Skript zur Vorlesung

Nichtparametrische Regressionsschätzung

von Prof. Dr. Michael Kohler

Sommersemester 2011

Inhaltsverzeichnis

1 Einführung **3**

1.1 Historische Vorbemerkungen 3

1.2 Regressionsanalyse 4

1.3 Regressionsschätzung 6

1.4 Anwendung in der Mustererkennung 7

1.5 Inhalt dieser Vorlesung 10

2 Ein Slow-Rate-Resultat **11**

3 Konvergenzgeschwindigkeit des Kernschätzers **17**

4 Minimax-Konvergenzraten **24**

4.1 Motivation 24

4.2 Eine untere Minimax-Konvergenzrate 25

5 Datenabhängige Wahl von Parametern **35**

5.1 Motivation 35

5.2 Unterteilung der Stichprobe 35

5.3 Kreuzvalidierung 40

6 Hilfsmittel aus der Theorie empirischer Prozesse **42**

6.1 Motivation 42

6.2 Uniforme Exponentialungleichungen 43

6.3 Abschätzung von Überdeckungszahlen 46

7 Analyse von Kleinste-Quadrate-Schätzer **57**

Kapitel 1

Einführung

1.1 Historische Vorbemerkungen

Einige Daten zur Regressionsschätzung:

1632 Galileo Galileo bearbeitet ein Problem der linearen Regression (ihm liegen Messwerte vor, die nach Theorie auf einer Geraden liegen müssen, aufgrund von Messfehlern aber nicht auf einer Geraden liegen).

1805 A. M. Legendre und C. F. Gauß schlagen unabhängig voneinander die Methode der Kleinsten-Quadrate vor.

ca. 1900 Sir F. Galton und sein Schüler K. Pearson führen den Begriff der Regression ein (im Rahmen von Untersuchungen zum Zusammenhang der Korpergröße von Vätern und Söhnen. Dabei haben sehr große (bzw. sehr kleine) Väter etwas kleinere (bzw. etwas größere) Söhne, d.h. die Körpergröße "schreitet zurück" in Richtung des durchschnittlichen Wertes).

Lange Zeit wurden ausschließlich parametrische Verfahren verwendet (bei denen die Bauart der zur schätzenden Regressionsfunktion als bekannt voraus gesetzt wird und nur von endlich vielen unbekannten Parametern abhängt).

1964 E. A. Nadaraya und G. S. Watson schlagen den Kernschätzer vor (ein nichtparametrisches Verfahren).

1.2 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.

Beispiele:

a) Y = Wert einer Immobilie,
X = Beschreibung der Immobilie.

Ziel ist hier primär die *Interpretation* des Zusammenhangs zwischen X und Y.

b) Y = prozentualer Anteil an Körperfett (exakte Messung benötigt Volumen einer Person)
X = Vektor einfach messbarer Größen wie z.B. elektrischer Widerstand der Haut, Größe, Gewicht und Alter.

Ziel ist hier primär die *Vorhersage von Werten* (d.h. ausgehend vom Wert von X soll der Wert von Y vorhergesagt werden).

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 1.1 *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*

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

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

$$\mathbf{E}\left\{|f(X) - Y|^2\right\} = \mathbf{E}\left\{|m(X) - Y|^2\right\} + \int_{\mathbb{R}^d} |f(x) - m(x)|^2 \mathbf{P}_X(dx). \tag{1.1}$$

Wegen

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

folgt daraus die Behauptung.

Zum Nachweis von (1.1) 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 (1.1) impliziert.

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

$$\begin{aligned} \mathbf{E}\left\{|f(X) - Y|^2\right\} &= \mathbf{E}\left\{|(f(X) - m(X)) + (m(X) - Y)|^2\right\} \\ &= \mathbf{E}\left\{|f(X) - m(X)|^2\right\} + \mathbf{E}\left\{|m(X) - Y|^2\right\}, \end{aligned} \tag{1.2}$$

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 (1.2) folgt nun die Behauptung. □

Bemerkung. Gemäß dem obigen Beweis (siehe (1.1)) 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.

1.3 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.

1.4 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:

$$\mathbf{P}\{g^*(X) \neq Y\} = \min_{g:\mathbb{R}^d \to \{0,1\}} \mathbf{P}\{g(X) \neq Y\}. \tag{1.3}$$

Es gilt:

Lemma 1.2 *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 (1.3).

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:

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

□

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 1.1 *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 1.1.

Gemäß Beweis von Lemma 1.2 gilt:

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

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

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

Mit Lemma 1.2 folgt daraus

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

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

1.5 Inhalt dieser Vorlesung

Ziel dieser Vorlesung ist die Herleitung mathematischer Aussagen zur Regressionsschätzung, die möglichst allgemein (und damit in möglichst vielen Anwendungen) gelten. Dabei werden nichtparametrische Verfahren untersucht, die keine Annahmen an die Bauart der zu schätzenden Regressionsfunktion machen.

In der Vorlesung "Mathematische Statistik", WS 10/11, wurde bereits gezeigt:

Es existieren Schätzverfahren m_n mit

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \to 0 \quad (n \to \infty) \tag{1.4}$$

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

Z.B. gilt diese Aussage für den sogenannten **Kernschätzer**

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

mit naivem Kern $K = 1_{S_1(0)}$ (wobei $S_1(0)$ die Kugel um 0 mit Radius 1 ist) und Bandbreite $h_n > 0$, die so gewählt ist, dass gilt:

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

In dieser Vorlesung untersuchen wir primär Fragen zur Geschwindigkeit, mit der in (1.4) die Konvergenz gegen Null erfolgt.

Kapitel 2

Ein Slow-Rate-Resultat

In diesem Kapitel 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 2.1 *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:

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

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 2.1 benötigen wir das folgende deterministische Lemma.

