Ingo Bürk

Skript vom Sommersemester 2012

Nichtparametrische Statistik

Mathematics, rightly viewed, possesses not only truth, but supreme beauty.
Bertrand Russell

Universität Stuttgart
2012

Dieses Skript entstand im Rahmen der Vorlesung „Nichtparametrische Statistik“ bei Prof. Ingo Steinwart als Vorlesungsmitschrieb.

Es kann nicht garantiert werden, dass dieses Dokument fehlerfrei ist und der Autor übernimmt für möglicherweise entstandene Schäden jeglicher Art keine Haftung. Dieser Mitschrieb ist kein offizielles Dokument der Universität Stuttgart, Mitarbeiter eben dieser tragen daher ebenfalls keine Verantwortung.

Inhaltsverzeichnis

Vorwort 5

1 Einführung 7
- 1.1 Grundannahmen . . . 7
- 1.2 Problemstellungen . . . 10
- 1.3 Einige klassische Verfahren am Beispiel der binären Klassifikation . . . 14
- 1.4 Konzentrationsungleichungen . . . 16

2 Dichteschätzung 23
- 2.1 Histogrammregel . . . 23
- 2.2 Kernregeln . . . 39

3 Regression 53
- 3.1 Empirische Risikominimierung . . . 53
- 3.2 Eigenschaften der Least-Squares-Verlustfunktion . . . 60
- 3.3 Histogrammregel für Regression . . . 64
- 3.4 Kernregel für Regression . . . 73

4 Klassifikation 75
- 4.1 Eigenschaften des Klassifikationsproblems . . . 75
- 4.2 Das No-Free-Lunch-Theorem . . . 79
- 4.3 Histogramme für Klassifikation . . . 86

Abbildungsverzeichnis 91

Stichwortverzeichnis 93

Vorwort

Die nichtparametrische Statistik, auch parameterfreie Statistik genannt, beschäftigt sich mit parameterfreien statistischen Modellen und Tests. Dies bedeutet, dass über die Struktur der zugrunde liegenden Daten a priori keine Annahmen gemacht werden, sondern erst aus den Daten gewonnen werden. Statistische Tests ohne solche Annahmen haben den Vorteil, auch dann anwendbar zu sein, wenn die notwendigen Voraussetzungen der parametrischen Statistik nicht erfüllt sind oder hierüber zumindest Unklarheit herrscht.

1 Einführung

Im ersten Kapitel wollen wir uns mit den grundlegenden Begriffen und Annahmen auseinandersetzen, die wir in der Regel zugrundelegen werden.

1.1 Grundannahmen

In der *parametrischen Statistik* betrachteten wir eine Familie $(P_\vartheta)_{\vartheta\in\Theta}$ von Wahrscheinlichkeitsmaßen auf X. Ferner hatten wir einen Datensatz $D = (x_1, \dots, x_n) \in X^n$ gegeben, den wir als i. i. d. gemäß eines P_ϑ^n angenommen haben, wobei $\vartheta \in \Theta$ unbekannt ist. Unser Ziel war es dann, mit Hilfe von D etwas über das unbekannte ϑ herauszufinden. Hierbei hatten wir stets $\Theta \subset \mathbf{R}^m$ angenommen.

In der *nichtparametrischen Statistik* wollen wir prinzipiell die selben Dinge machen, doch wir wollen nun auf Parameter wie ϑ verzichten. Statt X werden wir in Zukunft entweder $Z = X$ oder $Z = X \times Y$ betrachten, wobei $Y \subset \mathbf{R}$ ist. Dementsprechend betrachten wir einen Datensatz $D = (z_1, \dots, z_n) \in Z^n$, der i. i. d. gemäß einer (fast) völlig unbekannten Verteilung P^n auf Z^n ist. Mit Hilfe von D wollen wir dann Aussagen über P treffen.

Der offensichtliche Vorteil der nichtparametrischen Statistik ist, dass die Methoden wesentlich allgemeiner einsetzbar sind, da auf schwierig zu überprüfende Annahmen verzichtet werden kann. Auf der anderen Seite sind die Ergebnisse schlechter interpretierbar und die Schlüsse sind in der nichtparametrischen Statistik schwächer als in der parametrischen Statistik.

Im Folgenden werden wir Informationen, die wir mit Hilfe von D über P sammeln, in einer Funktion $f_D : X \to \mathbf{R}$ kodieren. Ferner schreiben wir $\mathscr{L}_0(X)$ für die Menge aller messbaren Funktionen $f : X \to \mathbf{R}$.

Definition 1.1.1 **Lernmethode**
Eine Lernmethode $\mathscr{L}$ ist eine Folge $(L_n)_{n\geq 1}$ von messbaren Abbildungen

$$L_n : Z^n \to \mathscr{L}_0(X) \quad \text{mit} \quad D \mapsto f_D,$$

Wir wollen uns nun zunächst mit der Frage auseinandersetzen, wie wir beurteilen können, welche Funktionen f_D gut sind.

Definition 1.1.2 Verlustfunktion
Ein $L : X \times Y \times \mathbf{R} \to [0,\infty)$ heißt Verlustfunktion, falls L messbar ist.

Häufig werden auch die Fälle einer unüberwachte Verlustfunktion $L : X \times \mathbf{R} \to [0,\infty)$ oder einer überwachten Verlustfunktion $L : Y \times \mathbf{R} \to [0,\infty)$ betrachtet.

Eine Verlustfunktion L misst gewissermaßen durch die Werte $f(x,y,f(x))$, wie gut f die gesuchten Eigenschaften beschreibt.

In der parametrischen Statistik hatten wir auch Verlustfunktionen betrachtet, dort hatten diese jedoch die Form $L(\vartheta,\hat{\vartheta})$, wobei ϑ den wahren Parameter und $\hat{\vartheta}$ den geschätzten Parameter darstellt. Wir sehen, dass sich die Verlustfunktionen in der nichtparametrischen Statistik hiervon unterscheiden.

Definition 1.1.3 Risiko
Sei L eine Verlustfunktion und P ein Wahrscheinlichkeitsmaß auf X oder $X \times Y$. Dann ist das Risiko einer messbaren Abbildung $f : X \to \mathbf{R}$ definiert durch

$$R_{L,P}(f) = \int_{X\times Y} L(x,y,f(x)) \, \mathrm{d}P(x,y),$$

wobei im ersten Fall natürlich nur über X integriert wird.

Man kann leicht einsehen, dass dieses Integral immer definiert ist. Falls (z_i) eine i. i. d. Folge zukünftiger Beobachtungen ist, gilt, falls $R_{L,P}(f) < \infty$ ist, die Identität

$$R_{L,P}(f) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} L(x_i,y_i,f(x_i))$$

P^∞-fast sicher, wobei $z_i = x_i$ oder $z_i = (x_i,y_i)$ ist. Dies gilt nach dem starken Gesetz der großen Zahlen. Da uns P nicht bekannt ist, können wir $R_{L,P}$ jedoch nicht berechnen.

Definition 1.1.4 Bayes-Risiko/-Entscheidungsfunktion
Sei L eine Verlustfunktion und P ein Wahrscheinlichkeitsmaß auf X oder $X \times Y$. Dann heißt das kleinstmögliche Risiko

$$R^*_{L,P} := \inf\{R_{L,P}(f) \mid f : X \to \mathbf{R} \text{ messbar}\}$$

Bayes-Risiko. Eine messbare Funktion $f^*_{L,P} : X \to \mathbf{R}$ mit $R^*_{L,P} = R_{L,P}(f^*_{L,P})$ heißt Bayes-Entscheidungsfunktion.

Im Allgemeinen existiert eine solche Bayes-Entscheidungsfunktion jedoch nicht und wenn sie existiert, so ist sie im Allgemeinen nicht eindeutig.

Eine High-Level-Beschreibung unseres Ziels ist nun die Frage, welche Lernmethoden uns garantieren, dass das Überschussrisiko $R_{L,P}(f_D) - R^*_{L,P}$ klein ist.

1.2 Problemstellungen

Beim *überwachten Lernen* haben wir einen Eingaberaum X, einen Ausgaberaum $Y \subset \mathbf{R}$ und ein Wahrscheinlichkeitsmaß P auf $X \times Y$, welches die Eingabe-/Ausgabe-Relation beschreibt. Unser Ziel ist das Finden einer Abbildung $f_D : X \to \mathbf{R}$ bzw. $f_D : X \to Y$ mit Hilfe von D, so dass f_D die Eingabe-/Ausgabe-Relation gut im Sinne eines geeigneten Risikos beschreibt.

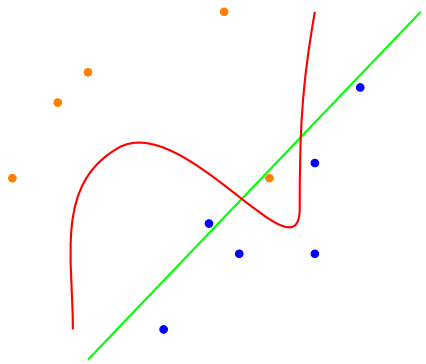

Abbildung 1.1: Darstellung der Problematik der binären Klassifikation aus Beispiel 1.2.1.

Beispiel 1.2.1 *Binäre Klassifikation*
Es sei $Y = \{-1, 1\}$. Unser Ziel ist es, ein f_D zu finden, so dass $f_D(x) = y$ möglichst häufig eintritt. Wir betrachten $X \subset \mathbf{R}^2$ und Abbildung 1.1. Die orangen Punkte stehen für Paare mit $y_i = -1$ und analog die blauen Punkte für Paare mit $y_i = 1$. Es sei $f = \mathbf{1}_A - \mathbf{1}_B$, wobei $A \cap B = \emptyset$ und $A \cup B = X$ ist. Dabei stehen A und B für die Menge der Punkte, welche die jeweilige Funktion jeweils einem Label zuordnet.

Die rote Funktion macht keine Fehler auf D. Die Frage ist jedoch, ob dies auch für zukünftige Daten der Fall ist? Im Allgemeinen werden wir dies nicht wissen, die grüne Funktion könnte, im Sinne des gleich einzuführenden Risikos, durchaus die optimale Funktion sein. Dieses Problem heißt *Overfitting*.

Die grüne Funktion macht die Annahme der Linearität, das heißt, dass das optimale f im Sinne des Risikos linear ist. Dies kann durchaus falsch sein, falls zum Beispiel die rote Funktion optimal wäre. Die Annahme führt also zu systematischen Fehlern. Die nichtparametrische Statistik versucht diese Fehler zu vermeiden. Wenn eine Entscheidungsfunktion f_D wenig an D angepasst ist, sprechen wir von *Underfitting*.

Beide Phänomene konkurrieren miteinander. Wir werden im Zuge der Vorlesung Methoden kennenlernen, um beide sinnvoll auszubalancieren. Typische Anwendungen der binären Klassifikation sind Spamfilter, Diagnose-Systeme, *Fraud Detection* (Betrugserkennung) oder Methoden der Bildauswertung, es gibt jedoch noch wesentlich mehr Anwendungen.

Formalisierung: Eine mögliche Verlustfunktion $L : Y \times \mathbf{R} \to [0, \infty)$ ist gegeben durch

$$L(y, t) := \mathbf{1}_{(-\infty, 0]}(y \cdot \operatorname{sign} t),$$

wobei wir $\operatorname{sign} 0 := 1$ setzen. Wir sehen, dass $L(y,t) = 1$ genau dann eintritt, wenn $\operatorname{sign} t \neq y$ ist und $L(y,t) = 0$ genau dann, wenn $\operatorname{sign} t = y$ ist. Mit anderen Worten wird eine richtige Vorhersage von y durch $\operatorname{sign} t$ nicht bestraft, während eine falsche Vorhersage mit 1 bestraft wird. Damit können wir das Risiko berechnen vermöge

$$\begin{aligned} R_{L,P}(f) &= \int_{X\times Y} L(y, f(x)) \, \mathrm{d}P(x,y) \\ &= P(\{(x,y) \in X \times Y : y \neq \operatorname{sign} f(x)\}). \end{aligned}$$

Das heißt, dass das Risiko von f die Wahrscheinlichkeit einer falschen Vorhersage durch $\operatorname{sign} \circ f$ beschreibt. Dann beschreibt $R^*_{L,P}$ die kleinstmögliche dieser Wahrscheinlichkeiten. //

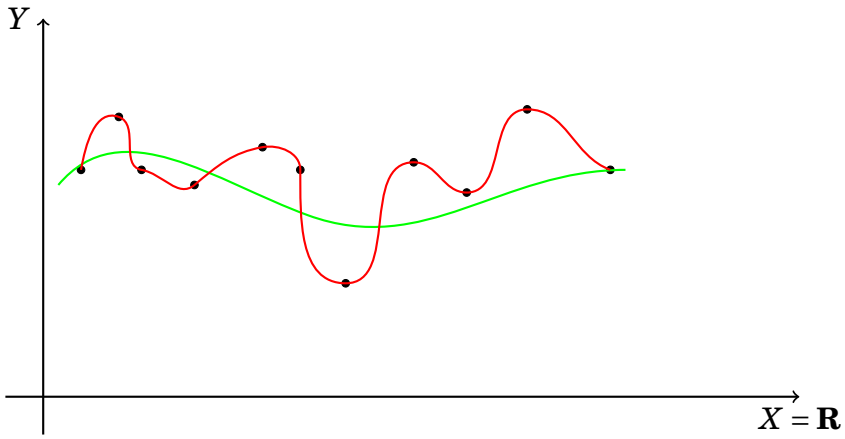

Abbildung 1.2: Darstellung für Beispiel 1.2.2.

Beispiel 1.2.2 *Regression*
Sei nun $Y = [-M, M]$ oder $Y = \mathbf{R}$ und betrachte Abbildung 1.2. Eine mögliche Frage wäre, ob die grüne Linie under- oder die rote Linie overfitted ist. Wir wollen mit Hilfe von D eine Funktion f_D finden, so dass $f_D(x) \approx y$ für neue $(x,y) \in X \times Y$ gilt.

Eine mögliche Formalisierung ist das Betrachten von $L : X \times \mathbf{R} \to [0,\infty)$ mit $L(y,t) := (y-t)^2$. Dies ist die so genannte Least-Squares-Verlustfunktion. Die zugehörige Risikofunktion ist dann gegeben durch

$$R_{L,P}(f) = \int_{X\times Y} (y - f(x))^2 \, \mathrm{d}P(x,y).$$

Dies findet unter anderem in der Finanzindustrie, Preisbestimmung, Ausfallrisiken und der binären Klassifikation Anwendung. //

Beispiel 1.2.3 *Dichteschätzung*
Es sei μ ein bekanntes und σ-endliches Maß auf X, zum Beispiel das Lebesguemaß für $X = \mathbf{R}^d$. Die Verteilung P ist μ-absolut stetig und die entsprechende Dichte von P bezüglich μ sei $h\colon X \to [0,\infty)$. Unser Ziel ist es, diese Dichte h zu schätzen.

Schätzen wir eine gewisse Menge an Daten durch die Dichte einer $\mathcal{N}(\mu,\sigma^2)$-Verteilung, so führt dies im Allgemeinen zum Underfitting. Alternativ könnten wir für jeden Datenpunkt einen kleinen Dichtekreis schätzen, dies führt im Allgemeinen jedoch zum Overfitting.

Wir kommen nun zur Formalisierung. Die grundlegende Idee ist es, den Abstand des wahren h und der Schätzung f durch eine Norm zu beschreiben. Sei zum Beispiel $L\colon X \times \mathbf{R} \to [0,\infty)$ mit $(x,t) \mapsto |h(x) - t|$. Hierbei ist zu beachten, dass wir L nicht kennen, da uns h unbekannt ist. Das Risiko ist nun gegeben durch

$$R_{L,\mu}(f) = \int |h(x) - f(x)| \,\mathrm{d}\mu(x) = \|h - f\|_{L_1(\mu)}.$$

Um dieses Risiko in P anzugeben, können wir alternativ auch

$$\tilde{L}(x,t) := \left|1 - \frac{t}{h(x)}\right|$$

betrachten, was uns wiederum zu

$$R_{\tilde{L},P}(f) = \int \left|1 - \frac{f(x)}{h(x)}\right| h(x) \,\mathrm{d}\mu(x) = R_{L,\mu}(f)$$

führt. ⫽

Beispiel 1.2.4 *Level Set Estimation*
Es seien μ, h und P wie in Beispiel 1.2.3 definiert und es sei $\rho \geq 0$. Unser Ziel ist es nun, $\{h > \rho\}$ beziehungsweise $\{h \geq \rho\}$ zu schätzen.

Dies wird zum Beispiel eingesetzt, um Ausreißer zu identifizieren (engl. *Anomaly Detection*), wobei Punkte $x \in \{h < \rho\}$ als Ausreißer betrachtet werden. Andere Beispiele wären Überwachungssysteme, wobei je nach Anwendung andere Formalisierungen nötig sind.

Wir betrachten $L\colon X \times \mathbf{R} \to [0,\infty)$ vermöge

$$(x,t) \mapsto \mathbf{1}_{(-\infty,0)}((h(x) - \rho)\operatorname{sign} t) = \begin{cases} 1 & \text{für } t \geq 0 \wedge h(x) < \rho \\ 1 & \text{für } t < 0 \wedge h(x) > \rho \\ 0 & \text{sonst} \end{cases}.$$

Die offensichtliche Ähnlichkeit zur binären Klassifikation ist nicht nur oberflächlich. Falls $\mu(\{h = \rho\}) = 0$ ist, so gilt

$$R_{L,\mu}(f) = \mu(\{h \geq \rho\} \triangle \{f \geq 0\}),$$

wobei $A \triangle B$ die symmetrische Differenz zwischen zwei Mengen A und B bezeichnet. ⫽

Beispiel 1.2.5 *Clusteranalyse*
Es seien μ, h, P und ρ definiert wie in Beispiel 1.2.4. Das Ziel ist es nun, die topologischen Zusammenhangskomponenten von $\{h \geq \rho\}$ beziehungsweise $\{h > \rho\}$ zu finden. Hierbei gilt es die Unverträglichkeit von topologischen und maßtheoretischen Begriffen zu beachten.

Clusteranalyse findet Anwendungen in vielen Bereichen wie der Bildverarbeitung, Bioinformatik oder Medizin. Die Formalisierung arbeitet jedoch nicht mit Verlustfunktionen. ∥

1.3 Einige klassische Verfahren am Beispiel der binären Klassifikation

Beispiel 1.3.1 *Histogrammregel*
Es sei $X = \mathbf{R}^d$ und $Y = \{-1, 1\}$. Wir partitionieren X nun in (Hyper-)Würfel Q_j, die jeweils die Länge h besitzen. Für $x \in X$ existiert genau ein Q_j mit $x \in Q_j$. Wir definieren dann

$$f_D(x) := \operatorname{sign}\left(\sum_{x_i \in Q_j} y_i\right).$$

Mit anderen Worten ist $f_D(x)$ gewissermaßen ein Mehrheitsvotum unter den Labels, die im selben Würfel Q_j liegen.

Dieses Verfahren besitzt den freien Parameter h. Große Werte führen im Allgemeinen zum Underfitting, während kleine Werte zum Overfitting führen. Eine mögliche Frage ist, ob wir h unabhängig von P wählen können, so dass beides vermieden wird. //

Beispiel 1.3.2 *Nearest Neighbor*
Wieder sei $X = \mathbf{R}^d$ und $Y = \{-1, 1\}$. Wir fixieren ein $k \in \mathbf{N}$ und bestimmen für $x \in X$ die k Punkte $(x_i, y_i) \in D$, die x am nächsten liegen, zum Beispiel bezüglich der euklidischen Norm. Dann wird ein Mehrheitsvotum über die k nächsten Nachbarn festgelegt vermöge

$$f_D(x) = \operatorname{sign}\left(\sum_{i=1}^{k} y_i\right). \quad //$$

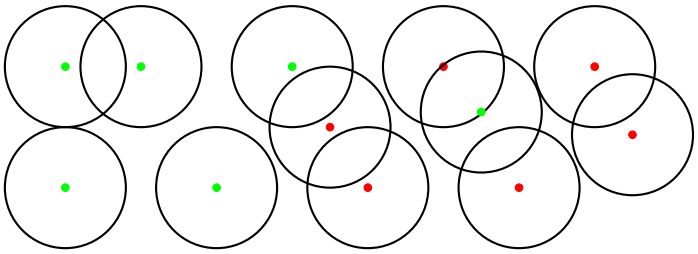

Abbildung 1.3: Darstellung für Beispiel 1.3.3. Der Parameter h bestimmt den Radius der Kreise.

Beispiel 1.3.3 *Moving Window und Kernregeln*
Es sei $K : [0, \infty) \to [0, \infty)$ eine monoton fallende Funktion und $h > 0$ fixiert. Für $x \in X$ definieren wir dann

$$f_D(x) := \operatorname{sign}\left(\sum_{i=1}^{n} y_i K(h^{-1} \|x - x_i\|)\right).$$

Wählt man $K = \mathbf{1}_{[0,1]}$, so nennt man dieses Verfahren moving window rule. Dies ist in Abbildung 1.3 dargestellt. //

1.4 Konzentrationsungleichungen

Unser Ziel in diesem Abschnitt ist die Abschätzung der Abweichung von

$$\frac{1}{n}\sum_{i=1}^{n}\xi_i \quad \text{und} \quad \mathbf{E}\xi_1$$

für unabhängige und identisch verteilte Zufallsvariablen.

Satz 1.4.1 **Markovungleichung**
Sei $(\Omega,\mathscr{A},P)$ ein Wahrscheinlichkeitsraum. Für messbare $f:\Omega\to\mathbf{R}$ und $t>0$ gilt dann

$$P(\{\omega\in\Omega:|f(\omega)|\geq t\})\leq\frac{\mathbf{E}_P|f|}{t}.$$

Beweis: Der Beweis wurde in [WTSkript11] geführt. □

Lemma 1.4.2
Für $x>-1$ gilt

$$(1+x)\ln(1+x)-x\geq\frac{3}{2}\frac{x^2}{x+3}.$$

Beweis: Es sei $f(x):=(1+x)\ln(1+x)-x$ und $g(x):=\frac{3}{2}\frac{x^2}{x+3}$, dann ist $f(x)\geq g(x)$ zu zeigen. Dazu betrachten wir die Ableitungen $f'(x)=\ln(1+x)$, $f''(x)=\frac{1}{1+x}$, $g'(x)=\frac{3}{2}\frac{x^2+6x}{(x+3)^2}$ und $g''(x)=\frac{27}{(x+3)^3}$. Es gilt offenbar $f(0)=0=f'(0)$ und $g(0)=0=g'(0)$. Außerdem gilt $f''(x)\geq g''(x)$, denn

$$\frac{1}{1+x}\geq\frac{27}{(x+3)^3}\iff x^2(x+9)\geq 0,$$

wie man durch leichtes Nachrechnen zeigen kann. Für $x\geq 0$ folgt

$$\begin{aligned}f'(x)=f'(x)-f'(0)&=\int_0^x f''(t)\,\mathrm{d}t\\&\geq\int_0^x g''(t)\,\mathrm{d}t=g'(x)-g'(0)=g'(x).\end{aligned}$$

Wenden wir dieses Argument nochmals an, so erhalten wir $f(x)\geq g(x)$. Für $x\in(-1,0]$ gilt ähnlich hierzu

$$\begin{aligned}-f'(x)=f'(0)-f'(x)&=\int_x^0 f''(t)\,\mathrm{d}t\\&\geq\int_x^0 g''(t)\,\mathrm{d}t=g'(0)-g'(x)=-g'(x).\end{aligned}$$

Auch hier folgt durch nochmaliges Anwenden der Argumentation schließlich $f(x)\geq g(x)$. □

Satz 1.4.3 **Bernsteins Ungleichung**
Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, $B > 0$, $\sigma > 0$ und $n \geq 1$. Ferner seien $\xi_1, \ldots, \xi_n$ unabhängige Zufallsvariablen mit

i) $\mathbf{E}_P \xi_i = 0$

ii) $\|\xi_i\|_\infty \leq B$

iii) $\mathbf{E}_P \xi_i^2 \leq \sigma^2$

für alle $i \in \{1, \ldots, n\}$. Für $\tau > 0$ gilt dann

$$P\left(\frac{1}{n}\sum_{i=1}^{n} \xi_i \geq \sqrt{\frac{2\sigma^2\tau}{n}} + \frac{2B\tau}{3n}\right) \leq e^{-\tau}.$$

Beweis: Für $t \geq 0$ und $\varepsilon > 0$ gilt mit der Markovungleichung

$$\begin{aligned} P\left(\frac{1}{n}\sum_{i=1}^{n} \xi_i \geq \varepsilon\right) &= P\left(\exp\left(t\sum_{i=1}^{n} \xi_i\right) \geq e^{t\varepsilon n}\right) \\ &\leq \frac{\mathbf{E}_P \exp\left(t\sum_{i=1}^{n} \xi_i\right)}{e^{t\varepsilon n}} = e^{-t\varepsilon n} \prod_{i=1}^{n} \mathbf{E}_P e^{t\xi_i}. \end{aligned}$$

Wir betrachten nun die Faktoren in diesem Produkt. Da ξ_i beschränkt ist, ist auch $e^{t\xi_i}$ beschränkt, wodurch der Satz von Lebesgue anwendbar wird. Damit erhalten wir

$$\mathbf{E}_P e^{t\xi_i} = \mathbf{E}_P \sum_{k=0}^{\infty} \frac{t^k}{k!}\xi_i^k = \sum_{k=0}^{\infty} \frac{t^k}{k!}\mathbf{E}_P \xi_i^k = 1 + 0 + \sum_{k=2}^{\infty} \frac{t^k}{k!}\mathbf{E}_P \xi_i^k$$

und wegen $\xi_i^k = \xi_i^2 \xi_i^{k-2}$ erhalten wir mit iii) und ii)

$$\begin{aligned} &\leq 1 + \sum_{k=2}^{\infty} \frac{t^k}{k!}\sigma^2 B^{k-2} = 1 + \frac{\sigma^2}{B^2}\sum_{k=2}^{\infty} \frac{t^k}{k!}B^k \\ &= 1 + \frac{\sigma^2}{B^2}\left(e^{tB} - tB - 1\right). \end{aligned}$$

Ferner gilt $1 + x \leq e^x$ und damit folgt für $x := \frac{\sigma^2}{B^2}\left(e^{tB} - tB - 1\right)$ nun

$$\begin{aligned} P\left(\frac{1}{n}\sum_{i=1}^{n} \xi_i \geq \varepsilon\right) &\leq e^{-t\varepsilon n}\left(1 + \frac{\sigma^2}{B^2}\left(e^{tB} - tB - 1\right)\right)^n \\ &\leq \exp\left(-t\varepsilon n + \frac{\sigma^2 n}{B^2}\left(e^{tB} - tB - 1\right)\right). \end{aligned}$$

Wir setzen nun $h(t) := -t\varepsilon n + \frac{\sigma^2 n}{B^2}\left(e^{tB} - tB - 1\right)$ und minimieren h. Man sieht leicht, dass $h'(t) = -\varepsilon n + \frac{\sigma^2 n}{B^2}\left(Be^{tB} - B\right)$ ist und dann gilt $h'(t) = 0$ genau dann, wenn

$$\frac{\sigma^2 n}{B} e^{tB} = \varepsilon n + \frac{\sigma^2 n}{B}$$

gilt. Lösen wir dies nach t, so erhalten wir $t^* = \frac{1}{B}\ln\left(1 + \frac{B\varepsilon}{\sigma^2}\right)$. Da $\lim_{t\to\pm\infty} h(t) = \infty$ gilt, muss t^* ein Minimum von h sein. Für $y := \frac{\varepsilon B}{\sigma^2}$ ist $t^* = \frac{1}{B}\ln(1+y)$ und

$$\begin{aligned}
h(t^*) &= -t^*\varepsilon n + \frac{\sigma^2 n}{B^2}\left(e^{t^* B} - t^* B - 1\right) \\
&= -\frac{\varepsilon n}{B}\ln(1+y) + \frac{\sigma^2}{B^2}(1 + y - \log(1+y) - 1) \\
&= \frac{\sigma^2 n}{B^2}(-y\ln(1+y) + y - \ln(1+y)) \\
&= -\frac{\sigma^2 n}{B^2}((1+y)\ln(1+y) - y) \\
&\leq -\frac{\sigma^2 n}{B^2}\frac{3}{2}\frac{y^2}{y+3} \\
&= -\frac{3n\varepsilon^2}{2\varepsilon B + 6\sigma^2},
\end{aligned}$$

wobei die Abschätzung mit Lemma 1.4.2 erfolgt. Dann gilt

$$P\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i \geq \varepsilon\right) \leq e^{-\frac{3n\varepsilon^2}{2\varepsilon B + 6\sigma^2}}.$$

Wir setzen nun $\tau := \frac{3n\varepsilon^2}{2\varepsilon B + 6\sigma^2}$ und durch quadratische Ergänzung und Abschätzen erhalten wir

$$\begin{aligned}
\varepsilon &= \sqrt{\frac{2\sigma^2\tau}{n} + \frac{B^2\tau^2}{9n^2}} + \frac{B\tau}{3n} \\
&\leq \sqrt{\frac{2\sigma^2\tau}{n}} + \sqrt{\frac{B^2\tau^2}{9n^2}} + \frac{B\tau}{3n} \\
&= \sqrt{\frac{2\sigma^2\tau}{n}} + \frac{2B\tau}{3n}.
\end{aligned}$$

□

Satz 1.4.4 **Hoeffding-Ungleichung**
Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, $a < b$, $n > 1$ und $\xi_1, \ldots, \xi_n : \Omega \to [a,b]$ seien unabhängige Zufallsvariablen. Für $\tau > 0$ gilt dann

$$P\left(\frac{1}{n}\sum_{i=1}^{n}(\xi_i - \mathbf{E}_P\xi_i) \geq (b-a)\sqrt{\frac{\tau}{2n}}\right) \leq e^{-\tau}.$$

Beweis: Siehe [Devroye96, Satz 8.1, S. 122] oder [Devroye00, Satz 2.1, S. 6]. □

Lemma 1.4.5 **Union Bound**
Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und $f_1, \ldots, f_m : \Omega \to \mathbf{R}$ seien messbare Abbildungen. Für $t \in \mathbf{R}$ gilt dann

$$P\left(\sup_{i=1,\ldots,m} f_i \geq t\right) \leq \sum_{i=1}^{m} P(f_i \geq t).$$

Beweis: Die Aussage folgt wegen $\{\sup_i f_i \geq t\} \subset \bigcup_{i=1}^{m} \{f_i \geq t\}$. □

Für die Ungleichungen von Bernstein und Hoeffding existieren zweiseitige Versionen. Dazu betrachtet man $\xi_1, \ldots, \xi_n, -\xi_1, \ldots, -\xi_n$ und den Union Bound für die Summen im ersten und zweiten Block. Dann folgt

$$\left|\frac{1}{n}\sum_{i=1}^{n} \xi_i\right| = \sup\left\{\frac{1}{n}\sum_{i=1}^{n} \xi_i, \frac{1}{n}\sum_{i=1}^{n} -\xi_i\right\}.$$

Im Falle der Bernstein-Ungleichung folgt

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} \xi_i\right| \geq \sqrt{\frac{2\sigma^2\tau}{n}} + \frac{2B\tau}{3n}\right) \leq 2e^{-\tau},$$

im Falle der Hoeffding-Ungleichung folgt

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} (\xi_i - \mathbf{E}\xi_i)\right| \geq (b-a)\sqrt{\frac{\tau}{2n}}\right) \leq 2e^{-\tau}.$$

Ist $\mathcal{F}$ eine endliche Menge von Funktionen $f : X \to \mathbf{R}$ und $\varepsilon := (b-a)\sqrt{\frac{\tau}{2n}}$, so ergibt die Kombination der Hoeffding-Ungleichung und dem Union Bound die Abschätzung

$$P^n\left((x_1, \ldots, x_n) \in X^n : \sup_{f \in \mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} f(x_i) - \mathbf{E}_p f\right| \geq \varepsilon\right) \leq \sum_{f \in \mathcal{F}} 2e^{-\tau} = e^{\ln(2|\mathcal{F}|)-\tau}.$$

Mit einer Variablentransformation ergibt sich

$$P^n\left((x_1, \ldots, x_n) \in X^n : \sup_{f \in \mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} f(x_i) - \mathbf{E}_P f\right| \geq (b-a)\sqrt{\frac{\ln(2|\mathcal{F}|)+\tau}{2n}}\right) \leq e^{-\tau}.$$

Wir sehen also, dass die Größe der Funktionenklasse $\mathcal{F}$ nur einen logarithmischen Einfluss auf die Abschätzung hat. Dies erlaubt es uns, recht große $\mathcal{F}$ zu betrachten ohne die Abschätzung wesentlich zu verändern. Der Fall $|\mathcal{F}| = \infty$ ist jedoch noch nicht möglich.

