Nichtparametrische Statistik
Skript zur Vorlesung im Wintersemester 2010/11

Markus Reiß
Humboldt-Universität zu Berlin
mreiss@mathematik.hu-berlin.de

VORLÄUFIGE FASSUNG: 18. März 2011

Inhaltsverzeichnis

1 Einführung

1.1 Statistische Modellierung

Aufgabe der Statistik ist es, auf Grund von zufälligen Beobachtungen Rückschlüsse auf zugrundeliegende Modellparameter zu ziehen. Zur mathematischen Formalisierung benötigen wir daher zunächst ein Beobachtungsmodell. Wir bezeichnen daher als *statistisches Modell* einen Messraum $(\mathcal{X}, \mathscr{F})$ versehen mit einer Familie $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ von Wahrscheinlichkeitsmaßen, wobei $\Theta \neq \varnothing$ eine beliebige Parametermenge bezeichnet. *Beobachtungen* in diesem Modell sind beliebige Zufallsvariablen Y. Wie gewohnt, spielt der zugrundeliegende Raum häufig keine Rolle, und wir benutzen nur, dass die Beobachtung Y eine von ϑ abhängige Verteilung besitzt. Sind $X_1, \ldots, X_n$ unabhängige, identisch verteilte Zufallsvariablen unter jedem P_ϑ, so heißt $(X_1, \ldots, X_n)$ eine *mathematische Stichprobe* vom Umfang n.

Ein *Schätzer* $\hat\vartheta$ des unbekannten Parameters ϑ ist eine messbare Funktion der Beobachtungen Y, insbesondere also wiederum eine Zufallsvariable. Allgemeiner wird ein abgeleiteter Parameter $g(\vartheta)$ für eine Funktion g geschätzt durch eine messbare Funktion $\hat g$ der Beobachtungen. Wir messen den Fehler dieses Schätzers mittels einer nicht-negativen *Verlustfunktion* $\ell(\hat g, g(\vartheta))$ und bezeichnen als *Risiko* oder weniger genau *Fehler* dieses Schätzers bei Vorliegen des wahren, aber unbekannten Parameters ϑ den mittleren Verlust

$$R(\hat g, \vartheta) := \mathbb{E}_\vartheta[\ell(\hat g, g(\vartheta))] := \int_{\mathcal{X}} \ell(\hat g(Y(x)), g(\vartheta))\ \mathbb{P}_\vartheta(dx).$$

Beachte, dass das Risiko eine Funktion von ϑ ist, es also im Allgemeinen sinnlos ist, von dem besten Schätzer $\hat g$ im Modell zu sprechen, da für verschiedene $\vartheta \in \Theta$ Schätzer ganz unterschiedlich große Fehler besitzen können. Wir werden Vergleichskriterien im Laufe der Vorlesung kennenlernen. Schließlich sei noch darauf hingewiesen, dass die gesamte Modellierung in der Statistik vor der Datenauswertung stattfinden muss. *Daten* sind realisierte Beobachtungen $Y(x)$ für ein $x \in \mathcal{X}$ und führen zu realisierten Schätzern und somit zu konkreten *Schätzwerten* $\hat g(Y(x))$.

1.1 Beispiel. Es sei $X_1, \ldots, X_n$ eine $N(\mu, 1)$-verteilte mathematische Stichprobe mit unbekanntem Mittelwert $\mu \in \mathbb{R}$. Diese kann zum Beispiel modelliert werden als Identität auf dem Raum $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n}, (N(\mu\mathbf{1}, E_n))_{\mu\in\mathbb{R}})$ mit Einsvektor $\mathbf{1} = (1, 1, \ldots, 1)^\top \in \mathbb{R}^n$ und Einheitsmatrix $E_n \in \mathbb{R}^{n\times n}$. Das Stichprobenmittel $\hat\mu := \frac{1}{n}\sum_{i=1}^n X_i$ oder auch der Stichprobenmedian $\tilde\mu := \mathrm{med}(X_1, \ldots, X_n)$ sind natürliche Schätzer für μ. Allerdings ist auch eine konstante $\bar\mu := \pi/3$ ein zulässiger Schätzer. Häufig wird ein quadratischer Verlust $\ell(x, y) := (x - y)^2$ betrachtet, der zum quadratischen Fehler (*MSE: mean squared error*) führt: $R(\hat\mu, \mu) = \mathbb{E}_\mu[(\hat\mu - \mu)^2]$. Eine einfache Rechnung ergibt $R(\hat\mu, \mu) = \frac{1}{n}$ sowie $R(\bar\mu, \mu) = (\mu - \pi/3)^2$, für $\tilde\mu$ sind die Ausdrücke

komplizierter. Für $\mu = \pi/3$ ist $\bar{\mu}$ sicherlich der beste Schätzer, allerdings ist er für die meisten anderen Werte von μ sehr schlecht.

Ist der abgeleitete Parameter $g(\vartheta)$ reellwertig, so heißt ein Schätzer $\hat{g}_n$ *unverzerrt* oder *erwartungstreu* (*unbiased*), falls $\mathbb{E}_\vartheta[\hat{g}] = g(\vartheta)$ für alle $\vartheta \in \Theta$ gilt. Der *Bias* $\mathbb{E}_\vartheta[\hat{g}] - g(\vartheta)$ misst die Verzerrung. Für den MSE ist die *Bias-Varianz-Zerlegung* von grundlegender Bedeutung:

$$\mathbb{E}_\vartheta[(\hat{g}-g(\vartheta))^2] = \mathbb{E}_\vartheta[((\hat{g}-\mathbb{E}_\vartheta[\hat{g}])+(\mathbb{E}_\vartheta[\hat{g}]-g(\vartheta)))^2] = \underbrace{(\mathbb{E}_\vartheta[\hat{g}] - g(\vartheta))^2}_{\text{quadrierter Bias}} + \underbrace{\mathrm{Var}_\vartheta(\hat{g})}_{\text{Varianz}}. \tag{1.1}$$

1.2 Parametrische und nichtparametrische Statistik

Die sogenannte parametrische Statistik betrachtet den Fall endlichdimensionaler Parameter, das heißt $\Theta \subseteq \mathbb{R}^k$. Auf Grund der differenzierbaren Struktur des $\mathbb{R}^k$ und einfacher Kompaktheitsargumente gibt es in der parametrischen Statistik starke Aussagen über Konstruktion und Eigenschaften von Schätzern. Häufig wird eine asymptotische Perspektive eingenommen, beispielsweise der Fall wachsenden Stichprobenumfangs oder fallenden Rauschniveaus. Wir erwähnen kurz ein Hauptresultat der Likelihood-Theorie.

Es sei $X_1, \dots, X_n$ eine mathematische Stichprobe, die bezüglich einer Lebesguedichte f_ϑ auf $\mathbb{R}$ verteilt sei mit $\vartheta \in \Theta \subseteq \mathbb{R}^k$ unbekannt. Mit $L(\vartheta, x) := f_\vartheta(x)$ wird die Likelihoodfunktion bezeichnet. Der Maximum-Likelihoodschätzer ist definiert als

$$\hat{\vartheta}_n := \mathrm{argmax}_{\vartheta\in\Theta} \prod_{i=1}^n L(\vartheta, X_i) = \mathrm{argmax}_{\vartheta\in\Theta} \sum_{i=1}^n \log(L(\vartheta, X_i)),$$

sofern dies wohldefiniert ist. Falls nun Θ eine offene Menge ist und $L(\vartheta, x)$ bezüglich ϑ differenzierbar ist mit Ableitung (Gradient) $\dot{L}(\vartheta, x)$, erhalten wir $\sum_{i=1}^n \dot{L}(\hat{\vartheta}_n, X_i)/L(\hat{\vartheta}_n, X_i) = 0$ als Schätzgleichung. Mit dem Gesetz der großen Zahlen und dem zentralen Grenzwertsatz kann man unter weiteren Regularitätsbedingungen (z.B. höherer Differenzierbarkeitsordnung) folgende Asymptotik für $n \to \infty$ nachweisen:

$$\sqrt{n}(\hat{\vartheta}_n - \vartheta) \overset{\mathbb{P}_\vartheta}{\Rightarrow} N(0, I^{-1}(\vartheta)), \quad \vartheta \in \Theta.$$

Dabei bezeichnet $I(\vartheta)$ die sogenannte Fisher-Informationsmatrix. Insbesondere ist in regulären Modellen die stochastische Konvergenzordnung $(\hat{\vartheta}_n - \vartheta) = \mathcal{O}_{\mathbb{P}_\vartheta}(n^{-1/2})$ bestmöglich. Die Rate $n^{-1/2}$ ist in den meisten parametrischen Modellen vom Umfang n typisch und leitet sich aus Varianten des (mehrdimensionalen) zentralen Grenzwertsatzes her.

In der nichtparametrischen Statistik ist die Parametermenge Θ unendlichdimensional, es wird keine einfache Parametrisierung des Modells vorgenommen. Häufig ist der unbekannte Parameter die Dichte der Beobachtungen selbst (Dichteschätzung) oder ein unbekannter funktionaler Zusammenhang. Bei der *Regression* werden die Beobachtungen modelliert durch

$$Y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

mit statistischen Fehlern (ε_i) (meist $\mathbb{E}[\varepsilon_i] = 0$) und deterministischen oder zufälligen *Designpunkten* $x_i \in D \subseteq \mathbb{R}^d$. Lineare Regression behandelt im einfachsten Fall lineare Regressionsfunktionen $f \in \mathcal{F} = \{g : D \to \mathbb{R} \mid g(x) = a^\top x + b,\, a \in \mathbb{R}^d,\, b \in \mathbb{R}\}$, während in der nichtparametrischen Regression die Funktionsklasse $\mathcal{F}$ aus allen Funktionen $g : D \to \mathbb{R}$ besteht, die gewisse allgemeine Bedingungen wie Stetigkeit, Monotonie oder Differenzierbarkeit erfüllen. Wegen fehlender Kompaktheits- oder Differenzierbarkeitseigenschaften im Parameterraum bedarf es in der nichtparametrischen Statistik neuer Methoden und mathematischer Analysen.

Selbst wenn es a priori gute Gründe gibt, ein parametrisches Modell anzunehmen, dienen nichtparametrische Verfahren häufig dazu, Modellmisspezifikationen anhand der Daten aufzudecken (goodness-of-fit-Tests). In der Praxis gibt es immer mehr hochdimensionale Daten und Modelle, zum Beispiel in der Bildverarbeitung, bei der Genanalyse oder dem Data-Mining. Wenn nicht gleichzeitig enorme Stichprobenumfänge vorliegen, greift die Asymptotik der parametrischen Statistik nicht und fast immer kommen nichtparametrische Verfahren zum Einsatz.

Schließlich seien noch ein paar Literaturhinweise gegeben. Zum Hintergrund parametrischer Schätztheorie siehe Lehmann and Casella (1998). Nichtparametrische Schätzmethoden und ihre mathematische Analyse werden inzwischen in vielen Büchern behandelt. Zur eher praktisch orientierten Dichteschätzung ist Silverman (1986) ein Klassiker, Wand and Jones (1995) ein etwas aktuelleres praxisorientiertes Lehrbuch zur Kernschätzung. Eine umfassende Monographie zur nichtparametrischen Regression haben Györfi, Kohler, Krzyżak, and Walk (2002) vorgelegt, Härdle (1991) behandelt dieses Thema mit Anwendungsbezug. Der Modellwahlansatz ist umfassend und gut aufbereitet in Massart (2007) zu finden. Aus einem Vorlesungsskript für Mathematiker hervorgegangen und für Theorievermittlung am empfehlenswertesten ist Tsybakov (2009). Eine umfassende Einführung in aktuelle nichtparametrische Methoden und insbesondere ihre Anwendungen im Gebiet des Statistischen Lernens gibt Hastie, Tibshirani, and Friedman (2001), während Efromovich (1999) eher breit auf unterschiedliche statistische Anwendungen eingeht. Schließlich sei Wasserman (2006) für eine breite und aktuelle Übersicht mit intuitiven Erklärungen (aber meist ohne Beweise) empfohlen.

2 Dichteschätzung

2.1 Modell und empirische Verteilung

Wir werden folgendes Modell für die Dichteschätzung betrachten, das als Grundlage für vielseitige Verallgemeinerungen und spezifische Anwendungen dient.

2.1 Definition. Es sei $\mathscr{F}_d := \{f : \mathbb{R}^d \to [0,\infty) \text{ messbar} \mid \int f = 1\}$ die Menge aller Lebesguedichten auf $\mathbb{R}^d$. Für ein unbekanntes $f \in \mathscr{F}_d$ beobachten wir eine Stichprobe $X_1, \dots, X_n \sim f$ i.i.d. vom Umfang n. Mit P_f und $\mathbb{E}_f$ wird die Wahrscheinlichkeit bzw. der Erwartungswert in diesem Modell bezeichnet. Für einen Schätzer $\hat{f}_n$ von f werden wir meist eines der folgenden Risiken betrachten:

Punktweises (quadratisches) Risiko: $R_x(\hat{f}_n, f) := \mathbb{E}_f[(\hat{f}_n(x) - f(x))^2]$ für ein $x \in \mathbb{R}^d$;

Quadratisches Risiko (MISE): $R_D(\hat{f}_n, f) := \mathbb{E}_f[\int_D (\hat{f}_n(x) - f(x))^2 dx]$ für eine messbare Teilmenge $D \subseteq \mathbb{R}^d$ (sofern $\hat{f}_n, f \in L^2(D)$);

Gleichmäßiges Risiko: $R_{D,\infty}(\hat{f}_n, f) := \mathbb{E}_f[\|\hat{f}_n(x) - f(x)\|_{L^\infty(D)}]$ für eine messbare Teilmenge $D \subseteq \mathbb{R}^d$ (sofern $\hat{f}_n, f \in L^\infty(D)$).

Grundidee jeder Dichteschätzung ist es, die empirische Verteilung von $(X_1, \dots, X_n)$ zu verwenden. Beachte dazu, dass im eindimensionalen Fall $d = 1$ die *empirische Verteilungsfunktion*

$$\hat{F}_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i \leqslant x), \quad x \in \mathbb{R},$$

die wahre Verteilungsfunktion F punktweise *erwartungstreu* und *konsistent* schätzt: $\mathbb{E}[\hat{F}_n(x)] = F(x)$ und $\lim_{n\to\infty} \hat{F}_n(x) = F(x)$ gilt fast sicher für alle $x \in \mathbb{R}^d$. Der ▶ Übung Satz von Glivenko-Cantelli sichert sogar gleichmäßige Konvergenz $\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| = 0$ fast sicher. Darüberhinaus liefert der zentrale Grenzwertsatz die Konvergenzrate $n^{-1/2}$: $\sqrt{n}(\hat{F}_n(x) - F(x)) \to N(0, F(x)(1-F(x))$.

Wenn wir nun wissen, dass eine Dichte f existiert, so gilt natürlich $F' = f$ (ggf. im schwachen Sinn), allerdings ist der naive Ansatz $\hat{f}_n(x) := \hat{F}_n'(x)$ nicht möglich, da die empirische Verteilungsfunktion nicht (im Funktionensinn) differenzierbar ist. Jedoch ist $\hat{F}_n$ die Verteilungsfunktion des *empirischen Maßes*

$$\hat{\mu}_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i},$$

wobei δ_x das Punkt- oder Diracmaß in x bezeichnet. Das empirische Maß ist auch im d-dimensionalen ein wohldefiniertes zufälliges Maß auf den Borelmengen von $\mathbb{R}^d$. Es sei bemerkt, dass unter unserer i.i.d.-Annahme $\hat{\mu}_n$

(wie auch $\hat{F}_n$) eine suffiziente Statistik ist und damit kein Informationsverlust beim Übergang von den Beobachtungen $X_1, \dots, X_n$ zu $\hat{\mu}_n$ auftritt. Man kann $\hat{\mu}_n$ als Ableitung von $\hat{F}_n$ im Distributionensinn interpretieren. Um jedoch zu einem funktionswertigen Schätzer $\hat{f}_n$ der Dichte f zu gelangen, muss $\hat{\mu}_n$ noch geglättet werden.

2.2 Kernschätzer

2.2 Definition. Eine messbare Funktion $K : \mathbb{R}^d \to \mathbb{R}$ mit $\int_{\mathbb{R}^d} K(x)\,dx = 1$ heißt *Kern* oder *Kernfunktion*. Man setzt für einen Kern K und eine *Bandweite* $h > 0$

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

so dass K_h wiederum Kernfunktion ist. Allgemeiner, jedoch nicht hier, werden auch reguläre Bandweitenmatrizen $H \in \mathbb{R}^{d\times d}$ sowie $K_H(x) = |\det(H^{-1})|K(H^{-1}x)$ betrachtet (der skalare Fall entspricht $H = \mathrm{diag}(h, \dots, h)$).

Kernfunktionen werden benutzt, um das empirische Maß zu glätten. Dies ist dieselbe Idee wie die der Diracfolgen in der Analysis. Für $h \to 0$ konvergiert K_h gegen δ_0 in dem Sinne, dass für Faltungen $g * K_h(x) := \int g(x-y)K_h(y)dy$ unter Regularitätsbedingungen an K und die Funktion $g : \mathbb{R}^d \to \mathbb{R}$ gilt

$$\lim_{h\to 0} g * K_h(x) = g(x) = g * \delta_0(x).$$

2.3 Definition. Für einen Kern K und eine Bandweite h definiert man den Kerndichteschätzer

$$\hat{f}_{n,h}(x) := K_h * \hat{\mu}_n(x) := \int_{\mathbb{R}^d} K_h(x-y)\hat{\mu}_n(dy) = \frac{1}{n}\sum_{i=1}^{n} K_h(x-X_i), \quad x \in \mathbb{R}^d\,.$$

Die Abhängigkeit von der Kernfunktion K wird in der Notation meist unterdrückt.

2.4 Beispiele.

(a) $K(x) = \mathbf{1}([-1/2, 1/2]^d)(x)$ ist Kern mit $K_h(x) = h^{-d}\mathbf{1}([-h/2, h/2]^d)$, so dass

$$\hat{f}_{n,h}(x) = \frac{1}{nh^d}\sum_{i=1}^{n} \mathbf{1}([-h/2, h/2]^d)(x-X_i) = \#\{i : |X_i - x|_\infty \leqslant h/2\}/(nh^d)$$

(für $d = 1$ heißt K *Rechteckkern* und $\hat{f}_{n,h}$ *Fensterschätzer*).

(b) Für $d = 1$ ist $K(x) = (1 - |x|)\mathbf{1}([-1, 1])(x)$ der *Dreieckskern*.

(c) Für $d = 1$ heißt $K(x) = \frac{3}{4\sqrt{5}}(1-x^2/5)\mathbf{1}_{[-\sqrt{5},\sqrt{5}]}(x)$ *Epanechnikov-Kern*. Dieser hat theoretisches Interesse, da der zugehörige Kernschätzer eine gewisse Optimalitätseigenschaft besitzt, vergleiche Silverman (1986).

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

(e) Eindimensionale Kerne $K_1, \dots, K_d$ können zum d-dimensionalen Produktkern $K(x) = \prod_{i=1}^d K_i(x_i)$, $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, kombiniert werden.

(f) Der *sinc-Kern* $K(x) = \pi^{-d}\prod_{i=1}^d \sin(x_i)/x_i$ ist ein Beispiel eines nicht überall positiven (und nur uneigentlich integrierbaren, bei Null durch $K(0) = \pi^{-d}$ stetig zu ergänzenden) Kerns, der wegen seiner Fourierdarstellung $\mathcal{F}K(u) := \int_{\mathbb{R}^d} K(x)e^{i\langle x,u\rangle}dx = \mathbf{1}([-1,1]^d)(u)$ wichtig ist. ▶ Übung Es gilt nämlich

$$\mathcal{F}\hat{f}_{n,h}(u) = \mathcal{F}(K_h * \hat{\mu}_n)(u) = \mathcal{F}K_h(u)\mathcal{F}\hat{\mu}_n(u) = \mathcal{F}K(hu)\hat{\varphi}_n(u)$$

mit der *empirischen charakteristischen Funktion* $\hat{\varphi}_n(u) := \frac{1}{n}\sum_{j=1}^n e^{i\langle u,X_j\rangle}$. Daher ergibt sich beim sinc-Kern gerade der *spektrale cut-off-Schätzer*:

$$\hat{f}_{n,h}(x) = \mathcal{F}^{-1}\Big(\hat{\varphi}_n\mathbf{1}([-h^{-1},h^{-1}]^d)\Big)(x).$$

2.5 Lemma. ▶ Übung *Ist die Kernfunktion K eine Wahrscheinlichkeitsdichte (d.h. nichtnegativ), so ist der Kerndichteschätzer wiederum eine Wahrscheinlichkeitsdichte. Ist K keine Wahrscheinlichkeitsdichte, so ist* $\max(\hat{f}_{n,h}, 0)$ *stets eine Verbesserung von $\hat{f}_{n,h}$ für die oben angegebenen Risiken, allerdings ist auch dieser Schätzer im Allgemeinen keine Wahrscheinlichkeitsdichte.*

2.6 Bemerkung. Selbst im Fall des Rechteckkerns darf der Kerndichteschätzer nicht mit einem ▶ Übung *Histogramm* verwechselt werden. ▶ Übung Verallgemeinerungen des Kernschätzers bilden die *lokalen Polynomschätzer* sowie *kNN-Schätzer* (*kth nearest neighbour*).

2.3 Bias-Varianz-Dilemma

Der mittlere quadratische Fehler eines Kerndichteschätzers lässt sich leicht bestimmen.

2.7 Satz. *Für den Kerndichteschätzer $\hat{f}_{n,h}$ gilt:*

$$R_x(\hat{f}_{n,h}, f) = (K_h * f - f)(x)^2 + \frac{1}{n}\big((K_h^2 * f)(x) - (K_h * f)^2(x)\big),$$
$$R_D(\hat{f}_{n,h}, f) = \int_D \Big((K_h * f - f)(x)^2 + \frac{1}{n}\big((K_h^2 * f)(x) - (K_h * f)^2(x)\big)\Big)\,dx.$$

Beweis. Nach der Bias-Varianz-Zerlegung (1.1) folgt

$$\begin{aligned}\mathbb{E}_f[(\hat{f}_{n,h}(x) - f(x))^2] &= (\mathbb{E}_f[\hat{f}_{n,h}(x) - f(x)])^2 + \operatorname{Var}_f(\hat{f}_{n,h}(x)) \\ &= \Big(\int K_h(x-y) f(y)\, dy - f(x) \Big)^2 + \frac{1}{n} \operatorname{Var}_f(K_h(x - X_1)) \\ &= (K_h * f - f)(x)^2 + \frac{1}{n}\big((K_h^2 * f)(x) - (K_h * f)^2(x)\big).\end{aligned}$$

Integration über $x \in D$ ergibt nach dem Satz von Tonelli das zweite Ergebnis. □

Wir wollen uns die oberen Fehlerschranken in Hinblick auf ihre Größenordnungen anschauen. Wir beschränken uns beispielhaft auf das punktweise Risiko. Der Bias-Term $(K_h * f - f)(x)^2$ ist unabhängig vom Stichprobenumfang n und konvergiert für $h \to 0$ im Allgemeinen (z.B. falls f stetig bei x und K mit kompaktem Träger) gegen Null. Der Varianzterm hingegen ist von der Ordnung n^{-1} im Stichprobenumfang, und für hinreichend reguläre Kernfunktion K und Dichte f folgt für $h \to 0$

$$K_h^2 * f(x) = h^{-d} \int_{\mathbb{R}^d} K^2(w) f(x - hw)\, dw = \mathcal{O}(h^{-d}),$$

während $K_h * f(x) = \mathcal{O}(1)$ gilt. Uns offenbart sich das *Bias-Varianz-Dilemma*: je kleiner die Bandweite h gewählt wird, desto unverzerrter ist die Schätzung, desto größer ist jedoch andererseits ihre Varianz. Dies ist auch intuitiv einsichtig, weil der Schätzer bei kleinerem h weniger stark geglättet wird, so dass zwar Details besser aufgelöst werden, es aber auch zu vermehrten Oszillationen kommt.

Die Bandweite h sollte vom Statistiker idealerweise so gewählt werden, dass das gesamte Risiko minimal ist:

$$h^* := \operatorname{argmin}_{h>0} R_x(\hat{f}_{n,h}, f). \tag{2.1}$$

Betrachtet man jedoch Bias- und Varianz-Term, so stellt man fest, dass diese nicht nur von den bekannten Größen n und K, sondern auch von der unbekannten und gerade zu schätzenden Dichtefunktion f abhängen. Die Bandweite h^* ist also in praxi nicht bekannt und kann nur als theoretische Messlatte dienen bei der Wahl der Bandweite durch den Statistiker. Man nennt h^* aus naheliegenden Gründen *Orakel-Bandweite*. Im folgenden werden wir zunächst den Minimax-Ansatz zur Bandweitenwahl untersuchen.

2.4 Glattheitsklassen und asymptotisches Risiko

2.8 Lemma. *Der Kern K liege in $L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ und f sei beschränkt auf $\mathbb{R}^d$ sowie stetig bei $x \in \mathbb{R}^d$. Wählt man eine Folge $h_n \to 0$ mit $h_n^d n \to \infty$ für $n \to \infty$, so ist $\hat{f}_{n,h_n}(x)$ ein konsistenter Schätzer von $f(x)$.*

Beweis. Wir wenden Satz 2.7 an und schließen mit dominierter Konvergenz:

$$K_{h_n} * f(x) = \int_{\mathbb{R}^d} f(x - h_n z) K(z)\, dz \to \int_{\mathbb{R}^d} f(x) K(z)\, dz = f(x),$$
$$h_n^d K_{h_n}^2 * f(x) = \int_{\mathbb{R}^d} f(x - h_n z) K(z)^2\, dz \to f(x) \int_{\mathbb{R}^d} K(z)^2\, dz.$$

Mit $h_n^{-d} n^{-1} \to 0$ schließen wir, dass $R_x(\hat{f}_{n,h_n}, f)$ gegen Null konvergiert. □

Dies lässt viel Freiheit für die asymptotische Wahl der Bandweite und kann zu beliebig langsamer Konvergenz führen. Sofern die Dichtefunktion f beliebig aus $\mathscr{F}_d$ sein kann, haben wir jedoch keinen Ansatzpunkt für eine geeignete Wahl der Bandweite h: f kann sowohl stark oszillierend als auch von sehr geringer Variation sein. Eine natürliche Annahme ist daher, von f eine gewisse Regularität vorauszusetzen. Wir beschreiben diese Vorkenntnis durch Normschranken in Glattheitsklassen.

2.9 Definition. Setze $\langle x \rangle := \max\{m \in \mathbb{N} \mid m < x\}$ mit strikter Ungleichung. Ist $\beta \in \mathbb{N}^d$ ein Multiindex, so bezeichnet $g^{(\beta)}$ für $g : \mathbb{R}^d \to \mathbb{R}$ die Ableitung $\frac{\partial^{|\beta|} g}{\partial x_1^{\beta_1} \cdots \partial x_d^{\beta_d}}$ mit $|\beta| = \sum_{k=1}^d \beta_k$.

Für $\alpha > 0$ und eine offene Menge $D \subseteq \mathbb{R}^d$ sagen wir, dass $f : \mathbb{R}^d \to \mathbb{R}$ in $C^\alpha(D)$ liegt, sofern f $\langle\alpha\rangle$-mal stetig differenzierbar auf D ist und jede Ableitung der Ordnung $\beta \in \mathbb{N}^d$ mit $|\beta| = \langle\alpha\rangle$ die Hölder-Bedingung

$$\sup_{x,y \in D,\, x \neq y} \frac{|f^{(\beta)}(x) - f^{(\beta)}(y)|}{|x - y|^{\alpha - \langle\alpha\rangle}} < \infty$$

erfüllt. Als *Hölderklasse* $\mathscr{H}_D(\alpha; R, L)$ mit Parametern $\alpha, R, L > 0$ auf $D \subseteq \mathbb{R}^d$ bezeichnen wir die Menge

$$\Big\{ f \in C^\alpha(D) \,\Big|\, \sup_{x \in D} |f(x)| \leqslant R,\ \max_{|\beta| = \langle\alpha\rangle} \sup_{x,y \in D,\, x \neq y} \frac{|f^{(\beta)}(x) - f^{(\beta)}(y)|}{|x - y|^{\alpha - \langle\alpha\rangle}} \leqslant L \Big\}.$$

Im einfachsten Fall $\alpha \leqslant 1$ und $D = \mathbb{R}^d$ kann die Hölderannahme direkt zur Beschränkung des Bias-Terms verwendet werden:

$$\begin{aligned} |(f * K_h - f)(x)| &= \Big| \int (f(\xi) - f(x)) K_h(x - \xi)\, d\xi \Big| \\ &\leqslant \int L |\xi - x|^\alpha |K_h(x - \xi)|\, d\xi \\ &= L h^\alpha \int |w|^\alpha |K(w)|\, dw. \end{aligned}$$

Sofern das letzte Integral endlich ist, ergibt sich also die Ordnung $\mathcal{O}(h^\alpha)$ und zwar gleichmäßig über alle $f \in \mathscr{H}_{\mathbb{R}^d}(\alpha; R, L)$. Für $\alpha > 1$ lässt sich diese Rate verbessern, sofern die Kernfunktion eine polynomiale Exaktheitsbedingung erfüllt.

2.10 Definition. Ein Kern $K : \mathbb{R}^d \to \mathbb{R}$ ist von der *Ordnung* $m \in \mathbb{N}_0$, sofern für alle Multiindizes $\beta \in \mathbb{N}^d$ mit $|\beta| \in \{1, \dots, m\}$ gilt

$$\int_{\mathbb{R}^d} x^\beta K(x)\, dx = 0 \qquad (x^\beta := x_1^{\beta_1} \cdots x_d^{\beta_d}).$$

2.11 Beispiele.

(a) $K(x) = \mathbf{1}([-1/2, 1/2]^d)(x)$ besitzt die Ordnung 1, jedoch nicht die Ordnung 2. Dies gilt auch für den Dreieckskern, den Epanechnikov-Kern und den Gauß-Kern und allgemein für nichtnegative Kerne, weil für sie $\int x_1^2 K(x)\, dx$ stets strikt positiv ist.

(b) Der quadratische Kern $K(x) = \frac{9-15x^2}{8}\mathbf{1}([-1,1])(x)$ erfüllt $\int K = 1$, $\int x^{2m-1}K(x)\, dx = 0$ für $m \in \mathbb{N}$ (aus Symmetrie) und $\int x^2 K(x)dx = 0$, $\int x^4 K(x)dx = \frac{9}{40} - \frac{15}{56} \neq 0$. Also ist K ein Kern der Ordnung 3. ▶ Übung Man kann allgemein zeigen, dass für jedes $p \in \mathbb{N}$ genau ein Polynom P vom Grad höchstens p existiert, so dass $K(x) = P(x)\mathbf{1}_{[-1,1]}(x)$ ein Kern der Ordnung p ist.

(c) Der sinc-Kern K ist eigentlich prädestiniert, als Kern beliebiger Ordnung (sogenannter *Superkern*) zu dienen, da Momente durch Ableitungen bei Null im Fourierbereich berechnet werden und $\mathcal{F}K$ dort konstant ist. Leider ist jedoch $\int x^p K(x)dx$ nicht wohldefiniert als Lebesgue-Integral. Ist jedoch $g \in C^\infty(\mathbb{R}^d)$ eine Funktion mit kompaktem Träger und $g(0) = 1$, $D^\alpha g(0) = 0$, $|\alpha| \geqslant 1$ (ein Beispiel ist $g(x) = \exp(2(1 - (x^2+1)/(x^2-1)^2))\mathbf{1}_{[-1,1]}(x)$), so gilt für $K = \mathcal{F}^{-1}g$ die Kerneigenschaft $\int K = \mathcal{F}K(0) = 1$ sowie für $|\beta| \geqslant 1$

$$\int x^\beta K(x)\, dx = \Big(\int K(x) D_u^\beta e^{i\langle u,x\rangle} i^{-|\beta|} dx \Big)\Big|_{u=0} = i^{-|\beta|} D_u^\beta \mathcal{F}K(0) = 0.$$

Damit ist ein solches K ein Superkern.

