
&= \sum_{k=0}^{\infty} c\cdot\exp\left(-\frac{n\cdot\delta_n}{16^2\cdot c}\cdot 2^k\right)\\
&\leq \sum_{k=0}^{\infty} c\cdot\exp\left(-\frac{n\cdot\delta_n}{16^2\cdot c}\cdot(k+1)\right)\\
&\leq \tilde{c}\cdot\exp(-\tilde{c}\cdot n\cdot\delta_n)\to 0 \quad (n\to\infty),
\end{aligned}$$

wobei wir im letzten Schritt erneut die Voraussetzung $n\cdot\delta_n\to\infty$ $(n\to\infty)$ benutzt haben. □

Korollar 3.8. *Ist in Satz 3.7 $d=1$ und $\mathcal{F}_n=\mathcal{F}$ die Menge aller monoton wachsender Funktionen $f:\mathbb{R}\to[0,1]$, so gilt:*

$$\begin{aligned}
&\mathbf{P}\left[\frac{1}{n}\sum_{i=1}^{n}|m_n(x_i)-m(x_i)|^2 > c_2\cdot\left(\left(\frac{\log(n)}{n}\right)^{2/3}+\min_{f\in\mathcal{F}_n}\frac{1}{n}\sum_{i=1}^{n}|f(x_i)-m(x_i)|^2\right)\right]\\
&\to 0 \quad (n\to\infty)
\end{aligned}$$

für ein $c_2>0$.

Beweis. Gemäß Lemma 3.5 gilt

$$\mathcal{N}_2(\delta,\mathcal{F}_n,x_1^n)\leq\left(n+\frac{1}{\delta}\right)^{1/\delta},$$

was

$$\begin{aligned}
&\log\mathcal{N}_2\left(u,\left\{f-m_n^*\,:\,f\in\mathcal{F}_n,\|f-m_n^*\|_n^2\leq\delta\right\},x_1^n\right)\\
&\leq\log\mathcal{N}_2\left(u,\left\{f-m_n^*\,:\,f\in\mathcal{F}_n\right\},x_1^n\right)\\
&=\log\mathcal{N}_2\left(u,\left\{f\,:\,f\in\mathcal{F}_n\right\},x_1^n\right)\leq\frac{1}{u}\cdot\log\left(n+\frac{1}{u}\right)
\end{aligned}$$

impliziert. Daraus folgt für $\delta\geq 128\cdot L/n$:

$$c_1\cdot\int_{\delta/(128\cdot L)}^{\sqrt{\delta}}\left(\log\mathcal{N}_2\left(u,\left\{f-m_n^*\,:\,f\in\mathcal{F}_n,\|f-m_n^*\|_n^2\leq\delta\right\},x_1^n\right)\right)^{1/2}du$$

$$
\begin{aligned}
&\leq c_1 \cdot \int_{\delta/(128 \cdot L)}^{\sqrt{\delta}} \sqrt{\frac{1}{u}} \cdot \sqrt{\log(2 \cdot n)}\, du \\
&= c_1 \cdot \sqrt{\log(2 \cdot n)} \cdot 2 \cdot u^{1/2} \Big|_{u=\delta/(128 \cdot L)}^{\sqrt{\delta}} \\
&\leq 2 \cdot c_1 \cdot \sqrt{\log(2 \cdot n)} \cdot \delta^{1/4}.
\end{aligned}
$$

Also folgt (3.14) für $\delta_n \geq \frac{128 \cdot L}{n}$ aus

$$
\begin{aligned}
\sqrt{n} \cdot \delta \geq 2 \cdot c_1 \cdot \sqrt{\log(2 \cdot n)} \cdot \delta^{1/4} \quad &\Leftrightarrow \quad \delta^{3/4} \geq 2 \cdot c_1 \cdot \left(\frac{\log(2 \cdot n)}{n} \right)^{1/2} \\
&\Leftrightarrow \quad \delta \geq (2 \cdot c_1)^{4/3} \cdot \left(\frac{\log(2 \cdot n)}{n} \right)^{2/3}.
\end{aligned}
$$

Daher ist (3.14) für

$$
\delta_n = c_3 \cdot \left(\frac{\log(2 \cdot n)}{n} \right)^{2/3}
$$

mit $c_3 > 0$ genügend groß erfüllt, und trivialerweise gilt in diesem Fall auch (3.13). Mit Satz 3.7 folgt nun die Behauptung. □

Bemerkung.

a) Ist $m : \mathbb{R} \to \mathbb{R}$ in Korollar 3.8 monoton wachsend und nimmt m nur Werte in $[0, 1]$ an, so gilt in Korollar 3.8

$$
\mathbf{P} \left[\frac{1}{n} \sum_{i=1}^{n} |m_n(x_i) - m(x_i)|^2 > c_2 \cdot \left(\frac{\log(n)}{n} \right)^{2/3} \right] \to 0 \quad (n \to \infty).
$$

b) Durch eine aufwendigere Abschätzung der Überdeckungszahl oben lässt sich der logarithmische Faktor oben vermeiden.

c) Die Berechnung des Schätzers in Korollar 3.8 kann durch Lösen eines Minimierungsproblems mit Nebenbedingungen erfolgen.

Kapitel 4

Regressionsschätzung bei zufälligem Design

4.1 Einführung

4.1.1 Regressionsanalyse

(X, Y) sei eine $\mathbb{R}^d \times \mathbb{R}$-wertige Zufallsvariable mit $\mathbf{E}|Y| < \infty$.

Analysiert werden soll die Abhängigkeit des Wertes von Y vom Wert von X.

Ziele dabei (je nach Anwendung):

I) Interpretation des Zusammenhangs zwischen X und Y
z.B.
X = Alter einer amerikanischen Walddrossel,
Y = Gewicht.

II) Vorhersagen des Wertes von Y zu beobachteten Wert von X
z.B.
$\left.\begin{array}{l} X = \text{ Ortsvektor} \\ Y = \text{Geschwindigkeit eines Partikels} \end{array}\right\}$ (in einem Strömungsfeld)
oder
X = Einstellung eines Motors
Y = Abgase

Betrtachtet wird dazu die sogenannte *Regressionsfunktion* $m : \mathbb{R}^d \to \mathbb{R}$ definiert

durch

$$m(x) = \mathbf{E}\{Y|X = x\} \quad (x \in \mathbb{R}^d).$$

Anschaulich:
$m(x)$ ist der durchschnittliche Wert von Y unter der Bedingung $X = x$.

Formal:
m ist diejenige Borel-messbare Funktion $m : \mathbb{R}^d \to \mathbb{R}$ mit

$$\forall B \in \mathcal{B}_d : \quad \int_B m(x)\, \mathbf{P}_X(dx) = \int_{X^{-1}(B)} Y \, d\mathbf{P}.$$

Diese ist $\mathbf{P}_X$-f.ü. eindeutig (vgl. Vorlesung Wahrscheinlichkeitstheorie).

Die Regressionsfunktion hat die folgende Optimalitätseigenschaft:
Lemma 4.1. *Ist (X, Y) eine $\mathbb{R}^d \times \mathbb{R}$-wertige Zufallsvariable mit $\mathbf{E}Y^2 < \infty$, so gilt für $m : \mathbb{R}^d \to \mathbb{R}$, $m(x) = \mathbf{E}\{Y|X = x\}$ die Beziehung*

$$\mathbf{E}\left\{|m(X) - Y|^2\right\} = \min_{f:\mathbb{R}^d \to \mathbb{R} \text{ messbar}} \mathbf{E}\left\{|f(X) - Y|^2\right\}.$$

Beweis. Wir zeigen, dass für beliebiges (messbares) $f : \mathbb{R}^d \to \mathbb{R}$ gilt:

$$\mathbf{E}\left\{|f(X) - Y|^2\right\} = \mathbf{E}\left\{|m(X) - Y|^2\right\} + \int_{\mathbb{R}^d} |f(x) - m(x)|^2 \mathbf{P}_X(dx). \tag{4.1}$$

Wegen

$$\int_{\mathbb{R}^d} |f(x) - m(x)|^2 \mathbf{P}_X(dx) \geq 0$$

folgt daraus die Behauptung.

Zum Nachweis von (4.1) beachten wir, dass wegen $\mathbf{E}Y^2 < \infty$ nach der Jensenschen Ungleichung gilt:

$$\mathbf{E}\{|m(X)|^2\} = \mathbf{E}\{|\mathbf{E}\{Y|X\}|^2\} \leq \mathbf{E}\{\mathbf{E}\{|Y|^2|X\}\} = \mathbf{E}Y^2 < \infty.$$

Ist nun $\mathbf{E}\{|f(X)|^2\} = \infty$, so folgt

$$\mathbf{E}\left\{|f(X) - Y|^2\right\} = \infty = \int_{\mathbb{R}^d} |f(x) - m(x)|^2 \mathbf{P}_X(dx)$$

(da z.B. $\mathbf{E}\{|f(X)|^2\} \leq 2 \cdot \mathbf{E}\left\{|f(X) - m(X)|^2\right\} + 2 \cdot \mathbf{E}\left\{|m(X)|^2\right\}$ gilt), was (4.1) impliziert.

Ist dagegen $\mathbf{E}\{|f(X)|^2\} < \infty$, so gilt

$$\begin{aligned}\mathbf{E}\left\{|f(X)-Y|^2\right\} &= \mathbf{E}\left\{|(f(X)-m(X))+(m(X)-Y)|^2\right\} \\ &= \mathbf{E}\left\{|f(X)-m(X)|^2\right\} + \mathbf{E}\left\{|m(X)-Y|^2\right\}, \quad (4.2)\end{aligned}$$

da

$$\begin{aligned}&\mathbf{E}\left\{(f(X)-m(X))\cdot(m(X)-Y)\right\} \\ &= \mathbf{E}\left\{\mathbf{E}\left\{(f(X)-m(X))\cdot(m(X)-Y)\middle|X\right\}\right\} \\ &= \mathbf{E}\left\{(f(X)-m(X))\cdot\mathbf{E}\left\{m(X)-Y\middle|X\right\}\right\} \\ &= \mathbf{E}\left\{(f(X)-m(X))\cdot(m(X)-\mathbf{E}\left\{Y\middle|X\right\})\right\} \\ &= \mathbf{E}\left\{(f(X)-m(X))\cdot(\mathbf{E}\left\{Y\middle|X\right\}-\mathbf{E}\left\{Y\middle|X\right\})\right\} \\ &= 0.\end{aligned}$$

Hierbei wurde beim zweiten Gleichheitszeichen benutzt, dass nach Cauchy-Schwarz gilt

$$\begin{aligned}&\mathbf{E}\left\{|(f(X)-m(X))\cdot(m(X)-Y)|\right\} \\ &\leq \sqrt{\mathbf{E}\{|f(X)-m(X)|^2\}}\cdot\sqrt{\mathbf{E}\{|m(X)-Y|^2\}} < \infty\end{aligned}$$

und damit $(f(X)-m(X))\cdot(m(X)-Y)$ integrierbar ist.

Aus (4.2) folgt nun die Behauptung. □

Bemerkung. Gemäß dem obigen Beweis (siehe (4.1)) gilt für das sogenannte *L_2-Risiko* einer beliebigen (messbaren) Funktion:

$$\mathbf{E}\left\{|f(X)-Y|^2\right\} = \mathbf{E}\left\{|m(X)-Y|^2\right\} + \int_{\mathbb{R}^d} |f(x)-m(x)|^2 \mathbf{P}_X(dx).$$

Damit ist der mittlere quadratische Vorhersagefehler einer Funktion darstellbar als Summe des L_2-Risikos der Regressionsfunktion (unvermeidbarer Fehler) und des sogenannten *L_2-Fehlers*

$$\int |f(x)-m(x)|^2 \mathbf{P}_X(dx),$$

der entsteht aufgrund der Verwendung von f anstelle von m bei der Vorhersage bzw. Approximation des Wertes von Y.

4.1.2 Regressionsschätzung

In Anwendungen ist üblicherweise die Verteilung von (X,Y) unbekannt, daher kann $m(x) = \mathbf{E}\{Y|X=x\}$ nicht berechnet werden. Oft ist es aber möglich,

Werte von (X, Y) zu beobachten. Ziel ist dann, daraus die Regressionsfunktion zu schätzen. Im Hinblick auf die Minimierung des L_2-Risikos sollte dabei der L_2-Fehler der Schätzfunktion möglichst klein sein.

Formal führt das auf folgende Problemstellung:

(X, Y), (X_1, Y_1), (X_1, Y_2), ... seien unabhängige identisch verteilte $\mathbb{R}^d \times \mathbb{R}$-wertige Zufallsvariablen mit $\mathbf{E}Y^2 < \infty$. $m : \mathbb{R}^d \to \mathbb{R}$ definiert durch $m(x) = \mathbf{E}\{Y|X = x\}$ sei die zugehörige Regressionsfunktion.

Gegeben ist die Datenmenge

$$\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}\,.$$

Gesucht ist eine Schätzung

$$m_n(\cdot) = m_n(\cdot, \mathcal{D}_n) : \mathbb{R}^d \to \mathbb{R}$$

von m, für die

$$\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx)$$

möglichst klein ist.

Klassischer Ansatz: **Parametrische Regression**
Bauart von m ist bekannt und hängt nur von endlich vielen Parametern ab, schätze diese: z.B. durch lineare Regression

Im Folgenden: **Nichtparametrische Regression:**
Regressionsfunktion ist nicht durch endlich viele Parameter beschreibbar.

Wir untersuchen dabei z.B. den sogenannten Kernschätzer (Nadaraya, Watson)

$$m_n(x) = \frac{\sum_{i=1}^n Y_i \cdot K\left(\frac{x - X_i}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{x - X_j}{h_n}\right)} \quad \left(\text{ wobei } \frac{0}{0} := 0\right)$$

mit sogenanntem **Kern** $K : \mathbb{R}^d \to \mathbb{R}$ und sogenannter **Bandbreite** $h_n > 0$

z.B. $K = I_{S_0(1)} = I_{\{z \in \mathbb{R}^d : \|z\| \leq 1\}}$.

Deutung: Schätzwert ist Mittelwert aller der Y_i's, für die X_i nahe bei x ist.

4.1.3 Anwendung in der Mustererkennung

(X, Y) sei $\mathbb{R}^d \times \{0, 1\}$-wertige Zufallsvariable.

In der Mustererkennung beschäftigt man sich mit dem folgenden Vorhersageproblem:

Zu beobachtetem Wert von X möchte man den zugehörigen Wert von Y vorhersagen.

Bsp.: Erkennung von Werbeemails:

$$\begin{aligned} X &= \text{Text der Email bzw. Charakteristika des Textes} \\ Y &= \begin{cases} 1, & \text{falls es sich um eine Werbeemail handelt,} \\ 0, & \text{sonst.} \end{cases} \end{aligned}$$

Gesucht ist eine Funktion $g^* : \mathbb{R}^d \to \{0, 1\}$, für die die Wahrscheinlichkeit einer falschen Vorhersage möglichst klein ist, d.h. für die gilt:

$$\mathbf{P}\left\{g^*(X) \neq Y\right\} = \min_{g:\mathbb{R}^d \to \{0,1\}} \mathbf{P}\left\{g(X) \neq Y\right\}. \tag{4.3}$$

Es gilt:
Lemma 4.2. *Für $g^* : \mathbb{R}^d \to \{0, 1\}$ definiert durch*

$$g^*(x) = \begin{cases} 1, & \mathbf{P}\{Y = 1 | X = x\} > \mathbf{P}\{Y = 0 | X = x\}, \\ 0, & \textit{sonst.} \end{cases}$$

gilt (4.3).

Beweis. Sei $g : \mathbb{R}^d \to \{0, 1\}$ beliebig. Dann gilt für jedes $x \in \mathbb{R}^d$

$$\mathbf{P}\{g(X) \neq Y | X = x\} = 1 - \mathbf{P}\{g(X) = Y | X = x\} = 1 - \mathbf{P}\{g(x) = Y | X = x\},$$

und mit der Definition von g^* folgt daraus

$$\begin{aligned} &\mathbf{P}\{g(X) \neq Y | X = x\} - \mathbf{P}\{g^*(X) \neq Y | X = x\} \\ &= \mathbf{P}\{g^*(x) = Y | X = x\} - \mathbf{P}\{g(x) = Y | X = x\} \\ &\geq 0. \end{aligned}$$

Somit:

$$\mathbf{P}\{g^*(X) \neq Y\} = \int_{\mathbb{R}^d} \mathbf{P}\{g^*(X) \neq Y | X = x\} \mathbf{P}_X(dx)$$

$$\begin{aligned} &\leq \int_{\mathbb{R}^d} \mathbf{P}\{g(X) \neq Y | X = x\} \mathbf{P}_X(dx) \\ &= \mathbf{P}\{g(X) \neq Y\}. \end{aligned}$$

□

Bem.: Im obigen Beweis (wie auch im Beweis von Satz 4.3 unten) ist die Umformung

$$\mathbf{P}\{g(X) = Y | X = x\} = \mathbf{P}\{g(x) = Y | X = x\}$$

zwar intuitiv einleuchtend, i.A. aber mathematisch nicht korrekt. Sie lässt sich aber vermeiden, sofern man wie folgt argumentiert:

$$\begin{aligned} &\mathbf{P}\{g(X) \neq Y\} \\ &= \mathbf{P}\{g(X) = 0, Y = 1\} + \mathbf{P}\{g(X) = 1, Y = 0\} \\ &= \mathbf{E}\left\{\mathbf{P}\{g(X) = 0, Y = 1 | X\} + \mathbf{P}\{g(X) = 1, Y = 0 | X\}\right\} \\ &= \mathbf{E}\left\{1_{\{g(X)=0\}} \cdot \mathbf{P}\{Y = 1 | X\} + 1_{\{g(X)=1\}} \cdot \mathbf{P}\{Y = 0 | X\}\right\} \\ &= \int \left(1_{\{g(x)=0\}} \cdot \mathbf{P}\{Y = 1 | X = x\} + 1_{\{g(x)=1\}} \cdot \mathbf{P}\{Y = 0 | X = x\}\right) \mathbf{P}_X(dx) \\ &= \int \mathbf{P}\{Y \neq g(x) | X = x\} \mathbf{P}_X(dx), \end{aligned}$$

und daraus dann analog zu den angegebenen Beweisen die Behauptung schließt.

Wegen

$$\mathbf{P}\{Y = 1 | X = x\} + \mathbf{P}\{Y = 0 | X = x\} = 1$$

$\mathbf{P}_X$-f.ü. können wir g^* auch durch

$$g^*(x) = \begin{cases} 1, & \mathbf{P}\{Y = 1 | X = x\} > \frac{1}{2}, \\ 0, & \text{sonst} \end{cases}$$

definieren.

Die sogenannte **aposteriori Wahrscheinlichkeit**

$$\mathbf{P}\{Y = 1 | X = x\} = \mathbf{E}\left\{I_{\{Y=1\}} \middle| X = x\right\} =: m(x)$$

lässt sich als Regressionsfunktion zum Zufallsvektor $(X, I_{\{Y=1\}})$ auffassen. Approximiert man diese (z.B. mittels Regressionsschätzung) durch eine Funktion

$$\bar{m} : \mathbb{R}^d \to \mathbb{R}$$

und definiert man dann die sogenannte **Plug-In-Schätzfunktion** $\bar{g}$ durch

$$\bar{g}(x) = \begin{cases} 1, & \bar{m}(x) > \frac{1}{2}, \\ 0, & \text{sonst} \end{cases} = \begin{cases} 1, & \bar{m}(x) > 1 - \bar{m}(x), \\ 0, & \text{sonst}, \end{cases}$$

so gilt:

Satz 4.3. *Mit den obigen Bezeichnungen gilt:*

$$\begin{aligned} 0 &\leq \mathbf{P}\{\bar{g}(X) \neq Y\} - \mathbf{P}\{g^*(X) \neq Y\} \leq 2 \cdot \int |\bar{m}(x) - m(x)| \mathbf{P}_X(dx) \\ &\leq 2 \cdot \sqrt{\int |\bar{m}(x) - m(x)|^2 \mathbf{P}_X(dx)}. \end{aligned}$$

Damit führt ein "gutes" Regressionsschätzverfahren automatisch zu einem "guten" Mustererkennungsverfahren.

Beweis von Satz 4.3.

Gemäß Beweis von Lemma 4.2 gilt:

$$\begin{aligned} &\mathbf{P}\{\bar{g}(X) \neq Y | X = x\} - \mathbf{P}\{g^*(X) \neq Y | X = x\} \\ &= \mathbf{P}\{g^*(x) = Y | X = x\} - \mathbf{P}\{\bar{g}(x) = Y | X = x\} \\ &= m(x) \cdot I_{\{g^*(x)=1\}} + (1 - m(x)) \cdot I_{\{g^*(x)=0\}} \\ &\quad - \big(m(x) \cdot I_{\{\bar{g}(x)=1\}} + (1 - m(x)) \cdot I_{\{\bar{g}(x)=0\}}\big) \\ &= m(x) \cdot I_{\{g^*(x)=1\}} + (1 - m(x)) \cdot I_{\{g^*(x)=0\}} \\ &\quad - \big(\bar{m}(x) \cdot I_{\{g^*(x)=1\}} + (1 - \bar{m}(x)) \cdot I_{\{g^*(x)=0\}}\big) \\ &\quad + \Big\{\bar{m}(x) \cdot I_{\{g^*(x)=1\}} + (1 - \bar{m}(x)) \cdot I_{\{g^*(x)=0\}} \\ &\quad - \big(\bar{m}(x) \cdot I_{\{\bar{g}(x)=1\}} + (1 - \bar{m}(x)) \cdot I_{\{\bar{g}(x)=0\}}\big) \Big\} \\ &\quad + \bar{m}(x) \cdot I_{\{\bar{g}(x)=1\}} + (1 - \bar{m}(x)) \cdot I_{\{\bar{g}(x)=0\}} \\ &\quad - \big(m(x) \cdot I_{\{\bar{g}(x)=1\}} + (1 - m(x)) \cdot I_{\{\bar{g}(x)=0\}}\big) \\ &\leq 2 \cdot |\bar{m}(x) - m(x)|, \end{aligned}$$

da die Definition von $\bar{g}$ impliziert, dass gilt:

$$\Big\{ \dots \Big\} \leq 0.$$

Mit Lemma 4.2 folgt daraus

$$\begin{aligned} 0 &\leq \mathbf{P}\{\bar{g}(X) \neq Y\} - \mathbf{P}\{g^*(X) \neq Y\} \\ &= \int (\mathbf{P}\{\bar{g}(X) \neq Y | X = x\} - \mathbf{P}\{g^*(X) \neq Y | X = x\}) \, \mathbf{P}_X(dx) \\ &\leq 2 \cdot \int |\bar{m}(x) - m(x)| \, \mathbf{P}_X(dx). \end{aligned}$$

Mit der Ungleichung von Cauchy-Schwarz folgt daraus die Behauptung. □

4.2 Der Satz von Stone

Der oben bereits eingeführte Kernschätzer ist Beispiel für lokalen Durchschnittsschätzer

$$m_n(x) = \sum_{i=1}^{n} W_{n,i}(x) \cdot Y_i, \tag{4.4}$$

wobei

$$W_{n,i}(x) = W_{n,i}(x, X_1, \ldots, X_n)$$

von den x-Werten der gegebenen Daten abhängende Gewichte sind.