Satz 1.4.6 **Bounded Difference Inequality (McDiarmid)**
Sei $g\colon X^n \to \mathbf{R}$ eine messbare Abbildung mit beschränkten Differenzen, das heißt für alle $i \in \{1,\dots,n\}$ existiert ein $c_i > 0$ mit

$$\sup_{\substack{x_1,\dots,x_n \in X \\ x_i' \in X}} \left| g(x_1,\dots,x_n) - g(x_1,\dots,x_{i-1},x_i',x_{i+1},\dots,x_n) \right| \le c_i, \tag{*}$$

dann gilt für alle $\varepsilon > 0$ die Abschätzung

$$P^n\left((x_1,\dots,x_n) \in X^n : g(x_1,\dots,x_n) - \mathbf{E}_{P^n} g \ge \varepsilon\right) \le \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

Beweis: Siehe [Devroye96, Seite 8]. □

Die Ungleichung in (*) besagt, dass das Ändern einer Komponente die Funktion g nicht stark ändert. Aus dem stochastischen Blickwinkel bedeutet dies, dass einzelne Beobachtungen keinen großen Einfluss auf g haben.

Beispiel 1.4.7 *Arithmetisches Mittel*
Wir betrachten die Abbildung $g(x_1,\dots,x_n) := \frac{1}{n}\sum_{i=1}^n x_i$ mit $x_i \in [a,b]$. Dann gilt

$$\begin{aligned}
\left| g(x_1,\dots,x_n) - g(x_1,\dots,x_{i-1},x_i',x_{i+1},\dots,x_n) \right| &= \left| \frac{1}{n}\sum_{j=1}^n x_j - \frac{1}{n}\sum_{j=1}^n x_j - \frac{1}{n}x_i + \frac{1}{n}x_i' \right| \\
&= \frac{1}{n}\left| x_i - x_i' \right| \\
&\le \frac{b-a}{n} =: c_i.
\end{aligned}$$

Damit können wir Satz 1.4.6 anwenden und erhalten für die Schranke auf der rechten Seite

$$\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n \frac{1}{n^2}(b-a)^2}\right) = \exp\left(-\frac{2\varepsilon^2 n}{(b-a)^2}\right).$$

Setzen wir nun $\tau := \frac{2\varepsilon^2 n}{(b-a)^2}$, so erhalten wir $\varepsilon = (b-a)\sqrt{\frac{\tau}{2n}}$ und sehen, dass wir die Hoeffding-Ungleichung erhalten haben. Gewissermaßen ist die Ungleichung von McDiarmid also eine Verallgemeinerung der Hoeffding-Ungleichung. //

Beispiel 1.4.8
Sei $\mathscr{F}$ eine abzählbare Menge von messbaren Funktionen $f : X \to [a,b]$. Für den Datensatz $D := (x_1,\dots,x_n) \in X^n$ schreiben wir $\mathbf{E}_D f = \frac{1}{n}\sum_{i=1}^n f(x_i)$. Dann gilt für $g(D) := g(x_1,\dots,x_n) := \sup_{f\in\mathscr{F}} |\mathbf{E}_D f - \mathbf{E}_P f|$ und $D' := (x_1,\dots,x_{i-1},x_i',x_{i+1},x_n)$ mit Hilfe der umgekehrten Dreiecksungleichung die Abschätzung

$$\begin{aligned}\big|g(x_1,\dots,x_n) - g(x_1,\dots,x_{i-1},x_i',x_{i+1},\dots,x_n)\big| &= \left|\sup_{f\in\mathscr{F}} |\mathbf{E}_D f - \mathbf{E}_P f| - \sup_{f\in\mathscr{F}} |\mathbf{E}_{D'} f - \mathbf{E}_P f|\right| \\ &\le \sup_{f\in\mathscr{F}} |\mathbf{E}_D f - \mathbf{E}_P f - \mathbf{E}_{D'} f + \mathbf{E}_P f| \\ &\le \sup_{f\in\mathscr{F}} \frac{1}{n}\big|f(x_i) - f(x_i')\big| \\ &\le \frac{b-a}{n} =: c_i.\end{aligned}$$

Damit ist Satz 1.4.6 wieder anwendbar und wir erhalten

$$P^n\left(D \in X^n : \sup_{f\in\mathscr{F}} |\mathbf{E}_D f - \mathbf{E}_P f| \ge \mathbf{E}_{P^n} \sup_{f\in\mathscr{F}} |\mathbf{E}_D f - \mathbf{E}_P f| + (b-a)\sqrt{\frac{\tau}{2n}}\right) \le e^{-\tau}.$$

Die Hoeffding-Ungleichung und der Union Bound ergaben

$$\varepsilon = \sqrt{\frac{\log(2|\mathscr{F}|) + \tau}{2n}}(b-a),$$

im Vergleich hierzu haben wir die Größe von $\mathscr{F}$ also an einer anderen Stelle, nämlich in dem zusätzlichen Term $\mathbf{E}_{P^n} \sup_{f\in\mathscr{F}} |\mathbf{E}_D f - \mathbf{E}_P f|$. Solche Terme werden wir später noch abschätzen können.

Es sei gesagt, dass auf die Abzählbarkeit von $\mathscr{F}$ verzichtet werden kann, wenn man einige technische Modifikationen vornimmt. //

Satz 1.4.9 **Talagrands Ungleichung**
Sei P ein Wahrscheinlichkeitsmaß auf X, $\mathscr{F}$ eine abzählbare Menge messbarer Funktionen $f : X \to [-B,B]$ mit

i) $\mathbf{E}_P f = 0$ und

ii) $\mathbf{E}_P f^2 \le \sigma^2$

für alle $f \in \mathscr{F}$. Dann gilt für $g : X^n \to \mathbf{R}$ mit $g(D) := \sup_{f\in\mathscr{F}} |\mathbf{E}_D f - \mathbf{E}_P f|$ die Abschätzung

$$P^n\left(D \in X^n : g(D) \ge \mathbf{E}_{P^n} g + \sqrt{\frac{2\tau\sigma^2}{n} + \frac{2B\mathbf{E}_{P^n} g}{n}} + \frac{2\tau B}{3n}\right) \le e^{-\tau}.$$

Für Satz 1.4.9 gilt zu beachten, dass wir den Wurzelterm abschätzen können durch

$$\sqrt{\frac{2\tau\sigma^2}{n} + \frac{2B\mathbf{E}_{P^n} g}{n}} \leq \sqrt{\frac{2\tau\sigma^2}{n}} + \sqrt{\frac{2B\mathbf{E}_{P^n} g}{n}} \leq \sqrt{\frac{2\tau\sigma^2}{n}} + \gamma\mathbf{E}_{P^n} g + \frac{\tau B}{\gamma n}$$

für alle $\gamma > 0$.

Beweis: Eine frühere Version wurde 1996 von Michel Talagrand bewiesen. Die heutige Version, wie wir den Satz formuliert haben, wurde 2002 von Olivier Bousquet bewiesen. Ein großer Teil der Arbeit wurde auch von Michel Ledoux, Pascal Massart und Emanuel Rio geleistet. Eine vollständige Version des Beweises umfasst etwa 14 Seiten. □

2
Dichteschätzung

In diesem Kapitel widmen wir uns der Schätzung von Wahrscheinlichkeitsdichten, wie wir es im vorangegangenen Kapitel schon angesprochen haben. Wir werden hierbei im Wesentlichen zwei Verfahren kennenlernen.

Im Folgenden sei immer $X \subset \mathbf{R}^d$ eine abgeschlossene Teilmenge, $\mu := \lambda^d$ das Lebesgue-Maß mit $\mu(X) > 0$ und P sei ein absolut stetiges Wahrscheinlichkeitsmaß bezüglich μ mit Dichte h, so dass $P(X) = 1$ gilt. Hierbei sind P, h und gegebenfalls auch X unbekannt.

2.1 Histogrammregel

Wir wollen diese Regel nicht in aller Allgemeinheit vorstellen, sondern als einen einfachen Einstieg in die Vorgehensweise geben. Diese einfache Histogrammregel besitzt jedoch eine große praktische Relevanz.

Für $s > 0$ setzen wir $[0,s)^d := \{x \in \mathbf{R}^d : 0 \le x_i < s \text{ für } i = 1,\dots,d\}$.

Definition 2.1.1 **Würfelpartition**
Eine Partition $\mathscr{A} = (A_j)_{j\ge1}$ von $\mathbf{R}^d$ heißt Würfelpartition der Weite $s > 0$ genau dann, wenn für alle $j \ge 1$ ein $z_j \in \mathbf{R}^d$ existiert, so dass $A_j = z_j + [0,s)^d$ ist. Wir nennen jede Menge A_j eine Zelle.

Definition 2.1.2 **$\mathscr{A}$-Histogramm**
Sei $\mathscr{A} = (A_j)_{j\ge1}$ eine Würfelpartition der Weite $s > 0$. Ferner sei Q ein Wahrscheinlichkeitsmaß auf $\mathbf{R}^d$. Die Funktion

$$h_{Q,\mathscr{A}} : \mathbf{R}^d \to [0,\infty),$$
$$x \mapsto \frac{1}{s^d} \sum_{j=1}^{\infty} Q(A_j) \mathbf{1}_{A_j}(x)$$

heißt dann $\mathscr{A}$-Histogramm von Q.

Lemma 2.1.3
Ist $h_{Q,\mathscr{A}}$ ein $\mathscr{A}$-Histogramm der Weite s, so gelten die folgenden Aussagen:

i) $h_{Q,\mathscr{A}}$ ist die Dichte eines Wahrscheinlichkeitsmaßes, das heißt $h_{Q,\mathscr{A}}$ ist messbar und es gilt $\int h_{Q,\mathscr{A}}\, \mathrm{d}\mu = 1$.

ii) Für $x \in A_j$ gilt

$$h_{Q,\mathscr{A}}(x) = \frac{Q(A_j)}{\mu(A_j)} = \frac{Q(A_j)}{s^d}.$$

Beweis: Die Messbarkeit ist offensichtlich. Dann gilt

$$\begin{aligned}\int_{\mathbf{R}^d} h_{Q,\mathscr{A}}\, \mathrm{d}\mu &= \frac{1}{s^d} \int_{\mathbf{R}^d} \sum_{j=1}^{\infty} Q(A_j)\mathbf{1}_{A_j}\, \mathrm{d}\mu = \frac{1}{s^d} \sum_{j=1}^{\infty} \int Q(A_j)\mathbf{1}_{A_j}\, \mathrm{d}\mu = \frac{1}{s^d} \sum_{j=1}^{\infty} Q(A_j)\mu(A_j) \\ &= \sum_{j=1}^{\infty} Q(A_j) \\ &= 1.\end{aligned}$$

Die zweite Aussage ist klar, denn für $x \in A_j$ gilt $\mathbf{1}_{A_l}(x) = 0$ für alle $l \neq j$. □

Die wichtigsten Beispiele für $\mathscr{A}$-Histogramme sind:

i) Ist $Q = P$, so gilt für $x \in A_j$

$$h_{P,\mathscr{A}}(x) = \frac{P(A_j)}{\mu(A_j)} = \frac{1}{s^d} \int_{A_j} h\, \mathrm{d}\mu.$$

Mit anderen Worten ist $h_{P,\mathscr{A}}$ in A_j also die über A_j gemittelte Dichte.

ii) Ist $Q = D$ für $D \in \mathbf{R}^d$, wobei wir dies als empirisches Maß verstehen, so gilt

$$\mathbf{E}_D \mathbf{1}_A = D(A) = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_A(x_i).$$

Für $x \in A_j$ gilt dann

$$h_{D,\mathscr{A}}(x) = \frac{D(A_j)}{\mu(A_j)} = \frac{1}{ns^d} \sum_{i=1}^{n} \mathbf{1}_{A_j}(x_i).$$

Die Summe in der Gleichung zählt die Treffer von D in A_j. Die Abbildung $D \mapsto h_{D,\mathscr{A}}$ heißt Histogrammregel zur Weite s.

Als nächstes wollen wir uns mit der Frage beschäftigen, wie gut $h_{D,\mathscr{A}}$ die unbekannte Dichte h approximiert. Eine typische Strategie, die uns noch öfters begegnen wird, ist es, diese Frage in zwei kleinere Probleme zu zerlegen:

i) Wie gut approximiert $h_{D,\mathscr{A}}$ die Funktion $h_{P,\mathscr{A}}$? – Hierbei werden wir ausnutzen, dass beide Funktionen die selbe Struktur besitzen.

ii) Wie gut approximiert $h_{P,\mathscr{A}}$ die Dichte h? – Diese Frage ist im Gegensatz zur ersten Frage nicht mehr stochastischer Natur.

Lemma 2.1.4
Ist $\mathscr{A} = (A_j)$ eine Würfelpartition der Weite $s > 0$, so gelten die folgenden Aussagen:

i) $$\int_{\mathbf{R}^d} \left| h_{D,\mathscr{A}} - h_{P,\mathscr{A}} \right| \, \mathrm{d}\mu = \sum_{j=1}^{\infty} \left| \mathbf{E}_D \mathbf{1}_{A_j} - \mathbf{E}_P \mathbf{1}_{A_j} \right|$$

ii) $$\left\| h_{D,\mathscr{A}} - h_{P,\mathscr{A}} \right\|_\infty = \frac{1}{s^d} \sup_{j \geq 1} \left| \mathbf{E}_D \mathbf{1}_{A_j} - \mathbf{E}_P \mathbf{1}_{A_j} \right|$$

Beide Normen $\|\cdot\|_1$ und $\|\cdot\|_\infty$ des Abstandes von $h_{D,\mathscr{A}}$ und $h_{P,\mathscr{A}}$ lassen sich also durch $\left| \mathbf{E}_D \mathbf{1}_{A_j} - \mathbf{E}_P \mathbf{1} A_j \right|$ abschätzen.

Beweis: Für die erste Eigenschaft gilt

$$\begin{aligned}
\int_{\mathbf{R}^d} \left| h_{D,\mathscr{A}} - h_{P,\mathscr{A}} \right| \, \mathrm{d}\mu &= \frac{1}{s^d} \int_{\mathbf{R}^d} \left| \sum_{j=1}^{\infty} (D(A_j) - P(A_j)) \mathbf{1}_{A_j} \right| \, \mathrm{d}\mu \\
&= \frac{1}{s^d} \sum_{j=1}^{\infty} \int_{A_j} \left| D(A_j) - P(A_j) \right| \, \mathrm{d}\mu \\
&= \sum_{j=1}^{\infty} \left| D(A_j) - P(A_j) \right| \\
&= \sum_{j=1}^{\infty} \left| \mathbf{E}_D \mathbf{1}_{A_j} - \mathbf{E}_P \mathbf{1} A_j \right| .
\end{aligned}$$

Für den zweiten Teil betrachten wir mit Lemma 2.1.3

$$\begin{aligned}
\left\| h_{D,\mathscr{A}} - h_{P,\mathscr{A}} \right\|_\infty &= \sup_{j \geq 1} \sup_{x \in A_j} \left| h_{D,\mathscr{A}}(x) - h_{P,\mathscr{A}}(x) \right| \\
&= \sup_{j \geq 1} \sup_{x \in A_j} \left| \frac{D(A_j)}{\mu(A_j)} - \frac{P(A_j)}{\mu(A_j)} \right| \\
&= \frac{1}{s^d} \sup_{j \geq 1} \left| D(A_j) - P(A_j) \right| .
\end{aligned}$$

□

Satz 2.1.5 **Orakelungleichung für Histogrammregel**
Sei $\mathscr{A} = (A_j)$ eine Würfelpartition der Weite $s \in (0,1]$. Ferner sei $r \geq 2$ und $B_{l_\infty^d} := [-1,1]^d$. Wir setzen $c_d := (3d+2)2^{2d+2}$, dann gilt für $n \geq 1$ und $\tau > 0$ die Abschätzung

$$P^n\left(D \in X^n : \left\|h_{D,\mathscr{A}} - h\right\|_{L_1(\mu)} \leq \sqrt{c_d}\sqrt{\frac{r^d\left(\log\left(\frac{r}{s}\right)+\tau\right)}{s^d n}} + c_d\frac{r^d\left(\log\left(\frac{r}{s}\right)+\tau\right)}{s^d n} + \ldots \right.$$
$$\left. \ldots + 2P\left(\mathbf{R}^d \setminus rB_{l_\infty^d}\right) + \left\|h_{P,\mathscr{A}} - h\right\|_{L_1(\mu)}\right) \qquad (*)$$
$$\geq 1 - e^{-\tau}.$$

Die Abschätzung in Satz 2.1.5 beschreibt den stochastischen Fehler (*), der entsteht, wenn wir h mit $h_{D,\mathscr{A}}$ statt $h_{P,\mathscr{A}}$ approximieren. Das Histogramm $h_{P,\mathscr{A}}$ kennt P, statt es mit D schätzen zu müssen, wie wir es tun. Es kann also gewissermaßen als Lösung eines allwissenden Orakels angesehen werden, daher rührt der Name.

Beweis: Sei $J := \left\{j \in \mathbf{N} : A_j \cap rB_{l_\infty^d} \neq \emptyset\right\}$. Da $\mathscr{A}$ eine Partition ist, gilt $rB_{l_\infty^d} \subset \bigcup_{j\in J} A_j$.

Ferner gilt $\bigcup_{j\in J} A_j \subset 2rB_{l_\infty^d}$, denn für $x \in A_j$ mit $j \in J$ gilt, dass es ein $x' \in A_j \cap rB_{l_\infty^d}$ gibt. Dann folgt

$$\|x\|_\infty \leq \left\|x - x'\right\|_\infty + \left\|x'\right\|_\infty \leq s + r \leq 2r,$$

also gilt $x \in 2rB_{l_\infty^d}$.

Weiter gilt $|J| \leq 4^d\left(\frac{r}{s}\right)^d$, insbesondere ist J also endlich. Dazu benutzen wir ein Volumenvergleichselement. Mit der eben gezeigten Eigenschaft gilt

$$|J|\,s^d = \sum_{j\in J}\mu(A_j) = \mu\left(\bigcup_{j\in J} A_j\right)$$
$$\leq \mu\left(2rB_{l_\infty^d}\right) = 4^d r.$$

Wir wollen nun die Bernsteinungleichung für $|\mathbf{E}_D\mathbf{1}_A - \mathbf{E}_P\mathbf{1}_A|$ anwenden. Hierzu betrachten wir für eine messbare Menge $A \subset \mathbf{R}^d$ die Abbildung $\xi_i := \mathbf{1}_A\pi_i - \mathbf{E}_P\mathbf{1}_A$, wobei π_i die i-te Projektion ist. Wir müssen nun die Voraussetzungen für die Anwendung der Bernsteinungleichung verifizieren. Es gilt:

i) $\mathbf{E}_{P^n}\xi_i = 0$.

ii) $\|\xi_i\|_\infty \leq 1$.

iii) $\mathbf{E}_{P^n}\xi_i^2 \leq \mathbf{E}_P\mathbf{1}_A^2 = \mathbf{E}_P\mathbf{1}_A = P(A)$.

iv) (ξ_i) sind unabhängig bezüglich P^n.

Durch Anwendung der Bernsteinungleichung folgt daher nun

$$P^n\left(D \in X^n : |\mathbf{E}_D \mathbf{1}_A - \mathbf{E}_P \mathbf{1}_A| \geq \sqrt{\frac{2P(A)\tau}{n}} + \frac{2\tau}{3n}\right) \leq 2e^{-\tau}.$$

Wir setzen nun $A_0 := \bigcup_{j \notin J} A_j$ und $\sqrt{\frac{2P(A)\tau}{n}} + \frac{2\tau}{3n} =: \varepsilon_A$ für den Term in der Abschätzung. Mit dem Union Bound erhalten wir nun

$$\begin{aligned} P^n\Big(D \in X^n : \Big|\mathbf{E}_D \mathbf{1}_{A_j} - \mathbf{E}_P \mathbf{1}_{A_j}\Big| \geq \varepsilon_{A_j} \text{ für ein } j \in J \cup \{0\}\Big) \\ &\leq \sum_{j \in J \cup \{0\}} P^n\left(D \in X^n : \Big|\mathbf{E}_D \mathbf{1}_{A_j} - \mathbf{E}_P \mathbf{1} A_j\Big| \geq \varepsilon_{A_j}\right) \\ &\leq \sum_{j \in J \cup \{0\}} 2e^{-\tau} \\ &\leq 2(|J|+1)e^{-\tau}. \end{aligned}$$

Damit erhalten wir durch Komplementbildung und einer Variablentransformation

$$\begin{aligned} P^n\Big(D \in X^n : \Big|\mathbf{E}_D \mathbf{1}_{A_j} - \mathbf{E}_P \mathbf{1}_{A_j}\Big| < \sqrt{\frac{2P(A_j)(\log(2(|J|+1)+\tau)}{n}} + \dots \\ \dots + \frac{2\log(2|J|+2)+2\tau}{3n} \text{ für alle } j \in J \cup \{0\}\Big) \\ \geq 1 - e^{-\tau}. \end{aligned}$$

Sei nun $D \in X^n$ ein Datensatz, der dieser Abschätzung in $P^n(\dots)$ genügt. Dann gilt mit Lemma 2.1.4 und $L := 2\log(2|J|+2)+2\tau$ die Abschätzung

$$\begin{aligned} \left\|h_{D,\mathscr{A}} - h_{P,\mathscr{A}}\right\|_{L_1(\mu)} &= \sum_{j \in J} \Big|\mathbf{E}_D \mathbf{1}_{A_j} - \mathbf{E}_P \mathbf{1}_{A_j}\Big| + \sum_{j \notin J} \Big|\mathbf{E}_D \mathbf{1}_{A_j} - \mathbf{E}_P \mathbf{1}_{A_j}\Big| \\ &\leq \sum_{j \in J} \sqrt{\frac{2P(A_j)L}{n}} + |J|\frac{L}{3n} + \sum_{j \notin J} \mathbf{E}_D \mathbf{1}_{A_j} + \sum_{j \notin J} \mathbf{E}_P \mathbf{1}_{A_j} \\ &= \sqrt{\frac{L}{n}} \sum_{j \in J} \sqrt{P(A_j)} + \frac{1}{3} \cdot \frac{L}{n} |J| + \mathbf{E}_D \mathbf{1}_{A_0} + \mathbf{E}_P \mathbf{1}_{A_0} \\ &\leq \sqrt{\frac{L}{n}} \sum_{j \in J} \sqrt{P(A_j)} + \frac{1}{3} \cdot \frac{L}{n} |J| + \mathbf{E}_P \mathbf{1}_{A_0} + \sqrt{\frac{L}{n}} \sqrt{P(A_0)} + \frac{1}{3} \cdot \frac{L}{n} + \mathbf{E}_P \mathbf{1}_{A_0} \\ &= \sqrt{\frac{L}{n}} \sum_{j \in J \cup \{0\}} \sqrt{P(A_j)} + \frac{1}{3} \cdot \frac{L}{n}(|J|+1) + 2\mathbf{E}_P \mathbf{1}_{A_0}. \end{aligned}$$

Wir führen nun eine kurze Nebenrechnung aus. Mit der Hölder-Ungleichung gilt

$$\begin{aligned} \|f\|_{\frac{1}{2}} := \left(\int |f|^{\frac{1}{2}} \, \mathrm{d}\nu\right)^2 &= \left(\int |\mathbf{1}|^{\frac{1}{2}} |f|^{\frac{1}{2}} \, \mathrm{d}\nu\right)^2 \\ &\leq \int |\mathbf{1}| \, \mathrm{d}\nu \int |f| \, \mathrm{d}\nu \\ &= \nu(\Omega)\|f\|_1. \end{aligned}$$

Ist ν das Zählmaß auf $J \cup \{0\}$, so können wir die obige Abschätzung fortführen durch

$$\begin{aligned}\left\|h_{D,\mathscr{A}} - h_{P,\mathscr{A}}\right\|_{L_1(\mu)} &\leq \dots \\ &\leq \sqrt{\frac{L}{n}}\sqrt{|J|+1}\sum_{j\in J\cup\{0\}} P(A_j) + \frac{1}{3}\cdot\frac{L}{n}(|J|+1) + 2P(A_0).\end{aligned}$$

Als nächstes müssen wir den Term $\frac{L}{n}(|J|+1)$ abschätzen. Da $|J| \geq 1$ gilt, beobachten wir $|J|+1 \leq 2|J|$. Damit folgt

$$\begin{aligned}L|J| &\leq (2\log(4|J|) + 2\tau)|J| \\ &\leq 2\left(\log\left(2^{2+2d}\left(\frac{r}{s}\right)^d\right) + \tau\right)4^d\left(\frac{r}{s}\right)^d \\ &\leq 2\left(\log\left(\frac{r}{s}\right)^{2+3d} + \tau\right)4^d\left(\frac{s}{s}\right)^d \\ &= 2^{2d+1}(2+3d)\left(\log\left(\frac{r}{s}\right) + \tau\right)\left(\frac{r}{s}\right)^d\end{aligned}$$

und mit $c_d := (3d+2)2^{2d+2}$ folgt weiter

$$\leq \frac{1}{2}c_d\left(\log\left(\frac{r}{s}\right) + \tau\right)\left(\frac{r}{s}\right)^d.$$

Insgesamt erhalten wir damit also

$$\begin{aligned}\left\|h_{D,\mathscr{A}} - h_{P,\mathscr{A}}\right\|_{L_1(\mu)} &\leq \sqrt{\frac{2L|J|}{n}} + \frac{2L|J|}{n} + 2P(A_0) \\ &\leq \sqrt{c_d}\sqrt{\frac{r^d\left(\log\left(\frac{r}{s}\right)+\tau\right)}{s^d n}} + c_d\frac{r^d\left(\log\left(\frac{r}{s}\right)+\tau\right)}{s^d n} + 2P(A_0).\end{aligned}$$

Ferner gilt

$$A_0 = \bigcup_{j\notin J} A_j = \bigcup_{j\colon A_j\cap rB_{l_\infty^d}\neq\emptyset} A_j \subset \mathbf{R}^d \setminus rB_{l_\infty^d}.$$

Die Behauptung folgt dann durch Kombination der Abschätzungen und

$$\left\|h_{D,\mathscr{A}} - h\right\|_{L_1} \leq \left\|h_{D,\mathscr{A}} - h_{P,\mathscr{A}}\right\|_{L_1} + \left\|h_{P,\mathscr{A}} - h\right\|_{L_1}.$$ □

Satz 2.1.6
Sei ν ein σ-endliches Maß auf $(\mathbf{R}^d, \mathscr{B}^d)$ und $p \in [1,\infty)$. Dann liegt der Raum

$$C_c(\mathbf{R}^d) := \left\{ f : \mathbf{R}^d \to \mathbf{R} \text{ stetig mit kompaktem Träger} \right\},$$

wobei $\operatorname{supp} f := \overline{\{f \neq 0\}}$ der Träger ist, dicht in $\mathscr{L}_p(\nu)$, das heißt für alle $f \in \mathscr{L}_P(\nu)$ und alle $\varepsilon > 0$ existiert ein $g \in C_c(\mathbf{R}^d)$ mit $\|f - g\|_{\mathscr{L}_p(\nu)} \leq \varepsilon$.