2.12 Bemerkung. Manchmal wird zusätzlich die Bedingung $\int |x|^{m+1}|K(x)|dx < \infty$ für einen Kern der Ordnung m gefordert. Dies garantiert eine endliche Schranke im folgenden Lemma.

2.13 Lemma. *Es gelte $f \in \mathscr{F}_d \cap \mathscr{H}_U(\alpha; R, L)$ für eine Umgebung U von x und K besitze die Ordnung $\langle\alpha\rangle$ sowie einen kompakten Träger. Dann gilt für hinreichend kleines $h > 0$*

$$|(f * K_h - f)(x)| \leqslant h^\alpha L \tfrac{d^{\langle\alpha\rangle}}{\langle\alpha\rangle!} \int |w|^\alpha |K(w)|\, dw.$$

Beweis. Wir benutzen die Taylorentwicklung um x für alle y in einer Kugel $B \subseteq U$ um x

$$f(y) = f(x) + \sum_{|\beta|<\langle\alpha\rangle} f^{(\beta)}(x)\frac{(y-x)^\beta}{\beta!} + \sum_{|\beta|=\langle\alpha\rangle} f^{(\beta)}(\tau_y)\frac{(y-x)^\beta}{\beta!}$$

mit einer Zwischenstelle $\tau_y = x+\rho(y-x)$, $\rho \in [0,1]$, und $\beta! = \beta_1! \cdots \beta_d!$. Wegen des kompakten Trägers von K erstreckt sich das Integral $\int f(y)K_h(x-y)dy$ für hinreichend kleines h nur über B. Die Kerneigenschaft $\int K_h = 1$ sowie die Ordnung von K und damit von K_h ergeben somit

$$\begin{aligned}
|(f * K_h - f)(x)| &= \Big| \int_{\mathbb{R}^d} (f(y) - f(x)) K_h(x-y)\, dy \Big| \\
&\leqslant \sum_{|\beta|<\langle\alpha\rangle} \Big| \int_{\mathbb{R}^d} f^{(\beta)}(x) \frac{(y-x)^\beta}{\beta!} K_h(x-y)\, dy \Big| \\
&\quad + \sum_{|\beta|=\langle\alpha\rangle} \Big| \int_{\mathbb{R}^d} f^{(\beta)}(\tau_y) \frac{(y-x)^\beta}{\beta!} K_h(x-y)\, dy \Big| \\
&= \sum_{|\beta|=\langle\alpha\rangle} \Big| \int_{\mathbb{R}^d} (f^{(\beta)}(\tau_y) - f^{(\beta)}(x)) \frac{(y-x)^\beta}{\beta!} K_h(x-y)\, dy \Big| \\
&\leqslant \sum_{|\beta|=\langle\alpha\rangle} \int_{\mathbb{R}^d} L|y-x|^{\alpha-\langle\alpha\rangle} \frac{|(y-x)^\beta|}{\beta!} |K_h(x-y)|\, dy \\
&\leqslant Lh^\alpha \int_{\mathbb{R}^d} |z|^\alpha |K(z)|\, dz \sum_{|\beta|=\langle\alpha\rangle} \frac{1}{\beta!}.
\end{aligned}$$

Die letzte Summe lässt sich exakt bestimmen über einen Potenzreihenansatz [1]. Aus $e^{x_1+\cdots+x_d} = \sum_\beta x^\beta/\beta!$ folgt $e^{dx} = \sum_{m=0}^\infty \sum_{|\beta|=m} x^m/\beta!$. Da andererseits $e^{dx} = \sum_{m=0}^\infty x^m d^m/m!$ gilt, ergibt ein Koeffizientenvergleich $\sum_{|\beta|=m} 1/\beta! = d^m/m!$. □

2.14 Bemerkung. Wie der Beweis zeigt, gilt das Resultat auch, falls K keinen kompakten Träger besitzt, jedoch $U = \mathbb{R}^d$ betrachtet wird.

2.15 Korollar. *Unter den Voraussetzungen des vorangegangenen Satzes ist der Bias des Kerndichteschätzers von der Ordnung $\mathcal{O}(h^\alpha)$; genauer gilt:*

$$|\mathbb{E}_f[\hat{f}_{n,h}(x) - f(x)]| \leqslant CLh^\alpha$$

mit $C = d^{\langle\alpha\rangle} \int |w|^\alpha |K(w)|\, dw/\langle\alpha\rangle!$.

Beweis. Dies folgt unmittelbar aus der Bias-Darstellung und dem vorangegangenen Satz. □

Da wir die wahre Dichte f nicht kennen, aber voraussetzen, dass sie in der Klasse $\mathscr{H}_D(\alpha; L, R)$ mit bekanntem $\alpha, L, R > 0$ liegt, können wir die Bandweite h so wählen, dass das maximale Risiko über diese Klasse möglichst klein wird (sogenannter *Minimax-Ansatz*).

[1] Dank an Martin Wahl für diesen Trick!

2.16 Satz. *Es seien $\alpha, L, R > 0$, $D \subseteq \mathbb{R}^d$ offen und $K \in L^2(\mathbb{R}^d)$ ein Kern der Ordnung $\langle\alpha\rangle$ mit kompaktem Träger. $C = C(\alpha, d, K)$ bezeichne die Konstante aus Korollar 2.15. Für jedes $x \in D$ gilt bei hinreichend kleinem $h > 0$*

$$\sup_{f \in \mathscr{F}_d \cap \mathscr{H}_D(\alpha;L,R)} R_x(\hat{f}_{n,h}, f) \leqslant h^{2\alpha} C^2 L^2 + n^{-1} h^{-d} R \|K\|_{L^2}^2.$$

Die rechte Seite wird minimal bei der Wahl

$$h^\star = \left(n^{-1} (2\alpha)^{-1} R d \|K\|_{L^2}^2 C^{-2} L^{-2} \right)^{1/(2\alpha+d)},$$

und es folgt

$$\sup_{f \in \mathscr{F}_d \cap \mathscr{H}_D(\alpha;L,R)} R_x(\hat{f}_{n,h^\star}, f) \leqslant \left(n^{-1} R \|K\|_{L^2}^2 \right)^{2\alpha/(2\alpha+d)} \left(2\alpha C^2 L^2 d^{-1} \right)^{d/(2\alpha+d)}.$$

Insbesondere gilt für den maximalen Fehler

$$\sup_{f \in \mathscr{F}_d \cap \mathscr{H}_D(\alpha;L,R)} R_x(\hat{f}_{n,h^\star}, f) = \mathcal{O}(R^{2\alpha/(2\alpha+d)} L^{2d/(2\alpha+d)} n^{-2\alpha/(2\alpha+d)}).$$

Beweis. Einsetzen und Nachrechnen. □

2.17 Beispiel. Für Lipschitz-stetiges f und $d = 1$ ist das quadratische Risiko von der Ordnung $\mathcal{O}(n^{-2/3})$, für $f \in C^2(\mathbb{R})$ erhalten wir $\mathcal{O}(n^{-4/5})$. Im Grenzfall $\alpha \to \infty$ kann sich $\mathcal{O}(n^{-1})$ ergeben, also die gewöhnliche parametrische Konvergenzrate. Dabei ist zu beachten, dass die Schranke L natürlich von α abhängt und nicht notwendigerweise beschränkt bleibt.

Je größer die Dimension d ist, desto schlechter ist die Konvergenzrate (*Fluch der Dimension*, vergleiche auch Tabelle 4.2 in Silverman (1986)). Am exakten Ergebnis kann man auch erkennen, dass der Kern K möglichst um die Null herum konzentriert sein sollte mit kleiner L^2-Norm (unter der Restriktion durch $\int K = 1$ und die Ordnung $\langle\alpha\rangle$).

Wenden wir uns dem MISE zu, so lassen sich alle Resultate übertragen durch Integration über das betrachtete Gebiet D. Allerdings lassen sich für $D = \mathbb{R}^d$ und $f, K \in L^2(\mathbb{R}^d)$ bessere und transparentere Abschätzungen durch Übergang in den Spektralbereich gewinnen. Hauptwerkzeug ist dabei die *Plancherel-Gleichung* (vgl. Werner (2007), jedoch mit anderer 2π-Normierung)

$$\int_{\mathbb{R}^d} |\mathcal{F}g(u)|^2 du = (2\pi)^d \int_{\mathbb{R}^d} |g(x)|^2 dx \text{ für beliebiges } g \in L^2(\mathbb{R}^d).$$

Wendet man diese auf die Spektraldarstellung des Kerndichteschätzers an (vergleiche Beispiel 2.4), so ergibt sich mit der charakteristischen Funktion $\varphi(u) = \mathcal{F}f(u)$

$$\int_{\mathbb{R}^d} (\hat{f}_{n,h}(x) - f(x))^2 dx = (2\pi)^{-d} \int_{\mathbb{R}^d} |\mathcal{F}K(hu)\hat{\varphi}_n(u) - \varphi(u)|^2 du \qquad (2.2)$$

Aus ▶ Übung $\mathbb{E}_f[\hat{\varphi}_n(u)] = \varphi(u)$ und $\mathbb{E}_f[|\hat{\varphi}_n(u) - \varphi(u)|^2] = n^{-1}(1 - |\varphi(u)|^2)$ erhalten wir die Bias-Varianz-Zerlegung im Spektralbereich

$$R_{\mathbb{R}^d}(\hat{f}_{n,h}, f) = (2\pi)^{-d}\Big(\|(\mathcal{F}K(h\bullet)-1)\varphi\|_{L^2}^2+n^{-1}\Big(\|\mathcal{F}K(h\bullet)\|_{L^2}^2-\|\mathcal{F}K(h\bullet)\varphi\|_{L^2}^2\Big)\Big).$$

Wegen $\int K = 1$ gilt $\mathcal{F}K(0) = 1$, und wir sehen, dass der Bias-Term für $h \to 0$ gegen Null konvergiert (sofern die Vertauschung von Grenzwert und Integral zulässig ist). Der Varianzterm ist kleiner als $n^{-1}\|\mathcal{F}K(h\bullet)\|_{L^2}^2 = n^{-1}h^{-d}\|\mathcal{F}K\|_{L^2}^2$ und damit wiederum von der Ordnung $\mathcal{O}(n^{-1}h^{-d})$. Wir fassen zusammen.

2.18 Lemma. *Für den MISE des Kerndichteschätzers mit $f, K \in L^2$ gilt*

$$R_{\mathbb{R}^d}(\hat{f}_{n,h}, f) = (2\pi)^{-d}\Big(\|(\mathcal{F}K(h\bullet)-1)\mathcal{F}f\|_{L^2}^2+n^{-1}h^{-d}\|\mathcal{F}K\|_{L^2}^2-n^{-1}\|\mathcal{F}K(h\bullet)\mathcal{F}f\|_{L^2}^2\Big).$$

2.19 Bemerkung. Diese Abschätzung kann auch vollständig im Ortsbereich hergeleitet und formuliert werden:

$$R_{\mathbb{R}^d}(\hat{f}_{n,h}, f) = \|K_h * f - f\|_{L^2}^2 + n^{-1}h^{-d}\|K\|_{L^2}^2 - n^{-1}\|K_h * f\|_{L^2}^2.$$

Beachte auch hier, dass der letzte Term $\mathcal{O}(n^{-1})$ für $n \to \infty$ und $h \to 0$ ist und damit eine kleinere Größenordnung als der zweite Term besitzt.

Während wir den Approximationsfehler zuvor mittels Taylorentwicklung abgeschätzt haben, sehen wir nun, dass dieser klein ist, wenn $|\varphi(u)|$ dort klein ist, wo $\mathcal{F}K(hu)$ weit von 1 abweicht. Da $\mathcal{F}K$ stetig ist mit $\lim_{u\to\pm\infty} \mathcal{F}K(u) = 0$ für $K \in L^1(\mathbb{R})$ (Riemann-Lebesgue-Lemma), sollte $|\varphi(u)|$ für $u \to \pm\infty$ hinreichend schnell abfallen.

2.20 Definition. Der L^2-*Sobolevraum* der Ordnung $s \geqslant 0$ ist definiert als

$$H^s(\mathbb{R}^d) := \Big\{g \in L^2(\mathbb{R}^d) \,\Big|\, \int_{\mathbb{R}^d}(1 + |u|^2)^s|\mathcal{F}g(u)|^2du < \infty\Big\}.$$

Dies ist ein Hilbertraum bezüglich dem Skalarprodukt

$$\langle g, h\rangle_s := \int_{\mathbb{R}^d}(1 + |u|^2)^s\mathcal{F}g(u)\overline{\mathcal{F}h(u)}\,du.$$

Für $s \in \mathbb{N}$ kann die Sobolevnorm auch mit Hilfe schwacher Ableitungen direkt definiert werden: es gilt $f \in H^s(\mathbb{R}^d)$, wenn f s-mal schwach differenzierbar ist sowie f und alle Ableitungen quadrat-integrierbar sind. Insbesondere liegt eine s-fach klassisch differenzierbare Funktion $g \in L^2(\mathbb{R})$ mit $g^{(s)} \in L^2(\mathbb{R})$ in $H^s(\mathbb{R})$.

2.21 Beispiel. Die Laplace-Dichte $f(x) = \frac{1}{2}e^{-|x|}$ besitzt die Fouriertransformierte $\mathcal{F}f(u) = (1 + u^2)^{-1}$, so dass f in $H^s(\mathbb{R})$ liegt für alle $s < 3/2$. Wegen $f'(x) = -\operatorname{sgn}(x)\frac{1}{2}e^{-|x|}$ im schwachen Sinn und $f' \in L^2(\mathbb{R})$ sieht man direkt zumindest, dass $f \in H^1(\mathbb{R})$ gilt, obgleich $f \notin C^1(\mathbb{R})$.

Folgendes Resultat ist klassisch, siehe z.B. Werner (2007).

2.22 Satz (Soboleveinbettungssatz)**.** *Für $s > \alpha + d/2$ gilt $H^s(\mathbb{R}^d) \hookrightarrow C^\alpha(\mathbb{R}^d)$: jedes $f \in H^s(\mathbb{R}^d)$ besitzt eine Version in $C^\alpha(\mathbb{R}^d)$.*

Für Sobolevklassen ist der Biasterm im MISE des Kerndichteschätzers leicht abzuschätzen.

2.23 Satz. *Es sei $K \in L^1 \cap L^2(\mathbb{R}^d)$ ein Kern der Ordnung $\langle s \rangle$ mit $\mathcal{F}K \in C^{\langle s \rangle + 1}(\mathbb{R}^d)$ und beschränkten Ableitungen der Ordnung $\langle s \rangle + 1$. Für den MISE des Kerndichteschätzers mit $f \in H^s(\mathbb{R}^d)$ für $s > 0$ gilt*

$$R_{\mathbb{R}^d}(\hat{f}_{n,h}, f) \leqslant C^2 \|f\|_s^2 h^{2s} + n^{-1} h^{-d} \|K\|_{L^2}^2$$

mit $C = (2\pi)^{-d/2}(\|K\|_{L^1} + 1)(\sum_{|\beta| = \langle s \rangle + 1} \frac{\|\mathcal{F}K^{(\beta)}\|_\infty}{\beta!(\|K\|_{L^1}+1)})^{s/(\langle s \rangle + 1)}$. Für nichtnegative Kerne sowie $d = 1$ und $s \in \mathbb{N}$ ergibt sich $C = \frac{\|\mathcal{F}K^{(s)}\|_\infty}{s!\sqrt{2\pi}}$.

2.24 Bemerkung. Nach der Fouriertheorie folgt $\mathcal{F}K \in C^{\langle s \rangle + 1}(\mathbb{R}^d)$ mit gleichmäßig beschränkten Ableitungen aus der Momentenbedingung $\int |K(x)||x|^{\langle s \rangle + 1}\, dx < \infty$, insbesondere also für Kerne mit kompaktem Träger. ▶ Übung Ein einfacherer Beweis ist möglich für Kerne mit kompaktem Träger im Fourierbereich, zum Beispiel für den sinc-Kern.

Beweis. Die Ordnung von K impliziert im Fourierbereich $\mathcal{F}K^{(\beta)}(0) = 0$ für $|\beta| \in \{1, \dots, \langle s \rangle\}$. Mittels Taylorentwicklung von $\mathcal{F}K$ um Null erhalten wir daher

$$\begin{aligned}
&\|(\mathcal{F}K(h\bullet) - 1)\mathcal{F}f\|_{L^2}^2 \\
&= \int |\mathcal{F}K(hu) - 1|^2 (1 + |u|^2)^{-s} (1 + |u|^2)^s |\mathcal{F}f(u)|^2\, du \\
&\leqslant \|f\|_s^2 \sup_{u \in \mathbb{R}^d} |\mathcal{F}K(hu) - 1|^2 (1 + |u|^2)^{-s} \\
&\leqslant \|f\|_s^2 \sup_{u \in \mathbb{R}^d} \Big(\Big(|hu|^{\langle s \rangle + 1} \sum_{|\beta| = \langle s \rangle + 1} \frac{\|\mathcal{F}K^{(\beta)}\|_\infty}{\beta!} \Big) \wedge (\|K\|_{L^1} + 1) \Big)^2 |u|^{-2s} \\
&= \|f\|_s^2 h^{2s} \sup_{v \in \mathbb{R}^d} \Big(\Big(|v|^{\langle s \rangle + 1 - s} \sum_{|\beta| = \langle s \rangle + 1} \frac{\|\mathcal{F}K^{(\beta)}\|_\infty}{\beta!} \Big) \wedge (\|K\|_{L^1} + 1)|v|^{-s} \Big)^2 \\
&= \|f\|_s^2 h^{2s} (\|K\|_{L^1} + 1)^2 \Big(\sum_{|\beta| = \langle s \rangle + 1} \frac{\|\mathcal{F}K^{(\beta)}\|_\infty}{\beta!(\|K\|_{L^1} + 1)} \Big)^{2s/(\langle s \rangle + 1)}.
\end{aligned}$$

Wir schließen mittels Lemma 2.18, wobei wir den letzten Summanden dort durch Null abschätzen. Die Vereinfachung folgt aus $\|K\|_{L^1} = 1$ für nichtnegative Kerne K durch Einsetzen. □

Minimieren des MISE bezüglich h ergibt die Konvergenzrate $\mathcal{O}(n^{-2s/(2s+d)})$.

2.25 Korollar. *Unter den Voraussetzungen und in der Notation des vorigen Satzes gilt für $M > 0$ und mit $h^\star = (n^{-1}(2s)^{-1}d\|K\|_{L^2}^2 C^{-2}M^{-2})^{1/(2s+d)}$*

$$\sup_{f\in\mathscr{F}_d\cap H^s(\mathbb{R}^d),\,\|f\|_s\leqslant M} R_{\mathbb{R}^d}(\hat{f}_{n,h^\star}, f) \leqslant \left(n^{-1}\|K\|_{L^2}^2\right)^{2s/(2s+d)} \left(2sC^2M^2d^{-1}\right)^{d/(2s+d)}.$$

Insbesondere ist für Sobolevbälle der Ordnung $s > 0$ mit Radius M der maximale MISE von der Ordnung $\mathcal{O}(M^{2d/(2s+d)}n^{-2s/(2s+d)})$ in n und M.

3 Nichtparametrische Regression

3.1 Modell und Signal in weißem Rauschen

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

3.1 Definition. Gegeben sei eine Stichprobe $(X_i, Y_i)_{i=1,\dots,n}$ von i.i.d. Beobachtungen mit $X_i \in D \subseteq \mathbb{R}^d$, $Y_i \in \mathbb{R}$. Mit $f(x) := \mathbb{E}[Y_i \,|\, X_i = x]$, $x \in D$, werde die *Regressionsfunktion* bezeichnet (die existieren möge), so dass äquivalent folgendes Modell beobachtet wird:

$$(X_i, Y_i) \text{ mit } Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

mit $(\varepsilon_i)_{i=1,\dots,n}$ i.i.d. und $\mathbb{E}[\varepsilon_i \,|\, X_i] = 0$. Dies ist das *Regressionsmodell mit zufälligem Versuchsplan (Design)* und Ziel ist statistische Inferenz für die Regressionsfunktion f. Die Y_i werden als *Antwortvariablen* (*response variables*), die X_i als *Regressor-, Prädiktor- oder Kovariablen* (*regressor/predictor variables, covariates*) und die ε_i als *Fehlervariablen* (*error variables*) bezeichnet.

Sind $x_1, \dots, x_n \in D \subseteq \mathbb{R}^d$ deterministisch, so wird das analoge Beobachtungsmodell

$$Y_i = f(x_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

mit $(\varepsilon_i)_{i=1,\dots,n}$ i.i.d. und $\mathbb{E}[\varepsilon_i] = 0$ als Regressionsmodell mit *deterministischem Versuchsplan (Design)* bezeichnet. Die weiteren Bezeichnungen stimmen überein. Ist die Regressionsfunktion f nicht endlich-dimensional parametrisiert, so spricht man von *nichtparametrischer Regression.*

Das prototypische Regressionsmodell ist gegeben durch äquidistante Beobachtungen auf dem Einheitsintervall und normalverteilte Fehler

$$Y_i = f(i/n) + \varepsilon_i, \quad i = 1, \dots, n, \quad \varepsilon_i \sim N(0, \sigma^2) \text{ i.i.d.} \tag{3.1}$$

Ein kontinuierliches Analogon dazu ist gegeben durch das Modell eines *Signals in weißem Rauschen*:

$$dY_t = f(t) + \frac{\sigma}{\sqrt{n}}\, dW_t, \quad t \in [0,1], \tag{3.2}$$

mit einer Standard-Brownschen Bewegung W (die verallgemeinerte Ableitung von W wird auch weißes Rauschen genannt). Das heißt, in diesem Modell beobachtet man insbesondere die Inkremente $Y_b - Y_a = \int_a^b f(t)dt + \frac{\sigma}{\sqrt{n}}(W_b - W_a)$ für $0 \leqslant a < b \leqslant 1$. Man beobachtet also auch den Differenzenquotienten

$$\bar{Y}_i := \frac{Y_{i/n} - Y_{(i-1)/n}}{1/n} = n \int_{(i-1)/n}^{i/n} f(t)\, dt + \bar{\varepsilon}_i$$

mit $\bar{\varepsilon}_i = \sigma\sqrt{n}(W_{i/n} - W_{(i-1)/n}) \sim N(0, \sigma^2)$, $i = 1, \ldots, n$ i.i.d. Besitzt das Signal f nun eine gewisse Hölderregularität $\alpha > 0$, so gilt

$$\bar{f}(i/n) := n \int_{(i-1)/n}^{i/n} f(t)\, dt = f(i/n) + \mathcal{O}(n^{-\alpha}).$$

Insbesondere können wir also aus dem Signal in weißem Rauschen das Regressionsmodell approximativ zurückgewinnen:

$$\bar{Y}_i = \bar{f}(i/n) + \varepsilon_i, \quad i = 1, \ldots, n, \quad \varepsilon_i \sim N(0, \sigma^2) \text{ i.i.d.}$$

Diese Äquivalenz im statistischen Sinn ist von Le Cam mathematisch exakt formuliert worden und Brown and Low (1996) haben in der Tat bewiesen, dass für Funktionenklassen $\mathcal{F} \subseteq \mathscr{H}_D(\alpha; R, L)$ mit $\alpha > 1/2$ die statistischen Experimente generiert durch (3.1) und (3.2) asymptotisch äquivalent sind. Das Modell des Signals in weißem Rauschen lässt im Allgemeinen sehr viel transparentere Konstruktionen und Beweise zu, wie wir im folgenden noch sehen werden.

3.2 Lokal-polynomiale Schätzer

Da wir schon einige Möglichkeiten zur Dichteschätzung kennengelernt haben, schauen wir zunächst, wie sich diese Ideen auf die Schätzung der Regressionsfunktion übertragen lassen. Beachte dazu, dass wir mit der bedingten Dichte $f_{Y|X=x}$ von Y_i gegeben $X_i = x$ und der gemeinsamen Dichte $f_{X,Y}$ folgende Darstellung erhalten:

$$f(x) = \mathbb{E}[Y_i \,|\, X_i = x] = \int_{\mathbb{R}} y f_{Y|X=x}(y)\, dy = \int_{\mathbb{R}} y \frac{f_{X,Y}(x,y)}{\int f_{X,Y}(x,\eta) d\eta}\, dy.$$

Benutzen wir nun einen Kernschätzer zur Schätzung von $f_{X,Y}$ der Form

$$\hat{f}_{X,Y}(x,y) = \frac{1}{n} \sum_{i=1}^{n} K^x_{h_x}(x - X_i) K^y_{h_y}(y - Y_i)$$

mit Kernen K^x, K^y und Bandweiten $h_x, h_y > 0$, so ergibt Einsetzen (plug-in-Ansatz)

$$\begin{aligned}\hat{f}_{n,h_x,h_y}(x) &= \frac{\frac{1}{n}\sum_{i=1}^n K^x_{h_x}(x-X_i)\int yK^y_{h_y}(y-Y_i)dy}{\frac{1}{n}\sum_{i=1}^n K^x_{h_x}(x-X_i)\int K^y_{h_y}(\eta-Y_i)d\eta}\\ &= \frac{\sum_{i=1}^n K^x_{h_x}(x-X_i)\int yK^y_{h_y}(y-Y_i)dy}{\sum_{i=1}^n K^x_{h_x}(x-X_i)}.\end{aligned}$$

Da wir ein Integral bezüglich y schätzen wollen, zeigt es sich, dass Glätten bezüglich y nicht notwendig ist und der Grenzwert für $h_y \to 0$ wohldefiniert ist:

$$\lim_{h_y\to 0}\hat{f}_{n,h_x,h_y}(x) = \frac{\sum_{i=1}^n Y_i K^x_{h_x}(x-X_i)}{\sum_{i=1}^n K^x_{h_x}(x-X_i)}.$$

Dieser Schätzer ist auch intuitiv. Betrachte beispielsweise den Rechteckkern ($d = 1$). Dann wird $f(x)$ geschätzt durch ein lokales Mittel derjenigen Y_i, deren Kovariablen X_i Abstand maximal $h_x/2$ zu x besitzen. Bei allgemeineren Kernen handelt es sich um ein gewichtetes Mittel. Dies ergibt auch im Fall eines deterministischen Designs einen sinnvollen und sehr gebräuchlichen Kernschätzer.

3.2 Definition. Es seien K ein Kern im $\mathbb{R}^d$ und $h > 0$ eine Bandweite. Im Regressionsmodell mit deterministischem oder zufälligem Design (x_i) bezeichnet dann

$$\hat{f}^{NW}_{n,h}(x) := \frac{\sum_{i=1}^n Y_i K_h(x-x_i)}{\sum_{i=1}^n K_h(x-x_i)}$$

den *Nadaraya-Watson-Schätzer* der Regressionsfunktion f. Dieser ist wohldefiniert für alle $x \in D$ mit $\sum_{i=1}^n K_h(x-x_i) \neq 0$.

Im Modell des Signals im weißen Rauschen definiert man analog (wegen $\int K_h(x-t)dt = 1$ ist der Nenner gleich Eins):

$$\hat{f}^{NW}_{n,h}(x) := \int_0^1 K_h(x-t)\,dY_t.$$

Eine Verallgemeinerung dieser Konstruktion ergibt sich Im Fall nichtnegativer Kernfunktionen, wenn man den Nadaraya-Watson-Schätzer über ein Minimierungsprinzip charakterisiert ▶ Übung :

$$\hat{f}^{NW}_{n,h}(x) = \operatorname{argmin}_{y\in\mathbb{R}}\Big(\sum_{i=1}^n (Y_i-y)^2 K_h(x-x_i)\Big).$$

Wenn man die $K_h(x-x_i)$ als lokale Gewichte ansieht, ist der Nadaraya-Watson-Schätzer also die beste lokale Approximation mit der Konstanten y an die Daten. Oft wird es sich als nützlich erweisen, f lokal durch ein Polynom zu approximieren, weil dies einen besseren "fit" bei glatten Funktionen f erlaubt. Der Übersichtlichkeit halber betrachten wir im folgenden nur den skalaren Fall $d = 1$.

3.3 Definition. Es seien $K : \mathbb{R} \to \mathbb{R}$ ein Kern und $h > 0$ eine Bandweite. Für $m \in \mathbb{N}_0$ setze

$$P(z) := \Big(1, z, z^2, \ldots, z^m\Big)^\top$$

sowie im Regressionsmodell mit deterministischem oder zufälligem Design (x_i)

$$\hat{\vartheta}_{n,h}(x) := \operatorname{argmin}_{\vartheta \in \mathbb{R}^{m+1}} \sum_{i=1}^{n} \Big(Y_i - \langle \vartheta, P(x_i - x)\rangle\Big)^2 K_h(x - x_i).$$

Dann heißt

$$\hat{f}_{n,h}^{LPm}(x) := \langle \hat{\vartheta}_{n,h}(x), P(0)\rangle = (\hat{\vartheta}_{n,h}(x))_1$$

lokal-polynomialer Schätzer vom Grad m. Im Fall $m = 0$ ist dies der Nadaraya-Watson-Schätzer, im Fall $m = 1$ spricht man von einem *lokal-linearen Schätzer.*

3.4 Lemma. *Es gilt*

$$\hat{f}_{n,h}^{LPm}(x) = \sum_{i=1}^{n} Y_i w_i(x)$$

mit lokalen Gewichten

$$w_i(x) := n^{-1}\big(B(x)^{-1} P(x_i - x)\big)_1 K_h(x - x_i),$$

sofern die $(m+1) \times (m+1)$*-Matrix*

$$B(x) := \frac{1}{n} \sum_{i=1}^{n} P(x_i - x) P(x_i - x)^\top K_h(x - x_i)$$

invertierbar ist.

3.5 Bemerkung. Bei einem nicht-negativen Kern K ist $B(x)$ stets positiv-semidefinit wegen

$$\langle B(x)v, v\rangle = v^\top B(x)^\top v = \frac{1}{n} \sum_{i=1}^{n} (v^\top P(x_i - x))^2 K_h(x - x_i) \geqslant 0 \quad \forall v \in \mathbb{R}^{m+1}.$$

In diesem Fall ist $B(x)$ invertierbar genau dann, wenn $B(x)$ strikt positiv definit ist, was häufig gefordert wird.