Beim Kernschätzer:

$$W_{n,i}(x) = \frac{K\left(\frac{x - X_i}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{x - X_j}{h_n}\right)}$$

Satz 4.4. *(Stone (1977)).*
Sei m_n ein lokaler Durchschnittsschätzer definiert durch (4.4).
Für jede beliebige Verteilung von X gelte:

(i)

$$\exists\, c \in \mathbb{R}_+ \, \forall f : \mathbb{R}^d \to \mathbb{R}_+ \text{ messbar mit } E\{f(X)\} < \infty \quad \forall n \in \mathbb{N} :$$
$$E\{\sum_{i=1}^{n} |W_{n,i}(X)| \cdot f(X_i)\} \leq c \cdot E\{f(X)\}.$$

(ii)

$$\exists\, D \geq 1 \, \forall n : P\{\sum_{i=1}^{n} |W_{n,i}(X)| \leq D\} = 1.$$

("'Beschränktheit der Gewichte"')

(iii)

$$\forall a > 0 : \; E\left\{ \sum_{i=1}^{n} |W_{n,i}(X)| \cdot I_{\{\|X_i - X\| > a\}} \right\} \to 0 \, (n \to \infty)$$

("'Lokale Entscheidung"')

(iv)

$$\sum_{i=1}^{n} W_{n,i}(X) \to^{P} 1$$

(v)

$$E\{\sum_{i=1}^{n} W_{n,i}(X)^2\} \to 0 \, (n \to \infty)$$

("'Einzelnes Gewicht hat keinen zu großen Einfluss"').

Dann gilt:

$$E \int |m_n(x) - m(x)|^2 P_X(dx) \to 0 \quad (n \to \infty)$$

für ***alle*** *Verteilungen von* (X, Y) *mit* $E\{Y^2\} < \infty$

(d.h. Schätzer ist ***universell*** *konsistent).*

Beweis:

$$\begin{aligned}
& E \int |m_n(x) - m(x)|^2 P_X(dx) \\
& \quad \overset{\text{Fubini}}{=} E\{|m_n(X) - m(X)|^2\} \\
& \quad \overset{(4.2)}{=} E\{|\sum_{i=1}^{n} W_{n,1}(X) \cdot Y_i - m(X)|^2\} \\
& \quad \overset{(a+b+c)^2 \le 3a^2+3b^2+3c^2}{\le} 3 \cdot E\{|\sum_{i=1}^{n} W_{n,i}(X) \cdot (Y_i - m(X_i))|^2\} \\
& \qquad + 3 \cdot E\{|\sum_{i=1}^{n} W_{n,i}(X) \cdot (m(X_i) - m(X))|^2\} \\
& \qquad + 3 \cdot E\{|\sum_{i=1}^{n} W_{n,i}(X) \cdot m(X) - m(X)|^2\} \\
& =: 3I_n + 3J_n + 3L_n.
\end{aligned}$$

Gemäß (IV) gilt

$$\sum_{i=1}^{n} W_{n,i}(X) \cdot m(X) - m(X) = m(X) \cdot \left(\sum_{i=1}^{n} W_{n,1}(X) - 1 \right) \to^{P} 0.$$

Daraus und aus (ii) und $E(|m(X)|^2) < \infty$ folgt mit dem Satz von der majorisierten Konvergenz (auf Teilteilfolgen angewendet)

$$L_n \to 0 \quad (n \to \infty). \tag{4.5}$$

Für J_n gilt:

$$\begin{aligned}
J_n &\leq E\{(\sum_{i=1}^{n} |W_{n,i}(X)| \cdot |m(X_i) - m(X)|)^2\} \\
&= E\{\left(\sum_{j=1}^{n} |W_{n,j}(X)| \right)^2 \cdot \left(\sum_{i=1}^{n} \frac{|W_{n,i}(X)|}{\sum_{j=1}^{n} |W_{n,j}(X)|} \cdot |m(X_i) - m(X)| \right)^2 \} \\
&\overset{\text{Jensen}}{\leq} E\{\sum_{j=1}^{n} |W_{n,j}(X)| \cdot \sum_{i=1}^{n} |W_{n,i}(X)| \cdot |m(X_i) - m(X)|^2\} \\
&\overset{\text{(ii)}}{\leq} D \cdot E\{\sum_{i=1}^{n} |W_{n,i}(X)| \cdot |m(X_i) - m(X)|^2\}.
\end{aligned}$$

Sei $\varepsilon > 0$ beliebig. Dann existiert beschränktes und gleichmäßig stetiges $\tilde{m}$ mit $E\{|m(X) - \tilde{m}(X)|^2\} < \varepsilon$.

Nun gilt:

$$\begin{aligned}
J_n \leq & 3D \cdot E\{\sum_{i=1}^{n} |W_{n,i}(X)| \cdot |m(X_i) - \tilde{m}(X_i)|^2\} \\
& + 3D \cdot E\{\sum_{i=1}^{n} |W_{n,i}(X)| \cdot |\tilde{m}(X_i) - \tilde{m}(X)|^2\} \\
& + 3D \cdot E\{\sum_{i=1}^{n} |W_{n,i}(X)| \cdot |\tilde{m}(X) - m(X)|^2\} \\
=: & 3D \cdot J_{n,1} + 3D \cdot J_{n,2} + 3D \cdot J_{n,3}.
\end{aligned}$$

Mit (i) folgt

$$J_{n,1} \leq c \cdot E\{|m(X) - \tilde{m}(X)|^2\} \overset{\text{s.o.}}{<} c \cdot \varepsilon,$$

und gemäß (ii) gilt

$$J_{n,3} \leq D \cdot E\{|\tilde{m}(X) - m(X)|^2\} \overset{\text{s.o.}}{<} D \cdot \varepsilon.$$

Für $J_{n,2}$ gilt für $\delta > 0$ beliebig:

$$\begin{aligned} J_{n,2} =& E\{\sum_{i=1}^{n} |W_{n,i}(X)| \cdot |\tilde{m}(X_i) - \tilde{m}(X)|^2 \cdot I_{\{\|X_i - X\| > \delta\}}\} \\ &+ E\{\sum_{i=1}^{n} |W_{n,i}(X)| \cdot |\tilde{m}(X_i) - \tilde{m}(X)|^2 \cdot I_{\{\|X_i - X\| \leq \delta\}}\} \\ \overset{\text{(ii)}}{\leq}& 4 \cdot \|\tilde{m}\|_\infty^2 \cdot E\{\sum_{i=1}^{n} |W_{n,i}(X)| \cdot I_{\{\|X_i - X\| > \delta\}}\} \\ &+ D \cdot \sup_{u,v \in \mathbb{R}^d : \|u-v\| \leq \delta} |\tilde{m}(u) - \tilde{m}(v)| \end{aligned}$$

Mit (iii) folgt

$$\bar{\lim}_{n\to\infty} J_{n,2} \leq D \cdot \sup_{u,v \in \mathbb{R}^d : \|u-v\| \leq \delta} |\tilde{m}(u) - \tilde{m}(v)|,$$

woraus wir wegen $\tilde{m}$ gleichmäßig stetig mit $\delta \downarrow 0$ erhalten:

$$J_{n_2} \to 0 \quad (n \to \infty).$$

Also gilt

$$\limsup_{n\to\infty} J_n \leq (c \cdot D) \cdot \varepsilon,$$

und mit $\varepsilon \downarrow 0$ erhalten wir

$$J_n \to 0 \quad (n \to \infty). \tag{4.6}$$

Für I_n gilt:

$$\begin{aligned} I_n =& \sum_{i,j=1}^{n} E\{W_{n,i}(X) \cdot W_{n,j}(X) \cdot (Y_i - m(X_i)) \cdot (Y_j - m(X_j))\} \\ =& \sum_{i=1}^{n} E\{W_{n,i}(X)^2 \cdot (Y_i - m(X_i))^2\}, \end{aligned}$$

da für $i \neq j$:

$$\begin{aligned} & E\{W_{n,i}(X) \cdot W_{n,j}(X) \cdot (Y_i - m(X_i)) \cdot (Y_j - m(X_j))\} \\ =& E\{E\{\ldots | X_1, \ldots, X_n, Y_i\}\} \\ =& E\{W_{n,i}(X) \cdot W_{n,j}(X) \cdot (Y_i - m(X_i)) \cdot \underbrace{E\{Y_j - m(X_j) | X_1, \ldots, X_n, Y_i\}}_{\substack{\text{Unabhängigkeit} \\ = E\{Y_j - m(X_j)|X_j\} \\ = m(X_j) - m(X_j) = 0}} \\ =& 0. \end{aligned}$$

Also erhalten wir für $\sigma^2(x) = E\{|Y - m(X)|^2 | X = x\}$:

$$\begin{aligned} I_n =& \sum_{i=1}^{n} E\{E\{W_{n,i}(X)^2 \cdot (Y_j - m(X_j))^2 | X_1, \ldots, X_n\}\} \\ =& \sum_{i=1}^{n} E\{W_{n,i}(X)^2 \cdot \underbrace{E\{(Y_i - m(X_i))^2 | X_1, \ldots, X_n\}}_{\substack{\text{Unabhängigkeit} \\ = E\{(Y_i - m(X_i))^2 | X_i\} \\ = \sigma^2(X_i)}} \\ =& E\{\sum_{i=1}^{n} W_{n,i}(X)^2 \cdot \sigma^2(X_i)\}. \end{aligned}$$

Sei $\varepsilon > 0$ beliebig. Dann existiert $f : \mathbb{R}^d \to \mathbb{R}$ beschränkt mit

$$E\{|\sigma^2(X) - f(X)| \leq \varepsilon.$$

Damit:

$$\begin{aligned} I_n =& E\{\sum_{i=1}^{n} W_{n,i}(X)^2 \cdot f(X_i)\} \\ &+ E\{\sum_{i=1}^{n} W_{n,i}(X)^2 \cdot (\sigma^2(X_i) - f(X_i))\} \\ \overset{\text{(ii)}}{\leq} & \|f\|_\infty \cdot E\{\sum_{i=1}^{n} W_{n,i}(X)^2\} + D \cdot E\{\sum_{i=1}^{n} |W_{n,i}(X)| \cdot |\sigma^2(X_i) - f(X_i)|\}. \end{aligned}$$

Mit (i) und (v) folgt

$$\begin{aligned} \limsup_{n\to\infty} I_n &\leq 0 + D \cdot E\{|\sigma^2(X) - f(X)|\} \\ &\leq D \cdot c \cdot \epsilon, \end{aligned}$$

woraus wir mit $\varepsilon \downarrow 0$

$$I_n \to 0 \quad (n \to \infty) \tag{4.7}$$

erhalten.
Aus (4.5) - (4.7) folgt nun die Behauptung. □

4.3 Universelle Konsistenz des Kernschätzers

Im Folgenden sei

$$m_n(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right) \cdot Y_i}{\sum_{j=1}^{n} K\left(\frac{x-X_j}{h_n}\right)}$$

der Kernschätzer mit Kern $K : \mathbb{R}^d \to \mathbb{R}_+$ und Bandbreite $h_n > 0$.

Satz 4.5. *Sei $S_{0,R}$ ein Ball mit Radius $R > 0$ und Mittelpunkt 0. Sei $K = I_{S_{0,R}}$ der sogenannte naive Kern und für die Bandbreite gelte*

$$h_n \to 0 \quad (0 \to \infty) \text{ und } n \cdot h_n^d \to \infty \quad (n \to \infty).$$

Dann gilt für den Kernschätzer

$$E\int |m_n(x)-m(x)|^2 P_X(dx) \to 0 \quad (n\to\infty)$$

für ***alle*** *Verteilungen von (X,Y) mit $EY^2<\infty$, d.h. der Kernschätzer ist universell konsistent.*

Im Beweis benötigen wir:

Lemma 4.6. *Sei B eine $b(n,p)$-verteilte ZV mit $n\in\mathbb{N}, \quad p\in(0,1]$.*

D.g.:

a) $E\{\frac{1}{1+B}\}\le\frac{1}{(n+1)\cdot p}$.

b) $E\{\frac{1}{B}\cdot I_{\{B>0\}}\}\le\frac{2}{(n+1)\cdot p}$.

Beweis:
a)

$$\begin{aligned}
E\{\frac{1}{1+B}\} =& \sum_{k=0}^{n}\frac{1}{1+k}\cdot\binom{n}{k}\cdot p^k(1-p)^{n-k}\\
\overset{\binom{n+1}{k+1}=\frac{n+1}{k+1}\cdot\binom{n}{k}}{=}& \frac{1}{n+1}\sum_{k=0}^{n}\binom{n+1}{k+1}\cdot p^{k+1}\cdot(1-p)^{n-k}\cdot\frac{1}{p}\\
=&\frac{1}{(n+1)\cdot p}\cdot\sum_{l=1}^{n+1}\binom{n+1}{l}\cdot p^l\cdot(1-p)^{n+1-l}\\
\le&\frac{1}{(n+1)\cdot p}\cdot\sum_{l=0}^{n+1}\underbrace{\binom{n+1}{l}\cdot p^l\cdot(1-p)^{n+1-l}}_{\text{Zähldichte } b(n+1,p)} = \frac{1}{(n+1)\cdot p}\cdot 1.
\end{aligned}$$

b)

$$\begin{aligned}
&E\{\frac{1}{B}\cdot I_{\{B>0\}}\} = E\{\frac{2}{B+B}\cdot I_{\{B>0\}}\}\\
\le&E\{\frac{2}{B+1}\cdot I_{\{B>0\}}\}\le 2\cdot E\{\frac{1}{B+1}\}\overset{a)}{\le}\frac{2\cdot 1}{(n+1)\cdot p} \quad \square
\end{aligned}$$

Beweis von Satz 4.5
Es genügt zu zeigen, dass für

$$W_{n,i}(x) = \frac{K\left(\frac{x-X_i}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{x-X_j}{h_n}\right)}$$

die Bedingungen (i)-(v) aus Satz 4.4 gelten.

Nachweis von (i)
Sei $f : \mathbb{R}^d \to \mathbb{R}_+$ beliebig mit $E\{f(X)\} < \infty$. Zu zeigen ist, dass für eine von f unabhängige Konstante $c > 0$ gilt:

$$E\{\sum_{i=1}^n |W_{n,i}(X)| \cdot f(X_i)\} \leq c \cdot Ef(X).$$

Wegen

$$\begin{aligned}
&E\{\sum_{i=1}^n |W_{n,i}(X)| \cdot f(X_i)\} \\
=&E\{\sum_{i=1}^n \frac{K\left(\frac{X-X_i}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{X-X_j}{h_n}\right)} \cdot f(X_i)\} \\
=&n \cdot E\{\frac{K\left(\frac{X-X_1}{h_n}\right)}{K\left(\frac{X-X_1}{h_n}\right) + \sum_{j=2}^n K\left(\frac{X-X_j}{h_n}\right)} \cdot f(X_1)\} \\
\overset{\text{Fubini}}{=}&n \cdot \int f(u) \cdot E\{\int \frac{K\left(\frac{x-u}{h_n}\right)}{K\left(\frac{x-u}{h_n}\right) + \sum_{j=2}^n K\left(\frac{x-X_j}{h_n}\right)} P_X(dx)\} P_X(du)
\end{aligned}$$

genügt es dazu zu zeigen:

$\exists\, c > 0 \quad \forall n \in \mathbb{N} \quad \forall u \in \mathbb{R}^d$:

$$E\{\int \frac{K\left(\frac{x-u}{h_n}\right)}{K\left(\frac{x-u}{h_n}\right) + \sum_{j=2}^n K\left(\frac{x-X_j}{h_n}\right)} P_X(dx)\} \leq \frac{c}{n}.$$

Dazu überdecken wir

$$S_{0,R}$$

durch Kugeln

$$x_1 + S_{0,R/2}, \ldots, x_M + S_{0,R/2}$$

vom Radius $R/2$ (wobei $x + A := \{x + z : z \in A\}$ für $A \subseteq \mathbb{R}^d$).

Dann gilt für $x \in u + h_n \cdot x_k + S_{0,R \cdot h_n/2}$:

$$u + h_n \cdot x_k + S_{0,R \cdot h_n/2} \subseteq x + S_{0,R \cdot h_n},$$

woraus für alle $z \in \mathbb{R}^d$ folgt:

$$K\left(\frac{x-z}{h_n}\right) = I_{\{z \in x + S_{0,R \cdot h_n}\}} \geq I_{\{z \in u + h_n \cdot x_k + S_{0,R \cdot h_n/2}\}} \tag{4.8}$$

Weiter gilt wegen

$$S_{0,R} \subseteq \bigcup_{k=1}^{M} x_k + S_{0,R/2}$$

auch

$$S_{0,R \cdot h_n} \subseteq \bigcup_{k=1}^{M} h_n \cdot x_k + S_{0,R \cdot h_n/2},$$

woraus folgt

$$K\left(\frac{x-u}{h_n}\right) = I_{\{x \in u + S_{0,R \cdot h_n}\}} \leq \sum_{k=1}^{M} I_{\{x \in u + h_n \cdot x_k + S_{0,R h_n/2}\}} \tag{4.9}$$

für alle $x, u \in \mathbb{R}^d$.

Damit:

$$
\begin{aligned}
& E\{\int \frac{K\left(\frac{x-u}{h_n}\right)}{K\left(\frac{x-u}{h_n}\right)+\sum_{j=2}^n K\left(\frac{x-X_j}{h_n}\right)} P_X(dx)\} \\
\overset{K(z)\in\{0,1\}}{=} & \{\int \frac{K\left(\frac{x-u}{h_n}\right)}{1+\sum_{j=2}^n K\left(\frac{x-X_j}{h_n}\right)} P_X(dx)\} \\
\overset{(4.9)}{\leq} & \sum_{k=1}^M E\{\int_{u+h_n\cdot x_k+S_{0,R\cdot h_n/2}} \frac{1}{1+\sum_{j=2}^n \underbrace{K\left(\frac{x-X_j}{h_n}\right)}_{=I_{\{X_j\in x+S_{0,R\cdot h_n}\}}}} P_X(dx)\} \\
\overset{(4.8)}{\leq} & \sum_{k=1}^M E\{\int_{u+h_n\cdot x_k+S_{0,R\cdot h_n/2}} \frac{1}{1+\sum_{j=2}^n I_{\{X_j\in u+h_n\cdot x_k+S_{0,R\cdot h_n/2}\}}} P_X(dx)\} \\
= & \sum_{k=1}^M P_X(u+h_n\cdot x_k+S_{0,R\cdot h_n/2})\cdot E\{\frac{1}{1+\underbrace{\sum_{j=2}^n I_{\{X_j\in u+h_n\cdot x_k+S_{0,R\cdot h_n/2}\}}}_{b(n,p)-\text{verteilt mit } p=P_X(\ldots)}} P_X(dx)\}. \\
\overset{\text{Lemma 4.6}}{\leq} & \sum_{k=1}^M \frac{1}{n} = \frac{1\cdot M}{n},
\end{aligned}
$$

womit (i) gezeigt ist.

Nachweis von (ii): Klar, da $0 \leq \sum_{i=1}^n W_{n,i}(X) \leq 1$.

Nachweis von (iii)

$$
\begin{aligned}
W_{n,i}(X) = & \frac{K\left(\frac{X-X_i}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{X-X_j}{h_n}\right)} = \frac{I_{\{\|X-X_i\|\leq R\cdot h_n\}}}{\sum_{j=1}^n I_{\{\|X-X_j\|\leq R\cdot h_n\}}} \\
= & 0 \quad \text{falls } \|X-X_i\| > R\cdot h_n.
\end{aligned}
$$

Sei $a>0$ beliebig. Wegen $h_n \to 0\,(n\to\infty)$ existiert n_0 mit

$$R\cdot h_n < a \quad \text{für alle } n \geq n_0.$$

Damit gilt für $n \geq n_0$:

$$\begin{aligned}\sum_{i=1}^{n} |W_{n_i}(X)| \cdot I_{\{\|X_i - X\| > a\}} &\leq \sum_{i=1}^{n} |W_{n_i}(X)| \cdot I_{\{\|X_i - X_i\| > R \cdot h_n\}} \\ &\overset{\text{s.o.}}{=} \sum_{i=1}^{n} 0 = 0,\end{aligned}$$

womit (iii) gezeigt ist.

Nachweis von (iv)
Sei $\varepsilon > 0$ beliebig. Dann gilt:

$$\begin{aligned}P\{|1 - \sum_{i=1}^{n} W_{n,i}(X)| > \varepsilon\} &\leq P\{\sum_{i=1}^{n} W_{n,i}(X) \neq 1\} \\ = P\{\sum_{i=1}^{n} K\left(\frac{X - X_i}{h_n}\right) = 0\} \\ = P\{\sum_{i=1}^{n} I_{\{X_i \in X + S_{0,R \cdot h_n}\}} = 0\} \\ \overset{\text{Unabh. + Fubini}}{=} \int P\{\sum_{i=1}^{n} I_{\{X_i \in x + S_{0,R \cdot h_n}\}} = 0\} \, P_X(dx) \\ \overset{\text{Unabh.}}{=} \int \prod_{i=1}^{n} P\{X_i \notin x + S_{0,R \cdot h_n}\} \, P_X(dx) \\ \overset{\text{identische Verteiltheit}}{=} \int (1 - P_X(x + S_{0,R \cdot h_n}))^n \, P_X(dx)\end{aligned}$$

Sei S beliebig (beschränkte) Kugel um 0. Dann folgt weiter:

$$
\begin{aligned}
& P\{1 - \sum_{i=1}^{n} W_{n,i}(X)| > \varepsilon\} \\
& \leq \int_S (1 - P_X(x + X_{0,R\cdot h_n}))^n P_X(dx) + P_X(S^c) \\
& \overset{1-x\leq e^{-x}}{\leq} \int_S e^{-n\cdot P_X(x+S_{0,R\cdot h_n})} P_X(dx) + P_{X(S^c)} \\
& \leq \max_{z\in\mathbb{R}}(z\cdot e^{-z}) \cdot \int_S \frac{1}{n\cdot P_X(x+S_{0,R\cdot h_n})} P_X(dx) + P_X(S^c).
\end{aligned}
$$

Wir zeigen im Folgenden:

$$
\int_S \frac{1}{n\cdot P_X(x+S_{0,R\cdot h_n})} P_X(dx) \leq \frac{\tilde{c}}{n\cdot h_n^d} \tag{4.10}
$$

für eine von S abhängige Konstante $\tilde{c} > 0$.