Lemma 2.1 *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

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

$$\begin{aligned} &= \left(1-\frac{1}{2n}\right)^n \cdot \sum_{i=n}^{\infty} 2\cdot(a_i - a_{i+1}) \\ &= \left(1-\frac{1}{2n}\right)^n \cdot 2 \cdot a_n \\ &\geq a_n \end{aligned}$$

für n genügend groß, da

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

□

Beweis von Satz 2.1:

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

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

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

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 $C = (C_1, C_2, \ldots)$:

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

Im Falle $\mu_n(A_j) = 0$ gilt $X_1 \notin A_j$, ..., $X_n \notin A_j$, was impliziert, dass (X_1, Y_1), ..., (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

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

Wegen

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

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 2.1 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. □

Kapitel 3

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 3.1 *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 3.1 *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

$$S := supp(\mathbf{P}_X) \quad \textit{ist beschränkt,} \tag{3.1}$$

$$\mathbf{Var}\{Y|X=x\} \leq \sigma^2 \quad \textit{für alle } x \in S \tag{3.2}$$

und

$$|m(x) - m(z| \leq C \cdot \|x-z\|^p \quad \textit{für alle } x,z \in S \tag{3.3}$$

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 3.2 *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 $S_{h_n/2}(z_1)$, ..., $S_{h_n/2}(z_{l_n})$ mit Radius $h_n/2$ so, dass gilt

$$S \subseteq \cup_{l=1}^{l_n} S_{h_n/2}(z_l). \tag{3.4}$$

Wegen

$$S_{h_n/2}(z_l) \subseteq S_{h_n}(x) \tag{3.5}$$

für $x \in S_{h_n/2}(z_l)$ gilt dann

$$\begin{aligned}
\int_S \frac{1}{n \cdot \mathbf{P}_X(S_{h_n}(x))} \mathbf{P}_X(dx) &\overset{(3.4)}{\leq} \sum_{l=1}^{l_n} \int_{S_{h_n/2}(z_l)} \frac{1}{n \cdot \mathbf{P}_X(S_{h_n}(x))} \mathbf{P}_X(dx) \\
&\overset{(3.5)}{\leq} \sum_{l=1}^{l_n} \int_{S_{h_n/2}(z_l)} \frac{1}{n \cdot \mathbf{P}_X(S_{h_n/2}(z_l))} \mathbf{P}_X(dx) \\
&= \sum_{l=1}^{l_n} \frac{1}{n \cdot \mathbf{P}_X(S_{h_n/2}(z_l))} \cdot \mathbf{P}_X(S_{h_n/2}(z_l)) \\
&\leq \frac{l_n}{n} \leq \frac{\hat{c}}{n \cdot h_n^d}.
\end{aligned}$$

□

Beweis von Satz 3.1: 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

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

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

$$\begin{aligned}&|\hat{m}_n(x)-m(x)|^2\\&=\left|\frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)\cdot(m(X_i)-m(x))}{\sum_{j=1}^n K\left(\frac{x-X_j}{h_n}\right)}\right|^2\cdot I_{B_n(x)}+|m(x)|^2\cdot I_{B_n(x)^c}\\&\le\frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)\cdot|m(X_i)-m(x)|^2}{\sum_{j=1}^n K\left(\frac{x-X_j}{h_n}\right)}\cdot I_{B_n(x)}+|m(x)|^2\cdot I_{B_n(x)^c}\\&\overset{(3.3)}{\le}\frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)\cdot C^2\cdot\|X_i-x\|^{2p}}{\sum_{j=1}^n K\left(\frac{x-X_j}{h_n}\right)}\cdot I_{B_n(x)}+|m(x)|^2\cdot I_{B_n(x)^c}\\&\le C^2\cdot h_n^{2p}+|m(x)|^2\cdot I_{B_n(x)^c},\end{aligned}$$

bzw.

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

Mit

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

und Lemma 3.2 folgt daraus

$$\begin{aligned}
&\mathbf{E}\int |\hat{m}_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\leq C^2 \cdot h_n^{2p} + \sup_{z\in S} |m(z)|^2 \cdot \int \frac{1}{e} \cdot \frac{1}{n \cdot \mathbf{P}_{X_1}\left(S_{h_n}(x)\right)} \mathbf{P}_X(dx) \\
&\leq C^2 \cdot h_n^{2p} + \sup_{z\in S} |m(z)|^2 \cdot \frac{1}{e} \cdot \frac{\hat{c}}{n \cdot h_n^d}. \qquad (3.6)
\end{aligned}$$

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

$$\begin{aligned}
& \mathbf{E}\left\{|m_n(x) - \hat{m}_n(x)|^2 \middle| X_1, \ldots, X_n\right\} \\
\leq \quad & \mathbf{E}\left\{ \left| \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right) \cdot (Y_i - m(X_i))}{\sum_{j=1}^n K\left(\frac{x-X_j}{h_n}\right)} \right|^2 \middle| X_1, \ldots, X_n \right\} \\
= \quad & \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)^2 \cdot \mathbf{E}\left\{|Y_i - m(X_i)|^2 \middle| X_1, \ldots, X_n\right\}}{\left(\sum_{j=1}^n K\left(\frac{x-X_j}{h_n}\right)\right)^2}
\end{aligned}$$

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

$\sum_{j=1}^n K\left(\frac{x-X_j}{h_n}\right)$ ist $b(n, \mathbf{P}_X(S_{h_n}(x)))$-verteilt. Nach Lemma 4.4 aus der Vorlesung Mathematische Statistik im WS 10/11 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 3.2

$$\begin{aligned} &\mathbf{E}\left\{\int |m_n(x) - \hat{m}_n(x)|^2 \mathbf{P}_X(dx)\right\} \\ &= \int \mathbf{E}\left\{\mathbf{E}\left\{|m_n(x) - \hat{m}_n(x)|^2 \middle| X_1, \ldots, X_n\right\}\right\} \mathbf{P}_X(dx) \\ &\leq \sigma^2 \cdot \int \mathbf{E}\left\{\frac{1}{\sum_{j=1}^n K\left(\frac{x-X_j}{h_n}\right)} \cdot I_{\{n\cdot\mu_n(S_{h_n}(x))>0\}}\right\} \mathbf{P}_X(dx) \\ &\leq \sigma^2 \cdot \int \frac{2}{(n+1)\cdot \mathbf{P}_X(S_{h_n}(x))} \mathbf{P}_X(dx) \\ &\leq \sigma^2 \cdot 2 \cdot \frac{\hat{c}}{n\cdot h_n^d}. \end{aligned} \tag{3.7}$$

Aus (3.6) und (3.7) folgt nun die Behauptung. □

Um unter den Voraussetzungen in Satz 3.1 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

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

möglichts 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 3.1 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 3.1 *Unter den Voraussetzung von Satz 3.1 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.