Beweis: Der Beweis wird hier nicht geführt. □

Aus technischen Gründen betrachten wir im Folgenden verallgemeinerte Histogrammregeln. Hierzu seien Q und ν zwei σ-endliche Maße mit $Q \ll \nu$. Wir setzen dann

$$h_{Q,\nu,\mathscr{A}}(x) := \sum_{j=1}^{\infty} \frac{Q(A_j)}{\nu(A_j)} \mathbf{1}_{A_j}$$

und vereinbaren $\frac{0}{0} := 0$. Für $Q := P$ und $\nu := \mu$ erhalten wir damit wieder $h_{P,\mathscr{A}} = h_{Q,\nu,\mathscr{A}}$. Auf A_j mit $\nu(A_j) > 0$ gilt

$$h_{Q,\nu,\mathscr{A}}(x) = \frac{Q(A_j)}{\nu(A_j)} = \frac{1}{\nu(A_j)} \int_{A_j} h \, \mathrm{d}\nu,$$

wobei h eine Dichte von Q bezüglich ν ist.

Satz 2.1.7 **Approximationsfehler**
Seien Q, ν und h wie oben und $\nu \neq 0$. Für jedes $\varepsilon > 0$ gibt es dann ein $s_\varepsilon > 0$, so dass für alle $s \in (0, s_\varepsilon]$ und alle Würfelpartitionen der Weite s die folgende Abschätzung gilt:

$$\left\| h_{Q,\nu,\mathscr{A}} - h \right\|_{L_1(\nu)} \leq \varepsilon.$$

Ist h ferner α-Hölder-stetig, das heißt für alle $x, x' \in \mathbf{R}^d$ gilt

$$\left| h(x) - h(x') \right| \leq c \left\| x - x' \right\|_{l_\infty^d}^{\alpha},$$

so gilt für alle $s \leq \left(\frac{\varepsilon}{c}\right)^{\frac{1}{\alpha}}$ sogar

$$\left\| h_{P,\mathscr{A}} - h \right\|_{\infty} \leq \varepsilon.$$

Beweis: Sei $\varepsilon > 0$ und wir betrachten zunächst die erste Behauptung. Mit Satz 2.1.6 existiert ein $f \in C_c(\mathbf{R}^d)$ mit

$$\|h - f\|_{L_1(\mu)} \leq \frac{\varepsilon}{3}. \qquad (*)$$

Da f einen kompakten Träger besitzt, gibt es ein $r \geq 1$ mit den Eigenschaften $\operatorname{supp} f \subset rB_{l_\infty^d}$ und $\nu\left(2rB_{l_\infty^d}\right) > 0$. Ferner ist f stetig und $\operatorname{supp} f$ kompakt, also ist f gleichmäßig stetig. Das heißt, es existiert $\delta \in (0,1]$ mit

$$\left\|x - x'\right\|_{l_\infty^d} \leq \delta \Rightarrow \left|f(x) - f(x')\right| \leq \frac{1}{3} \frac{\varepsilon}{\nu\left(2rB_{l_\infty^d}\right)}.$$

Beachte, dass $\delta \leq 1 \leq r$ gilt. Setzen wir nun $s_\varepsilon := \delta$ und sei $\mathscr{A} = (A_j)$ eine Würfelpartition der Weite $s \in (0, s_\varepsilon]$. Für $x \in \mathbf{R}^d$ existiert dann genau ein $j \geq 1$ mit $x \in A_j$. Wir setzen dann $A(x) := A_j$. Es gilt nun für $x \in \mathbf{R}^d$ mit $\nu(A(x)) > 0$

$$h_{Q,\nu,\mathscr{A}}(x) = \frac{Q(A(x))}{\nu(A(x))} = \frac{1}{\nu(A(x))} \int_{A(x)} h(x') \, \mathrm{d}\nu(x').$$

Wir setzen $\overline{f} \colon \mathbf{R}^d \to \mathbf{R}$ mit

$$\overline{f}(x) := \frac{1}{\nu(A(x))} \int_{A(x)} f(x') \, \mathrm{d}\nu(x'),$$

falls $\nu(A(x)) > 0$ ist und $\overline{f}(x) := 0$ sonst. Wir zeigen zunächst

$$x \in \mathbf{R}^d \setminus 2rB_{l_\infty^d} \Rightarrow A(x) \subset \mathbf{R}^d \setminus rB_{l_\infty^d}. \tag{**}$$

Dazu sei $x' \in A(x)$, dann gilt

$$2r < \|x\|_\infty \leq \left\|x - x'\right\|_\infty + \left\|x'\right\|_\infty \leq s + \left\|x'\right\|_\infty \leq r + \left\|x'\right\|_\infty,$$

also $\left\|x'\right\|_\infty > r$ und damit die Behauptung (**). Es gilt nun

$$\left\|h - h_{Q,\nu,\mathscr{A}}\right\|_{L_1(\nu)} \leq \underbrace{\|h - f\|_{L_1(\nu)}}_{=:\mathrm{I}} + \underbrace{\left\|f - \overline{f}\right\|_{L_1(\nu)}}_{=:\mathrm{II}} + \underbrace{\left\|\overline{f} - h_{Q,\nu,\mathscr{A}}\right\|_{L_1(\nu)}}_{=:\mathrm{III}}.$$

Dabei ist I klar nach (*). Für II sei $x \notin 2rB_{l_\infty^d}$, dann folgt $f(x) = 0$ und falls $\nu(A(x)) > 0$ ist, so gilt auch

$$\overline{f}(x) = \frac{1}{\nu(A(x))} \int_{A(x)} f(x') \, \mathrm{d}\nu(x') \overset{(**)}{=} 0.$$

Für $\nu(A(x)) = 0$ folgt $\overline{f}(x) = 0$ per Definition. Damit gilt

$$\mathrm{II} = \int_{2rB_{l_\infty^d}} \left|f(x) - \overline{f}(x)\right| \, \mathrm{d}\nu(x).$$

Für $x \in 2rB_{l_\infty^d}$ mit $\nu(A(x)) > 0$ folgt

$$\begin{aligned}
\left|f(x) - \overline{f}(x)\right| &= \left|\frac{1}{\nu(A(x))} \int_{A(x)} f(x) \, \mathrm{d}\nu(x') - \frac{1}{\nu(A(x))} \int_{A(x)} f(x') \, \mathrm{d}\nu(x)\right| \\
&\leq \frac{1}{\nu(A(x))} \int_{A(x)} \left|f(x) - f(x')\right| \, \mathrm{d}\nu(x') \\
&\leq \frac{\varepsilon}{3} \frac{1}{\nu(2rB_{l_\infty^d})},
\end{aligned}$$

da $A(x)$ die Weite $s \le s_\varepsilon = \delta$ hat. Damit folgt für A_j mit $\nu(A_j \cap 2rB_{l_\infty^d}) > 0$

$$\int_{A_j \cap 2rB_{l_\infty^d}} \left| f(x) - \overline{f}(x) \right| \, \mathrm{d}\nu(x) \le \frac{\varepsilon}{3} \frac{\nu(A_j \cap 2rB_{l_\infty^d})}{\nu(2rB_{l_\infty^d})}.$$

Für A_j mit $\nu(A_j \cap 2rB_{l_\infty^d}) = 0$ gilt die Abschätzung trivialerweise. Damit gilt

$$\begin{aligned} \mathrm{II} &= \left\| f - \overline{f} \right\|_{L_1(\nu)} = \sum_{j=1}^{\infty} \int_{A_j \cap 2rB_{l_\infty^d}} \left| f - \overline{f} \right| \, \mathrm{d}\nu \le \sum_{j=1}^{\infty} \frac{\varepsilon}{3} \frac{\nu(A_j \cap 2rB_{l_\infty^d})}{\nu(2rB_{l_\infty^d})} \\ &\le \frac{\varepsilon}{3}. \end{aligned}$$

Kommen wir nun zu III. Es gilt

$$\begin{aligned} \left\| \overline{f} - h_{Q,\nu,\mathscr{A}} \right\|_{L_1} &= \sum_{\nu(A_j)>0} \int_{A_j} \left| \overline{f} - h_{Q,\nu,\mathscr{A}} \right| \, \mathrm{d}\nu \\ &= \sum_{\nu(A_j)>0} \int_{A_j} \left| \frac{1}{\nu(A_j)} \int_{A_j} f \, \mathrm{d}\nu - \frac{1}{\nu(A_j)} \int_{A_j} h \, \mathrm{d}\nu \right| \, \mathrm{d}\nu \\ &= \sum_{\nu(A_j)>0} \left| \int_{A_j} f \, \mathrm{d}\nu - \int_{A_j} h \, \mathrm{d}\nu \right| \\ &\le \sum_{\nu(A_j)>0} \int_{A_j} |f - h| \, \mathrm{d}\nu \\ &\le \int_{\mathbf{R}^d} |f - h| \, \mathrm{d}\nu \\ &\le \frac{\varepsilon}{3}. \end{aligned}$$

Damit folgt die erste Behauptung. Für die zweite Behauptung sei $s \le \left(\frac{\varepsilon}{c}\right)^{\frac{1}{\alpha}}$ und $x, x' \in A_j$. Dann gilt

$$\left| h(x) - h(x') \right| \le c \left\| x - x' \right\|_{l_\infty^d}^{\alpha} \le c s^{\alpha} \le \varepsilon.$$

Für $x \in \mathbf{R}^d$ folgt nun

$$\begin{aligned} \left| h(x) - h_{P,\mathscr{A}}(x) \right| &= \left| \frac{1}{s^d} \int_{A(x)} h(x) \, \mathrm{d}\mu(x') - \frac{1}{s^d} \int_{A(x)} h(x') \, \mathrm{d}\mu(x') \right| \\ &\le \frac{1}{s^d} \int_{A(x)} \underbrace{\left| h(x) - h(x') \right|}_{\le \varepsilon} \, \mathrm{d}\mu(x') \\ &\le \varepsilon. \end{aligned}$$

□

Satz 2.1.8 **Konsistenz der Histogrammregel**
Für $n \geq 1$ sei $\mathscr{A}_n$ eine Würfelpartition der Weite s_n. Es gelten die folgenden Eigenschaften:

i) $s_n \to 0$ – dies sorgt dafür, dass der Approxiationsfehler verschwindet.

ii) $\frac{ns_n}{\log n} \to \infty$ – die s_n sollen nicht zu schnell gegen 0 konvergieren, so dass wir noch genügend viele Daten pro Zelle haben.

Dann ist das Verfahren $X^n \to \mathscr{L}_1(\nu)$ mit $D \mapsto h_{D,\mathscr{A}_n}$ universell konsistent, das heißt $\|h_{D,\mathscr{A}_n} - h\|_{L_1(\mu)} \to 0$ in Wahrscheinlichkeit P^∞ für $n \to \infty$, das heißt für alle $\varepsilon, \delta > 0$ existiert ein $n_0 \geq 1$, so dass für alle $n \geq n_0$ gilt:

$$P^n\left(D \in X^n : \|h_{D,\mathscr{A}_n} - h\|_{L_1(\mu)} \leq \varepsilon\right) \geq 1 - \delta.$$

Die nötigen Voraussetzungen sind zum Beispiel erfüllt für $s_n := n^{-\gamma}$ mit $\gamma \in \left(0, \frac{1}{d}\right)$.

Beweis: Ohne Einschränkung sei $s_n \leq 1$. Wir setzen $r_n := \left(\frac{s_n^d n}{\log n}\right)^{\frac{1}{2d}} \to \infty$ und wir können ebenfalls ohne Einschränkung $r_n \geq 2$ annehmen. Damit wird Satz 2.1.5 anwendbar. Wegen $s_n \to 0$ gilt mit Satz 2.1.7 ferner $\|h_{P,\mathscr{A}_n} - h\|_{L_1(\mu)} \leq \varepsilon$ und wegen $r_n \to \infty$ gilt $2P\left(\mathbf{R}^d \setminus r_n B_{l_d^\infty}\right) \leq \varepsilon$ für alle $n \geq n_0$. Für $\tau := -\log \delta$ und wegen $r_n \to \infty$ und $s_n \to 0$ gilt $\log \frac{r_n}{s_n} \geq \tau$ für $n \geq n_0$. Wir betrachten nun

$$\alpha_n := \frac{r_n^d \log\left(\frac{r_n}{s_n}\right)}{s_n^d n} = \frac{\left(s_n^d n\right)^{\frac{1}{2}}}{s_n^d n} \cdot \frac{\log\left(\frac{r_n}{s_n}\right)}{(\log n)^{\frac{1}{2}}} = \left(s_n^d n\right)^{-\frac{1}{2}} \frac{\log \frac{r_n}{s_n}}{(\log n)^{\frac{1}{2}}}.$$

Außerdem gilt

$$\begin{aligned}\log \frac{r_n}{s_n} &= \log \frac{1}{s_n} \left(\frac{s_n^d n}{\log n}\right)^{\frac{1}{2d}} = \log s_n^{-\frac{1}{2}} n^{\frac{1}{2d}} - \frac{1}{2d} \log \log n \\ &\leq \log\left(s_n^d n\right)^{\frac{1}{2d}}.\end{aligned}$$

Ferner gilt ohne Einschränkung $s_n^d n \geq 1$ und damit $s_n^{-d} \leq n$. Damit folgt weiter

$$\ldots \leq \log n^{\frac{1}{d}} = \frac{1}{d} \log n.$$

Insgesamt folgt daher für $n \geq n_0$

$$\alpha_n \leq \frac{1}{d}\left(s_n^d n\right)^{-\frac{1}{2}} (\log n)^{\frac{1}{2}} = \frac{1}{d}\left(\frac{\log n}{s_n^d n}\right)^{\frac{1}{2}} \longrightarrow 0.$$

Wir wenden Satz 2.1.5 nun an und erhalten mit Wahrscheinlichkeit P^n von $\geq 1 - \delta$

$$\|h_{D,\mathscr{A}_n} - h\|_{L_1(\mu)} \leq \sqrt{c_d}\sqrt{2\varepsilon} + 2c_d\varepsilon + \varepsilon + \varepsilon. \qquad \square$$

Eine gute Eigenschaft des Satzes ist es, dass es von h und P unabhängig ist. Dies entspricht dem Sinne der nicht-parametrischen Statistik. Auf der anderen Seite haben wir keine Vorstellung von der Konvergenzrate.

Um eine Konvergenzrate zu bekommen, müssen wir über die Informationen über die Konvergenzgeschwindigkeiten der folgenden Terme haben:

- $P\left(\mathbf{R}^d \setminus rB_{l_\infty^d}\right) \to 0$ für $r \to \infty$
- $\left\|h_{P,\mathscr{A}_n} - h\right\|_{L_1(\mu)} \to 0$ für $s \to 0$

Beide Terme werden wir in Form von Annahmen verarbeiten. Es zeigt sich, dass man ohne Annahmen keine Konvergenzgeschwindigkeit bestimmen kann.

Lemma 2.1.9 **Vergleich von Normen für Histogrammregeln**
Für eine Würfelpartition $\mathscr{A}$ der Weite $s \le 1$ und $r \ge 1$ gilt

$$\left\|h_{P,\mathscr{A}} - h\right\|_{L_1(\mu)} \le 4^d r^d \left\|h_{P,\mathscr{A}} - h\right\|_\infty + 2P\left(\mathbf{R}^d \setminus rB_{l_\infty^d}\right).$$

Beweis: Aus dem Beweis von Satz 2.1.7 wissen wir für $x \notin 2rB_{l_\infty^d}$, dass $A(x) \subset \mathbf{R}^d \setminus rB_{l_\infty^d}$ gilt. Damit folgt

$$\left\|h_{P,\mathscr{A}} - h\right\|_{L_1(\mu)} = \underbrace{\int_{2rB_{l_\infty^d}} \left|h_{P,\mathscr{A}} - h\right| \,\mathrm{d}\mu}_{=:\mathrm{I}} + \underbrace{\int_{\mathbf{R}^d \setminus 2rB_{l_\infty^d}} \left|h_{P,\mathscr{A}} - h\right| \,\mathrm{d}\mu}_{=:\mathrm{II}}.$$

Dann gilt

$$\mathrm{I} \le \mu\left(2rB_{l_\infty^d}\right) \left\|h_{P,\mathscr{A}} - h\right\|_\infty \le 4^d r^d \left\|h_{P,\mathscr{A}} - h\right\|_\infty.$$

Für den zweiten Teil gilt

$$\begin{aligned}\mathrm{II} &\le \int_{R^d \setminus 2rB_{l_\infty^d}} h_{P,\mathscr{A}} \,\mathrm{d}\mu + \underbrace{\int_{\mathbf{R}^d \setminus 2rB_{l_\infty^d}} h \,\mathrm{d}\mu}_{=P\left(\mathbf{R}^d \setminus 2rB_{l_\infty^d}\right)} \\ &\le \int_{R^d \setminus 2rB_{l_\infty^d}} h_{P,\mathscr{A}} \,\mathrm{d}\mu + P\left(\mathbf{R}^d \setminus rB_{l_\infty^d}\right)\end{aligned}$$

und

$$\begin{aligned}\int_{\mathbf{R}^d \setminus 2rB_{l_\infty^d}} h_{P,\mathscr{A}} \, \mathrm{d}\mu &= \int_{\mathbf{R}^d \setminus 2rB_{l_\infty^d}} \sum_{j=1}^{\infty} \frac{P(A_j)}{s^d} \mathbf{1}_{A_j}(x) \, \mathrm{d}\mu(x) \\ &= \sum_{j=1}^{\infty} \frac{P(A_j)}{s^d} \int_{\mathbf{R}^d \setminus 2rB_{l_\infty^d}} \mathbf{1}_{A_j}(x) \, \mathrm{d}\mu(x) \\ &\leq \sum_{A_j \cap \left(\mathbf{R}^d \setminus 2rB_{l_\infty^d}\right) \neq \emptyset} \\ &= \sum_{A_j \in \mathbf{R}^d \setminus rB_{l_\infty^d}} P(A_j) \\ &= P\left(\mathbf{R}^d \setminus rB_{l_\infty^d}\right). \end{aligned}$$

□

Für $(a_n),(b_n) \subset (0,\infty)$ schreiben wir $a_n \overline{\leq} b_n$ genau dann, wenn ein $c > 0$ existiert, so dass $a_n \leq c b_n$ für alle $n \geq 0$ gilt. Ferner schreiben wir $a_n \sim b_n$ genau dann, wenn $a_n \overline{\leq} b_n$ und $b_n \overline{\leq} a_n$ gilt.

Satz 2.1.10 **Konvergenzraten für Histogrammregeln**
Die Dichte h sei α-Hölder-stetig. Wir betrachten die folgenden drei Fälle:

i) $P\left(\mathbf{R}^d \setminus rB_{l_\infty^d}\right) \leq c r^{-\gamma d}$ für alle $r \geq 1$

ii) $P\left(\mathbf{R}^d \setminus rB_{l_\infty^d}\right) \leq c e^{-a r^\gamma}$ für alle $r \geq 1$

iii) $P\left(\mathbf{R}^d \setminus rB_{l_\infty^d}\right) = 0$ für alle $r \geq r_0$

Hierbei sind a, c und γ Konstanten. Für diese Fälle betrachten wir die Folge s_n der entsprechend zugehörigen Form:

i) $s_n \sim \left(\frac{\log n}{n}\right)^{\frac{1+\gamma}{(1+\gamma)(2\alpha+d)-\alpha}}$

ii) $s_n \sim \left(\frac{\log n}{n}\right)^{\frac{1}{2\alpha+d}} (\log n)^{-\frac{d}{\gamma}\frac{1}{2\alpha+d}}$

iii) $s_n \sim \left(\frac{\log n}{n}\right)^{\frac{1}{2\alpha+d}}$

Dann gelten in der Abschätzung $P^n\left(D \in X^n : \left\|h_{D,\mathscr{A}_n} - h\right\|_{L_1(\mu)} \leq \varepsilon_n\right) \geq 1 - \frac{1}{n}$ für $n \geq 1$ mit einer Würfelpartition $\mathscr{A}$ der Weite s_n die folgende Konvergenzrate:

i) $\varepsilon_n \sim \left(\frac{\log n}{n}\right)^{\frac{\alpha\gamma}{(1+\gamma)(2\alpha+d)}}$

ii) $\varepsilon_n \sim \left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+d}} (\log n)^{\frac{d}{\gamma}\frac{\alpha+d}{2\alpha+d}}$

iii) $\varepsilon_n \sim \left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+d}}$

Beweis: Der Beweis wird die Behauptung eigentlich nur für $n \geq n_0$ zeigen, aber es gilt

$$\left\|h_{D,\mathscr{A}_n} - h\right\|_{L_1(\mu)} \leq \left\|h_{D,\mathbf{A}_n}\right\|_{L_1(\mu)} + \|h\|_{L_1(\mu)} \leq 2,$$

durch Vergrößern der Konstanten lässt es sich daher für alle n zeigen. Nach Satz 2.1.5, Satz 2.1.7 und Lemma 2.1.9 gilt mit Wahrscheinlichkeit $P^n \geq 1 - e^{-\tau}$

$$\left\|h_{D,\mathscr{A}_n} - h\right\|_{L_1(\mu)} \leq \sqrt{c_d}\sqrt{\frac{r^d\left(\log\left(\frac{r}{s}\right)+\tau\right)}{s^d n}} + c_d\frac{r^d\left(\log\left(\frac{r}{s}\right)+\tau\right)}{s^d n} + \ldots$$
$$\ldots + 4P\left(\mathbf{R}^d \setminus rB_{l_\infty^d}\right) + 4^d r^d \left\|h_{P,\mathscr{A}} - h\right\|_{L_1(\mu)}.$$

Wir betrachten $\tau := \tau_n := \log n$ und müssen nun mit Hilfe unserer Annahmen an das Tailverhalten von P die Folgen (r_n) und (s_n) möglichst optimal bestimmen.

Wir betrachten zunächst den ersten Fall, also $P\left(\mathbf{R}^d \setminus rB_{l_\infty^d}\right) \leq cr^{-\gamma d}$. Wenn die rechte Seite gegen 0 konvergiert, so konvergiert der erste Term langsamer als der zweite. Daraus folgt, dass wir den zweiten Term ignorieren können. Dies ergibt ohne Konstanten

$$\sqrt{\frac{r^d\left(\log\left(\frac{r_n}{s_n}\right)+\log n\right)}{s_n^d n}} + r_n^{-\gamma d} + r_n^d s_n^\alpha. \tag{*}$$

Wir lassen bei unseren Betrachtungen $\log\frac{r_n}{s_n}$ weg. Wir werden später sehen, dass sich dies rechtfertigt. Zunächst wollen wir s_n durch Betrachten des ersten und dritten Terms bestimmen. Eigentlich müssten wir nun nach s ableiten, doch dies würde zu vielen weiteren Konstanten führen. Da wir lediglich an der Asymptotik interessiert sind, setzen wir den ersten und dritten Term gleich und lösen nach s_n auf. Man kann einsehen, dass dies das asymptotische Verhalten nicht ändert. Dies führt uns auf

$$\frac{r_n^{\frac{d}{2}}(\log n)}{\left(s_n^d n\right)^{\frac{1}{2}}} = r_n^d s_n^\alpha,$$

was sich zu $s_n = \left(\frac{\log n}{n r_n^d}\right)^{\frac{1}{2\alpha+d}}$ lösen lässt. Dies setzen wir in den ersten und dritten Term ein, wodurch sich beide zu

$$r_n^d\left(\frac{\log n}{n r_n^d}\right)^{\frac{\alpha}{2\alpha+d}}$$

ergeben. Dies ergibt als rechte Seite

$$r_n^d\left(\frac{\log n}{n r_n^d}\right)^{\frac{\alpha}{2\alpha+d}} + r_n^{-\gamma d}.$$

Zur Bestimmung von r_n setzen wir analog

$$r_n^{-\gamma d} = r_n^d \left(\frac{\log n}{n r_n^d} \right)^{\frac{\alpha}{2\alpha+d}}.$$

Lösen wir auch dies, so erhalten wir

$$r_n = \left(\frac{n}{\log n} \right)^{\frac{\alpha}{d(1+\gamma)(2\alpha+d)-\alpha d}}.$$

Offenbar ist $(1+\gamma)(2\alpha+d) > \alpha$ und daher ist $r_n \to \infty$, wobei diese Konvergenz im Wesentlichen polynomial ist. Nun setzen wir diesen Ausdruck für r_n in den obigen Ausdruck für s_n ein und erhalten nach einiger Rechnung

$$s_n = \left(\frac{\log n}{n} \right)^{\frac{1+\gamma}{(1+\gamma)(2\alpha+d)-\alpha}}.$$

Beachte, dass $s_n \to 0$ gilt und auch dies im Wesentlichen polynomial passiert. Dadurch ist $\log \frac{r_n}{s_n} \sim \log n$, was unsere Annahme von oben rechtfertigt. Zur Bestimmung der Konvergenzrate müssen wir schließlich s_n und r_n in (*) einsetzen und erhalten

$$\varepsilon_n \sim r_n^{-\gamma d} = \left(\frac{\log n}{n} \right)^{\frac{\alpha\gamma}{(1+\gamma)(2\alpha+d)-\alpha}}.$$

Die anderen Fälle verlaufen weitestgehend analog. □

Ein Vorteil des Satzes ist, dass wir nur schwache, nicht-parametrische Voraussetzungen benötigen. Ferner sind die Konvergenzraten, die wir gesehen haben, im Wesentlichen optimal. Für die Menge $\mathscr{H}_\alpha$ aller Wahrscheinlichkeitsmaße P mit $P([-1,1]^d) = 1$ und α-Hölder-stetiger Dichte gilt zum Beispiel

$$\inf_{\mathscr{L}} \sup_{h \in \mathscr{H}_\alpha} \mathbf{E}_{D \sim P^n} \|h_D - h\|_{L_1(\mu)} \geq 2^{-\frac{\alpha}{2\alpha+d}},$$

wobei $\mathscr{L}$ die Menge aller Dichteschätzer $D \mapsto h_D$ bezeichnet. In Satz 2.1.10 waren wir also nur um einen log-Term schlechter. Ferner führen andere Folgen s_n auch zu suboptimalen Raten.

Auf der anderen Seite hängt die Konstante bei ε_n im Allgemeinen stark von P ab und kann beliebig groß werden. Um die Raten zu erzielen, benötigen wir ferner Wissen über das Tailverhalten und die Hölderstetigkeit, wir haben im Allgemeinen jedoch nichts von beidem. Die Frage wäre also, wie wir diese Raten erzielen können, ohne dieses Wissen zu haben. Dies führt zum Begriff der Adaptivität.

Überdies wäre es interessant zu wissen, was wir sagen können, wenn h wesentlich glatter oder unstetig ist.

Ferner sind die Raten in Satz 2.1.10 selbst für moderate d ziemlich schlecht, für $d = 3$ und $\alpha = 1$ erhalten wir zum Beispiel $n^{-\frac{1}{5}}$. Man spricht hier vom *curse of dimensionality*.