Wegen $n \cdot h_n^d \to \infty$ folgt daraus mit $S \uparrow \mathbb{R}^d$ die Behauptung.

Zum Nachweis von (4.10) wählen wir

$$
z_1 + S_{0,R\cdot h_n/2}, \ldots, z_{M_n} + S_{0,R\cdot h_n/2}
$$

so, dass

$$
S \subseteq \bigcup_{j=1}^{M_n} z_j + S_{0,R\cdot h_n/2}
$$

und

$$
M_n \leq \frac{\tilde{c}}{h_n^d}
$$

(z.B. M_n= maximal Anzahl von Punkten in S, die paarweise Abstand $\geq R\cdot h_n/2$ haben (womit Volumina von Kugeln um diese Punkte mit Radius $R\cdot h_n/4$ in um "'$R\cdot h_n/4$"' -vergrößerte Kugel S enthalten und paarweise disjunkt sind)).

Dann gilt:

$$
\begin{aligned}
&\int_S \frac{1}{n \cdot P_X(x + S_{0,R \cdot h_n})} P_X(dx) \\
&\leq \sum_{j=1}^{M_n} \int_{z_j + S_{0,R \cdot h_n/2}} \frac{1}{n \cdot P_X(x + S_{0,R \cdot h_n})} P_X(dx) \\
&\leq \sum_{j=1}^{M_n} \int_{z_j + S_{0,R \cdot h_n/2}} \frac{1}{n \cdot P_X(z_j + S_{0,R \cdot h_n/2})} P_X(dx) \\
&\text{da für } x \in z_{j+S_{0,R \cdot h_n/2}} \text{ gilt } x + S_{0,R \cdot h_n} \supseteq z_j + S_{0,R \cdot h_n/2} \\
&= \sum_{j=1}^{M_n} P_X(z_j + S_{0,R \cdot h_n/2}) \cdot \frac{1}{n \cdot P_X(z_j + S_{0,R \cdot h_n/2})} \\
&\leq \frac{M_n}{n} \leq \frac{\tilde{c}}{n \cdot h_n^d} \quad \rightsquigarrow (4.10) \rightsquigarrow \text{ (iv)}.
\end{aligned}
$$

Nachweis von (v):
Es gilt:

$$
\begin{aligned}
\sum_{i=1}^n W_{n,i}^2(x) &= \sum_{i=1}^n \frac{K\left(\frac{x - X_i}{h_n}\right)^2}{\left(\sum_{j=1}^n K\left(\frac{x - X_j}{h_n}\right)\right)^2} = \sum_{i=1}^n \frac{(I_{\{X_i \in x + S_{0,R \cdot h_n}\}})^2}{\left(\sum_{j=1}^n I_{\{X_j \in x + S_{0,R \cdot h_n}\}}\right)^2} \\
&= \sum_{i=1}^n \frac{I_{\{X_i \in x + S_{0,R \cdot h_n}\}}}{\left(\sum_{j=1}^n I_{X_j \in x + S_{0,R \cdot h_n}}\right)^2} \\
&= \frac{1}{\sum_{j=1}^n I_{\{X_j \in x + S_{0,R \cdot h_n}\}}} \cdot I_{\{\sum_{j=1}^n I_{\{X_j \cdot \in x + S_{0,R \cdot h_n}\}} > 0\}}.
\end{aligned}
$$

Damit folgt für jede (beschränkte) Kugel S um 0:

$$
\begin{aligned}
&E\{\sum_{i=1}^{n} W_{n_i}(X)^2\} \\
\overset{\text{Fubini}}{=}& \int E\{\sum_{i=1}^{n} W_{n_i}(x)^2\} P_X(dx) \\
\overset{\text{s.o.}}{=}& P_X(S^c) + \int_S E\{\frac{1}{\sum_{j=1}^{n} I_{\{X_j \in x+S_{0,R\cdot h_n}\}}} \cdot I_{\{\sum_{j=1}^{n} I_{\{X_j \in x+S_{0,R\cdot h_n}\}}>0\}}\} P_X(dx) \\
\overset{\text{Lemma 4.6}}{\leq}& P_X(S^c) + \int_S \frac{2}{(n+1)\cdot P_X(x+S_{0,R\cdot h_n})} P_X(dx) \\
\overset{(4.10)}{\leq}& P_X(S^c) + \frac{n}{n+1} \cdot 2 \cdot \frac{\tilde{c}}{n \cdot h_n^d} \\
\to& P_X(S^c) \quad \text{für } n \to \infty, \text{ da } n \cdot h_n^d \to \infty \, (n \to \infty).
\end{aligned}
$$

Mit $S \uparrow \mathbb{R}^d$ folgt (v). □

4.4 Ein Slow-Rate-Resultat

In diesem Unterkapitel zeigen wir, dass ohne Regularitätsvoraussetzungen an die zugrunde liegende Verteilung in der nichtparametrischen Regression eine nichttriviale Aussage zur Konvergenzgeschwindigkeit nicht herleitbar ist.

Die folgt aus:
Satz 4.7. *Sei $(m_n)_{n\in\mathbb{N}}$ eine beliebige Folge von Schätzfunktionen. Dann existiert zu jeder monoton gegen Null fallenden Folge $(a_n)_{n\in\mathbb{N}}$ nichtnegativ reeller Zahlen eine Verteilung von (X, Y) mit den Eigenschaften*

1. *$X \sim U[0,1]$,*
2. *$Y = m(X)$,*
3. *m ist $\{0,1\}$-wertig*

für die darüberhinaus gilt:

$$\limsup_{n\to\infty} \frac{\mathbf{E}\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx)}{a_n} \geq 1.$$

Bemerkung. Wendet man Satz 4.7 mit $\sqrt{a_n}$ statt a_n an, so sieht man, dass in Satz 4.7 gilt:

$$\limsup_{n\to\infty} \frac{\mathbf{E}\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx)}{a_n} = \infty.$$

D.h., selbst wenn (X, Y) fehlerfrei und X auf $[0, 1]$ gleichverteilt ist, so existiert dennoch für jeden Regressionsschätzer eine Verteilung von (X, Y), für die der erwartete L_2-Fehler des Schätzers beliebig langsam gegen Null konvergiert.

Im Beweis von Satz 4.7 benötigen wir das folgende deterministische Lemma.

Lemma 4.8. *Zu jeder Folge $(a_n)_{n\in\mathbb{N}}$ mit*

$$\frac{1}{4} \geq a_1 \geq a_2 \geq \cdots \geq a_n \to 0 \quad (n \to \infty)$$

existiert eine Zähldichte $(p_j)_{j\in\mathbb{N}}$ so, dass für alle genügend großen n gilt:

$$\sum_{j=1}^{\infty} (1 - p_j)^n \cdot p_j \geq a_n.$$

Beweis. Setze

$$p_1 = 1 - 2a_1 \geq 0 \quad \text{und} \quad k_1 = 1$$

und wähle dann $p_2, p_3, \ldots$ und $1 = k_1 < k_2 < k_3 < \ldots$ so, dass für alle $n \in \mathbb{N}$ gilt:

$$\sum_{i=k_n+1}^{k_{n+1}} p_i = 2 \cdot (a_n - a_{n+1}) \quad (\geq 0)$$

und

$$0 \leq p_i \leq \frac{1}{2n} \quad \text{für } i > k_n.$$

Dann folgt

$$p_j \geq 0 \quad \text{und} \quad \sum_{j=1}^{\infty} p_j = p_1 + \sum_{n=1}^{\infty} 2 \cdot (a_n - a_{n+1}) = p_1 + 2 \cdot a_1 = 1,$$

wobei die vorletzte Gleichheit wegen $a_n \to 0$ $(n \to \infty)$ und der daraus folgenden Beziehung

$$\sum_{n=1}^{N} (a_n - a_{n+1}) = a_1 - a_{N+1} \to a_1 \quad (N \to \infty)$$

gilt.

Weiterhin erhalten wir

$$\begin{aligned}
\sum_{j=1}^{\infty}(1-p_j)^n \cdot p_j &\geq \sum_{j\in\mathbb{N}:p_j\leq 1/(2n)} (1-p_j)^n \cdot p_j \\
&\geq \left(1-\frac{1}{2n}\right)^n \cdot \sum_{j\in\mathbb{N}:p_j\leq 1/(2n)} p_j \\
&\geq \left(1-\frac{1}{2n}\right)^n \cdot \sum_{j=k_n+1}^{\infty} p_j \\
&= \left(1-\frac{1}{2n}\right)^n \cdot \sum_{i=n}^{\infty} 2\cdot(a_i - a_{i+1}) \\
&= \left(1-\frac{1}{2n}\right)^n \cdot 2 \cdot a_n \\
&\geq a_n
\end{aligned}$$

für n genügend groß , da

$$\left(1-\frac{1}{2n}\right)^n \cdot 2 = \sqrt{\left(1-\frac{1}{2n}\right)^{2n}} \cdot 2 \to \sqrt{\frac{1}{e}} \cdot 2 > 1 \quad (n\to\infty).$$

□

Beweis von Satz 4.7:

1. Schritt: Wir definieren uns in Abhängigkeit von einer Zähldichte $(p_j)_{j\in\mathbb{N}}$ und eines Parameters $c = (c_j)_{j\in\mathbb{N}} \in \{-1,1\}^{\mathbb{N}}$ eine Verteilung von (X,Y).

Dazu gehen wir folgendermaßen vor: Wir wählen

$$X \sim U[0,1] \quad \text{und} \quad Y = m^{(c)}(X),$$

wobei wir zur Definition von $m^{(c)}$ zunächst in Abhängigkeit der Zähldichte $(p_j)_{j\in\mathbb{N}}$ das Intervall $[0,1]$ in Intervalle A_j der Länge p_j partitionieren und dann setzen:

$$m^{(c)}(x) = \begin{cases} 1, & \text{falls } x \in A_j,\ c_j = 1, \\ -1, & \text{falls } x \in A_j,\ c_j = -1 \end{cases}$$

$(j \in \mathbb{N})$.

2. Schritt: Wir schätzen

$$\mathbf{E}\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx)$$

für die Verteilung aus dem 1. Schritt nach unten ab.

Setze dazu

$$\tilde{m}_n(x) = \frac{1}{p_j} \int_{A_j} m_n(z)\, \mathbf{P}_X(dz) \quad \text{für } x \in A_j,$$

d.h. $\tilde{m}_n$ ist die L_2-Projektion von m_n auf die Menge aller bzgl. $(A_j)_{j\in\mathbb{N}}$ stückweise konstanten Funktionen.

Dann gilt

$$\begin{aligned}
&\int_{A_j} |m_n(x) - m^{(c)}(x)|^2 \mathbf{P}_X(dx) \\
&= \int_{A_j} |m_n(x) - \tilde{m}_n(x)|^2 \mathbf{P}_X(dx) + \int_{A_j} |\tilde{m}_n(x) - m^{(c)}(x)|^2 \mathbf{P}_X(dx),
\end{aligned}$$

da wegen $\tilde{m}_n - m^{(c)}$ konstant auf A_j für $x_j \in A_j$ beliebig gilt

$$\begin{aligned}
&\int_{A_j} (m_n(x) - \tilde{m}_n(x)) \cdot (\tilde{m}_n(x) - m^{(c)}(x))\, \mathbf{P}_X(dx) \\
&= (\tilde{m}_n(x_j) - m^{(c)}(x_j)) \cdot \int_{A_j} (m_n(x) - \tilde{m}_n(x))\, \mathbf{P}_X(dx) \\
&= (\tilde{m}_n(x_j) - m^{(c)}(x_j)) \cdot \left(\int_{A_j} m_n(x)\, \mathbf{P}_X(dx) - \int_{A_j} m_n(x)\, \mathbf{P}_X(dx) \right) \\
&= (\tilde{m}_n(x_j) - m^{(c)}(x_j)) \cdot 0 \\
&= 0.
\end{aligned}$$

Damit folgt

$$\begin{aligned}
\int_{A_j} |m_n(x) - m^{(c)}(x)|^2 \mathbf{P}_X(dx) &\geq \int_{A_j} |\tilde{m}_n(x) - m^{(c)}(x)|^2 \mathbf{P}_X(dx), \\
&= |\tilde{m}_n(x_j) - c_j|^2 \cdot p_j
\end{aligned}$$

für $x_j \in A_j$ beliebig aber fest.

Wir verwenden nun $\tilde{m}_n$, um c_j vorherzusagen, und setzen dazu

$$\hat{c}_{n,j} = \begin{cases} 1, & \text{falls } \tilde{m}_n(x_j) = \frac{1}{p_j} \cdot \int_{A_j} m_n(z)\, \mathbf{P}_X(dz) \geq 0, \\ -1, & \text{sonst.} \end{cases}$$

Im Falle $c_j = 1$ und $\hat{c}_{n,j} = -1$ (was $\tilde{m}_n(x_j) < 0$ impliziert) gilt dann

$$|\tilde{m}_n(x_j) - c_j| = c_j - \tilde{m}_n(x_j) \geq c_j - 0 = 1,$$

und im Falle $c_j = -1$ und $\hat{c}_{n,j} = 1$ (was $\tilde{m}_n(x_j) \geq 0$ impliziert) gilt

$$|\tilde{m}_n(x_j) - c_j| = \tilde{m}_n(x_j) - c_j \geq 0 - c_j = 1.$$

Daraus folgt

$$|\tilde{m}_n(x_j) - c_j|^2 \geq I_{\{\hat{c}_{n,j} \neq c_j\}}$$

und insgesamt

$$\int_{A_j} |m_n(x) - m^{(c)}(x)|^2 \mathbf{P}_X(dx) \geq p_j \cdot I_{\{\hat{c}_{n,j} \neq c_j\}}.$$

Damit ergibt sich nun

$$\begin{aligned}
&\mathbf{E} \int |m_n(x) - m^{(c)}(x)|^2 \mathbf{P}_X(dx) \\
&= \sum_{j=1}^{\infty} \mathbf{E} \int_{A_j} |m_n(x) - m^{(c)}(x)|^2 \mathbf{P}_X(dx) \\
&\geq \sum_{j=1}^{\infty} p_j \cdot \mathbf{P}\left\{\hat{c}_{n,j} \neq c_j\right\} \\
&\geq \sum_{j=1}^{\infty} \mathbf{P}\left\{\hat{c}_{n,j} \neq c_j, \mu_n(A_j) = 0\right\} \cdot p_j =: R_n(c),
\end{aligned}$$

wobei

$$\mu_n(A_j) = \frac{|\{1 \leq i \leq n \,:\, X_i \in A_j\}|}{n}$$

die empirische Verteilung zu $X_1, \ldots, X_n$ ist.

Hier wurde also der Fehler des Regressionsschätzers nach unten abgeschätzt durch den "Fehler" einer Vorhersagefunktion für c_j.

3. Schritt: Als nächstes schätzen wir

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \quad \text{bzw.} \quad R_n(c)$$

nach unten ab, indem wir c zufällig aus $\{-1, 1\}^{\mathbb{N}}$ wählen und über das Resultat mitteln.

Dazu seien $C_1, C_2, \ldots$ unabhängig identisch verteilte Zufallsvariablen mit

$$\mathbf{P}\{C_1 = 1\} = \frac{1}{2} = \mathbf{P}\{C_1 = -1\},$$

die unabhängig von $X_1, \ldots, X_n$ sind. Dann gilt für $C = (C_1, C_2, \ldots)$:

$$\begin{aligned}
\mathbf{E}\{R_n(C)\} &= \sum_{j=1}^{\infty} \mathbf{P}\{\hat{c}_{n,j} \neq C_j, \mu_n(A_j) = 0\} \cdot p_j \\
&= \sum_{j=1}^{\infty} \mathbf{E}\left\{\mathbf{P}\left\{\hat{c}_{n,j} \neq C_j, \mu_n(A_j) = 0 \middle| X_1, \ldots, X_n\right\}\right\} \cdot p_j \\
&= \sum_{j=1}^{\infty} \mathbf{E}\left\{I_{\{\mu_n(A_j)=0\}} \cdot \mathbf{P}\left\{\hat{c}_{n,j} \neq C_j \middle| X_1, \ldots, X_n\right\}\right\} \cdot p_j.
\end{aligned}$$

Im Falle $\mu_n(A_j) = 0$ gilt $X_1 \notin A_j, \ldots, X_n \notin A_j$, was impliziert, dass (X_1, Y_1), $\ldots$, (X_n, Y_n) (und damit auch $\hat{c}_{n,j}$) unabhängig von C_j ist. In diesem Fall gilt aber

$$\begin{aligned}
&\mathbf{P}\left\{\hat{c}_{n,j} \neq C_j \middle| X_1, \ldots, X_n\right\} \\
&= \mathbf{E}\left\{\mathbf{P}\left\{\hat{c}_{n,j} \neq C_j \middle| (X_1, Y_1), \ldots, (X_n, Y_n)\right\} \middle| X_1, \ldots, X_n\right\} \\
&= \mathbf{E}\left\{\frac{1}{2} \middle| X_1, \ldots, X_n\right\} = \frac{1}{2},
\end{aligned}$$

und wir erhalten

$$\begin{aligned}
\mathbf{E}\{R_n(C)\} &= \sum_{j=1}^{\infty} \frac{1}{2} \cdot \mathbf{P}\{\mu_n(A_j) = 0\} \cdot p_j \\
&= \sum_{j=1}^{\infty} \frac{1}{2} \cdot \mathbf{P}\{X_1 \notin A_j, \ldots, X_n \notin A_j\} \cdot p_j \\
&= \frac{1}{2} \cdot \sum_{j=1}^{\infty} (1 - p_j)^n \cdot p_j.
\end{aligned}$$

Wegen

$$R_n(C) \leq \sum_{j=1}^{\infty} \mathbf{P}\{\mu_n(A_j) = 0\} \cdot p_j = \sum_{j=1}^{\infty} (1 - p_j)^n \cdot p_j$$

gilt darüberhinaus

$$\frac{R_n(C)}{\mathbf{E}\{R_n(C)\}} \leq \frac{\sum_{j=1}^{\infty} (1 - p_j)^n \cdot p_j}{\frac{1}{2} \cdot \sum_{j=1}^{\infty} (1 - p_j)^n \cdot p_j} \leq 2.$$

Damit ist das Lemma von Fatou anwendbar, und wir erhalten

$$\mathbf{E}\left\{\limsup_{n\to\infty} \frac{R_n(C)}{\mathbf{E}\{R_n(C)\}}\right\} \geq \limsup_{n\to\infty} \mathbf{E}\left\{\frac{R_n(C)}{\mathbf{E}\{R_n(C)\}}\right\} = 1.$$

Da nun der Wert im Mittel größer oder gleich Eins ist, muss insbesondere irgendeiner der (zufälligen) Werte ebenfalls größer oder gleich Eins sein. Also existiert ein $c \in \{-1,1\}^{\mathbb{N}}$ mit

$$\limsup_{n\to\infty} \frac{R_n(c)}{\frac{1}{2}\cdot\sum_{j=1}^{\infty}(1-p_j)^n\cdot p_j} = \limsup_{n\to\infty} \frac{R_n(c)}{\mathbf{E}\{R_n(C)\}} \geq 1.$$

Mit Lemma 4.8 angewandt auf $a_n/2$, wobei wir den Anfang der Folge abändern so dass die Werte alle kleiner oder gleich $1/4$ sind, folgt daraus die Behauptung. □

4.5 Konvergenzgeschwindigkeit des Kernschätzers

Ziel im Folgenden ist die Abschätzung des erwarteten L_2-Fehlers

$$\mathbf{E}\int |m_n(x)-m(x)|^2 \mathbf{P}_X(dx)$$

im Falle des sogenannten Kernschätzers

$$m_n(x) = \frac{\sum_{i=1}^n Y_i \cdot K\left(\frac{x-X_i}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{x-X_j}{h_n}\right)}$$

mit naivem Kern $K = 1_{S_1(0)}$ und Bandbreite $h_n > 0$.

Dabei machen wir die folgenden Regularitätsannahmen an die zugrundeliegende Verteilung:

1. Beschränktheitsannahme an X.
2. Beschränktheitsannahme an
$$\begin{aligned}\mathbf{Var}\{Y|X=x\} &= \mathbf{E}\left\{(Y-\mathbf{E}\{Y|X=x\})^2\middle|X=x\right\}\\ &= \mathbf{E}\left\{Y^2\middle|X=x\right\} - \left(\mathbf{E}\left\{Y\middle|X=x\right\}\right)^2.\end{aligned}$$
3. Glattheitsannahme an die Regressionsfunktion.

Zur Formalisierung der ersten Bedingungen fordern wir, dass der sogenannte *Support von X bzw. $\mathbf{P}_X$* definert durch

$$supp(\mathbf{P}_X) = \left\{x\in\mathbb{R}^d \middle| \forall \epsilon > 0 : \mathbf{P}_X(S_\epsilon(x)) > 0\right\}$$

beschränkt ist. Dieser hat die folgenden beiden Eigenschaften:

Lemma 4.9. *Ist $supp(\mathbf{P}_X)$ der Support der $\mathbb{R}^d$-wertigen Zufallsvariablen X, so gilt:*

a) $\mathbf{P}\{X \in supp(\mathbf{P}_X)\} = 1$.

b) $supp(\mathbf{P}_X)$ *ist abgeschlossen.*

Beweis. a) Wegen

$$S_{\epsilon/2}(z) \subseteq S_\epsilon(x) \quad \text{für jedes } z \in S_{\epsilon/2}(x)$$

folgt für $z \in S_{\epsilon/2}(x)$ aus $\mathbf{P}(S_\epsilon(x)) = 0$ immer $\mathbf{P}(S_{\epsilon/2}(z)) = 0$. Unter Verwendung dieser Beziehung sehen wir

$$\begin{aligned} supp(\mathbf{P}_X)^c &= \left\{x \in \mathbb{R}^d \middle| \exists \epsilon > 0 : \mathbf{P}_X(S_\epsilon(x)) = 0\right\} \\ &\subseteq \bigcup_{x \in supp(\mathbf{P}_X)^c \cap \mathbb{Q}^d, \epsilon \in \mathbb{Q}_+ \setminus \{0\}, \mathbf{P}_X(S_\epsilon(x))=0} S_\epsilon(x). \end{aligned}$$

Die rechte Seite ist eine abzählbare Vereinigung von $\mathbf{P}_X$-Nullmengen, und damit ist auch $supp(\mathbf{P}_X)^c$ eine $\mathbf{P}_X$-Nullmenge.

b) Ist $x \notin supp(\mathbf{P}_X)$, so gilt

$$\mathbf{P}_X(S_\epsilon(x)) = 0$$

für ein $\epsilon > 0$. Nach dem Beweis von a) impliziert dies aber $S_{\epsilon/2}(x) \subseteq supp(\mathbf{P}_X)^c$, also ist $supp(\mathbf{P}_X)^c$ offen. □

Nun gilt:
Satz 4.10. *Sei*

$$m_n(x) = \frac{\sum_{i=1}^n Y_i \cdot K\left(\frac{x-X_i}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{x-X_j}{h_n}\right)}$$

der Kernschätzer mit naivem Kern $K = 1_{S_1(0)}$ und Bandbreite $h_n > 0$.