Kapitel 4

Minimax-Konvergenzraten

4.1 Motivation

Gemäß dem letzten Kapitel 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

$$\sup_{(X,Y)\in\mathcal{D}} \mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \tag{4.1}$$

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 (4.1) minimal wird, d.h. genauer, dass (4.1) asymptotisch wie

$$\inf_{\tilde{m}_n} \sup_{(X,Y)\in\mathcal{D}} \mathbf{E} \int |\tilde{m}_n(x) - m(x)|^2 \mathbf{P}_X(dx) \tag{4.2}$$

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 (4.2) der zu erwartende L_2-Fehler.

Die obigen Überlegungen formalisieren wir in

Definition 4.1 *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 $p \in (0, 1]$, $C_1, C_2 > 0$ und ist $\mathcal{D}$ die Klasse aller Verteilungen von (X, Y) mit $X \in [0, 1]^d$ *f.s.*, $\sup_{x\in[0,1]^d} \mathbf{Var}\{Y|X = x\} \leq c_1$, $\sup_{x\in[0,1]^d} |m(x)| \leq c_2$ und $|m(x) - m(z)| \leq c_3 \cdot \|x - z\|^p$ 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.

4.2 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 3.1 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.

Zur Definition von $\tilde{\mathcal{D}}$ verwenden wir:

Definition 4.2 *Sei $p = k + \beta$ für ein $k \in \mathbb{N}_0$ und $0 < \beta \leq 1$. Sei $C > 0$. Eine Funktion $f : \mathbb{R}^d \to \mathbb{R}$ heißt (p, C)-**glatt**, falls für jedes $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ mit $\sum_{j=1}^d \alpha_j = k$ die partielle Ableitung*

$$\frac{\partial^k f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}$$

existiert und für diese gilt:

$$\left| \frac{\partial^k f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}(x) - \frac{\partial^k f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}(z) \right| \leq C \cdot \|x - z\|^\beta$$

für alle $x, z \in \mathbb{R}^d$.

Bem. Für $p \leq 1$ gilt:

$$m \quad (p, C)\text{-glatt} \quad \Leftrightarrow \quad \forall x, z \in \mathbb{R}^d : \, |m(x) - m(z)| \leq C \cdot \|x - z\|^p.$$

Im Fall $p \leq 1$ betrachten wir als Unterklasse von $\mathcal{D}$:

Definition 4.3 *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. *$|m(x)| \leq 1$ für $x \in [0,1]^d$.*

Das Hauptresultat von Kapitel 4 ist

Satz 4.1 *Seien $p, C > 0$ und $\mathcal{D}^{(p,C)}$ definiert wie oben. Dann ist*

$$\left(n^{-\frac{2p}{2p+d}} \right)_{n \in \mathbb{N}} \tag{4.3}$$

eine untere Minimax-Konvergenzrate für $\mathcal{D}^{(p,C)}$.

Im Falle $p \leq 1$ ist damit (4.3) die optimale Minimax-Konvergenzrate für die Klasse $\mathcal{D}$ aus Abschnitt 4.1.

Im Beweis von Satz 4.1 benötigen wir:

Lemma 4.1 *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 $N^{(1)}$, $\ldots$, $N^{(l)}$ 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

$$\begin{aligned}
&\mathbf{P}\{g(Z) \neq C\} \\
&= \mathbf{P}\{g(C\cdot u + N) \neq C\} \\
&= \mathbf{P}\{g(C\cdot u + N) \neq C, C=1\} + \mathbf{P}\{g(C\cdot u + N) \neq C, C=-1\} \\
&= \mathbf{P}\{g(-u+N) = -1, C=1\} + \mathbf{P}\{g(u+N)=1, C=-1\} \\
&= \mathbf{P}\{g(-u+N)=-1\} \cdot \mathbf{P}\{C=1\} + \mathbf{P}\{g(u+N)=1\} \cdot \mathbf{P}\{C=-1\} \\
&= \frac{1}{2} \cdot \mathbf{P}\{g(-u+N)=-1\} + \frac{1}{2} \cdot \mathbf{P}\{g(u+N)=1\}.
\end{aligned}$$

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

$$\begin{aligned} L^* &= \mathbf{P}\left\{g^*(Z) \neq C\right\} \\ &= \mathbf{P}\left\{g^*(Cu+N) \neq C, C=1\right\} + \mathbf{P}\left\{g^*(Cu+N) \neq C, C=-1\right\} \\ &= \frac{1}{2}\cdot\mathbf{P}\left\{g^*(u+N) = -1\right\} + \frac{1}{2}\cdot\mathbf{P}\left\{g^*(-u+N) = 1\right\} \\ &= \frac{1}{2}\cdot\mathbf{P}\left\{< u+N, u > \quad \leq 0\right\} + \frac{1}{2}\cdot\mathbf{P}\left\{< -u+N, u > \quad > 0\right\} \\ &= \frac{1}{2}\cdot\mathbf{P}\left\{\|u\|^2 + < u, N > \quad \leq 0\right\} + \frac{1}{2}\cdot\mathbf{P}\left\{-\|u\|^2 + < u, N > \quad > 0\right\} \\ &= \frac{1}{2}\cdot\mathbf{P}\left\{< u, N > \quad \leq -\|u\|^2\right\} + \frac{1}{2}\cdot\mathbf{P}\left\{< u, N > \quad > \|u\|^2\right\}. \end{aligned}$$

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 Konvexkombination von unabhängigen standardnormalverteilten Zufallsvariablen selbst 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 4.1: Wir beweisen Satz 4.1 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

$$\left| \left(\frac{d}{dx}\right)^k m^{(c_n)}(x) - \left(\frac{d}{dx}\right)^k m^{(c_n)}(z) \right|$$

$$
\begin{aligned}
&= |c_{n,i}| \cdot \left| \left(\frac{d}{dx}\right)^k g_{n,i}(x) - \left(\frac{d}{dx}\right)^k g_{n,i}(z) \right| \\
&= 1 \cdot M_n^{-p} \cdot M_n^k \cdot C \cdot 2^{\beta-1} \left| M_n(x - a_{n,i}) - M_n(z - a_{n,i}) \right|^\beta \\
&\leq C \cdot 2^{\beta-1} \cdot |x - z|^\beta \leq C \cdot |x - z|^\beta.
\end{aligned}
$$