Unser letztes Ziel sind nun Raten für die $\|\cdot\|_\infty$-Approximation. Dazu stellen wir folgenden Satz vor:

Satz 2.1.11 **$\|\cdot\|_\infty$-Orakelungleichung für Histogrammregel**
Sei $P([-1,1]^d) = 1$ und h beschränkt. Wir setzen $c := 2\|h\|_\infty(d+1)$, dann gilt für eine Würfelpartition der Weite $s \le 1$

$$P^n\left(D \in X^n : \left\|h_{D,\mathscr{A}} - h\right\|_\infty \le \sqrt{c}\sqrt{\frac{\log\left(\frac{4}{s}\right)+\tau}{s_n^d n}} + c\frac{\log\left(\frac{4}{s}\right)+\tau}{s_n^d n} + \left\|h_{P,\mathscr{A}} - h\right\|_\infty\right) \ge 1 - e^{-\tau}.$$

Beweis: Wir setzen $J := \left\{j \in \mathbf{N} : A_j \cap B_{l_\infty^d} \neq \emptyset\right\}$. Aus dem Beweis von Satz 2.1.5 wissen wir $|J| \le \left(\frac{4}{s}\right)^d$ und

$$P^n\left(D \in X^n : \left|\mathbf{E}_D \mathbf{1}_{A_j} - \mathbf{E}_P \mathbf{1}_{A_j}\right| \ge \sqrt{\frac{2P(A_j)\tau}{n}} + \frac{2\tau}{3n}\right) \le 2e^{-\tau}.$$

Mit dem Union Bound und $\mu(A_j) = s^d$ folgt

$$P^n\left(D \in X^n : \frac{1}{s^d}\sup_{j\in J}\left|\mathbf{E}_D \mathbf{1}_{A_j} - \mathbf{E}_P \mathbf{1}_{A_j}\right| \ge \sqrt{\frac{2P(A_j)(\log(2|J|)+\tau)}{\mu(A_j)s^d n}} + \frac{2\log(2|J|)+2\tau}{3ns^d}\right) \le e^{-\tau}.$$

Wir betrachten nun die linke Seite in der Wahrscheinlichkeit. Für $j \notin D$ gilt $\mathbf{E}_P \mathbf{1}_{A_j} = 0$ und $\mathbf{E}_D \mathbf{1}_{A_j} = 0$ für P^n-fast sicher alle $D \in X^n$. Damit gilt mit Lemma 2.1.4 schließlich P^n-fast sicher

$$\begin{aligned}\frac{1}{s^d}\sup_{j\in J}\left|\mathbf{E}_D \mathbf{1}_{A_j} - \mathbf{E}_P \mathbf{1}_{A_j}\right| &= \frac{1}{s^d}\sup_{j\ge 1}\left|\mathbf{E}_D \mathbf{1}_{A_j} - \mathbf{E}_P \mathbf{1}_{A_j}\right| \\ &= \left\|h_{D,\mathscr{A}} - h_{P,\mathscr{A}}\right\|_\infty.\end{aligned}$$

Auf der rechten Seite gilt

$$\frac{P(A_j)}{\mu(A_j)} = \frac{1}{\mu(A_j)}\int_{A_j} h \, \mathrm{d}\mu \le \frac{\|h\|_\infty}{\mu(A_j)}\int_{A_j} 1 \, \mathrm{d}\mu = \|h\|_\infty$$

und $\log(2|J|) \le \log\left(2\left(\frac{4}{s}\right)^d\right) \le (d+1)\log\frac{4}{s}$. Wegen

$$\left\|h_{D,\mathscr{A}} - h\right\|_\infty \le \left\|h_{D,\mathscr{A}} - h_{P,\mathscr{A}}\right\|_\infty + \left\|h_{P,\mathscr{A}} - h\right\|_\infty$$

folgt dann die Behauptung durch Kombination der Argumente. □

Satz 2.1.12 **$\|\cdot\|_\infty$-Konvergenzraten für Histogrammregel**
Sei $P([-1,1]^d) = 1$ und h sei α-Hölder-stetig. Wir betrachten $s_n \sim \left(\frac{\log n}{n}\right)^{\frac{1}{2\alpha+d}}$, dann gibt es für alle $n \ge 1$ eine Konstante k mit

$$P^n\left(D \in X^n : \left\|h_{D,\mathscr{A}} - h\right\|_\infty \le k\left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+d}}\right) \ge 1 - \frac{1}{n}.$$

Beweis: Der Beweis ist eine Kombination aus Satz 2.1.11 und Satz 2.1.7, analog zum dritten Fall im Beweis von Satz 2.1.10. □

Im dritten Fall von Satz 2.1.10 gilt also nicht nur $\|\cdot\|_{L_1}$-Konvergenz, sondern sogar Konvergenz bezüglich der Supremumsnorm.

2.2 Kernregeln

In diesem Abschnitt wollen wir analog zu den Histogrammregeln die Kernregeln untersuchen.

Definition 2.2.1 **Kernfunktion**
Eine beschränkte, monoton fallende Funktion $K : [0,\infty) \to [0,\infty)$ mit $\int_{\mathbf{R}^d} K(\|x\|)\, \mathrm{d}\mu(x) =: \kappa \in (0,\infty)$ heißt d-dimensionale Kernfunktion.

Lemma 2.2.2
Die Wahl der Norm $\|\cdot\|$ in Definition 2.2.1 spielt keine Rolle, das heißt alle Normen erfüllen $\kappa \in (0,\infty)$ unter den selben Voraussetzungen.

Beweis: Seien $\|\cdot\|$ und $\|\cdot\|'$ Normen, welche $\kappa \in (0,\infty)$ erfüllen. Dann weiß man, dass eine Konstante c mit $\|x\| \le c\,\|x\|'$ für alle $x \in \mathbf{R}$ existiert. Dann folgt

$$\int_{\mathbf{R}^d} K(\|x\|')\, \mathrm{d}\mu(x) \le \int_{\mathbf{R}^d} K\left(\frac{1}{c}\,\|x\|\right) \mathrm{d}\mu(x) = c^d \int_{\mathbf{R}^d} K(\|x\|)\, \mathrm{d}\mu(x) < \infty. \qquad \square$$

Lemma 2.2.3
Eine beschränkte, monoton fallende Funktion $h : [0,\infty) \to [0,\infty)$ ist eine d-dimensionale Kernfunktion genau dann, wenn

$$\int_0^\infty K(r) r^{d-1}\, \mathrm{d}r \in (0,\infty).$$

Beweis: Nach Lemma 2.2.2 genügt es, die Integrationsbedingung für Kernfunktionen für die euklidische Norm $\|\cdot\|_{l_2^d}$ zu betrachten. Dann gilt

$$\int_{\mathbf{R}^d} K\left(\|x\|_{l_2^d}\right)\, \mu(\mathrm{d}x) = d\,\tau_d \int_0^\infty K(r) r^{d-1}\, \mathrm{d}r,$$

wobei $\tau_d = \frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)}$ das Volumen von $B_{\|\cdot\|_2^d}$ ist. $\square$

An dieser Stelle wollen wir einige Beispiel für Funktionen geben, die für alle $d \ge 1$ Kernfunktionen sind:

- Moving Window: $K = \mathbf{1}_{[0,1]}$
- Dreiecksfunktion: $K(r) = (1-r)\mathbf{1}_{[0,1]}(r)$

- Epanechnikov: $K(r) = (1-r^2)\mathbf{1}_{[0,1]}(r)$
- Gaussian: $K(r) = e^{-r^2}$

Wir wollen nun eine Kernfunktion K und das Wahrscheinlichkeitsmaß kombinieren, um eine Dichte zu erzeugen. Dazu nehmen wir im Folgenden $\kappa = 1$ an, was in der Regel einen Vorfaktor für K erfordert. Dieser hängt von der Norm ab und ist im Wesentlichen für die Implementierung relevant.

Definition 2.2.4 **Glättung**
Sei K eine d-dimensionale Kernfunktion und Q ein Wahrscheinlichkeitsmaß auf $\mathbf{R}^d$. Dann heißt für $s > 0$

$$\begin{aligned} h_{Q,K,s}(x) &:= h_{Q,s}(x) \\ &:= \frac{1}{s^d} \int_{\mathbf{R}^d} K\left(\frac{\|x - x'\|}{s}\right) \mathrm{d}Q(x') \end{aligned}$$

eine K-Glättung von Q.

Lemma 2.2.5
Ist $h_{Q,s}$ eine K-Glättung von Q, so gilt $\int h_{Q,s}(x)\,\mathrm{d}\mu(x) = 1$, also ist $h_{Q,s}$ eine Dichte.

Beweis: Es ist

$$\begin{aligned} \int_{\mathbf{R}^d} h_{Q,s}(x)\,\mathrm{d}\mu(x) &= \int_{\mathbf{R}^d}\int_{\mathbf{R}^d} \frac{1}{s^d} K\left(\frac{\|x-x'\|}{s}\right) \mathrm{d}Q(x')\,\mathrm{d}\mu(x) \\ &= \int_{\mathbf{R}^d}\int_{\mathbf{R}^d} \frac{1}{s^d} K\left(\frac{\|x-x'\|}{s}\right) \mathrm{d}\mu(x)\,\mathrm{d}Q(x') \\ &= \int_{\mathbf{R}^d}\int_{\mathbf{R}^d} K(\|x\|)\,\mathrm{d}\mu(x)\,\mathrm{d}Q(x') \\ &= 1. \end{aligned}$$

□

Um etwas Schreibarbeit zu sparen, schreiben wir von nun an $K_s \colon \mathbf{R}^d \to [0,\infty)$ mit $x \mapsto \frac{1}{s^d} K(\|x\|)$.

Beachte, dass K_s eine Dichte ist, es gilt also $\|K_s\|_{L_1(\mu)} = 1$. Dann ist $h_{Q,s}$ die Dichte des Maßes, welches durch Faltung von dem zu K_s gehörigem Maß und Q entsteht. Mit anderen Worten ist $h_{Q,s}$ die Dichte von $\nu_s * Q = Q * \nu_s$, wobei ν_s das zu K_s gehörige Maß ist.

Beispiel 2.2.6
Für $Q = P = h\,\mathrm{d}\mu$ gilt $h_{Q,s} = K_s * h = h * K_s$. Dies folgt aus der Definition der Faltung, vergleiche hierzu [WTSkript11, Abschnitt II.5]. //

Beispiel 2.2.7
Für $Q = D$, also das empirische Maß, gilt

$$h_{D,s}(x) = \frac{1}{ns^d} \sum_{i=1}^{n} K\left(\frac{\|x - x_i\|}{s}\right),$$

was eine Mischung von solchen K_s um die Punkte x_i ist. //

Lemma 2.2.8 **(Un-)Gleichungen für die Faltung**
Die Struktur $(L_1(\mu), +, \cdot, *)$ ist eine kommutative Banachalgebra, das heißt insbesondere $(f+g)*h = f*h + g*h$, $f*g = g*f$, $(\alpha f)*g = f*(\alpha g) = \alpha(f*g)$, $(f*g)*h = f*(g*h)$ und so weiter für entsprechende Funktionen. Wichtiger ist die Eigenschaft

$$\|f*g\|_{L_1(\mu)} \le \|f\|_{L_1(\mu)} \|g\|_{L_1(\mu)}.$$

Ferner gelten die folgenden Eigenschaften:

- Ist $f \in L_p$ und $g \in L_q$ mit $1 < p < \infty$ und $p^{-1} + q^{-1} = 1$, so ist $f*g$ stetig und $\{|f*g| \ge \varepsilon\}$ ist für alle $\varepsilon > 0$ kompakt. Wir sagen, dass $f*g$ im Unendlich verschwindet und schreiben $f*g \in C_0(\mathbf{R}^d)$. Insbesondere ist $f*g$ also beschränkt.

  Für $p = 1$ und $q = \infty$ gilt $f*g \in C_b(\mathbf{R}^d)$, die Faltung liegt also im Raum der stetigen und beschränkten Funktionen. Ferner ist $f*g$ gleichmäßig stetig. Da $K_s \in L_\infty$ und $h \in L_1$ ist, bedeutet dies, dass $h_{P,s} = h * K_s$ gleichmäßig stetig und beschränkt ist. Ist ferner $g \in C_c(\mathbf{R}^d)$, so gilt auch $f*g \in C_0(\mathbf{R}^d)$.
- Young-Ungleichung: Für $1 \le p, q \le \infty$ und r mit $p^{-1} + q^{-1} + r^{-1} = 1$ und $f \in L_p$, $g \in L_q$ folgt $f*g \in L_r$ und $\|f*g\|_r \le \|f\|_p \|g\|_q$.

Beweis: Wir werden den Beweis hier nicht führen. □

Die Abbildung $D \mapsto h_{D,s}$ gibt uns nun einen Dichteschätzer. Diesen wollen wir in Hinblick auf Konsistenz und Konvergenzraten untersuchen. Dazu machen wir die Zerlegung

$$\|h_{D,s} - h\|_{L_1(\mu)} \le \|h_{D,s} - h_{P,s}\|_{L_1(\mu)} + \|h_{P,s} - h\|_{L_1(\mu)}.$$

Dies entspricht im Grunde dem Vorgehen, das wir bereits bei den Histogrammregeln gesehen haben.

Satz 2.2.9 **Approximationsfehler**
Sei K eine d-dimensionale Kernfunkion mit $\kappa = 1$. Dann gibt es zu jedem $\varepsilon > 0$ ein $s_\varepsilon > 0$, so dass für $s \in (0, s_\varepsilon]$ die Abschätzung

$$\left\| h_{P,s} - h \right\|_{L_1(\mu)} \leq \varepsilon$$

gilt. Ist h ferner α-Hölder-stetig mit $\int K(r) r^{\alpha+d-1} \, \mathrm{d}r < \infty$, so gibt es eine Konstante $C > 0$ mit

$$\left\| h_{P,s} - h \right\|_\infty \leq C s^\alpha.$$

Beweis: Sei $f \in C_c(\mathbf{R}^d)$ mit $\|h - f\|_{L_1(\mu)} \leq \frac{\varepsilon}{3}$, dann gilt

$$\begin{aligned}
\left\| h_{P,s} - h \right\|_{L_1(\mu)} &= \int_{\mathbf{R}^d} |h * K_s - h| \, \mathrm{d}\mu \\
&\leq \int_{\mathbf{R}^d} |h * K_s - f * K_s| \, \mathrm{d}\mu + \int_{\mathbf{R}^d} |f * K_s - f| \, \mathrm{d}\mu + \int_{\mathbf{R}^d} |f - h| \, \mathrm{d}\mu \\
&\leq \frac{2\varepsilon}{3} + \int_{\mathbf{R}^d} |f * K_s - f| \, \mathrm{d}\mu.
\end{aligned}$$

Nun gibt es ein $M > 0$ mit $\operatorname{supp} f \subset M \cdot B_{\|\cdot\|}$ und ein $r > 0$ mit

$$\int_{\mathbf{R}^d \setminus r B_{\|\cdot\|}} K(\|x\|) \, \mathrm{d}\mu(x) \leq \frac{\varepsilon}{g \|f\|_1}.$$

Wir definieren $L \colon \mathbf{R}^d \to [0, \infty)$ mit $L(x) := \mathbf{1}_{[-r,r]}(\|x\|) K(\|x\|)$ und $L_s \colon \mathbf{R}^d \to [0, \infty)$ mit $L_s(x) := \frac{1}{s^d} L\left(\frac{x}{s}\right)$. Damit folgt

$$\begin{aligned}
\int |f * K_s - f| \, \mathrm{d}\mu &\leq \int |f * K_s - f * L_s| \, \mathrm{d}\mu + \int |f * L_s - f| \, \mathrm{d}\mu \\
&\leq \|f\|_{L_1} \|K_s - L_s\|_{L_1} + \int \left| f * L_s - f \int L_s \, \mathrm{d}\mu \right| \, \mathrm{d}\mu + \int \left| f * \int (L_s - K_s) \, \mathrm{d}\mu \right| \, \mathrm{d}\mu \\
&\leq 2 \|f\|_{L_1} \|K_s - L_s\|_{L_1} + \int \left| f * L_s - f \int L_s \, \mathrm{d}\mu \right| \, \mathrm{d}\mu.
\end{aligned}$$

Ferner gilt

$$\begin{aligned}
\|K_s - L_s\|_{L_1} &= \int \frac{1}{s^d} \left| \mathbf{1}_{[-r,r]}\left(\frac{normx}{s}\right) K\left(\frac{normx}{s}\right) - \left(\frac{normx}{s}\right) \right| \, \mathrm{d}\mu(x) \\
&= \int \left| \mathbf{1}_{[-r,r]}(\|x\|) K(\|x\|) - K(\|x\|) \right| \, \mathrm{d}\mu(x) \\
&= \int_{\mathbf{R}^d \setminus r B_{\|\cdot\|}} K(\|x\|) \, \mathrm{d}\mu(x) \\
&\leq \frac{\varepsilon}{g \|f\|_1}.
\end{aligned}$$

Schließlich gilt für $s \leq 1$

$$\begin{aligned} I &:= \int \left| f * L_s - f \int L_s \, \mathrm{d}\mu \right| \, \mathrm{d}\mu \\ &= \int \left| \int (f(x-x') - f(x)) L_s(x') \, \mathrm{d}\mu(x') \right| \, \mathrm{d}\mu(x) \\ &\leq \int_{(r+M)B_{\|\cdot\|}} \int_{\mathbf{R}^d} \left| f(x-x') - f(x) \right| L_s(x') \, \mathrm{d}\mu(x') \, \mathrm{d}\mu(x). \end{aligned}$$

Für $\varepsilon' := \frac{\varepsilon}{g(r+M)^d \mu(B_{\|\cdot\|})} > 0$ gilt, da f gleichmäßig stetig ist, ferner, dass ein $s_0 > 0$ existiert, so dass für alle $s \leq s_0$ und $\|x'\| \leq rs$ gilt:

$$\left| f(x-x') - f(x) \right| \leq \varepsilon'.$$

Damit erhalten wir

$$\begin{aligned} \int_{\mathbf{R}^d} \left| f(x-x') - f(x) \right| L_s(x') \, \mathrm{d}\mu(x') &\leq \varepsilon' \int_{rsB_{\|\cdot\|}} L_s(x') \, \mathrm{d}\mu(x') \\ &\leq \varepsilon' \int_{\mathbf{R}^d} K_s \, \mathrm{d}\mu \\ &= \varepsilon' \end{aligned}$$

und damit $I \leq \int_{(r+M)B_{\|\cdot\|}} \varepsilon' \, \mathrm{d}\mu = \frac{\varepsilon}{9}$. Damit folgt die erste Behauptung.

Für die zweite Behauptung betrachten wir

$$\begin{aligned} \left| h_{P,s}(x) - h(x) \right| &= \left| \frac{1}{s^d} \int_{\mathbf{R}^d} K\left(\frac{\|x-x'\|}{s} \right) h(x') \, \mathrm{d}\mu(x') - h(x) \right| \\ &= \left| \int_{\mathbf{R}^d} K(\|x'\|) h(x+sx') \, \mathrm{d}\mu(x') - h(x) \right| \\ &= \left| \int_{\mathbf{R}^d} K(\|x'\|)(h(x+sx') - h(x)) \, \mathrm{d}\mu(x') \right| \\ &\leq \int_{\mathbf{R}^d} K(\|x'\|) c'(s\|x'\|)^\alpha \, \mathrm{d}\mu(x') \\ &\leq c \int_{\mathbf{R}^d} K(\|x'\|_{l_2^d}) s^\alpha \|x'\|_{l_2^d}^\alpha \, \mathrm{d}\mu(x') \\ &\leq c s^\alpha \int_{\mathbf{R}^d} K(r) r^{\alpha+d-1} \, \mathrm{d}r. \end{aligned}$$

Da das letzte Integral nach Voraussetzung endlich ist, sind wir fertig. □

In der Analysis haben wir gesehen, dass $h_{P,s} \to h$ punktweise konvergiert, wenn K ein kompakter Träger und h gleichmäßig stetig ist.

Als nächstes wollen wir

$$\|h_{D,s} - h_{P,s}\|_{L_1} = \int \left| h_{D,s}(x) - h_{P,s}(x) \right| \, \mathrm{d}\mu(x)$$

erreichen, doch der Union Bound funktioniert über endlich viele x nicht auf diese Art und Weise. Wir versuchen daher $y_1, \dots, y_m \in \mathbf{R}^d$ derart zu finden, dass für jedes $x \in \mathbf{R}^d$ ein y_i existiert, so dass $h_{D,s}(x)$ gut durch $h_{D,s}(y_i)$ approximiert wird. Dazu müssen wir jedoch klären, wie sich dieses „gut“ zu der Zahl m verhält.

Für einen metrischen Raum (X,d) schreiben wir

$$B_d(x,\varepsilon) := \{x' \in X : d(x,x') \le \varepsilon\}.$$

Definition 2.2.10 **Überdeckungszahl**
Sei (X,d) ein metrischer Raum und $A \subset X$. Für $\varepsilon > 0$ heißt

$$\mathcal{N}(A,d,\varepsilon) := \min\{n \ge 1 : \exists_{x_1,\dots,x_n \in X} A \subset \bigcup_{i=1}^{n} B_d(x_i,\varepsilon)\}$$

dann die ε-Überdeckungszahl von A.

Beachte, dass $\min \emptyset = \infty$ ist.

An dieser Stelle wollen wir einige Dinge anmerken:

- Wir könnten für unsere Betrachtungen auf $x_1, \dots, x_n \in A$ und offene Kugeln betrachten.
- A ist präkompakt genau dann, wenn $\mathcal{N}(A,d,\varepsilon) < \infty$ für alle $\varepsilon > 0$ gilt.
- $\mathcal{N}(A,d,\varepsilon)$ wächst für $\varepsilon \to 0$. Für uns wird ε das oben genannte „gut“ messen, während die Überdeckungszahl den Parameter m beschreibt.
- Für verschiedene A können unterschiedliche Wachstumsverhalten auftreten. Dies führt von der qualitativen Präkompaktheitsdefinition zu einem quantitativen Begriff.
- Typische Verhaltensweisen sind zum Beispiel $\mathcal{N}(A,d,\varepsilon) \sim \varepsilon^{-p}$, $\log \mathcal{N}(A,d,\varepsilon) \sim \varepsilon^{-p}$ oder $\log \mathcal{N}(A,d,\varepsilon) \sim \left(\log \frac{1}{\varepsilon}\right)^p$ für $\varepsilon \to 0$.
- Ist $E \supset A$ ein Banachraum mit Norm $\|\cdot\|$, so schreiben wir $\mathcal{N}(A, \|\cdot\|, \varepsilon)$.
- Für $A \subset B$ gilt offenbar $\mathcal{N}(A,d,\varepsilon) \le \mathcal{N}(B,d,\varepsilon)$.

Lemma 2.2.11
Seien (X,d) und (Y,e) metrische Räume und $T\colon X \to Y$ eine α-Hölderstetige Abbildung mit Konstante c. Für $A \subset X$ gilt dann für alle $\varepsilon > 0$

$$\mathcal{N}(T(A),e,c\varepsilon^\alpha) \le \mathcal{N}(A,d,\varepsilon).$$

Beweis: Sei $x_1, \dots, x_n$ ein ε-Netz von A, das heißt es gilt $A \subset \bigcup_{j=1}^n B_d(x_j,\varepsilon)$. Wir setzen $y_i := T(x_i)$ und zeigen, dass dies ein $c\varepsilon^\alpha$-Netz von $T(A)$ ergibt. Dazu zeigen wir zunächst

$T(B_d(x_i,\varepsilon)) \subset B_e(y_i, c\varepsilon^\alpha)$. Kombinieren wir dies mit der Netzeigenschaft, so erhalten wir die Behauptung.

Sei also $y \in T(B_d(x_i,\varepsilon))$, dann existiert ein $x \in B_d(x_i,\varepsilon)$ mit $T(x) = y$. Damit folgt $e(y,y_i) = e(T(x),T(x_i)) \leq cd^\alpha(x,x_i) \leq c\varepsilon^\alpha$. □

In Banachräumen gilt für $c \neq 0$ also insbesondere

$$\mathcal{N}(cA, \|\cdot\|, \varepsilon) = \mathcal{N}\left(A, \|\cdot\|, \frac{\varepsilon}{c}\right).$$

Satz 2.2.12
Seien $\|\cdot\|$ und $\|\cdot\|'$ Normen auf $\mathbf{R}^d$. Dann gibt es eine Konstante $c > 0$, so dass für alle $\varepsilon \in (0,1]$ gilt:

$$\mathcal{N}(B_{\|\cdot\|}, \|\cdot\|', \varepsilon) \leq c\varepsilon^{-d}.$$

Es gilt zu beachten, dass für unendlichdimensionale Räume nichts Vergleichbares gilt.

Beweis: Wir betrachten zunächst $\|\cdot\| = \|\cdot\|' = \|\cdot\|_{l_\infty^d}$. Sei nun $\varepsilon \in (0,1]$, dann existiert ein $l \geq 1$ mit $\frac{1}{l+1} < \varepsilon \leq \frac{1}{l}$. Wir zerlegen $B_{l_\infty^d}$ in Würfel der Länge $\frac{2}{l+1} =: 2s$. Hierfür benötigen wir $\left(\frac{2}{2s}\right)^d = (l+1)^d$ Würfel.

Wir setzen nun $x_1, \ldots, x_{(l+1)^d}$ als die Mittelpunkte der Würfel, also ist jeder Würfel eine $\|\cdot\|_{l_\infty^d}$-Kugel mit Radius s. Dann gilt

$$\begin{aligned}\mathcal{N}(B_{\|\cdot\|_\infty}, \|\cdot\|_\infty, \varepsilon) \leq \mathcal{N}\left(B_{\|\cdot\|_\infty}, \|\cdot\|_\infty, \frac{1}{l+1}\right) &\leq (l+1)^d \leq 2^d l^d \\ &\leq 2^d \varepsilon^{-d}.\end{aligned}$$

Wir betrachten nun $B_{\|\cdot\|}$. Dann existiert ein $k \geq 1$ mit $B_{\|\cdot\|} \subset kB_{\|\cdot\|_\infty}$. Damit folgt dann

$$\mathcal{N}(B_{\|\cdot\|}, \|\cdot\|_\infty, \varepsilon) \leq \mathcal{N}\left(B_{\|\cdot\|_\infty}, \|\cdot\|_\infty, \frac{\varepsilon}{k}\right).$$

Schließlich ist $\mathrm{id}\colon l_\infty^d \to (\mathbf{R}^d, \|\cdot\|')$ Lipschitzstetig mit Konstante $k' \geq 1$ und mit Lemma 2.2.11 folgt daher $\mathcal{N}(B_{\|\cdot\|}, \|\cdot\|', \varepsilon) \leq \mathcal{N}\left(B_{\|\cdot\|}, \|\cdot\|_\infty, \frac{\varepsilon}{k'}\right)$. □

Satz 2.2.13
Sei K eine d-dimensionale α-Hölderstetige Kernfunktion und $s > 0$. Für $x \in \mathbf{R}^d$ definieren wir

$$k_{x,s} := s^{-d} K\left(\frac{\|x - \cdot\|}{s}\right) = K_s(x - \cdot).$$

Für $x, x' \in \mathbf{R}^d$ gilt dann

$$\sup_{y \in \mathbf{R}^d} \left| k_{x,s}(y) - k_{x',s}(y) \right| \leq \frac{c}{s^{\alpha+d}} \left\| x - x' \right\|^{\alpha}.$$

Mit anderen Worten ist die Abbildung $\mathbf{R}^d \to l_\infty(\mathbf{R}^d)$ in den Raum der beschränkten Funktionen α-Hölderstetig mit der Konstanten $\frac{c}{s^{\alpha+d}}$.

Beweis: Es ist

$$\begin{aligned} \left| k_{x,s}(y) - k_{x',s}(y) \right| &= \frac{1}{s^d} \left| K\left(\frac{\|x-y\|}{s} \right) - K\left(\frac{\left\|x'-y\right\|}{s} \right) \right| \\ &\leq \frac{c}{s^d} \left| \frac{\|x-y\|}{s} - \frac{\left\|x'-y\right\|}{s} \right|^{\alpha} \\ &\leq \frac{c}{s^{\alpha+d}} \left\| x - x' \right\|^{\alpha}. \end{aligned}$$

□

Korollar 2.2.14
Es sei K eine d-dimensionale α-Hölderstetige Kernfunktion und $s \in (0,1]$. Wir definieren $k_{x,s}$ wie in Lemma 2.2.13, also $k_{x,s} := K_s(x - \cdot)$, und betrachten $\mathcal{K}_s := \left\{ k_{x,s} : x \in rB_{\|\cdot\|} \right\} \subset l_\infty(\mathbf{R}^d)$ für $r \geq 1$. Dann gibt es eine von r und s unabhängige Konstante c mit

$$\mathcal{N}(\mathcal{K}_s, \|\cdot\|_\infty, \varepsilon) \leq c r^d s^{-d-\frac{d^2}{\alpha}} \varepsilon^{-\frac{d}{\alpha}}$$

für alle $\varepsilon \in (0,1]$.

Beweis: $\mathcal{K}_s$ ist das Bild der Hölderstetigen Abbildung $rB_{\|\cdot\|} \to l_\infty(\mathbf{R}^d)$ nach Lemma 2.2.13 mit der Konstanten $\frac{c}{s^{\alpha+d}}$. Mit Lemma 2.2.11 und Lemma 2.2.12 folgt nun

$$\begin{aligned} \mathcal{N}(\mathcal{K}_s, \|\cdot\|_\infty, \varepsilon) &\leq \mathcal{N}\left(rB_{\|\cdot\|}, \|\cdot\|, \left(\frac{s^{\alpha+d}}{c} \varepsilon \right)^{\frac{1}{\alpha}} \right) \\ &= \mathcal{N}\left(B_{\|\cdot\|}, \|\cdot\|, \left(\frac{s^{\alpha+d} \varepsilon}{c r^\alpha} \right)^{\frac{1}{\alpha}} \right) \\ &\leq c' \left(\frac{s^{\alpha+d} \varepsilon}{r^\alpha} \right)^{-\frac{d}{\alpha}}. \end{aligned}$$

□

Durch Logarithmieren der Abschätzung aus Korollar 2.2.14 erhalten wir

$$\log \mathcal{N}(\mathcal{K}_s, \|\cdot\|_\infty, \varepsilon) \leq \left(d + \frac{d^2}{\alpha} \right) \log \frac{r}{s} + \frac{d}{\alpha} \log \frac{1}{\varepsilon} + \log c'.$$

Ist $\frac{r}{s} \geq 2$, so gibt es ein c_2 mit

$$\log \mathcal{N}(\mathcal{K}_s, \|\cdot\|_\infty, \varepsilon) \leq c_2 \log \frac{r}{s\varepsilon}.$$

Da ferner $\mathrm{id}\colon L_\infty(\mathbf{R}^d) \to L_2(P)$ lipschitzstetig ist, folgt schließlich

$$\log \mathcal{N}(\mathcal{K}_s, \|\cdot\|_{L_2(P)}, \varepsilon) \leq c_2 \log \frac{r}{s\varepsilon}, \tag{*}$$

wobei c_2 nicht vom Wahrscheinlichkeitsmaß P abhängt. Dies gilt sogar für wesentlich allgemeinere Kernfunktionen, insbesondere auch für gewisse unstetige Kernfunktionen. Die folgende Aussage könnte auch für Kernfunktionen, die (*) erfüllen, bewiesen werden. Dies ist jedoch deutlich anspruchsvoller.