Seien $C > 0$, $p \in (0, 1]$ und $\sigma > 0$. Dann gilt für jede Verteilung von (X, Y) mit

$$S := supp(\mathbf{P}_X) \quad \textit{ist beschränkt,} \tag{4.11}$$

$$\mathbf{Var}\{Y|X = x\} \leq \sigma^2 \quad \textit{für alle } x \in S \tag{4.12}$$

und

$$|m(x) - m(z| \leq C \cdot \|x - z\|^p \quad \textit{für alle } x, z \in S \tag{4.13}$$

die folgende Abschätzung für den erwarteten L_2-Fehler des Kernschätzers:

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_1 \cdot \frac{\sigma^2 + \sup_{z \in S} |m(z)|^2}{n \cdot h_n^d} + C^2 \cdot h_n^{2p}.$$

Hierbei ist c_1 eine nur von d und dem Durchmesser von $S = supp(\mathbf{P}_X)$ abhängende Konstante.

Im Beweis benötigen wir:
Lemma 4.11. *Ist $S = supp(\mathbf{P}_X)$ beschränkt, so gilt für eine nur von d und dem Durchmesser von S abhängende Konstante $\hat{c}$:*

$$\int_S \frac{1}{n \cdot \mathbf{P}_X(S_{h_n}(x))} \mathbf{P}_X(dx) \leq \frac{\hat{c}}{n \cdot h_n^d}.$$

Beweis. Wähle $l_n \leq \hat{c}/h_n^d$ Kugeln $S_{h_n/2}(z_1), \ldots, S_{h_n/2}(z_{l_n})$ mit Radius $h_n/2$ so, dass gilt

$$S \subseteq \cup_{l=1}^{l_n} S_{h_n/2}(z_l). \tag{4.14}$$

Wegen

$$S_{h_n/2}(z_l) \subseteq S_{h_n}(x) \tag{4.15}$$

für $x \in S_{h_n/2}(z_l)$ gilt dann

$$\begin{aligned}
\int_S \frac{1}{n \cdot \mathbf{P}_X(S_{h_n}(x))} \mathbf{P}_X(dx) &\overset{(4.14)}{\leq} \sum_{l=1}^{l_n} \int_{S_{h_n/2}(z_l)} \frac{1}{n \cdot \mathbf{P}_X(S_{h_n}(x))} \mathbf{P}_X(dx) \\
&\overset{(4.15)}{\leq} \sum_{l=1}^{l_n} \int_{S_{h_n/2}(z_l)} \frac{1}{n \cdot \mathbf{P}_X(S_{h_n/2}(z_l))} \mathbf{P}_X(dx) \\
&= \sum_{l=1}^{l_n} \frac{1}{n \cdot \mathbf{P}_X(S_{h_n/2}(z_l))} \cdot \mathbf{P}_X(S_{h_n/2}(z_l)) \\
&\leq \frac{l_n}{n} \leq \frac{\hat{c}}{n \cdot h_n^d}.
\end{aligned}$$

□

Beweis von Satz 4.10: Setze

$$\hat{m}_n(x) = \mathbf{E}\left\{m_n(x) \middle| X_1, \ldots, X_n\right\} = \frac{\sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right) \cdot m(X_i)}{\sum_{j=1}^n K\left(\frac{x - X_j}{h_n}\right)}.$$

Wegen

$$\begin{aligned}
&\mathbf{E}\left\{|m_n(x) - m(x)|^2 \middle| X_1, \ldots, X_n\right\} \\
&= \mathbf{E}\left\{|m_n(x) - \mathbf{E}\left\{m_n(x) \middle| X_1, \ldots, X_n\right\}|^2 \middle| X_1, \ldots, X_n\right\} \\
&\quad + \left|\mathbf{E}\left\{m_n(x) \middle| X_1, \ldots, X_n\right\} - m(x)\right|^2
\end{aligned}$$

erhalten wir unter Verwendung des Satzes von Fubini und der Definition der bedingten Erwartung analog zur Bias-Varianz-Zerlegung aus der Statistik die folgende Darstellung unseres Fehlers:

$$\begin{aligned}
&\mathbf{E}\int |m_n(x)-m(x)|^2 \mathbf{P}_X(dx)\\
&= \mathbf{E}\left\{\int \mathbf{E}\left\{|m_n(x)-m(x)|^2 \middle| X_1,\dots,X_n\right\} \mathbf{P}_X(dx)\right\}\\
&= \mathbf{E}\left\{\int |m_n(x)-\hat{m}_n(x)|^2 \mathbf{P}_X(dx)\right\} + \mathbf{E}\left\{\int |\hat{m}_n(x)-m(x)|^2 \mathbf{P}_X(dx)\right\}.
\end{aligned}$$

Hierbei ist der erste bzw. zweite Term auf der rechten Seite oben die erwartete integrierte Varianz bzw. der erwartete integrierte Bias des Schätzers.

Als erstes schätzen wir den erwarteten integrierten Bias des Schätzers ab. Dazu setzen wir

$$\mu_n(A) = \frac{|\{1 \le i \le n \,:\, X_i \in A\}|}{n}$$

und

$$B_n(x) = \{n \cdot \mu_n(S_{h_n}(x)) > 0\}.$$

Beachtet man, dass $K((x-X_i)/h_n) > 0$ nur gelten kann, sofern $\|x-X_i\| \le h_n$ ist, so erhält man unter Verwendung der Ungleichung von Jensen

$$\begin{aligned}
&|\hat{m}_n(x)-m(x)|^2\\
&= \left|\frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)\cdot(m(X_i)-m(x))}{\sum_{j=1}^n K\left(\frac{x-X_j}{h_n}\right)}\right|^2 \cdot I_{B_n(x)} + |m(x)|^2 \cdot I_{B_n(x)^c}\\
&\le \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)\cdot|m(X_i)-m(x)|^2}{\sum_{j=1}^n K\left(\frac{x-X_j}{h_n}\right)} \cdot I_{B_n(x)} + |m(x)|^2 \cdot I_{B_n(x)^c}\\
&\overset{(4.13)}{\le} \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)\cdot C^2\cdot\|X_i-x\|^{2p}}{\sum_{j=1}^n K\left(\frac{x-X_j}{h_n}\right)} \cdot I_{B_n(x)} + |m(x)|^2 \cdot I_{B_n(x)^c}\\
&\le C^2\cdot h_n^{2p} + |m(x)|^2 \cdot I_{B_n(x)^c},
\end{aligned}$$

bzw.

$$\begin{aligned}
&\mathbf{E}\int |\hat{m}_n(x)-m(x)|^2 \mathbf{P}_X(dx)\\
&\le C^{2p}\cdot h_n^{2p} + \sup_{z\in S}|m(z)|^2 \cdot \int \mathbf{P}\{n\cdot\mu_n(S_{h_n}(x))=0\}\mathbf{P}_X(dx).
\end{aligned}$$

Mit

$$\begin{aligned}
& \mathbf{P}\{n \cdot \mu_n(S_{h_n}(x)) = 0\} \\
= \quad & \mathbf{P}\left\{X_1 \notin S_{h_n}(x), \ldots, X_n \notin S_{h_n}(x)\right\} \\
= \quad & \mathbf{P}\left\{X_1 \notin S_{h_n}(x)\right\} \cdots \mathbf{P}\left\{X_n \notin S_{h_n}(x)\right\} \\
= \quad & \left(1 - \mathbf{P}_{X_1}\left(S_{h_n}(x)\right)\right)^n \\
\overset{1+x \le e^x}{\le} \quad & e^{-n \cdot \mathbf{P}_{X_1}(S_{h_n}(x))} \\
= \quad & n \cdot \mathbf{P}_{X_1}\left(S_{h_n}(x)\right) \cdot e^{-n \cdot \mathbf{P}_{X_1}(S_{h_n}(x))} \cdot \frac{1}{n \cdot \mathbf{P}_{X_1}\left(S_{h_n}(x)\right)} \\
\le \quad & \max_{z \ge 0}\left(z \cdot e^{-z}\right) \cdot \frac{1}{n \cdot \mathbf{P}_{X_1}\left(S_{h_n}(x)\right)} \\
\le \quad & \frac{1}{e} \cdot \frac{1}{n \cdot \mathbf{P}_{X_1}\left(S_{h_n}(x)\right)}
\end{aligned}$$

und Lemma 4.11 folgt daraus

$$\begin{aligned}
& \mathbf{E} \int |\hat{m}_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
& \le C^2 \cdot h_n^{2p} + \sup_{z \in S} |m(z)|^2 \cdot \int \frac{1}{e} \cdot \frac{1}{n \cdot \mathbf{P}_{X_1}\left(S_{h_n}(x)\right)} \mathbf{P}_X(dx) \\
& \le C^2 \cdot h_n^{2p} + \sup_{z \in S} |m(z)|^2 \cdot \frac{1}{e} \cdot \frac{\hat{c}}{n \cdot h_n^d}. \qquad (4.16)
\end{aligned}$$

Im Folgenden wird nun die integrierte Varianz abgeschätzt. Hierzu gilt unter Beachtung der Unabhängigkeit der Daten

$$\begin{aligned}
& \mathbf{E}\left\{|m_n(x) - \hat{m}_n(x)|^2 \middle| X_1, \ldots, X_n\right\} \\
\le \quad & \mathbf{E}\left\{ \left| \frac{\sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right) \cdot (Y_i - m(X_i))}{\sum_{j=1}^n K\left(\frac{x - X_j}{h_n}\right)} \right|^2 \middle| X_1, \ldots, X_n \right\} \\
= \quad & \frac{\sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right)^2 \cdot \mathbf{E}\left\{|Y_i - m(X_i)|^2 \middle| X_1, \ldots, X_n\right\}}{\left(\sum_{j=1}^n K\left(\frac{x - X_j}{h_n}\right)\right)^2} \\
\overset{K(z) \in \{0,1\}}{=} \quad & \frac{\sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right) \cdot \mathbf{E}\left\{|Y_i - m(X_i)|^2 \middle| X_i\right\}}{\left(\sum_{j=1}^n K\left(\frac{x - X_j}{h_n}\right)\right)^2} \\
\le \quad & \sup_{z \in S} \mathbf{Var}\{Y | X = z\} \cdot \frac{1}{\sum_{j=1}^n K\left(\frac{x - X_j}{h_n}\right)} \cdot I_{\{n \cdot \mu_n(S_{h_n}(x)) > 0\}}.
\end{aligned}$$

$\sum_{j=1}^n K\left(\frac{x-X_j}{h_n}\right)$ ist $b(n, \mathbf{P}_X(S_{h_n}(x)))$-verteilt. Nach Lemma 4.6 gilt daher

$$\mathbf{E}\left\{\frac{1}{\sum_{j=1}^n K\left(\frac{x-X_j}{h_n}\right)} \cdot I_{\{n\cdot\mu_n(S_{h_n}(x))>0\}}\right\} \leq \frac{2}{(n+1)\cdot \mathbf{P}_X(S_{h_n}(x))}.$$

Damit erhalten wir unter Beachtung von Lemma 4.11

$$\begin{aligned}
&\mathbf{E}\left\{\int |m_n(x)-\hat{m}_n(x)|^2 \mathbf{P}_X(dx)\right\} \\
&= \int \mathbf{E}\left\{\mathbf{E}\left\{|m_n(x)-\hat{m}_n(x)|^2 \middle| X_1,\ldots,X_n\right\}\right\} \mathbf{P}_X(dx) \\
&\leq \sigma^2 \cdot \int \mathbf{E}\left\{\frac{1}{\sum_{j=1}^n K\left(\frac{x-X_j}{h_n}\right)} \cdot I_{\{n\cdot\mu_n(S_{h_n}(x))>0\}}\right\} \mathbf{P}_X(dx) \\
&\leq \sigma^2 \cdot \int \frac{2}{(n+1)\cdot \mathbf{P}_X(S_{h_n}(x))} \mathbf{P}_X(dx) \\
&\leq \sigma^2 \cdot 2 \cdot \frac{\hat{c}}{n\cdot h_n^d}. \qquad (4.17)
\end{aligned}$$

Aus (4.16) und (4.17) folgt nun die Behauptung. □

Um unter den Voraussetzungen in Satz 4.10 einen möglichst kleinen Fehler zu erhalten, muss man h_n so wählen, dass

$$c_1 \cdot \frac{\sigma^2 + \sup_{z\in S}|m(z)|^2}{n\cdot h_n^d} + C^2 \cdot h_n^{2p}$$

möglichst klein wird. Dabei darf h_n nicht zu klein sein, damit der Varianz-Term

$$\frac{1}{n\cdot h_n^d}$$

möglichst klein wird, andererseits darf h_n aber auch nicht zu groß sein, damit der Bias-Term

$$C^2 \cdot h_n^{2p}$$

nicht zu groß wird.

Zur Bestimmung des im Hinblick auf die Minimierung der Fehlerabschätzung in Satz 4.10 optimalen h_n betrachten wird die Minimierung von

$$f(u) = \frac{A}{n\cdot u^d} + C^2 u^{2p}.$$

Nullsetzen der Ableitung führt auf

$$0 = f'(u) = \frac{-d \cdot A}{n} \cdot u^{-(d+1)} + C^2 \cdot 2p \cdot u^{2p-1}$$

bzw.

$$u^{d+2p} = \frac{d \cdot A}{2p \cdot C^2 \cdot n}$$

bzw.

$$u = \left(\frac{d \cdot A}{2p \cdot C^2 \cdot n} \right)^{1/(2p+d)}$$

sowie

$$\begin{aligned} \min_{u \in \mathbb{R}_+} f(u) &= f\left(\left(\frac{d \cdot A}{2p \cdot C^2 \cdot n} \right)^{1/(2p+d)} \right) \\ &= \frac{A}{n} \cdot \left(\frac{2p \cdot C^2 \cdot n}{d \cdot A} \right)^{d/(2p+d)} + C^2 \cdot \left(\frac{d \cdot A}{2p \cdot C^2 \cdot n} \right)^{2p/(2p+d)} \\ &= \left(\frac{A}{n} \right)^{2p/(2p+d)} \cdot C^{2d/(2p+d)} \cdot \left(\frac{2p}{d} \right)^{d/(2p+d)} \\ &\quad + C^{2d/(2p+d)} \cdot \left(\frac{A}{n} \right)^{2p/(2p+d)} \cdot \left(\frac{d}{2p} \right)^{2p/(2p+d)}. \end{aligned}$$

Damit folgt:

Korollar 4.12. *Unter den Voraussetzung von Satz 4.10 wird die dort angegebene Schranke für den Fehler minimal für*

$$h_n = \left(\frac{d \cdot c_1 \cdot (\sigma^2 + \sup_{z \in S} |m(z)|^2)}{2p \cdot C^2 \cdot n} \right)^{1/(2p+d)},$$

und mit dieser Bandbreite erhält man

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq \bar{c} \cdot \left(\frac{\sigma^2 + \sup_{z \in S} |m(z)|^2}{n} \right)^{2p/(2p+d)} \cdot C^{2d/(2p+d)}.$$

Bemerkung: Die obere rechte Seite ist monoton wachsend in σ und C und monoton fallend in n.

4.6 Minimax-Konvergenzraten

4.6.1 Motivation

Gemäß dem letzten Abschnitt gilt für den Kernschätzer m_n im Falle einer Lipschitz-stetigen Regressionsfunktion und beschränkten Daten

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) = O\left(n^{-\frac{2}{2+d}}\right).$$

Es stellt sich die Frage, ob man diese Rate durch Wahl eines anderen Schätzverfahrens verbessern kann bzw. was unter den obigen Voraussetzungen die optimale Konvergenzrate ist.

Um dies genauer zu formulieren, betrachten wir für eine feste Klasse $\mathcal{D}$ von Verteilungen von (X, Y) den maximal erwarteten L_2-Fehler

$$\sup_{(X,Y)\in\mathcal{D}} \mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \tag{4.18}$$

innerhalb dieser Klasse, wobei der Regressionsschätzer eine Stichprobe (X_1, Y_1), $\ldots$, (X_n, Y_n) der Verteilung von (X, Y) bekommt. Ziel im Folgenden ist es, m_n so zu wählen, dass (4.18) minimal wird, d.h. genauer, dass (4.18) asymptotisch wie

$$\inf_{\tilde{m}_n} \sup_{(X,Y)\in\mathcal{D}} \mathbf{E} \int |\tilde{m}_n(x) - m(x)|^2 \mathbf{P}_X(dx) \tag{4.19}$$

gegen Null konvergiert, wobei obiges Infimum über alle Regressionsschätzer $\tilde{m}_n$ gebildet wird.

Dies lässt sich als Zwei-Parteien-Spiel deuten: Wir spielen gegen die Natur. Im 1. Schritt wählt die Natur eine Verteilung aus $\mathcal{D}$ und gibt uns eine Stichprobe dieser Verteilung. Anschließend wählen wir einen Schätzer um die zugehörige Regressionsfunktion zu schätzen. Dabei verfolgt die Natur das Ziel, dass die Schätzung möglichst schlecht wird, und wir verfolgen das Ziel, dass diese möglichst gut wird. Spielen nun beide Spieler optimal, so ist gerade (4.19) der zu erwartende L_2-Fehler.

Die obigen Überlegungen formalisieren wir in

Definition 4.13. *Sei $\mathcal{D}$ eine Klasse von Verteilungen von (X, Y) und $(a_n)_{n\in\mathbb{N}}$ eine Folge positiver reeller Zahlen.*

a) $(a_n)_{n\in\mathbb{N}}$ *heißt* **untere Minimax-Konvergenzrate für** $\mathcal{D}$*, falls gilt*

$$\liminf_{n\to\infty} \inf_{m_n} \sup_{(X,Y)\in\mathcal{D}} \frac{\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx)}{a_n} = C_1 > 0.$$

b) $(a_n)_{n\in\mathbb{N}}$ *heißt* **obere Minimax-Konvergenzrate für** $\mathcal{D}$*, falls für ein Schätzverfahren* m_n *gilt*

$$\limsup_{n\to\infty} \sup_{(X,Y)\in\mathcal{D}} \frac{\mathbf{E}\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx)}{a_n} = C_2 < \infty.$$

c) $(a_n)_{n\in\mathbb{N}}$ *heißt* **optimale Minimax-Konvergenzrate für** $\mathcal{D}$*, falls* $(a_n)_{n\in\mathbb{N}}$ *sowohl untere als auch obere Minimax-Konvergenzrate für* $\mathcal{D}$ *ist.*

Aus Kapitel 3 wissen wir: Ist $p \in (0,1]$, sind $c_1, c_2, c_3 > 0$ und ist $\mathcal{D}$ die Klasse aller Verteilungen von (X,Y) mit $X \in [0,1]^d$ *f.s.*, $\sup_{x\in[0,1]^d} \mathbf{Var}\{Y|X=x\} \le c_1$, $\sup_{x\in[0,1]^d} |m(x)| \le c_2$ und $|m(x) - m(z)| \le c_3 \cdot \|x - z\|^p$ für alle $x, z \in [0,1]^d$, so ist

$$\left(n^{-\frac{2p}{2p+d}}\right)_{n\in\mathbb{N}}$$

obere Minimax-Konvergenzrate für $\mathcal{D}$.

Im Folgenden zeigen wir, dass dies sogar die optimale Minimax-Konvergenzrate für $\mathcal{D}$ ist, so dass der Kernschätzer in diesem Sinne sogar ein "optimales" Schätzverfahren ist.

4.6.2 Eine untere Minimax-Konvergenzrate

Um nachzuweisen, dass $\left(n^{-\frac{2p}{2p+d}}\right)_{n\in\mathbb{N}}$ optimale Minimax-Konvergenzrate für $\mathcal{D}$ ist, genügt es aufgrund von Korollar 4.12 für $\tilde{\mathcal{D}} \subseteq \mathcal{D}$ geeignet zu zeigen, dass $\left(n^{-\frac{2p}{2p+d}}\right)_{n\in\mathbb{N}}$ eine untere Minimax-Konvergenzrate für $\tilde{\mathcal{D}}$ ist.

Im Fall $p \le 1$ betrachten wir als Unterklasse von $\mathcal{D}$:
Definition 4.14. *Für* $p, C > 0$ *sei* $\mathcal{D}^{(p,C)}$ *die Klasse aller Verteilungen von* (X,Y) *mit:*

1. $X \sim U([0,1]^d)$
2. $Y = m(X) + N$ *wobei* $N \sim N(0,1)$ *und* X, N *unabhängig*
3. m (p,C)*-glatt.*
4. $|m(x)| \le 1$ *für* $x \in [0,1]^d$.

Das Hauptresultat von diesem Abschnitt ist

Satz 4.15. *Seien $p, C > 0$ und $\mathcal{D}^{(p,C)}$ definiert wie oben. Dann ist*

$$\left(n^{-\frac{2p}{2p+d}}\right)_{n\in\mathbb{N}} \tag{4.20}$$

eine untere Minimax-Konvergenzrate für $\mathcal{D}^{(p,C)}$.

Im Falle $p \leq 1$ ist damit (4.20) die optimale Minimax-Konvergenzrate für die Klasse $\mathcal{D}$ aus Abschnitt 4.6.1.

Im Beweis von Satz 4.15 benötigen wir:
Lemma 4.16. *Sei $u \in \mathbb{R}^l$ und sei C eine $\{-1,1\}$-wertige Zufallsvariable mit*

$$\mathbf{P}\{C = 1\} = \frac{1}{2} = \mathbf{P}\{C = -1\}.$$

Sei N eine $\mathbb{R}^l$-wertige standardnormalverteilte Zufallsvariable unabhängig von C, d.h. es gilt $N = (N^{(1)}, \ldots, N^{(l)})$ wobei $N^{(1)}$, …, $N^{(l)}$ reellwertige unabhängig standardnormalverteilte Zufallsvariablen sind, die unabhängig von C sind. Setze

$$Z = C \cdot u + N$$

und betrachte das Problem, ausgehend von Z den Wert von C vorherzusagen. Dann gilt

$$L^* := \min_{g:\mathbb{R}^l\to\{-1,1\}} \mathbf{P}\{g(Z) \neq C\} = \Phi(-\|u\|),$$

wobei Φ die Verteilungsfunktion von $N(0,1)$ ist.