(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:

$$\liminf_{n\to\infty} \inf_{m_n} \sup_{(X,Y)\in\bar{\mathcal{D}}_n^{(p,C)}} \frac{M_n^{2p}}{C^2} \cdot \mathbf{E} \int |m_n(x) - m^{(c_n}(x)|^2 dx > 0. \tag{4.4}$$

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 $\{m^{(c_n)} : c_n \in C_n\}$ 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

$$\begin{aligned}
&\int |m_n(x) - m^{(c_n)}(x)|^2 dx \\
&\geq \int |\hat{m}_n(x) - m^{(c_n)}(x)|^2 dx \\
&= \sum_{j=1}^{M_n} \int_{A_{n,j}} |\hat{c}_{n,j} \cdot g_{n,j}(x) - c_{n,j} \cdot g_{n,j}(x)|^2 dx \\
&= \sum_{j=1}^{M_n} |\hat{c}_{n,j} - c_{n,j}|^2 \cdot \int_{A_{n,j}} g_{n,j}^2(x)\, dx \\
&= \int g^2(x)\, dx \cdot \frac{1}{M_n^{2p+1}} \cdot \sum_{j=1}^{M_n} |\hat{c}_{n,j} - c_{n,j}|^2.
\end{aligned}$$

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

$$\int |m_n(x) - m^{(c_n)}(x)|^2 dx \geq \int g^2(x)\, dx \cdot \frac{1}{M_n^{2p+1}} \cdot \sum_{j=1}^{M_n} I_{\{\tilde{c}_{n,j} \neq c_{n,j}\}},$$

also folgt (4.4) aus

$$\liminf_{n\to\infty} \inf_{\hat{c}_n} \sup_{c \in \mathcal{C}_n} \frac{1}{M_n} \sum_{j=1}^{M_n} \mathbf{P}\left\{\tilde{c}_{n,j} \neq c_{n,j}\right\} > 0. \tag{4.5}$$

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

$$\begin{aligned} &\inf_{\hat{c}_n} \sup_{c_n \in \mathcal{C}_n} \frac{1}{M_n} \sum_{j=1}^{M_n} \mathbf{P}\left\{\hat{c}_{n,j} \neq c_{n,j}\right\} \\ &\geq \inf_{\hat{c}_n} \frac{1}{M_n} \sum_{j=1}^{M_n} \mathbf{P}\left\{\hat{c}_{n,j} \neq C_{n,j}\right\}. \end{aligned}$$

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

$$\mathbf{P}\left\{\hat{c}_{n,j} \neq C_{n,j}\right\} \geq \mathbf{P}\left\{\bar{C}_{n,j} \neq C_{n,j}\right\} = \mathbf{P}\left\{\bar{C}_{n,1} \neq C_{n,1}\right\}$$

und wir erhalten

$$\inf_{\hat{c}_n} \sup_{c_n \in \mathcal{C}_n} \frac{1}{M_n} \sum_{j=1}^{M_n} \mathbf{P}\left\{\hat{c}_{n,j} \neq c_{n,j}\right\} \geq \mathbf{P}\left\{\bar{C}_{n,1} \neq C_{n,1}\right\}.$$

Also genügt es zu zeigen:

$$\liminf_{n\to\infty} \mathbf{P}\left\{\bar{C}_{n,1} \neq C_{n,1}\right\} > 0. \tag{4.6}$$

4. Schritt: Nachweis von (4.6).

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

$$(Y_{i_1}, \ldots, Y_{i_l}) = C_{n,1} \cdot (g_{n,1}(X_{i_1}), \ldots, g_{n,1}(X_{i_l})) + (N_{i_1}, \ldots, N_{i_l}). \tag{4.7}$$

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 (4.7) 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 4.1

$$\begin{aligned} \mathbf{P}\left\{\bar{C}_{n,1} \neq C_{n,1} \middle| X_1, \ldots, X_n\right\} &= \Phi\left(-\sqrt{\sum_{r=1}^{l} g_{n,1}^2(X_{i_r})}\right) \\ &= \Phi\left(-\sqrt{\sum_{i=1}^{n} g_{n,1}^2(X_i)}\right), \end{aligned}$$

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

$$\begin{aligned}
\mathbf{P}\left\{\bar{C}_{n,1} \neq C_{n,1}\right\} &= \mathbf{E}\left\{\Phi\left(-\sqrt{\sum_{i=1}^{n} g_{n,1}^2(X_i)}\right)\right\} \\
&\geq \Phi\left(-\sqrt{\mathbf{E}\left\{\sum_{i=1}^{n} g_{n,1}^2(X_i)\right\}}\right) \\
&= \Phi\left(-n \cdot \int g_{n,1}^2(x)dx\right) \\
&= \Phi\left(-n \cdot M_n^{-(2p+1)} \cdot C^2 \int \bar{g}^2(x)dx\right) \\
&\geq \Phi\left(-\int \bar{g}^2(x)dx\right),
\end{aligned}$$

da

$$M_n = \lceil (C^2 \cdot n)^{\frac{1}{2p+1}} \rceil \geq (C^2 \cdot n)^{\frac{1}{2p+1}}.$$

□

Kapitel 5

Datenabhängige Wahl von Parametern

5.1 Motivation

Die Bandbreite des Kernschätzers in Korollar 3.1, dessen L_2-Fehler gemäß Satz 4.1 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 3.1 beschrieben durch p und C) unbekannt ist.

Nötig ist daher eine datenabhängige Wahl der Bandbreite, die wir in diesem Kapitel untersuchen.