Beweis. Wir können $\hat{\vartheta}_{n,h}$ als Minimierer der quadratischen Form

$$F(\vartheta) := -\frac{2}{n} \sum_{i=1}^{n} Y_i \langle \vartheta, P(x_i - x)\rangle K_h(x - x_i) + \langle B(x)\vartheta, \vartheta\rangle$$

schreiben. Als notwendige Bedingung erhalten wir die Normalengleichungen

$$0 = \nabla F(\hat{\vartheta}_{n,h}) = -\frac{2}{n}\sum_{i=1}^{n} Y_i P(x_i - x) K_h(x - x_i) + 2B(x)\hat{\vartheta}_{n,h}.$$

Falls $B(x)$ invertierbar ist, so folgt eindeutig

$$\hat{\vartheta}_{n,h} = n^{-1}\sum_{i=1}^{n} Y_i B(x)^{-1} P(x_i - x) K_h(x - x_i)$$

und damit die Behauptung. □

3.6 Beispiel. Betrachte den lokal-linearen Schätzer mit Rechteckkern $K(x) = \mathbf{1}([-1/2, 1/2])(x)$. Setze $I_h := \{i : |x_i - x| \leqslant h/2\}$. Dann gilt

$$B(x) = \frac{1}{nh}\sum_{i\in I_h} P(x_i - x)P(x_i - x)^\top = \frac{1}{nh}\begin{pmatrix} \sum_{i\in I_h} 1 & \sum_{i\in I_h}(x_i - x) \\ \sum_{i\in I_h}(x_i - x) & \sum_{i\in I_h}(x_i - x)^2 \end{pmatrix}.$$

Aus der Cauchy-Schwarz-Ungleichung folgt $(\sum_i A_i)^2 \leqslant \sum_i A_i^2 \sum_i 1$ mit strikter Ungleichung, wenn der Vektor (A_i) nicht kollinear mit dem 1-Vektor ist. Daher ist $\det(B(x))$ strikt positiv, sobald es mindestens zwei verschiedene Designpunkte x_i, x_j in einer $h/2$-Umgebung von x gibt. Wir setzen

$$d(x) := \sum_{j\in I_h}(x_j - x)^2 \sum_{j\in I_h} 1 - \Big(\sum_{j\in I_h}(x_j - x)\Big)^2$$

und erhalten

$$B(x)^{-1} = \frac{nh}{d(x)}\begin{pmatrix} \sum_{j\in I_h}(x_j - x)^2 & -\sum_{j\in I_h}(x_j - x) \\ -\sum_{j\in I_h}(x_j - x) & \sum_{j\in I_h} 1 \end{pmatrix}.$$

Als Gewichte $w_i(x)$ ergeben sich somit

$$w_i(x) = d(x)^{-1}\Big(\sum_{j\in I_h}(x_j - x)^2 - \sum_{j\in I_h}(x_j - x)(x_i - x)\Big)\mathbf{1}(i \in I_h).$$

Der Faktor $d(x)^{-1}$ führt zur Normalisierung $\sum_i w_i(x) = 1$. Es gilt ferner $\sum_i x_i w_i(x) = x$ wegen

$$\begin{aligned}\sum_{i=1}^{n}(x_i - x)w_i(x) &= \frac{\sum_j (x_j - x)^2 \sum_i (x_i - x) - \sum_j (x_j - x)\sum_i (x_i - x)^2}{d(x)} \\ &= 0.\end{aligned}$$

Dies bedeutet, dass bei exakt linearen Daten $Y_i = ax_i + b$, $i \in I_h$, der Schätzer $\hat{f}_{n,h}^{LP1}(x) = ax + b$ erfüllt, was auch der Intention der Definition eines lokal linearen Schätzers entspricht. Im Falle eines äquidistanten Designs $x_i = i/n$

auf $[0,1]$ und für $x = k/n$, $k = 1, \ldots, n-1$, $h < \min(x, 1-x)/2$ fällt der lineare Term aus Symmetriegründen weg: $\sum_{j \in I}(x_j - x) = 0$. Es ergibt sich $w_i(x) = (\#I_h)^{-1}\mathbf{1}(i \in I_h)$ und damit der entsprechende Nadaraya-Watson-Schätzer. Der lokal-lineare Schätzer hat Vorzüge gegenüber dem Nadaraya-Watson-Schätzer, wenn das Design unregelmäßig ist oder aber am Rand des Intervalls D geschätzt werden soll. Beachte dazu auch die Fehleranalyse im nächsten Abschnitt.

3.7 Definition. Wir führen die Menge aller reellen Polynome vom maximalen Grad m ein:

$$\mathrm{Pol}_m := \Big\{ \mathbb{R} \ni x \mapsto \sum_{k=0}^{m} a_k x^k \,\Big|\, a_k \in \mathbb{R} \Big\}.$$

3.8 Lemma. *Die Gewichte $w_i(x)$ eines lokal-polynomialen Schätzers vom Grad m erfüllen für alle Polynome $p \in \mathrm{Pol}_m$*

$$\sum_{i=1}^{n} p(x_i) w_i(x) = p(x),$$

sofern die Matrix $B(x)$ invertierbar ist. Insbesondere folgt

$$\sum_{i=1}^{n} w_i(x) = 1, \quad \sum_{i=1}^{n} (x - x_i)^k w_i(x) = 0,\ k = 1, \ldots, m.$$

Beweis. Wir schreiben $p(x_i) = \sum_{k=0}^{m} \frac{p^{(k)}(x)}{k!}(x_i - x)^k = \langle \rho, P(x_i - x) \rangle$ mit $\rho = (p^{(k)}(x)/k!)_{k=0,\ldots,m}$. Damit gilt

$$\begin{aligned} \sum_{i=1}^{n} p(x_i) w_i(x) &= \frac{1}{n} \sum_{i=1}^{n} \langle \rho, P(x_i - x) \rangle \langle B(x)^{-1} P(x_i - x), P(0) \rangle K_h(x - x_i) \\ &= \rho^\top B(x) B(x)^{-1} P(0) = \rho_1. \end{aligned}$$

Mit $\rho_1 = p(x)$ folgt die Behauptung. Die Folgerung ergibt sich durch Betrachten des Polynoms $p(y) = (x - y)^k$ für festes x. □

3.3 Fehleranalyse

3.9 Definition. Im Regressionsmodell heißt jeder Schätzer der Form

$$\hat{f}_n(x) = \sum_{i=1}^{n} Y_i w_i(x)$$

mit Gewichten $w_i \in \mathbb{R}$, die nur von x und vom Design (x_i) abhängen, *linearer Schätzer*. Gilt $d = 1$ und

$$\sum_{i=1}^{n} w_i(x) = 1, \quad \sum_{i=1}^{n} (x - x_i)^k w_i(x) = 0,\ k = 1, \ldots, m,$$

so besitzt $\hat{f}_n$ die Eigenschaft der *polynomialen Reproduktion der Ordnung m*.

Im Modell des Signals im weißen Rauschen ist der Nadaraya-Watson-Schätzer $\hat{f}_{n,h}^{NW}(x) := \int_0^1 K_h(x-t)\,dY_t$ ein linearer Schätzer und besitzt die polynomiale Reproduktionsordnung m (im Sinne von $\int (x-t)^k K_h(x-t)dt = 0$, $k = 1, \dots, m$) des Kerns K. Mittels Bias-Varianz-Zerlegung und Lemma 2.13 erhalten wir für $f \in \mathscr{H}_D(\alpha; \infty, L)$ und $m \geqslant \langle \alpha \rangle$ direkt

$$\begin{aligned}
&\mathbb{E}_f[(\hat{f}_{n,h}^{NW}(x) - f(x))^2] \\
&= \Big(\int_0^1 K_h(x-t) f(t) dt - f(x) \Big)^2 + \mathbb{E}_f \Big[\Big(\int_0^1 K_h(x-t) \tfrac{\sigma}{\sqrt{n}} dW_t \Big)^2 \Big] \\
&= (K_h * f - f)(x)^2 + \frac{\sigma^2}{n} \int_0^1 K_h(x-t)^2 dt \\
&\leqslant \Big(h^\alpha \tfrac{L}{\langle \alpha \rangle !} \int |w|^\alpha |K(w)|\, dw \Big)^2 + \sigma^2 n^{-1} h^{-1} \|K\|_{L^2}^2.
\end{aligned}$$

Wir haben also dasselbe Bias-Varianz-Dilemma wie bei der Dichteschätzung, sogar die Größenordnung der Terme ist identisch. Insbesondere wird bei geeigneter Wahl der Bandweite h die Konvergenzrate ebenfalls $\mathcal{O}(n^{-\alpha/(2\alpha+1)})$ betragen. Im allgemeinen Regressionsmodell müssen wir zur Fehlerabschätzung mehr arbeiten, aber unter Bedingungen an die Gewichte werden wir ähnliche Resultate erzielen.

3.10 Lemma. *Betrachte für $d = 1$ einen linearen Schätzer $\hat{f}_n$ mit polynomialer Reproduktion der Ordnung m im Regressionsmodell mit deterministischem Design (x_i). Es gelte $\sigma^2 := \mathbb{E}[\varepsilon_i^2] < \infty$. Dann erhalten wir für Bias und Varianz:*

$$|\mathbb{E}_f[\hat{f}_n(x) - f(x)]| \leqslant \inf_{p_m \in \mathrm{Pol}_m,\, p_m(0)=0} \Big| \sum_{i=1}^n w_i(x)(f(x_i) - f(x) - p_m(x_i - x)) \Big|,$$

$$\mathrm{Var}_f(\hat{f}_n) = \sigma^2 \sum_{i=1}^n w_i(x)^2.$$

3.11 Bemerkung. Bei zufälligem Design (X_i) folgen die Aussagen bedingt auf das Ereignis $\{X_i = x_i,\, i = 1, \dots, n\}$, also zum Beispiel

$$|\mathbb{E}_f\left[\hat{f}_n(x) - f(x) \,\middle|\, (X_i) = (x_i)\right]| \leqslant \inf_{p_m} \Big| \sum_{i=1}^n w_i(x)(f(x_i) - f(x) - p_m(x_i - x)) \Big|,$$

wobei $w_i(x)$ das Gewicht bei der Realisierung $(X_i) = (x_i)$ bezeichnet. Im folgenden werden wir nur deterministisches Design betrachten. Resultate für zufälliges Design ergeben sich durch Mittelung über die (x_i) gemäß der Verteilung von (X_i).

Beweis. Es sei $p_m \in \mathrm{Pol}_m$ mit $p_m(0) = 0$. Dann folgt aus der polynomialen Reproduktion $\sum_i p_m(x_i - x) w_i(x) = 0$ und somit

$$\mathbb{E}_f[\hat{f}_n(x)] - f(x) = \sum_{i=1}^n (f(x_i) - f(x)) w_i(x) - \sum_{i=1}^n p_m(x_i - x) w_i(x).$$

Dies impliziert die Abschätzung für den Betrag. Aus der i.i.d.-Verteilung der (ε_i) folgt

$$\mathrm{Var}_f(\hat{f}_n(x)) = \sum_{i=1}^n \mathrm{Var}_f(Y_i) w_i(x)^2 = \sum_{i=1}^n \mathrm{Var}(\varepsilon_i) w_i(x)^2 = \sigma^2 \sum_{i=1}^n w_i(x)^2.$$

□

Wir betrachten wiederum Hölderklassen, um die Regularität der Regressionsfunktion f zu beschreiben. Im Gegensatz zum Dichteschätzproblem müssen wir weder fordern, dass f eine Dichte ist, noch, dass f gleichmäßig beschränkt ist. Wir betrachten nur den eindimensionalen Fall.

3.12 Definition. Als *Hölderklasse* mit Parametern $\alpha, L > 0$ auf $I \subseteq \mathbb{R}$ offen bezeichnen wir die Menge

$$\mathscr{H}_I(\alpha; L) := \Big\{ f \in C^\alpha(I) \,\Big|\, \sup_{x,y \in I,\, x \neq y} \frac{|f^{(\langle\alpha\rangle)}(x) - f^{(\langle\alpha\rangle)}(y)|}{|x-y|^{\alpha - \langle\alpha\rangle}} \leqslant L \Big\}.$$

3.13 Satz. *Es sei $\hat{f}_{n,h}(x)$ ein linearer Schätzer im Regressionsmodell mit $d = 1$ und mit deterministischem Design (x_i). Die Gewichte $w_{i,n,h}(x)$ mögen für $h > 0$ und mit Konstanten $C_1, C_2 > 0$ folgende Eigenschaften erfüllen:*

(a) $\sum_{i=1}^n |w_{i,n,h}(x)| \leqslant C_1$;

(b) $\sum_{i=1}^n w_{i,n,h}(x)^2 \leqslant C_2^2 n^{-1} h^{-1}$;

(c) $w_{i,n,h}(x) = 0$ *für alle i mit* $|x_i - x| > h$.

Weiterhin sei $\sigma^2 := \mathbb{E}[\varepsilon_i^2]$ endlich sowie $x \in \mathbb{R}$, $U(x)$ eine offene Umgebung von x und $\alpha, L > 0$. Reproduziert $\hat{f}_{n,h}(x)$ Polynome der Ordnung $m := \langle\alpha\rangle$, so gilt für hinreichend kleines $h > 0$ die obere Schranke

$$\sup_{f \in \mathscr{H}_{U(x)}(\alpha, L)} \mathbb{E}_f[(\hat{f}_{n,h}(x) - f(x))^2] \leqslant (C_1/m!)^2 L^2 h^{2\alpha} + \frac{C_2^2 \sigma^2}{nh}.$$

Wählt man $h_n = n^{-1/(2\alpha+1)}$, so ergibt sich die Rate

$$\sup_{f \in \mathscr{H}_{U(x)}(\alpha, L)} \mathbb{E}_f[(\hat{f}_{n,h_n}(x) - f(x))^2] = \mathcal{O}(n^{-2\alpha/(2\alpha+1)}).$$

3.14 Bemerkung. Ein entsprechendes Resultat gilt auch für $d \geqslant 2$ mit Rate $n^{-2\alpha/(2\alpha+d)}$, sofern in Eigenschaft (b) statt h^{-1} der Term h^{-d} gefordert wird (was natürlich ist).

Beweis. Nach dem vorangegangenen Lemma und Eigenschaft (b) gilt für die Varianz

$$\mathrm{Var}_f(\hat{f}_{n,h}(x)) = \sigma^2 \sum_{i=1}^{n} w_{i,n,h}(x)^2 \leqslant \frac{C_2^2 \sigma^2}{nh}.$$

Eine Taylorentwicklung zeigt mittels vorangegangenem Lemma und $m = \langle \alpha \rangle$ für den Biasterm (mit $\tau_i \in [x_i, x]$ werde eine Zwischenstelle bezeichnet, vergleich auch Lemma 2.13):

$$\begin{aligned} |\mathbb{E}_f[\hat{f}_n(x) - f(x)]| &\leqslant \inf_{p_m} \Big| \sum_{i=1}^{n} w_i(x)(f(x_i) - f(x) - p_m(x_i - x)) \Big| \\ &\leqslant \Big| \sum_{i=1}^{n} w_i(x)(f^{(m)}(\tau_i) - f^{(m)}(x))(x_i - x)^m / m! \Big| \\ &\leqslant C_1 L h^\alpha / m!. \end{aligned}$$

Im letzten Schritt haben wir die Dreiecksungleichung, Eigenschaft (a) sowie die Abschätzungen $|\tau_i - x| \leqslant h$ und $|x_i - x| \leqslant h$ für diejenigen i mit $w_i(x) \neq 0$ aus Eigenschaft (c) verwendet. Die Bias-Varianz-Zerlegung ergibt die behauptete Ungleichung und Einsetzen die angegebene Kovergenzrate. □

3.15 Beispiel. Betrachte den Nadaraya-Watson-Schätzer $\hat{f}_{n,h}^{NW}$ mit nichtnegativem und stetigem Kern K. Dann gilt

$$w_{i,n,h}(x) = \frac{K_h(x - x_i)}{\sum_{j=1}^{n} K_h(x - x_j)} \geqslant 0,$$

sofern $\hat{f}_{n,h}^{NW}(x)$ wohldefiniert ist. Wir erhalten

$$\sum_{i=1}^{n} |w_{i,n,h}(x)| = \sum_{i=1}^{n} w_{i,n,h}(x) = 1$$

und somit die polynomiale Reproduktion vom Grad $m = 0$ und Eigenschaft (a) im vorigen Satz. Liegt der Träger von K in $[-1, 1]$, so folgt Eigenschaft (c). Eigenschaft (b) wird impliziert, wenn

$$w_{i,n,h}(x) = \frac{K_h(x - x_i)}{\sum_{j=1}^{n} K_h(x - x_j)} \leqslant C_2^2/(nh)$$

gilt (benutze $\sum_i w_i^2 \leqslant \max_i w_i \sum_i w_i$). Für den Zähler gilt $K_h(x - x_i) \leqslant h^{-1}\|K\|_\infty$. Damit der Nenner die richtige Größenordnung besitzt, fordern wir ein *reguläres Design*, nämlich

$$\exists C > 0 \; \forall x \in I, \, n \geqslant 1 : \; \mathrm{dist}(y, \{x_i \,|\, i = 1, \ldots n\}) \leqslant C/n. \tag{3.3}$$

Beachte dabei, dass das Design von n abhängt, wir der Kürze halber aber x_i statt $x_i^{(n)}$ schreiben. Aus $K \geqslant 0$ und $\int K = 1$ folgt mit einem Stetigkeitsargument, dass es ein offenes Intervall (A, B) und $\varepsilon > 0$ gibt mit $K(x) > \varepsilon$ für $x \in (A, B)$. Wir schließen

$$\sum_{j=1}^{n} K_h(x - x_j) \geqslant \sum_{j:\, hA \leqslant x - x_j \leqslant hB} h^{-1}\varepsilon \geqslant h^{-1}\varepsilon \frac{h(B-A)}{2C/n},$$

so dass wir $C_2^2 = 2C\|K\|_\infty/(\varepsilon(B-A))$ wählen können. Zusammenfassend erhalten wir die Aussage, dass die Risikoabschätzung im vorangegangenen Satz für den Nadaraya-Watson-Schätzer gilt, sofern der Kern K stetig ist und Träger in $[-1, 1]$ besitzt und das Design regulär ist im Sinne von (3.3). Da nur Konstanten reproduziert werden, gilt dies jedoch nur für Hölderregularität $\alpha \leqslant 1$. Wie wir gesehen haben, fällt der Nadaraya-Watson-Schätzer manchmal mit dem lokal linearen Schätzer zusammen (beim Rechteckkern und bei äquidistantem Design). In diesem Fall gilt das Resultat sogar für alle $\alpha \leqslant 2$. Beachte auch, dass nicht-negative Gewichte maximal lineare Polynome reproduzieren können, genauso wie nicht-negative Kerne maximal die Ordnung 1 besitzen.

Man kann die Bedingungen im vorigen Satz asymptotisch auch für allgemeine lokal-polynomiale Schätzer und reguläres Design nachweisen. Darüberhinaus erhält man analog zum Dichteschätzproblem Abschätzungen des MISE für Sobolevklassen. All dies wird im Buch von Tsybakov (2004) umfassend dargestellt. Viele praktische Aspekte beim Einsatz lokaler Polynome werden von Fan and Gijbels (1996) und Hastie, Tibshirani, and Friedman (2001) diskutiert.

3.4 Projektionsschätzer

Eine weitere wichtige Methode zum Schätzen der Regressionsfunktion f (und der Dichtefunktion, s.u.) ist es, diese durch eine Linearkombination $f = \sum_{k=1}^{K} c_k \varphi_k$ von Funktionen $(\varphi_k)_{k=1,\dots,K}$ zu approximieren und die Koeffizienten $(c_k)_{k=1,\dots,K} \subseteq \mathbb{R}^K$ (parametrisch) zu schätzen. Mit wachsender Anzahl n von Beobachtungen wird man auch K größer wählen, so dass der Approximationsfehler asymptotisch klein wird. Die Auswahl der Funktionenfamilie (φ_k) liegt in der Hand des Statistikers und sollte möglichst so gewählt werden, dass die unbekannte Funktion f gut approximiert wird. Aktuell werden dazu häufig sehr umfangreiche Familien betrachtet, jedoch mit der Forderung, dass nur wenige Koeffizienten (c_k) ungleich Null sind (*dictionaries* (φ_k) und *sparse representations* (c_k)). Besonders einfach in Durchführung und Analyse ist jedoch die Schätzung, wenn $(\varphi_k)_{k=1,\dots,n}$ ein Orthonormalsystem bezüglich dem empirischen Skalarprodukt

$$\langle \varphi, \psi \rangle_n := \frac{1}{n} \sum_{i=1}^{n} \varphi(x_i)\psi(x_i)$$

bilden, weil dann ein Kleinster-Quadrate-Schätzer bei orthogonalem Design verwendet werden kann (beachte, dass jede Basis mittels Gram-Schmidt-Verfahren orthonormalisiert werden kann).

3.16 Definition. Im Regressionsmodell sei $(\varphi_k)_{k=1,\dots,K}$, $1 \leqslant K \leqslant n$, ein Orthonormalsystem von Funktionen in $L^2(D, \|\bullet\|_n)$. Dann heißt

$$\hat{f}_{n,K}(x) := \sum_{k=1}^{K} \hat{c}_k \varphi_k(x) \text{ mit } \hat{c}_k := \frac{1}{n} \sum_{i=1}^{n} Y_i \varphi_k(x_i)$$

Orthogonalreihen- oder Projektionsschätzer von f.

3.17 Bemerkung. Basisunabhängig kann $\hat{f}_{n,K}$ als Orthogonalprojektion P_K von $(Y_i)_{i=1,\dots,n}$ auf den von $(\varphi_k(x_i))_{i=1,\dots,n}$, $k = 1, \dots, K$, aufgespannten K-dimensionalen Unterraum aufgefasst werden. Die Werte $\hat{f}_{n,K}(x)$ für $x \neq x_i$ ergeben sich durch die Werte von φ_k dort.

Im Modell des Signals in weißem Rauschen ist das Design gewissermaßen kontinuierlich und gleichmäßig auf $[0,1]$, so dass der Projektionsschätzer für ein Orthogonalsystem $(\varphi_k)_{k=1,\dots,K}$, $K \in \mathbb{N}$, von $L^2([0,1])$ gegeben ist durch

$$\hat{f}_{n,K}(x) := \sum_{k=1}^{K} \hat{c}_k \varphi_k(x), \quad \hat{c}_k := \int_0^1 \varphi_k(t)\, dY_t.$$

Im *Folgenraummodell* der empirischen Koeffizienten verfügen wir daher unter der Annahme $f \in L^2([0,1])$ äquivalent über die Beobachtungen

$$\hat{c}_k := c_k + \varepsilon_k, \quad k = 1, \dots, K, \text{ mit } c_k := \langle f, \varphi_k \rangle,\ \varepsilon_k \sim N(0, \sigma^2/n) \text{ i.i.d.}$$

Die Abschätzung des MISE ist hier besonders einfach (beachte $P_K f = \sum_{k=1}^{K} f_k \varphi_k$ und benutze Orthogonalität):

$$\mathbb{E}_f \left[\|\hat{f}_{n,K} - f\|_{L^2}^2 \right] = \mathbb{E}_f \left[\sum_{k=1}^{K} (\hat{c}_k - c_k)^2 + \|P_K f - f\|_{L^2}^2 \right] = \|f - P_K f\|_{L^2}^2 + \frac{\sigma^2}{n} K.$$

Diesmal ergibt sich das Bias-Varianz-Dilemma in der Wahl von K. Je größer die Dimension K des Ansatzraumes ist, desto kleiner ist der Approximationsfehler $\|f - P_K f\|_{L^2}^2$, desto größer jedoch die Varianz $\frac{\sigma^2}{n} K$.

3.18 Lemma. *Es seien $\hat{f}_{n,K}$ der Projektionsschätzer im Regressionsmodell mit deterministischem Design, $f \in L^2(D)$ und $\sigma^2 := \mathbb{E}[\varepsilon_i^2] < \infty$. Dann gilt folgende Bias-Varianz-Zerlegung für $\hat{f}_{n,K}$ in der empirischen Norm:*

$$\mathbb{E}_f[\|\hat{f}_{n,K} - f\|_n^2] = \|f - P_K f\|_n^2 + \sigma^2 K n^{-1},$$

wobei P_K die Orthogonalprojektion auf $\operatorname{span}(\varphi_1, \dots, \varphi_K)$ *bezüglich dem empirischen Skalarprodukt bezeichnet.*

Im Regressionsmodell mit zufälligem Design, $f \in L^2(D)$ und $\sigma^2(x) := \mathrm{Var}_f(Y_i \,|\, X_i = x) < \infty$ $\mathbb{P}^{X_i}$-f.s. ergibt sich entsprechend

$$\mathbb{E}_f[\|\hat{f}_{n,K} - f\|_n^2] = \mathbb{E}_f[\|f - P_K f\|_n^2] + \frac{1}{n}\sum_{k=1}^{K} \mathbb{E}_f[\sigma^2(X_i)\varphi_k^2(X_i)].$$

Beweis. Bei deterministischem Design erweitern wir $(\varphi_k)_{1\leqslant k\leqslant K}$ im Fall $K < n$ zu einer Orthonormalbasis $(\varphi_k)_{k=1,\dots,n}$ von $L^2(D, \|\bullet\|_n)$ und erhalten

$$\begin{aligned}
&\mathbb{E}_f[\|\hat{f}_{n,K} - f\|_n^2] \\
&= \mathbb{E}_f\Big[\sum_{k=1}^{n} \langle \hat{f}_{n,K} - f, \varphi_k\rangle_n^2\Big] \\
&= \sum_{k=1}^{K} \mathbb{E}_f[\langle \hat{f}_{n,K} - f, \varphi_k\rangle_n^2] + \sum_{k=K+1}^{n} \mathbb{E}_f[\langle f, \varphi_k\rangle_n^2] \\
&= \sum_{k=1}^{K} \mathrm{Var}_f\Big(\frac{1}{n}\sum_{i=1}^{n} Y_i\varphi_k(x_i)\Big) + \sum_{k=K+1}^{n} \langle f, \varphi_k\rangle_n^2 \\
&= \|f - P_K f\|_n^2 + \sum_{k=1}^{K} \frac{\sigma^2}{n}\|\varphi_k\|_n^2 \\
&= \|f - P_K f\|_n^2 + K\frac{\sigma^2}{n}.
\end{aligned}$$

Bei zufälligem Design erhalten wir zunächst durch Bedingung auf das Design

$$\mathbb{E}_f[\|\hat{f}_{n,K} - f\|_n^2 \,|\, (X_i)] = \|f - P_K f\|_n^2 + \frac{K}{n^2}\sum_{i=1}^{n} \mathrm{Var}_f(Y_i\varphi_k(X_i) \,|\, X_i).$$

Integrieren über die Verteilung von X_i liefert

$$\mathbb{E}_f[\|\hat{f}_{n,K} - f\|_n^2] = \mathbb{E}_f[\|f - P_K f\|_n^2] + \frac{1}{n}\sum_{k=1}^{K} \mathbb{E}_f[\sigma^2(X)\varphi_k^2(X)].$$

□

3.19 Bemerkungen.

(a) Für $K = n$ interpolieren wir nur die Daten: der Biasterm ist gleich Null, aber der Varianzterm gleich σ^2. Für kleine Werte von K kann man dieses Resultat auch als Quantifizierung einer möglichen Modellmisspezifikation lesen. Setzen wir beispielsweise eine lineare Regressionsgerade an, so schätzen wir im Prinzip die beste lineare Approximation an unsere allgemeine Regressionsfunktion, und der entsprechende Approximationsfehler erscheint im Risiko.

(b) Beachte, dass bei all diesen Rechnungen die empirische Norm die natürliche Verlustfunktion bezeichnet, da die Schätzung durch Orthogonalprojektion in dieser Geometrie erfolgt. Im Fall von zufälligem Design ist dann jedoch sowohl der empirische Verlust als auch die Projektion zufällig. Es wird eine starke gleichmäßige Konzentrationseigenschaft der durch die empirische Verteilung von n bestimmten Norm $\|\bullet\|_n$ um die durch die wahre Verteilung der (X_i) bestimmte Norm $\|\bullet\|_X$ benötigt, um den sogenannten *Vorhersagefehler (prediction error)* $\|\hat{f}_{n,K} - f\|_X^2$ zu kontrollieren, vergleiche Kapitel 11 in Györfi, Kohler, Krzyżak, and Walk (2002).

Ein wesentlicher Vorteil der Projektionsschätzung ist es, dass für das entsprechende Anwendungsproblem maßgeschneiderte Basisfunktionen verwendet werden können. Um die Ergebnisse mit den Kernmethoden zu vergleichen, betrachten wir hier nur den wichtigen Fall regulärer Funktionen f, die auf klassische Approximationstheorie führen.

3.20 Definition. Es sei $(V_K)_{K\geqslant 1} \subseteq L^2(D)$, $D \subseteq \mathbb{R}^d$, eine Folge endlichdimensionaler Unterräume mit wachsender Dimension $d_K := \dim(V_K) \uparrow \infty$. Diese heißen *Approximationsräume der Ordnung* m, falls für alle $K \geqslant 1$, $f \in \mathscr{H}_D(s, L)$, $s \leqslant m+1$, mit einer Konstanten $C_s > 0$ gilt

$$\|f - P_K f\|_{L^2} = \inf_{f_K \in V_K} \|f - f_K\|_{L^2} \leqslant C_s L d_K^{-s/d}.$$

3.21 Bemerkungen.

(a) Beachte, dass das Infimum gerade von der Orthogonalprojektion $P_{V_K} f$ von f auf V_K angenommen wird.

(b) Ungleichungen dieser Art heißen auch *direkte Ungleichungen* oder *Jackson-Ungleichungen*. Sie geben die Approximationsgüte von Unterräumen wie $C^s(D)$ durch endliche Unterräume an. Dies kann in den allgemeinen Zusammenhang mit Entropiezahlen kompakter Mengen eingeordnet werden.

(c) Da wir die L^2-Norm als Verlust betrachten, ist im obigen Beispiel der L^2-Sobolevraum $H^s(D)$ und nicht $C^s(D)$ die natürliche (und beste) Wahl. Wir verzichten hier jedoch auf die Theorie der Sobolevräume auf Teilmengen des $\mathbb{R}^d$.

3.22 Beispiele.

(a) Es sei $D = [0,1]$ und $V_K := \{\varphi : [0,1] \to \mathbb{R} \,|\, \varphi = \sum_{m=1}^K c_m \mathbf{1}_{[(m-1)/K, m/K)}\}$ der Raum der stückweise konstanten Funktionen. Dann gilt für $f \in \mathscr{H}_{[0,1]}(s, L)$, $s \leqslant 1$:

$$\sum_{m=1}^K \int_{(m-1)/K}^{m/K} (f(x) - f(m/K))^2 dx \leqslant K \int_0^{1/K} L^2 x^{2(s\wedge 1)} dx \leqslant L^2 K^{-2s}.$$

Damit besitzen $(V_K)_{K\geqslant 1}$ die Approximationsordnung $m = 0$. Eine größere Approximationsordnung ist nicht zu erwarten, selbst eine lineare Funktion wird nur mit einem Fehler der Ordnung K^{-1} stückweise konstant approximiert. Durch $\varphi := \mathbf{1}_{[0,1]}$, $\psi_{j,k} = 2^{j/2}(\mathbf{1}_{[k2^{-j},(k+1/2)2^{-j})} - \mathbf{1}_{((k+1/2)2^{-j},(k+1)2^{-j})})$, $j \geqslant 1$, $k = 0,\ldots,2^j-1$, wird eine Orthonormalbasis der geschachtelten V_K mit $K = 2^j$ definiert, die bekannte *Haarbasis*. Man nennt die $\psi_{j,k}$ auch *Haar-Wavelets*.

(b) Es sei $D = [0,1]^d$ und $V_K := \{\varphi : [0,1] \to \mathbb{R} \,|\, \varphi = \sum_{m\in\{1,\ldots,K\}^d} c_m \mathbf{1}_{[(m-1)/K,m/K)}\}$, wobei $[(m-1)/K, m/K) \subseteq \mathbb{R}^d$ die entsprechenden Würfel bezeichnet (interpretiere $1 = (1,\ldots,1)$). Dann gilt für $f \in \mathscr{H}_{[0,1]^d}(s,L)$, $s \leqslant 1$:

$$\sum_{m\in\{1,\ldots,K\}^d} \int_{[(m-1)/K,m/K)} (f(x) - f(m/K))^2 dx \leqslant L^2 K^{-2s}.$$

Wegen $d_K = K^d$ besitzen $(V_K)_{K\geqslant 1}$ in Dimension $d \geqslant 2$ ebenfalls die Approximationsordnung $m = 0$.

(c) Betrachte in $L^2([0,1])$ die *Splineräume* der Ordnung $M \in \mathbb{N}$

$$V_K := \{\varphi \in C^{M-1}([0,1]) \,|\, \forall m = 1,\ldots,K : \varphi|_{[(m-1)/K,m/K]} \text{ ist Polynom vom Grad } \leqslant M\}$$

und die Räume der stückweisen Polynome vom Grad $M \in \mathbb{N}$

$$W_K := \{\varphi \in L^2([0,1]) \,|\, \forall m = 1,\ldots,K : \varphi|_{[(m-1)/K,m/K]} \text{ ist Polynom vom Grad } \leqslant M\}$$

Es gilt natürlich $V_K \subseteq W_K$. Da jedes $\varphi \in W_K$ durch die Werte $\varphi(m/K), \varphi'(m/K-), \ldots, \varphi^{(M)}(m/K-)$ an den *Knoten* m/K eindeutig festgelegt ist, gilt in beiden Fällen $d_K = \mathcal{O}(K)$. Eine Taylorentwicklung der allgemeinen Funktion $f \in C^s([0,1])$ mit $s \leqslant m+1$ um die Knotenpunkte zeigt (vergleiche z.B. Lemma 3.10), dass die Räume W_K der stückweisen Polynome vom Grad M auch die Approximationsordnung M besitzen. Dasselbe gilt auch für die Splineräume V_K, der Nachweis ist jedoch etwas aufwändiger, siehe z.B. De Boor (2001). Spline-Ansatzfunktionen im Mehrdimensionalen werden im Rahmen der *finiten Elemente* behandelt und besitzen analoge Approximationseigenschaften.