Beweis. Für $g : \mathbb{R}^l \to \{-1, 1\}$ beliebig gilt wegen N, C unabhängig

$$\begin{aligned}
&\mathbf{P}\left\{g(Z) \neq C\right\} \\
&= \mathbf{P}\left\{g(C \cdot u + N) \neq C\right\} \\
&= \mathbf{P}\left\{g(C \cdot u + N) \neq C, C = 1\right\} + \mathbf{P}\left\{g(C \cdot u + N) \neq C, C = -1\right\} \\
&= \mathbf{P}\left\{g(-u + N) = -1, C = 1\right\} + \mathbf{P}\left\{g(u + N) = 1, C = -1\right\} \\
&= \mathbf{P}\left\{g(-u + N) = -1\right\} \cdot \mathbf{P}\left\{C = 1\right\} + \mathbf{P}\left\{g(u + N) = 1\right\} \cdot \mathbf{P}\left\{C = -1\right\} \\
&= \frac{1}{2} \cdot \mathbf{P}\left\{g(-u + N) = -1\right\} + \frac{1}{2} \cdot \mathbf{P}\left\{g(u + N) = 1\right\}.
\end{aligned}$$

Sei φ die Dichte von N, d.h. für $v = (v^{(1)}, \ldots, v^{(l)})$ gilt

$$\varphi(v) = \prod_{i=1}^{l} \frac{1}{\sqrt{2 \cdot \pi}} \cdot e^{-\frac{|v^{(i)}|^2}{2}} = (2 \cdot \pi)^{-l/2} \cdot e^{-\|v\|^2/2}.$$

Dann hat $u + N$ die Dichte $\varphi(v - u)$, und $-u + N$ hat die Dichte $\varphi(v + u)$ (wie man z.B. durch Ableiten der jeweiligen Verteilungsfunktion sieht).

Damit folgt

$$\begin{aligned}
&\mathbf{P}\{g(Z) \neq C\} \\
&= \frac{1}{2} \cdot \int I_{\{g(z)=-1\}} \cdot \varphi(z-u)\, dz + \frac{1}{2} \cdot \int I_{\{g(z)=1\}} \cdot \varphi(z+u)\, dz \\
&= \frac{1}{2} \cdot \int \left(I_{\{g(z)=-1\}} \cdot \varphi(z-u) + I_{\{g(z)=1\}} \cdot \varphi(z+u) \right) dz.
\end{aligned}$$

Der obige Ausdruck wird minimal für

$$g^*(z) = \begin{cases} 1, & \text{falls } \varphi(z-u) > \varphi(z+u), \\ -1, & \text{sonst.} \end{cases}$$

Wegen

$$\begin{aligned}
\varphi(z-u) > \varphi(z+u) &\Leftrightarrow (2\cdot\pi)^{-l/2} \cdot e^{-\|z-u\|^2/2} > (2\cdot\pi)^{-l/2} \cdot e^{-\|z+u\|^2/2} \\
&\Leftrightarrow \|z+u\|^2 > \|z-u\|^2 \\
&\Leftrightarrow < z, u > \quad > 0
\end{aligned}$$

gilt

$$g^*(z) = \begin{cases} 1, & \text{falls } < z, u > \quad > 0, \\ -1, & \text{sonst} \end{cases}$$

und wir erhalten analog zu oben

$$\begin{aligned}
L^* &= \mathbf{P}\{g^*(Z) \neq C\} \\
&= \mathbf{P}\{g^*(Cu+N) \neq C, C=1\} + \mathbf{P}\{g^*(Cu+N) \neq C, C=-1\} \\
&= \frac{1}{2} \cdot \mathbf{P}\{g^*(u+N) = -1\} + \frac{1}{2} \cdot \mathbf{P}\{g^*(-u+N) = 1\} \\
&= \frac{1}{2} \cdot \mathbf{P}\{< u+N, u > \quad \leq 0\} + \frac{1}{2} \cdot \mathbf{P}\{< -u+N, u > \quad > 0\} \\
&= \frac{1}{2} \cdot \mathbf{P}\left\{\|u\|^2 + < u, N > \quad \leq 0\right\} + \frac{1}{2} \cdot \mathbf{P}\left\{-\|u\|^2 + < u, N > \quad > 0\right\} \\
&= \frac{1}{2} \cdot \mathbf{P}\left\{< u, N > \quad \leq -\|u\|^2\right\} + \frac{1}{2} \cdot \mathbf{P}\left\{< u, N > \quad > \|u\|^2\right\}.
\end{aligned}$$

Ist nun $u = 0$, so folgt

$$L^* = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 0 = \frac{1}{2} = \Phi(-\|u\|).$$

Ist $\|u\| \neq 0$, so ist

$$< \frac{u}{\|u\|}, N >$$

als Linearkombination von unabhängigen standardnormalverteilten Zufallsvariablen wegen $\mathbf{E}\{< \frac{u}{\|u\|}, N >\} = 0$ und $\mathbf{Var}\{< \frac{u}{\|u\|}, N >\} = \|u\|^2/\|u\|^2 = 1$ standardnormalverteilt, und es folgt

$$\begin{aligned} L^* &= \frac{1}{2} \cdot \mathbf{P}\left\{< \frac{u}{\|u\|}, N > \quad \leq -\|u\|\right\} + \frac{1}{2} \cdot \mathbf{P}\left\{< \frac{u}{\|u\|}, N > \quad > \|u\|\right\} \\ &= \frac{1}{2} \cdot \Phi(-\|u\|) + \frac{1}{2} \cdot (1 - \Phi(\|u\|)) \\ &= \Phi(-\|u\|). \end{aligned}$$

□

Beweis von Satz 4.15: Wir beweisen Satz 4.15 nur für $d = 1$, der allgemeine Fall wird in den Übungen behandelt.

1. Schritt: In Abhängigkeit von n definieren wir Unterklassen von $\mathcal{D}^{(p,C)}$.

Dazu setzen wir

$$M_n = \lceil (C^2 \cdot n)^{\frac{1}{2p+1}} \rceil$$

(mit $\lceil x \rceil = \inf\{z \in \mathbb{Z} : z \geq x\}$) und partitionieren $[0,1]$ in M_n äquidistante Intervalle $A_{n,j}$ der Länge $1/M_n$. $a_{n,j}$ sei der Mittelpunkt von $A_{n,j}$.

Sodann wählen wir ein beschränktes $\bar{g} : \mathbb{R} \to \mathbb{R}$ mit

$$supp(\bar{g}) \subseteq (-1/2, 1/2), \quad \int \bar{g}^2(x)\,dx > 0 \quad \text{und} \quad \bar{g} \ (p, 2^{\beta-1})\text{-glatt}$$

(wobei wir die letzte Bedingung durch Reskalierung einer genügend oft differenzierbaren Funktion erfüllen können), und setzen dann

$$g(x) = C \cdot \bar{g}(x) \quad (x \in \mathbb{R}).$$

Dann gilt

$$supp(g) \subseteq (-1/2, 1/2), \quad \int g^2(x)\,dx = C^2 \cdot \int \bar{g}^2(x)\,dx > 0$$

und

$$g \ (p, C \cdot 2^{\beta-1})\text{-glatt}.$$

Für $c_n = (c_{n,1}, \ldots, c_{n,M_n}) \in \{-1,1\}^{M_n} =: \mathcal{C}_n$ setzen wir

$$m^{(c_n)}(x) = \sum_{j=1}^{M_n} c_{n,j} \cdot g_{n,j}(x)$$

wobei

$$g_{n,j}(x) = M_n^{-p} \cdot g(M_n(x - a_{n,j})).$$

Dann ist $m^{(c_n)}$ (p, C)-glatt, wie wir wie folgt sehen:

(i) Für $x, z \in A_{n,i}$ gilt

$$\begin{aligned}
&\left| \left(\frac{d}{dx}\right)^k m^{(c_n)}(x) - \left(\frac{d}{dx}\right)^k m^{(c_n)}(z) \right| \\
&= |c_{n,i}| \cdot \left| \left(\frac{d}{dx}\right)^k g_{n,i}(x) - \left(\frac{d}{dx}\right)^k g_{n,i}(z) \right| \\
&= 1 \cdot M_n^{-p} \cdot M_n^k \cdot C \cdot 2^{\beta-1} \left| M_n(x - a_{n,i}) - M_n(z - a_{n,i}) \right|^\beta \\
&\leq C \cdot 2^{\beta-1} \cdot |x - z|^\beta \leq C \cdot |x - z|^\beta.
\end{aligned}$$

(ii) Für $x \in A_{n,i}$ und $z \in A_{n,j}$ mit $i \neq j$ seien $\tilde{x}$ bzw. $\tilde{z}$ die Punkte am Rand von $A_{n,i}$ bzw. $A_{n,j}$ in Richtung von z bzw. x. Da $g_{n,i}$ und $g_{n,j}$ (p, C)-glatt sind (s.o.) und am Rand verschwinden gilt dann

$$\left(\frac{d}{dx}\right)^k g_{n,i}(\tilde{x}) = 0 = \left(\frac{d}{dx}\right)^k g_{n,j}(\tilde{z}).$$

Unter Verwendung des Resultates aus Schritt (i) folgt dann

$$\begin{aligned}
&\left| \left(\frac{d}{dx}\right)^k m^{(c_n)}(x) - \left(\frac{d}{dx}\right)^k m^{(c_n)}(z) \right| \\
&= \left| c_{n,i} \cdot \left(\frac{d}{dx}\right)^k g_{n,i}(x) - c_{n,j} \cdot \left(\frac{d}{dx}\right)^k g_{n,j}(z) \right| \\
&\leq |c_{n,i}| \cdot \left| \left(\frac{d}{dx}\right)^k g_{n,i}(x) \right| + |c_{n,j}| \cdot \left| \left(\frac{d}{dx}\right)^k g_{n,j}(z) \right| \\
&= \left| \left(\frac{d}{dx}\right)^k g_{n,i}(x) - \left(\frac{d}{dx}\right)^k g_{n,i}(\tilde{x}) \right| + \left| \left(\frac{d}{dx}\right)^k g_{n,j}(z) - \left(\frac{d}{dx}\right)^k g_{n,j}(\tilde{z}) \right| \\
&\leq C \cdot 2^{\beta-1} \cdot |x - \tilde{x}|^\beta + C \cdot 2^{\beta-1} \cdot |z - \tilde{z}|^\beta \\
&= C \cdot 2^\beta \cdot \left(\frac{1}{2} \cdot |x - \tilde{x}|^\beta + \frac{1}{2} \cdot |z - \tilde{z}|^\beta \right) \\
&\leq C \cdot 2^\beta \cdot \left(\frac{|x - \tilde{x}|}{2} + \frac{|z - \tilde{z}|}{2} \right)^\beta \\
&\leq C \cdot (|x - \tilde{x}| + |z - \tilde{z}|)^\beta \leq C \cdot |x - z|^\beta,
\end{aligned}$$

wobei die vorletzte Ungleichung mit Hilfe der Ungleichung von Jensen aus der Konkavität von $u \mapsto u^\beta$ auf $\mathbb{R}_+ \setminus \{0\}$ folgt.

Damit ist die Klasse $\bar{\mathcal{D}}_n^{(p,C)}$ aller Verteilungen von (X, Y) mit

1. $X \sim U[0,1]$,
2. $Y = m^{(c_n)}(X) + N$ für ein $c_n \in \mathcal{C}_n$ und ein $N \sim N(0,1)$, wobei X und N unabhängig sind

für genügend großes n eine Unterklasse von $\mathcal{D}^{(p,C)}$, und es genügt zu zeigen:

$$\liminf_{n\to\infty} \inf_{m_n} \sup_{(X,Y)\in\bar{\mathcal{D}}_n^{(p,C)}} \frac{M_n^{2p}}{C^2} \cdot \mathbf{E} \int_0^1 |m_n(x) - m^{(c_n}(x)|^2 dx > 0. \tag{4.21}$$

2. Schritt: Wir verwenden Regressionsschätzer, um den Parameter $c_n \in \mathcal{C}_n$ einer Verteilung $(X,Y) \in \bar{\mathcal{D}}_n^{(p,C)}$ zu schätzen.

Dazu sei m_n ein beliebiger Regressionsschätzer. Nach Konstruktion sind die Supports der $g_{n,j}$ disjunkt, also sind die $\{g_{n,j} : j \in \mathbb{N}\}$ in L_2 orthogonal. Daher ist die orthogonale Projektion von m_n auf $\{m^{(c_n)} : c_n \in \mathbb{R}^{M_n}\}$ gegeben durch

$$\hat{m}_n(x) = \sum_{j=1}^{M_n} \hat{c}_{n,j} \cdot g_{n,j}(x)$$

wobei

$$\hat{c}_{n,j} = \frac{\int_{A_{n,j}} m_n(x) \cdot g_{n,j}(x)\, dx}{\int_{A_{n,j}} g_{n,j}^2(x)\, dx}.$$

Für $c_n \in \mathcal{C}_n$ beliebig gilt nun

$$\begin{aligned}
&\int_0^1 |m_n(x) - m^{(c_n)}(x)|^2 dx \\
&\geq \int_0^1 |\hat{m}_n(x) - m^{(c_n)}(x)|^2 dx \\
&= \sum_{j=1}^{M_n} \int_{A_{n,j}} |\hat{c}_{n,j} \cdot g_{n,j}(x) - c_{n,j} \cdot g_{n,j}(x)|^2 dx \\
&= \sum_{j=1}^{M_n} |\hat{c}_{n,j} - c_{n,j}|^2 \cdot \int_{A_{n,j}} g_{n,j}^2(x)\, dx \\
&= \int g^2(x)\, dx \cdot \frac{1}{M_n^{2p+1}} \cdot \sum_{j=1}^{M_n} |\hat{c}_{n,j} - c_{n,j}|^2.
\end{aligned}$$

Setze

$$\tilde{c}_{n,j} = \begin{cases} 1, & \text{falls } \hat{c}_{n,j} \geq 0, \\ -1, & \text{sonst.} \end{cases}$$

Dann gilt

$$|\hat{c}_{n,j} - c_{n,j}| \geq \frac{1}{2} \cdot |\tilde{c}_{n,j} - c_{n,j}| = I_{\{\tilde{c}_{n,j} \neq c_{n,j}\}},$$

wie man leicht durch Betrachtung der beiden Fälle $\tilde{c}_{n,j} = 1$, $c_{n,j} = -1$ und $\tilde{c}_{n,j} = -1$, $c_{n,j} = 1$ sieht.

Damit erhalten wir

$$\int_0^1 |m_n(x) - m^{(c_n)}(x)|^2 dx \geq \int g^2(x)\, dx \cdot \frac{1}{M_n^{2p+1}} \cdot \sum_{j=1}^{M_n} I_{\{\tilde{c}_{n,j} \neq c_{n,j}\}},$$

also folgt (4.21) aus

$$\liminf_{n\to\infty} \inf_{\tilde{c}_n} \sup_{c \in \mathcal{C}_n} \frac{1}{M_n} \sum_{j=1}^{M_n} \mathbf{P}\left\{\tilde{c}_{n,j} \neq c_{n,j}\right\} > 0. \tag{4.22}$$

3. Schritt: Wir wählen $c_n \in \mathcal{C}_n$ zufällig.

Seien $C_{n,1}, \ldots, C_{n,M_n}$ unabhängig identisch verteilte reelle Zufallsvariablen mit

$$\mathbf{P}\{C_{n,1} = 1\} = \frac{1}{2} = \mathbf{P}\{C_{n,1} = -1\},$$

die unabhängig von (X_1, N_1), $\ldots$, (X_n, N_n) sind. Setze

$$C_n = (C_{n,1}, \ldots, C_{n,M_n})\,.$$

Dann gilt

$$\begin{aligned} &\inf_{\tilde{c}_n} \sup_{c_n \in \mathcal{C}_n} \frac{1}{M_n} \sum_{j=1}^{M_n} \mathbf{P}\left\{\tilde{c}_{n,j} \neq c_{n,j}\right\} \\ &\geq \inf_{\tilde{c}_n} \frac{1}{M_n} \sum_{j=1}^{M_n} \mathbf{P}\left\{\tilde{c}_{n,j} \neq C_{n,j}\right\}. \end{aligned}$$

Die optimale Vorhersagefunktion ist

$$\bar{C}_{n,j} = \begin{cases} 1, & \text{falls } \mathbf{P}\{C_{n,j} = 1 | (X_1, Y_1), \ldots, (X_n, Y_n)\} \geq \frac{1}{2}, \\ -1, & \text{sonst.} \end{cases}$$

Aus Symmetriegründen gilt daher

$$\mathbf{P}\left\{\tilde{c}_{n,j} \neq C_{n,j}\right\} \geq \mathbf{P}\left\{\bar{C}_{n,j} \neq C_{n,j}\right\} = \mathbf{P}\left\{\bar{C}_{n,1} \neq C_{n,1}\right\}$$

und wir erhalten

$$\inf_{\tilde{c}_n} \sup_{c_n \in \mathcal{C}_n} \frac{1}{M_n} \sum_{j=1}^{M_n} \mathbf{P}\left\{\tilde{c}_{n,j} \neq c_{n,j}\right\} \geq \mathbf{P}\left\{\bar{C}_{n,1} \neq C_{n,1}\right\}.$$

Also genügt es zu zeigen:

$$\liminf_{n\to\infty} \mathbf{P}\left\{\bar{C}_{n,1} \neq C_{n,1}\right\} > 0. \tag{4.23}$$

4. Schritt: Nachweis von (4.23).

Wir verwenden

$$\mathbf{P}\left\{\bar{C}_{n,1} \neq C_{n,1}\right\} = \mathbf{E}\left\{\mathbf{P}\left\{\bar{C}_{n,1} \neq C_{n,1} \middle| X_1, \ldots, X_n\right\}\right\}.$$

Seien $X_{i_1}, \ldots, X_{i_l}$ diejenigen X_i mit $X_i \in A_{n,1}$. Dann gilt

$$(Y_{i_1}, \ldots, Y_{i_l}) = C_{n,1} \cdot (g_{n,1}(X_{i_1}), \ldots, g_{n,1}(X_{i_l})) + (N_{i_1}, \ldots, N_{i_l}). \tag{4.24}$$

Alle Y_j mit $X_j \notin A_{n,1}$ hängen nur von $C_{n,2}, \ldots, C_{n.M_n}$ sowie

$$\{(X_r, N_r) : r \notin \{i_1, \ldots, i_l\}\}$$

ab und sind damit unabhängig von den Daten in (4.24) gegeben $X_1, \ldots, X_n$. Bedingt man nun auf alle diese Zufallsvariablen ebenfalls noch, so folgt unter Beachtung von

$$g_{n,1}(X_j) = 0 \quad \text{für } X_j \notin A_{n,1}$$

mit Lemma 4.16

$$\begin{aligned} \mathbf{P}\left\{\bar{C}_{n,1} \neq C_{n,1} \middle| X_1, \ldots, X_n\right\} &= \Phi\left(-\sqrt{\sum_{r=1}^{l} g_{n,1}^2(X_{i_r})}\right) \\ &= \Phi\left(-\sqrt{\sum_{i=1}^{n} g_{n,1}^2(X_i)}\right), \end{aligned}$$

wobei Φ die Verteilungsfunktion zu $N(0,1)$ ist.

Man sieht (z.B. durch Berechnung der 2. Ableitung) leicht, dass

$$x \mapsto \Phi(-\sqrt{x})$$

konvex ist. Anwendung der Ungleichung von Jensen liefert

$$\begin{aligned}
\mathbf{P}\left\{\bar{C}_{n,1} \neq C_{n,1}\right\} &= \mathbf{E}\left\{\Phi\left(-\sqrt{\sum_{i=1}^{n} g_{n,1}^2(X_i)}\right)\right\} \\
&\geq \Phi\left(-\sqrt{\mathbf{E}\left\{\sum_{i=1}^{n} g_{n,1}^2(X_i)\right\}}\right) \\
&= \Phi\left(-n \cdot \int_0^1 g_{n,1}^2(x)dx\right) \\
&= \Phi\left(-n \cdot M_n^{-(2p+1)} \cdot C^2 \int_0^1 \bar{g}^2(x)dx\right) \\
&\geq \Phi\left(-\int_0^1 \bar{g}^2(x)dx\right),
\end{aligned}$$

da

$$M_n = \lceil (C^2 \cdot n)^{\frac{1}{2p+1}} \rceil \geq (C^2 \cdot n)^{\frac{1}{2p+1}}.$$

□

4.7 Datenabhängige Wahl von Parametern

4.7.1 Motivation

Die Bandbreite des Kernschätzers in Korollar 4.12, dessen L_2-Fehler gemäß Satz 4.15 mit optimaler Geschwindigkeit gegen Null konvergierte, hing von p, C, σ^2 und dem Maximalwert des Betrages der Regressionsfunktion ab. Eine solche Wahl der Bandbreite ist in Anwendungen nicht möglich, da dort insbesondere die Glattheit der Regressionsfunktion (in Korollar 4.12 beschrieben durch p und C) unbekannt ist.

Nötig ist daher eine datenabhängige Wahl der Bandbreite, die wir in diesem Kapitel untersuchen.

4.7.2 Unterteilung der Stichprobe

Seien (X, Y), (X_1, Y_1), (X_2, Y_2), ... unabhängige identisch verteilte $\mathbb{R}^d \times \mathbb{R}$-wertige Zufallsvariablen mit $\mathbf{E}\{Y^2\} < \infty$. Setze $m(x) = \mathbf{E}\{Y|X = x\}$. Seien

$$\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$$

die gegebenen Daten. Wir gehen im Folgenden davon aus, dass wir eine endliche Parametermenge $\mathcal{P}_n$ und für jedes $h \in \mathcal{P}_n$ einen Schätzer

$$m_n^{(h)}(x) = m_n^{(h)}(x, \mathcal{D}_n)$$

von $m(x)$ gegeben haben (z.B. $m_n^{(h)}$ ist Kernschätzer mit Bandbreite h). Unser Ziel ist, in Abhängigkeit der gegebenen Daten

$$\hat{h} = \hat{h}(\mathcal{D}_n) \in \mathcal{P}_n$$

so zu bestimmen, dass approximativ gilt:

$$\int |m_n^{(\hat{h})}(x) - m(x)|^2 \mathbf{P}_X(dx) \approx \min_{h \in \mathcal{P}_n} \int |m_n^{(h)}(x) - m(x)|^2 \mathbf{P}_X(dx).$$

Bei der sogenannten *Unterteilung der Stichprobe* gehen wir zur datenabhängigen Wahl von h wie folgt vor:

Zuerst unterteilen wir unsere Stichprobe in Lerndaten

$$\mathcal{D}_{n_l} = \{(X_1, Y_1), \ldots, (X_{n_l}, Y_{n_l})\}$$

und Testdaten

$$\{(X_{n_l+1}, Y_{n_l+1}), \ldots, (X_{n_l+n_t}, Y_{n_l+n_t})\},$$

wobei $n_l, n_t \geq 1$ mit $n_l + n_t = n$. Dann berechnen wir für jeden Parameter $h \in \mathcal{P}_n$ mit Hilfe der Lerndaten den Schätzer

$$m_{n_l}^{(h)}(\cdot) = m_{n_l}^{(h)}(\cdot, \mathcal{D}_{n_l}),$$

berechnen dessen empirisches L_2-Risiko auf den Testdaten, d.h.

$$\frac{1}{n_t} \sum_{i=n_l+1}^{n} |Y_i - m_{n_l}^{(h)}(X_i)|^2, \tag{4.25}$$

und wählen dasjenige $\hat{h} \in \mathcal{P}_n$, für das (4.25) minimal wird, d.h. wir setzen

$$\hat{h} = \hat{h}(\mathcal{D}_n) = \arg\min_{h \in \mathcal{P}_n} \frac{1}{n_t} \sum_{i=n_l+1}^{n} |Y_i - m_{n_l}^{(h)}(X_i)|^2. \tag{4.26}$$

Sodann verwenden wir

$$m_n(x) = m_{n_l}^{(h)}(x, \mathcal{D}_{n_l}) \tag{4.27}$$

als Regressionsschätzer. Für diesen gilt:

Satz 4.17. *Sei $0 < L < \infty$. Es gelte*

$$|Y| \le L \quad f.s. \quad und \quad \max_{h \in \mathcal{P}_n} \|m_{n_l}^{(h)}(\cdot)\|_\infty \le L.$$

Sei m_n definiert durch (4.26) und (4.27). Dann gilt für jedes $\delta > 0$:

$$\begin{aligned}
&\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\le (1+\delta) \cdot \min_{h \in \mathcal{P}_n} \mathbf{E} \int |m_{n_l}^{(h)}(x) - m(x)|^2 \mathbf{P}_X(dx) + c \cdot \frac{1 + \log |\mathcal{P}_n|}{n_t},
\end{aligned}$$

wobei $c = L^2 \cdot (\frac{32}{\delta} + 70 + 38 \cdot \delta)$.