Satz 2.2.15 **Orakelungleichung für Kernregeln**
Sei K eine β-Hölderstetige Kernfunktion mit $\int_0^\infty K(r) r^{\beta+d-1}\,\mathrm{d}r < \infty$. Dann gibt es eine Konstante $c \geq 1$, so dass für alle $s \leq 1$, $r \geq 2$, $\tau \geq 1$ und $n \geq 1$ die folgende Abschätzung gilt:

$$\begin{aligned} P^n\Big(D \in X^n : \big\|h_{D,s} - h_{P,s}\big\|_{L_1(\mu)} \leq c\sqrt{\frac{r^d\left(\tau + \log\left(\frac{r}{s}n\right)\right)}{s^d n}} + c\frac{r^d\left(\tau + \log\left(\frac{r}{s}n\right)\right)}{s^d n} + \dots \\ \dots + 2P\left(\mathbf{R}^d \setminus rB_{\|\cdot\|}\right) + \left(\frac{s}{r}\right)^\beta\Big) \\ \geq 1 - e^{-\tau}. \end{aligned}$$

Beweis: Wir definieren $k_{x,s} := s^{-d} K\left(\frac{\|x-\cdot\|}{s}\right)$ und $f_{x,s} := k_{x,s} - \mathbf{E}_P k_{x,s}$. Beachte, dass dann $\mathbf{E}_D f_{x,s} = \mathbf{E}_D k_{x,s} - \mathbf{E}_P k_{x,s} = h_{D,s}(x) - h_{P,s}(x)$ ist. Damit gilt

$$\big\|h_{D,s}(x) - h_{P,s}(x)\big\|_{L_1(\mu)} = \int \left|\mathbf{E}_D f_{x,s}\right| \,\mathrm{d}\mu(x).$$

Jetzt schätzen wir zunächst $\mathbf{E}_D f_{x,s}$ für ein x mit der Bernsteinungleichung ab. Dazu vergewissern wir uns, dass die Voraussetzungen hierfür erfüllt sind:

- $\mathbf{E} f_{x,s} = 0$ ist klar.
- Es ist $\big\|f_{x,s}\big\|_\infty \leq 2\big\|k_{x,s}\big\|_\infty \leq 2s^{-d}\|K\|_\infty = 2d^{-d}K(0)$.
- Ferner gilt $\mathbf{E}_P f_{x,s}^2 \leq \mathbf{E}_P k_{x,s}^2 = s^{-2d}\int_{\mathbf{R}^d} K^2\left(\frac{\|x-x'\|}{s}\right)\mathrm{d}P(x')$.

Mit der zweiseitigen Bernsteinungleichung folgt dann

$$P^n\left(D \in X^n : \left|\mathbf{E}_D f_{x,s}\right| \geq \sqrt{\frac{2\tau\int K^2\left(\frac{\|x-x'\|}{s}\right)\mathrm{d}P(x')}{s^{2d}n}} + \frac{4K(0)\tau}{3s^d n}\right) \leq 2e^{-\tau} \tag{*}$$

für jedes $x \in \mathbf{R}^d$. Sei nun $\mathcal{K}_s' := \{f_{x,s} : x \in rB_{\|\cdot\|}\}$. Ferner seien $y_1, \ldots, y_m \in rB_{\|\cdot\|}$ so gewählt, dass $k_{y_1,s}, \ldots, k_{y_m,s}$ ein ε-Netz von $\mathcal{K}_s'$ bezüglich $\|\cdot\|_\infty$ ist und $m = \mathcal{N}\left(\mathcal{K}_s', \|\cdot\|_\infty, \frac{\varepsilon}{2}\right)$. Es gilt $\log 2m \le c_1 \log \frac{r}{s\varepsilon}$, wobei c_1 eine von r, s und ε unabhängige Konstante ist. Aus (*) und dem Union Bound (vgl. den Beweis von Satz 2.1.5) folgt

$$P^n\left(D \in X^n : \sup_{j=1,\ldots,m} \left|\mathbf{E}_D f_{y_j,s}\right| < \sqrt{\frac{2\int K^2\left(\frac{\|x-x'\|}{s}\right) \mathrm{d}P\,(\tau + \log(2m))}{s^{2d}n}} + \frac{4K(0)(\tau+\log(2m))}{s^d n}\right) \ge 1 - e^{-\tau}. \tag{**}$$

Wir wollen nun die linke und rechte Seite mit x statt den y_j haben. Sei dazu $D \in X^n$ mit der entsprechenden Abschätzung aus (**). Für $x \in rB_{\|\cdot\|}$ folgt, dass es ein y_j gibt mit $\left\|k_{x,s} - k_{y_j,s}\right\|_\infty \le \varepsilon$. Damit erhalten wir

$$\begin{aligned}
\left|\left|\mathbf{E}_D f_{x,s}\right| - \left|\mathbf{E}_D f_{y_j,s}\right|\right| &\le \left|\mathbf{E}_D f_{x,s} - \mathbf{E}_D f_{y_j,s}\right| \\
&\le \left|\mathbf{E}_D k_{x,s} - \mathbf{E}_D k_{y_j,s}\right| + \left|\mathbf{E}_P k_{x,s} - \mathbf{E}_P k_{y_j,s}\right| \\
&\le \left\|k_{x,s} - k_{y_j,s}\right\|_{L_1(D)} + \left\|k_{x,s} - k_{y_j,s}\right\|_{L_1(P)} \\
&\le 2\varepsilon.
\end{aligned}$$

Damit folgt nun

$$\left|\mathbf{E}_D f_{x,s}\right| \le \left|\mathbf{E}_D f_{y_j,s}\right| + 2\varepsilon. \tag{***}$$

Wir setzen nun $a := \sqrt{\frac{2(\tau+\log(2m))}{n}}$, dann gilt

$$\begin{aligned}
\left|a\sqrt{\frac{\int K^2\left(\frac{\|x-x'\|}{s}\right) \mathrm{d}P(x')}{s^{2d}}} - a\sqrt{\frac{\int K^2\left(\frac{norm y_j - x'}{s}\right) \mathrm{d}P(x')}{s^{2d}}}\right| &= \left|\left\|ak_{x,s}\right\|_{L_2(P)} - \left\|ak_{y_j,s}\right\|_{L_2(P)}\right| \\
&\le a\left\|k_{x,s} - k_{y_j,s}\right\|_{L_2(P)} \\
&\le \sqrt{\frac{2(\tau+\log(2m))}{n}}\,\varepsilon.
\end{aligned} \tag{****}$$

Jetzt bauen wir diese Dinge zusammen. Es ist

$$\begin{aligned}
\left|\mathbf{E}_D f_{x,s}\right| &\overset{(***)}{\le} \left|\mathbf{E}_D f_{y_j,s}\right| + 2\varepsilon \\
&\le \sqrt{\frac{2\int K^2\left(\frac{\|y_j-x'\|}{s}\right) d\,dP(x')(\tau+\log(2m))}{s^{2d}n}} + \frac{4K(0)(\tau+\log(2m))}{s^d n} + 2\varepsilon \\
&\overset{(****)}{\le} \sqrt{\frac{2\int K^2\left(\frac{\|x-x'\|}{s}\right) \mathrm{d}P(x')(\tau+\log(2m))}{s^{2d}n}} + \frac{4K(0)(\tau+\log(2m))}{s^d n} + \ldots \\
&\qquad \ldots + 2\varepsilon + \sqrt{\frac{2(\tau+\log(2m))}{n}} \cdot \varepsilon
\end{aligned}$$

für alle $x \in rB_{\|\cdot\|}$. Für D folgt nun

$$\begin{aligned}\int_{rB_{\|\cdot\|}} \left|\mathbf{E}_D f_{x,s}\right| \,\mathrm{d}\mu(x) \le \int_{rB_{\|\cdot\|}} \sqrt{\frac{2\int k_{x,s}^2(x')\,\mathrm{d}P(x')(\tau+\log(2m))}{n}}\,\mathrm{d}\mu(x) + \dots \\ \dots + r^d\mu(B_{\|\cdot\|})\frac{4K(0)(\tau+\log(2m))}{s^d n} + \dots \\ \dots + r^d\mu(B_{\|\cdot\|})\left(\sqrt{\frac{2(\tau+\log(2m))}{n}} + 2\right)\varepsilon.\end{aligned}$$

Wir betrachten nun das Integral. Mit $\|f\|_{\frac{1}{2}} \le \nu(\Omega)\|f\|_1$ (vgl. Beweis von Satz 2.1.5) folgt

$$\begin{aligned}\int_{rB_{\|\cdot\|}} \sqrt{\frac{2\int k_{x,s}^2(x')\,\mathrm{d}P(x')(\tau+\log(2m))}{n}}\,\mathrm{d}\mu(x) \le \sqrt{\mu(rB_{\|\cdot\|})}\cdot\dots \\ \dots\cdot\sqrt{\int_{rB_{\|\cdot\|}} \frac{(2\tau+\log(2m))\int k_{x,s}^2(x')\,\mathrm{d}P(x')}{n}\,\mathrm{d}\mu(x)}.\end{aligned}$$

Ferner ist

$$\begin{aligned}\int_{rB_{\|\cdot\|}}\int k_{x,s}^2(x')\,\mathrm{d}P(x')\,\mu(\mathrm{d}x) &= \int_{\mathbf{R}^d}\int_{rB_{\|\cdot\|}} \frac{K^2\left(\frac{\|x-x'\|}{s}\right)}{s^{2d}}\,\mathrm{d}\mu(x)\,\mathrm{d}P(x') \\ &\le \int_{\mathbf{R}^d}\int_{\mathbf{R}^d} s^{-2d}K^2\left(\frac{\|x\|}{s}\right)\,\mathrm{d}\mu(x)\,\mathrm{d}P(x') \\ &= s^{-d}\int_{\mathbf{R}^d} K^2(\|x\|)\,\mathrm{d}\mu(x) \\ &= c_2 s^{-d}.\end{aligned}$$

Wir setzen nun $\varepsilon = \frac{1}{n}$ und erhalten $\log 2m \le c_1 \log \frac{r}{s} n$. Damit erhalten wir

$$\begin{aligned}\int_{rB_{\|\cdot\|}} \left|\mathbf{E}_D f_{x,s}\right| \,\mathrm{d}\mu(x) &\le c_3\left(\sqrt{\frac{r^d(\tau+\log(2m))}{s^d n}} + \frac{r^d(\tau+\log(2m))}{s^d n} + \varepsilon r^d\sqrt{\frac{\tau+\log(2m)}{n}}\right) \\ &\le c_4\left(\sqrt{\frac{r^d\left(\tau+\log\frac{r}{s}n\right)}{s^d n}} + \frac{r^d(\tau+\log\frac{r}{s}n)}{s^d n}\right). \qquad (*)\end{aligned}$$

Jetzt mssen wir noch das entsprechende Integral über der Menge $\mathbf{R}^d \setminus rB_{\|\cdot\|}$ abschätzen. Nach Definition gilt

$$\int_{\mathbf{R}^d\setminus rB_{\|\cdot\|}} \left|\mathbf{E}_D f_{x,s}\right| \,\mathrm{d}\mu(x) \le \int_{\mathbf{R}^d\setminus rB_{\|\cdot\|}} \mathbf{E}_D k_{x,s}\,\mathrm{d}\mu(x) + \int_{\mathbf{R}^d\setminus rB_{\|\cdot\|}} \mathbf{E}_P k_{x,s}\,\mathrm{d}\mu(x).$$

Wir betrachten jetzt $\int_{\mathbf{R}^d \setminus rB_{\|\cdot\|}} \mathbf{E}_Q k_{x,s} \, \mathrm{d}\mu(x)$ für ein beliebiges Wahrscheinlichkeitsmaß Q, um beide Ausdrücke simultan zu behandeln. Dann erhalten wir

$$\begin{aligned}
\int_{\mathbf{R}^d \setminus rB_{\|\cdot\|}} \mathbf{E}_Q k_{x,s} \, \mathrm{d}\mu(x) &= \int_{\mathbf{R}^d \setminus rB_{\|\cdot\|}} \int_{\mathbf{R}^d} \frac{1}{s^d} K\left(\frac{\|x - x'\|}{s}\right) \mathrm{d}Q(x') \, \mathrm{d}\mu(x) \\
&= \int_{\mathbf{R}^d} \int_{\mathbf{R}^d} \frac{1}{s^d} K(\|x\|) \mathbf{1}_{\mathbf{R}^d \setminus rB_{\|\cdot\|}}(sx + x') \, \mathrm{d}\mu(x) \, \mathrm{d}Q(x') \\
&= \int_{\mathbf{R}^d} K(\|x\|) \int_{\mathbf{R}^d} \mathbf{1}_{\mathbf{R}^d \setminus rB_{\|\cdot\|}}(sx + x') \, \mathrm{d}Q(x') \, \mathrm{d}\mu(x) \\
&\leq \int_{t_0 B_{\|\cdot\|}} K(\|x\|) \int_{\mathbf{R}^d} \mathbf{1}_{\mathbf{R}^d \setminus rB_{\|\cdot\|}}(sx + x') \, \mathrm{d}Q(x') \, \mathrm{d}\mu(x) + \ldots \\
&\qquad \ldots + \int_{\mathbf{R}^d \setminus t_0 B_{\|\cdot\|}} K(\|x\|) \, \mathrm{d}\mu(x).
\end{aligned}$$

Wir setzen $t_0 := \frac{r}{2s}$. Dann ist $\mathbf{1}_{\mathbf{R}^d \setminus rB_{\|\cdot\|}}(sx + x') = 1$ genau dann, wenn $\|sx + x'\| > r$ ist. In diesem Fall ist $\|x'\| > r - s\|x\| \geq r - st_0 = \frac{r}{2}$, falls zudem $x \in t_0 B_{\|\cdot\|}$ ist. Damit folgt

$$\begin{aligned}
\int_{\mathbf{R}^d \setminus rB_{\|\cdot\|}} \mathbf{E}_Q k_{x,s} \, \mathrm{d}\mu(x) &\leq \int_{t_0 B_{\|\cdot\|}} K(\|x\|) Q\left(\mathbf{R}^d \setminus \frac{r}{s} B_{\|\cdot\|}\right) \mathrm{d}\mu(x) + \ldots \\
&\qquad \ldots + \int_{\mathbf{R}^d \setminus t_0 B_{\|\cdot\|}} K(\|x\|) \, \mathrm{d}\mu(x) \\
&\leq Q\left(\mathbf{R}^d \setminus \frac{r}{s} B_{\|\cdot\|}\right) + c_5 \int_{t_0}^{\infty} K(t) t^{d-1} \, \mathrm{d}t \\
&\leq Q\left(\mathbf{R}^d \setminus \frac{r}{s} B_{\|\cdot\|}\right) + c_5 \int_{t_0}^{\infty} K(t) t_0^{-\beta} t^{d+\beta-1} \, \mathrm{d}t \\
&\leq Q\left(\mathbf{R}^d \setminus \frac{r}{s} B_{\|\cdot\|}\right) + c_6 t_0^{-\beta} \\
&\leq Q\left(\mathbf{R}^d \setminus \frac{r}{s} B_{\|\cdot\|}\right) + c_7 \left(\frac{s}{r}\right)^{\beta}.
\end{aligned}$$

Mit der Hoeffdingungleichung gilt

$$P^n\left(D \in X^n : D\left(\mathbf{R}^d \setminus \frac{r}{s} B_{\|\cdot\|}\right) - P\left(\mathbf{R}^d \setminus \frac{r}{s} B_{\|\cdot\|}\right) < \sqrt{\frac{\tau}{2n}}\right) \geq 1 - e^{-\tau}.$$

Damit ist die Wahrscheinlichkeit für

$$\int_{\mathbf{R}^d \setminus rB_{\|\cdot\|}} \left|\mathbf{E}_D f_{x,s}\right| \, \mathrm{d}\mu(x) \leq 2P\left(\mathbf{R}^d \setminus rB_{\|\cdot\|}\right) + \sqrt{\frac{\tau}{2n}} + 2c_7 \left(\frac{s}{r}\right)^{\beta}.$$

größer gleich $1-e^{-\tau}$. Fassen wir alles zusammen, so erhalten wir mit einer Wahrscheinlichkeit $\geq 1-2e^{-\tau}$

$$\begin{aligned}\left\|h_{D,s}-h_{P,s}\right\|_{L_1(\mu)} &= \int_{\mathbf{R}^d}\left|\mathbf{E}_D f_{x,s}\right| \,\mathrm{d}\mu(x)\\ &\leq c_5\left(\sqrt{\frac{r^d(\tau+\log(2m))}{s^d n}}+\frac{r^d(\tau+\log(2m))}{s^d n}\right)+\ldots\\ &\qquad \ldots+2P\left(\mathbf{R}^d\setminus\frac{r}{2}B_{\|\cdot\|}\right)+\sqrt{\frac{\tau}{2n}}+c_8\left(\frac{s}{r}\right)^{\beta}.\end{aligned}$$

Nach einer Transformation $\frac{r}{2}\to r$ und $\log(2)-\tau\to-\tau$ folgt dann die Behauptung. □

Der Beweis wäre einfacher geworden, wenn wir statt der Abschätzung „$\int\sqrt{\int}\leq\sqrt{\int\int}$“ ein elementareres Argument verwendet hätten. Dies hätte jedoch die Faktoren r^{2d} statt r^d zur Folge gehabt.

Satz 2.2.16 **Konsistenz und Konvergenzraten für Kernregel**
Sei K eine β-Hölderstetige Kernfunktion mit $\int_0^\infty K(r)r^{\beta+d-1}\,\mathrm{d}r<\infty$ und (s_n) sei eine Nullfolge. Wir betrachten das Verfahren $X^n\to\mathscr{L}_0(X)$ mit $D\mapsto h_{D,s_n}$. Dann gelten die folgenden Aussagen:

i) Ist $\frac{ns^d}{\log n}\to\infty$, so ist das Verfahren konsistent.

ii) Ist die Dichte h α-Hölderstetig und zum Beispiel $P(\mathbf{R}^d\setminus r_0B_{\|\cdot\|})=0$ für ein $r_0>0$ und $s_n\sim\left(\frac{\log n}{n}\right)^{\frac{1}{2\alpha+d}}$, so gilt

$$P^n\left(D\in X^n:\left\|h_{D,s_n}-h\right\|_{L_1(\mu)}\leq c\left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+d}}\right)\geq 1-\frac{1}{n}.$$

iii) Für ein anderes Tailverhalten erhalten wir entsprechend andere Raten wie bei der Histogrammregel (vgl. Satz 2.1.10).

Beweis: Der Beweis verläuft im Wesentlichen analog zum Beweis von Satz 2.1.10. □

Zum Schluss wollen wir noch eine Schätzung für die L_∞-Norm herleiten. Den Approximationsfehler haben wir in Satz 2.2.9 bereits behandelt, es fehlt also lediglich noch eine Orakelungleichung.

Satz 2.2.17 $\|\cdot\|_\infty$**-Orakelungleichung für Kernregel**
Sei K eine d-dimensionale Kernfunktion für welche es Konstanten $c,\gamma>0$ gibt mit $\mathscr{N}(\mathscr{K}_1,\|\cdot\|_{L_2(Q)},\varepsilon)\leq c\varepsilon^{-\gamma}$ für alle $\varepsilon\in(0,1]$ und allen Wahrscheinlichkeitsmaßen Q auf $\mathbf{R}^d$.

Ferner sei die Dichte von h beschränkt. Dann existiert eine Konstante c_1, die von c, γ, h und K abhängt, so dass für alle $s > 0$, $\tau > 0$ und $n \geq 1$ gilt

$$P^n\left(D \in X^n : \left\|h_{D,s} - h_{P,s}\right\|_\infty \leq \sqrt{\frac{c_1}{ns^d}\log\frac{c_1}{s}} + \frac{c_1}{ns^d}\log\frac{c_1}{s} + \ldots \right.$$
$$\left. \ldots + \frac{c_1\sqrt{\tau}}{\sqrt{ns^d}} + \frac{c_1\tau}{ns^d}\right)$$
$$\geq 1 - e^{-\tau}.$$

Beweis: Wir wollen den Beweis lediglich skizzieren. Ohne Wissen in empirischen Prozessen ist der Beweis sehr aufwendig. Wir betrachten $\mathscr{F}_s := \{f_{x,s} : x \in \mathbf{R}^d\}$ wie bisher, dann gilt

$$\left\|h_{D,s} - h_{P,s}\right\|_\infty = \sup_{x\in\mathbf{R}^d} \left|\mathbf{E}_D f_{x,s}\right|.$$

Mit L_∞-Überdeckungszahlen könnte man nun ähnlich wie im Beweis von Satz 2.2.15 vorgehen. Für $L_2(Q)$-Überdeckungszahlen, die im Allgemeinen einfacher zu bekommen sind, gilt dies nicht. In diesem Fall müssen wir Talagrands Ungleichung verwenden. Hierfür verwenden wir das Varianz-Bound und erhalten

$$\mathbf{E}_P f_{x,s}^2 \leq \mathbf{E}_P k_{x,s}^2 \leq s^{-d}\int K^2\left(\frac{\left\|x - x'\right\|}{s}\right) h(x') s^{-d}\, \mathrm{d}\mu(x') \leq c\left\|h\right\|_\infty s^{-d},$$

da $K \in L_\infty \cap L_1$ ist und daher auch $K \in L_2$ ist. Mit der Talagrandungleichung folgt dann, dass mit Wahrscheinlichkeit P^n nicht kleiner als $1 - e^{-\tau}$ gilt

$$\left\|h_{D,s} - h_{P,s}\right\|_\infty \leq 4\mathbf{E}_{D\sim P^n}\left\|h_{D,s} - h_{P,s}\right\|_\infty + \sqrt{\frac{2\tau c\left\|h\right\|_\infty}{ns^d}} + \frac{2\tau}{ns^d}.$$

Wir müssen nun noch $\mathbf{E}_{D\sim P^n}\left\|h_{D,s} - h_{P,s}\right\|_\infty$ abschätzen, dies ist jedoch der aufwändige Teil des Beweises. Dazu führt man zunächst eine Symmetrisierung durch und verwendet dann Dudleys Entropie-Integral. □

3
Regression

In diesem Kapitel wollen wir die Konsistenz und Konvergenzraten für Regressionsmethoden analog zum vorausgehenden Kapitel untersuchen.

3.1 Empirische Risikominimierung

Von nun an sei $L\colon X \times \mathbf{R} \to [0,\infty)$ messbar mit $Y \subset \mathbf{R}$. Wir nennen dies eine Verlustfunktion. Ferner sei P ein Wahrscheinlichkeitsmaß auf $X \times Y$. Das zur Verlustfunktion gehörende Risiko ist gegeben durch $R_{L,P}(f) := \int_{X\times Y} L(y, f(x))\, \mathrm{d}P(x,y)$. Insbesondere ist das Bayes-Risiko gegeben durch

$$R^*_{L,P} := \inf\{R_{L,P} \mid f\colon X \to \mathbf{R} \text{ messbar}\}$$

und das Überschuss-Risiko ist gegeben durch

$$R_{L,P}(f) - R^*_{L,P}.$$

Die Bayes-Entscheidungsfunktion $f^*_{L,P}\colon X \to \mathbf{R}$ erfüllt $R_{L,P}(f^*_{L,P}) = R^*_{L,P}$. Beachte jedoch, dass diese Funktion im Allgemeinen weder existiert noch eindeutig ist.

Für einen Datensatz $D \in (X \times Y)^n$ ist das empirische Risiko gegeben durch

$$R_{L,D}(f) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)),$$

wobei insbesondere $R^*_{L,D} = 0$ gilt, falls für alle $y \in Y$ ein $t \in \mathbf{R}$ mit $L(y,t) = 0$ existiert.

Definition 3.1.1 **Empirische Risikominimierung (ERM)**
Sei $\mathscr{F}$ eine Menge messbarer Funktionen $X \to \mathbf{R}$. Eine Lernmethode $\mathscr{L}$ heißt dann empirische Risikominimierung (ERM) genau dann, wenn für die von $\mathscr{L}$ erzeugten Entscheidungsfunktionen f_D gilt, dass $f_D \in \mathscr{F}$ ist und für alle $D \in (X \times Y)^n$ und $n \geq 1$ gilt

$$R_{L,D}(f_D) = \inf_{f \in \mathscr{F}} R_{L,D}(f). \tag{*}$$

Weder Existenz noch Eindeutigkeit solch einer empirischen Risikominimierung ist im Allgemeinen gegeben. Ist $\mathscr{F}$ endlich, so ist die Existenz jedoch gesichert. Die fehlende Eindeutigkeit ist in der Praxis jedoch irrelevant.

Zunächst wollen wir uns wieder mit Orakelungleichungen beschäftigen. Dazu setzen wir $R^*_{L,P,\mathscr{F}} := \inf\{R_{L,P}(f) : f \in \mathscr{F}\}$. Der Approximationsfehler ist dann gegeben durch $R^*_{L,P,\mathscr{F}} - R^*_{L,P}$.

Satz 3.1.2 **Orakelungleichung für ERM**
Sei L eine Verlustfunktion und $\mathscr{F} \subset \mathscr{L}_0(X)$ endlich, so dass es ein $B > 0$ gibt mit $L(y, f(x)) \leq B$ für alle $f \in \mathscr{F}$ und $(x, y) \in X \times Y$. Dann gilt für jede empirische Risikominimierung über $\mathscr{F}$ und alle $n \geq 1$ und $\tau > 0$ die Abschätzung

$$P^n\left(D \in (X \times Y)^n : R_{L,P}(f_D) < R^*_{L,P,\mathscr{F}} + B\sqrt{\frac{2\tau + 2\log(2|\mathscr{F}|)}{n}}\right) \geq 1 - e^{-\tau}.$$

Hierbei ist zu beachten, dass $|\mathscr{F}|$ wie bei der Histogrammregel lediglich einen logarithmischen Einfluss hat.

Beweis: Sei $f^* \in \mathscr{F}$ mit $R_{L,P}(f^*) = R^*_{L,P,\mathscr{F}}$. Damit gilt

$$\begin{aligned} R_{L,P}(f_D) - R^*_{L,P,\mathscr{F}} &= R_{L,P}(f_D) - R_{L,D}(f_D) + R_{L,D}(f_D) - R_{L,P}(f^*) \\ &\leq \left|R_{L,P}(f_D) - R_{L,D}(f_D)\right| + \left|R_{L,D}(f^*) - R_{L,P}(f^*)\right| \\ &\leq 2\sup_{f \in \mathscr{F}} \left|R_{L,P}(f) - R_{L,D}(f)\right|. \end{aligned}$$

Durch Anwenden der Hoeffding-Ungleichung und einem Union Bound erhalten wir

$$\begin{aligned} P^n\left(D : R_{L,P}(f_D) - R^*_{L,P,\mathscr{F}} \geq B\sqrt{\frac{2\tau}{n}}\right) &\leq P^n\left(D : \sup_{f\in\mathscr{F}} \left|R_{L,P}(f) - R_{L,D}(f)\right| \geq B\sqrt{\frac{\tau}{2n}}\right) \\ &\leq \sum_{f\in\mathscr{F}} P^n\left(D : \left|R_{L,P}(f) - R_{L,D}(f)\right| \geq B\sqrt{\frac{\tau}{2n}}\right) \\ &\leq \sum_{f\in\mathscr{F}} 2e^{-\tau} \\ &= 2\left|\mathscr{F}\right| e^{-\tau}. \end{aligned}$$

Durch Umformen erhalten wir dann die Aussage. □

Ist $\mathscr{F}$ unendlich, so kann der Beweis mit Hilfe von Überdeckungszahlen modifiziert werden.