(d) Weitere Beispiele von Approximationsräumen werden durch Basisfunktionen definiert, wie z.B. Polynome, stückweise Polynome oder trigonometrische Polynome (vergleiche dazu Definition 3.31).

Im Modell des Signals im weißen Rauschen erhalten wir bei Funktionen $f \in C^s(D)$ und Projektionsschätzern auf entsprechende Approximationsräume unmittelbar die Abschätzung, dass der MISE von der Ordnung $\mathcal{O}(d_K^{-2s/d} + \sigma^2 n^{-1} d_K)$ ist, also bei optimaler Wahl der Dimension d_K von der Ordnung $\mathcal{O}(n^{-2s/(2s+d)})$.

3.23 Satz. *Es sei $\hat{f}_{n,K}$ der Projektionsschätzer im Regressionsmodell mit äquidistantem Design in $D = [0,1]^d$ auf Approximationsräume (V_K) mit Approximationsordnung m. Mit $d_K := \dim(V_K)$ gelte für alle $K \geqslant 1$, $f \in \mathscr{H}_D(s,L)$, $1 \leqslant s \leqslant m+1$, mit einer Konstanten $C_s' \geqslant 1$*

$$\|(f - P_K f)'\|_\infty \leqslant C_s' L d_K^{-(s-1)/d}.$$

Im Fall $\sigma^2 := \mathbb{E}[\varepsilon_i^2] < \infty$ gilt dann folgende Abschätzung des Risikos in empirischer Norm:

$$\sup_{f \in \mathscr{H}_D(s,L)} \mathbb{E}_f[\|\hat{f}_{n,K} - f\|_n^2] \leqslant C_s^2 L^2 d_K^{-2s/d}(1 + \mathcal{O}(C_s'^2 (d_K/n)^{1/d})) + \sigma^2 d_K n^{-1}.$$

Unter Vernachlässigung des Terms $\mathcal{O}(C_s'^2 (d_K/n)^{1/d})$ ergibt Optimierung in der Dimension (so der Wert angenommen wird)

$$d_K^* := \Big(\frac{2sC_s^2 L^2 n}{\sigma^2 d}\Big)^{d/(2s+d)},$$

und wir erhalten

$$\sup_{f \in \mathscr{H}_D(s,L)} \mathbb{E}_f[\|\hat{f}_{n,K} - f\|_n^2] \leqslant \frac{2s+d}{2s}\Big(\frac{\sigma^2}{n}\Big)^{2s/(2s+d)} \times \\ \times \Big(\frac{2sC_s^2 L^2}{d}\Big)^{d/(2s+d)}(1 + \mathcal{O}(n^{-2s/(2ds+d^2)})).$$

Beweis. Nach dem vorangegangenen Lemma reicht es, den Bias $\|f - P_K f\|_n$ abzuschätzen:

$$\begin{aligned}
&\Big|\|f - P_K f\|_n^2 - \|f - P_K f\|_{L^2}^2\Big| \\
&= \Big|\sum_{m \in \{1,\dots,n^{1/d}\}^d} \int_{[(m-1)/n^{1/d}, m/n^{1/d})} \big((f - P_K f)(m/n^{1/d})^2 - (f - P_K f)(x)^2\big)\, dx\Big| \\
&\leqslant \sqrt{d} n^{-1/d} \|((f - P_K f)^2)'\|_\infty \\
&\leqslant 2\sqrt{d} n^{-1/d} \|f - P_K f\|_\infty \|(f - P_K f)'\|_\infty.
\end{aligned}$$

Ist die Funktion $|f - P_K f|$ minimal bei $x_m \in D$, so gilt

$$\begin{aligned}
\sup_{x \in D} |(f - P_K f)(x)| &= \Big|(f - P_K f)(x_m) + \int_0^1 (f - P_K f)'(x_m + t(x - x_m))\, dt\Big| \\
&\leqslant \|f - P_K f\|_{L^2} + \|(f - P_K f)'\|_\infty.
\end{aligned}$$

Die Approximationseigenschaft der (V_K) und die zusätzliche Annahme an (V_K) liefern daher

$$\begin{aligned}\|f - P_K f\|_n^2 &\leqslant C_s^2 L^2 d_K^{-2s/d} + 2\sqrt{d} L^2 (C_s C_s' + C_s'^2) n^{-1/d} d_K^{-(2s-1)/d} \\ &= C_s^2 L^2 d_K^{-2s/d} (1 + \mathcal{O}(C_s'^2 (d_K/n)^{1/d})).\end{aligned}$$

Dies ergibt die angegebene Risikoabschätzung, das Risiko für d_K^* folgt durch Einsetzen. □

3.24 Beispiel. Die Annahme an die Ableitung von $f - P_K$ ist für die gewöhnlichen Approximationsräume erfüllt, so sie aus differenzierbaren Funktionen bestehen. Betrachte beispielsweise die Räume W_K der stückweisen Polynome vom Grad M. Dort erfüllt das Taylorpolynom f_K M-ten Grades von $f \in C^s$, $s \geqslant M$, bei einem $x_0 \in D$ gerade, dass f_K' das Taylorpolynom $(M-1)$-ten Grades an $f' \in C^{s-1}$ ist. Die Standardabschätzungen implizieren dann direkt die Annahme.

3.5 Weitere Schätzmethoden

Eine weitere wichtige Schätzidee im Regressionsmodell beruht auf einer penalisierten Kleinste-Quadrate-Schätzung:

$$\hat{f}_n := \operatorname{argmin}_{g \in \mathcal{G}} \Big(\sum_{i=1}^n |Y_i - g(x_i)|^2 + \operatorname{Pen}(g) \Big)$$

mit einem geeigneten Strafterm (*penalty*) $\operatorname{Pen}(g)$ und Optimierung über eine geeignete Funktionenklasse $\mathcal{G}$. Die Idee ist, einen Schätzer zu konstruieren, der die Daten gut approximiert und gleichzeitig adäquate Eigenschaften besitzt. Beispielsweise wird eine kleine Norm in einer Glattheitsklasse von Funktionen erreicht durch einen *roughness penalty* $\operatorname{Pen}(g) = \lambda \int_D g^{(m)}(x)^2 dx$ mit $\lambda > 0, m \in \mathbb{N}$. Im skalaren Fall und mit $\mathcal{G} = C^m(D)$ ergibt sich als Schätzer dann eine Spline-Funktion mit Knoten (x_i) vom Grad $2m-1$, ein sogenannter *smoothing spline*. Diese Schätzmethode ist in der Praxis oft sehr erfolgreich, zumindest bei guter Wahl von λ. Die Fehleranalyse bedarf anspruchsvoller Entropiemethoden. Wir verweisen auf Hastie, Tibshirani, and Friedman (2001) und Györfi, Kohler, Krzyżak, and Walk (2002) für die Details.

3.6 Übertragung auf Dichteschätzung

Natürlich lassen sich analog auch Projektionsschätzer im Dichteschätzproblem konstruieren. Im Detail ergeben sich andere Eigenschaften, aber im Großen und Ganzen ist die Theorie die gleiche.

3.25 Definition. Betrachte das Schätzproblem auf einer offenen Teilmenge $D \subseteq \mathbb{R}^d$ und setze $f|_D \in L^2(D)$ voraus. Es sei $S_K \subseteq L^2(D) \cap C(D)$ ein K-dimensionaler linearer Unterraum und $\varphi_1, \dots \varphi_K$ bezeichne eine $L^2(D)$-Orthonormalbasis von S_K. Dann ist der *Projektionsschätzer* auf S_K, basierend auf den Beobachtungen $X_1, \dots, X_n$, definiert als

$$\hat{f}_{n,K}(x) := \sum_{k=1}^{K} \langle \hat{\mu}_n, \varphi_k \rangle \varphi_k(x) := \sum_{k=1}^{K} \Big(\frac{1}{n} \sum_{i=1}^{n} \varphi_k(X_i) \Big) \varphi_k(x), \quad x \in D.$$

3.26 Lemma. *Es gilt $\hat{f}_{n,K} \in S_K$ sowie $\int \varphi(x) \hat{f}_{n,K}(x) dx = \int \varphi(x) \hat{\mu}_n(dx)$ für alle $\varphi \in S_K$. Diese Eigenschaften charakterisieren $\hat{f}_{n,K}$ unabhängig von der Wahl der Orthonormalbasis (φ_m).*

3.27 Bemerkung. Wäre $\hat{\mu}_n$ eine L^2-Funktion, so wäre der Projektionsschätzer einfach die Orthogonalprojektion auf S_K und damit offensichtlich unabhängig von der Basiswahl.

Beweis. Als Linearkombination der φ_k liegt $\hat{f}_{n,K}$ offensichtlich in S_K. Einsetzen und die Darstellung $\varphi = \sum_k \langle \varphi, \varphi_k \rangle \varphi_k$ ergeben

$$\begin{aligned} \int_D \varphi(x) \hat{f}_{n,K}(x) dx &= \sum_{k=1}^{K} \Big(\frac{1}{n} \sum_{i=1}^{n} \varphi_k(X_i) \Big) \langle \varphi, \varphi_k \rangle_{L^2(D)} \\ &= \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i) \\ &= \int_D \varphi(x) \hat{\mu}_n(dx). \end{aligned}$$

Jede Funktion g in S_K mit $\int \varphi(x) g(x) dx = \int \varphi(x) \hat{\mu}_n(dx)$ für alle $\varphi \in S_K$ erfüllt auch $\int \varphi(x)(\hat{f}_{n,K}(x) - g(x)) dx = 0$, so dass $(\hat{f}_{n,K}(x) - g(x)) \perp_{L^2} S_K$ und somit $\hat{f}_{n,K}(x) - g(x) = 0$ gilt. Also ist $\hat{f}_{n,K}$ eindeutig durch diese Eigenschaften festgelegt. □

3.28 Satz. *Es sei $f|_D \in L^2(D)$. Dann gilt folgende Bias-Varianz-Zerlegung für $\hat{f}_{n,K}$ in $L^2(D)$:*

$$R_D(\hat{f}_{n,K}, f) = \|f - P_{S_K} f\|^2_{L^2(D)} + n^{-1} \Big(\int_D \sum_{k=1}^{K} \varphi_k(x)^2 f(x)\, dx - \|P_{S_K} f\|^2_{L^2(D)} \Big),$$

wobei $P_{S_K} : L^2(D) \to S_K$ die L^2-Orthogonalprojektion auf S_K bezeichnet und kurz f für die Einschränkung $f|_D$ steht.

Beweis. Wir erweitern $(\varphi_k)_{1\leqslant k\leqslant K}$ zu einer Orthonormalbasis $(\varphi_k)_{k\geqslant 1}$ von $L^2(D)$ und erhalten mit der Parsevalgleichung für $R_D(\hat{f}_{n,K}, f)$:

$$\begin{aligned}
&\mathbb{E}_f\Big[\int_D (\hat{f}_{n,K}(x) - f(x))^2 dx\Big] \\
&= \mathbb{E}_f\Big[\sum_{k=1}^{\infty} \langle \hat{f}_{n,K} - f, \varphi_k\rangle^2_{L^2(D)}\Big] \\
&= \sum_{k=1}^{K} \mathbb{E}_f[\langle \hat{f}_{n,K} - f, \varphi_k\rangle^2_{L^2(D)}] + \sum_{k=K+1}^{\infty} \mathbb{E}_f[\langle f, \varphi_k\rangle^2_{L^2(D)}] \\
&= \sum_{k=1}^{K} \mathrm{Var}_f(\langle \hat{f}_{n,K}, \varphi_k\rangle_{L^2(D)}) + \sum_{k=K+1}^{\infty} \langle f, \varphi_k\rangle^2_{L^2(D)} \\
&= \sum_{k=1}^{K} n^{-1}\Big(\int_D \varphi_k(x)^2 f(x)\,dx - \Big(\int_D \varphi_k(x) f(x)\,dx\Big)^2\Big) + \|f - P_{S_K} f\|^2_{L^2(D)} \\
&= \|f - P_{S_K} f\|^2_{L^2(D)} + n^{-1}\Big(\int_D \sum_{k=1}^{K} \varphi_k(x)^2 f(x)\,dx - \|P_{S_K} f\|^2_{L^2(D)}\Big).
\end{aligned}$$

□

Wiederum sehen wir das Bias-Varianz-Dilemma, sofern wir als Parameter die Dimension K von S_K asymptotisch wachsen lassen. Sofern die Approximationsräume aufsteigend sind, also $S_K \subseteq S_{K+1}$ gilt, und $\bigcup_{K\geqslant 1} S_K$ dicht in $L^2(D)$ liegt, konvergiert der Biasterm $\|f - P_{S_K} f\|_{L^2(D)}$ für $K \to \infty$ gegen Null. Der Varianzterm wird dominiert von $n^{-1}\int_D \sum_{k=1}^{K} \varphi_k(x)^2 f(x)\,dx$ und wegen $\int \varphi_k^2 = 1$ sowie $f \geqslant 0$ ist dieser Term von der Ordnung Kn^{-1}, sofern f beschränkt ist. Halten wir dies als Konsistenzaussage fest.

3.29 Satz. *Der Projektionsschätzer $\hat{f}_{n,K_n}$ ist konsistent in $L^2(D)$, falls f beschränkt ist, $S_{K_n} \subseteq S_{K_n+1}$ für alle $n \geqslant 1$ gilt, $\bigcup_{n\geqslant 1} S_{K_n}$ dicht in $L^2(D)$ liegt und $K_n \to \infty$ für $n \to \infty$ gilt mit $K_n/n \to 0$.*

3.30 Bemerkung. Für unbeschränkte Dichten f kann der Varianzterm durch $n^{-1}\sup_{x\in D}\sum_{k=1}^{K} \varphi_k(x)^2$ abgeschätzt werden. Falls K so langsam mit n wächst, dass dieser Term gegen Null konvergiert, bleibt das Konsistenzresultat für allgemeine Dichten in $L^2(D)$ gültig.

Wie wir sehen, ist der Biasterm allgemein durch den Approximationsfehler $\|f - P_{S_K} f\|_{L^2(D)}$ gegeben, und es ist Aufgabe des Statistikers, gute endlich-dimensionale Approximationsräume S_K für die Dichte f zu finden. Für Glattheitsklassen steht das Instrumentarium der numerischen Analysis und Approximationstheorie zur Verfügung mit Splines, finiten Elementen oder Wavelets. Einfacher sind Ansatzräume mit Polynomen oder trigonometrischen Funktionen. Wir beschränken uns hier auf eine Anwendung mit der

Fourierbasis $\varphi_k(x) = e^{2\pi ikx}$ im komplex-wertigen Funktionenraum $L^2([0,1])$. Genauer werden Projektionsschätzer im Rahmen von Regressionsproblemen untersucht werden.

3.31 Definition. Der periodische L^2-Sobolevraum der Ordnung $s > 0$ auf $[0,1]$ ist definiert als

$$H^s_{per}([0,1]) := \Big\{ f \in L^2([0,1]) \,\Big|\, \sum_{k\in\mathbb{Z}} (1+k^2)^s |\langle f, \varphi_k\rangle_{L^2([0,1])}|^2 < \infty \Big\}.$$

Wir definieren (für diese Vorlesung) die entsprechende Sobolevnorm

$$\|f\|_s := \Big(\sum_{k\in\mathbb{Z}} (1+k^2)^s |\langle f, \varphi_k\rangle_{L^2([0,1])}|^2 \Big)^{1/2}.$$

Man kann wiederum zeigen, dass für $s \in \mathbb{N}$ genau die s-mal schwach differenzierbaren Funktionen f mit quadrat-integrierbaren Ableitungen in $H^s_{per}([0,1])$ liegen, wobei am Intervallrand die Argumente modulo 1 genommen werden. Für ungerade $K \in \mathbb{N}$ setze

$$S_K := \operatorname{span}(\varphi_k,\, k = -(K-1)/2, \ldots, (K-1)/2).$$

Für Dichten f in $H^s_{per}([0,1])$ erhalten wir so die Biasabschätzung

$$\begin{aligned} \|f - P_{S_K} f\|^2_{L^2(D)} &= \sum_{|k| \geqslant (K+1)/2} |\langle f, \varphi_k\rangle|^2 \\ &\leqslant (1 + (K+1)^2/4)^{-s} \|f\|^2_{H^s_{per}} \\ &= \mathcal{O}(K^{-2s}). \end{aligned}$$

Wegen $\|\varphi_k\|_\infty \leqslant 1$ ist der Varianzterm von der Ordnung $\mathcal{O}(n^{-1}K)$. Wählt man $K = \lfloor n^{1/(2s+1)} \rfloor$, was raten-asymptotisch optimal ist, so ergibt sich für $f \in H^s_{per}([0,1])$ ein MISE des Projektionsschätzers von der Ordnung $\mathcal{O}(n^{-2s/(2s+d)})$. Diese Rate gilt gleichmäßig über die Sobolevklasse

$$\Big\{ f \in \mathcal{F}_1 \cap H^s_{per}([0,1]) \,\Big|\, \sum_{k\in\mathbb{Z}} (1+k^2)^s |\langle f, \varphi_k\rangle_{L^2([0,1])}|^2 \leqslant R \Big\}$$

für jedes feste $R > 0$. Die Konvergenzrate ist dieselbe wie für Dichten f in $H^s(\mathbb{R})$ und Kerndichteschätzer.

Mit der d-dimensionalen Fourierbasis erzielt man die Konvergenzrate $\mathcal{O}(n^{-2s/(2s+d)})$ des MISE für Funktionen in Sobolevklassen der Ordnung $s > 0$. Andere Schätzmethoden liefern im wesentlichen dieselbe Konvergenzrate. Dies führt auf die spannende Frage, ob diese Konvergenzrate universell ist in dem Sinne, dass keine Folge von Schätzern asymptotisch einen geringeren MISE aufweisen kann. Wenn man bedenkt, dass Schätzer beliebige messbare Funktionen der Beobachtungen sind, ist es bemerkenswert, dass man in der Tat eine solche untere Schranke in vielen Situationen beweisen kann.

4 Untere Schranken

4.1 Allgemeine Strategie

Wir haben gesehen, dass Kerndichte- und Projektionsschätzer bei geeigneter Bandweitenwahl für punktweises und L^2-Risiko mit gewissen Raten gegen die wahre Dichte f konvergieren. Ziel ist es nun, zu zeigen, dass diese Raten optimal sind. Dazu mache man sich zunächst klar, dass es keinen Sinn ergibt, das Risiko für eine feste Dichte f von unten abzuschätzen: dann besitzt der konstante Schätzer $\hat{f}_n := f$ das Risiko Null. Ein Standardansatz ist daher, das maximale Risiko über eine feste Parametermenge zu betrachten. Dafür können in der Tat untere Schranken hergeleitet werden.

Die Strategie für den Beweis unterer Schranken beruht stets auf denselben Konzepten. Zunächst ist eine Klasse $\mathscr{G}$ von Parametern (Funktionen) f vorgegeben. Zu jedem $f \in \mathscr{G}$ bezeichne $\mathbb{P}_{f,n}$ die Verteilung der Beobachtungen im n-ten Modell (z.B. n i.i.d. Beobachtungen gemäß einem $\mathbb{P}_f$). Das Risiko ergibt sich als Erwartungswert über eine (Semi-)Metrik d und wir betrachten für jeden Schätzer $\hat{f}_n$ in Modell n das maximale quadratische Risiko über $\mathscr{G}$:

$$R_{\mathscr{G}}(\hat{f}_n) := \sup_{f \in \mathscr{G}} \mathbb{E}_{f,n}[d(\hat{f}_n, f)^2].$$

Bislang haben wir dieses Risiko für bestimmte Schätzer und Hölder- bzw. Sobolevklassen von oben abgeschätzt. Nun wollen wir eine untere Schranke für dieses Risiko für beliebige Schätzer beweisen.

4.1 Definition. Es sei $(v_n)_{n \geqslant 1}$ eine Nullfolge und es gebe Schätzer $\hat{f}_n$ im n-ten Modell, so dass

$$\limsup_{n \to \infty} v_n^{-2} \sup_{f \in \mathscr{G}} \mathbb{E}_{f,n}[d(\hat{f}_n, f)^2] < \infty$$

gilt. Dann heißt (v_n) *optimale Konvergenzrate* im Minimaxsinn über $\mathscr{G}$, falls gleichzeitig

$$\liminf_{n \to \infty} v_n^{-2} \inf_{\hat{\vartheta}_n} \sup_{f \in \mathscr{G}} \mathbb{E}_{f,n}[d(\hat{\vartheta}_n, f)^2] > 0$$

gilt, wobei sich das Infimum über alle Schätzer (d.h. messbaren Funktionen) $\hat{\vartheta}_n$ in Modell n erstreckt.

4.2 Bemerkung. Wir betrachten hier den quadrierten Abstand, natürlich sind aber auch andere Momente oder Funktionen des Abstands denkbar und gebräuchlich. Man beachte, dass die optimale Konvergenzrate nur die Ordnung angibt, in diesem Sinne also $v_n = n^{-s/(2s+1)}$ und $v_n' = 7n^{-s/(2s+1)}$ gleichbedeutend sind, während $v_n'' = (n/\log n)^{-s/(2s+1)}$ eine langsamere Konvergenzrate angibt. Sind wir auch an asymptotischen Konstanten interessiert (den Grenzwerten), so wird die Situation um einiges schwieriger. In gewissen Fällen können sogenannte Pinsker-Konstanten nachgewiesen werden, vergleiche Tsybakov (2004).

Wir diskutieren nun, wie der Nachweis einer unteren Schranke auf ein Testproblem reduziert werden kann, und geben in diesem allgemeinen Rahmen bereits untere Schranken an.

Beachte zunächst, dass es ausreicht, untere Schranken für die stochastische Konvergenz nachzuweisen. Wegen

$$\forall \alpha > 0 : \; v_n^{-2}\, \mathbb{E}_{f,n}[d(\hat{\vartheta}_n, f)^2] \geqslant \alpha^2\, \mathbb{P}_{f,n}(d(\hat{\vartheta}_n, f) \geqslant \alpha v_n)$$

genügt es für ein $\alpha > 0$ folgende Asymptotik nachzuweisen:

$$\liminf_{n\to\infty} \inf_{\hat{\vartheta}_n} \sup_{f\in\mathscr{G}} \mathbb{P}_{f,n}(d(\hat{\vartheta}_n, f) \geqslant \alpha v_n) > 0.$$

Die Reduktion auf ein Testproblem besteht nun darin, eine endliche Teilmenge $\{f_1, \dots, f_M\} \subseteq \mathscr{G}$ zu betrachten, mit

$$\forall k \neq l : d(f_k, f_l) > 2\alpha v_n.$$

Damit erhalten wir für jeden Schätzer $\hat{\vartheta}_n$

$$\begin{aligned}\sup_{f\in\mathscr{G}} \mathbb{P}_{f,n}(d(\hat{\vartheta}_n, f) \geqslant \alpha v_n) &\geqslant \max_{j=1,\dots,M} \mathbb{P}_{f_j,n}(d(\hat{\vartheta}_n, f_j) \geqslant \alpha v_n) \\ &\geqslant \max_{j=1,\dots,M} \mathbb{P}_{f_j,n}(\psi_n^* \neq j),\end{aligned}$$

wobei $\psi_n^* := \operatorname{argmin}_{j=1,\dots,M} d(\hat{\vartheta}_n, f_j)$ den auf $\hat{\vartheta}_n$ beruhenden Minimum-Distanz-Test zwischen den M Hypothesen $H_j : f = f_j$ bezeichnet. Können wir nun nachweisen, dass

$$\liminf_{n\to\infty} \inf_{\psi_n} \max_{j=1,\dots,M} \mathbb{P}_{f_j,n}(\psi_n \neq j) > 0$$

gilt für alle Tests ψ_n in Modell n, so folgt insbesondere die untere Schranke für das Schätzproblem.

4.3 Satz. *Es seien $\mathbb{P}_1, \dots, \mathbb{P}_M$ Wahrscheinlichkeitsmaße auf $(\mathcal{X}, \mathscr{F})$, die bezüglich einem Maß μ absolut-stetig sind mit Dichten $p_1, \dots, p_M$ (man kann z.B. stets $\mu = \sum_{i=1}^M \mathbb{P}_i$ betrachten). Dann gilt für jeden Test $\psi : \mathcal{X} \to \{1, \dots, M\}$ zwischen den M Hypothesen*

$$\max_{j=1,\dots,M} \mathbb{P}_j(\psi \neq j) \geqslant \frac{1}{M} \sum_{j=1}^M \mathbb{P}_j(\psi \neq j) \geqslant 1 - \frac{1}{M} \int_{\mathcal{X}} \max_{j=1,\dots,M} p_j(x)\, \mu(dx)$$

Beweis. Die erste Ungleichung gilt, weil der Durchschnitt stets kleiner oder gleich dem Maximum ist. Für die zweite Ungleichung beachte

$$\begin{aligned}\frac{1}{M} \sum_{j=1}^M \mathbb{P}_j(\psi \neq j) &= 1 - \frac{1}{M} \sum_{j=1}^M \int \mathbf{1}(\psi(x) = j) p_j(x)\, \mu(dx) \\ &= 1 - \frac{1}{M} \int \Big(\sum_{j=1}^M \mathbf{1}(\psi(x) = j) p_j(x) \Big)\, \mu(dx).\end{aligned}$$

Nun beachte, dass der letzte Integrand maximal gleich $\max_j p_j(x)$ ist, so dass

$$\max_{j=1,\dots,M} \mathbb{P}_j(\psi \neq j) \geqslant 1 - \frac{1}{M} \int \max_{j=1,\dots,M} p_j(x)\, \mu(dx)$$

folgt, wie behauptet. □

Verwenden wir diesen Satz, so erhalten wir insgesamt die Abschätzung

$$\inf_{\hat{\vartheta}_n} \sup_{f \in \mathscr{G}} v_n^{-2}\, \mathbb{E}_{f,n}[d(\hat{\vartheta}_n, f)^2] \geqslant \alpha^2 \Big(1 - \frac{1}{M_n} \int_{\mathcal{X}} \max_{j=1,\dots,M_n} p_{j,n}(x)\, \mu_n(dx)\Big),$$

wobei wir auf der rechten Seite die mögliche Abhängigkeit von n der Größen angeben. Beachte, dass wir bereits einen großen Schritt gemacht haben, indem auf der rechten Seite kein Infimum mehr erscheint. Es bleibt, das Integral möglichst gut abzuschätzen, wozu wir im folgenden Abstandsmaße für Wahrscheinlichkeiten untersuchen werden. Beachte, dass die größte Herausforderung ist, die endliche Teilmenge $f_{1,n}, \dots, f_{M_n,n}$ a priori so geschickt zu wählen, dass obige Ungleichungen nicht zu Verlusten in der Rate führen. Im Bayesschen Sinne sollte die Gleichverteilung auf $f_{1,n}, \dots, f_{M_n,n}$ zu einer ungünstigsten a priori-Verteilung im Modell n führen (zumindest approximativ).

4.4 Definition. Für zwei Maße μ und ν auf $(\mathcal{X}, \mathscr{F})$ ist der *Totalvariations-Abstand* gegeben durch

$$\|\mu - \nu\|_{TV} := \sup_{A \in \mathscr{F}} |\mu(A) - \nu(A)|.$$

4.5 Bemerkung. Die Totalvariationsnorm $\|\mu\|_{TV} := \sup_{A \in \mathscr{F}} |\mu(A)|$ ist eine Norm im Raum aller endlichen signierten Maße und spielt eine große Rolle in Maßtheorie und Funktionalanalysis, siehe z.B. Elstrodt (2007), Werner (2007).

4.6 Lemma. *Es seien $\mathbb{P}$ und $\mathbb{Q}$ Wahrscheinlichkeitsmaße auf $(\mathcal{X}, \mathscr{F})$ mit Dichten p und q bezüglich einem Maß μ. Dann gilt:*

(a) $0 \leqslant \|\mathbb{P} - \mathbb{Q}\|_{TV} \leqslant 1$;

(b) $\|\mathbb{P} - \mathbb{Q}\|_{TV} = \frac{1}{2} \int_{\mathcal{X}} |p(x) - q(x)|\, \mu(dx) = \int_{\mathcal{X}} \max(p(x), q(x))\, \mu(dx) - 1.$

Beweis. Teil (a) folgt direkt aus der Definition wegen $\mathbb{P}(\mathcal{X}) = \mathbb{Q}(\mathcal{X}) = 1$. Wegen $|A-B| = 2\max(A,B) - A - B$ sowie $\int p d\mu = \int q d\mu = 1$ brauchen wir in Teil (b) nur die erste Identität zu zeigen. Setzt man $A^* := \{x \in \mathcal{X} \,|\, p(x) > q(x)\}$, so gilt $A^* \in \mathscr{F}$ sowie wegen $\int (p-q) d\mu = 0$

$$\begin{aligned} \mathbb{P}(A^*) - \mathbb{Q}(A^*) &= \int_{A^*} (p(x) - q(x))\, \mu(dx) - \frac{1}{2} \int_{\mathcal{X}} (p(x) - q(x))\, \mu(dx) \\ &= \frac{1}{2} \int_{\mathcal{X}} |p(x) - q(x)| \mu(dx). \end{aligned}$$

Dies zeigt $\|\mathbb{P}-\mathbb{Q}\|_{TV} \geqslant \frac{1}{2}\int_{\mathcal{X}}|p(x)-q(x)|\,\mu(dx)$. Andererseits gilt für beliebiges $A \in \mathscr{F}$

$$\begin{aligned}|\mathbb{P}(A)-\mathbb{Q}(A)| &= \Big|\int_{\mathcal{X}}(p(x)-q(x))(\mathbf{1}(A\cap A^*)+\mathbf{1}(A\setminus A^*))\,\mu(dx)\Big| \\ &\leqslant \max\Big(\mathbb{P}(A^*)-\mathbb{Q}(A^*), \mathbb{Q}(\mathcal{X}\setminus A^*)-\mathbb{P}(\mathcal{X}\setminus A^*)\Big).\end{aligned}$$

Aus $\mathbb{P}(\mathcal{X})=\mathbb{Q}(\mathcal{X})=1$ folgt, dass die beiden Argumente von max gleich sind und insbesondere mit $\frac{1}{2}\int_{\mathcal{X}}|p(x)-q(x)|\mu(dx)$ zusammenfallen. Dies impliziert die behauptete Identität. □

Aus statistischer Sicht hat der Totalvariationsabstand eine natürliche Interpretation als Minimaxfehler beim Testen.

4.7 Korollar. *Es seien P_0 und $\mathbb{P}_1$ Wahrscheinlichkeitsmaße auf $(\mathcal{X},\mathscr{F})$. Dann gilt*

$$\inf_{\psi}\max_{j=0,1}\mathbb{P}_j(\psi\neq j) \geqslant \inf_{\psi}\frac{1}{2}\Big(\mathbb{P}_0(\psi=1)+\mathbb{P}_1(\psi=0)\Big) = \frac{1}{2}\Big(1-\|\mathbb{P}_0-\mathbb{P}_1\|_{TV}\Big),$$

wobei sich das Infimum über beliebige Tests (messbare Funktionen $\psi:\mathcal{X}\to\{0,1\}$) zwischen den Hypothesen $H_0:\mathbb{P}=\mathbb{P}_0$ und $H_1:\mathbb{P}=\mathbb{P}_1$ erstreckt. Die untere Schranke wird vom Neyman-Pearson-Test $\psi_{NM}(x):=\mathbf{1}(p_1(x)>p_0(x))$ erreicht, wobei p_0,p_1 die Dichten von $\mathbb{P}_0$, $\mathbb{P}_1$ bezüglich irgendeinem Maß μ bezeichnen.