Beweis. Wir verwenden die Fehlerzerlegung

$$\begin{aligned}
&\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&= \mathbf{E}\left\{ |m_n(X) - Y|^2 \middle| \mathcal{D}_n \right\} - \mathbf{E}\left\{ |m(X) - Y|^2 \right\} \\
&= \Big(\mathbf{E}\left\{ |m_n(X) - Y|^2 \middle| \mathcal{D}_n \right\} - \mathbf{E}\left\{ |m(X) - Y|^2 \right\} \\
&\qquad -(1+\delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^{n} \left\{ |m_n(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2 \right\} \Big) \\
&\quad +(1+\delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^{n} \left\{ |m_n(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2 \right\} \\
&=: T_{1,n} + T_{2,n}.
\end{aligned}$$

Nach Definition des Schätzers ist

$$\frac{1}{n_t} \sum_{i=n_l+1}^{n} |m_n(X_i) - Y_i|^2 = \min_{h \in \mathcal{P}_n} \frac{1}{n_t} \sum_{i=n_l+1}^{n} |m_{n_l}^{(h)}(X_i) - Y_i|^2,$$

woraus folgt

$$\begin{aligned}
\mathbf{E}\{T_{2,n}\} &= \mathbf{E}\left\{ (1+\delta) \cdot \min_{h \in \mathcal{P}_n} \frac{1}{n_t} \sum_{i=n_l+1}^{n} \left\{ |m_{n_l}^{(h)}(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2 \right\} \right\} \\
&\le (1+\delta) \cdot \min_{h \in \mathcal{P}_n} \mathbf{E}\left\{ \frac{1}{n_t} \sum_{i=n_l+1}^{n} \left\{ |m_{n_l}^{(h)}(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2 \right\} \right\} \\
&= (1+\delta) \cdot \min_{h \in \mathcal{P}_n} \mathbf{E} \int |m_{n_l}^{(h)}(x) - m(x)|^2 \mathbf{P}_X(dx).
\end{aligned}$$

Also genügt es, im Folgenden noch zu zeigen:

$$\mathbf{E}\{T_{1,n}\} \leq c \cdot \frac{1+\log|\mathcal{P}_n|}{n_t}. \tag{4.28}$$

Zum Nachweis von (4.28) beachten wir, dass für $s > 0$ gilt:

$$\begin{aligned}
&\mathbf{P}\left\{T_{1,n} > s \middle| \mathcal{D}_{n_l}\right\} \\
&= \mathbf{P}\Bigg\{\mathbf{E}\left\{|m_n(X)-Y|^2 \middle| \mathcal{D}_n\right\} - \mathbf{E}\left\{|m(X)-Y|^2\right\} \\
&\qquad -(1+\delta)\cdot\frac{1}{n_t}\sum_{i=n_l+1}^{n}\left\{|m_n(X_i)-Y_i|^2 - |m(X_i)-Y_i|^2\right\} > s \middle| \mathcal{D}_{n_l}\Bigg\} \\
&\leq \mathbf{P}\Bigg\{\exists h \in \mathcal{P}_n : \mathbf{E}\left\{|m_{n_l}^{(h)}(X)-Y|^2 \middle| \mathcal{D}_{n_l}\right\} - \mathbf{E}\left\{|m(X)-Y|^2\right\} \\
&\qquad -(1+\delta)\cdot\frac{1}{n_t}\sum_{i=n_l+1}^{n}\left\{|m_{n_l}^{(h)}(X_i)-Y_i|^2 - |m(X_i)-Y_i|^2\right\} > s \middle| \mathcal{D}_{n_l}\Bigg\} \\
&\leq |\mathcal{P}_n| \cdot \max_{h\in\mathcal{P}_n} \mathbf{P}\Bigg\{\mathbf{E}\left\{|m_{n_l}^{(h)}(X)-Y|^2 \middle| \mathcal{D}_{n_l}\right\} - \mathbf{E}\left\{|m(X)-Y|^2\right\} \\
&\qquad -(1+\delta)\cdot\frac{1}{n_t}\sum_{i=n_l+1}^{n}\left\{|m_{n_l}^{(h)}(X_i)-Y_i|^2 - |m(X_i)-Y_i|^2\right\} > s \middle| \mathcal{D}_{n_l}\Bigg\}.
\end{aligned}$$

Beachtet man, dass für $h \in \mathcal{P}_n$ fest gilt

$$\begin{aligned}
\sigma^2 &:= \mathbf{Var}\left\{|m_{n_l}^{(h)}(X)-Y|^2 - |m(X)-Y|^2 \middle| \mathcal{D}_{n_l}\right\} \\
&\leq \mathbf{E}\left\{\left(|m_{n_l}^{(h)}(X)-Y|^2 - |m(X)-Y|^2\right)^2 \middle| \mathcal{D}_{n_l}\right\} \\
&= \mathbf{E}\left\{\left(m_{n_l}^{(h)}(X)-m(X)\right)^2 \cdot \left(m_{n_l}^{(h)}(X)+m(X)-2Y\right)^2 \middle| \mathcal{D}_{n_l}\right\} \\
&\leq 16L^2 \cdot \mathbf{E}\left\{\left(m_{n_l}^{(h)}(X)-m(X)\right)^2 \middle| \mathcal{D}_{n_l}\right\} \\
&= 16L^2 \cdot \left(\mathbf{E}\left\{|m_{n_l}^{(h)}(X)-Y|^2 \middle| \mathcal{D}_{n_l}\right\} - \mathbf{E}\left\{|m(X)-Y|^2\right\}\right),
\end{aligned}$$

so folgt

$$\begin{aligned}
&\mathbf{P}\Bigg\{\mathbf{E}\left\{|m_{n_l}^{(h)}(X)-Y|^2 \middle| \mathcal{D}_{n_l}\right\} - \mathbf{E}\left\{|m(X)-Y|^2\right\} \\
&\qquad -(1+\delta)\cdot\frac{1}{n_t}\sum_{i=n_l+1}^{n}\left\{|m_{n_l}^{(h)}(X_i)-Y_i|^2 - |m(X_i)-Y_i|^2\right\} > s \middle| \mathcal{D}_{n_l}\Bigg\}
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathbf{P}\Bigg\{ (1+\delta)\cdot\left(\mathbf{E}\left\{|m_{n_l}^{(h)}(X)-Y|^2\middle|\mathcal{D}_{n_l}\right\}-\mathbf{E}\left\{|m(X)-Y|^2\right\}\right) \\
&\qquad -(1+\delta)\cdot\frac{1}{n_t}\sum_{i=n_l+1}^{n}\left\{|m_{n_l}^{(h)}(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\right\} \\
&\qquad > s+\delta\cdot\left(\mathbf{E}\left\{|m_{n_l}^{(h)}(X)-Y|^2\middle|\mathcal{D}_{n_l}\right\}-\mathbf{E}\left\{|m(X)-Y|^2\right\}\right)\Bigg|\mathcal{D}_{n_l}\Bigg\} \\
&\leq \mathbf{P}\Bigg\{ \mathbf{E}\left\{|m_{n_l}^{(h)}(X)-Y|^2\middle|\mathcal{D}_{n_l}\right\}-\mathbf{E}\left\{|m(X)-Y|^2\right\} \\
&\qquad -\frac{1}{n_t}\sum_{i=n_l+1}^{n}\left\{|m_{n_l}^{(h)}(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\right\} \\
&\qquad > \frac{s}{1+\delta}+\frac{\delta}{1+\delta}\cdot\frac{\sigma^2}{16L^2}\Bigg\}.
\end{aligned}
$$

Mit der Ungleichung von Bernstein lässt sich die letzte Wahrscheinlichkeit nach oben abschätzen durch

$$
\begin{aligned}
&\exp\left(-\frac{n_t\cdot\left(\frac{s}{1+\delta}+\frac{\delta}{1+\delta}\cdot\frac{\sigma^2}{16L^2}\right)^2}{2\sigma^2+\frac{2}{3}\cdot 8L^2\cdot\left(\frac{s}{1+\delta}+\frac{\delta}{1+\delta}\cdot\frac{\sigma^2}{16L^2}\right)}\right) \\
&\leq \exp\left(-\frac{n_t\cdot\left(\frac{s}{1+\delta}+\frac{\delta}{1+\delta}\cdot\frac{\sigma^2}{16L^2}\right)^2}{\left(\frac{s}{1+\delta}+\frac{\delta}{1+\delta}\cdot\frac{\sigma^2}{16L^2}\right)\cdot 32L^2\cdot\frac{1+\delta}{\delta}+\frac{16L^2}{3}\cdot\left(\frac{s}{1+\delta}+\frac{\delta}{1+\delta}\cdot\frac{\sigma^2}{16L^2}\right)}\right) \\
&\leq \exp\left(-\frac{n_t\cdot\frac{s}{1+\delta}}{32L^2\cdot\frac{1+\delta}{\delta}+\frac{16L^2}{3}}\right) \\
&\leq \exp\left(-\frac{n_t\cdot s}{c}\right),
\end{aligned}
$$

da

$$
\begin{aligned}
(1+\delta)\cdot\left(32L^2\cdot\frac{1}{\delta}+32L^2+\frac{16L^2}{3}\right) &\leq (1+\delta)\cdot\left(32L^2\cdot\frac{1}{\delta}+38L^2\right) \\
&= L^2\left(\frac{32}{\delta}+70+38\cdot\delta\right)=c.
\end{aligned}
$$

Damit erhalten wir für $u>0$ beliebig:

$$
\mathbf{E}\left\{T_{1,n}\right\} \leq \int_0^\infty \mathbf{P}\{T_{1,n}>s\}\,ds
$$

$$\begin{aligned} &\leq u + \int_u^\infty \mathbf{P}\{T_{1,n} > s\}\, ds \\ &\overset{s.o.}{\leq} u + \int_u^\infty |\mathcal{P}_n| \cdot \exp\left(-\frac{n_t \cdot s}{c}\right) ds \\ &= u + |\mathcal{P}_n| \cdot \frac{c}{n_t} \cdot \exp\left(-\frac{n_t \cdot u}{c}\right). \end{aligned}$$

Mit

$$u = \frac{c \cdot \log |\mathcal{P}_n|}{n_t}$$

folgt

$$\mathbf{E}\{T_{1,n}\} \leq \frac{c \cdot \log |\mathcal{P}_n|}{n_t} + \frac{c}{n_t} = c \cdot \frac{1 + \log |\mathcal{P}_n|}{n_t},$$

w.z.z.w. □

Korollar 4.18. *Die Verteilung von (X, Y) erfülle*

(i) *$supp(X)$ beschränkt,*

(ii) *$|Y| \leq L$ f.s. für ein $L > 0$,*

(iii) *$\exists p \in [0,1], C > 0 \quad \forall x, z \in supp(X) : |m(x) - m(z)| \leq C \cdot \|x - z\|^p$.*

Sei m_n der Kernschätzer mit naivem Kern, wobei die datenabhängige Bandbreite aus der Menge

$$\{2^k \quad : \quad k \in \{-n, \ldots, n\}\}$$

mit Hilfe des Verfahrens der Unterteilung der Stichprobe gewählt wird, und $n_l \approx n_t \approx n/2$ gelte.

Dann folgt

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) = O\left(n^{-\frac{2p}{2p+d}}\right).$$

Beweis: Folgt unmittelbar aus Satz 4.17 und Korollar 4.12. □

4.7.3 Kreuzvalidierung

Nachteile der Unterteilung der Stichprobe sind:

1. Nach Wahl des Parameters wird der Schätzer nur noch mit einem Teil der Daten berechnet.

2. Der Schätzer hängt von der zufälligen Unterteilung der Stichprobe ab (und zusätzlicher Zufall vergrößert einen mittleren quadratischen Fehler immer).

Beides versucht die sogenannte *Kreuzvalidierung* zu vermeiden. Bei der sogenannten *k-fachen Kreuzvalidierung* mit $k \in \{2, \dots, n\}$ (wobei wir oBdA $n/k \in \mathbb{N}$ voraussetzen, um die Schreibweise zu vereinfachen), wird die Datenmenge

$$\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$$

in k gleich große Teile unterteilt. Sei $\mathcal{D}_{n,k}^{(l)}$ die Datenmenge ohne den l-ten Teil, also

$$\mathcal{D}_{n,k}^{(l)} = \left\{(X_1, Y_1), \dots, (X_{(l-1)\cdot\frac{n}{k}}, Y_{(l-1)\cdot\frac{n}{k}}), (X_{l\cdot\frac{n}{k}+1}, Y_{l\cdot\frac{n}{k}+1}), \dots, (X_n, Y_n)\right\}.$$

Sei

$$m_{n-\frac{n}{k},l}^{(p)}(x) = m_{n-\frac{n}{k},l}^{(p)}(x; \mathcal{D}_{n,k}^{(l)})$$

der Schätzer berechnet mit den Daten $\mathcal{D}_{n,k}^{(l)}$ und Parameter $p \in \mathcal{P}_n$. Bei der k-fachen Kreuzvalidierung wählen wir den Parameter durch Minimierung des Mittels der empirischen L_2-Risikos aller dieser Schätzer berechnet jeweils auf den weggelassenen Daten, d.h. wir wählen

$$\hat{p} = \arg\min_{p \in \mathcal{P}_n} \frac{1}{k} \sum_{l=1}^{k} \frac{1}{\frac{n}{k}} \sum_{i=(l-1)\cdot\frac{n}{k}+1}^{l\cdot\frac{n}{k}} \left| Y_i - m_{n-\frac{n}{k},l}^{(p)}(X_i) \right|^2$$

und setzen

$$m_n(x) = m_n^{(\hat{p})}(x; \mathcal{D}_n).$$

Im Spezialfall von $k = n$, d.h. bei n-facher Kreuzvalidierung, spricht man auch von *Kreuzvalidierung*. Hier ist der Schätzer gegeben durch

$$\hat{p} = \arg\min_{p \in \mathcal{P}_n} \frac{1}{n} \sum_{i=1}^{n} \left| Y_i - m_{n-1}^{(p)}\left(X_i; (X_1, Y_1), \dots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), \dots, (X_n, Y_n)\right) \right|^2$$

und

$$m_n(x) = m_n^{(\hat{p})}(x; \mathcal{D}_n).$$

4.8 Hilfsmittel aus der Theorie empirischer Prozesse

4.8.1 Motivation

Sei $\mathcal{F}_n$ eine Klasse von Funktionen $f : \mathbb{R}^d \to \mathbb{R}$ und

$$m_n(\cdot) = \arg\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2$$

der zugehörige Kleinste-Quadrate-Schätzer der Regressionsfunktion

$$m(\cdot) = \arg\min_{f:\mathbb{R}^d \to \mathbb{R}} \mathbf{E}\{|f(X) - Y|^2\}.$$

Ziel im Folgenden ist die Abschätzung von dessen L_2-Fehler:

$$\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) = \mathbf{E}\left\{|m_n(X) - Y|^2 \middle| \mathcal{D}_n\right\} - \mathbf{E}\left\{|m(X) - Y|^2\right\}.$$

Die Idee dazu ist, dass eines empirische Variante dieses Fehlers einfach abgeschätzt werden kann, da nach Definition des Schätzers gilt:

$$\begin{aligned} Z_n &:= \frac{1}{n} \sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 - \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2 \\ &= \min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2 - \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2, \end{aligned}$$

woraus folgt

$$\begin{aligned} \mathbf{E}\{Z_n\} &\leq \min_{f \in \mathcal{F}_n} \mathbf{E}\left\{|f(X) - Y|^2\right\} - \mathbf{E}\left\{|m(X) - Y|^2\right\} \\ &= \min_{f \in \mathcal{F}_n} \int |f(x) - m(x)|^2 \mathbf{P}_X(dx). \end{aligned}$$

Im Weiteren schätzen wir die Differenz zwischen dem L_2-Fehler und einem Vielfachen der obigen empirischen Variante desselben ab.

4.8.2 Uniforme Exponentialungleichungen

Nötig in Abschnitt 4.8.1 sind Abschätzungen für Ausdrücke wie

$$\mathbf{E}\left\{|m_n(X) - Y|^2 \middle| \mathcal{D}_n\right\} - \frac{1}{n} \sum_{i=1}^{n} |m_n(X_i) - Y_i|^2.$$

Ein Problem dabei ist, dass innerhalb des Erwartungswertes bzw. der Summe eine *zufällige* Funktion $m_n \in \mathcal{F}_n$ steht. Dieses Problem wird man los, indem man den obigen Ausdruck nach oben abschätzt durch

$$\sup_{f \in \mathcal{F}_n} \left\{ \mathbf{E}\left\{ |f(X) - Y|^2 \right\} - \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2 \right\}.$$

Für Abschätzungen von Ausdrücken dieser Bauart benötigen wir ein Maß für die "Komplexität" des Funktionenraumes $\mathcal{F}_n$, das wir in der nächsten Definition einführen.

Definition 4.19. *Sei $\epsilon > 0$, sei $\mathcal{G}$ eine Menge von Funktionen $g : \mathbb{R}^l \to \mathbb{R}$, sei $1 \leq p < \infty$ und sei ν ein Wahrscheinlichkeitsmaß auf $\mathbb{R}^l$. Für $g : \mathbb{R}^l \to \mathbb{R}$ sei*

$$\|g\|_{L_p(\nu)} := \left\{ \int |g(x)|^p \nu(dx) \right\}^{\frac{1}{p}}.$$

a) *Jede endliche Menge von Funktionen $g_1, \ldots, g_N : \mathbb{R}^l \to \mathbb{R}$ mit*

$$\forall g \in \mathcal{G}\ \exists j = j(g) \in \{1, \ldots, N\} : \ \|g - g_j\|_{L_p(\nu)} < \epsilon$$

heißt **ϵ-Überdeckung von $\mathcal{G}$** *bzgl. $\|\cdot\|_{L_p(\nu)}$.*

b) *Die* **ϵ-Überdeckungszahl von $\mathcal{G}$** *bzgl. $\|\cdot\|_{L_p(\nu)}$ mit Bezeichnung*

$$\mathcal{N}(\epsilon, \mathcal{G}, \|\cdot\|_{L_p(\nu)})$$

wird definiert als minimale Kardinalität aller ϵ-Überdeckung von $\mathcal{G}$ bzgl. $\|\cdot\|_{L_p(\nu)}$. Im Falle, dass keine endliche ϵ-Überdeckung von $\mathcal{G}$ bzgl. $\|\cdot\|_{L_p(\nu)}$ existiert setzen wir $\mathcal{N}(\epsilon, \mathcal{G}, \|\cdot\|_{L_p(\nu)}) = \infty$.

c) *Seien $z_1^n = (z_1, \ldots, z_n)$ n Punkte in $\mathbb{R}^l$. Sei ν_n die zugehörige empirische Verteilung, also*

$$\nu_n(A) = \frac{1}{n} \sum_{i=1}^{n} I_A(z_i) \quad (A \subseteq \mathbb{R}^l),$$

so dass

$$\|g\|_{L_p(\nu_n)} = \left\{ \frac{1}{n} \sum_{i=1}^{n} |g(z_i)|^p \right\}^{\frac{1}{p}}.$$

Dann heißt jede ϵ-Überdeckung von $\mathcal{G}$ bzgl. $\|\cdot\|_{L_p(\nu_n)}$ auch **L_p-ϵ-Überdeckung von $\mathcal{G}$ auf** *z_1^n, und für die ϵ-Überdeckungszahl von $\mathcal{G}$ bzgl. $\|\cdot\|_{L_p(\nu_n)}$ wird die Notation*

$$\mathcal{N}_p(\epsilon, \mathcal{G}, z_1^n)$$

verwendet.

Satz 4.20. *(Pollard (1984)).*

Seien Z, Z_1, ..., Z_n unabhängig identisch verteilte $\mathbb{R}^l$-wertige Zufallsvariablen. Sei $B > 0$ und sei $\mathcal{G}$ eine Klasse von Funktionen $g : \mathbb{R}^l \to [0, B]$. Dann gilt für jedes $n \in \mathbb{N}$ und jedes $\epsilon > 0$:

$$\begin{aligned}&\mathbf{P}\left\{\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n} g(Z_i) - \mathbf{E}\{g(Z)\}\right| > \epsilon\right\}\\ &\leq 8 \cdot \mathbf{E}\left\{\mathcal{N}_1(\epsilon/8, \mathcal{G}, Z_1^n)\right\} \cdot \exp\left(-\frac{n\cdot\epsilon^2}{128\cdot B^2}\right),\end{aligned}$$

wobei $Z_1^n = (Z_1, \ldots, Z_n)$.

Bemerkung: Hierbei vernachläsigen wir eventuell auftretende Messbarkeitsprobleme (die beim Supremum und bei der Überdeckungszahl auftreten können).