5.2 Unterteilung der Stichprobe

Seien $(X, Y), (X_1, Y_1), (X_2, Y_2), \ldots$ 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.

$$\frac{1}{n_t} \sum_{i=n_l+1}^{n} |Y_i - m_{n_l}^{(h)}(X_i)|^2, \tag{5.1}$$

und wählen dasjenige $\hat{h} \in \mathcal{P}_n$, für das (5.1) minimal wird, d.h. wir setzen

$$\hat{h} = \hat{h}(\mathcal{D}_n) = \arg\min_{h \in \mathcal{P}_n} \frac{1}{n_t} \sum_{i=n_l+1}^{n} |Y_i - m_{n_l}^{(h)}(X_i)|^2. \tag{5.2}$$

Sodann verwenden wir

$$m_n(x) = m_{n_l}^{(h)}(x, \mathcal{D}_{n_l}) \tag{5.3}$$

als Regressionsschätzer. Für diesen gilt:

Satz 5.1 *Sei* $0 < L < \infty$. *Es gelte*

$$|Y| \leq L \quad f.s. \quad \textit{und} \quad \max_{h \in \mathcal{P}_n} \|m_{n_l}^{(h)}(\cdot)\|_\infty \leq L.$$

Sei m_n definiert durch (5.2) und (5.3). Dann gilt für jedes $\delta > 0$:

$$\begin{aligned}
&\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\leq (1+\delta) \cdot \min_{h \in \mathcal{P}_n} \mathbf{E} \int |m_{n_l}^{(h)}(x) - m(x)|^2 \mathbf{P}_X(dx) + c \cdot \frac{1 + \log |\mathcal{P}_n|}{n_t},
\end{aligned}$$

wobei $c = L^2 \cdot (\frac{32}{\delta} + 70 + 38 \cdot \delta)$.

Beweis. Wir verwenden die Fehlerzerlegung

$$\begin{aligned}
&\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&= \mathbf{E}\left\{|m_n(X) - Y|^2 \big| \mathcal{D}_n\right\} - \mathbf{E}\left\{|m(X) - Y|^2\right\} \\
&= \Big(\mathbf{E}\left\{|m_n(X) - Y|^2 \big| \mathcal{D}_n\right\} - \mathbf{E}\left\{|m(X) - Y|^2\right\} \\
&\qquad -(1+\delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^{n} \left\{|m_n(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2\right\} \Big) \\
&\quad +(1+\delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^{n} \left\{|m_n(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2\right\} \\
&=: T_{1,n} + T_{2,n}.
\end{aligned}$$

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

$$\begin{aligned}
\mathbf{E}\{T_{2,n}\} &= \mathbf{E}\left\{(1+\delta) \cdot \min_{h \in \mathcal{P}_n} \frac{1}{n_t} \sum_{i=n_l+1}^{n} \left\{|m_{n_l}^{(h)}(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2\right\}\right\} \\
&\leq (1+\delta) \cdot \min_{h \in \mathcal{P}_n} \mathbf{E}\left\{\frac{1}{n_t} \sum_{i=n_l+1}^{n} \left\{|m_{n_l}^{(h)}(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2\right\}\right\} \\
&= (1+\delta) \cdot \min_{h \in \mathcal{P}_n} \mathbf{E} \int |m_{n_l}^{(h)}(x) - m(x)|^2 \mathbf{P}_X(dx).
\end{aligned}$$

Also genügt es, im Folgenden noch zu zeigen:

$$\mathbf{E}\{T_{1,n}\} \leq c \cdot \frac{1 + \log |\mathcal{P}_n|}{n_t}. \tag{5.4}$$

Zum Nachweis von (5.4) beachten wir, dass für $s > 0$ gilt:

$$
\begin{aligned}
&\mathbf{P}\left\{T_{1,n} > s \middle| \mathcal{D}_{n_l}\right\} \\
&= \mathbf{P}\Bigg\{ \mathbf{E}\left\{|m_n(X) - Y|^2 \middle| \mathcal{D}_n\right\} - \mathbf{E}\left\{|m(X) - Y|^2\right\} \\
&\qquad\qquad -(1+\delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^{n} \left\{|m_n(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2\right\} > s \middle| \mathcal{D}_{n_l} \Bigg\} \\
&\leq \mathbf{P}\Bigg\{ \exists h \in \mathcal{P}_n \,:\, \mathbf{E}\left\{|m_{n_l}^{(h)}(X) - Y|^2 \middle| \mathcal{D}_{n_l}\right\} - \mathbf{E}\left\{|m(X) - Y|^2\right\} \\
&\qquad\qquad -(1+\delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^{n} \left\{|m_{n_l}^{(h)}(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2\right\} > s \middle| \mathcal{D}_{n_l} \Bigg\} \\
&\leq |\mathcal{P}_n| \cdot \max_{h \in \mathcal{P}_n} \mathbf{P}\Bigg\{ \mathbf{E}\left\{|m_{n_l}^{(h)}(X) - Y|^2 \middle| \mathcal{D}_{n_l}\right\} - \mathbf{E}\left\{|m(X) - Y|^2\right\} \\
&\qquad\qquad -(1+\delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^{n} \left\{|m_{n_l}^{(h)}(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2\right\} > s \middle| \mathcal{D}_{n_l} \Bigg\}.
\end{aligned}
$$

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

$$
\begin{aligned}
&\mathbf{P}\Bigg\{ \mathbf{E}\left\{|m_{n_l}^{(h)}(X) - Y|^2 \middle| \mathcal{D}_{n_l}\right\} - \mathbf{E}\left\{|m(X) - Y|^2\right\} \\
&\qquad\qquad -(1+\delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^{n} \left\{|m_{n_l}^{(h)}(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2\right\} > s \middle| \mathcal{D}_{n_l} \Bigg\} \\
&\leq \mathbf{P}\Bigg\{ (1+\delta) \cdot \left(\mathbf{E}\left\{|m_{n_l}^{(h)}(X) - Y|^2 \middle| \mathcal{D}_{n_l}\right\} - \mathbf{E}\left\{|m(X) - Y|^2\right\}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\quad -(1+\delta)\cdot\frac{1}{n_t}\sum_{i=n_l+1}^{n}\left\{|m_{n_l}^{(h)}(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\right\}\\
&> s+\delta\cdot\left(\mathbf{E}\left\{|m_{n_l}^{(h)}(X)-Y|^2\middle|\mathcal{D}_{n_l}\right\}-\mathbf{E}\left\{|m(X)-Y|^2\right\}\right)\Bigg|\mathcal{D}_{n_l}\Bigg\}\\
\leq \mathbf{P}\Bigg\{&\mathbf{E}\left\{|m_{n_l}^{(h)}(X)-Y|^2\middle|\mathcal{D}_{n_l}\right\}-\mathbf{E}\left\{|m(X)-Y|^2\right\}\\
&-\frac{1}{n_t}\sum_{i=n_l+1}^{n}\left\{|m_{n_l}^{(h)}(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\right\}\\
&> \frac{s}{1+\delta}+\frac{\delta}{1+\delta}\cdot\frac{\sigma^2}{16L^2}\Bigg\}.
\end{aligned}
$$