Beweis. Aus Satz 4.3 folgt

$$\begin{aligned}\inf_{\psi}\max_{j=0,1}\mathbb{P}_j(\psi\neq j) &\geqslant \inf_{\psi}\frac{1}{2}\Big(\mathbb{P}_0(\psi=1)+\mathbb{P}_1(\psi=0)\Big) \\ &\geqslant 1-\frac{1}{2}\int\max(p_0(x),p_1(x))\mu(dx),\end{aligned}$$

wobei ein Blick auf den Beweis oder einfaches Einsetzen zeigt, dass für ψ_{NM} die untere Schranke in der zweiten Ungleichung angenommen wird. Aus Teil (b) im vorangegangenen Lemma folgt daher die Behauptung. □

4.8 Beispiele.

(a) Gilt $\mathbb{P}_0=\mathbb{P}_1$, so sind Hypothese und Alternative nicht unterscheidbar. Die Summe der Fehlerwahrscheinlichkeiten erster und zweiter Art sind für jeden Test gleich Eins. Ein randomisierter Test ψ, der sich mit Wahrscheinlichkeit $1/2$ für H_0 bzw. H_1 entscheidet, erreicht auch $\max_{j=0,1}\mathbb{P}_j(\psi\neq j)=1/2$.

(b) Sind $\mathbb{P}_0$ und $\mathbb{P}_1$ singulär, gibt es also ein $A\in\mathscr{F}$ mit $\mathbb{P}_0(A)=1, \mathbb{P}_1(A)=0$, so erhalten wir $\|\mathbb{P}_0-\mathbb{P}_1\|_{TV}=1$ und der Neyman-Pearson-Test entscheidet fast sicher korrekt zwischen Hypothese und Alternative.

Für unser Ziel, untere Schranken im Schätzproblem herzuleiten, ergibt sich aus Lemma 4.6 im Fall einer Reduktion auf ein einfaches Testproblem ($M = 2$):

$$\inf_{\hat{\vartheta}_n} \sup_{f \in \mathscr{G}} \mathbb{E}_{f,n}[d(\hat{\vartheta}_n, f)^2] \geqslant \frac{\alpha^2 v_n^2}{2}\Big(1 - \|\mathbb{P}_{f_1,n} - \mathbb{P}_{f_2,n}\|_{TV}\Big).$$

Diese Ungleichung ist bereits sehr nützlich, allerdings ist in vielen Fällen die Totalvariationsnorm sehr schwierig direkt abzuschätzen. Wir führen daher ein weiteres Abstandsmaß ein.

4.9 Definition. Für zwei Wahrscheinlichkeitsmaße $\mathbb{P}$ und $\mathbb{Q}$ auf demselben Messraum $(\mathcal{X}, \mathscr{F})$ heißt

$$\mathrm{KL}(\mathbb{P} \mid \mathbb{Q}) = \begin{cases} \int_{\mathcal{X}} \log\big(\frac{d\mathbb{P}}{d\mathbb{Q}}(x)\big)\ \mathbb{P}(dx), & \text{falls } \mathbb{P} \ll \mathbb{Q}, \\ +\infty, & \text{sonst} \end{cases}$$

Kullback-Leibler-Divergenz (oder auch *Kullback-Leibler-Abstand*, *relative Entropie*) von $\mathbb{P}$ bezüglich $\mathbb{Q}$.

4.10 Lemma. $\mathrm{KL}(\mathbb{P} \mid \mathbb{Q})$ *ist keine Metrik (nicht einmal die Symmetrisierung* $\mathrm{KL}(\mathbb{P} \mid \mathbb{Q}) + \mathrm{KL}(\mathbb{Q} \mid \mathbb{P})$ *ist eine). Es gelten jedoch folgende Eigenschaften:*

(a) Gilt $\mathbb{P} \ll \mathbb{Q}$ *und existieren Dichten* p *und* q *von* $\mathbb{P}$ *bzw.* $\mathbb{Q}$ *bezüglich einem Maß* μ*, so gilt*

$$\mathrm{KL}(\mathbb{P} \mid \mathbb{Q}) = \int_{pq>0} \log\big(\frac{p(x)}{q(x)}\big) p(x)\, \mu(dx);$$

(b) $\mathrm{KL}(\mathbb{P} \mid \mathbb{Q}) \in [0, +\infty]$ *sowie* $\mathrm{KL}(\mathbb{P} \mid \mathbb{Q}) = 0 \iff \mathbb{P} = \mathbb{Q}$*;*

(c) für Produktmaße gilt $\mathrm{KL}(\mathbb{P}_1 \otimes \mathbb{P}_2 \mid \mathbb{Q}_1 \otimes \mathbb{Q}_2) = \mathrm{KL}(\mathbb{P}_1 \mid \mathbb{Q}_1) + \mathrm{KL}(\mathbb{P}_2 \mid \mathbb{Q}_2)$*, insbesondere* $\mathrm{KL}(\mathbb{P}^{\otimes n} \mid \mathbb{Q}^{\otimes n}) = n\,\mathrm{KL}(\mathbb{P} \mid \mathbb{Q})$.

Beweis. ▶ Übung □

4.11 Satz (Pinsker)**.** *Folgende Ungleichungen gelten für Wahrscheinlichkeitsmaße* $\mathbb{P}$ *und* $\mathbb{Q}$*:*

(a) $\|\mathbb{P} - \mathbb{Q}\|_{TV} \leqslant \sqrt{\mathrm{KL}(\mathbb{P} \mid \mathbb{Q})/2}$*;*

(b) $\int (\log(d\mathbb{P}/d\mathbb{Q}))_- dP \leqslant \|\mathbb{P} - \mathbb{Q}\|_{TV}$*, falls* $\mathbb{P} \ll \mathbb{Q}$
(man setzt $A_- := \max(-A, 0)$*).*

4.12 Bemerkung. Aus Teil (a) folgt insbesondere, dass $\mathrm{KL}(\mathbb{P}_n \mid \mathbb{P}) \to 0$ die Konvergenz $\mathbb{P}_n \to \mathbb{P}$ in Totalvariationsnorm impliziert. ▶ Übung Man kann zeigen, dass die Umkehrung nicht gilt, auch nicht unter der Voraussetzung $\mathbb{P}_n \ll \mathbb{P}$ für alle n.

Beweis. Für Teil (a) genügt es ebenfalls, nur den Fall $\mathbb{P} \ll \mathbb{Q}$ zu behandeln. Betrachte dazu die Funktion

$$h(z) := z\log(z) - z + 1, \quad z \geqslant 0,$$

wobei man $z\log(z) = 0$ für $z = 0$ stetig ergänzt. ▶ Übung Man kann zeigen, dass

$$\forall\, z \geqslant 0 :\ (\tfrac{4}{3} + \tfrac{2}{3}z)h(z) \geqslant (z-1)^2$$

gilt. Daraus schließen wir mittels Lemma 4.6 für μ-Dichten p und q sowie mit der Cauchy-Schwarz-Ungleichung:

$$\begin{aligned}
\|\mathbb{P} - \mathbb{Q}\|_{TV} &= \frac{1}{2}\int |p(x) - q(x)|\,\mu(dx) \\
&\leqslant \frac{1}{2}\int_{q>0} q(x)\Big((\tfrac{4}{3} + \tfrac{2p(x)}{3q(x)})h(p(x)/q(x))\Big)^{1/2}\,\mu(dx) \\
&\leqslant \frac{1}{2}\Big(\int (\tfrac{4q(x)}{3} + \tfrac{2p(x)}{3})\,\mu(dx)\Big)^{1/2}\Big(\int_{q>0} q(x)h(p(x)/q(x))\,\mu(dx)\Big)^{1/2} \\
&= \frac{1}{\sqrt{2}}\Big(\int_{pq>0} p(x)\log(p(x)/q(x))\,\mu(dx)\Big)^{1/2} \\
&= \sqrt{\mathrm{KL}(\mathbb{P} \mid \mathbb{Q})/2}.
\end{aligned}$$

Um Teil (b) nachzuweisen, setze $A := \{x \in \mathcal{X} \mid q(x) \geqslant p(x) > 0\}$. Wegen $\int_{\mathcal{X}} \log(d\,\mathbb{P}/d\,\mathbb{Q})_- d\,\mathbb{P} = \int_A \log(d\,\mathbb{Q}/d\,\mathbb{P}) d\,\mathbb{P}$ folgt aus $\mathbb{P}(A) = 0$ offensichtlich $\int_{\mathcal{X}} \log(d\,\mathbb{P}/d\,\mathbb{Q})_- d\,\mathbb{P} = 0$. Nehme also $\mathbb{P}(A) > 0$ an und schließe mittels Jensenscher Ungleichung

$$\begin{aligned}
\int_A \log(d\,\mathbb{P}/d\,\mathbb{Q})_- d\,\mathbb{P} &= \mathbb{P}(A)\int_A \log(q(x)/p(x))\frac{p(x)}{\mathbb{P}(A)}\,\mu(dx) \\
&\leqslant \mathbb{P}(A)\log\Big(\int_A \frac{q(x)}{p(x)}\frac{p(x)}{\mathbb{P}(A)}\,\mu(dx)\Big) \\
&= \mathbb{P}(A)\log(\mathbb{Q}(A)/\,\mathbb{P}(A)) \\
&\leqslant \mathbb{Q}(A) - \mathbb{P}(A) = \|\mathbb{Q} - \mathbb{P}\|_{TV}.
\end{aligned}$$

□

Indem wir die Abschätzung aus Teil (a) des Satzes verwenden, erhalten wir auch eine untere Schranke bei Reduktion auf ein einfaches Testproblem in Kullback-Leibler-Version. Wir fassen zusammen.

4.13 Satz. *Es seien $f_1, f_2 \in \mathscr{G}$ mit $d(f_1, f_2) > 2\alpha v_n$ für ein $\alpha > 0$. Für die zugehörigen Verteilungen gelte mit einem $\delta > 0$*

$$\|\mathbb{P}_{f_1,n} - \mathbb{P}_{f_2,n}\|_{TV} \leqslant 1 - 2\delta \text{ oder } \mathrm{KL}(\mathbb{P}_{f_1,n} \mid \mathbb{P}_{f_2,n}) \leqslant 2(1-2\delta)^2.$$

Dann erhalten wir die untere Schranke

$$\inf_{\hat{\vartheta}_n} \sup_{f \in \mathscr{G}} \mathbb{E}_{f,n}[d(\hat{\vartheta}_n, f)^2] \geqslant \alpha^2 v_n^2 \delta.$$

Im Fall eines Produktmodells $\mathbb{P}_{f_1,n} = \mathbb{P}_{f_1}^{\otimes n}$, $\mathbb{P}_{f_2,n} = \mathbb{P}_{f_2}^{\otimes n}$ *genügt die Bedingung* $\mathrm{KL}(\mathbb{P}_{f_1} \mid \mathbb{P}_{f_2}) \leqslant 2(1-2\delta)^2 n^{-1}$.

Es wird sich herausstellen, dass für punktweises Risiko diese Reduktion ausreicht, während für den MISE oder gleichmäßiges Risiko eine Reduktion auf ein Testproblem mit mehreren Hypothesen notwendig ist. Nach Satz 4.3 bedarf es dazu einer Abschätzung des Maximums der jeweiligen Dichten. Für $M = 2$ war dafür der Totalvariationsabstand das natürliche Werkzeug. Im Fall $M > 2$ zeigt sich, dass unmittelbar die Kullback-Leibler-Divergenz verwendet werden sollte.

4.14 Satz. *Es seien* $f_1, \dots, f_M \in \mathscr{G}$ *mit* $d(f_k, f_l) > 2\alpha v_n$ *für ein* $\alpha > 0$ *und alle* $k \neq l$. *Es bezeichne* $\mathbb{P}_0$ *ein Wahrscheinlichkeitsmaß mit* $\mathbb{P}_{f_j,n} \ll \mathbb{P}_0$ *für alle* j *und setze*

$$\delta := \max\Big(\frac{1}{M}\sum_{j=1}^{M} \mathrm{KL}(\mathbb{P}_{f_j,n} \mid \mathbb{P}_0), \frac{1}{M}\Big).$$

Dann erhalten wir die untere Schranke

$$\inf_{\hat{\vartheta}_n} \sup_{f \in \mathscr{G}} \mathbb{E}_{f,n}[d(\hat{\vartheta}_n, f)^2] \geqslant \alpha^2 v_n^2 \Big(1 - \frac{2\delta + \sqrt{2\delta}}{\log(M) + \log(\delta + \sqrt{\delta/2})}\Big).$$

Für $\delta \geqslant 1/2$ *gilt die gröbere Abschätzung*

$$\inf_{\hat{\vartheta}_n} \sup_{f \in \mathscr{G}} \mathbb{E}_{f,n}[d(\hat{\vartheta}_n, f)^2] \geqslant \alpha^2 v_n^2 \Big(1 - \frac{4\delta}{\log(M)}\Big).$$

Beweis. Mittels Jensenscher Ungleichung, angewendet auf die monotone und konvexe Funktion $H(z) := z(\log z)_+$, $z \geqslant 0$, erhalten wir unter Benutzung der Ungleichungen aus Satz 4.11

$$\begin{aligned}
H\Big(\mathbb{E}_0\Big[\max_{j=1,\dots,M} p_j\Big]\Big) &\leqslant \mathbb{E}_0\Big[H\Big(\max_{j=1,\dots,M} p_j\Big)\Big] \\
&= \mathbb{E}_0\Big[\max_{j=1,\dots,M} H(p_j)\Big] \\
&\leqslant \sum_{j=1}^{M} \mathbb{E}_0[H(p_j)] \\
&= \sum_{j=1}^{M} \Big(\mathrm{KL}(\mathbb{P}_j \mid \mathbb{P}_0) + \int p_j(x)(\log p_j(x))_- \, \mathbb{P}_0(dx)\Big)
\end{aligned}$$

$$\leqslant \sum_{j=1}^{M} \Big(\mathrm{KL}(\mathbb{P}_j \mid \mathbb{P}_0) + \|\mathbb{P}_j - \mathbb{P}_0\|_{TV} \Big)$$
$$\leqslant \sum_{j=1}^{M} \Big(\mathrm{KL}(\mathbb{P}_j \mid \mathbb{P}_0) + \sqrt{\mathrm{KL}(\mathbb{P}_j \mid \mathbb{P}_0)/2} \Big).$$

Mit der Jensenschen Ungleichung folgt

$$\frac{1}{M} \sum_{j=1}^{M} \sqrt{\mathrm{KL}(\mathbb{P}_j \mid \mathbb{P}_0)/2} \leqslant \Big(\frac{1}{M} \sum_{j=1}^{M} \mathrm{KL}(\mathbb{P}_j \mid \mathbb{P}_0)/2 \Big)^{1/2} \leqslant \sqrt{\delta/2}.$$

Die Inverse von H erfüllt $H^{-1}(x) \leqslant 2x/\log x$ für $x > 1$ (wende H auf beide Seiten an), so dass wir wegen $M(\delta + \sqrt{\delta/2}) > 1$ erhalten

$$\mathbb{E}_0 \Big[\max_{j=1,\dots,M} p_j \Big] \leqslant H^{-1}(M\delta + M\sqrt{\delta/2}) \leqslant M \frac{2\delta + \sqrt{2\delta}}{\log(M) + \log(\delta + \sqrt{\delta/2})}.$$

Daher liefert Satz 4.3 die behauptete Abschätzung. Die Vereinfachung ergibt sich aus $\sqrt{2\delta} \leqslant 2\delta$ für $2\delta \geqslant 1$. □

4.2 Dichteschätzprobleme

Wir wenden nun das allgemeine Schema zum Beweis unterer Schranken auf das Dichteschätzproblem mit punktweisen Risiko bzw. MISE an. Jeweils der wichtigste Schritt ist das Auffinden ungünstigster Alternativen. Beim punktweisen Risiko in einem Punkt x ist es natürlich, Dichten zu betrachten, die sich nur in einer kleinen Umgebung von x unterscheiden, um nicht zusätzliche Informationen aus weiter entfernt liegenden Beobachtungen preiszugeben. In der Tat führt eine lokale Störung um x von einer gegebenen Dichte für Hölderklassen zum Erfolg.

4.15 Satz. *Betrachte das Dichteschätzproblem für die Hölderklasse $\mathscr{H}_D(\alpha, L, R)$ mit $\alpha, L, R > 0$ auf einem Gebiet $D \subseteq \mathbb{R}^d$ und $x_0 \in D$. Dann gilt die asymptotische untere Schranke*

$$\liminf_{n\to\infty} n^{2\alpha/(2\alpha+d)} \inf_{\hat{\vartheta}_n} \sup_{f \in \mathscr{H}_D(\alpha,L,R)} \mathbb{E}_{f,n}[|\hat{\vartheta}_n - f(x_0)|^2] > 0,$$

wobei sich das Infimum über beliebige Schätzer $\hat{\vartheta}_n$ basierend auf n Beobachtungen erstreckt. Die Rate $n^{-\alpha/(2\alpha+d)}$ ist die optimale Minimaxrate für dieses Problem.

Beweis. Es sei $f \in \mathscr{H}_D(\alpha, L', R')$ mit $L' < L$, $R' < R$ sowie $f(x_0) > 0$ beliebig gegeben. Weiterhin sei $\varphi : \mathbb{R}^d \to \mathbb{R}$ eine glatte (mindestens α-reguläre) Funktion mit Träger in $[-1,1]^d$ und $\varphi(0) > 0$, $\int \varphi = 0$. Dann ist für $\gamma \in (0, f(x_0)/\|\varphi\|_\infty)$ und hinreichend kleines $h > 0$

$$f_h(x) := f(x) + \gamma\varphi(h^{-1}(x - x_0)), \quad x \in \mathbb{R}^d,$$

wiederum eine Dichte. Da $\varphi(h^{-1}\bullet)$ durch $\|\varphi\|_\infty$ beschränkt ist sowie eine Hölderkonstante L_h von der Ordnung $h^{-\alpha}$ besitzt, liegt f_h in der Klasse $\mathscr{H}_D(\alpha, L, R)$ für $\gamma = \gamma_0 h^\alpha$ mit $\gamma_0 > 0$ hinreichend klein. Beachte, dass für den punktweisen Abstand d dann gilt

$$d(f, f_h) = |f(x_0) - f_h(x_0)| = h^\alpha \gamma_0 \varphi(0).$$

Andererseits gilt für den Kullback-Leibler-Abstand der Verteilungen $\mathbb{P}_h^{\otimes n}$ und $\mathbb{P}^{\otimes n}$ von n i.i.d. Beobachtungen gemäß f_h bzw. f:

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_h^{\otimes n} \mid \mathbb{P}^{\otimes n}) &= n\,\mathrm{KL}(\mathbb{P}_h \mid \mathbb{P}) \\
&= n \int_{\mathbb{R}^d} \log\Big(1 + \frac{\gamma\varphi(h^{-1}(x - x_0))}{f(x)}\Big) f_h(x)\,dx \\
&\leqslant n\gamma \int_{\mathbb{R}^d} \varphi(h^{-1}(x - x_0))\Big(1 + \frac{\gamma\varphi(h^{-1}(x - x_0))}{f(x)}\Big)\,dx \\
&= n\gamma^2 \int_{\mathbb{R}^d} \frac{\varphi(h^{-1}(x - x_0))^2}{f(x)}\,dx \\
&\leqslant n\gamma^2 \max_{|x - x_0|_\infty \leqslant h} \big(f(x)^{-1}\big) h^d \|\varphi\|_{L^2}^2 \\
&= n h^{2\alpha + d} \gamma_0^2 \|\varphi\|_{L^2}^2 \big(\min_{|x|_\infty \leqslant h} f(x_0 + x)\big)^{-1}.
\end{aligned}$$

Wählen wir nun $h = n^{-1/(2\alpha+d)} h_0$ mit $h_0 > 0$ hinreichend klein, so ist das Minimum aus Stetigkeitsgründen echt positiv sowie der gesamte letzte Ausdruck kleiner als eins für alle $n \geqslant 1$. Gemäß Satz 4.13 können wir also $\delta = (2 + \sqrt{2})/4$ wählen und erhalten

$$\inf_{\hat\vartheta_n} \sup_{f \in \mathscr{H}_D(\alpha, L, R)} \mathbb{E}_{f,n}[|\hat\vartheta_n - f(x_0)|^2] \geqslant \delta d(f, f_h)^2/5 = n^{-2\alpha/(2\alpha+d)} \delta h_0^{2\alpha} \gamma_0^2 \varphi(0)^2/5,$$

was zu zeigen war. Die Optimalität der Rate ergibt sich aus Satz 2.16. □

▶ Übung Betrachten wir dieselben Alternativen, um den MISE abzuschätzen, so erhalten wir nur eine untere Schranke der Ordnung $n^{-1/2}$, also die parametrische Rate. Da die Verlustfunktion durch Integration über ein Gebiet gegeben ist, müssen wir $M > 2$ Alternativen auf dem gesamten Gebiet betrachten. Dafür wird folgendes Lemma aus der Informationstheorie sehr nützlich sein, das die Komplexität hochdimensionaler 0-1-Vektoren bezüglich dem sogenannten Hammingabstand abschätzt.

4.16 Lemma (Varshamov-Gilbert, 1962). *Es sei $E := \{0,1\}^m$ mit $m \geqslant 8$ die Menge der m-dimensionalen 0-1-Vektoren. Führe den Hammingabstand $\rho(\varepsilon, \varepsilon') := \sum_{i=1}^m \mathbf{1}(\varepsilon_i \neq \varepsilon_i')$ ein.*

Dann gibt es eine Teilmenge $\{\varepsilon^{(0)}, \dots, \varepsilon^{(J)}\} \subseteq E$ vom Umfang $J \geqslant 2^{m/8}$ mit $\varepsilon^{(0)} = (0, \dots, 0)$, so dass $\rho(\varepsilon^{(k)}, \varepsilon^{(\ell)}) \geqslant m/8$ für alle $k \neq \ell$ gilt.

Beweis. Setze $D = \lfloor m/8 \rfloor$, $\varepsilon^{(0)} := (0, \dots, 0)$ und

$$E_1 := \{\varepsilon \in E \,|\, \rho(\varepsilon, \varepsilon^{(0)}) > D\}.$$

Wähle nun ein $\varepsilon^{(1)} \in E_1$ beliebig und fahre entsprechend iterativ fort. Das heißt, wir setzen für $j \geqslant 2$

$$E_j := \{\varepsilon \in E_{j-1} \,|\, \rho(\varepsilon, \varepsilon^{(j-1)}) > D\}$$

und wählen $\varepsilon^{(j)}$ beliebig aus E_j, solange E_j nicht die leere Menge ist. J bezeichne den letzten Index j mit $E_j \neq \varnothing$. Nach Konstruktion gilt für alle $0 \leqslant k < \ell \leqslant J$ die Ungleichung

$$\rho(\varepsilon^{(k)}, \varepsilon^{(\ell)}) \geqslant D + 1 > m/8.$$

Um $J \geqslant 2^{m/8}$ nachzuweisen, analysieren wir die Kardinalität n_j der im j-ten Schritt ausgeschlossenen Vektoren:

$$n_j := \#(E_j \setminus E_{j+1}) \leqslant \#\{\varepsilon \in E \,|\, \rho(\varepsilon, \varepsilon^{(j)}) \leqslant D\} = \sum_{i=0}^{D} \binom{m}{i}.$$

Nach Definition gilt $\sum_{j=0}^{J} n_j = \#E = 2^m$, so dass

$$(J+1) \sum_{i=0}^{D} \binom{m}{i} \geqslant 2^m \text{ bzw. } \sum_{i=0}^{D} \binom{m}{i} 2^{-m} \geqslant (J+1)^{-1}$$

folgt. Beachte nun, dass die linke Seite in der zweiten Darstellung die Wahrscheinlichkeit $\mathbb{P}(S_m \leqslant D)$ angibt für eine $\mathrm{Bin}(m, 1/2)$-verteilte Zufallsvariable S_m. Aus der Hoeffding-Ungleichung folgt nach Beispiel 5.10(a)

$$\mathbb{P}(S_m \leqslant D) = \mathbb{P}(S_m - m/2 \leqslant D - m/2) \leqslant 2e^{-2(D-m/2)^2/m} \leqslant 2e^{-9m/32}.$$

Wegen $e^{-9m/32} < 2^{-m/3}$ erhalten wir $J + 1 \geqslant 2^{m/3-1} \geqslant 2^{m/8} + 1$ für $m \geqslant 8$. □

4.17 Satz. *Betrachte das Dichteschätzproblem für die Sobolevklasse* $\mathscr{G}(s, R) := \{f \in H^s(\mathbb{R}^d) \,|\, f \text{ ist Dichte}, \|f\|_s \leqslant R\}$ *mit* $s \in \mathbb{N}$, $R > 0$. *Für den MISE auf dem Einheitswürfel* $[0,1]^d$ *gilt dann die asymptotische untere Schranke*

$$\liminf_{n\to\infty} n^{2s/(2s+d)} \inf_{\hat{\vartheta}_n} \sup_{f \in \mathscr{G}(s,R)} \mathbb{E}_{f,n} \Big[\int_{[0,1]^d} |\hat{\vartheta}_n(x) - f(x)|^2 dx \Big] > 0,$$

wobei sich das Infimum über beliebige $L^2([0,1]^d)$*-wertige Schätzer* $\hat{\vartheta}_n$ *basierend auf* n *Beobachtungen erstreckt. Die Rate* $n^{-s/(2s+d)}$ *ist die optimale Minimaxrate für dieses Problem.*

4.18 Bemerkung. Die Aussage des Satzes bleibt auch korrekt für nichtganzzahlige $s > 0$ sowie beliebige Mengen $D \subseteq \mathbb{R}^d$ mit nichtleerem Innern anstelle von $D = [0,1]^d$. In der angegebenen Situation ist der Beweis jedoch transparenter.

Beweis. Wir beschränken uns zunächst auf den eindimensionalen Fall $d = 1$. Betrachte eine Dichte $f \in H^s(\mathbb{R})$ mit $\|f\|_s < R$ sowie $c_f := \inf_{x\in[0,1]} f(x) > 0$. Verwende wiederum eine glatte (mindestens s-reguläre) Funktion $\varphi : \mathbb{R} \to \mathbb{R}$ mit Träger in $[0,1]$ und $\int \varphi = 0$, $\varphi \neq 0$. Für ein später spezifiertes $m \geqslant 1$ setze für jedes $\varepsilon \in E = \{0,1\}^m$

$$f_\varepsilon(x) = f(x) + \gamma \sum_{i=1}^m \varepsilon_i \varphi(mx - (i-1)).$$

Jedes f_ε ist wiederum eine Dichte für $\gamma < c_f/\|\varphi\|_\infty$. Beachte, dass die lokalen Störungen $\varphi(m\bullet - (i-1))$ disjunkte Träger für verschiedene i besitzen. Da für $\beta \leqslant s$ gilt

$$\begin{aligned}\int \Big(D^\beta \sum_{i=1}^m \varepsilon_i \varphi(mx-(i-1))\Big)^2 dx &\leqslant \sum_{i=1}^m \varepsilon_i \int m^{2s}\varphi^{(\beta)}(mx-(i-1))^2 dx \\ &\leqslant m^{2s}\|\varphi^{(\beta)}\|_{L^2}^2,\end{aligned}$$

erhalten wir $\|f_\varepsilon\|_s \leqslant R$ für $\gamma = \gamma_0 m^{-s}$ mit $\gamma_0 > 0$ hinreichend klein. Wir betrachten nun als Alternativen die Dichten $f_j := f_{\varepsilon^{(j)}}$, $j = 0, \dots, J$, mit den $\varepsilon^{(j)}$ und $J \geqslant 2^{m/8}$ aus dem Lemma von Varshamov-Gilbert. Dann gilt für den Abstand zwischen zwei Dichten $f_k \neq f_\ell$

$$\int_{[0,1]} (f_k(x) - f_\ell(x))^2 dx = \gamma^2 \sum_{i:\varepsilon_i^{(k)} \neq \varepsilon_i^{(\ell)}} \int \varphi(mx-(i-1))^2 dx \geqslant \gamma^2 \|\varphi\|_{L^2}^2/8.$$

Um Satz 4.14 anwenden zu können, schätzen wir nun den Kullback-Leibler-Abstand zwischen den Dichten f_j und der Dichte $f_0 = f$ ab (d.h. zwischen den zugehörigen Verteilungen):

$$\begin{aligned}\mathrm{KL}(f_j \,|\, f_0) &= \int \log(f_j/f_0) f_j \\ &\leqslant \int f_j(f_j - f_0)/f_0 \\ &= \int (f_j - f_0)^2/f_0 \\ &= \gamma^2 \sum_{i=1}^m \varepsilon_i^{(j)} \int \frac{\varphi(mx-(i-1))^2}{f_0(x)}\, dx \\ &\leqslant \gamma^2 c_f^{-1} \|\varphi\|_{L^2}^2.\end{aligned}$$

Damit gilt für die Verteilungen bei n i.i.d. Beobachtungen

$$\mathrm{KL}(\mathbb{P}_j^{\otimes n} \mid \mathbb{P}_0^{\otimes n}) \leqslant nm^{-2s}\gamma_0^2 c_f^{-1}\|\varphi\|_{L^2}^2, \quad j = 1, \dots J.$$

Wegen $\log J \geqslant m\log(2)/8$ ergibt die Wahl $m = \lfloor m_0 n^{1/(2s+1)} \rfloor$ mit $m_0 > 0$ hinreichend klein

$$\delta := \frac{1}{J}\sum_{j=1}^{J} \mathrm{KL}(\mathbb{P}_j^{\otimes n} \mid \mathbb{P}_0^{\otimes n}) \leqslant \frac{\log(J)}{8}.$$

Wenden wir nun Satz 4.14 an, so erhalten wir

$$\inf_{\hat\vartheta_n} \sup_f \mathbb{E}_{f,n}[\|\hat\vartheta_n - f\|_{L^2}^2] \geqslant \gamma^2\|\varphi\|_{L^2}^2/65 = \lfloor m_0 n^{1/(2s+1)} \rfloor^{-2s}\gamma_0^2\|\varphi\|_{L^2}^2/65,$$

was es im Fall $d = 1$ zu beweisen galt.

Für $d \geqslant 2$ und $x \in \mathbb{R}^d$ betrachte die Dichten

$$f_\varepsilon(x) = f(x) + \gamma \sum_{i_1,\dots,i_d=1}^{m} \varepsilon_i \varphi(mx - (i - \mathbf{1})), \quad i = (i_1, \dots i_d),$$

mit $\varepsilon \in \{0,1\}^{m^d}$ und $\mathbf{1} = (1, \dots, 1)$. Alle weiteren Abschätzungen bleiben gleich, wir gewinnen jedoch aus $J \geqslant 2^{m^d}/8$ die Möglichkeit $m = \lfloor m_0 n^{1/(2s+d)} \rfloor$ zu wählen, so dass immer noch $\delta \leqslant \frac{\log(J)}{8}$ gilt. Dies führt in denselben Schritten auf die untere Schranke der Ordnung $n^{-2s/(2s+d)}$.

Die Optimalität ergibt sich aus Korollar 2.25, wo sogar für den MISE auf ganz $\mathbb{R}^d$ die obere Schranke $\mathcal{O}(n^{-2s/(2s+d)})$ hergeleitet wurde. □

▶ Übung Für das gleichmäßige Risiko auf $[0,1]^d$ bei Schätzern für Dichten in α-Hölderklassen kann man die Minimaxrate $(n/\log n)^{-\alpha/(2\alpha+d)}$ nachweisen, man verliert also einen logarithmischen Faktor und muss die Regularität im Hölder- anstatt im Sobolevsinn messen. Weitere interessante Fragestellungen zum Beweis unterer Schranken werden von Tsybakov (2004) am Regressionsmodell erörtert.

5 Wahl des Glättungsparameters

5.1 Unverzerrte Risikoschätzung und Block-Stein-Schätzer

Um die wesentlichen Punkte herauszuarbeiten, werden wir hier im Modell des Signals in weißem Rauschen arbeiten:

$$dY_t = f(t)\,dt + \frac{\sigma}{\sqrt{n}}\,dW_t, \quad t \in [0,1].$$

Wie oben gesehen, erhalten wir durch Wahl einer Orthonormalbasis $(\varphi_k)_{k\geqslant 1}$ in $L^2([0,1])$ das äquivalente Folgenraummodell

$$y_k := \int_0^1 \varphi_k(t)\, dY_t = f_k + \frac{\sigma}{\sqrt{n}}\zeta_k, \quad k \geqslant 1,$$

mit wahren Koeffizienten $f_k := \langle f, \varphi_k\rangle$ und Rauschtermen $\zeta_k \sim N(0,1)$ i.i.d. Im Regressionsmodell $Y_i = f(i/n) + \varepsilon_i$ mit $\varepsilon_i \sim N(0,\sigma^2)$ würde man in analoger Weise mit $\tilde{y}_k := \frac{1}{n}\sum_{i=1}^n Y_i \varphi_k(i/n)$ arbeiten.