Beweis. Analog zu Satz 2.2 aus der Vorlesung Mathematische Statistik im WS 14/15. □

Bei der Anwendung des obigen Satzes tritt das Problem auf, dass die rechte Seite für $\epsilon \leq 1/\sqrt{n}$ nicht gegen Null konvergiert, was nicht zufriedenstellend ist hinsichtlich der optimalen Konvergenzrate von

$$n^{-\frac{2p}{2p+d}}$$

aus Abschnitt 4.6. Schneller gegen Null konvergierende obere Schranken lassen sich aber herleiten, sofern wir die Differenz zwischen Erwartungswerten und Vielfachen des Stichprobenmittels abschätzen, denn es gilt

$$\begin{aligned}&\mathbf{E}\left\{|m_n(X) - Y|^2 \middle| \mathcal{D}_n\right\} - \mathbf{E}\left\{|m(X) - Y|^2\right\}\\ &\qquad -2 \cdot \frac{1}{n}\sum_{i=1}^{n}\left(|m_n(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2\right)\\ &> t\\ \Leftrightarrow\quad &\mathbf{E}\left\{|m_n(X) - Y|^2 \middle| \mathcal{D}_n\right\} - \mathbf{E}\left\{|m(X) - Y|^2\right\}\\ &\qquad -\frac{1}{n}\sum_{i=1}^{n}\left(|m_n(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2\right)\\ &> \frac{1}{2}\cdot\left(t + \mathbf{E}\left\{|m_n(X) - Y|^2 \middle| \mathcal{D}_n\right\} - \mathbf{E}\left\{|m(X) - Y|^2\right\}\right)\end{aligned}$$

sowie

Satz 4.21. *(Lee, Bartlett and Williamson (1996)).*

Seien (X,Y), (X_1,Y_1), …, (X_n,Y_n) unabhängig identisch verteilte $\mathbb{R}^d\times\mathbb{R}$-wertige Zufallsvariablen mit $|Y|\le B$ f.s. für ein $B\ge 1$. Sei $\mathcal{F}$ eine Klasse von Funktionen $f:\mathbb{R}^d\to[-B,B]$. Dann gilt für $n\in\mathbb{N}$, $\alpha,\beta>0$ und $0<\epsilon\le 1/2$ beliebig:

$$\begin{aligned}\mathbf{P}\Bigg\{&\exists f\in\mathcal{F}:\mathbf{E}\left\{|f(X)-Y|^2\right\}-\mathbf{E}\left\{|m(X)-Y|^2\right\}\\&\qquad-\frac{1}{n}\sum_{i=1}^{n}\left(|f(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\right)\\&\qquad>\epsilon\cdot\left(\alpha+\beta+\mathbf{E}\left\{|f(X)-Y|^2\right\}-\mathbf{E}\left\{|m(X)-Y|^2\right\}\right)\Bigg\}\\\le\ &14\cdot\sup_{x_1^n}\mathcal{N}_1\left(\frac{\beta\cdot\epsilon}{20\cdot B},\mathcal{F},x_1^n\right)\cdot\exp\left(-\frac{\epsilon^2(1-\epsilon)\cdot\alpha\cdot n}{214\cdot(1+\epsilon)\cdot B^4}\right).\end{aligned}$$

Beweis: erfolgt im Seminar im WS 15/16. □

Im Folgenden: Herleitung von Abschätzungen für Überdeckungszahlen.

4.8.3 Abschätzung von Überdeckungszahlen

Definition 4.22. *Sei $\epsilon>0$, sei $\mathcal{G}$ eine Menge von Funktionen $g:\mathbb{R}^l\to\mathbb{R}$, sei $1\le p<\infty$ und sei ν ein Wahrscheinlichkeitsmaß auf $\mathbb{R}^l$. Für $g:\mathbb{R}^l\to\mathbb{R}$ sei*

$$\|g\|_{L_p(\nu)}:=\left\{\int|g(x)|^p\nu(dx)\right\}^{\frac{1}{p}}.$$

a) *Jede endliche Menge von Funktionen $g_1,\dots,g_N\in\mathcal{G}$ mit*

$$\|g_i-g_j\|_{L_p(\nu)}\ge\epsilon\quad\text{für alle } 1\le i<j\le N$$

heißt **ϵ-Packung von $\mathcal{G}$** *bzgl. $\|\cdot\|_{L_p(\nu)}$.*

b) *Die* **ϵ-Packzahl von $\mathcal{G}$** *bzgl. $\|\cdot\|_{L_p(\nu)}$*

$$\mathcal{M}\left(\epsilon,\mathcal{G},\|\cdot\|_{L_p(\nu)}\right)$$

ist definiert als die maximale Kardinalität aller ϵ-Packungen von $\mathcal{G}$ bzgl. $\|\cdot\|_{L_p(\nu)}$. Hierbei setzen wir $\mathcal{M}\left(\epsilon,\mathcal{G},\|\cdot\|_{L_p(\nu)}\right)=\infty$, falls für jedes $n\in\mathbb{N}$ eine ϵ-Packung von $\mathcal{G}$ bzgl. $\|\cdot\|_{L_p(\nu)}$ mit n Elementen existiert.

c) *Die L_p-ϵ-***Packzahl von** $\mathcal{G}$ **auf** z_1^n *ist*

$$\mathcal{M}_p\left(\epsilon, \mathcal{G}, z_1^n\right) = \mathcal{M}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_p(\nu_n)}\right),$$

wobei ν_n die empirische Verteilung zu $z_1^n = (z_1, \ldots, z_n) \in (\mathbb{R}^l)^n$ ist.

Lemma 4.23. *Ist $\epsilon > 0$, $\mathcal{G}$ eine Menge von Funktionen $g : \mathbb{R}^l \to \mathbb{R}$, $1 \leq p < \infty$ und ist ν ein Wahrscheinlichkeitsmaß auf $\mathbb{R}^l$, so gilt:*

$$\mathcal{M}\left(2 \cdot \epsilon, \mathcal{G}, \|\cdot\|_{L_p(\nu)}\right) \leq \mathcal{N}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_p(\nu)}\right) \leq \mathcal{M}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_p(\nu)}\right).$$

Beweis. a) Ist $g_1, \ldots, g_N$ eine $2 \cdot \epsilon$-Packung von $\mathcal{G}$ bzgl. $\|\cdot\|_{L_p(\nu)}$, so enthält jede offene Kugel mit Radius ϵ höchstens eines der $g_1, \ldots, g_N$, und damit besteht jede ϵ-Überdeckung von $\mathcal{G}$ bzgl. $\|\cdot\|_{L_p(\nu)}$ aus mindestens N Funktionen.

b) Ist $g_1, \ldots, g_N$ eine ϵ-Packung von $\mathcal{G}$ bzgl. $\|\cdot\|_{L_p(\nu)}$ maximaler Größe, so ist für jedes $g \in \mathcal{G}$

$$g_1, \ldots, g_N, g$$

keine ϵ-Packung. Folglich existiert für jedes $g \in \mathcal{G}$ ein $j = j(g) \in \{1, \ldots, N\}$ mit

$$\|g - g_j\|_{L_p(\nu)} < \epsilon.$$

Damit ist aber $g_1, \ldots, g_N$ eine ϵ-Überdeckung von $\mathcal{G}$ bzgl. $\|\cdot\|_{L_p(\nu)}$. □

Zur Herleitung einer Abschätzung für Überdeckungszahlen betrachten wir zuerst den Spezialfall, dass die Funktionen alle Indikatorfunktionen sind.

Sind $f = I_A$, $g = I_B$ für $A, B \subseteq \mathbb{R}^d$, und sind $z_1, \ldots, z_n \in \mathbb{R}^d$, so gilt

$$\begin{aligned}\left\{\frac{1}{n}\sum_{i=1}^{n}|f(z_i) - g(z_i)|^p\right\}^{\frac{1}{p}} &\leq \max_{i=1,\ldots,n}|f(z_i) - g(z_i)| \\ &= \begin{cases} 1, & \text{falls } A \cap \{z_1, \ldots, z_n\} \neq B \cap \{z_1, \ldots, z_n\} \\ 0, & \text{sonst.}\end{cases}\end{aligned}$$

Ist also $\mathcal{G} = \{1_A : A \in \mathcal{A}\}$ für $\mathcal{A} \subseteq \mathcal{P}(\mathbb{R}^d)$ und $0 < \epsilon < 1$, so gilt:

$$\mathcal{N}_p(\epsilon, \mathcal{G}, z_1^n) \leq \left|\{A \cap \{z_1, \ldots, z_n\} \,:\, A \in \mathcal{A}\}\right|.$$

Definition 4.24. *Sei $\mathcal{A}$ eine Klasse von Mengen $A \subseteq \mathbb{R}^d$ und sei $n \in \mathbb{N}$.*

a) *Für $z_1, \ldots, z_n \in \mathbb{R}^d$ ist*

$$s\left(\mathcal{A}, \{z_1, \ldots, z_n\}\right) := \left|\{A \cap \{z_1, \ldots, z_n\} \quad : \quad A \in \mathcal{A}\}\right|$$

die Anzahl der Teilmengen von $\{z_1, \ldots, z_n\}$, die durch Mengen aus $\mathcal{A}$ "herausgegriffen" werden können.

b) *Sei G eine endlichen Teilmenge von $\mathbb{R}^d$. Man sagt, $\mathcal{A}$* **zerlegt** *(shatters) G, falls*

$$s(\mathcal{A}, G) = 2^{|G|},$$

d.h., falls jede Teilmenge von G in der Form $A \cap G$ für ein $A \in \mathcal{A}$ dargestellt werden kann.

c) *Der n-te* **Zerlegungskoeffizient von** $\mathcal{A}$

$$S(\mathcal{A}, n) := \max_{z_1,\ldots,z_n \in \mathbb{R}^d} s\left(\mathcal{A}, \{z_1, \ldots, z_n\}\right)$$

ist die maximale Anzahl verschiedener Teilmengen von n Punkten in $\mathbb{R}^d$, die durch Mengen aus $\mathcal{A}$ herausgegriffen werden können.

Beispiele: a) Die Menge aller Intervalle der Form $(-\infty, a]$, $a \in \mathbb{R}$, zerlegt einelementige Teilmengen von $\mathbb{R}$, aber keine zweielementigen.

b) Die Menge aller Intervalle der Form $(a, b]$, $a, b \in \mathbb{R}$, zerlegt zweielementige Teilmengen von $\mathbb{R}$, aber keine dreielementigen.

c) Die Menge aller Halbebenen in $\mathbb{R}^2$ kann drei (geeignet gewählte) Punkte in $\mathbb{R}^2$ zerlegen.

d) Die Menge aller konvexen Mengen in $\mathbb{R}^2$ kann n (geeignet gewählte) Punkte in $\mathbb{R}^2$ zerlegen für jedes $n \in \mathbb{N}$.

Da ein Mengensystem, dass eine Menge G nicht zerlegt, auch keine Obermenge von G zerlegen kann, gilt:

$$S(\mathcal{A}, k) < 2^k \quad \Rightarrow \quad S(\mathcal{A}, n) < 2^n \text{ für alle } n > k.$$

Das größte n mit $S(\mathcal{A}, n) = 2^n$ ist die sogenannte VC-Dimension von $\mathcal{A}$.
Definition 4.25. *Sei $\mathcal{A}$ eine Klasse von Teilmengen von $\mathbb{R}^d$ mit $\mathcal{A} \neq \emptyset$. Die* **VC-Dimension** *(Vapnik-Chervonenkis-Dimension) $V_{\mathcal{A}}$ von $\mathcal{A}$ wird definiert durch*

$$V_{\mathcal{A}} = \sup\left\{n \in \mathbb{N} \quad : \quad S(\mathcal{A}, n) = 2^n\right\},$$

d.h. $V_{\mathcal{A}}$ ist die maximale Anzahl von Punkten, die durch $\mathcal{A}$ zerlegt werden.

Beispiel: a) $\mathcal{A} = \{(-\infty, a] \ : \ a \in \mathbb{R}\} \Rightarrow V_A = 1$

b) $\mathcal{A} = \{(a, b] \ : \ a, b \in \mathbb{R}\} \Rightarrow V_A = 2$

c) $\mathcal{A} = \{A \ : \ A \text{ konvex}\} \Rightarrow V_A = \infty$

Das nächste Theorem impliziert:

Entweder gilt $S(\mathcal{A}, n) = 2^n$ für alle $n \in \mathbb{N}$, oder $S(\mathcal{A}, n)$ wächst höchstens polynomiell in n.

Satz 4.26. *(Vapnik und Chervonenkis (1971)).*

Sei $\mathcal{A}$ eine Menge von Teilmengen von $\mathbb{R}^d$ mit VC-Dimension $V_\mathcal{A}$. Dann gilt für alle $n \in \mathbb{N}$:

$$S(\mathcal{A}, n) \leq \sum_{i=0}^{V_\mathcal{A}} \binom{n}{i}.$$

Korollar 4.27. *Ist $\mathcal{A}$ eine Menge von Teilmengen von $\mathbb{R}^d$ mit VC-Dimension $V_\mathcal{A}$, so gilt:*

a)

$$S(\mathcal{A}, n) \leq (n+1)^{V_\mathcal{A}} \quad \textit{für alle } n \in \mathbb{N}.$$

b)

$$S(\mathcal{A}, n) \leq \left(\frac{e \cdot n}{V_\mathcal{A}}\right)^{V_\mathcal{A}} \quad \textit{für alle } n \geq V_\mathcal{A}.$$

Beweis: a) Nach Satz 4.26 und dem binomischen Lehrsatz gilt:

$$\begin{aligned} S(\mathcal{A}, n) &\leq \sum_{i=0}^{V_\mathcal{A}} \binom{n}{i} = \sum_{i=0}^{V_\mathcal{A}} n \cdot (n-1) \cdot \cdots \cdot (n-i+1) \cdot \frac{1}{i!} \\ &\leq \sum_{i=0}^{V_\mathcal{A}} n^i \cdot \frac{V_\mathcal{A}!}{(V_\mathcal{A} - i)!} \cdot \frac{1}{i!} \\ &= \sum_{i=0}^{V_\mathcal{A}} n^i \cdot \binom{V_\mathcal{A}}{i} = (n+1)^{V_\mathcal{A}}. \end{aligned}$$

b) Ist $V_\mathcal{A}/n \leq 1$, so gilt nach Satz 4.26:

$$\begin{aligned} \left(\frac{V_\mathcal{A}}{n}\right)^{V_\mathcal{A}} \cdot S(\mathcal{A}, n) &\leq \sum_{i=0}^{V_\mathcal{A}} \left(\frac{V_\mathcal{A}}{n}\right)^{V_\mathcal{A}} \cdot \binom{n}{i} \\ &\leq \sum_{i=0}^{n} \left(\frac{V_\mathcal{A}}{n}\right)^{i} \cdot \binom{n}{i} \end{aligned}$$

$$= \left(1 + \frac{V_{\mathcal{A}}}{n}\right)^n \leq e^{V_{\mathcal{A}}},$$

wobei die letzte Ungleichung aus $1 + x \leq e^x$ $(x \in \mathbb{R})$ folgt. Dies impliziert

$$S(\mathcal{A}, n) \leq \left(\frac{n}{V_{\mathcal{A}}}\right)^{V_{\mathcal{A}}} \cdot e^{V_{\mathcal{A}}} = \left(\frac{e \cdot n}{V_{\mathcal{A}}}\right)^{V_{\mathcal{A}}}.$$

□

Beweis von Satz 4.26: O.B.d.A. gilt $V_{\mathcal{A}} < n$, da sonst die rechte Seite mit 2^n trivialerweise größer oder gleich als die linke Seite ist.

Seien $z_1, \ldots, z_n \in \mathbb{R}^d$ beliebig. Wir zeigen:

$$|\{A \cap \{z_1, \ldots, z_n\} \quad : \quad A \in \mathcal{A}\}| \leq \sum_{i=0}^{V_{\mathcal{A}}} \binom{n}{i}.$$

Dazu: Seien $F_1, \ldots, F_k$ mit $k = \binom{n}{V_{\mathcal{A}}+1}$ alle $(V_{\mathcal{A}}+1)$-elementigen Teilmengen von $\{z_1, \ldots, z_n\}$. Nach Definition von $V_{\mathcal{A}}$ existiert zu jedem $i \in \{1, \ldots, k\}$ ein $H_i \subseteq F_i$ mit

$$A \cap F_i \neq H_i \quad \text{für alle } A \in \mathcal{A}$$

(da $\mathcal{A}$ die Menge F_i wegen $|F_i| > V_{\mathcal{A}}$ nicht zerlegt).

Aus $H_i \subseteq F_i \subseteq \{z_1, \ldots, z_n\}$ folgt

$$(A \cap \{z_1, \ldots, z_n\}) \cap F_i \neq H_i \quad \text{für alle } A \in \mathcal{A}.$$

Also gilt

$$\begin{aligned} &\{A \cap \{z_1, \ldots, z_n\} \quad : \quad A \in \mathcal{A}\} \\ &\subseteq \{C \subseteq \{z_1, \ldots, z_n\} \quad : \quad C \cap F_i \neq H_i \text{ für alle } i \in \{1, \ldots, k\}\} =: \mathcal{C}_0. \end{aligned}$$

Also genügt es zu zeigen:

$$|\mathcal{C}_0| \leq \sum_{i=0}^{V_{\mathcal{A}}} \binom{n}{i}.$$

Dies ist einfach, falls $H_i = F_i$ für alle $i \in \{1, \ldots, k\}$. Denn $F_1, \ldots, F_k$ sind alle Teilmengen der Kardinalität $V_{\mathcal{A}} + 1$ von $\{z_1, \ldots, z_n\}$, und für $C \subseteq \{z_1, \ldots, z_n\}$ folgt aus

$$C \cap F_i \neq H_i = F_i \quad \text{für alle } i \in \{1, \ldots, k\},$$

dass C höchstens $V_{\mathcal{A}}$ viele Elemente enthalten kann, was impliziert:

$$|\mathcal{C}_0| \leq \sum_{i=0}^{V_{\mathcal{A}}} \binom{n}{i}.$$

Im Folgenden führen wir den allgemeinen Fall darauf zurück.

Dazu setzen wir

$$H_i' = (H_i \cup \{z_1\}) \cap F_i.$$

Wegen $H_i \subseteq F_i$ wird hier H_i im Falle $z_1 \in F_i$ und $z_1 \notin H_i$ um z_1 erweitert, andernfalls bleibt H_i gleich.

Sodann definieren wir

$$\mathcal{C}_1 := \{C \subseteq \{z_1, \ldots, z_n\} \quad : \quad C \cap F_i \neq H_i' \text{ für alle } i \in \{1, \ldots, k\}\}.$$

Wir zeigen nun

$$|\mathcal{C}_0| \leq |\mathcal{C}_1|. \tag{4.29}$$

Dazu genügt es zu zeigen

$$|\mathcal{C}_0 \setminus \mathcal{C}_1| \leq |\mathcal{C}_1 \setminus \mathcal{C}_0|,$$

und dazu wiederum zeigen wir, dass die Abbildung

$$f : \mathcal{C}_0 \setminus \mathcal{C}_1 \to \mathcal{C}_1 \setminus \mathcal{C}_0, \quad f(C) = C \setminus \{z_1\}$$

wohldefiniert und injektiv ist.

Sei $C \in \mathcal{C}_0 \setminus \mathcal{C}_1$. Dann gilt

$$C \cap F_i \neq H_i \text{ für alle } i \in \{1, \ldots, k\}$$

und

$$C \cap F_{i_0} = H_{i_0}' \text{ für ein } i_0 \in \{1, \ldots, k\}.$$

Also gilt für ein $i_0 \in \{1, \ldots, k\}$:

$$H_{i_0}' = C \cap F_{i_0} \neq H_{i_0}.$$

Nach Definition von H_i' unterscheidet sich dieses aber höchstens um z_1 von H_i, also folgt aus der obigen Beziehung

$$z_1 \in H_{i_0}' = C \cap F_{i_0} \subseteq C.$$

Damit gilt aber für $C \in \mathcal{C}_0 \setminus \mathcal{C}_1$ immer $z_1 \in C$, so dass die obige Abbildung f - sofern wohldefiniert - immer injektiv ist.

Noch zu zeigen: f ist wohldefiniert, d.h. für $C \in \mathcal{C}_0 \setminus \mathcal{C}_1$ gilt immer:

$$C \setminus \{z_1\} \in \mathcal{C}_1 \setminus \mathcal{C}_0.$$

Dazu beachten wir:

1. Wie oben schon gesehen, folgt aus $C \in \mathcal{C}_0 \setminus \mathcal{C}_1$ immer $H'_{i_0} = H_{i_0} \cup \{z_1\}$, $z_1 \notin H_{i_0}$ und $C \cap F_{i_0} = H'_{i_0}$, was impliziert

   $$C \setminus \{z_1\} \cap F_{i_0} = (C \cap F_{i_0}) \setminus \{z_1\} = H'_{i_0} \setminus \{z_1\} = H_{i_0}.$$

   Dies zeigt $C \setminus \{z_1\} \notin \mathcal{C}_0$.

2. Ist nun $z_1 \notin F_1$, so gilt $H_i = H'_i$, was wegen $C \in \mathcal{C}_0$ impliziert

   $$(C \setminus \{z_1\}) \cap F_i = C \cap F_i \neq H_i = H'_i.$$

   Ist dagegen $z_1 \in F_i$, so folgt $z_1 \in H'_i$, was

   $$C \setminus \{z_1\} \cap F_i \neq H'_i$$

   impliziert, da die linke Seite z_1 nicht enthält, die rechte Seite aber schon.

   Also gilt in beiden Fällen $C \setminus \{z_1\} \in \mathcal{C}_1$.

Damit ist (4.29) gezeigt.

Erweitert man nun analog H'_i um $z_2, z_3, \ldots, z_n$, so erhält man

$$|\mathcal{C}_0| \leq |\mathcal{C}_1| \leq \cdots \leq |\mathcal{C}_n|,$$

und bei $\mathcal{C}_n$ erfüllen alle Mengen $H_i^{(n)}$ die Bedingungen des Spezialfalles zu Beginn des Beweises, woraus die Behauptung folgt. □

Zur Abschätzung von Packzahlen einer Menge $\mathcal{G}$ von Funktionen $g : \mathbb{R}^l \to \mathbb{R}$ ist die Betrachtung der VC-Dimension der Menge

$$\mathcal{G}^+ := \left\{ \left\{ (z,t) \in \mathbb{R}^l \times \mathbb{R} \,:\, t \leq g(z) \right\} \quad : \quad g \in \mathcal{G} \right\}$$

aller Untergraphen von $\mathcal{G}$ hilfreich. Genauer gilt:

Satz 4.28. *Sei $B > 0$ und sei $\mathcal{G}$ eine Menge von Funktionen $g : \mathbb{R}^l \to [0, B]$ mit $V_{\mathcal{G}^+} \geq 2$. Dann gilt für jedes Wahrscheinlichkeitsmaß ν auf $\mathbb{R}^l$ und $0 < \epsilon < B/4$ beliebig:*

$$\mathcal{M}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_1(\nu)}\right) \leq 3 \cdot \left(\frac{2 \cdot e \cdot B}{\epsilon} \cdot \log \frac{3 \cdot e \cdot B}{\epsilon} \right)^{V_{\mathcal{G}^+}}.$$

Beweis. Wir zeigen

$$\mathcal{M}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_1(\nu)}\right) \leq 3 \cdot S\left(\mathcal{G}^+, \left\lfloor \frac{B}{\epsilon} \cdot \log\left(2 \cdot \mathcal{M}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_1(\nu)}\right)\right)\right\rfloor\right). \qquad (4.30)$$

Dies impliziert die Behauptung, denn im Falle

$$\left\lfloor \frac{B}{\epsilon} \cdot \log\left(2 \cdot \mathcal{M}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_1(\nu)}\right)\right)\right\rfloor < V_{\mathcal{G}^+}$$

ist diese trivialerweise erfüllt, und im Falle

$$\left\lfloor \frac{B}{\epsilon} \cdot \log\left(2 \cdot \mathcal{M}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_1(\nu)}\right)\right)\right\rfloor \geq V_{\mathcal{G}^+}$$

folgt mit Korollar 4.27 b) aus (4.30)

$$\mathcal{M}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_1(\nu)}\right) \leq 3 \cdot \left(\frac{e \cdot B}{\epsilon \cdot V_{G^+}} \cdot \log\left(2 \cdot \mathcal{M}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_1(\nu)}\right)\right)\right)^{V_{\mathcal{G}^+}},$$

und aus letzterem folgt mit der elementar (aber mühsam) nachrechenbaren Beziehung

$$x \leq 3 \cdot \left(\frac{a}{b} \cdot \log(2 \cdot x)\right)^b \quad \Longrightarrow \quad x \leq 3 \cdot (2 \cdot a \cdot \log(3 \cdot a))^b$$

die Behauptung von Satz 4.21.