Die Orakelungleichung liefert eine Konvergenzrate, die sich wie $\frac{1}{\sqrt{n}}$ verhält. Dies ist im Allgmeinen optimal, häufig jedoch nicht. Wir wollen zunächst ein einfaches Beispiel vorstellen: Sei $\mathscr{F}$ endlich mit $R^*_{L,P,\mathscr{F}} = 0$, d. h. es existiert ein $f^* \in \mathscr{F}$ mit $R_{L,P}(f^*) = 0$. Weiterhin gelten die Annahmen aus Satz 3.1.2. Dann gilt $L(y, f^*(x)) = 0$ für P-fast alle $(x, y) \in X \times Y$ und es folgt $R_{L,D}(f^*) = 0$ für P^n-fast alle $D \in (X \times Y)^n$. Dann erfüllt jede empirische Risikominimierung $R_{L,D}(f_D) = 0$. Wir betrachten nun ERM und wollen $R_{L,P}(f_D) \overset{\leq}{=} \frac{1}{n}$ zeigen. Dazu sei $r > 0$ und wir setzen $h_f := L \circ f$, $L := L(y, f(x))$ und

$$g_{f,r} := \frac{\mathbf{E}_P h_f - h_f}{\mathbf{E}_P h_f + r}.$$

Unser Ziel ist es, auf diese Funktion die Bernstein-Ungleichung und den Union Bound anzuwenden. Für $f \in \mathscr{F}$ mit $\mathbf{E}_P h_f \neq 0$ gilt

$$\mathbf{E} g^2_{f,r} \leq \frac{\mathbf{E}_P h_f^2}{(\mathbf{E}_P h_f + r)^2} \leq \frac{\mathbf{E}_P h_f^2}{2\mathbf{E}_P h_f \cdot r} \leq \frac{B\mathbf{E}_P h_f}{2\mathbf{E}_P h_f \cdot r} \leq \frac{B}{2r}.$$

Für $f \in \mathscr{F}$ mit $\mathbf{E}_P h_f = 0$ gilt

$$\mathbf{E} g^2_{f,r} \leq \frac{\mathbf{E}_P h_f^2}{(\mathbf{E}_P h_f + r)^2} \leq \frac{B\mathbf{E}_P h_f}{(\mathbf{E}_P h_f + r)^2} = 0 \leq \frac{B}{2r}.$$

Ferner gilt

$$\left\|\mathbf{E} g_{f,r}\right\|_\infty = \frac{\left\|\mathbf{E}_P h_f - h_f\right\|_\infty}{\mathbf{E}_P h_f + r} \leq \frac{B}{r}$$

und $\mathbf{E}_P g_{f,r} = 0$. Damit können wir die Bernstein-Ungleichung und Union Bound anwenden und erhalten

$$P^n\left(D : \sup_{f\in\mathscr{F}} \mathbf{E}_D g_{f,r} \geq \sqrt{\frac{B\tau}{nr}} + \frac{2B\tau}{3nr}\right) \leq \left|\mathscr{F}\right| e^{-\tau}.$$

Da $f_D \in \mathscr{F}$ ist, folgt

$$P^n\left(D : \mathbf{E}_P h_{f_D} - \mathbf{E}_D h_{f_D} \geq \left(\mathbf{E}_P h_{f_D} + r\right)\left(\sqrt{\frac{B\tau}{nr}} + \frac{2B\tau}{3nr}\right)\right) \leq |\mathscr{F}|\, e^{-\tau}.$$

Für solche Datensätze D wie in der Ungleichung gilt

$$\left(1 - \sqrt{\frac{B\tau}{nr}} - \frac{2B\tau}{3nr}\right)\mathbf{E}_P h_{f_D} \geq r\left(\sqrt{\frac{B\tau}{nr}} + \frac{2B\tau}{3nr}\right).$$

Wir setzen nun $r := \frac{4B\tau}{n}$ und erhalten dann

$$R_{L,P}(f_D) \geq \frac{8B\tau}{n}.$$

Lemma 3.1.3
Für $q \in (1,\infty)$ setzen wir $q' \in (1,\infty)$ durch $q^{-1} + q'^{-1} = 1$. Für $a, b \geq 0$ gelten dann folgende Eigenschaften:

i) $ab \leq \dfrac{a^q}{q} + \dfrac{b^{q'}}{q'}$.

ii) $q^{\frac{1}{q}} a^{\frac{2}{q}} (q')^{\frac{1}{q'}} b^{\frac{2}{q'}} \leq (a+b)^2$.

Beweis: Ohne Einschränkung sei $a > 0$. Wir setzen nun

$$h_a(b) := \frac{a^q}{q} + \frac{b^{q'}}{q'} - ab$$

für $b \geq 0$. Dann gilt $h_a'(b) = b^{q'-1} - a$ und h_a hat ein globales Minimum bei $b^* = a^{\frac{1}{q'-1}}$. Dann folgt

$$h_a(b^*) = \frac{a^q}{q} + \frac{a^{\frac{q'}{q'-1}}}{q'} - a^{1+\frac{1}{q'-1}} = a^q - a^q = 0.$$

Damit folgt die erste Aussage. Die zweite Aussage folgt hieraus für $\tilde{a} := (qa^2)^{\frac{1}{q}}$ und $\tilde{b} := (q'b^2)^{\frac{1}{q'}}$, sowie $a^2 + b^2 \leq (a+b)^2$. □

Satz 3.1.4 **Verbesserte Orakelungleichung für ERM**
Sei $L : X \times \mathbf{R} \to [0,\infty)$ eine Verlustfunktion, $\mathscr{F} \subset \mathscr{L}_0(x)$ endlich und P eine Verteilung auf $X \times Y$, so dass es ein $f_{L,P}^*$, also mit $R_{L,P}(f_{L,P}^*) = R_{L,P}^*$, gibt. Ferner sollen Konstanten $B > 0$, $\theta \in [0,1]$ und $V \geq B^{2-\theta}$ existieren mit

i) $\left\| L \circ f - L \circ f_{L,P}^* \right\|_\infty \leq B$

ii) $\mathbf{E}_P\left(L\circ f - L\circ f^*_{L,P}\right)^2 \le V\left(\mathbf{E}_P(L\circ f - L\circ f^*_{L,P})\right)^\theta$

für alle $f\in\mathscr{F}$. Dann gilt für ERM über $\mathscr{F}$

$$P^n\Bigg(D\in(X\times Y)^n : R_{L,P}(f_D) - R^*_{L,P} < 6(R^*_{L,P,\mathscr{F}} - R^*_{L,P}) + \ldots \\ \ldots + 4\left(\frac{8V(\tau+\log(1+|\mathscr{F}|))}{n}\right)^{\frac{1}{2-\theta}}\Bigg) \\ \ge 1-e^{-\tau}$$

für alle $n\ge 1$ und $\tau>0$.

Der Faktor 6 vor dem Approximationsfehler wird uns nicht stören, da dieser später gegen 0 konvergiert. Ferner sind die Raten von der Ordnung $n^{-\frac{1}{2-\theta}}$, also zwischen $\frac{1}{\sqrt{n}}$ und $\frac{1}{n}$.

Die Bedingung ii) heißt *Variance Bound* und wird benutzt, um die Varianzabschätzung zur Anwendung der Bernsteinungleichung zu erhalten. Sie gilt immer für $\theta=0$.

Historisch wurde das Beispiel vor Satz 3.1.4 lange als Kuriosum angesehen. Dies änderte sich erst ab 1996 und führte dann auf *fast-rates*-Arbeiten.

Beweis: Für $n<8\tau$ ist nichts zu beweisen, da dann

$$R_{L,P}(f_D) - R^*_{L,P} \le B \le V\frac{1}{2-\theta} \le \left(\frac{8V\tau}{n}\right)^{\frac{1}{2-\theta}}$$

gilt. Im Folgenden sei daher $n\ge 8\tau$. Für $f\in\mathscr{F}$ setzen wir $h_f = L\circ f - L\circ f^*_{L,P}$. Ferner sei $f_0\in\mathscr{F}$ mit $R_{L,P}(f_0) = R^*_{L,P,\mathscr{F}}$. Da $R_{L,D}(f_D)\le R_{L,D}(f_0)$ nach Definition der empirischen Risikominimierung gilt, folgt für alle $D\in(X\times Y)^n$

$$\begin{aligned} R_{L,P}(f_D) - R_{L,P}(f_0) &= \mathbf{E}_P h_{f_D} - \mathbf{E}_P h_{f_0} \\ &\le \underbrace{\mathbf{E}_P h_{f_D} - \mathbf{E}_D h_{f_D}}_{=:\mathrm{I}} + \underbrace{\mathbf{E}_D h_{f_0} - \mathbf{E}_P h_{f_0}}_{=:\mathrm{II}}. \end{aligned}$$

Wir kümmern uns zunächst um eine Abschätzung für II. Im ersten Fall sei $\theta>0$, nach ii) gilt dann

$$\mathbf{E}_P\left(h_{f_0} - \mathbf{E}_P h_{f_0}\right)^2 \le \mathbf{E}_P h^2_{f_0} \le V\left(\mathbf{E}_P h_{f_0}\right)^\theta.$$

Für $q := \frac{2}{2-\theta}$, $q' := \frac{2}{\theta}$, $a := \left(\frac{2^{1-\theta}\theta^\theta V\tau}{n}\right)^{\frac{1}{2}}$ und $b := \left(\frac{2\mathbf{E}_P h_{f_0}}{\theta}\right)^{\frac{\theta}{2}}$ gilt nach Lemma 3.1.3

$$\begin{aligned} \sqrt{\frac{2\tau V\left(\mathbf{E}_P h_{f_0}\right)^\theta}{n}} &\le \left(1-\frac{\theta}{2}\right)\left(\frac{2^{1-\theta}\theta^\theta V\tau}{n}\right)^{\frac{1}{2-\theta}} + \mathbf{E}_P h_{f_0} \\ &\le \left(\frac{2V\tau}{n}\right)^{\frac{1}{2-\theta}} + \mathbf{E}_P h_{f_0}. \end{aligned}$$

Da ferner $\|h_{f_0} - \mathbf{E}h_{f_0}\|_\infty \leq 2B$ gilt, folgt mit der Ungleichung von Bernstein

$$P^n\left(D \in (X \times Y)^n : \mathbf{E}_D h_{f_0} - \mathbf{E}_P h_{f_0} < \mathbf{E}_P h_{f_0} + \left(\frac{2V\tau}{n}\right)^{\frac{1}{2-\tau}} + \frac{4B\tau}{3n}\right) \geq 1 - e^{-\tau}. \tag{*}$$

Im zweiten Fall sei $\theta = 0$. Dann folgt obige Abschätzung unmittelbar aus der Hoeffding-Ungleichung und $\|h_{f_0}\|_\infty \leq B \leq \sqrt{V} = V^{\frac{1}{2-\theta}}$.

Wir wollen uns nun um den Ausdruck I kümmern. Dazu definieren wir

$$g_{f,r} = \frac{\mathbf{E}_P h_f - h_f}{\mathbf{E}_P h_f + r}$$

für $f \in \mathscr{F}$ und $r > 0$. Unser Ziel ist es, auf $g_{f,r}$ die Bernstein-Ungleichung anzuwenden. Dazu müssen wir die entsprechenden Voraussetzungen verifizieren. $\mathbf{E}_P g_{f,r} = 0$ ist klar und es gilt

$$\|g_{f,r}\|_\infty \leq \frac{\|\mathbf{E}_P h_f - h_f\|_\infty}{r} \leq \frac{2B}{r}.$$

Für $\theta > 0$ gilt ferner mit Lemma 3.1.3

$$\begin{aligned}\mathbf{E}_P g_{f,r}^2 &\leq \frac{\mathbf{E}_P h_f^2}{(\mathbf{E}_P h_f + r)^2} \leq \frac{(2-\theta)^{\frac{2-\theta}{2}}\theta^{\frac{\theta}{2}}\mathbf{E}_P h_f^2}{2r^{2-\theta}(\mathbf{E}_P h_f)^\theta} \\ &\leq V r^{\theta-2},\end{aligned}$$

für $\theta = 0$ gilt $\mathbf{E}_P g_{f,r}^2 \leq \frac{\mathbf{E}_P h_f^2}{r^2} \leq \frac{V}{r^2} = V r^{\theta-2}$. Damit können wir die Bernstein-Ungleichung anwenden und erhalten

$$P^n\left(D \in (X \times Y)^n : \sup_{f \in \mathscr{F}} \mathbf{E}_P g_{f,r} < \sqrt{\frac{2V\tau}{nr^{2-\tau}}} + \frac{4B\tau}{3nr}\right) \geq 1 - |\mathscr{F}|e^{-\tau}. \tag{**}$$

Sei nun D ein Datensatz, der die Bedingungen in (*) und (**) erfüllt. Die Wahrscheinlichkeit für ein soches D ist $\geq 1 - (1 + |\mathscr{F}|)e^{-\tau}$. Aus (**) und $f_D \in \mathscr{F}$ folgt

$$\mathbf{E}_P h_{f_D} - \mathbf{E}_D h_{f_D} < \mathbf{E}_P h_{f_D}\left(\frac{2V\tau}{nr^{2-\theta}} + \frac{4B\tau}{3nr}\right) + \sqrt{\frac{2V\tau r^\theta}{n}} + \frac{4B\tau}{3n}.$$

Kombinieren wir dies mit (*) und (**), so erhalten wir

$$\mathbf{E}_P h_{f_D} - \mathbf{E}_P h_{f_0} < \mathbf{E}_P h_{f_0} + \mathbf{E}_P h_{f_D}\left(\sqrt{\frac{2V\tau}{nr^{2-\theta}}} + \frac{4B\tau}{3nr}\right) + \sqrt{\frac{2V\tau r^\theta}{n}} + \left(\frac{2V\tau}{n}\right)^{\frac{1}{2-\theta}} + \frac{8B\tau}{3n}.$$

Wir setzen nun $r := \left(\frac{8V\tau}{n}\right)^{\frac{1}{2-\theta}}$ und erhalten damit $\sqrt{\frac{2V\tau r^\theta}{n}} = \frac{1}{2}$, beziehungsweise äquivalent hierzu $\sqrt{\frac{2V\tau r^\theta}{n}} = \frac{r}{2}$. Damit erhalten wir

$$\frac{4B\tau}{3nr} = \frac{1}{6}\frac{8\tau}{n}\frac{B}{r} \leq \frac{1}{6}\left(\frac{8\tau}{n}\right)^{\frac{1}{2-\theta}}\frac{V^{\frac{1}{2-\theta}}}{r} = \frac{1}{6}$$

und nach der letzten Abschätzung damit $\frac{8B\tau}{3n} \leq \frac{r}{3}$. Ferner gilt $2 \leq 4^{\frac{1}{2-\theta}}$ und damit

$$\left(\frac{2V\tau}{n}\right)^{\frac{1}{2-\theta}} = \left(\frac{8V\tau}{n}\right)^{\frac{1}{2-\tau}} \left(\frac{1}{4}\right)^{\frac{1}{2-\theta}} \leq \frac{r}{2}.$$

Damit folgt

$$\mathbf{E}_P h_{f_D} < 2\mathbf{E}_P h_{f_0} + \mathbf{E}_P h_{f_D} \left(\frac{1}{2} + \frac{1}{6}\right) + \frac{r}{2} + \frac{r}{2} + \frac{r}{3} = \frac{4}{3}\left(\frac{8V\tau}{n}\right)^{\frac{1}{2-\theta}}.$$

Einfaches Umformen und eine Transformation in τ führen dann zur Behauptung. □

3.2 Eigenschaften der Least-Squares-Verlustfunktion

In der Vorlesung zu den stochastischen Prozessen haben wir faktorisierte bedingte Erwartungen $\mathbf{E}(Y \mid X = x)$ und Wahrscheinlichkeiten $P(A \mid X = x)$ kennengelernt. Wieder ist zum Beispiel $P(\cdot \mid X = x)$ im Allgemeinen kein Wahrscheinlichkeitsmaß. Dieses Problem wird durch den folgenden Satz behoben.

Satz 3.2.1 **Reguläre bedingte Erwartungen**
Sei $(X, \mathscr{A})$ ein Messraum und $\mathscr{B}$ die Borelsche σ-Algebra auf einer abgeschlossenen Menge $Y \subset \mathbf{R}$. Dann gibt es zu jedem Wahrscheinlichkeitsmaß P auf $X \times Y$ eine Abbildung $P(\cdot \mid \cdot)\colon \mathscr{B} \times X \to [0,1]$ mit den folgenden Eigenschaften:

i) $P(\cdot \mid X)\colon \mathscr{B} \to [0,1]$ ist ein Wahrscheinlichkeitsmaß für alle $x \in X$.

ii) $P(B \mid \cdot)\colon X \to [0,1]$ ist messbar für alle $B \in \mathscr{B}$.

iii) Für alle $A \in \mathscr{A}$ und $B \in \mathscr{B}$ gilt

$$P(A \times B) = \int_A P(B \mid x)\, \mathrm{d}P_X(x), \tag{*}$$

wobei das Wahrscheinlichkeitsmaß P_X auf X durch $P_X(A) := P(A \times Y)$ mit $A \in \mathscr{A}$ definiert ist.

Beweis: Der Beweis verwendet die Tatsache, dass Y ein vollständiger, separabler metrischer Raum ist. Wir werden den Beweis an dieser Stelle nicht führen. □

Die Gleichung (*) ermöglicht es, P zu desintegrieren, d. h. die Integration bezüglich P in zwei Integrationsschritte aufzuspalten. Genauer gilt

$$\mathbf{E}_P f = \int_{X \times Y} f \, \mathrm{d}P = \int_X \int_Y f(x,y)\, P(\mathrm{d}y \mid y)\, \mathrm{d}P_X(x). \tag{**}$$

Ist $P = P_X \otimes P_Y$ ein Produktmaß, so gilt $P(\cdot \mid x) = P_Y$ für P_X-fast alle $x \in X$ und (**) wird zum Satz von Fubini.

$P(\cdot \mid \cdot)$ ist P_X-fast sicher eindeutig und wird reguläre bedingte Wahrscheinlichkeit genannt.

Im Folgenden werden wir einige Notationen und Annahmen verwenden, die wir nun einführen wollen. Zunächst sei L stets die Least-Squares-Verlustfunktion und P ein Wahrscheinlichkeitsmaß auf $X \times Y$ für $Y \subset \mathbf{R}$ abgeschlossen. Das Maß P soll zudem folgende Bedingungen erfüllen:

- $|P|_2^2 := \int_X \int_Y y^2 P(\mathrm{d}y \mid x)\, \mathrm{d}P_X(x) < \infty$ – mit anderen Worten ist das durchschnittliche zweite Moment also endlich. Diese Bedingung ist für beschränkte Y offenbar erfüllt.

- Die Abbildung $\mathbf{E}(Y \mid x) := \int_Y y P(\mathrm{d}y \mid x)$ für $x \in X$ heißt reguläre bedingte Erwartung. Sie ist messbar und wie wir gleich sehen in $L_2(P_X)$. Ferner ist sie P_X-fast sicher eindeutig.

Lemma 3.2.2 **Bayes-Entscheidungsfunktion**
Unter den obigen Voraussetzungen gelten für $f^*_{L,P}(x) := \mathbf{E}(Y \mid x)$ mit $x \in X$ die folgenden Aussagen:

i) $R_{L,P}(f^*_{L,P}) = R^*_{L,P}$

ii) Ist $f \in \mathscr{L}_0(P_X)$ mit $R_{L,P}(f) = R^*_{L,P}$, so gilt P_X-fast sicher $f = f^*_{L,P}$.

Mit anderen Worten ist $f^*_{L,P}$ die einzige Bayes-Entscheidungsfunktion bezüglich L für die Verteilung P.

Beweis: Für $f \in \mathscr{L}_0(P_X)$ gilt

$$R_{L,P}(f) = \int_X \int_Y (y - f(x))^2 \, P(\mathrm{d}y \mid x) \, P_X(\mathrm{d}x).$$

Falls wir ein f finden, für welches das innere Integral für jedes x minimal ist, so muss diese Funktion eine Bayes-Entscheidungsfunktion sein. Es gilt

$$\begin{aligned}\int_Y (y - f(x))^2 \, P(\mathrm{d}y \mid x) &= \int_Y y^2 \, P(\mathrm{d}y \mid x) - 2f(x) \int_Y y \, P(\mathrm{d}y \mid x) + f^2(x) \\ &=: a - tb + t^2.\end{aligned}$$

Diese Funktion in t ist minimal für $t^* = b$. Mit anderen Worten minimiert $f(x) = \int_Y y \, P(\mathrm{d}y \mid x) = \mathbf{E}(Y \mid x)$ das innere Integral, woraus wir die erste Behauptung folgern können. Da $t^* = b$ der einzige Minimierer ist, folgt auch die zweite Aussage. □

Das nächste Ziel ist es nun, das Überschussrisiko $R_{L,P}(f) - R^*_{L,P}$ genauer zu bestimmen.

Lemma 3.2.3 **Überschussrisiko**
Unter den obigen Voraussetzungen gilt

$$R_{L,P}(f) - R^*_{L,P} = \left\| f - f^*_{L,P} \right\|^2_{L_2(P_X)}.$$

Eine Funktion f mit kleinem Überschussrisiko ist also ein guter Schätzer für $f^*_{L,P}$. Die Least-Squares-Regression ist also eine Schätzung für $\mathbf{E}(Y \mid x)$. Beachte ferner $f^*_{L,P} \in L_2(P_X)$.

Beweis: Unter Verwendung von $\mathbf{E}(Y \mid x) = f^*_{L,P}$ gilt

$$\begin{aligned} R_{L,P}(f) - R^*_{L,P} &= \int_{X\times Y} y^2 - 2yf(x) + f^2(x) \, \mathrm{d}P(x,y) - \int_{X\times Y} y^2 - 2yf^*_{L,P} + (f^*_{L,P})^2(x) \, \mathrm{d}P(x,y) \\ &= \int_X f^2(x) - (f^*_{L,P})^2(x) - 2f(x)\mathbf{E}(Y \mid x) + 2f^*_{L,P}(x)\mathbf{E}(Y \mid x) \, \mathrm{d}P_X(x) \\ &= \int_X (f(x) - f^*_{L,P}(x))^2 \, \mathrm{d}P_X(x). \end{aligned}$$

□

Lemma 3.2.4 **Supremum-/Varianz-Bounds**
Für $M > 0$ und $Y \subset [-M,M]$ abgeschlossen und $f : X \to [-M,M]$ messbar gelten die folgenden Aussagen:

i) $\left\| L \circ f - L \circ f^*_{L,P} \right\|_\infty \leq 4M^2$

ii) $\mathbf{E}_P \left(L \circ f - L \circ f^*_{L,P} \right)^2 \leq 16M^2 \mathbf{E}_P \left(L \circ f - L \circ f^*_{L,P} \right)$

Wir haben ein Supremum-Bound im Sinne von Satz 3.1.4 für $B = 4M^2$ und einen Varianz-Bound für $\theta = 1$ und $V = 16M^2$. Insbesondere bekommen wir dann im Satz 3.1.4 eine Rate der Ordnung $\frac{1}{n}$.

Beweis: Für i) betrachten wir $L(y, f(x)) \in [0, 4M^2]$ und $L(y, f^*_{L,P}(x)) \in [0, 4M^2]$. Dann folgt

$$\left| L(y, f(x)) - L(y, f^*_{L,P}(x)) \right| \leq 4M^2.$$

Für ii) gilt durch etwas Rechnerei und einer analogen Abschätzung dann

$$\begin{aligned} \left(L \circ f - L \circ f^*_{L,P} \right)^2 &= \left(f(x) + f^*_{L,P}(x) - 2y \right)^2 \left(f(x) - f^*_{L,P}(x) \right)^2 \\ &\leq 16M^2 \left(f(x) - f^*_{L,P}(x) \right)^2. \end{aligned}$$

Damit folgt mit Lemma 3.2.3

$$\begin{aligned}\mathbf{E}_P\left(L\circ f - L\circ f^*_{L,P}\right)^2 &\leq 16M^2\left\|f - f^*_{L,P}\right\|^2_{L_2(P_X)}\\ &\leq 16M^2\left(R_{L,P}(f) - R^*_{L,P}\right)\\ &\leq 16M^2\mathbf{E}_P\left(L\circ f - L\circ f^*_{L,P}\right).\end{aligned}$$

□

3.3 Histogrammregel für Regression

Definition 3.3.1
Sei $\mathscr{A} = (A_j)_{j \geq 1}$ eine Würfelpartition der Weite s und Q ein Wahrscheinlichkeitsmaß auf $\mathbf{R}^d \times \mathbf{R}$ mit $|Q|_2 < \infty$. Die Funktion

$$\begin{aligned} h_{Q,s} \colon \mathbf{R}^d &\to \mathbf{R} \\ x &\mapsto \sum_{j=1}^{\infty} \mathbf{1}_{A_j}(x) \cdot \frac{1}{Q(A_j \times R)} \int_{A_j \times \mathbf{R}} y \, \mathrm{d}Q(x,y) \end{aligned}$$

heißt $\mathscr{A}$-Histogramm von Q. Dabei vereinbaren wir $\frac{0}{0} := 0$.

Ist D das empirische Maß bezüglich $((x_1, y_1), \ldots, (x_n, y_n))$, so gilt

$$D(A_j \times \mathbf{R}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{A_j}(x_i)$$

und

$$\int_{A_j} y \, \mathrm{d}D(x,y) = \frac{1}{n} \sum_{i=1}^{n} y_i \mathbf{1}_{A_j}(x_i).$$

Damit erhalten wir dann das empirische Histogramm

$$h_{D,s}(x) = \frac{\sum_{i=1}^{n} y_i \mathbf{1}_{A_j}(x_i)}{\sum_{i=1}^{n} \mathbf{1}_{A_j}(x_i)}$$

für $x \in A_j$.

Lemma 3.3.2
Sei $\mathscr{A}$ eine Würfelpartition der Weite s, dann gelten die folgenden Aussagen:

i) $R_{L,P}(h_{D,s}) - R_{L,P}(h_{P,s}) = \left\| h_{D,s} - h_{P,s} \right\|^2_{L_2(D_x)}$

ii) $h_{P,s} = \sum_{j=1}^{\infty} \mathbf{1}_{A_j} \frac{1}{P_X(A_j)} \int_{A_j} f^*_{L,P}(x) \, \mathrm{d}P_X(x)$

Beweis: Wir werden den Beweis an dieser Stelle nicht führen. □

Satz 3.3.3 **Histogrammregel ist empirische Risikominimierung**
Betrachte die Menge $\mathscr{F} := \left\{\sum_{j=1}^{\infty} c_j \mathbf{1}_{A_j} : c_j \in \mathbf{R}\right\}$ aller Funktionen, die auf jeder Zelle konstant sind. Dann gilt für jedes Wahrscheinlichkeitsmaß Q mit $|Q|_2 < \infty$

$$R_{L,Q}(h_{Q,s}) = \inf_{f \in \mathscr{F}} R_{L,Q}(f).$$

Beweis: Für $f \in \mathscr{F}$ gilt

$$R_{L,Q}(f) = \sum_{j=1}^{\infty} \int_{A_j \times \mathbf{R}} (y - f(x))^2 \, \mathrm{d}Q(x,y) = \sum_{j=1}^{\infty} \int_{A_j \times \mathbf{R}} (y - c_j)^2 \, \mathrm{d}Q(x,y).$$

Für $j \geq 1$ ist nun

$$\begin{aligned} h(c_j) &:= \int_{A_j \times \mathbf{R}} (y - c_j)^2 \, \mathrm{d}Q(x,y) \\ &= \int_{A_j \times \mathbf{R}} y^2 \, \mathrm{d}Q(x,y) - 2c_j \int_{A_j \times \mathbf{R}} y \, \mathrm{d}Q(x,y) + c_j^2 Q(A_j \times \mathbf{R}). \end{aligned}$$

Minimieren wir h, so erhalten wir ein eindeutiges, globales Minimum bei

$$c_j^* = \frac{\int_{A_j \times \mathbf{R}} y \, \mathrm{d}Q(x,y)}{Q(A_j \times \mathbf{R})}.$$

Damit minimiert die Histogrammregel das Risiko über $\mathscr{F}$ auf jeder Zelle. □

Ist mehr über Q bekannt, so kann ein kleineres $\mathscr{F}$ betrachtet werden. Ist zum Beispiel $Q([-1,1]^d \times [-1,1]) = 1$, so genügt es, die Menge

$$\mathscr{F} := \left\{ \sum_{A_j \cap [-1,1]^d \neq \emptyset} c_j \mathbf{1}_{A_j} : c_j \in [-1,1] \right\}$$

zu betrachten.

Gilt $P([-1,1]^d \times [-1,1]) = 1$, so gilt für alle $D \in \mathbf{R}^d \times \mathbf{R}$ ebenfalls $D([-1,1]^d \times [-1,1]) = 1$. Damit können wir wieder obige Funktionenklasse betrachten.

Im Folgenden werden wir nur solche P betrachten, da wir bei der Dichteschätzung schon gesehen haben, wie man mit X-Tails umgeht. Die technische Arbeit wollen wir uns hier daher ersparen.