Ziel dieses Kapitels ist es, ein daten-getriebenes Glättungsverfahren zu finden, das unabhängig von der a priori-Kenntnis der Glattheit der unbekannten Funktion f auskommt, was auf sogenannte *adaptive Schätzer* (*adaptive, data-driven or unsupervised estimators*) von f führt.

Im Weiteren werden wir uns lineare, koeffizientenweise Schätzer der Form

$$\hat{f}_\lambda = \sum_{k\geqslant 1} \hat{f}_k \varphi_k \text{ mit } \hat{f}_k := \lambda_k y_k$$

anschauen, wobei die Folge $\lambda = (\lambda_k)_{k\geqslant 1}$ reellwertige *Gewichte* (oder *Filter*) angibt. Als mittleren quadratischen Fehler erhalten wir mit Bias-Varianz-Zerlegung

$$MISE(\lambda, f) := \mathbb{E}_f[\|\hat{f}_\lambda - f\|_{L^2}^2] = \sum_{k\geqslant 1}\Big((1-\lambda_k)^2 f_k^2 + \frac{\sigma^2}{n}\lambda_k^2\Big),$$

so dass wir $\lambda \in \ell^2$ voraussetzen wollen und im Allgemeinen auch $\lambda_k \in [0,1]$ betrachten werden (andere Wahlen sind offenbar suboptimal). Theoretisch optimal in einer Klasse $\Lambda \subseteq \ell^2$ von Gewichten wäre die Wahl

$$\lambda^{\text{Orakel}} := \operatorname{argmin}_{\lambda\in\Lambda} MISE(\lambda, f).$$

Da f unbekannt ist, ist diese *Orakelwahl* dem Statistiker nicht gestattet, sie dient jedoch als Messlatte und Vergleichskriterium für eine empirische Wahl.

Naheliegende Idee ist es, den unbekannten Wert $MISE(\lambda, f)$ zu schätzen. Der quadratische Verlust ist gerade

$$\|\hat{f}_\lambda - f\|_{L^2}^2 = \sum_{k\geqslant 1}\Big(\lambda_k^2 y_k^2 - 2\lambda_k y_k f_k + f_k^2\Big).$$

Nun hängt der Summand f_k^2 nicht von λ ab, und wir können $y_k f_k$ durch $y_k^2 - \frac{\sigma^2}{n}$ erwartungstreu (unverzerrt) schätzen:

$$\mathbb{E}_f[y_k^2 - \tfrac{\sigma^2}{n} - y_k f_k] = \mathbb{E}[f_k \tfrac{\sigma}{\sqrt{n}}\zeta_k + \tfrac{\sigma^2}{n}(\zeta_k^2 - 1)] = 0.$$

Setze daher

$$\mathcal{I}(\lambda) := \sum_{k\geqslant 1}\Big(\lambda_k^2 y_k^2 - 2\lambda_k(y_k^2 - \tfrac{\sigma^2}{n})\Big),$$

so dass $\mathbb{E}_f[\mathcal{I}(\lambda)] = MISE(\lambda, f) - \|f\|_{L^2}^2$ gilt und $\mathcal{I}(\lambda)$ ein *unverzerrter Schätzer des Risikos* (*unbiased risk estimator*) bis auf den von λ unabhängigen Term $\|f\|_{L^2}^2$ darstellt. Wir betrachten also

$$\tilde{\lambda} := \operatorname{argmin}_{\lambda \in \Lambda} \mathcal{I}(\lambda), \quad \tilde{f} := \hat{f}_{\tilde{\lambda}}.$$

Beachte, dass $\tilde{\lambda}$ als Minimalstelle eines zufälligen Kriteriums bereits schwierig zu analysieren ist und dass $\tilde{f}$ wegen der stochastischen Abhängigkeit zwischen den (y_k) und $\tilde{\lambda}$ noch komplexer ist.

5.1 Beispiele.

(a) Projektionsschätzer auf $S_K := \operatorname{span}(\varphi_k,\, k = 1, \dots, K)$: wir betrachten

$$\Lambda_{proj} := \{\lambda \in \ell^2 \,|\, \lambda_k = \mathbf{1}(k \leqslant K), \quad K \in \{1, 2, \dots, K_{max}\}\}.$$

Anstatt mit $\lambda \in \Lambda_{proj}$ indizieren wir mit K die entsprechenden Größen und erhalten

$$\hat{f}_{k,K} := y_k \mathbf{1}(k \leqslant K), \quad \mathcal{I}(K) = \sum_{k=1}^{K} (y_k^2 - 2(y_k^2 - \tfrac{\sigma^2}{n})) = 2K\tfrac{\sigma^2}{n} - \sum_{k=1}^{K} y_k^2.$$

Das Minimum von $\mathcal{I}$ wird angenommen bei

$$\tilde{K} = \operatorname{argmin}_{1 \leqslant K \leqslant K_{max}} \Big(\sum_{k=1}^{K_{max}} (y_k - \hat{f}_{k,K})^2 + 2K\tfrac{\sigma^2}{n} \Big).$$

Beachte, dass $\tilde{K}$ damit den empirischen Verlust des Schätzers (*RSS: residual sum of squares*) plus einen Strafterm minimiert, der gleich zweimal der Varianz ist. Dies entspricht *Mallows C_p-Krierium* für die Modellwahl bei linearen Modellen.

(b) Endlich-dimensionales Modell: wir betrachten für bekanntes $d \in \mathbb{N}$

$$\Lambda_{const} := \{\lambda \in \ell^2 \,|\, \lambda_k = t\mathbf{1}(1 \leqslant k \leqslant d), \quad t \in [0, 1]\}.$$

Der Einfachheit halber nehmen wir hier auch $f = \sum_{k=1}^{d} f_k \varphi_k$ an, so dass in der Tat ein endlich-dimensionales Problem vorliegt. Mit Indizierung durch t ergibt sich

$$\hat{f}_{k,t} := t y_k, \quad \mathcal{I}(t) = \sum_{k=1}^{d} \big(t^2 y_k^2 - 2t(y_k^2 - \tfrac{\sigma^2}{n})\big) = (t^2 - 2t)|y|^2 + 2td\tfrac{\sigma^2}{n},$$

wobei wir $|x|^2 := \sum_{k=1}^{d} x_k^2$ für $x \in \{y, f\}$ setzen und im folgenden mit Vektornotation arbeiten (beachte $|f|^2 = \|f\|_{L^2}^2$). Wir erhalten durch Minimerung über $t \in [0, 1]$

$$\tilde{t} := \Big(1 - \frac{d\sigma^2}{n|y|^2}\Big)_+ \text{ mit } A_+ := \max(A, 0).$$

Dies liefert den sogenannten *Stein-Schätzer* mit positivem Anteil

$$\tilde{f} = \hat{f}_{\tilde{t}} = \Big(1 - \frac{d\sigma^2}{n|y|^2}\Big)_+ y =: \hat{f}_{S+}.$$

Aus der mathematischen Statistik ist der entsprechende James-Stein-Schätzer bekannt:

$$\hat{f}_{JS+} = \Big(1 - \frac{(d-2)\sigma^2}{n|y|^2}\Big)_+ y.$$

Mit dem Steinschen Lemma beweist man wiederum (Tsybakov 2004, Lemma 3.10):

$$\forall\, d \geqslant 1 :\ \mathbb{E}_f\left[|\hat{f}_{S+} - f|^2\right] \leqslant \frac{d\sigma^2|f|^2}{d\sigma^2 + n|f|^2} + 4\frac{\sigma^2}{n}.$$

Das Überraschende ist nicht nur, dass wir eine solche Schranke konkret angeben können, sondern auch dass diese Schranke sehr nah am Orakelrisiko liegt:

$$\min_{t\in[0,1]} \mathbb{E}_f\left[|\hat{f}_t - f|^2\right] = \min_{t\in[0,1]} \sum_{k=1}^{d} \left((1-t)^2 f_k^2 + \tfrac{\sigma^2}{n} t^2\right) = \frac{d\sigma^2|f|^2}{d\sigma^2 + n|f|^2},$$

wobei das Minimum bei der Orakelwahl $t^{Orakel} := \frac{n|f|^2}{d\sigma^2+n|f|^2}$ angenommen wird. Damit erhalten wir folgende *Orakelungleichung* für den endlich-dimensionalen Fall:

$$\mathbb{E}_f[|\tilde{f} - f|^2] \leqslant \underbrace{\min_{t\in[0,1]} \mathbb{E}_f[|\hat{f}_t - f|^2]}_{\text{Orakelrisiko}} + \underbrace{4\frac{\sigma^2}{n}}_{\text{Kosten durch Risikoschätzung}}.$$

Beachte, dass diese Orakelungleichung explizit (nicht-asymptotisch mit einfachen Konstanten) ist und für jedes beliebige $f = \sum_{k=1}^{d} f_k \varphi_k$ gilt.

(c) Blockweise konstante Gewichte: betrachte für eine Partition $\{1, 2, \ldots, K_{max}\} = \bigcup_{j=1}^{J} B_j$

$$\Lambda_{Block} = \Big\{\lambda \in \ell^2 \,|\, \lambda_k = \sum_{j=1}^{J} t_j \mathbf{1}(k \in B_j), \quad 0 \leqslant t_j \leqslant 1, j = 1, \ldots, J\Big\}.$$

Damit lassen sich nun recht allgemeine Gewichte approximieren (s.u.) und gleichzeitig das endlich-dimensionale Resultat aus (b) verwenden. Wir erhalten einen blockweise konstanten Koeffizientenschätzer $\tilde{f} = \sum_{k=1}^{K_{max}} \tilde{f}_k \varphi_k$ mit

$$\tilde{f}_k := \sum_{j=1}^{J} \tilde{\lambda}_{(j)} \mathbf{1}(k \in B_j) y_k, \text{ wobei } \tilde{\lambda}_{(j)} := \Big(1 - \frac{|B_j|\sigma^2}{n|y|^2_{(j)}}\Big)_+$$

mit der Kardinalität $|B_j|$ von B_j und $|y|^2_{(j)} = \sum_{k\in B_j} y_k^2$. Wir nennen $\tilde f$ *Block-Stein-Schätzer*. Wie zuvor erhalten wir eine Orakelungleichung

$$\mathbb{E}_f[|\tilde f - f|^2] \leqslant \min_{\lambda\in\Lambda_{Block}} \mathbb{E}_f[|\hat f_\lambda - f|^2] + 4J\frac{\sigma^2}{n}.$$

5.2 Lemma. *Betrachte Blöcke $B_j = \{b_{j-1}+1,\dots,b_j\}$, $j = 1,\dots,J$, mit $0 = b_0 < b_1 < \cdots < b_J = K_{max}$ sowie*

$$\Lambda_{Block} = \Big\{\lambda\in\ell^2 \,|\, \lambda_k = \sum_{j=1}^J t_j \mathbf{1}(k\in B_j),\quad 0\leqslant t_j\leqslant 1, j = 1,\dots,J\Big\},$$
$$\Lambda_{mon} = \Big\{\lambda\in\ell^2 \,|\, \lambda_k\in[0,1],\, \lambda_{k+1}\leqslant\lambda_k,\, k\geqslant 1, \lambda_{K_{max}} = 0\Big\}.$$

Dann gilt für alle $f\in L^2$:

$$\min_{\lambda\in\Lambda_{Block}} MISE(\lambda, f) \leqslant \Big(1\vee \max_{1\leqslant j\leqslant J-1}\frac{|B_{j+1}|}{|B_j|}\Big)\min_{\lambda\in\Lambda_{mon}} MISE(f,\lambda) + |B_1|\frac{\sigma^2}{n}.$$

Beweis. Zu $\lambda\in\Lambda_{mon}$ wähle $\bar\lambda\in\Lambda_{Block}$ mit $\bar\lambda_k = \sum_{j=1}^J \bar\lambda_{(j)}\mathbf{1}(k\in B_j)$ und $\bar\lambda_{(j)} := \max_{k\in B_j}\lambda_k$. Dann gilt stets $\bar\lambda_k\geqslant\lambda_k$ und somit

$$MISE(\bar\lambda, f) = \sum_{k\geqslant 1}\Big((1-\bar\lambda_k)^2 f_k^2 + \frac{\sigma^2}{n}\bar\lambda_k^2\Big) \leqslant \sum_{k\geqslant 1}\Big((1-\lambda_k)^2 f_k^2 + \frac{\sigma^2}{n}\bar\lambda_k^2\Big).$$

Es genügt also, folgendes zu zeigen:

$$\sum_{k=1}^{K_{max}}\bar\lambda_k^2 \leqslant \max_{1\leqslant j\leqslant J-1}\frac{|B_{j+1}|}{|B_j|}\sum_{k=1}^{K_{max}}\lambda_k^2 + |B_1|.$$

Dies folgt durch Indexverschiebung in den Blöcken wegen $\bar\lambda_{(j)} = \lambda_{b_{j-1}+1}\leqslant\lambda_k$ für alle $k\in B_{j-1}$:

$$\begin{aligned}
\sum_{k=1}^{K_{max}}\bar\lambda_k^2 &\leqslant |B_1| + \sum_{j=2}^J |B_j|\bar\lambda_{(j)}^2\\
&\leqslant |B_1| + \max_{1\leqslant j\leqslant J-1}\frac{|B_{j+1}|}{|B_j|}\sum_{j=2}^J |B_{j-1}|\bar\lambda_{(j)}^2\\
&\leqslant |B_1| + \max_{1\leqslant j\leqslant J-1}\frac{|B_{j+1}|}{|B_j|}\sum_{j=2}^J\sum_{k\in B_{j-1}}\lambda_k^2\\
&= |B_1| + \max_{1\leqslant j\leqslant J-1}\frac{|B_{j+1}|}{|B_j|}\sum_{k=1}^{K_{max}}\lambda_k^2.
\end{aligned}$$

□

Wir erhalten also insgesamt folgendes Resultat.

5.3 Satz. *Der Block-Stein-Schätzer $\tilde{f}$ zu den Blöcken $B_j = \{b_{j-1} + 1, \dots, b_j\}$, $j = 1, \dots, J$, mit $0 = b_0 < b_1 < \cdots < b_J = K_{max}$ erfüllt die Orakelungleichung bezüglich monotoner Gewichte*

$$\mathbb{E}_f[|\tilde{f} - f|^2] \leqslant \max_{1 \leqslant j \leqslant J-1} \frac{|B_{j+1}|}{|B_j|} \min_{\lambda \in \Lambda_{mon}} MISE(f, \lambda) + (|B_1| + 4J)\frac{\sigma^2}{n}.$$

5.4 Beispiele.

(a) Dyadische Blöcke $b_j = 2^j$, $j = 1, \dots, J$, ergeben die Orakelungleichung

$$\min_{\lambda \in \Lambda_{Block}} MISE(\lambda, f) \leqslant 2 \min_{\lambda \in \Lambda_{mon}} MISE(f, \lambda) + (2 + 4J)\frac{\sigma^2}{n}.$$

(b) Für *schwach geometrische Blöcke* mit $|B_1| = \lfloor \rho \rfloor$, $|B_{j+1}| = \lfloor |B_j|(1 + \frac{1}{\rho}) \rfloor$, $j = 1, \dots, J-1$, mit $\rho := \log(n/\sigma^2)$ gilt

$$\frac{|B_{j+1}|}{|B_j|} \leqslant (1 + \tfrac{1}{\rho})$$

sowie mit $J = \lfloor \rho^2 \rfloor$ und $n/\sigma^2 \to \infty$ ($A \sim B$ bedeute $A = O(B)$ und $B = O(A)$):

$$\begin{aligned} K_{max} &= \sum_{j=1}^{J} |B_j| \sim \rho \sum_{j=1}^{J} (1 + \tfrac{1}{\rho})^{j-1} \\ &= \rho^2\big((1 + \tfrac{1}{\rho})^J - 1\big) = \rho^2\Big(\exp\big(\rho^2 \log(1 + \tfrac{1}{\rho})\big) - 1\Big) \\ &\sim \rho^2 e^{\rho} \sim \frac{n}{\sigma^2} \log^2(n/\sigma^2). \end{aligned}$$

Dies liefert eine *asymptotisch exakte Orakelungleichung* in dem Sinne, dass der Faktor vor dem Orakelrisiko gerade $1 + o(1)$ für $n \to \infty$ ist.

Das letzte Beispiel liefert unmittelbar folgende Korollare, wobei K_{max} in Λ_{mon} wie in Λ_{Block} gewählt sei.

5.5 Korollar. *Der Block-Stein-Schätzer $\tilde{f}$ mit schwach geometrischen Blöcken erfüllt die für $n \to \infty$ asymptotisch exakte Orakelungleichung bezüglich monotoner Gewichte*

$$\mathbb{E}_f[|\tilde{f} - f|^2] \leqslant \Big(1 + \frac{1}{\log(n/\sigma^2)}\Big) \min_{\lambda \in \Lambda_{mon}} MISE(f, \lambda) + \frac{\sigma^2(\log(n/\sigma^2) + 4\log^2(n/\sigma^2))}{n}.$$

5.6 Korollar. *Es gelte $\Lambda \subseteq \Lambda_{mon}$ und $f \in L^2([0,1])$ besitze die Eigenschaft*

$$\lim_{n\to\infty} \frac{\min_{\lambda\in\Lambda} MISE_n(\lambda, f)}{\frac{\sigma^2}{n}\log^2(n/\sigma^2)} = \infty.$$

Dann folgt für den Block-Stein-Schätzer $\tilde{f}_n$ mit schwach geometrischen Blöcken

$$\limsup_{n\to\infty} \frac{\mathbb{E}_f[\|\tilde{f}_n - f\|_{L^2}^2]}{\min_{\lambda\in\Lambda} MISE_n(\lambda, f)} \leqslant 1.$$

Im Kontext der nichtparametrischen Schätzung unter Regularitätsannahmen an f folgt, dass der Block-Stein-Schätzer in der Tat ein adaptiver Schätzer ist, der die Minimaxraten für Glattheit $s > 0$ erreicht, ohne die Kenntnis von s vorauszusetzen.

Exemplarisch betrachten wir dazu die Familie der periodischen Sobolevräume $H^s_{per}([0,1])$, $s > 0$, aus Definition 3.31. Die Projektionsschätzer $\hat{f}_{n,K}$ auf $S_K = \operatorname{span}(\varphi_k, |k| \leqslant K)$ mit der Fourierbasis $(\varphi_k)_{k\geqslant 1}$ besitzen dann einen Bias kleiner als LK^{-s} für $f \in H^s_{per}$ mit $\|f\|_s \leqslant L$, und die ratenoptimale Wahl $K_s \sim (n/\sigma^2)^{1/(2s+1)}$ liefert den MISE $C_{s,L}(n/\sigma^2)^{-2s/(2s+1)}$ mit einer Konstanten $C_{s,L} > 0$, was auch unter periodischen Randbedingungen die optimale Minimaxrate darstellt. Die Fourierbasis wird mit $k \in \mathbb{Z}$ indiziert, aber alle Resultate bleiben sinngemäß erhalten , wenn die Indizes $k, k' \in \mathbb{Z}$ gemäß $|k| \leqslant |k'|$ geordnet werden. Beachte nun $\Lambda_{proj} \subseteq \Lambda_{mon}$ sowie, dass für jedes $s > 0$ der Dimensionsparameter K_s von kleinerer Größenordnung ist als n/σ^2 und die Minimaxrate polynomiell langsamer als $\frac{\sigma^2}{n}$ abklingt. Daher folgt aus den bisherigen Ergebnissen, dass der Block-Stein-Schätzer bei dyadischen Blöcken mit der Wahl $2^J = K_{max} \geqslant n/\sigma^2$ oder bei obigen schwach geometrischen Blöcken ein an die gesamte Familie $(H^s_{per}([0,1]))_{s>0}$ adaptiver ratenoptimaler Schätzer ist.

5.7 Satz. *Der Block-Stein-Schätzer $\tilde{f}_n$ bezüglich der Fourierbasis mit dyadischen Blöcken und $J = \lceil \log_2(n/\sigma^2) \rceil$ oder mit schwach geometrischen Blöcken ist adaptiv an die gesamte Familie $(H^s_{per}([0,1]))_{s>0}$ für $n \to \infty$:*

$$\forall s > 0, \forall f \in H^s_{per}([0,1]) : \mathbb{E}_f[\|\tilde{f}_n - f\|_{L^2}^2] = \mathcal{O}(n^{-2s/(2s+1)}).$$

5.8 Bemerkung. Die Konstante auf der rechten Seite hängt von s und $\|f\|_s$ ab und ist je nach Wahl der Blöcke durch $2C_{s,\|f\|_s}$ bzw. $C_{s,\|f\|_s}$ beschränkt. Da monotone Gewichte viel allgemeiner sind als Projektionsgewichte, ist die Konstante sogar bedeutend kleiner. In der Tat wird mit schwach geometrischen Blöcken sogar die asymptotisch minimax-optimale Konstante über $\|f\|_s$ erreicht:

$$\forall s, L > 0 : \lim_{n\to\infty} \sup_{\|f\|_s \leqslant L} \mathbb{E}_f[\|\tilde{f}_n - f\|_{L^2}^2](n/\sigma^2)^{2s/(2s+1)} = P_{s,L},$$

wobei $P_{s,L} = (L^2(2s+1))^{1/(2s+1)}\Big(\frac{s}{s+1}\Big)^{2s/(2s+1)}$ *Pinskerkonstante* heißt und diese optimal im Minimaxsinn ist, siehe Tsybakov (2004) für Details zum Pinskerschätzer und exakten Minimaxkonstanten.

Auf die Übertragung der Ergebnisse vom Modell des Signals im weißen Rauschen auf das Regressions- oder Dichteschätzproblem verzichten wir hier. ▶ Übung Im Regressionsmodell ist der Ansatz jedoch direkt zu übertragen, allerdings sind die Resultate wegen der Abhängigkeit vom Design komplexer.

5.2 Konzentrationsungleichungen

Im folgenden werden wir häufig Suprema über Familien von Zufallsvariablen abschätzen müssen. Dies ist im allgemeinen nur sehr grob möglich, Konzentrationsungleichungen sind jedoch ein wichtiges Hilfsmittel, um gute Abschätzungen zu erhalten. Man bedient sich der recht groben Abschätzung

$$\begin{aligned}\mathbb{P}\Big(\max_{k=1,\dots,N}(X_k - \mathbb{E}[X_k]) \geqslant \tau\Big) &= \mathbb{P}(\exists k \in \{1,\dots,N\} : X_k - \mathbb{E}[X_k] \geqslant \tau) \\ &\leqslant \sum_{k=1}^{N} \mathbb{P}(X_k - \mathbb{E}[X_k] \geqslant \tau), \qquad (5.1)\end{aligned}$$

hat jedoch eine starke Konzentration von X_k um den Erwartungswert der Form $\mathbb{P}(X_k - \mathbb{E}[X_k] \geqslant \kappa\sigma_k) \leqslant e^{-c\kappa}$, $c, \kappa > 0$, mit einem geeigneten Streuungsmaß σ_k von X_k. Dies führt auf

$$\mathbb{P}\Big(\max_{k=1,\dots,N}(X_k - \mathbb{E}[X_k]) \geqslant \tau\Big) \leqslant N \exp\Big(-c\tau / \max_k \sigma_k\Big).$$

Für $\tau = p\log(N)\max_k \sigma_k/c$ mit $p > 1$ erhält man dann die obere Schranke N^{1-p}. Damit ist $\max_{k=1,\dots,N}(X_k - \mathbb{E}[X_k])$ von der Größenordnung $\log(N)\max_k \sigma_k$. Wir bezahlen im wesentlichen nur den Faktor $\log(N)$, um das Maximum gleichmäßig abzuschätzen.

5.9 Satz (Hoeffding-Ungleichung (1963)). *Es sei $(M_n, \mathcal{F}_n)$ ein Martingal mit $M_0 = 0$. Falls für alle $n \geqslant 1$ positive Zahlen R_n existieren mit $|M_n - M_{n-1}| \leqslant R_n$ fast sicher, so gilt*

$$\mathbb{P}(|M_n| \geqslant \kappa) \leqslant 2\exp\Big(-\frac{\kappa^2}{2\sum_{i=1}^n R_i^2}\Big), \quad \kappa > 0.$$

Beweis. Da die Exponentialfunktion konvex ist, gilt für $\lambda > 0$, $|\delta| \leqslant R$

$$e^{\lambda\delta} \leqslant \tfrac{R-\delta}{2R}e^{-\lambda R} + \tfrac{R+\delta}{2R}e^{\lambda R}.$$

Angewendet auf die Zufallsvariable $M_n - M_{n-1}$, impliziert dies

$$\mathbb{E}[e^{\lambda(M_n - M_{n-1})} \,|\, \mathcal{F}_{n-1}] \leqslant (e^{-\lambda R_n} + e^{\lambda R_n})/2 < e^{\lambda^2 R_n^2/2}$$

(für die letzte Ungleichung betrachte z.B. die Potenzreihenentwicklung).

Mit der verallgemeinerten Markov-Ungleichung erhalten wir für $\lambda > 0$

$$\mathbb{P}(M_n \geqslant \kappa) \leqslant e^{-\lambda\kappa}\,\mathbb{E}[e^{\lambda M_n}].$$

Nach obiger Ungleichung gilt

$$\mathbb{E}[e^{\lambda M_n} \,|\, \mathcal{F}_{n-1}] = e^{\lambda M_{n-1}}\,\mathbb{E}[e^{\lambda(M_n - M_{n-1})} \,|\, \mathcal{F}_{n-1}] \leqslant e^{\lambda M_{n-1}} e^{\lambda^2 R_n^2/2}.$$

Iteriert man die bedingten Erwartungen, so ergibt sich

$$\mathbb{E}[e^{\lambda M_n}] \leqslant \prod_{i=1}^{n} e^{\lambda^2 R_i^2/2} = \exp\Big(\lambda^2 \sum_{i=1}^{n} R_i^2/2\Big).$$

Somit haben wir für beliebiges $\lambda > 0$

$$\mathbb{P}(M_n \geqslant \kappa) \leqslant \exp\Big(-\lambda\kappa + \lambda^2 \sum_{i=1}^{n} R_i^2/2\Big)$$

gezeigt, was für die optimale Wahl $\lambda = \kappa/\sum_{i=1}^n R_i$ auf

$$\mathbb{P}(M_n \leqslant \kappa) \leqslant \exp\Big(-\frac{\kappa^2}{2\sum_{i=1}^n R_i^2}\Big)$$

führt. Ein symmetrisches Argument für $-M_n$ liefert dieselbe Schranke, und die Behauptung folgt durch Addition beider Schranken. □

5.10 Beispiele.

(a) ▶ Übung Wir erhalten nicht-asymptotische Abschätzungen für große Abweichungen der Binomialverteilung, wenn wir eine Bernoulli-Kette $X_1, X_2, \ldots$ mit Erfolgswahrscheinlichkeit $p \in [0,1]$ betrachten und $S_n = \sum_{i=1}^n X_i$, $M_n = S_n - np$ mit $M_0 = 0$ setzen. Dann ist (M_n) ein Martingal bezüglich $\mathcal{F}_n := \sigma(X_1, \ldots, X_n)$ mit $|M_n - M_{n-1}| \leqslant \max(p, 1-p)$. Die Hoeffding-Ungleichung liefert daher

$$\begin{aligned}\mathbb{P}\left(|S_n - np| \geqslant \kappa\sqrt{np(1-p)}\right) &= \mathbb{P}\left(|M_n| \geqslant \kappa\sqrt{np(1-p)}\right)\\ &\leqslant 2\exp(-\kappa^2 \min(p,1-p)/(2\max(p,1-p))).\end{aligned}$$

Die standardisierte Binomialverteilung besitzt also exponentiell abfallende *tails* für $p \in (0,1)$ mit größtem Exponenten bei $p = 1/2$.

(b) Die empirische charakteristische Funktion $\hat{\varphi}_n(u) = \frac{1}{n}\sum_{k=1}^n e^{iX_k u}$ für (X_k) i.i.d. konvergiert für $n \to \infty$ und festes $u \in \mathbb{R}$ fast sicher gegen ihren Erwartungswert $\mathbb{E}[\hat{\varphi}_n(u)] = \varphi(u)$. Können wir sogar gleichmäßige Konvergenz auf einem Intervall $[A, B]$ beweisen und eine Konvergenzrate herleiten?

Wir betrachten das Einheitsintervall und $M_n(u) = \sum_{k=1}^n (\cos(X_k u) - \mathbb{E}[\cos(X_k u)])$, $M_0 = 0$. Dann ist M_n als Summe unabhängiger, zentrierter Zufallsvariablen ein Martingal. Wegen $|M_n(u) - M_{n-1}(u)| \leqslant 2$ liefert die Hoeffding-Ungleichung

$$\mathbb{P}(|M_n(u)| \geqslant \kappa/2) \leqslant 2\exp\Big(-\frac{(\kappa/2)^2}{8n}\Big).$$

Wähle nun für $J = J(n)$ äquidistante Punkte $u_j = j/J$, $j = 1, \ldots, J$. Dann gilt

$$\mathbb{P}\Big(\max_{1\leqslant j\leqslant J} |M_n(u_j)| \geqslant \kappa/2\Big) \leqslant 2J\exp\Big(-\frac{(\kappa/2)^2}{8n}\Big).$$

Es gilt für beliebige $u, v \in \mathbb{R}$

$$|\cos(X_k u) - \cos(X_k v)| \leqslant |X_k||u - v|.$$

Wenn nun $\mathbb{E}[|X_k|]$ endlich ist, impliziert dies $|M_n(u) - M_n(v)| \leqslant \sum_{k=1}^n (|X_k| + \mathbb{E}[|X_k|])|u - v|$ und wegen $\max_{u\in[0,1]} \min_j |u - u_j| \leqslant J^{-1}$

$$\mathbb{P}\Big(\sup_{u\in[0,1]} |M_n(u)| \geqslant \kappa\Big) \leqslant \mathbb{P}\Big(\max_{j=1,\ldots,J} |M_n(u_j)| + \sum_{k=1}^n (|X_k| + \mathbb{E}[|X_k|])J^{-1} \geqslant \kappa\Big).$$

Eine Anwendung der Markov-Ungleichung ergibt

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{u\in[0,1]} |M_n(u)| \geqslant \kappa\Big) \\
&\leqslant \mathbb{P}\Big(\max_{1\leqslant j\leqslant J} |M_n(u_j)| \geqslant \kappa/2\Big) + \mathbb{P}\Big(\sum_{k=1}^n (|X_k| + \mathbb{E}[|X_k|]) \geqslant J\kappa/2\Big) \\
&\leqslant 2J\exp\Big(-\frac{(\kappa/2)^2}{8n}\Big) + (J\kappa/2)^{-1} \sum_{k=1}^n \mathbb{E}[|X_k| + \mathbb{E}[|X_k|]] \\
&= 2J\exp\Big(-\frac{\kappa^2}{32n}\Big) + 4nJ^{-1}\kappa^{-1}\,\mathbb{E}[|X_k|].
\end{aligned}$$

Die Wahl $J = \sqrt{n/\kappa}\exp(\kappa^2/64n)$ liefert die Größenordnung

$$\mathbb{P}(\sup_{u\in[0,1]} |M_n(u)| \geqslant \kappa) = \mathcal{O}\Big(\sqrt{n/\kappa}\exp\Big(-\frac{\kappa^2}{64n}\Big)\Big).$$

Für $C > \sqrt{32}$ und $n \to \infty$ erhalten wir

$$\mathbb{P}\Big(\sup_{u\in[0,1]} |M_n(u)| \geqslant C\sqrt{n\log n}\Big) = \mathcal{O}\Big(n^{1/2}\exp\Big(-\frac{C^2\log n}{64}\Big)\Big) = o(1).$$

Mit einer analogen Abschätzung für den Imaginärteil von $\hat{\varphi}_n$ ergibt sich für hinreichend großes C ($C > \sqrt{32}$ reicht)

$$\lim_{n\to\infty} \mathbb{P}\Big(\max_{u\in[0,1]} |\hat{\varphi}_n(u) - \varphi(u)| \geqslant C\sqrt{\log(n)/n}\Big) = 0.$$

Wir haben also gezeigt, dass die empirische charakteristische Funktion für $X_k \in L^1$ gleichmäßig auf einem kompakten Intervall gegen φ konvergiert (in Wahrscheinlichkeit) mit der Rate $(\log(n)/n)^{1/2}$. Mit der Theorie empirischer Prozesse kann man sogar die Rate $n^{-1/2}$ wie im Fall punktweiser Konvergenz herleiten.