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:

$$
\begin{aligned}
\mathbf{E}\left\{T_{1,n}\right\} &\leq \int_0^\infty \mathbf{P}\{T_{1,n}>s\}\,ds\\
&\leq u+\int_u^\infty \mathbf{P}\{T_{1,n}>s\}\,ds
\end{aligned}
$$

$$\begin{aligned} &\overset{s.o.}{\leq} \quad u + \int_u^\infty |\mathcal{P}_n| \cdot \exp\left(-\frac{n_t \cdot s}{c}\right) ds \\ &= \quad u + |\mathcal{P}_n| \cdot \frac{c}{n_t} \cdot \exp\left(-\frac{n_t \cdot u}{c}\right). \end{aligned}$$

Mit

$$u = \frac{c \cdot \log |\mathcal{P}_n|}{n_t}$$

folgt

$$\mathbf{E}\left\{T_{1,n}\right\} \leq \frac{c \cdot \log |\mathcal{P}_n|}{n_t} + \frac{c}{n_t} = c \cdot \frac{1 + \log |\mathcal{P}_n|}{n_t},$$

w.z.z.w. □

Korollar 5.1 *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 5.1 und Korollar 3.1. □

5.3 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

$$m_n(x) = m_n^{(\hat{p})}(x; \mathcal{D}_n).$$

Kapitel 6

Hilfsmittel aus der Theorie empirischer Prozesse

6.1 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.

6.2 Uniforme Exponentialungleichungen

Nötig in Abschnitt 6.1 sind Abschätzungen für Ausdrücke wie

$$\mathbf{E}\left\{|m_n(X)-Y|^2\big|\mathcal{D}_n\right\} - \frac{1}{n}\sum_{i=1}^n |m_n(X_i)-Y_i|^2.$$

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 6.1 *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 6.1 *(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 10/11. □

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 Kapitel 4. 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 6.2 *(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:

$$\begin{aligned}
& \mathbf{P}\Bigg\{\exists f\in\mathcal{F}:\mathbf{E}\left\{|f(X)-Y|^2\right\} - \mathbf{E}\left\{|m(X)-Y|^2\right\} \\
& \qquad\qquad -\frac{1}{n}\sum_{i=1}^{n}\left(|f(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\right) \\
& \qquad > \epsilon\cdot\left(\alpha+\beta+\mathbf{E}\left\{|f(X)-Y|^2\right\} - \mathbf{E}\left\{|m(X)-Y|^2\right\}\right)\Bigg\} \\
& \le 14\cdot\sup_{x_1^n}\mathcal{N}_1\left(\frac{\beta\cdot\epsilon}{20\cdot B},\mathcal{F},x_1^n\right)\cdot\exp\left(-\frac{\epsilon^2(1-\epsilon)\cdot\alpha\cdot n}{214\cdot(1+\epsilon)\cdot B^4}\right).
\end{aligned}$$

Beweis: erfolgt im Seminar im WS 11/12. □

Im Folgenden: Herleitung von Abschätzungen für Überdeckungszahlen.

6.3 Abschätzung von Überdeckungszahlen

Definition 6.2 *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 $g_1, \ldots, g_N \in \mathcal{G}$ mit*

$$\|g_i - g_j\|_{L_p(\nu)} \geq \epsilon \quad \text{für alle } 1 \leq i < j \leq N$$

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 6.1 *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:

$$\mathcal{N}_p(\epsilon, \mathcal{G}, z_1^n) \leq |\{A \cap \{z_1, \ldots, z_n\} \,:\, A \in \mathcal{A}\}|\,.$$

Definition 6.3 *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*

$$s(\mathcal{A}, \{z_1, \ldots, z_n\}) := |\{A \cap \{z_1, \ldots, z_n\} \quad : \quad A \in \mathcal{A}\}|$$

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 $\mathcal{A}$**

$$S(\mathcal{A}, n) := \max_{z_1,\ldots,z_n \in \mathbb{R}^d} s(\mathcal{A}, \{z_1, \ldots, z_n\})$$

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 6.4 *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*

$$V_{\mathcal{A}} = \sup \{n \in \mathbb{N} \quad : \quad S(\mathcal{A}, n) = 2^n\},$$

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 6.3 *(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 6.1 *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 6.3 und dem binomischen Lehrsatz gilt:

$$\begin{aligned} S(\mathcal{A},n) &\leq \sum_{i=0}^{V_\mathcal{A}} \binom{n}{i} = \sum_{i=0}^{V_\mathcal{A}} n\cdot(n-1)\cdot\dots\cdot(n-i+1)\cdot\frac{1}{i!} \\ &\leq \sum_{i=0}^{V_\mathcal{A}} n^i \cdot \frac{V_\mathcal{A}!}{(V_\mathcal{A}-i)!}\cdot\frac{1}{i!} \\ &= \sum_{i=0}^{V_\mathcal{A}} n^i \cdot \binom{V_\mathcal{A}}{i} = (n+1)^{V_\mathcal{A}}. \end{aligned}$$

b) Ist $V_\mathcal{A}/n \leq 1$, so gilt nach Satz 6.3:

$$\begin{aligned} \left(\frac{V_\mathcal{A}}{n}\right)^{V_\mathcal{A}} \cdot S(\mathcal{A},n) &\leq \sum_{i=0}^{V_\mathcal{A}} \left(\frac{V_\mathcal{A}}{n}\right)^{V_\mathcal{A}} \cdot \binom{n}{i} \\ &\leq \sum_{i=0}^{n} \left(\frac{V_\mathcal{A}}{n}\right)^{i} \cdot \binom{n}{i} \\ &= \left(1+\frac{V_\mathcal{A}}{n}\right)^n \leq e^{V_\mathcal{A}}, \end{aligned}$$

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 6.3: 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 $z_1,\dots,z_n \in \mathbb{R}^d$ beliebig. Wir zeigen:

$$|\{A \cap \{z_1,\dots,z_n\} \quad : \quad A \in \mathcal{A}\}| \leq \sum_{i=0}^{V_\mathcal{A}} \binom{n}{i}.$$