Satz 3.3.4 **Approximationsfehler für Histogrammregel**
Sei P ein Wahrscheinlichkeitsmaß mit $|P|_2 < \infty$. Dann gibt es zu jedem $\varepsilon > 0$ ein $s_\varepsilon > 0$, so dass für alle $s \in (0, s_\varepsilon]$ und alle Würfelpartitionen der Weite s

$$R_{L,P}(h_{P,s}) - R_{L,P}^* \leq \varepsilon$$

gilt. Ist $f^*_{L,P}$ zudem α-Hölderstetig mit der Konstanten c, so gilt für alle Würfelpartitionen der Weite $s \geq 0$

$$\left\| h_{P,s} - f^*_{L,P} \right\|_{L_\infty(P_X)} \leq cs^\alpha.$$

Beweis: Für eine Würfelpartition $\mathscr{A}$ betrachten wir $\mathscr{F}_\mathscr{A} := \left\{ \sum_{j=1}^\infty c_j \mathbf{1}_{A_j} : c_j \in \mathbf{R} \right\}$. In Satz 2.1.7 haben wir gesehen: Für alle $h \in L_1^+(P_X)$ und $\varepsilon > 0$ existiert ein $s_\varepsilon > 0$, so dass für alle $s \in (0, s_\varepsilon]$ und Würfelpartitionen der Weite s ein $f \in \mathscr{F}_\mathscr{A}$ existiert, so dass $\|f - h\|_{L_1(P_X)} \leq \varepsilon$ gilt. Klar ist, dass die Voraussetzung $h \geq 0$ durch die übliche Zerlegung $h = h^+ - h^-$ fallengelassen werden kann. Sei nun $f^*_{L,P} \in L_2(P_X)$. Mit Satz 2.1.6 folgt dann, dass ein $h \in C_c(\mathbf{R}^d)$ mit $\left\| h - f^*_{L,P} \right\|_{L_2(P_X)} \leq \varepsilon$ existiert. Klar ist, dass $h \in L_1(P_X) \cap L_\infty$ gilt. Sei nun $\varepsilon > 0$, dann existiert mit obigen Überlegungen ein $s_\varepsilon > 0$, so dass für alle $s \in (0, s_\varepsilon]$ und alle Würfelpartitionen $\mathscr{A}$ der Weite s ein $f \in \mathscr{F}_\mathscr{A}$ existiert, so dass

$$\|f - h\|_{L_1(P_X)} \leq \frac{\varepsilon^2}{2\|h\|_\infty}$$

gilt. Ohne Einschränkung nehmen wir $\|f\|_\infty \leq \|h\|_\infty$ an, andernfalls betrachte man $\tilde{f} := \max\{-\|h\|_\infty, \min\{\|h\|_\infty, f\}\} \in \mathscr{F}_\mathscr{A}$. Damit erhalten wir

$$\begin{aligned}
\|f - h\|^2_{L_2(P_X)} &= \int |f - h|^2 \, \mathrm{d}P_X \\
&\leq \|f - h\|_\infty \|f - h\|_{L_1(P_X)} \\
&\leq 2\|h\|_\infty \|f - h\|_{L_1(P_X)} \\
&\leq \varepsilon^2.
\end{aligned}$$

Damit folgt $\left\| f^*_{L,P} - f \right\|_{L_2(P_X)} \leq 2\varepsilon$ und mit Satz 3.3.3 daher

$$\begin{aligned}
R_{L,P}(h_{P,s}) - R^*_{L,P} &\leq R_{L,P}(f) - R^*_{L,P} \\
&= \left\| f - f^*_{L,P} \right\|^2_{L_2(P_X)} \\
&\leq 4\varepsilon^2.
\end{aligned}$$

Für die zweite Aussage sei A_j nun eine Zelle mit $P_X(A_j) > 0$ und $x \in A_j$. Dann gilt mit Lemma 3.3.2

$$\begin{aligned}
\left| f^*_{L,P}(x) - h_{P,s}(x) \right| &= \left| f^*_{L,P}(x) - \frac{1}{P_X(A_j)} \int_{A_j} f^*_{L,P}(x') \, \mathrm{d}P_X(x') \right| \\
&= \frac{1}{P_X(A_j)} \left| \int_{A_j} f^*_{L,P}(x) - f^*_{L,P}(x') \, \mathrm{d}P_X(x') \right| \\
&\leq \frac{1}{P_X(A_j)} P_X(A_j) cs^\alpha \\
&= cs^\alpha.
\end{aligned}$$

□

Satz 3.3.5 **Least-Squares-ERM über unendlichen Mengen**
Sei $\mathscr{F} \subset \mathscr{L}_\infty(X)$ mit $\|f\|_\infty \le 1$ für alle $f \in \mathscr{F}$. Ferner existiere eine empirische Risikominimierung bezüglich der Least-Squares-Verlustfunktion. Die resultierenden Entscheidungsfunktionen werden mit f_D bezeichnet. Dann gilt für alle $n \ge 1$, $\tau > 0$ und $\varepsilon > 0$ mit Wahrscheinlichkeit $P^n(\ldots) \ge 1 - e^{-\tau}$

$$R_{L,P}(f_D) - R^*_{L,P} \le 5\left(R^*_{L,P,\mathscr{F}} - R^*_{L,P}\right) + 43\varepsilon + \frac{234(\tau + \log(1 + \mathscr{N}(\mathscr{F}, \|\cdot\|_\infty, \varepsilon)))}{n},$$

falls $P(\mathbf{R}^d \times [-1,1]) = 1$ ist.

Beachte, dass Satz 3.1.4 und Lemma 3.2.4 für endliche $\mathscr{F}$ ergeben haben, dass

$$R_{L,P}(f_D) - R^*_{L,P} \le 6\left(R^*_{L,P,\mathscr{F}} - R^*_{L,P}\right) + \frac{512(\tau + \log(1 + |\mathscr{F}|))}{n}$$

ist. Daher liefert Satz 3.3.5 für endliche $\mathscr{F}$ bessere Konstanten als Satz 3.1.4. Satz 3.3.5 lässt sich desweiteren ohne Probleme auf die Situation in Satz 3.1.4 verallgemeinern, falls die Verlustfunktion L im zweiten Argument lipschitz-stetig auf $[-1,1]$ mit einer von y unabhängigen Konstante ist.

Beweis: Wir betrachten $h_f := L \circ f - L \circ f^*_{L,P}$ für $f \in \mathscr{F}$. Wir führen nun einige Zwischenschritte durch:

i) Für f, f' mit $\|f\|_\infty \le 1$ und $\|f'\|_\infty \le 1$ gilt

$$\begin{aligned} \|h_f - h_{f'}\|_{L_\infty(P)} &= \|L \circ f - L \circ f'\|_\infty \qquad (*) \\ &= \sup_{\substack{x \in X \\ y \in [-1,1]}} \left|(y - f(x))^2 - (y - f'(x))^2\right| \\ &= \sup\left|-2yf(x) + 2yf'(x) + f^2(x) - (f'(x))^2\right| \\ &= \sup 2|y|\left|f(x) - f'(x)\right| + \sup\left|f(x) + f'(x)\right|\left|f(x) - f'(x)\right| \\ &\le 4\|f - f'\|_\infty. \end{aligned}$$

ii) Sei nun $\mathscr{C}$ ein ε-Netz von $\mathscr{F}$ bezüglich der $\|\cdot\|_\infty$-Norm. Dann gibt es zu jedem $D \in (X \times Y)^n$ ein $f \in \mathscr{C}$ mit $\|f_D - f\|_\infty \le \varepsilon$. Zusammen mit dem ersten Schritt folgt dann

$$\begin{aligned} \mathbf{E}_P h_{f_D} - \mathbf{E}_D h_{f_D} &\le \mathbf{E}_P h_f - \mathbf{E}_D h_f + \|h_{f_D} - h_f\|_{L_1(P)} + \|h_f - h_{f_D}\|_{L_1(D)} \\ &\le \mathbf{E}_P h_f - \mathbf{E}_D h_f + 8\varepsilon. \end{aligned}$$

iii) Wir betrachten nun $g_{f,r} := \frac{\mathbf{E}_P h_f - h_f}{\mathbf{E}_P h_f + r}$ für $f \in \mathscr{C}$ und $r > 0$. Im Beweis von Satz 3.1.4 haben wir gesehen, dass

$$P^n\left(D \in (X \times Y)^n : \sup_{f \in \mathscr{C}} \mathbf{E}_D g_{f,r} < \sqrt{\frac{2V\tau}{nr^{2-\theta}}} + \frac{4B\tau}{3nr}\right) \ge 1 - |\mathscr{C}|e^{-\tau}$$

gilt, wobei wir in Lemma 3.2.4 gesehen haben, dass $B = 4$, $V = 16$ und $\theta = 1$ ist. Sei nun D ein solcher Datensatz, dann gilt

$$\sup_{f \in \mathscr{C}} \frac{\mathbf{E}_P h_f - \mathbf{E}_D h_f}{\mathbf{E}_P h_f + r} < \sqrt{\frac{32\tau}{nr}} + \frac{16\tau}{3nr}.$$

Damit erhalten wir durch Umstellen

$$\mathbf{E}_P h_f - \mathbf{E}_D h_f < \left(\mathbf{E}_P h_f + r\right)\left(\sqrt{\frac{32\tau}{nr}} + \frac{16\tau}{3nr}\right)$$

für alle $f \in \mathscr{C}$. Setzen wir nun $r = \frac{128\tau}{n}$, so erhalten wir $\sqrt{\frac{32\tau}{nr}} + \frac{16\tau}{3nr} = \frac{1}{2} + \frac{1}{24} = \frac{13}{24}$ und daher

$$\mathbf{E}_P h_f - \mathbf{E}_D h_f < \frac{13}{24}\mathbf{E}_P h_f + \frac{208}{3}\frac{\tau}{n}$$

für alle $f \in \mathscr{C}$. Durch Anwenden des zweiten Schrittes erhalten wir

$$\begin{aligned}
\mathbf{E}_P h_{f_D} - \mathbf{E}_D h_{f_D} &< \frac{13}{24}\mathbf{E}_P h_f + \frac{208}{3}\frac{\tau}{n} + 8\varepsilon \\
&\leq \frac{13}{24}\left(\mathbf{E}_P h_{f_D} + 4\varepsilon\right) + \frac{208}{3}\frac{\tau}{n} + 8\varepsilon \\
&\leq \frac{13}{24}\mathbf{E}_P h_{f_D} + \frac{208}{3}\frac{\tau}{n} + \frac{117}{6}\varepsilon.
\end{aligned}$$

iv) Sei nun $f_0 \in \mathscr{F}$. Im Beweis von Satz 3.1.4 sahen wir

$$\mathbf{E}_D h_{f_0} - \mathbf{E}_P h_{f_0} < \mathbf{E}_P h_{f_0} + \frac{32\tau}{n} + \frac{16\tau}{3n} = \mathbf{E}_P h_{f_0} + \frac{112\tau}{3n}$$

mit Wahrscheinlichkeit $P^n(\ldots) \geq 1 - e^{-\tau}$.

v) Kombiniert man die letzten beiden Teilschritte, so gilt mit Wahrscheinlichkeit $P^n(\ldots) \geq 1 - (1 + |\mathscr{C}|)e^{-\tau}$

$$\begin{aligned}
\mathbf{E}_P h_{f_D} &\leq \mathbf{E}_P h_{f_D} - \mathbf{E}_D h_{f_D} + \mathbf{E}_D h_{f_0} - \mathbf{E}_P h_{f_0} + \mathbf{E}_P h_{f_0} \\
&\leq \frac{13}{24}\mathbf{E}_P h_{f_D} + 2\mathbf{E}_P h_{f_0} + \frac{320}{3}\frac{\tau}{n} + \frac{117}{6}\varepsilon.
\end{aligned}$$

Damit folgt nun

$$\begin{aligned}
R_{L,P}(f_D) - R^*_{L,P} &= \mathbf{E}_P h_{f_D} \\
&\leq \frac{48}{11}\mathbf{E}_P h_{f_0} + \frac{2560}{11}\frac{\tau}{n} + \frac{464}{11}\varepsilon \\
&\leq 5\mathbf{E}_P h_{f_0} + 234\frac{\tau}{n} + 43\varepsilon.
\end{aligned}$$

Sei nun $f_0 \in \mathscr{F}$ mit $R_{L,P}(f_0) - R^*_{L,P,\mathscr{F}} \leq \frac{\varepsilon}{48}$, dann folgt

$$R_{L,P}(f_D) - R^*_{L,P} \leq 5\left(R^*_{L,P,\mathscr{F}} - R^*_{L,P}\right) + \frac{234\tau}{n} + 43\varepsilon.$$

Durch eine Transformation von $\tau \to \tau - \log(1 + |\mathscr{C}|)$ erhalten wir dann schließlich die Behauptung. □

Korollar 3.3.6 **Orakelungleichung für Histogrammregel**
Sei P eine Verteilung mit $P([-1,1]^d \times [-1,1]) = 1$ und $\mathscr{A}$ eine Würfelpartition der Weite $s \in (0,1]$. Dann gilt für $n \geq 2$ und $\tau \geq 1$

$$R_{L,P}(h_{D,s}) - R^*_{L,P} \leq 5\left(R_{L,P}(h_{P,s}) - R^*_{L,P}\right) + \frac{320\tau}{n} + \frac{(2^d+1)\log n}{s^d n}.$$

Die Annahme an P ist nicht wirklich notwendig und wird hier nur verwendet, um Tail-Terme zu verhindern.

Beweis: Wir betrachten $\mathscr{F}_{\mathscr{A}} := \left\{\sum_{A_j \cap [-1,1]^d \neq \emptyset} c_j \mathbf{1}_{A_j} : c_j \in [-1,1]\right\}$. Nach Satz 3.3.3 ist die Histogrammregel eine empirische Risikominimierung über $\mathscr{F}_{\mathscr{A}}$. Nach Satz 3.3.5 müssen wir daher $\mathscr{N}(\mathscr{F}_{\mathscr{A}}, \|\cdot\|_\infty, \varepsilon)$ bestimmen. Aus dem Beweis von Satz 2.1.5 wissen wir nun, dass $\left|\{j : A_j \cap [-1,1]^d \neq \emptyset\}\right| \leq 2^d s^{-d}$ gilt und ferner $\mathscr{N}([-1,1], |\cdot|, \varepsilon) \leq 2\varepsilon^{-1}$. Damit erhalten wir $\mathscr{N}(\mathscr{F}_{\mathscr{A}}, \|\cdot\|_\infty, \varepsilon) \leq (2\varepsilon^{-1})$ und durch Logarithmieren

$$\log(1 + \mathscr{N}(\mathscr{F}_{\mathscr{A}}, \|\cdot\|_\infty, \varepsilon)) \leq (2^d+1)s^{-d}\log\frac{2}{\varepsilon}.$$

Für $\varepsilon := \frac{2}{n}$ folgt dann die Behauptung. □

Korollar 3.3.7 **Konsistenz und Raten**
Sei $P([-1,1]^d \times [-1,1]) = 1$ und $(s_n) \subset (0,1)$. Betrachte die Lernmethode $D \mapsto h_{D,s_n}$ für $D \in (X \times Y)^n$. Dann gilt:

i) Lernmethode ist konsistent, falls

$$s_n \to 0 \quad \text{mit} \quad \frac{\log n}{s_n^d \cdot n} \to 0.$$

ii) Ist $f^*_{L,P}$ α-Hölder-stetig und

$$s_n \sim \left(\frac{\log n}{n}\right)^{\frac{1}{2\alpha+d}},$$

so folgt $\exists c > 0$ mit

$$P^n\left(D : \underbrace{\left\|h_{D,s_n} - f^*_{L,P}\right\|_{L^2(P_X)}}_{=\sqrt{R_{L,P}(f_{D,s_n}) - R^*_{L,P}}} \leq c \cdot \left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+d}}\right) \geq 1 - \frac{1}{n}$$

für alle $n \geq 1$.

Beweis: Standardargument, vergleiche Satz 2.1.10. □

Gleiche Konvergenzrate geht für $L_\infty(P_X)$, falls z. B. P_X die Gleichverteilung ist.

Frage: Die Rate in (ii) wurde für

$$s_n \sim \left(\frac{\log n}{n}\right)^{\frac{1}{2\alpha+d}}$$

erzielt. Leider kennen wir α nicht.
$\rightsquigarrow$ Lernrate für uns jetzt nicht erzielbar.

Wie können wir die Lernrate erzielen ohne α zu kennen?
$\rightsquigarrow$ „Adaptivität“.

Ansatz: Mache vereinfachte Version der *Cross-Validierung* aus den Übungsaufgaben.

i) Für $n \geq 4$ betrachte die endliche Menge $S = s_n \subset (0,1]$.

ii) Splitte Datensatz in

$$D_1 := ((x_1, y_1), \ldots, (x_m, y_m)),$$
$$D_2 := ((x_{m+1}, y_{m+1}), \ldots, (x_n, y_n),),$$

für $m := \left\lfloor \frac{n}{2} \right\rfloor + 1$.

iii) Erzeuge $h_{D_1,s}$ für $s \in S$.

iv) Wähle

$$s^*_{D_2} := \operatorname*{arg\,min}_{s \in S} R_{L,D_2}(h_{D_1,s}).$$

v) Die Entscheidungsfunktion ist $h_{D_1,s^*_{D_2}}$.

Das Verfahren nennen wir TV-HR (Trainings-Validation).

Satz 3.3.8 Orakelungleichung für TV-HR
sei $P([-1,1]^d \times [-1,1]) = 1$ und $S \subset [0,1]$ endlich. Betrachte TV-HR. Dann gilt für $n \geq 4$, $\tau \geq 1$ mit Wahrscheinlichkeit $P^n(\cdot) \geq 1 - e^{-\tau}$:

$$R_{L,P}(h_{D_1,s^*_{D_2}}) - R^*_{L,P} \leq \inf_{s \in S}\left(25(R_{L,P}(h_{P,s}) - R^*_{L,P}) + \frac{10(s^d+1)\log n}{n}\right) + \frac{5072 \log(1+|S|)}{n}$$

- S sollte möglichst groß sein, um den ersten Term klein zu bekommen.
- S nicht zu groß für zweiten Term.

Gleich: Die richtige Wahl.

Beweisidee: 3. Schritt von TV-HR durch Satz 3.3.6 behandelt. 4. Schritt von TV-HR ist ERM bezücglich „S“ und Least-Squares.

Beachte: Für geeignete S wird der erste Term der im Wesentlichen von Satz 3.3.6 kommt, den 2. Term, der von ERM kommt dominieren. Kurz: „HR ist schwieriger als ERM“ $\Rightarrow$ 4. Schritt egal.

Beweis: Für $m := \lfloor \frac{n}{2} \rfloor + 1$ gilt $m \geq \frac{n}{2} \geq 2$. Nach Korollar 3.3.6 mit union bound gilt:

$$P^m\Big(D_1 : R_{L,P}(h_{D_1,s}) - R^*_{L,P} \leq 5 \cdot (R_{L,P}(h_{P,s}) - R^*_{L,P}) + \frac{320\tau}{m} + \frac{(2^d+1)\log n}{s^d \cdot m} \text{ f. a. } s \in S\Big) \geq 1 - |S|\,\mathrm{e}^{-\tau}. \tag{$*$}$$

Für D_1 fest gilt nach Satz 3.3.5:

$$P^m\Big(D_2 : R_{L,P}(h_{D_1,s^*_{D_2}}) - R^*_{L,P} \leq 5 \inf_{s \in S}(R_{L,P}(h_{D_1,s}) - R^*_{L,P}) + \frac{234(\tau + \log(1+|S|))}{n-m}\Big) \geq 1 - \mathrm{e}^{-\tau}. \tag{$**$}$$

Werden beide Abschätzungen kombiniert, so ergibt sich:

$$R_{L,P}(h_{D_1,s^*_{D_2}}) - R^*_{L,P} \leq \inf_{s \in S}\Big(25(R_{L,P}(h_{P,s}) - R^*_{L,P}) + \frac{1600\tau}{m} + \frac{5 \cdot (2^d+1)\log m}{s^d m}\Big) + \frac{234 \cdot (\tau + \log(1+|S|))}{n-m}$$

mit Wahrscheinlichkeit $1 - (1+|S|) \cdot \mathrm{e}^{-\tau}$. Mit $m \geq \frac{n}{2}$ und $n - m \geq \frac{n}{2} - 1 \geq \frac{n}{4}$ und Variablentransformation von τ folgt die Behauptung. □

Korollar 3.3.9 **Konsistenz und Raten von TV-HR**
Sei $P([-1,1]^d \times [-1,1]) = 1$ und S_n ein $n^{-1/d}$-Netz von $[0,1]$ mit $|S_n| \leq c \cdot n^{-1/d}$. Betrachte die TV-HR für $n \geq 4$ und S_n. Dann gilt:

i) Die Lernmethode ist universell konsistent.

ii) Ist $f^*_{L,P}$ α-Hölder-stetig, so gibt es $\kappa > 0$ mit

$$P^n\Big(D : \Big\|h_{D,s^*_{D_2}} - f^*_{L,P}\Big\|_{L^2(P_X)} \leq \kappa \cdot \Big(\frac{\log n}{n}\Big)^{\frac{\alpha}{2\alpha+d}}\Big) \geq 1 - \mathrm{e}^{-\tau},$$

für alle $n \geq 4$.

Kurz: TV-HR lernt mit richtiger Rate ohne α zu kennen. Wir müssen nichtmal wissen, dass es α gibt!

Beweis: Betrachte $s_n := n^{-\frac{1}{2d}}$. Aufgrund der Netzeigenschaft existiert $s_n^* \in S_n$ mit $\left|s_n - s_n^*\right| \leq n^{-1/d}$.

$$\begin{aligned} \Rightarrow \quad & s_n^* \leq s_n + n^{-1/d} \leq 2 \cdot n^{-\frac{1}{2d}} \\ \text{und} \quad & s_n^* \leq s_n - n^{-1/d} \geq \tfrac{1}{2} \cdot n^{-\frac{1}{2d}}, \end{aligned}$$

falls $n \geq 2^{d+2}$. Damit (rechter Seite von Satz 3.3.9 ohne Konstanten)

$$\inf_{s \in S}\left(R_{L,P}(h_{P,s}) - R_{L,P}^* + \frac{\log n}{s^d \cdot n}\right) \leq \underbrace{R_{L,P}(h_{P,s_n^*}) - R_{L,P}^*}_{\to 0, \text{ da } s_n^* \to 0} + \underbrace{\frac{\log n}{(s_n^*)^d \cdot n}}_{\sim \frac{\log n}{n} \to 0}$$

und $\frac{\log(1+|s_n|)}{n} \sim \frac{\log n}{n} \to 0$.
$\Rightarrow$ (i) mit Satz 3.3.9.
(ii) recht analog: $s_n := \left(\frac{\log n}{n}\right)^{\frac{1}{2\alpha+d}}$.
$\Rightarrow \mathbf{E} s_n^* \in S_n$ mit $\left|s_n^* - s_n\right| \leq n^{-1/d}$.
$\Rightarrow s_n^* \sim \left(\frac{\log n}{n}\right)^{\frac{1}{2\alpha+d}}$ Mit diesem s_n^* in Satz 3.3.8 wie eben folgt die Behauptung. □

- $n^{-1/d}$-Netz ist in der Praxis im Allgemeinen zu teuer. Denn das kleinste solche Netz sähe wie $(2k-1)\cdot n^{-1/d}$ f?r $k = 1, \ldots, \sim n^{1/d}$. Daher oft S_n endliche Menge aus $[0,1]$ mit $\min S_n = c \cdot n^{-1/d}$, $\max S = 1$ und s in S „geometrisch verteilt“, d.h. aufeinanderfolgende s_i, s_{i+1} haben gleichen Quotienten ($\frac{s_{i+1}}{s_i}$ = konstant).
- In der Praxis ist TV-HR oft instabil. Daher verwendet man Cross-Validierung.
- Die Halbierung in Schritt 2 ist nicht notwendig. Es geht für alle Aufteilungen $p \cdot n, (1-p) \cdot n$ für $p \in (0,1)$.
- Der ganze Ansatz geht auch für $P(\mathbf{R}^d \times [-M, M]) = 1$. Dann ist das Tail-Verhalten in X für die Raten wie bei Dichteschätzung wichtig. TV-HR ist auch in diesem Sinne adaptiv. Für Y-Tails müssen wir dann auch etwas mehr machen.
- Für $f_{L,P}^*$ α-Hölder-stetig und P_X Gleichverteilung kann auch $\|\cdot\|_\infty$ gezeigt werden. Dazu mussOrakelungleichung, die auf

$$\left\|h_{D,s} - h_{P,s}\right\|_\infty = \sup_{A_j \cap [-1,1]^d \neq \emptyset} \left|\frac{\mathbf{E}_D \pi_y \mathbb{1}_{A_j \times \mathbf{R}}}{D(A_j)} - \frac{\mathbf{E}_P \pi_y \mathbb{1}_{A_j \times \mathbf{R}}}{P(A_j)}\right|.$$

3.4 Kernregel für Regression

In diesem Abschnitt ist K eine d-dimensionale Kernfunktion mit

$$\int K(\|x\|)\mathrm{d}x = 1.$$

Definition 3.4.1 **Kernregel für Regression**
Sei Q ein Wahrscheinlichkeitsmaß auf $\mathbf{R}^d \times \mathscr{B}$ mit $|Q|_2 < \infty$. Für $s > 0$ heißt

$$h_{Q,K,s}(x) := h_{Q,s}(x) = \frac{\int y' \cdot K\left(\frac{\|x-x'\|}{s}\right)\mathrm{d}Q(x',y')}{\int K\left(\frac{\|x-x'\|}{s}\right)\mathrm{d}Q(x',y')}$$

(mit der Konvention $\frac{0}{0} = 0$) Kernregel für Q.

Ist Q das empirische Maß bezüglich D, so gilt

$$h_{D,s}(x) = \frac{\sum_{i=1}^n K(\frac{\|x-x_i\|}{s}) \cdot y_i}{\sum_{i=1}^n K(\frac{\|x-x_i\|}{s})}.$$

4
Klassifikation

⌜ ... ⌟

4.1 Eigenschaften des Klassifikationsproblems

...

Satz 4.1.1
...

...

Vorlesungen werden noch nachgetragen

Beweis: Wir schreiben $\eta(x) := P(y = 1 \mid x)$. Dann gilt

$$\begin{aligned} R_{L,P}(f) &= \int L(y, f(x)) \, \mathrm{d}P(x, y) \\ &= \iint L(y, f(x)) \, P(\mathrm{d}y \mid x) \, \mathrm{d}P_X(x) \\ &= \int \eta(x) L(1, f(x)) + (1 - \eta(x)) L(-1, f(x)) \, \mathrm{d}P_X(x) \\ &= \int \eta(x) \mathbf{1}_{(-\infty,0)}(f(x)) + (1 - \eta(x)) \mathbf{1}_{[0,\infty)}(f(x)) \, \mathrm{d}P_X(x). \end{aligned}$$

Da $\eta(x) > 1 - \eta(x)$ äquivalent zu $\eta(x) > \frac{1}{2}$ ist, sollte $f(x) \geq 0$ sein. Für $\eta(x) < \frac{1}{2}$ sollte analog $f(x) < 0$ sein. Daher ist eine Funktion mit $(2\eta - 1) \operatorname{sign} f \geq 0$ also P_X-fast sicher eine Bayes-Entscheidungsfunktion. Einsetzen eines solchen f in $R_{L,P}$ ergibt

$$R^*_{L,P} = \int \min\{\eta, 1 - \eta\} \, \mathrm{d}P_X.$$

Zur Überschussrisikoberechnung betrachten wir eine Fallunterscheidung. Es gibt die sechs Kombinationen der folgenden Situationen:

- $\eta(x) > \frac{1}{2}$, $\eta(x) = \frac{1}{2}$, $\eta(x) < \frac{1}{2}$
- $f(x) \geq 0$, $f(x) < 0$

Wir betrachten nun zum Beispiel $\eta(x) > \frac{1}{2}$ und $f(x) \geq 0$. Dann ist

$$\big|2\eta - 1\big| \, \mathbf{1}_{(-\infty,0]}((2\eta - 1)\operatorname{sign} f) = 0$$

und

$$\begin{aligned}\eta \mathbf{1}_{(-\infty,0)}(f) + (1-\eta)\mathbf{1}_{[0,\infty)}(f) - \min\{\eta, 1-\eta\} &= 0 + (1-\eta) - (1-\eta) \\ &= 0.\end{aligned}$$

Nun betrachten wir $\eta(x) \geq \frac{1}{2}$ und $f(x) < 0$. Dann haben wir

$$\big|2\eta - 1\big| \, \mathbf{1}_{(-\infty,0]}((2\eta - 1)\operatorname{sign} f) = 2\eta - 1$$

und

$$\begin{aligned}\eta \mathbf{1}_{(-\infty,0)}(f) + (1-\eta)\mathbf{1}_{[0,\infty)}(f) - \min\{\eta, 1-\eta\} &= \eta + 0 - (1-\eta) \\ &= 2\eta - 1.\end{aligned}$$

Die restlichen Fälle verlaufen analog. □

Korollar 4.1.2 **Klassifikation als Mengenschätzung**
Betrachte $X_{-1} := \{\eta < \frac{1}{2}\}$ und $X_1 := \{\eta > \frac{1}{2}\}$. Für $f : X \to \mathbf{R}$ gilt dann

$$\begin{aligned}R_{L,P}(f) - R^*_{L,P} &= \int_{X_1 \triangle \{f \geq 0\}} \big|2\eta - 1\big| \, \mathrm{d}P_X \\ &= \int_{X_{-1} \triangle \{f < 0\}} \big|2\eta - 1\big| \, \mathrm{d}P_X.\end{aligned}$$

Gute Klassifikatoren f sind also gute Mengenschätzer und umgekehrt.

Beweis: Es ist $\mathbf{1}_{(-\infty,0]}((\eta - 1)\operatorname{sign} f) = 1$ genau dann, wenn $(2\eta - 1)\operatorname{sign} f \leq 0$ ist. Dies ist äquivalen zu $\big(\eta \leq \frac{1}{2}$ und $f \geq 0\big)$ oder $\big(\eta \geq \frac{1}{2}$ und $f < 0\big)$. Dies ist wiederum äquivalent zu

$$\underbrace{\left(f \geq 0 \text{ und } \eta \leq \frac{1}{2}\right)}_{\{f \geq 0\} \setminus X_1} \text{ oder } \underbrace{\left(\eta > \frac{1}{2} \text{ und } f < 0\right)}_{X_1 \setminus \{f \geq 0\}} \text{ oder } \left(\eta = \frac{1}{2}\right).$$

Damit haben wir nun

$$\begin{aligned} R_{L,P}(f) - R^*_{L,P} &= \int \left|2\eta - 1\right| \mathbf{1}_{(-\infty,0]}((2\eta-1)\operatorname{sign} f) \\ &= \int_{X_1 \triangle \{f \geq 0\}} \left|2\eta - 1\right| \, \mathrm{d}P_X. \end{aligned}$$

□

Satz 4.1.3 **Schätzung von η ist Klassifikation**
Für $f : X \to \mathbf{R}$ gilt

$$R_{L,P}(2f-1) - R^*_{L,P} \leq 2 \int \left|\eta - f\right| \, \mathrm{d}P_X.$$

Ist f eine gute $\|\cdot\|_{L_1}$-Schätzung von η, so ist $2f-1$ ein guter Klassifikator. Die Umkehrung gilt im Allgemeinen nicht.