Die Bedingung $|M_n - M_{n-1}| \leqslant R_n$ ist häufig sehr restriktiv und führt auf große Werte R_n. Der Einfluss der (R_n) wird in folgender Ungleichung abgeschwächt durch Einführung eines Varianzterms. Zunächst wird der i.i.d.-Fall behandelt.

5.11 Satz (Bernstein-Ungleichung (1923)). *Es seien $(X_i)_{i\geqslant 1}$ unabhängige, identisch verteilte Zufallsvariablen mit $\mathbb{E}[X_i] = 0$ und $S_n := \sum_{i=1}^n X_i$. Falls $|X_i| \leqslant R$ fast sicher gilt, so folgt*

$$\mathbb{P}(|S_n| \geqslant \kappa) \leqslant 2\exp\Big(-\sup_{\lambda>0}\Big(\lambda\kappa - \operatorname{Var}(S_n)\frac{e^{\lambda R} - 1 - \lambda R}{R^2}\Big)\Big), \quad \kappa > 0.$$

Insbesondere gilt

$$\mathbb{P}(|S_n| \geqslant \kappa) \leqslant 2\exp\Big(-\frac{\kappa^2}{4(\operatorname{Var}(S_n) + \kappa R)}\Big), \quad \kappa > 0.$$

Beweis. Wiederum verwenden wir die verallgemeinerte Markovungleichung für $\lambda > 0$ und erhalten mit der Unabhängigkeit der X_i und dem Satz von Fubini

$$\mathbb{P}(S_n \geqslant \kappa) \leqslant e^{-\lambda\kappa}\prod_{i=1}^n \mathbb{E}[e^{\lambda X_i}] = e^{-\lambda\kappa}\Big(\sum_{k=0}^\infty \frac{\lambda^k\,\mathbb{E}[X_i^k]}{k!}\Big)^n.$$

Wir benutzen nun $\mathbb{E}[X_i] = 0$ sowie $|\mathbb{E}[X_i^k]| \leqslant \mathbb{E}[X_i^2]R^{k-2} = R^{k-2}\operatorname{Var}(S_n)/n$ für $k \geqslant 2$, so dass wegen $(1+A)^n \leqslant e^{An}$

$$\begin{aligned}\mathbb{P}(S_n \geqslant \kappa) &\leqslant e^{-\lambda\kappa}\Big(1 + \frac{1}{n}\sum_{k=2}^\infty \frac{\lambda^k R^{k-2}\operatorname{Var}(S_n)}{k!}\Big)^n \\ &\leqslant \exp\Big(-\lambda\kappa + \operatorname{Var}(S_n)R^{-2}(e^{\lambda R} - 1 - \lambda R)\Big)\end{aligned}$$

gilt. Dies zusammen mit einer symmetrischen Abschätzung für $-S_n$ gibt die erste Ungleichung, die zweite folgt mit der Wahl $\lambda := \kappa/(2\operatorname{Var}(S_n) + 2\kappa R)$, weil dann $\lambda R \leqslant 1/2$ impliziert $e^{\lambda R} - 1 - \lambda R \leqslant \lambda^2 R^2$. Folglich gilt

$$-\lambda\kappa + \operatorname{Var}(S_n)R^{-2}(e^{\lambda R} - 1 - \lambda R) \leqslant -\frac{\kappa^2}{2(\operatorname{Var}(S_n) + 2\kappa R)} + \frac{\kappa^2\operatorname{Var}(S_n)}{4(\operatorname{Var}(S_n) + \kappa R)^2},$$

und eine triviale Abschätzung des zweiten Zählers durch $\kappa^2(\mathrm{Var}(S_n) + \kappa R)$ liefert die Behauptung. □

5.12 Beispiel. ▶ Übung Im Fall der Binomialverteilung erhalten wir für $T_n = \sum_{i=1}^n (Y_i - \mathbb{E}[Y_i])$ und eine Bernoullikette $(Y_i)_{i\geqslant 1}$ mit Erfolgswahrscheinlichkeit p:

$$|Y_i - \mathbb{E}[Y_i]| \leqslant \max(p, 1-p), \quad \mathrm{Var}(T_n) = np(1-p).$$

Also impliziert die Bernsteinungleichung $\mathbb{P}(|T_n| \geqslant \kappa) \leqslant 2\exp(-\kappa^2/(4(np(1-p) + \kappa\max(p,1-p))))$. Für $S_n = \sum_{i=1}^n Y_i \sim \mathrm{Bin}(n,p)$ folgt durch Skalierung

$$\mathbb{P}\left(|S_n - np| \geqslant \kappa\sqrt{np(1-p)}\right) \leqslant 2\exp\left(-\frac{\kappa^2}{4 + 4\max(p,1-p)\kappa(np(1-p))^{-1/2}}\right).$$

Dies bedeutet, dass für große n die *tails* der standardisierten Binomialverteilung fast wie $e^{-\kappa^2/4}$ abfallen. Dies ist eine Verbesserung der Abschätzung mittels Hoeffding-Ungleichung in den Fällen $p < 1/3$ oder $p > 2/3$.

Im Martingalfall wird die Varianz durch die quadratische Variation ersetzt, die jedoch stochastisch ist. Zur Erinnerung: für ein quadratisch-integrierbares Martingal (M_n) ist die quadratische Variation $\langle M\rangle_n$ der Kompensator in der Doob-Zerlegung des Submartingals M_n^2, insbesondere gilt

$$\langle M\rangle_n = \sum_{k=1}^n \mathbb{E}[(M_k - M_{k-1})^2 \,|\, \mathcal{F}_{k-1}].$$

5.13 Satz (Bernstein-Ungleichung für Martingale)**.** *Es sei $(M_n, \mathcal{F}_n)$ ein Martingal mit $M_0 = 0$. Dann gilt für beliebige $\kappa, \beta, \rho > 0$*

$$\mathbb{P}(|M_n|) \geqslant \kappa) \leqslant 2\,\mathbb{P}(\langle M\rangle_n > \beta) + 2\,\mathbb{P}\Big(\max_{k=1,\dots,n} |M_k - M_{k-1}| > \rho\Big) + 2\exp\Big(-\sup_{\lambda>0}\Big(\kappa\lambda - \beta\frac{e^{\lambda\rho} - 1 - \lambda\rho}{\rho^2}\Big)\Big).$$

Insbesondere gilt

$$\mathbb{P}(|M_n| \geqslant \kappa) \leqslant 2\,\mathbb{P}(\langle M\rangle_n > \beta) + 2\,\mathbb{P}\Big(\max_{k=1,\dots,n} |M_k - M_{k-1}| > \rho\Big) + 2\exp\Big(-\frac{\kappa^2}{4(\beta + \kappa\rho)}\Big).$$

5.14 Bemerkungen.

(a) Die Ungleichung kann noch verschärft werden, indem auf der linken Seite $|M_n|$ durch $\max_{k\leqslant n}|M_k|$ ersetzt wird, vergleiche Theorem VII.6 in Shiryaev (1995).

(b) Die klassische Bernstein-Ungleichung ergibt sich mit $M_n = S_n$, $\langle M\rangle_n = \mathrm{Var}(S_n)$ und $\beta = \mathrm{Var}(S_n)$, $\rho = R$.

Beweis. Wir schreiben $h_\rho(\lambda) := (e^{\lambda\rho} - 1 - \lambda\rho)/\rho^2$ sowie

$$Z_n(\lambda) := \exp(\lambda M_n - h_\rho(\lambda)\langle M\rangle_n)\mathbf{1}\Big(\max_{k=1,\dots,n}|M_k - M_{k-1}| \leqslant \rho\Big), \quad n \geqslant 1,$$

und $Z_0(\lambda) := 1$. Wiederum verwenden wir die verallgemeinerte Markovungleichung für $\lambda > 0$ und erhalten

$$\begin{aligned}\mathbb{P}\Big(M_n \geqslant \kappa, \langle M\rangle_n \leqslant \beta, \max_{k=1,\dots,n}|M_k - M_{k-1}| \leqslant \rho\Big)\\ \leqslant e^{-\lambda\kappa + h_\rho(\lambda)\beta}\,\mathbb{E}\Big[Z_n(\lambda)\Big].\end{aligned}$$

Wie bei der Hoeffding-Ungleichung betrachten wir bedingte Erwartungen und verwenden $Z_n(\lambda) = Z_{n-1}(\lambda)R_n(\lambda)$ mit

$$R_n(\lambda) := \exp\Big(\lambda(M_n - M_{n-1}) - h_\rho(\lambda)(\langle M\rangle_n - \langle M\rangle_{n-1})\Big)\mathbf{1}(|M_n - M_{n-1}| \leqslant \rho).$$

Wir erhalten also $\mathbb{E}[Z_n(\lambda)\,|\,\mathcal{F}_{n-1}] = Z_{n-1}(\lambda)\,\mathbb{E}[R_n(\lambda)\,|\,\mathcal{F}_{n-1}]$. Beachte nun, dass für $x \in \mathbb{R}$ gilt

$$\begin{aligned}e^{\lambda x}\mathbf{1}(|x| \leqslant \rho) &= \Big(1 + \lambda x + x^2\sum_{k=2}^{\infty}\frac{\lambda^k x^{k-2}}{k!}\Big)\mathbf{1}(|x| \leqslant \rho)\\ &\leqslant \Big(1 + \lambda x + x^2\sum_{k=2}^{\infty}\frac{\lambda^k \rho^{k-2}}{k!}\Big)\mathbf{1}(|x| \leqslant \rho)\\ &= \Big(1 + \lambda x + x^2 h_\rho(\lambda)\Big)\mathbf{1}(|x| \leqslant \rho)\\ &\leqslant 1 + \lambda x + x^2 h_\rho(\lambda),\end{aligned}$$

wobei im letzten Schritt $h_\rho(\lambda) \geqslant \lambda^2/2$ und $1 + \lambda x + \lambda^2 x^2/2 \geqslant (1 + \lambda x/2)^2 \geqslant 0$ benutzt wurde. Wir erhalten

$$\begin{aligned}&\mathbb{E}[R_n(\lambda)\,|\,\mathcal{F}_{n-1}]\\ &\leqslant e^{-h_\rho(\lambda)(\langle M\rangle_n - \langle M\rangle_{n-1})}\,\mathbb{E}[1 + \lambda(M_n - M_{n-1}) + h_\rho(\lambda)(M_n - M_{n-1})^2\,|\,\mathcal{F}_{n-1}]\\ &= e^{-h_\rho(\lambda)(\langle M\rangle_n - \langle M\rangle_{n-1})}(1 + h_\rho(\lambda)(\langle M\rangle_n - \langle M\rangle_{n-1}).\end{aligned}$$

Auf Grund der Ungleichung $(1 + A)e^{-A} \leqslant 1$ für $A \geqslant 0$ ist die letzte Zeile kleiner gleich Eins. Also gilt $\mathbb{E}[Z_n(\lambda)] \leqslant \mathbb{E}[Z_0(\lambda)] = 1$. Insgesamt haben wir gezeigt

$$\mathbb{P}\Big(M_n \geqslant \kappa,\, \langle M\rangle_n \leqslant \beta,\, \max_{k=1,\dots,n}|M_k - M_{k-1}| \leqslant \rho\Big) \leqslant \exp\Big(-\sup_{\lambda>0}(\lambda\kappa - \beta h_\rho(\lambda))\Big),$$

so dass mit

$$\begin{aligned}\mathbb{P}(M_n \geqslant \kappa) \leqslant \mathbb{P}(\langle M\rangle_n > \beta) + \mathbb{P}(\max_{k=1,\dots,n}|M_k - M_{k-1}| > \rho)\\ + \mathbb{P}\Big(M_n \geqslant \kappa,\, \langle M\rangle_n \leqslant \beta,\, \max_{k=1,\dots,n}|M_k - M_{k-1}| \leqslant \rho\Big)\end{aligned}$$

sowie einer symmetrischen Abschätzung für $-M$ die erste Ungleichung folgt.

Wählt man $\lambda := \kappa/(2\beta + 2\kappa\rho)$, so ist $h_\rho(\lambda) \leqslant \lambda^2$, und es ergibt sich die zweite Ungleichung aus

$$-\lambda\kappa + \beta h_\rho(\lambda) \leqslant -\lambda\kappa + \beta\lambda^2 \leqslant -\frac{\kappa^2}{2(\beta + 2\kappa\rho)} + \frac{\kappa^2}{4(\beta + \kappa\rho)}.$$

□

Schließlich geben wir noch eine gleichmäßige Form der Bernstein-Ungleichung im i.i.d.-Fall an, die für statistische Belange von großer Bedeutung ist. Für einen Beweis mittels Entropiemethoden und weitere Resultate in diese Richtung siehe Massart (2007).

5.15 Satz (Talagrand-Ungleichung (1996)). *Es sei $\mathcal{P}$ eine abzählbare Menge und für jedes $p \in \mathcal{P}$ seien $X_1^{(p)}, X_2^{(p)}, \ldots, X_n^{(p)}$ unabhängige, identisch verteilte Zufallsvariablen auf demselben Wahrscheinlichkeitsraum mit $\mathbb{E}[X_i^{(p)}] = 0$ und $|X_i^{(p)}| \leqslant R$ fast sicher. Setze $S_n^{(p)} := \sum_{i=1}^n X_i^{(p)}$ und $\sigma^2 := \sup_{p\in\mathcal{P}} \operatorname{Var}(S_n^{(p)})$. Dann gibt es für beliebiges $\varepsilon > 0$ Konstanten $c_1, c_2(\varepsilon) > 0$ mit*

$$\mathbb{P}\left(\sup_{p\in\mathcal{P}} |S_n^{(p)}| \geqslant (1+\varepsilon)\,\mathbb{E}\left[\sup_{p\in\mathcal{P}} |S_n^{(p)}|\right] + c_1\sigma\kappa + c_2(\varepsilon)R\kappa^2\right) \leqslant \exp(-\kappa^2/2), \quad \kappa > 0.$$

5.16 Bemerkung. Ein analoges Konzentrationsresultat gilt auch für die Abschätzung von $\sup_{p\in\mathcal{P}} |S_n^{(p)}|$ nach unten. Das Bemerkenswerte an der Talagrand-Ungleichung ist, dass die obere Schranke allein im Erwartungswert von der Komplexität der Zufallsvariablen in der Familie $\mathcal{P}$ abhängt, bei σ^2 wird das Supremum nur auf die deterministischen Varianzen angewendet.

5.3 Bandweitenwahl durch Kreuzvalidierung

Wir haben gesehen, dass die optimale Wahl der Bandweite eines Kernschätzers im Sinne der Risikominimierung in (2.1) nicht möglich ist, da das Risiko von der unbekannten, zu schätzenden Funktion abhängt. Ein natürlicher statistischer Zugang ist daher, das Risiko selbst zu schätzen und dann eine (datengetriebene) Bandweite zu wählen, die das geschätzte Risiko minimiert. Kreuzvalidierung (*cross validation*) bietet eine einfache und effektive Möglichkeit, den MISE unverzerrt zu schätzen (sogenannte *unbiased risk estimation*). Wir werden uns hier auf das Dichteschätzproblem und Kerndichteschätzer konzentrieren.

Betrachte den Verlust in $L^2(\mathbb{R}^d)$ (ISE: *integrated square error*) bei der Kernschätzung $\hat{f}_{n,h}$ von f in Abhängigkeit von der Bandweite $h > 0$:

$$\begin{aligned} ISE_n(h) &:= \int_{\mathbb{R}^d} (\hat{f}_{n,h}(x) - f(x))^2 dx \\ &= \int_{\mathbb{R}^d} \hat{f}_{n,h}(x)^2 dx - 2\int_{\mathbb{R}^d} \hat{f}_{n,h}(x) f(x)\, dx + \int_{\mathbb{R}^d} f(x)^2 dx. \end{aligned}$$

Beachte, dass $ISE_n(h)$ eine zufällige Größe ist, deren Erwartungswert gerade $MISE_n(h) := R_{\mathbb{R}^d}(\hat{f}_{n,h}, f)$ ist. Jedes (zufällige) $\hat{h}$, das $ISE_n(h)$ minimiert, führt sogar zu einem besseren MISE als h_n^*, der Minimierer von $MISE_n(h)$ (klar?). Nun hängt $ISE_n(h)$ immer noch von der unbekannten Dichte f ab. Der letzte Summand ist jedoch unabhängig von h und spielt somit beim Minimieren keine Rolle. Weil $\hat{f}_{n,h}$ bekannt ist, müssen wir also nur den zweiten Summanden

$$I_n(h) := \int_{\mathbb{R}^d} \hat{f}_{n,h}(x) f(x)\, dx = \mathbb{E}_f[\hat{f}_{n,h}(X)]$$

schätzen (dabei wird die Erwartung nur bezüglich X genommen, einer unabhängigen Zufallsvariable, die wie X_i verteilt ist). Der erste natürliche Ansatz ist, den Erwartungswert durch das arithmetische Mittel zu ersetzen:

$$\mathbb{E}_f[\hat{f}_{n,h}(X_i)] \overset{?}{\approx} \frac{1}{n} \sum_{i=1}^{n} \hat{f}_{n,h}(X_i).$$

Dies ist jedoch kritisch, weil $\hat{f}_{n,h}$ ja selbst zufällig und abhängig von den X_i ist. Beachte, dass deshalb bereits die Erwartungswerte beider Seiten nicht übereinstimmen dürften.▶ Übung Ein solches Kriterium führt zum starken *Unterglätten*, weil nur der Fehler in der Stichprobe selbst verringert wird (*in-sample performance*), was naturgemäß für Dichten nahe am empirischen Maß der Fall ist.

Eine Möglichkeit, diese Abhängigkeiten zu vermeiden, ist es, die Stichprobe $(X_1, \ldots, X_n)$ in eine *Trainingsmenge* (*training set*) $(X_1, \ldots, X_m)$, $m < n$, und eine *Validierungsmenge* (*validation set*) $(X_{m+1}, \ldots, X_n)$ aufzuspalten (*sample splitting*). Man betrachtet nur Kerndichteschätzer $\hat{f}_{m,h}$ basierend auf den Trainingsdaten und bestimmt dessen Güte für verschiedene Bandweiten h anhand der Validierungsmenge. Im vorliegenden Fall würde man also die Approximation

$$\mathbb{E}_f[\hat{f}_{m,h}(X)] \overset{?}{\approx} \frac{1}{n-m} \sum_{i=m+1}^{n} \hat{f}_{m,h}(X_i)$$

anstreben. Nach dem Gesetz der großen Zahlen ist die Approximation für $n - m \to \infty$ konsistent. Ein offensichtlicher Nachteil dieses Verfahrens ist jedoch, dass wir nur $m < n$ Daten für die Schätzung verwenden und die Größe $n - m$ der Testmenge für eine gute Approximation hinreichend groß sein muss. Trotzdem liefert dieser Ansatz häufig gute Resultate.

Kreuzvalidierung (*cross validation*) beruht auf einer Verfeinerung der Idee der Stichprobenaufspaltung. Für $j = 1, \ldots, n$ führen wir die sogenannten *leave-one-out-Schätzer* ein:

$$\hat{f}_{n,h}^{(-j)}(x) := \frac{1}{n-1} \sum_{i \neq j} K_h(x - X_i).$$

Der Schätzer $\hat{f}_{n,h}^{(-j)}$ ist also der gewöhnliche Kernschätzer, aber basierend auf den Daten $\{X_1, \dots, X_n\} \setminus \{X_j\}$. Wir werden die Kreuzvalidierungsbandweite als Minimierer folgenden CV-Kriteriums gewinnen:

$$CV_n(h) := \int_{\mathbb{R}^d} \hat{f}_{n,h}(x)^2 dx - 2\hat{I}_n(h) \text{ mit } \hat{I}_n(h) := \frac{1}{n} \sum_{j=1}^{n} \hat{f}_{n,h}^{(-j)}(X_j).$$

Da in Theorie und Praxis eine Minimierung über alle Bandweite $h \in \mathbb{R}^+$ schwierig ist, beschränken wir uns zunächst auf endliche Teilmengen $\mathcal{H}_n \subseteq \mathbb{R}^+$, die von n abhängen dürfen und später spezifiziert werden, und definieren

$$\hat{h}_n := \operatorname{argmin}_{h \in \mathcal{H}_n} CV_n(h), \quad h_n := \operatorname{argmin}_{h \in \mathcal{H}_n} MISE_n(h).$$

Der adaptive Kernschätzer basierend auf Kreuzvalidierung ist dann

$$\hat{f}_n(x) := \hat{f}_{n,\hat{h}_n}(x).$$

Mit $CV_n(h)$ wird der MISE bis auf den unerheblichen Term $\int f^2$ unverzerrt geschätzt:

$$\begin{aligned}
&\mathbb{E}_f[CV_n(h)] = \int_{\mathbb{R}^d} \mathbb{E}_f[\hat{f}_{n,h}(x)^2]\, dx - 2\,\mathbb{E}_f[\hat{f}_{n,h}^{(-i)}(X_i)] \\
&= \int_{\mathbb{R}^d} \Big(K_h * f(x)^2 + \frac{1}{n}\big(K_h^2 * f(x) - K_h * f(x)^2\big) - 2K_h * f(x) f(x) \Big)\, dx.
\end{aligned}$$

Daher erwarten wir, dass $\hat{h}_n$ nahe bei h_n liegt, zumindest asymptotisch für $n \to \infty$. Ebenso sollte die Diskretisierung in $\mathcal{H}_n$ hinreichend fein sein, dass h_n und h_n^* asymptotisch zum selben Fehler führen. Ziel dieses Abschnitts ist folgender beeindruckende Satz von Stone (1984), dass Kreuzvalidierung asymptotisch für $n \to \infty$ zu demselben Fehler führt wie die optimale Orakel-Bandweite h_n^*, vergleiche auch die genauere Analyse in Hall and Marron (1987).

5.17 Satz. *Es sei $K \in L^1(\mathbb{R}^d)$ eine beschränkte Kernfunktion mit gleichmäßig Lipschitz-stetiger Fouriertransformierter:*

$$\exists L > 0 \, \forall u, u' \in \mathbb{R}^d : \; |\mathcal{F}K(u) - \mathcal{F}K(u')| \leqslant L|u - u'|.$$

Die Bandweitenmenge bei n Beobachtungen sei von der Form

$$\mathcal{H}_n = \{kn^{-d-1} \,|\, k \in \mathbb{N},\, n^{-1/d} \leqslant kn^{-d-1} \leqslant (\log n)^{-12/d}\}.$$

Sofern die Dichtefunktion f in einem Sobolevraum $H^s(\mathbb{R}^d)$ für irgendein $s > 0$ liegt, gilt für den kreuzvalidierten Kerndichteschätzer $\hat{f}_n$ von f

$$\lim_{n \to \infty} \frac{\|\hat{f}_n - f\|_{L^2}^2}{\inf_{h>0} MISE_n(h)} = 1 \text{ (stochastisch).}$$

Der Verlust von $\hat{f}_n$ ist also asymptotisch nicht größer als der Orakelfehler.

5.18 Bemerkung. Die exakte Form der Bandweitenmenge $\mathcal{H}_n$ ist relativ unerheblich. Wir werden im Beweis im wesentlichen nur verwenden, dass $\min \mathcal{H}_n \leqslant h_n^* \leqslant \max \mathcal{H}_n$ mit Orakelbandweite $h_n^* \sim n^{-1/(2s+d)}$ für alle $s > 0$ gilt, $\max \mathcal{H}_n$ logarithmisch fällt, die Diskretisierung hinreichend fein ist sowie die Kardinalität von $\mathcal{H}_n$ maximal polynomiell in n wächst.

Die Forderungen an den Kern sind recht schwach. Jeder Kern mit integrierbarer, gleichmäßig Lipschitz-stetiger Fouriertransformierter erfüllt sie (Riemann-Lebesgue-Lemma). Insbesondere werden sie auch vom Rechteckkern erfüllt, dessen Fouriertransformierte in $L^p(\mathbb{R})$ für $p > 1$, jedoch nicht in $L^1(\mathbb{R})$ liegt. ▶ Übung Im Beweis wird man sehen, dass die Lipschitzstetigkeit von $\mathcal{F}K$ sogar durch eine schwächere Hölderstetigkeit ersetzt werden kann, wenn man $\mathcal{H}_n$ gegebenenfalls etwas feiner wählt.

Man kann Funktionen $f \in L^2(\mathbb{R}^d)$ konstruieren, die in keinem $H^s(\mathbb{R}^d)$ liegen. Dies sind aber sehr artifizielle, äußerst irreguläre Funktionen, die in praxi nicht vorkommen dürften. Darüberhinaus zeigt der Beweis, dass wir nur benötigen, dass die Orakelbandweite h_n^* logarithmisch gegen Null konvergiert, so dass sogar eine Bedingung der Form $\int \log(1+|u|)^\sigma |\mathcal{F}f(u)|^2 du < \infty$ mit $\sigma > 0$ geeignet ausreicht.

Der Beweis des Satzes ist relativ aufwändig. Der Abstand zwischen $\hat{h}_n$, h_n und h_n^* sollte sinnvollerweise mit den zugehörigen Risiken $MISE_n(\bullet)$ gemessen werden. Man mache sich dabei klar, dass $MISE_n(\hat{h}_n)$ zufällig ist und nicht(!) den Fehler des adaptiven Kernschätzers $\hat{f}_n$ angibt. Wir beweisen zunächst einige Lemmata.

5.19 Lemma. *Die stochastische Konvergenz $MISE_n(\hat{h}_n)/MISE_n(h_n) \to 1$ für $n \to \infty$ folgt aus den beiden Konvergenzen*

$$\max_{h \in \mathcal{H}_n} \left(\frac{|ISE_n(h) - MISE_n(h)|}{MISE_n(h)} \right) \to 0 \text{ (stochastisch)},$$

$$\max_{h \in \mathcal{H}_n} \left(\frac{|(\hat{I}_n(h) - \hat{I}_n(h_n)) - (I_n(h) - I_n(h_n))|}{MISE_n(h)} \right) \to 0 \text{ (stochastisch)}.$$

Beweis. Nach Definition gilt stets $MISE_n(\hat{h}_n) \geqslant MISE_n(h_n)$ sowie $CV_n(\hat{h}_n) \leqslant CV_n(h_n)$. Daher erhalten wir

$$\frac{CV_n(\hat{h}_n) - CV_n(h_n) + MISE_n(h_n) - MISE(\hat{h}_n)}{MISE_n(\hat{h}_n)} \leqslant \frac{MISE_n(h_n)}{MISE_n(\hat{h}_n)} - 1 \leqslant 0.$$

Die Behauptung folgt nun aus der Abschätzung:

$$\begin{aligned} &\frac{CV_n(\hat{h}_n) - CV_n(h_n) + MISE_n(h_n) - MISE_n(\hat{h}_n)}{MISE_n(\hat{h}_n)} \\ &= \frac{ISE_n(\hat{h}_n) - MISE_n(\hat{h}_n)}{MISE_n(\hat{h}_n)} - \frac{ISE_n(h_n) - MISE_n(h_n)}{MISE_n(\hat{h}_n)} \end{aligned}$$

$$
\begin{aligned}
&+ 2\frac{\hat{I}_n(h_n) - I_n(h_n)}{MISE_n(\hat{h}_n)} - 2\frac{\hat{I}_n(\hat{h}_n) - I_n(\hat{h}_n)}{MISE_n(\hat{h}_n)} \\
&\geqslant -\frac{|ISE_n(\hat{h}_n) - MISE_n(\hat{h}_n)|}{MISE_n(\hat{h}_n)} - \frac{|ISE_n(h_n) - MISE_n(h_n)|}{MISE_n(h_n)} \\
&\quad - 2\frac{|(\hat{I}_n(\hat{h}_n) - \hat{I}_n(h_n)) - (I_n(\hat{h}_n) - I_n(h_n))|}{MISE_n(\hat{h}_n)} \\
&\geqslant -2\max_{h\in\mathcal{H}_n}\left(\frac{|ISE_n(h) - MISE_n(h)|}{MISE_n(h)}\right) \\
&\quad - 2\max_{h\in\mathcal{H}_n}\left(\frac{|(\hat{I}_n(h) - \hat{I}_n(h_n)) - (I_n(h) - I_n(h_n))|}{MISE_n(h)}\right). \qquad (5.2)
\end{aligned}
$$

□

Wir weisen zunächst für jedes h die Konzentration der jeweiligen Terme um ihren Erwartungswert nach und betrachten danach die Maxima.

5.20 Lemma. *Es gibt eine Konstante $c > 0$, die nur vom Kern K und der Dichte f abhängt, so dass für alle $n \geqslant 1$, $h > 0$, $\kappa > MISE_n(h)(h^{d/4}+n^{-1/2})$ gilt*

$$
\begin{aligned}
&\mathbb{P}(|ISE_n(h) - MISE_n(h)| \geqslant \kappa) \\
&\qquad \leqslant (6n+4)\exp\Big(-c\Big(\kappa/\big(MISE_n(h)(h^{d/4}+n^{-1/2})\big)\Big)^{2/3}\Big).
\end{aligned}
$$

5.21 Bemerkung. Die Größenordnungen in der Abschätzung des Lemmas sind nicht optimal. Für uns wesentlich ist bloß eine Exponentialungleichung von kleinerer Größenordnung als $MISE_n(h)$.