Dazu: Seien $F_1,\dots,F_k$ mit $k = \binom{n}{V_\mathcal{A}+1}$ alle $(V_\mathcal{A}+1)$-elementigen Teilmengen von $\{z_1,\dots,z_n\}$. Nach Definition von $V_\mathcal{A}$ existiert zu jedem $i \in \{1,\dots,k\}$ 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, …, 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

$$|\mathcal{C}_0| \leq |\mathcal{C}_1|. \tag{6.1}$$

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 (6.1) 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 6.4 *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

$$\mathcal{M}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_1(\nu)}\right) \leq 3 \cdot S\left(\mathcal{G}^+, \left\lfloor \frac{B}{\epsilon} \cdot \log\left(2 \cdot \mathcal{M}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_1(\nu)}\right)\right) \right\rfloor \right). \tag{6.2}$$

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 6.1 b) aus (6.2)

$$\mathcal{M}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_1(\nu)}\right) \leq 3 \cdot \left(\frac{e \cdot B}{\epsilon \cdot V_{G^+}} \cdot \log\left(2 \cdot \mathcal{M}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_1(\nu)}\right)\right) \right)^{V_{\mathcal{G}^+}},$$

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 6.2.

Zum Nachweis von (6.2) 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

$$m = \mathcal{M}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_1(\nu)}.\right)$$

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

$$G_f = \{(z,t) \,:\, t \le f(z)\} \quad \text{für } f \in \mathcal{G}.$$

Dann gilt (wobei die erste Gleichheit aus der Definition von s folgt):

$$\begin{aligned} & S(\mathcal{G}^+, k) \\ & \ge \mathbf{E}\left\{s\left(\mathcal{G}^+, R_1^k\right)\right\} \\ & \ge \mathbf{E}\left\{s\left(\{G_f \,:\, f \in \bar{\mathcal{G}}\}, R_1^k\right)\right\} \\ & \ge \mathbf{E}\left\{s\left(\left\{G_f \,:\, f \in \bar{\mathcal{G}} \text{ und } G_f \cap R_1^k \ne G_g \cap R_1^k \text{ für alle } g \in \bar{\mathcal{G}} \setminus \{f\}\right\}, R_1^k\right)\right\} \\ & = \mathbf{E}\left\{\sum_{f \in \bar{\mathcal{G}}} I_{\left\{G_f \cap R_1^k \ne G_g \cap R_1^k \text{ für alle } g \in \bar{\mathcal{G}} \setminus \{f\}\right\}}\right\} \\ & = \sum_{f \in \bar{\mathcal{G}}} \mathbf{P}\left\{G_f \cap R_1^k \ne G_g \cap R_1^k \text{ für alle } g \in \bar{\mathcal{G}} \setminus \{f\}\right\} \\ & = \sum_{f \in \bar{\mathcal{G}}} \left(1 - \mathbf{P}\left\{\exists g \in \bar{\mathcal{G}} \setminus \{f\} : G_f \cap R_1^k = G_g \cap R_1^k\right\}\right) \\ & \ge \sum_{f \in \bar{\mathcal{G}}} \left(1 - m \cdot \max_{g \in \bar{\mathcal{G}} \setminus \{f\}} \mathbf{P}\left\{G_f \cap R_1^k = G_g \cap R_1^k\right\}\right). \end{aligned}$$

Für beliebige $f, g \in \bar{\mathcal{G}}$ mit $f \ne g$ gilt wegen $R_1, \ldots, R_k$ unabhängig und identisch verteilt

$$\begin{aligned} & \mathbf{P}\left\{G_f \cap R_1^k = G_g \cap R_1^k\right\} \\ & = \mathbf{P}\left\{G_f \cap \{R_1\} = G_g \cap \{R_1\}, \ldots, G_f \cap \{R_k\} = G_g \cap \{R_k\}\right\} \end{aligned}$$

$$= \left(\mathbf{P}\left\{G_f \cap \{R_1\} = G_g \cap \{R_1\}\right\}\right)^k,$$

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 (6.2) gezeigt ist. □

Die Anwendung von Satz 6.4 benötigt eine Abschätzung von $V_{\mathcal{G}^+}$. Eine solche liefert:

Satz 6.5 *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 6.5, 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 6.5 erhalten wir

$$V_{\mathcal{G}^+} \leq r + 1.$$

Beweis von Satz 6.5: 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

$$\gamma_1 \cdot g(z_1) + \cdots + \gamma_{r+1} \cdot g(z_{r+1}) = 0 \tag{6.3}$$

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 (6.3). □

Kapitel 7

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

$$m_n(\cdot) = \arg \min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2 , \tag{7.1}$$

wobei $\mathcal{F}_n$ eine Menge von Funktionen $f : \mathbb{R}^d \to \mathbb{R}$ ist und wir annehmen, dass das Minimum in (7.1) existiert.

Für diesen Schätzer gilt:

Satz 7.1 *Für ein $L \geq 1$ gelte*

$$|Y| \leq L \quad f.s.$$

und

$$\|f\|_\infty = \sup_{x \in \mathbb{R}^d} |f(x)| \leq L \quad \textit{für alle } f \in \mathcal{F}_n.$$

Dann gilt für den Kleinste-Quadrate-Schätzer m_n definiert in (7.1):

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

$$\leq \frac{c_1}{n} + \frac{(c_2 + c_3 \log n) \cdot V_{\mathcal{F}_n^+}}{n} + 2 \inf_{f \in \mathcal{F}_n} \int |f(x) - m(x)|^2 \mathbf{P}_X(dx),$$

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 (7.1):

$$\begin{aligned} \mathbf{E}\{T_{2,n}\} &= 2 \cdot \mathbf{E}\left\{\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2 - \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2\right\} \\ &\leq 2 \cdot \inf_{f \in \mathcal{F}_n} \mathbf{E}\left\{\frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2 - \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2\right\} \\ &= 2 \cdot \inf_{f \in \mathcal{F}_n} \int |f(x) - m(x)|^2 \mathbf{P}_X(dx), \end{aligned}$$

also genügt es im Folgenden zu zeigen:

$$\mathbf{E}\{T_{1,n}\} \leq \frac{c_1}{n} + \frac{(c_2 + c_3 \log n) \cdot V_{\mathcal{F}_n^+}}{n}. \tag{7.2}$$