Beweis: Nach Satz 4.1.1 gilt

$$R^*_{L,P}(2f-1) - R^*_{L,P} = \int \underbrace{\left|2\eta - 1\right| \mathbf{1}_{(-\infty,0]}((2\eta-1)\operatorname{sign}(2f-1))}_{=:\mathrm{I}} \, \mathrm{d}P_X.$$

Wir schreiben nun $h := 2f-1$ und betrachten eine Fallunterscheidung. Im ersten Fall ist $\eta = \frac{1}{2}$ dann ist $\mathrm{I} = 0 \leq 2\left|\eta - f\right|$. Im zweiten Fall ist $(2\eta-1)\operatorname{sign} h > 0$, dann folgt ebenfalls $\mathrm{I} = 0 \leq 2\left|\eta - f\right|$.

Im dritten Fall ist $(2\eta-1)\operatorname{sign} h < 0$, dann folgt $\mathrm{I} = \left|2\eta - 1\right|$. Falls $\eta > \frac{1}{2}$ ist, so folgt $h < 0$ und

$$\left|2\eta - 1\right| = 2\eta - 1 < 2\eta - h - 1 = 2\eta - 2f \leq 2\left|\eta - f\right|.$$

Ist hingegen $\eta < \frac{1}{2}$, so ist $h \geq 0$ und der Rest folgt analog. □

Wir beobachten nun

$$\mathbf{E}(Y \mid x) = \int y \, P(\mathrm{d}y \mid x) = 1 \cdot \eta(x) + (-1)(1 - \eta(x)) = 2\eta(x) - 1.$$

Definition 4.1.4 **Plug-In-Verfahren**
Sei $\mathscr{L}$ eine Regressionsmethode, welche die Entscheidungsfunktion f_D erzeugt. Dann heiße eine Klassifikationsmethode, welche die Entscheidungsfunktion f_D bzw. $\operatorname{sign} f_D$ erzeugt, Plug-In-Klassifikator von $\mathscr{L}$.

Korollar 4.1.5 **Überschussrisiko von Plug-In**
Für $f : X \to \mathbf{R}$ gilt

$$R_{L,P}(f) - R^*_{L,P} \leq \|f - \mathbf{E}(Y \mid \cdot)\|_{L_2(P_X)}.$$

Eine solche Ungleichung heißt Kalibrierungsungleichung.

Insbesondere übertragen sich Orakelungleichungen, Konsistenz und Konvergenzraten der Regresstionsmethode auf die zugehörige Plug-In-Klassifikationsmethode.

Beweis: Es ist $2\int \left|f - \eta\right| \, \mathrm{d}P_X = \int \left|2f - 1 - (2\eta - 1)\right| \, \mathrm{d}P_X$ und

$$\begin{aligned} 2\eta - 1 &= \mathbf{E}(Y \mid \cdot) \\ &= \int |2f - 1 - \mathbf{E}(Y \mid \cdot)| \, \mathrm{d}P_X \\ &\leq \|h - \mathbf{E}(Y \mid \cdot)\|_{L_2}. \end{aligned}$$

Dann wenden wir Satz 4.1.3 an. □

4.2 Das No-Free-Lunch-Theorem

Bis jetzt haben wir Konvergenzraten nur unter Verteilungsannahmen bekommen. Die Frage ist also, ob dies zwingend ist. Wir werden dies nur für die Klassifikation besprechen, für die Regression folgt es dann automatisch. Ferner betrachten wir $\mathbf{E}_{D\sim P}\left(R_{L,P}(f_D) - R^*_{L,P}\right)$, statt das Überschussrisiko selbst zu betrachten. Beides ist jedoch nicht notwendig.

Definition 4.2.1 **Atom**
Sei μ ein Maß auf $(X, \mathscr{A})$. Dann heißt $A \in \mathscr{A}$ mit $\mu(A) > 0$ Atom genau dann, wenn für alle $B \in \mathscr{A}$ mit $B \subset A$ entweder $\mu(B) = 0$ oder $\mu(A \setminus B) = 0$ gilt.

Ferner heißt das Maß μ atomfrei, wenn alle $A \in \mathscr{A}$ keine Atome sind.

Atome lassen sich also nicht weiter in messbare Mengen mit positivem Maß zerlegen. So ist $(X, \#)$ nicht atomfrei, da jede einelementige Menge ein Atom ist. Das Lebesguemaß über $(X, \mathscr{B})$ ist jedoch atomfrei.

Satz 4.2.2 **Lyapunov**
Ist μ ein endliches, atomfreies Maß, so gilt $\{\mu(A) : A \in \mathscr{A}\} = [0, \mu(X)]$.

Satz 4.2.3 **No-Free-Lunch-Theorem (Devroy 1982)**
Sei $(a_n) \subset \left(0, \frac{1}{32}\right]$ eine fallende Nullfolge und $(X, \mathscr{A}, \mu)$ atomfrei, sowie $Y := \{-1, 1\}$. Dann gibt es zu jeder Lernmethode $\mathscr{L}$ ein P auf $X \times Y$ mit

i) $P_X = \mu$

ii) $R^*_{L,P} = 0$, d. h. $\eta \in \{0, 1\}$ P_X-fast sicher

iii) $\mathbf{E}_{D\sim P^n}\left(R_{L,P}(f_D) - R^*_{L,P}\right) \geq a_n$ für alle $n \geq 1$

Das No-Free-Lunch-Theorem besagt, dass es keine verteilungsfreien Konvergenzraten gibt. Da μ atomfrei ist, ist zu beachten, dass $|X| = \infty$ folgt. Auch Wissen über P_X kann das No-Free-Lunch-Theorem nicht verhindern. Ferner gilt es auch für $X = \mathbf{N}$, wenn μ nicht vorgegeben ist.

Lemma 4.2.4
Sei $a_n \searrow 0$ mit $a_1 \leq \frac{1}{16}$. Dann existiert $(p_n) \subset (0, 1)$ fallend mit

i) $\sum_{i=1}^{\infty} p_i = 1$

ii) $\sum_{i=n+1}^{\infty} p_i \geq \max\{8a_n, 32np_{n+1}\}$

Beweis: Da wir $p_n \geq p_{n+1}$ zeigen werden, reicht es

$$\sum_{i=n+1}^{\infty} p_i \geq \max\{8a_n, 32np_n\} \tag{*}$$

zu zeigen. Für $l < m$ definieren wir $H(m,l) := \sum_{i=l}^{m-1} \frac{1}{i}$. Wir setzen nun $n_1 = 1$, dann folgt $8a_n \leq 8 \cdot \frac{1}{16} \leq 2^{-1}$. Ferner gibt es ein $n_2 \geq n_1$ mit $H(n_2, n_1) \geq 32$. Da $H(\cdot, n_1)$ wachsend und (a_n) fallend ist, können wir ohne Einschränkung annehmen, dass $8a_{n_2} \leq 2^{-2}$ gilt. Aus den selben Gründen finden wir nun induktiv eine Folge n_k mit $n_k \geq n_{k-1}$ und $H(n_k, n_{k-1}) \geq H(n_{k-1}, n_{k-2})$ und $8a_{n_k} \leq 2^{-k}$. Ferner definieren wir nun

$$c_k := \frac{32}{2^k H(n_{k+1}, n_k)}$$

für $k \geq 1$, dann ist $c_k \searrow$ und es gilt

$$\frac{1}{32} \sum_{k=1}^{\infty} c_k H(n_{k+1}, n_k) = \sum_{k=1}^{\infty} \frac{1}{2^k} = 1.$$

Für $n \in [n_k, n_{k+1})$ setzen wir $p_n := \frac{c_k}{32n}$, dann ist $p_n \searrow$ und

$$\begin{aligned} \sum_{n=1}^{\infty} p_n &= \sum_{k=1}^{\infty} \sum_{i=n_k}^{n_{k+1}-1} \frac{c_k}{32i} = \sum_{k=1}^{\infty} \frac{c_k}{32} \sum_{i=n_k}^{n_{k+1}-1} i^{-1} = \frac{1}{32} \sum_{k=1}^{\infty} c_k H(n_{k+1}, n_k) \\ &= 1. \end{aligned}$$

Für $n \in [n_k, n_{k+1})$ gilt ferner

$$\sum_{i=n+1}^{\infty} p_i \geq \sum_{i=n_{k+1}}^{\infty} p_i = \sum_{i=k+1}^{\infty} \frac{c_i}{32} H(n_{i+1}, n_i) = \sum_{i=k+1}^{\infty} 2^{-i} = 2^{-k}.$$

Nun schätzen wir 2^{-k} wie folgt ab: aus $H(n_{k+1}, n_k) \geq H(n_k, n_1) \geq 32$ folgt

$$\frac{1}{2^k} \geq \frac{32}{2^k H(n_{k+1}, n_k)} = c_k = 32np_n$$

und außerdem $2^{-k} \geq 8a_{n_k} \geq 8a_n$, da $n \geq n_k$ und $a_n \searrow$ gilt. □

Bevor wir nun zum eigentlichen Beweis des No-Free-Lunch-Theorems kommen, wollen wir die Idee des Beweises zunächst vorstellen. Wir zerlegen X in Zellen mit den Wahrscheinlichkeiten p_k und η soll auf diesen Zellen konstant sein. Haben wir nun n Samples, die genau auf die ersten n Zellen verteilt sind, so könnten wir alle Daten in diesen Zellen exakt klassifizieren und da $p_k \searrow$ gilt, ist dies das bestmögliche Ergebnis. Wir werden daher zeigen, dass die

Masse in den Zellen danach, bei denen wir lediglich raten können, groß genug ist, um eine Fehlerabschätzung unmöglich zu machen.

Beweis von Satz 4.2.3: Ohne Einschränkung sei $f_D(x) \in \{-1,1\}$, andernfalls betrachten wir $\operatorname{sign} f_D$. Wir fixieren eine Folge (p_i) gemäß Lemma 4.2.4 und sukzessives Anwenden von Satz 4.2.2 liefert dann eine Partition $A_1, A_2, \ldots$ von X mit $\mu(A_j) = p_j$. Sei $\overline{\nu}$ das Wahrscheinlichkeitsmaß auf $\{0,1\}$ mit $\overline{\nu}(\{0\}) = \frac{1}{2}$. Wir betrachten $\nu := \bigotimes_{i=1}^{\infty} \overline{\nu}$ auf $\Omega := \{0,1\}^{\mathbf{N}}$. Wir konstruieren nun Wahrscheinlichkeitsmaße P. Für $\omega = (\omega_j)_{j\geq 1} \in \Omega$ schreiben wir

$$\eta_\omega(x) := \sum_{j=1}^{\infty} \omega_j \mathbf{1}_{A_j}(x).$$

Es gilt also $\eta_\omega(x) = \omega_j$ für $x \in A_j$. Sei P_ω das Wahrscheinlichkeitsmaß auf $X \times Y$, das zu $(P_\omega)_X = \mu$ und η_ω gehört. Dies wird ein zufälliges Maß. Nach Konstruktion gilt $(P_\omega)_X = \mu$ und

$$R^*_{L,P,\omega} = \int \min\{\eta_\omega, 1-\eta_\omega\} \, \mathrm{d}P_X = 0.$$

Der Ansatz ist es nun,

$$\int_\Omega \inf_{n\geq 1} \frac{1}{a_n} \underbrace{\int_{(X\times Y)^n} R_{L,P}(f_D) \, \mathrm{d}P_\omega^n(D)}_{=\mathbf{E}_{D\sim P_\omega^n}(R_{L,P}(f_D) - R^*_{L,P})} \, \mathrm{d}\nu(\omega) \geq \frac{1}{2} \tag{*}$$

zu zeigen. Falls dies gilt, so folgt die Existenz eines $\omega \in \Omega$ mit

$$\inf_{n\geq 1} \frac{1}{a_n} \mathbf{E}_{D\sim P_\omega^n} \left(R_{L,P}(f_D) - R^*_{L,P} \right) \geq \frac{1}{2}$$

und für $(2a_n)$ folgt dann die Behauptung. Dazu schätzen wir den Term in (*) ab:

$$\begin{aligned}
\int_\Omega \inf_{n\geq 1} \frac{1}{a_n} \int_{(X\times Y)^n} R_{L,P}(f_D) \, \mathrm{d}P_\omega^n(D) \, \mathrm{d}\nu(\omega) &\geq \int_\Omega \int_{(X\times Y)^\infty} \inf_{n\geq 1} \frac{1}{a_n} R_{L,P_\omega}(f_{D_n}) \, \mathrm{d}P_\omega^\infty(D) \, \mathrm{d}\nu(\omega) \\
&\geq \int_\Omega \int_{(X\times Y)^\infty} \mathbf{1}_{\bigcap_{n=1}^\infty \{R_{L,P_\omega}(f_{D_n}) \geq a_n\}} \, \mathrm{d}P_\omega^\infty(D) \, \mathrm{d}\nu(\omega) \\
&\geq 1 - \sum_{n=1}^{\infty} \int_\Omega \int_{(X\times Y)^\infty} \mathbf{1}_{\{R_{L,P_\omega}(f_{D_n}) < a_n\}} \, \mathrm{d}P_\omega^\infty(D) \, \mathrm{d}\nu(\omega).
\end{aligned}$$

Wir wollen nun versuchen, f_{D_n} durch eine Funktion zu ersetzen, die auf den Zellen konstant ist. Für $j \geq 1$ definieren wir

$$\overline{f}_{D_n}(j) := \begin{cases} 1 & \text{falls } \mu(\{f_{D_n} = 1\} \cap A_j) \geq \mu(\{f_{D_n} = -1\} \cap A_j) \\ -1 & \text{sonst} \end{cases}$$

und schreiben $E_{\omega,j}(f_{D_n}) := A_j \cap \{f_{D_n} \neq 2\eta_\omega - 1\}$. Auf A_j ist $\eta_\omega(x) = \omega_j$ und daher gilt

$$\overline{f}_{D_n}(j) \neq 2\omega_j - 1 \quad \Longrightarrow \quad \mu\left(\mathbf{E}_{\omega,j}\left(f_{D_n}\right)\right) \geq \frac{p_j}{2}.$$

Damit folgt

$$\mathbf{1}_{\{\mu(\mathbf{E}_{\omega,j}(f_{D_n}))\geq \frac{p_j}{2}\}} \geq \mathbf{1}_{\{\overline{f}_{D_n}(j)=2\omega_j-1\}}$$

und wegen $\mu(\mathbf{E}_{\omega,j}(f_{D_n})) \geq \frac{p_j}{2}\mathbf{1}_{\mu(\mathbf{E}_{\omega,j}(f_{D_n}))\geq\frac{p_j}{2}}$ folgt

$$\begin{aligned}
R_{L,P_\omega}(f_{D_n}) &= \int \left|2\eta_\omega - 1\right| \cdot \mathbf{1}_{(-\infty,0]}((2\eta_\omega - 1)f_{D_n}) \,\mathrm{d}\mu \\
&= \mu(\{f_{D_n} \neq 2\eta_\omega - 1\}) \\
&= \sum_{j=1}^{\infty} \mu(\mathbf{E}_{\omega,j}(f_{D_n})) \\
&\geq \frac{1}{2}\sum_{j=1}^{\infty} p_j \mathbf{1}_{\{\overline{f}_{D_n}(j)\neq 2\omega_j-1\}} \\
&\geq \frac{1}{2}\sum_{j\colon \forall_{i\leq n}\colon x_i\notin A_j} p_j \mathbf{1}_{\{\overline{f}_{D_n}(j)\neq 2\omega_j-1\}}.
\end{aligned}$$

Damit haben wir

$$\begin{aligned}
&\int_\Omega \int_{(X\times Y)^\infty} \mathbf{1}_{\{R_{L,P_\omega}(f_{D_n})<a_n\}} \,\mathrm{d}P_\omega^\infty(D) \,\mathrm{d}\nu(\omega) \\
&\qquad\qquad \leq \int_\Omega \int_{(X\times Y)^n} \mathbf{1}_{\left\{\sum_{j\colon x_i\notin A_j} p_j \mathbf{1}_{\{\overline{f}_D(j)\neq 2\omega_j-1\}} < 2a_n\right\}} \,\mathrm{d}P_\omega^n(D) \,\mathrm{d}\nu(\omega). \qquad (**)
\end{aligned}$$

Nun gibt es für $x \in X$ genau ein j mit $x \in A_j$, für welches wir $j(x)$ schreiben. Nach Konstruktion ist x_i von ω unabhängig und $y_i = 2\omega_j(x_i) - 1$. Daher hängt $\overline{f}_D$ nur von $D_X := (x_1,\ldots,x_n) \sim \mu^n$ und $\omega_{D_X} := \big(\omega_j(x_1),\ldots,\omega_j(x_n)\big)$ ab. Daher schreiben wir $\overline{f}_{D_X,\omega} := \overline{f}_D$. Es folgt

$$(**) = \int_{X^n} \int_\Omega \mathbf{1}_{\left\{\sum_{j\colon x_i\notin A_j} p_j \mathbf{1}_{\{\overline{f}_{D_X,\omega(j)}\neq 2\omega_j-1\}} < 2a_n\right\}} \,\mathrm{d}\nu(\omega) \,\mathrm{d}\mu^n(D_X).$$

Es bezeichne Ω_{D_X} das Kreuzprodukt, welches durch die Koordinaten $j(x_1),\ldots,j(x_n)$ gegeben ist. Analog ist $\Omega_{\neg D_X}$ das Kreuzprodukt der übrigen Koordinaten. Auf beiden betrachten wir

die Randverteilungen ν_{D_X} und $\nu_{\neg D_X}$. Dann ändert sich $\overline{f}_{D_X,\omega}$ nur in den Koordinaten, die von Ω_{D_X} beschrieben werden. Nun gilt

$$\begin{aligned}
&\int_\Omega \mathbf{1}_{\left\{\sum_{j:\, x_i \notin A_j} p_j \mathbf{1}_{\{\overline{f}_{D_X,\omega}(j) \neq 2\omega_j - 1\}} < 2a_n\right\}} \, \mathrm{d}\nu(\omega) \\
&\qquad = \int_{\Omega_{D_X}} \int_{\Omega_{\neg D_X}} \mathbf{1}_{\{\ldots\}} \, \mathrm{d}\nu_{\neg D_X}(\omega_{\neg D_X}) \, \mathrm{d}\nu_{D_X}(\omega_{D_X}) \\
&\qquad = \int_{\Omega_{D_X}} \int_{\Omega_{\neg D_X}} \mathbf{1}_{\left\{\sum_{j:\, x_i \notin A_j} p_j \mathbf{1}_{\{\omega_j = 1\}} < 2a_n\right\}} \, \mathrm{d}\nu_{\neg D_X}(\omega_{\neg D_X}) \, \mathrm{d}\nu_{D_X}(\omega_{D_X}) \\
&\qquad = \int_\Omega \mathbf{1}_{\left\{\sum_{j:\, x_i \notin A_j} p_j \omega_j < 2a_n\right\}} \, \mathrm{d}\nu(\omega) \\
&\qquad \leq \int_\Omega \mathbf{1}_{\left\{\sum_{j=n+1}^{\infty} p_j \omega_j < 2a_n\right\}} \, \mathrm{d}\nu(\omega) \\
&\qquad = \nu\left(\left\{\omega : -\sum_{j=n+1}^{\infty} p_j \omega_j > -2a_n\right\}\right) \\
&\qquad = \nu\left(\left\{\omega : \exp\left(2sa_n - s\sum_{j=n+1}^{\infty} p_j \omega_j\right) > 1\right\}\right) \\
&\qquad < \mathbf{E}_{\omega \sim \nu} \exp\left(2sa_n - s\sum_{j=n+1}^{\infty} p_j \omega_j\right) \\
&\qquad = e^{2sa_n} \mathbf{E}_{\omega \sim \nu} \exp\left(-s\sum_{j=n+1}^{\infty} p_j \omega_j\right) \\
&\qquad = e^{2sa_n} \prod_{j=n+1}^{\infty} \mathbf{E}_{\omega_j \sim \overline{\nu}} e^{-sp_j\omega_j} \\
&\qquad = e^{2sa_n} \prod_{j=n+1}^{\infty} \left(\frac{1}{2} + \frac{1}{2}e^{-sp_j}\right) \\
&\qquad \leq e^{2sa_n} \prod_{j=n+1}^{\infty} \frac{1}{2}\left(2 - sp_j + \frac{s^2 p_j^2}{2}\right) \\
&\qquad \leq e^{2sa_n} \prod_{j=n+1}^{\infty} \exp\left(-\frac{sp_j}{2} + \frac{s^2 p_j^2}{2}\right) \\
&\qquad = \exp\left(2sa_n - \frac{s}{2}\sum_{j=n+1}^{\infty} p_j + \frac{s^2}{4}\sum_{j=n+1}^{\infty} p_j^2\right) \\
&\qquad \leq \exp\left(2sa_n - \frac{s}{2}\sum_{j=n+1}^{\infty} p_j + \frac{s^2 p_{n+1}}{4}\sum_{j=n+1}^{\infty} p_j\right),
\end{aligned}$$

wobei wir $s > 0$ nun durch $s := \frac{\sum_{j=n+1}^{\infty} p_j - 4a_n}{p_{n+1} \sum_{j=n+1}^{\infty} p_j} > 0$ definieren, was sich aus der Konstruktion der p_j ergibt. Ferner sei $A := \sum_{j=n+1}^{\infty} p_j$, womit wir $s = \frac{A-4a_n}{p_{n+1}A}$ erhalten. Dann folgt

$$\begin{aligned}
&= \exp\left(-2\frac{A-4a_n}{p_{n+1}A} \cdot a_n - \frac{A-4a_n}{2p_{n+1}A}A + \left(\frac{A-4a_n}{p_{n+1}A}\right)\right) \\
&= \exp\left(\frac{-8a_n + 32a_n^2 + 2A^2 - 8Aa_n - A^2 + 8Aa_n - 16a_n^2}{4Ap_{n+1}}\right) \\
&= \exp\left(-\frac{1}{4}\frac{A - 8Aa_n + 16a_n^2}{p_{n+1}A}\right) \\
&= \exp\left(-\frac{1}{4}\frac{(A-4a_n)^2}{p_{n+1}A}\right).
\end{aligned}$$

Es gilt $A \geq 8a_n$ und die Funktion $x \mapsto (A-4x)^2 = (4x-A)^2$ besitzt ein Minimum bei $x = \frac{A}{4}$ und ist daher monoton fallend. Damit folgt $(A-4x)^2 \geq \left(A - \frac{A}{2}\right)^2 = \frac{A^2}{4}$ und es folgt weiter

$$\begin{aligned}
&\leq \exp\left(-\frac{A}{16p_{n+1}}\right) \\
&\leq e^{-2n}.
\end{aligned}$$

Insgesamt erhalten wir damit

$$\begin{aligned}
\int_\Omega \inf_{n\geq 1} \frac{1}{a_n} \int_{(X\times Y)^n} R_{L,P}(f_D)\, \mathrm{d}P_\omega^n(D)\, \mathrm{d}\nu(\omega) &\geq 1 - \sum_{n=1}^{\infty} \int_\Omega \int_{(X\times Y)^\infty} \mathbf{1}_{\{R_{L,P_\omega}(f_{D_n}) < a_n\}}\, \mathrm{d}P_\omega^\infty(D)\, \mathrm{d}\nu(\omega) \\
&> 1 - \sum_{n=1}^{\infty} e^{-2n} \\
&= \frac{e^2-2}{e^2-1} \\
&> \frac{1}{2}.
\end{aligned}$$

□

Es gibt das No-Free-Lunch-Theorem auch noch für andere Problemstellungen, zum Beispiel für $X \subset \mathbf{R}^d$ und η in C^∞, für $X \subset \mathbf{R}^2$ und η ist unimodal in x_0, das heißt für $\lambda > 0$ ist $\eta(\lambda x_0)$ monoton fallend, oder für $\eta \in \{0,1\}$, $X \subset \mathbf{R}^2$ und $\{\eta = 1\}$ kompakt und konvex mit $0 \in \{\eta = 1\}$.

Es gibt keine Superklassifikationsmethode: Ist $\mathscr{L}$ eine Klassifikationsmethode, so gibt es eine universell konsistente Klassifikationsmethode $\mathscr{L}'$ und ein P auf $X \times Y$ mit

$$\mathbf{E}_{D\sim P^n} R_{L_{\text{class}},P}(f_D) > \mathbf{E}_{D\sim P^n} R_{L_{\text{class}},P}(f'_D)$$

für alle $n \geq 1$.

Abschätzung von $R^*_{L_{\text{class}},P}$: Für jede Methode $\mathscr{L}$, die $R^*_{L_{\text{class}},P}$ schätzt und jedes $n \geq 1$ und $\varepsilon > 0$ gibt es ein P auf $X \times Y$ mit

$$\mathbf{E}_{D\sim P^n} \left| f_D - R^*_{L_{\text{class}},P} \right| \geq \frac{1}{4} - \varepsilon.$$

Offene Fragen: Eine Klassifikationsmethode $\mathscr{L}$ heißt *smart*, wenn $\mathbf{E}_{D\sim P^n} R_{L_{\text{class}},P}(f_D) \searrow$ in n für alle P auf $X \times Y$ gilt. Die Frage ist nun, ob es eine universell konsistente und smarte Klassifikationsmethode gibt.

4.3 Histogramme für Klassifikation

Wie bisher sei $Y = \{-1, 1\}$, aber nun sei $X \subset \mathbf{R}^d$.

Definition 4.3.1 Histogrammregel für Klassifikation
Sei $\mathscr{A}$ eine Würfelpartition der Weite s und Q ein Wahrscheinlichkeitsmaß auf $X \times Y$. Für $x \in X$ bezeichne $A(x)$ die Zelle von $\mathscr{A}$ mit $x \in A(x)$. Dann heißt die Abbildung $h_{Q,s} : X \to Y$ mit

$$h_{Q,s}(x) := \begin{cases} 1 & \text{für } f_{Q,s}(x) < 0 \\ 1 & \text{für } f_{Q,s}(x) \geq 0 \end{cases},$$

wobei $f_{Q,s}(x) := Q(A(x) \times \{1\}) - Q(A(x) \times \{-1\})$ ist, Histogramm von Q.

Für einen Datensatz D gilt

$$f_{D,s} := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{Y_i = 1\}} \mathbf{1}_{A(x)}(x_i) - \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{Y_i = -1\}} \mathbf{1}_{A(x)}(x_i).$$

$f_{D,s}(x)$ zählt die Samples (x_i, y_i), die in $A(x)$ fallen und positive bzw. negative Labels haben. Gibt es mehr negative Labels, so ist $h_{Q,s}(x) = -1$, ansonsten ist $h_{Q,s}(x) = 1$.

Satz 4.3.2 Histogrammregel ist Plug-In
Die Methode $D \mapsto h_{D,s}$ ist eine Plug-In-Klassifikationsmethode.

Beweis: Es gilt

$$\begin{aligned} f_{D,s}(x) &= \frac{1}{n} \sum_{i=1}^{n} y_i \mathbf{1}_{A(x)}(x_i) \\ &= D(A(x)) \cdot \underbrace{\frac{1}{n} \sum_{i=1}^{n} y_i \mathbf{1}_{A(x)}(x_i)}_{=: \overline{h}_{D,s}}, \end{aligned}$$

wobei $\overline{h}_{D,s}$ eine Histogrammregel für die Regression auf $A(x)$ ist. Es ist $D(f(x)) \geq 0$ und $h_{D,s}(x) = \operatorname{sign} f_{D,s}(x) = \operatorname{sign} \overline{h}_{D,s}(x)$. □

Zuletzt kompiliert: 27. Juni 2012

Literaturverzeichnis

[WTSkript11] I. Steinwart, *Wahrscheinlichkeitstheorie*, Mitschrieb der Vorlesung „Wahrscheinlichkeitstheorie“ von Ingo Bürk, Wintersemester 2010/2011

[Devroye96] L. Devroye, L. Györfi und G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer, 1996

[Devroye00] L. Devroye und G. Lugosi, *Combinatorial Methods in Density Estimation*, Springer, 2000

[Györfi02] L. Györfi, M. Kohler, A. Krzyzak und H. Walk, *A Distribution-Free Theory of Nonparametric Regression*, Springer, 2002

[Tsybakov08] A. Tsybakov, *Introduction to Nonparametric Estimation, Springer*, 2008

Abbildungsverzeichnis

1.1 Binäre Klassifikation 10
1.2 Regression 11
1.3 Moving Window Rule 14

Stichwortverzeichnis

Überdeckungszahl, 44

Atom, 79
    atomfreies Maß, 79

Bayes
    -Entscheidungsfunktion, 8
    -Risiko, 8
Bayes-Entscheidungsfunktion, 61
Bernstein
    -ungleichung, 17
Bounded Difference Inequality, 20

Empirische Risikominimierung, 54
ERM, 54

Glättung, 40

Histogramm, 23, 64
Histogrammregel, 24
Hoeffding
    -ungleichung, 18

Kernfunktion, 39
Konsistenz
    universelle, 32

Lernmethode, 7

Markov
    -ungleichung, 16
McDiarmid
    Inequality, 20

No-Free-Lunch-Theorem, 79

Orakelungleichung, 26, 37, 47, 51, 54, 56, 69

Plug-In-Klassifikator, 77

regulär bedingte Wahrscheinlichkeit, 60

Risiko, 8
    Überschuss-, 9
    Bayes-, 8

Talagrand
    -ungleichung, 21

Verlustfunktion, 8
    überwachte, 8
    unüberwacht, 8

Würfelpartition, 23

Young
    -Ungleichung, 41

Zelle, 23