Zum Nachweis von (4.30) wählen wir

$$\bar{\mathcal{G}} = \{g_1, \ldots, g_m\}$$

als ϵ-Packung von $\mathcal{G}$ in Bezug auf $\|\cdot\|_{L_1(\nu)}$ mit maximaler Kardinalität

$$m = \mathcal{M}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_1(\nu)\cdot}\right)$$

Weiter seien $Q_1, \ldots, Q_k, T_1, \ldots, T_k$ unbhängige Zufallsvariablen mit $Q_1, \ldots, Q_k$ identisch verteilt mit Verteilung ν und $T_1, \ldots, T_k$ identisch auf $[0, B]$ gleichverteilt. Wir setzen

$$\begin{aligned} R_i &= (Q_i, T_i) \quad (i = 1, \ldots, k) \\ R_1^k &= (R_1, \ldots, R_k) \end{aligned}$$

und

$$G_f = \{(z, t) \,:\, t \leq f(z)\} \quad \text{für } f \in \mathcal{G}.$$

Dann gilt (wobei die erste Gleichheit aus der Definition von s folgt):

$$
\begin{aligned}
&S(\mathcal{G}^+, k)\\
&\geq \mathbf{E}\left\{s\left(\mathcal{G}^+, R_1^k\right)\right\}\\
&\geq \mathbf{E}\left\{s\left(\{G_f \,:\, f \in \bar{\mathcal{G}}\}, R_1^k\right)\right\}\\
&\geq \mathbf{E}\left\{s\left(\left\{G_f \,:\, f \in \bar{\mathcal{G}} \text{ und } G_f \cap R_1^k \neq G_g \cap R_1^k \text{ für alle } g \in \bar{\mathcal{G}} \setminus \{f\}\right\}, R_1^k\right)\right\}\\
&= \mathbf{E}\left\{\sum_{f\in\bar{\mathcal{G}}} I_{\left\{G_f \cap R_1^k \neq G_g \cap R_1^k \text{ für alle } g\in\bar{\mathcal{G}}\setminus\{f\}\right\}}\right\}\\
&= \sum_{f\in\bar{\mathcal{G}}} \mathbf{P}\left\{G_f \cap R_1^k \neq G_g \cap R_1^k \text{ für alle } g \in \bar{\mathcal{G}} \setminus \{f\}\right\}\\
&= \sum_{f\in\bar{\mathcal{G}}} \left(1 - \mathbf{P}\left\{\exists g \in \bar{\mathcal{G}} \setminus \{f\} : G_f \cap R_1^k = G_g \cap R_1^k\right\}\right)\\
&\geq \sum_{f\in\bar{\mathcal{G}}} \left(1 - m \cdot \max_{g\in\bar{\mathcal{G}}\setminus\{f\}} \mathbf{P}\left\{G_f \cap R_1^k = G_g \cap R_1^k\right\}\right).
\end{aligned}
$$

Für beliebige $f, g \in \bar{\mathcal{G}}$ mit $f \neq g$ gilt wegen $R_1, \ldots, R_k$ unabhängig und identisch verteilt

$$
\begin{aligned}
&\mathbf{P}\left\{G_f \cap R_1^k = G_g \cap R_1^k\right\}\\
&= \mathbf{P}\left\{G_f \cap \{R_1\} = G_g \cap \{R_1\}, \ldots, G_f \cap \{R_k\} = G_g \cap \{R_k\}\right\}\\
&= \left(\mathbf{P}\left\{G_f \cap \{R_1\} = G_g \cap \{R_1\}\right\}\right)^k,
\end{aligned}
$$

sowie wegen T_1 auf $[0, B]$ gleichverteilt, $g(Q_1), f(Q_1) \in [0, B]$, Wahl von Q_1 und $\bar{\mathcal{G}}$ ϵ-Packung bzgl. $\|\cdot\|_{L_1(\nu)}$

$$
\begin{aligned}
&\mathbf{P}\left\{G_f \cap \{R_1\} = G_g \cap \{R_1\}\right\}\\
&= 1 - \mathbf{P}\left\{G_f \cap \{R_1\} \neq G_g \cap \{R_1\}\right\}\\
&= 1 - \mathbf{E}\left\{\mathbf{P}\left\{G_f \cap \{R_1\} \neq G_g \cap \{R_1\} \middle| Q_1\right\}\right\}\\
&= 1 - \mathbf{E}\left\{\mathbf{P}\left\{g(Q_1) < T_1 \leq f(Q_1) \text{ oder } f(Q_1) < T_1 \leq g(Q_1) \middle| Q_1\right\}\right\}\\
&= 1 - \mathbf{E}\left\{\frac{|f(Q_1) - g(Q_1)|}{B}\right\}\\
&= 1 - \frac{1}{B}\int |f(x) - g(x)| \nu(dx)\\
&\leq 1 - \frac{\epsilon}{B}.
\end{aligned}
$$

Daraus folgt unter Beachtung von $1 + x \leq e^x$ $(x \in \mathbb{R})$

$$
\mathbf{P}\left\{G_f \cap R_1^k = G_g \cap R_1^k\right\} \leq \left(1 - \frac{\epsilon}{B}\right)^k \leq \exp\left(-\frac{\epsilon \cdot k}{B}\right),
$$

was zusammen mit der oben hergeleiteten Beziehung impliziert

$$S(\mathcal{G}^+, k) \geq m \cdot \left(1 - m \cdot \exp\left(-\frac{\epsilon \cdot k}{B}\right)\right).$$

Wir setzen nun

$$k = \lfloor \frac{B}{\epsilon} \cdot \log(2 \cdot m) \rfloor.$$

Dann gilt

$$\begin{aligned}
&1 - m \cdot \exp\left(-\frac{\epsilon \cdot k}{B}\right) \\
&\geq 1 - m \cdot \exp\left(-\frac{\epsilon}{B} \cdot \left(\frac{B}{\epsilon} \cdot \log(2 \cdot m) - 1\right)\right) \\
&= 1 - m \cdot \frac{1}{2m} \cdot \exp\left(\frac{\epsilon}{B}\right) \\
&= 1 - \frac{1}{2} \cdot \exp\left(\frac{\epsilon}{B}\right) \\
&\geq 1 - \frac{1}{2} \cdot \exp\left(\frac{1}{4}\right) \geq \frac{1}{3}
\end{aligned}$$

und damit auch

$$S\left(\mathcal{G}^+, \lfloor \frac{B}{\epsilon} \cdot \log(2 \cdot m) \rfloor\right) \geq \frac{1}{3} \cdot m,$$

womit (4.30) gezeigt ist. □

Die Anwendung von Satz 4.28 benötigt eine Abschätzung von $V_{\mathcal{G}^+}$. Eine solche liefert:

Satz 4.29. *Sei $\mathcal{G}$ ein r-dimensionaler Vektorraum von reellwertigen Funktionen. Sei*

$$\mathcal{A} = \{\{z : g(z) \geq 0\} \quad : \quad g \in \mathcal{G}\}.$$

Dann gilt

$$V_{\mathcal{A}} \leq r.$$

Ist $\mathcal{G}$ wie in Satz 4.28, so gilt

$$\begin{aligned}
\mathcal{G}^+ &= \left\{\left\{(z,t) \in \mathbb{R}^l \times \mathbb{R} \quad : \quad t \leq g(z)\right\} \quad : \quad g \in \mathcal{G}\right\} \\
&\subseteq \left\{\left\{(z,t) \in \mathbb{R}^l \times \mathbb{R} \quad : \quad g(z) + \alpha \cdot t \geq 0\right\} \quad : \quad g \in \mathcal{G},\, \alpha \in \mathbb{R}\right\}
\end{aligned}$$

und mit Satz 4.29 erhalten wir

$$V_{\mathcal{G}^+} \leq r + 1.$$

Beweis von Satz 4.29: Seien $z_1, \ldots, z_{r+1}$ $(r+1)$ verschiedene Punkte aus dem Definitionsbereich der Funktionen in $\mathcal{G}$. Wir zeigen, dass diese Punkte nicht durch

$$\{\{z \,:\, g(z) \geq 0\} \quad : \quad g \in \mathcal{G}\}$$

zerlegt werden.

Dazu definieren wir

$$L : \mathcal{G} \to \mathbb{R}^{r+1}, \quad L(g) = (g(z_1), \ldots, g(z_{r+1}))^T .$$

Dann ist L lineare Abbildung, und das Bild $L\mathcal{G}$ des r-dimensionalen Vektorraumes $\mathcal{G}$ ist eine höchstens r-dimensionaler Unterraum von $\mathbb{R}^{r+1}$. Folglich existiert ein nichttrivialer Vektor, der senkrecht zu $L\mathcal{G}$ ist, d.h., es existieren $\gamma_1, \ldots, \gamma_{r+1} \in \mathbb{R}^{r+1}$ mit $\gamma_i \neq 0$ für ein i und

$$\gamma_1 \cdot g(z_1) + \cdots + \gamma_{r+1} \cdot g(z_{r+1}) = 0 \tag{4.31}$$

für alle $g \in \mathcal{G}$. OBdA gilt dabei sogar $\gamma_i < 0$ für ein $i \in \{1, \ldots, r+1\}$.

Existiert nun ein $g \in \mathcal{G}$ mit der Eigenschaft, dass

$$\{z \,:\, g(z) \geq 0\}$$

aus $\{z_1, \ldots, z_{r+1}\}$ genau die z_j herausgreift mit $\gamma_j \geq 0$, so hat $g(z_j)$ immer das gleiche Vorzeichen wie γ_j, d.h. es gilt

$$\gamma_j \cdot g(z_j) \geq j \quad (j \in \{1, \ldots, r+1\}).$$

Mit

$$\gamma_i \cdot g(z_i) > 0$$

folgt daraus aber

$$\gamma_1 \cdot g(z_1) + \cdots + \gamma_{r+1} \cdot g(z_{r+1}) > 0$$

im Widerspruch zu (4.31). □

4.9 Analyse von Kleinste-Quadrate-Schätzer

Im Folgenden seien (X, Y), (X_1, Y_1), (X_2, Y_2), ... unabhängige identisch verteilte $\mathbb{R}^d \times \mathbb{R}$-wertige Zufallsvariablen mit $\mathbf{E}\{Y^2\} < \infty$. Wir schätzen

$$m(\cdot) = \arg \min_{f:\mathbb{R}^d \to \mathbb{R}} \mathbf{E}\left\{|f(X) - Y|^2\right\}$$

durch einen Kleinste-Quadrate-Schätzer

$$m_n(\cdot) = \arg\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2 , \tag{4.32}$$

wobei $\mathcal{F}_n$ eine Menge von Funktionen $f : \mathbb{R}^d \to \mathbb{R}$ ist und wir annehmen, dass das Minimum in (4.32) existiert.

Für diesen Schätzer gilt:
Satz 4.30. *Für ein $L \geq 1$ gelte*

$$|Y| \leq L \quad f.s.$$

und

$$\|f\|_\infty = \sup_{x \in \mathbb{R}^d} |f(x)| \leq L \quad \text{für alle } f \in \mathcal{F}_n.$$

Dann gilt für den Kleinste-Quadrate-Schätzer m_n definiert in (4.32):

$$\begin{aligned} &\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\ &\leq \frac{c_1}{n} + \frac{(c_2 + c_3 \log n) \cdot V_{\mathcal{F}_n^+}}{n} + 2 \inf_{f \in \mathcal{F}_n} \int |f(x) - m(x)|^2 \mathbf{P}_X(dx), \end{aligned}$$

wobei $c_1, c_2, c_3 \in \mathbb{R}_+$ nur von L abhängende Konstanten sind.

Beweis. Setze

$$\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} .$$

Wir verwenden die Fehlerzerlegung

$$\begin{aligned} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) &= \mathbf{E}\left\{ |m_n(X) - Y|^2 \middle| \mathcal{D}_n \right\} - \mathbf{E}\left\{ |m(X) - Y|^2 \right\} \\ &= T_{1,n} + T_{2,n} \end{aligned}$$

mit

$$T_{2,n} = 2 \cdot \frac{1}{n} \sum_{i=1}^{n} \left(|m_n(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2 \right)$$

und

$$T_{1,n} = \mathbf{E}\left\{ |m_n(X) - Y|^2 \middle| \mathcal{D}_n \right\} - \mathbf{E}\left\{ |m(X) - Y|^2 \right\} - 2 \cdot T_{2,n}.$$

Für $T_{2,n}$ gilt nach (4.32):

$$\mathbf{E}\{T_{2,n}\} = 2 \cdot \mathbf{E}\left\{ \min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2 - \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2 \right\}$$

$$\begin{aligned} &\leq 2 \cdot \inf_{f \in \mathcal{F}_n} \mathbf{E}\left\{ \frac{1}{n}\sum_{i=1}^{n} |f(X_i) - Y_i|^2 - \frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2 \right\} \\ &= 2 \cdot \inf_{f \in \mathcal{F}_n} \int |f(x) - m(x)|^2 \mathbf{P}_X(dx), \end{aligned}$$

also genügt es im Folgenden zu zeigen:

$$\mathbf{E}\{T_{1,n}\} \leq \frac{c_1}{n} + \frac{(c_2 + c_3 \log n) \cdot V_{\mathcal{F}_n^+}}{n}. \tag{4.33}$$

Zum Nachweis von (4.33) sei $t \geq \frac{1}{n}$ beliebig. Analog zur Motivation von Satz 4.21 gilt dann:

$$\begin{aligned} &\mathbf{P}\{T_{1,n} > t\} \\ &= \mathbf{P}\Bigg\{ \mathbf{E}\left\{|m_n(X) - Y|^2 \middle| \mathcal{D}_n\right\} - \mathbf{E}\left\{|m(X) - Y|^2\right\} \\ &\qquad\qquad -\frac{1}{n}\sum_{i=1}^{n} \left(|m_n(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2\right) \\ &\qquad > \frac{1}{2} \cdot \left(t + \mathbf{E}\left\{|m_n(X) - Y|^2 \middle| \mathcal{D}_n\right\} - \mathbf{E}\left\{|m(X) - Y|^2\right\}\right) \Bigg\} \\ &\leq \mathbf{P}\Bigg\{ \exists f \in \mathcal{F}_n : \mathbf{E}\left\{|f(X) - Y|^2\right\} - \mathbf{E}\left\{|m(X) - Y|^2\right\} \\ &\qquad\qquad -\frac{1}{n}\sum_{i=1}^{n} \left(|f(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2\right) \\ &\qquad\qquad > \frac{1}{2} \cdot \left(t + \mathbf{E}\left\{|f(X) - Y|^2\right\} - \mathbf{E}\left\{|m(X) - Y|^2\right\}\right) \Bigg\}, \end{aligned}$$

wobei die letzte Abschätzung aus $m_n(\cdot) \in \mathcal{F}_n$ folgte.

Wenden wir auf den letzten Term Satz 4.21 mit $\alpha = \beta = t/2$, $\epsilon = 1/2$ und $B = L$ an, so erhalten wir

$$\begin{aligned} \mathbf{P}\{T_{1,n} > t\} &\leq 14 \cdot \sup_{x_1^n} \mathcal{N}_1\left(\frac{\frac{t}{2} \cdot \frac{1}{2}}{20 \cdot L}, \mathcal{F}_n, x_1^n\right) \cdot \exp\left(-\frac{\frac{1}{8} \cdot \frac{t}{2} \cdot n}{214 \cdot (1 + 1/2) \cdot L^4}\right) \\ &= 14 \cdot \sup_{x_1^n} \mathcal{N}_1\left(\frac{t}{80 \cdot L}, \mathcal{F}_n, x_1^n\right) \cdot \exp\left(-\frac{t \cdot n}{5136 \cdot L^4}\right). \end{aligned}$$

Mit Hilfe von Lemma 4.23 und Satz 4.28 (wobei wir den Wertebereich der Funktionen von $[-L, L]$ auf $[0, 2L]$ verschieben) lässt sich die Überdeckungszahl abschätzen durch

$$\begin{aligned}
\mathcal{N}_1\left(\frac{t}{80\cdot L}, \mathcal{F}_n, x_1^n\right) &\leq \mathcal{M}_1\left(\frac{t}{80\cdot L}, \mathcal{F}_n, x_1^n\right) \\
&\leq 3\cdot\left(\frac{2\cdot e\cdot(2L)}{t/(80L)}\cdot\log\frac{3\cdot e\cdot(2L)}{t/(80L)}\right)^{V_{\mathcal{F}_n^+}} \\
&\leq 3\cdot\left(480\cdot e\cdot L^2 n\right)^{2\cdot V_{\mathcal{F}_n^+}},
\end{aligned}$$

wobei wir in der letzten Zeile $t \geq 1/n$ und $\log(x) \leq x$ benutzt haben. Damit erhalten wir

$$\mathbf{P}\{T_{1,n} > t\} \leq 42\cdot\left(480\cdot e\cdot L^2 n\right)^{2\cdot V_{\mathcal{F}_n^+}}\cdot\exp\left(-\frac{t\cdot n}{5136\cdot L^4}\right),$$

und für beliebiges $\epsilon > 1/n$ folgt:

$$\begin{aligned}
\mathbf{E}\{T_{1,n}\} &\leq \int_0^\infty \mathbf{P}\{T_{1,n} > t\}\,dt \\
&\leq \int_0^\epsilon 1\,dt + \int_\epsilon^\infty \mathbf{P}\{T_{1,n} > t\}\,dt \\
&\leq \epsilon + \int_\epsilon^\infty 42\cdot\left(480\cdot e\cdot L^2 n\right)^{2\cdot V_{\mathcal{F}_n^+}}\cdot\exp\left(-\frac{t\cdot n}{5136\cdot L^4}\right)dt \\
&= \epsilon + 42\cdot\left(480\cdot e\cdot L^2 n\right)^{2\cdot V_{\mathcal{F}_n^+}}\cdot\exp\left(-\frac{t\cdot n}{5136\cdot L^4}\right)\cdot\frac{(-5136)\cdot L^4}{n}\Bigg|_{t=\epsilon}^{t=\infty} \\
&= \epsilon + 42\cdot\left(480\cdot e\cdot L^2 n\right)^{2\cdot V_{\mathcal{F}_n^+}}\cdot\frac{5136\cdot L^4}{n}\cdot\exp\left(-\frac{\epsilon\cdot n}{5136\cdot L^4}\right).
\end{aligned}$$

Der obige Ausdruck wird minimal für

$$\epsilon = \frac{5136\cdot L^4}{n}\cdot\log\left(42\cdot\left(480\cdot e\cdot L^2 n\right)^{2\cdot V_{\mathcal{F}_n^+}}\right),$$

und damit erhält man

$$\mathbf{E}\{T_{1,n}\} \leq \frac{5136\cdot L^4\cdot\left(\log(42) + 2\cdot V_{\mathcal{F}_n^+}\cdot\log(480\cdot e\cdot L^2 n)\right)}{n} + \frac{5136\cdot L^4}{n}.$$

Damit ist (4.33) gezeigt. □

Bemerkung 4.1: Ist $\mathcal{F}_n$ Teilmenge eines linearen Vektorraumes der Dimension K_n, so gilt nach Satz 4.29:

$$V_{\mathcal{F}_n^+} \leq K_n + 1,$$

und damit gilt nach Satz 4.30:

$$\mathbf{E}\int |m_n(x)-m(x)|^2\mathbf{P}_X(dx) = O\left(\frac{\log n\cdot K_n}{n} + \inf_{f\in\mathcal{F}_n}\int |f(x)-m(x)|^2\mathbf{P}_X(dx).\right)$$

Ist $supp(\mathbf{P}_X)\subseteq\mathbb{R}^d$ beschränkt und m (p,C)-glatt, so führt die Wahl von $\mathcal{F}_n$ als geeignet definierte stückweise Polynome bzgl. äquidistanter Partition auf

$$\inf_{f\in\mathcal{F}_n}\int |f(x)-m(x)|^2\mathbf{P}_X(dx) \overset{(!)}{=} O\left(\frac{1}{K_n^{2p/d}}\right)$$

und es folgt insgesamt:

$$\mathbf{E}\int |m_n(x)-m(x)|^2\mathbf{P}_X(dx) = O\left(\frac{\log n\cdot K_n}{n} + \frac{1}{K_n^{2p/d}}\right).$$

Minimierung dieser oberen Schranke bzgl. K_n führt auf

$$K_n \approx \left(\frac{n}{\log n}\right)^{\frac{d}{2p+d}}$$

und damit erhalten wir:

$$\mathbf{E}\int |m_n(x)-m(x)|^2\mathbf{P}_X(dx) = O\left(\left(\frac{\log n}{n}\right)^{\frac{2p}{2p+d}}\right).$$

Bemerkung 4.2: In Bemerkung 4.1 lässt sich der logarithmische Faktor durch Verwendung lokaler Überdeckungen vermeiden.

Bemerkung 4.3: Die Rate in Bemerkung 4.1 wird schlecht für d groß. Ein Ausweg ist, zusätzlich strukturelle Annahmen an die Bauart Regressionsfunktion zu machen. Z.B. ermöglicht die Annahme des sogenannten *additiven Modells*

$$m(x^{(1)},\ldots,x^{(d)}) = m_1(x^{(1)})+\cdots+m_d(x^{(d)}),$$

mit Hilfe des Prinzips der Kleinsten-Quadrate genauso aufgebaute Funktionen an die zu schätzende Regressionsfunktion anzupassen. Da dann die Komplexität des Funktionenraumes der im eindimensionalem Fall entspricht, erhält man in diesem Fall die entsprechende eindimensionale Rate.

Bemerkung 4.4: Sinnvoll ist Satz 4.30 (bzw. verwandte Abschätzungen mit Überdeckunsgzahlen statt VC-Dimension) vor allem im Falle nichtlinearer Funktionenrä ume, z.B. neuronaler Netze.