Zum Nachweis von (7.2) sei $t \geq \frac{1}{n}$ beliebig. Analog zur Motivation von Satz 6.2 gilt dann:

$$\begin{aligned} &\mathbf{P}\{T_{1,n} > t\} \\ &= \mathbf{P}\Bigg\{\mathbf{E}\left\{|m_n(X) - Y|^2 \middle| \mathcal{D}_n\right\} - \mathbf{E}\left\{|m(X) - Y|^2\right\} \end{aligned}$$

$$
\begin{aligned}
&\qquad -\frac{1}{n}\sum_{i=1}^{n}\left(|m_n(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\right)\\
&\qquad > \frac{1}{2}\cdot\left(t+\mathbf{E}\left\{|m_n(X)-Y|^2\middle|\mathcal{D}_n\right\}-\mathbf{E}\left\{|m(X)-Y|^2\right\}\right)\Bigg\}\\
\le\ \mathbf{P}\Bigg\{&\exists f\in\mathcal{F}_n:\ \mathbf{E}\left\{|f(X)-Y|^2\right\}-\mathbf{E}\left\{|m(X)-Y|^2\right\}\\
&\qquad -\frac{1}{n}\sum_{i=1}^{n}\left(|f(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\right)\\
&\qquad > \frac{1}{2}\cdot\left(t+\mathbf{E}\left\{|f(X)-Y|^2\right\}-\mathbf{E}\left\{|m(X)-Y|^2\right\}\right)\Bigg\},
\end{aligned}
$$

wobei die letzte Abschätzung aus $m_n(\cdot)\in\mathcal{F}_n$ folgte.

Wenden wir auf den letzten Term Satz 6.2 mit $\alpha=\beta=t/2$, $\epsilon=1/2$ und $B=L$ an, so erhalten wir

$$
\begin{aligned}
\mathbf{P}\{T_{1,n}>t\} &\le 14\cdot\sup_{x_1^n}\mathcal{N}_1\left(\frac{\frac{t}{2}\cdot\frac{1}{2}}{20\cdot L},\mathcal{F}_n,x_1^n\right)\cdot\exp\left(-\frac{\frac{1}{8}\cdot\frac{t}{2}\cdot n}{214\cdot(1+1/2)\cdot L^4}\right)\\
&= 14\cdot\sup_{x_1^n}\mathcal{N}_1\left(\frac{t}{80\cdot L},\mathcal{F}_n,x_1^n\right)\cdot\exp\left(-\frac{t\cdot n}{5136\cdot L^4}\right).
\end{aligned}
$$

Mit Hilfe von Lemma 6.1 und Satz 6.4 (wobei wir den Wertebereich der Funktionen von $[-L,L]$ auf $[0,2L]$ verschieben) lässt sich die Überdeckungszahl abschätzen durch

$$
\begin{aligned}
\mathcal{N}_1\left(\frac{t}{80\cdot L},\mathcal{F}_n,x_1^n\right) &\le \mathcal{M}_1\left(\frac{t}{80\cdot L},\mathcal{F}_n,x_1^n\right)\\
&\le 3\cdot\left(\frac{2\cdot e\cdot(2L)}{t/(80L)}\cdot\log\frac{3\cdot e\cdot(2L)}{t/(80L)}\right)^{V_{\mathcal{F}_n^+}}\\
&\le 3\cdot\left(480\cdot e\cdot L^2 n\right)^{2\cdot V_{\mathcal{F}_n^+}},
\end{aligned}
$$

wobei wir in der letzten Zeile $t\ge 1/n$ und $\log(x)\le x$ benutzt haben. Damit erhalten wir

$$
\mathbf{P}\{T_{1,n}>t\}\le 42\cdot\left(480\cdot e\cdot L^2 n\right)^{2\cdot V_{\mathcal{F}_n^+}}\cdot\exp\left(-\frac{t\cdot n}{5136\cdot L^4}\right),
$$

und für beliebiges $\epsilon>1/n$ folgt:

$$
\mathbf{E}\{T_{1,n}\} \le \int_0^\infty \mathbf{P}\{T_{1,n}>t\}\,dt
$$

$$
\begin{aligned}
&\leq \int_0^{\epsilon} 1\,dt + \int_{\epsilon}^{\infty} \mathbf{P}\{T_{1,n} > t\}\,dt \\
&\leq \epsilon + \int_{\epsilon}^{\infty} 42 \cdot (480 \cdot e \cdot L^2 n)^{2 \cdot V_{\mathcal{F}_n^+}} \cdot \exp\left(-\frac{t \cdot n}{5136 \cdot L^4}\right) dt \\
&= \epsilon + 42 \cdot (480 \cdot e \cdot L^2 n)^{2 \cdot V_{\mathcal{F}_n^+}} \cdot \exp\left(-\frac{t \cdot n}{5136 \cdot L^4}\right) \cdot \frac{(-5136) \cdot L^4}{n} \Bigg|_{t=\epsilon}^{t=\infty} \\
&= \epsilon + 42 \cdot (480 \cdot e \cdot L^2 n)^{2 \cdot V_{\mathcal{F}_n^+}} \cdot \frac{5136 \cdot L^4}{n} \cdot \exp\left(-\frac{\epsilon \cdot n}{5136 \cdot L^4}\right).
\end{aligned}
$$

Der obige Ausdruck wird minimal für

$$\epsilon = \frac{5136 \cdot L^4}{n} \cdot \log\left(42 \cdot (480 \cdot e \cdot L^2 n)^{2 \cdot V_{\mathcal{F}_n^+}}\right),$$

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 (7.2) gezeigt. □

Bemerkung 7.1: Ist $\mathcal{F}_n$ Teilmenge eines linearen Vektorraumes der Dimension K_n, so gilt nach Satz 6.5:

$$V_{\mathcal{F}_n^+} \leq K_n + 1,$$

und damit gilt nach Satz 7.1:

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) = O\left(\frac{\log n \cdot K_n}{n} + \inf_{f \in \mathcal{F}_n} \int |f(x) - m(x)|^2 \mathbf{P}_X(dx).\right)$$

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 7.2: In Bemerkung 7.1 lässt sich der logarithmische Faktor durch Verwendung lokaler Überdeckungen vermeiden.

Bemerkung 7.3: Die Rate in Bemerkung 7.1 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 7.4: Sinnvoll ist Satz 7.1 (bzw. verwandte Abschätzungen mit Überdeckunsgzahlen statt VC-Dimension) vor allem im Falle nichtlinearer Funktionenräume, z.B. neuronaler Netze.