Beweis. Beachte bei folgenden Rechnungen, dass $MISE_n(h)$ von unten sowohl durch den quadrierten Bias als auch durch den Varianzterm beschränkt ist, der die Ordnung $n^{-1}h^{-d}$ besitzt. Wir setzen $\psi_l(u) := e^{i\langle u, X_l\rangle} - \varphi(u)$, $l = 1, \dots, n$, und verwenden, dass $(\psi_l(u))$ i.i.d. sind mit

$$
\mathbb{E}[\psi_l(u)] = 0, \quad a(u,v) := \mathbb{E}[\psi_l(u)\psi_l(v)] = \varphi(u+v) - \varphi(u)\varphi(v)
$$

sowie $\sup_v \int |a(u,v)|^2 du \leqslant 4\|\varphi\|_{L^2}^2 = 4(2\pi)^d\|f\|_{L^2}^2$. Außerdem schreiben wir c_i, $i \geqslant 0$, für positive Konstanten, die nur von der Kernfunktion K und der Dichte f abhängen. Im Spektralbereich benutzen wir die Darstellung (2.2)

und erhalten mittels Fubini

$$
\begin{aligned}
&ISE_n(h) - MISE_n(h)\\
&= (2\pi)^{-d} \int_{\mathbb{R}^d} \big(|\mathcal{F}K(hu)\hat{\varphi}_n(u) - \varphi(u)|^2 - \mathbb{E}[|\mathcal{F}K(hu)\hat{\varphi}_n(u) - \varphi(u)|^2] \big)\, du\\
&= (2\pi)^{-d} \int_{\mathbb{R}^d} \big(|\mathcal{F}K(hu)|^2 \big(|(\hat{\varphi}_n(u) - \varphi(u))|^2 - \mathbb{E}[|\hat{\varphi}_n(u) - \varphi(u)|^2] \big)\\
&\qquad + 2\,\mathrm{Re}\,\big((\mathcal{F}K(hu)^2 - \mathcal{F}K(hu))\varphi(u)(\hat{\varphi}_n(u) - \varphi(u))) \big) \big)\, du\\
&= (2\pi)^{-d} \int_{\mathbb{R}^d} \Big(\frac{2|\mathcal{F}K(hu)|^2}{n^2} \sum_{k=2}^{n} \sum_{l=1}^{k-1} \mathrm{Re}(\psi_k(u)\overline{\psi_l(u)})\\
&\qquad + \frac{|\mathcal{F}K(hu)|^2}{n^2} \sum_{k=1}^{n} \Big(1 - 2\,\mathrm{Re}(\varphi(u)e^{-i\langle u, X_k\rangle}) + |\varphi(u)|^2 - (1 - |\varphi(u)|^2) \Big)\\
&\qquad + 2\,\mathrm{Re}\,\big((\mathcal{F}K(hu)^2 - \mathcal{F}K(hu))\varphi(u)(\hat{\varphi}_n(u) - \varphi(u))) \big) \Big)\, du\\
&= 2\frac{(2\pi)^{-d}}{n^2} \sum_{k=2}^{n} \int_{\mathbb{R}^d} |\mathcal{F}K(hu)|^2 \sum_{l=1}^{k-1} \mathrm{Re}(\psi_k(u)\psi_l(-u))\, du\\
&\quad - 2\frac{(2\pi)^{-d}}{n^2} \sum_{k=1}^{n} \int_{\mathbb{R}^d} |\mathcal{F}K(hu)|^2\, \mathrm{Re}(\varphi(u)\psi_k(-u))\, du\\
&\quad + 2\frac{(2\pi)^{-d}}{n} \sum_{k=1}^{n} \int_{\mathbb{R}^d} \mathrm{Re}\,\big((\mathcal{F}K(hu)^2 - \mathcal{F}K(hu))\varphi(u)\psi_k(u) \big)\, du\\
&=: S_1 + S_2 + S_3.
\end{aligned}
$$

Für den zweiten Summanden S_2 verwenden wir direkt die Hoeffding-Ungleichung. Aus

$$
2(2\pi)^{-d} n^{-2} \int_{\mathbb{R}^d} |\mathcal{F}K(hu)|^2 |\varphi(u)||\psi_k(-u)|\, du \leqslant 4n^{-2}h^{-d}\|K\|_{L^2}^2
$$

folgt

$$
\mathbb{P}(|S_2| \geqslant \kappa_2) \leqslant 2\exp(-\kappa_2^2/32n^{-3}h^{-2d}\|K\|_{L^2}^4), \quad \kappa_2 > 0. \tag{5.3}
$$

Beim dritten Summanden S_3 verwenden wir die klassische Bernstein-Ungleichung. Die Cauchy-Schwarz-Ungleichung liefert

$$
\begin{aligned}
&2\frac{(2\pi)^{-d}}{n} \Big| \int_{\mathbb{R}^d} \mathrm{Re}\,\big((\mathcal{F}K(hu)^2 - \mathcal{F}K(hu))\varphi(u)\psi_k(u) \big)\, du \Big|\\
&\leqslant 2\frac{(2\pi)^{-d}}{n} \Big(\int_{\mathbb{R}^d} |\mathcal{F}K(hu) - 1|^2 |\varphi(u)|^2 du \Big)^{1/2} \Big(\int_{\mathbb{R}^d} 4|\mathcal{F}K(hu)|^2\, du \Big)^{1/2}\\
&\leqslant \frac{4}{n} h^{-d/2} BIAS_n(h) \|K\|_{L^2},
\end{aligned} \tag{5.4}
$$

wobei $BIAS_n(h) = ((2\pi)^{-d}\int_{\mathbb{R}^d}|\mathcal{F}K(hu)-1|^2|\varphi(u)|^2du)^{1/2}$ den Biasterm bezeichnet. Ferner gilt

$$\begin{aligned}
&\mathrm{Var}\Big(2(2\pi)^{-d}\int_{\mathbb{R}^d}\mathrm{Re}\left((\mathcal{F}K(hu)^2-\mathcal{F}K(hu))\varphi(u)\psi_k(-u)\right)du\Big)\\
&\leqslant 4(2\pi)^{-2d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}|\mathcal{F}K(hu)^2-\mathcal{F}K(hu)||\varphi(u)|\bullet\\
&\qquad\qquad\bullet|\mathcal{F}K(-hv)^2-\mathcal{F}K(-hv)||\varphi(-v)||a(u,-v)|\,du\,dv\\
&\leqslant 4(2\pi)^{-2d}\|K\|_{L^1}\int_{\mathbb{R}^d}BIAS_n(h)\Big(\int_{\mathbb{R}^d}|\mathcal{F}K(hu)|^2|a(u,-v)|^2\,du\Big)^{1/2}\bullet\\
&\qquad\qquad\bullet|\mathcal{F}K(-hv)-1||\varphi(-v)|\,dv\\
&\leqslant 4(2\pi)^{-2d}BIAS_n(h)^2\|K\|_{L^1}\Big(\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}|\mathcal{F}K(hu)|^2|a(u,-v)|^2\,dv\,du\Big)^{1/2}\\
&\leqslant 8BIAS_n(h)^2h^{-d/2}\|K\|_{L^1}\|K\|_{L^2}\|f\|_{L^2}. \qquad (5.5)
\end{aligned}$$

Also erhalten wir für $\mathbb{P}(|S_3|\geqslant\kappa_3)$ die obere Schranke

$$2\exp\Big(\frac{-\kappa_3^2/4}{8n^{-1}h^{-d/2}BIAS_n(h)^2\|K\|_{L^1}\|K\|_{L^2}\|f\|_{L^2}+4\kappa_3n^{-1}h^{-d/2}BIAS_n(h)\|K\|_{L^2}}\Big)$$

für beliebiges $\kappa_3>0$. Mit einer Konstanten $c_0>0$ und einer Abschätzung der Form $2AB\leqslant A^2+B^2$ vereinfacht sich dies für $\kappa_3\geqslant h^{d/4}MISE_n(h)$ zu

$$\mathbb{P}(|S_3|\geqslant\kappa_3)\leqslant 2\exp\Big(-c_0\kappa_3/\big(h^{d/4}MISE_n(h)+n^{-1/2}MISE_n(h)\big)\Big). \qquad (5.6)$$

Zur Abschätzung des ersten Summanden S_1 führen wir die Notation

$$U_n := (2\pi)^{-d}\sum_{k=2}^{n}\int_{\mathbb{R}^d}|\mathcal{F}K(hu)|^2\sum_{l=1}^{k-1}\mathrm{Re}\left(\psi_k(u)\psi_l(-u)\right)du \qquad (5.7)$$

ein (U_n ist eine sogenannte *U-Statistik* und wir folgen nun Ideen von Houdré and Reynaud-Bouret (2003)). Bezüglich der Filtration $\mathcal{F}_n := \sigma(X_1,\ldots,X_n)$ ist U_n, $n\geqslant 1$, ein Martingal (setze $U_1:=0$); denn U_n ist $\mathcal{F}_n$-messbar und

$$\mathbb{E}[U_{n+1}-U_n\,|\,\mathcal{F}_n] = (2\pi)^{-d}\sum_{l=1}^{n}\int_{\mathbb{R}^d}|\mathcal{F}K(hu)|^2\,\mathrm{Re}\left(\mathbb{E}[\psi_{n+1}(u)]\psi_l(-u)\right)du = 0.$$

Außerdem gilt

$$|U_{n+1}-U_n| \leqslant (2\pi)^{-d}\sum_{l=1}^{n}\int_{\mathbb{R}^d}|\mathcal{F}K(hu)|^2|\psi_{n+1}(u)||\psi_l(-u)|\,du \leqslant 4nh^{-d}\|K\|_{L^2}^2.$$

Die Standardabweichung von U_n ist, wie wir sehen werden, von der Ordnung $nh^{-d/2}$ und wir erwarten, dass U_n eine Konzentrationsungleichung von dieser

Größenordnung erfüllt. Die Hoeffding-Ungleichung angewendet auf U_n liefert jedoch eine suboptimale Abschätzung von der Ordnung $n^{3/2}h^{-d}$ (wir brauchen $o(n^2 MISE_n)$, z.B. $o(nh^{-d})$). Beachte, dass wir bereits für $|U_{n+1} - U_n|$ eine viel bessere Konzentration mit der Hoeffding-Ungleichung für die Differenz, bedingt auf $X_{n+1} = x$ erreichen:

$$\mathbb{P}(|U_{n+1} - U_n| \geqslant \rho \,|\, X_{n+1} = x) \leqslant 2 \exp\Big(-\rho^2/(2nh^{-2d}\|K\|_{L^2}^4 4\|e^{i\bullet x} - \varphi\|_\infty^2)\Big).$$

Dies impliziert

$$\mathbb{P}(|U_{n+1} - U_n| \geqslant \rho) = \mathbb{E}[\mathbb{P}(|U_{n+1} - U_n| \geqslant \rho \,|\, X_{n+1})] \leqslant 2 \exp\Big(\frac{-\rho^2}{32nh^{-2d}\|K\|_{L^2}^4}\Big). \tag{5.8}$$

Wir streben daher an, die Bernstein-Ungleichung auf das Martingal (U_n) anzuwenden, und bestimmen dazu die quadratische Variation von U_n. Wir benutzen $\operatorname{Re}(w)\operatorname{Re}(z) = \operatorname{Re}(wz + w\bar{z})/2$ für $w, z \in \mathbb{C}$ sowie $|\mathcal{F}K(-hv)| = |\mathcal{F}K(hv)|$ und erhalten:

$$\begin{aligned}
&\langle U\rangle_{n+1} - \langle U\rangle_n = \mathbb{E}[(U_{n+1} - U_n)^2 \,|\, \mathcal{F}_n] \\
&= (2\pi)^{-2d}\, \mathbb{E}\Big[\Big(\sum_{l=1}^n \int_{\mathbb{R}^d} |\mathcal{F}K(hu)|^2 \operatorname{Re}(\psi_{n+1}(u)\psi_l(-u))\, du\Big)^2 \,\Big|\, \mathcal{F}_n\Big] \\
&= (2\pi)^{-2d} \sum_{l,l'=1}^n \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |\mathcal{F}K(hu)|^2 |\mathcal{F}K(hv)|^2 \\
&\qquad \tfrac{1}{2} \operatorname{Re}\Big(\psi_l(-u)\big(a(u,v)\psi_{l'}(-v) + a(u,-v)\psi_{l'}(v)\big)\Big)\, du\, dv \\
&= (2\pi)^{-2d} \sum_{l,l'=1}^n \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |\mathcal{F}K(hu)|^2 |\mathcal{F}K(hv)|^2 a(u,-v)\psi_l(-u)\psi_{l'}(v)\, du\, dv
\end{aligned}$$

Wegen $\int\int a(u,-v)\overline{g(u)}g(v)\, du\, dv \geqslant 0$ für alle $g \in L^1(\mathbb{R}^d)$ (folgt aus $a(u,-v) = \mathbb{E}[\psi_l(u)\overline{\psi_l(v)}]$) ist $(g,h) \mapsto \int\int a(u,-v)\overline{h(u)}g(v)\, du\, dv$ eine positiv-definite Bilinearform (Skalarprodukt), und Dualität ergibt eine Darstellung der Norm als Supremum über lineare Skalarprodukte:

$$\begin{aligned}
&\langle U\rangle_{n+1} - \langle U\rangle_n \\
&= \sup_{g \in L^1 \cap L^2(\mathbb{R}^d), g \neq 0} \Big|(2\pi)^{-d} \frac{\int\int a(u,-v)|\mathcal{F}K(hu)|^2 \sum_{l=1}^n \psi_l(v)\, dv\, \overline{g(u)}\, du}{\big(\int\int a(u,-v)\overline{g(u)}g(v)\, du\, dv\big)^{1/2}}\Big|^2 \\
&= \sup_{g \in \mathcal{G}} \Big|\sum_{l=1}^n X_l^{(g)}\Big|^2
\end{aligned}$$

mit einer abzählbaren in der $L^2(\mathbb{R}^d)$-Topologie dichten Teilmenge $\mathcal{G}$ von $\{g \in L^1 \cap L^2 : \int\int a(u,-v)\overline{g(u)}g(v)\,du\,dv = 1\}$ (der L^2 ist separabel) sowie für jedes $g \in \mathcal{G}$ unabhängigen Zufallsvariablen

$$X_l^{(g)} := (2\pi)^{-d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} a(u,-v)|\mathcal{F}K(hu)|^2 \psi_l(v)\overline{g(u)}\,dv\,du, \quad l = 1, \cdots, n.$$

Setzen wir noch $S_n^{(g)} := \sum_{l=1}^n X_l^{(g)}$, so liefert die Talagrand-Ungleichung (angewendet auf Real- und Imaginärteil) für $\varepsilon = 1$

$$\mathbb{P}\Big(\sup_{g\in\mathcal{G}}|S_n^{(g)}| \geqslant 2\,\mathbb{E}\Big[\sup_{g\in\mathcal{G}}|S_n^{(g)}|\Big] + c_1\sigma\beta + c_2 R\beta^2\Big) \leqslant \exp(-\beta^2/2), \quad \beta > 0,$$

mit Konstanten $c_1, c_2 > 0$. Mit der Abschätzung

$$\max\Big(\sigma^2, \mathbb{E}[\sup_{g\in\mathcal{G}}|S_n^{(g)}|]^2\Big) \leqslant \mathbb{E}[\sup_{g\in\mathcal{G}}|S_n^{(g)}|^2],$$

sowie $(A+B)^2 \leqslant 2(A^2+B^2)$ erhalten wir für alle $\beta > 0$

$$\mathbb{P}\Big(\langle U\rangle_n - \langle U\rangle_{n-1} \geqslant 2(2+c_1\beta)^2\,\mathbb{E}[\langle U\rangle_n - \langle U\rangle_{n-1}] + 2c_2^2R^2\beta^4\Big) \leqslant \exp(-\beta^2/2).$$

Wir berechnen

$$\begin{aligned}
\mathbb{E}[\langle U\rangle_n - \langle U\rangle_{n-1}] &= (2\pi)^{-2d} n \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |\mathcal{F}K(hu)|^2|\mathcal{F}K(hv)|^2|a(u,-v)|^2\,du\,dv \\
&\leqslant nh^{-d}\|K\|_{L^2}^2\|K\|_{L^1}^2\|f\|_{L^2}^2, \qquad (5.9)\\
R^2 &\leqslant (2\pi)^{-2d}4\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |\mathcal{F}K(hu)|^2|\mathcal{F}K(hv)|^2|a(u,-v)|\,du\,dv \\
&\leqslant 8h^{-3d/2}\|K\|_{L^2}^3\|K\|_{L^1}\|f\|_{L^2} \qquad (5.10)
\end{aligned}$$

und schließen mit neuen Konstanten $c_3, c_4 > 0$ für $\beta \geqslant 1$

$$\mathbb{P}\Big(\langle U\rangle_n - \langle U\rangle_{n-1} \geqslant c_3\beta^2 nh^{-d} + c_4\beta^4 h^{-3d/2}\Big) \leqslant \exp(-\beta^2/2).$$

Eine grobe Abschätzung liefert für $\bar\beta = n(c_3\beta^2 nh^{-d} + c_4\beta^4 h^{-3d/2})$ schließlich mit $c_5 > 0$

$$\mathbb{P}\Big(\langle U\rangle_n \geqslant \bar\beta\Big) \leqslant n\exp\Big(-c_5\big(\bar\beta/(n^2h^{-d}+nh^{-3d/2})\big)^{1/2}\Big), \quad \bar\beta \geqslant n^2h^{-d}+nh^{-3d/2}. \qquad (5.11)$$

Die Bernstein-Ungleichung für das Martingal (U_n) impliziert daher

$$\begin{aligned}
\mathbb{P}(|U_n| \geqslant \bar\kappa) &\leqslant n\exp\Big(-c_5\big(\bar\beta/(n^2h^{-d}+nh^{-3d/2})\big)^{1/2}\Big) \\
&\quad + 2n\exp\Big(-\rho^2/(32nh^{-2d}\|K\|_{L^2}^4)\Big) + 2\exp\Big(-\frac{\bar\kappa^2}{4(\bar\beta+\bar\kappa\rho)}\Big).
\end{aligned}$$

Wähle $\bar\beta = \bar\kappa\rho$ sowie $\rho = \bar\kappa^{1/3}(n^2h^{-d} + nh^{-2d})^{1/3}$ und betrachte nur noch Werte $\bar\kappa > \rho$. Dann folgt mit einer Konstanten $c_6 > 0$

$$\mathbb{P}(|U_n| \geqslant \kappa) \leqslant (4n+2)\exp\Big(- c_6\bar\kappa^{2/3}/\big(n^2h^{-d} + nh^{-2d}\big)^{1/3}\Big).$$

Es folgt schließlich für $S_1 = 2U_n/n^2$ mit einer Konstanten $c_7 > 0$

$$\mathbb{P}(|S_1| \geqslant \kappa_1) \leqslant 6n\exp\Big(- c_7\big(\kappa_1/(n^{-1}h^{-d/2} + n^{-3/2}h^{-d})\big)^{2/3}\Big). \qquad (5.12)$$

Die Behauptung folgt durch Addition der Abschätzungen (5.12), (5.3), (5.6) für S_1, S_2, S_3 mit $\kappa_i = \kappa/3$ und Zusammenfassen der Terme. □

5.22 Korollar. *Für die angegebene Bandweitenmenge $\mathcal{H}_n$ gilt mit stochastischer Konvergenz*

$$\lim_{n\to\infty} \max_{h\in\mathcal{H}_n} \frac{|ISE_n(h) - MISE_n(h)|}{MISE_n(h)} = 0.$$

Beweis. Nach der allgemeinen Abschätzung (5.1) folgt aus Lemma 5.20 für $\varepsilon > 0$

$$\begin{aligned}
&\mathbb{P}\Big(\max_{h\in\mathcal{H}_n} \frac{|ISE_n(h) - MISE_n(h)|}{MISE_n(h)} \geqslant \varepsilon\Big)\\
&\leqslant \sum_{h\in\mathcal{H}_n} 10n\exp\Big(- c\big(\varepsilon/(h^{d/4} + n^{-1/2})\big)^{2/3}\Big)\\
&\leqslant 10|\mathcal{H}_n|n\exp\Big(- c\big(\varepsilon/((\max\mathcal{H}_n)^{d/4} + n^{-1/2})\big)^{2/3}\Big).
\end{aligned}$$

Aus den Eigenschaften $(\max\mathcal{H}_n)^{d/4} = (\log n)^{-3}$ sowie $|\mathcal{H}_n| \leqslant n^{d+1}$ folgt die Ordnung $(\log n)^2$ im Exponential und damit die stochastische Konvergenz gegen Null. □

Unter der Voraussetzung, dass der Kern beschränkt ist, erhalten wir eine analoge Abschätzung für den zweiten Term in der Zerlegung (5.2). Beachte, dass sich die Abschätzungen im Beweis stark vereinfachen, falls $\mathcal{F}K \in L^1(\mathbb{R}^d)$ gilt, was aber beispielsweise den Rechteckkern ausschließen würde.

5.23 Lemma. *Die Kernfunktion $K \in L^1(\mathbb{R}^d)$ sei beschränkt. Dann gibt es eine Konstante $c > 0$, die nur vom Kern K und der Dichte f abhängt, so dass für alle $n \geqslant 1$, $h > 0$, $\kappa > MISE_n(h)(h^{d/4} + n^{-1/2})$ gilt*

$$\begin{aligned}
&\mathbb{P}(|(\hat I_n(h) - \hat I_n(h_n)) - (I_n(h) - I_n(h_n))| \geqslant \kappa)\\
&\leqslant (6n+2)\exp\Big(- c\Big(\kappa/\big((h^{d/4} + n^{-1/2})MISE_n(h) + h_n^{d/4}MISE_n(h_n)\big)\Big)^{2/3}\Big).
\end{aligned}$$

Beweis. Wiederum gehen wir in den Spektralbereich über und benutzen die Funktion $\psi_l(u) = e^{iuX_l} - \varphi(u)$ sowie $\Delta K_h := K_h - K_{h_n}$. Wir erhalten die Darstellung:

$$\begin{aligned}\hat{I}_n(h) - \hat{I}_n(h_n) &= \frac{1}{n}\sum_{k=1}^{n}(\hat{f}_{n,h}^{(-k)} - \hat{f}_{n,h_n}^{(-k)})(X_k)\\ &= \frac{1}{n}\sum_{k=1}^{n}(2\pi)^{-d}\int_{\mathbb{R}^d}\mathcal{F}(\hat{f}_{n,h}^{(-k)} - \hat{f}_{n,h_n}^{(-k)})(u)e^{-iuX_k}du\\ &= \frac{1}{n(n-1)}\sum_{k=1}^{n}\sum_{l\neq k}(2\pi)^{-d}\int_{\mathbb{R}^d}\mathcal{F}\Delta K_h(u)e^{iuX_l}e^{-iuX_k}du.\end{aligned}$$

Beachte, dass diese Identität im Allgemeinen nur fast sicher gilt, weil $\hat{f}_{n,h}^{(-k)}$ in $L^2(\mathbb{R}^d)$ nur fast überall festgelegt ist. Wir schließen:

$$\begin{aligned}&(\hat{I}_n(h) - \hat{I}_n(h_n)) - (I_n(h) - I_n(h_n))\\ &= \frac{1}{n(n-1)}\sum_{k=1}^{n}\sum_{l\neq k}(2\pi)^{-d}\int_{\mathbb{R}^d}\Big(\mathcal{F}\Delta K_h(u)(e^{iu(X_l-X_k)} - \varphi(u)\varphi(-u))\Big)\,du\\ &= \frac{2}{n(n-1)}\sum_{k=2}^{n}\sum_{l=1}^{k-1}(2\pi)^{-d}\int_{\mathbb{R}^d}\mathrm{Re}\Big(\mathcal{F}\Delta K_h(u)\psi_l(u)\psi_k(-u)\Big)\,du\\ &\quad + \frac{2}{n}\sum_{k=1}^{n}(2\pi)^{-d}\int_{\mathbb{R}^d}\mathrm{Re}\Big(\mathcal{F}\Delta K_h(u)\varphi(u)\psi_k(-u)\Big)\,du\\ &=: T_1 + T_2.\end{aligned}$$

Bei genauerem Hinsehen stellt man fest, dass der Term T_1 dieselbe Struktur wie der Term S_1 im vorigen Beweis hat sowie T_2 dieselbe Struktur wie S_3. Allerdings erscheinen die Fouriertransformierten des Kerns in anderer Weise.

Zur Behandlung von T_2 setzen wir

$$B := \{w \in \mathbb{R}^d \mid |\mathcal{F}K(w)| > 1/2\}$$

und bemerken, dass $|1-\mathcal{F}K(w)| \geqslant 1/2$ für $w \in B$ gilt. Wegen $\mathcal{F}K \in L^2(\mathbb{R}^d)$ ist das Lebesguemaß $\lambda(B)$ von B endlich. In Analogie zu (5.4) erhalten wir mit der Cauchy-Schwarz-Ungleichung

$$\begin{aligned}&\frac{2}{n}(2\pi)^{-d}\int_{\mathbb{R}^d}|\mathcal{F}\Delta K_h(u)||\varphi(u)||\psi_k(-u)|\,du\\ &\leqslant \frac{4}{n}(2\pi)^{-d}\Big(\Big(\int_{h^{-1}B}|\mathcal{F}\Delta K_h(u)|^2|\varphi(u)|^2du\Big)^{1/2}\lambda(h^{-1}B)^{1/2}\\ &\quad + \Big(\int_{\mathbb{R}^d\setminus h^{-1}B}|\mathcal{F}\Delta K_h(u)|^2du\Big)^{1/2}\Big(\int_{\mathbb{R}^d\setminus h^{-1}B}|\varphi(u)|^2du\Big)^{1/2}\Big)\end{aligned}$$

$$\leqslant \frac{4}{n}(2\pi)^{-d/2}\Big(\Big(2BIAS_n(h)^2 + 2BIAS_n(h_n)^2\Big)^{1/2} h^{-d/2}\lambda(B)^{1/2} + \|K\|_{L^2}(2h^{-d/2} + 2h_n^{-d/2})2BIAS_n(h)\Big).$$

Wegen $MISE_n(h) \geqslant MISE_n(h_n)$ erhalten wir mit einer Konstanten $c_1 > 0$

$$\frac{2}{n}(2\pi)^{-d}\int_{\mathbb{R}^d} |\mathcal{F}\Delta K_h(u)||\varphi(u)||\psi_k(-u)|\,du \leqslant c_1 n^{-1/2} MISE_n(h).$$

Mit entsprechenden Abschätzungen folgt wie in (5.5) für ein $c_2 > 0$

$$\mathrm{Var}\Big(2(2\pi)^{-d}\int_{\mathbb{R}^d} |\mathcal{F}\Delta K_h(u)||\varphi(u)||\psi_k(-u)|\,du\Big) \leqslant c_2 n^{1/2} MISE_n(h)^2.$$

Die Bernstein-Ungleichung liefert demnach

$$\mathbb{P}(|T_2| > \kappa_2) \leqslant 2\exp\Big(-\kappa_2^2/\Big(4(c_2 MISE_n(h) + c_1\kappa_2)n^{-1/2} MISE_n(h)\Big)\Big).$$

Zur Behandlung von T_1 setzt man wie im vorigen Beweis

$$\begin{aligned} U_n &:= \sum_{k=2}^{n}\sum_{l=1}^{k-1}(2\pi)^{-d}\int_{\mathbb{R}^d} \mathrm{Re}\Big(\mathcal{F}\Delta K_h(u)\psi_l(u)\psi_k(-u)\Big)\,du \\ &= \sum_{k=2}^{n}\sum_{l=1}^{k-1}\Big(\Delta K_h(X_l - X_k) - \mathbb{E}_f[\Delta K_h(X_l - X_k)\,|\,X_l] \\ &\quad - \mathbb{E}_f[\Delta K_h(X_l - X_k)\,|\,X_k] + \mathbb{E}_f[\Delta K_h(X_l - X_k)]\Big). \end{aligned}$$

Jeder Summand $\int \mathcal{F}\Delta K_h(u)\psi_l(u)\psi_k(-u)\,du$ ist gleichmäßig von der Größenordnung $h^{-d} + h_n^{-d}$, wie sofort aus der Darstellung im Ortsbereich und der Beschränktheit von K folgt. Wie in (5.8) liefert die Hoeffding-Ungleichung eine exponentielle Abschätzung von $|U_{n+1} - U_n|$ der Ordnung $n^{1/2}(h^{-d} + h_n^{-d})$. Vollkommen analog wendet man die Talagrand-Ungleichung auf die entsprechende quadratische Form an mit der Größenordnung $n(h^{-d} + h_n^{-d})$ für $\mathbb{E}[\langle U\rangle_{n+1} - \langle U\rangle_n]$, vergleiche (5.9). Für die gleichmäßige Schranke R in der Talagrand-Ungleichung schätzen wir teilweise im Fourier- und teilweise im Ortsbereich ab:

$$\begin{aligned} &(2\pi)^{-2d}\Big|\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \mathcal{F}\Delta K_h(u)\mathcal{F}\Delta K_h(-v)\psi_l(u)\psi_l(-v)a(u,-v)\,du\,dv\Big| \\ &\leqslant (2\pi)^{-2d}\Big(\Big|2\int_{\mathbb{R}^d} \mathcal{F}\Delta K_h(u)\varphi(u)\,du\Big|^2 \\ &\quad + \Big|\int_{\mathbb{R}^d} \mathcal{F}\Delta K_h(u)\psi_l(u)(\overline{\mathcal{F}\Delta K_h\psi_l} * \varphi)(u)\,du\Big|\Big) \\ &\leqslant 8(h^{-d} + h_n^{-d})\|K\|_{L^2}^2\|f\|_{L^2}^2 + \int_{\mathbb{R}^d}(\Delta K_h * (\delta_{X_l} - f))(x)^2 f(x)\,dx \end{aligned}$$

$$
\begin{aligned}
&\leqslant 8(h^{-d}+h_n^{-d})\|K\|_{L^2}^2\|f\|_{L^2}^2 + \|\Delta K_h\|_\infty \|\Delta K_h * (\delta_{X_l} - f)\|_{L^2}\|f\|_{L^2} \\
&\leqslant 8(h^{-d}+h_n^{-d})\|f\|_{L^2}\Big(\|K\|_{L^2}^2\|f\|_{L^2} \\
&\quad + \|K\|_\infty\Big((h^{-d/2}+h_n^{-d/2})\|K\|_{L^2} + \|f\|_{L^2} + BIAS_n(h) + BIAS_n(h_n)\Big)\Big).
\end{aligned}
$$

Diese Abschätzung zeigt, dass R^2 von der Ordnung $\mathcal{O}(h^{-3d/2}+h_n^{-3d/2})$ ist, vergleiche auch (5.10). Da sich die Größenordnungen der Abschätzungen nicht geändert haben, erhalten wir für T_1 die gleiche Abschätzung wie (5.12):

$$
\mathbb{P}(|T_1| \geqslant \kappa_1) \leqslant 6n \exp\Big(-c_3\big(\kappa_1/(n^{-1}(h^{-d/2}+h_n^{-d/2})+n^{-3/2}(h^{-d}+h_n^{-d}))\big)^{2/3}\Big),
$$

wobei $c_3 > 0$ eine Konstante ist und $\kappa_1 \geqslant n^{-1}(h^{-d/2}+h_n^{-d/2})+n^{-3/2}(h^{-d}+h_n^{-d})$ beliebig ist. Es bleibt, die Abschätzungen für T_1 und T_2 zusammenzuaddieren und zu vereinfachen. □

Beweis von Satz 5.17. Zunächst weisen wir $MISE_n(\hat{h}_n)/MISE_n(h_n) \to 1$ nach. Gemäß Lemma 5.19 reicht es dazu, die stochastische Konvergenz der beiden Maxima in (5.2) gegen Null zu zeigen. In Korollar 5.22 ist die Konvergenz des ersten Maximums bereits erbracht worden. Da Lemma 5.23 die gleichen Abschätzungen liefert, wenn man $h_n \in \mathcal{H}_n$ beachtet, erhalten wir vollkommen analog auch die Konvergenz des zweiten Maximums gegen Null.

Wie bereits angemerkt, ist $MISE(\hat{h}_n)$ nicht der Fehler von $\hat{f}_n$. Das Korollar 5.22 liefert aber gerade stochastische Konvergenz

$$
\lim_{n\to\infty} \max_{h\in\mathcal{H}_n} \frac{|ISE_n(h) - MISE_n(h)|}{MISE_n(h)} = 0.
$$

Zusammen mit $MISE_n(\hat{h}_n)/MISE_n(h_n) \to 1$ folgt daher

$$
\lim_{n\to\infty} \frac{|ISE_n(\hat{h}_n) - MISE_n(h_n)|}{MISE_n(h_n)} = 0 \text{ (stochastisch)} \tag{5.13}
$$

oder äquivalent $ISE_n(\hat{h}_n)/MISE_n(h_n) \to 1$ (stochastisch).

Es bleibt, die Abweichung von $MISE_n(h_n)$ in Bezug auf das eigentliche Orakelrisiko $MISE_n(h_n^*)$ zu ermitteln. Bei der Wahl von $\mathcal{H}_n$ beachte, dass wir nur Bandweiten $h > n^{-1/d}$ zu betrachten brauchen, weil für $h \leqslant n^{-1/d}$ der Varianzanteil in $MISE_n(h)$ nicht gegen Null konvergiert. Außerdem impliziert die Annahme $f \in H^s(\mathbb{R}^d)$ für ein $s > 0$, dass die Orakelbandweite h_n^* von der Ordnung $n^{-1/(2s+d)}$ ist und damit kleiner als $(\log n)^{-d/12}$. Dann gilt für den *Diskretisierungsfehler* $MISE_n(h_n) - \inf_{h\in[n^{-1/d},(\log n)^{-12/d}]} MISE_n(h)$, dass der Varianzanteil von der Ordnung

$$
|(n^{1/d}h_n)^{-d} - (n^{1/d}h_n^*)^{-d}| \leqslant d|n^{1/d}h_n - n^{1/d}h_n^*| \leqslant dn^{1/d-d-1}(\log n)^{-d/12}
$$

ist (beachte $|A^{-d}-B^{-d}| \leqslant d|A-B|$ für $A, B \geqslant 1$), was um Größenordnungen kleiner ist als $MISE_n(h_n^*) \geqslant n^{-1}$. Der Biasanteil ist bestimmt durch

$$\sup_{|\delta|<n^{-d-1}} \Big| \int_{\mathbb{R}^d} (|\mathcal{F}K(h_n u)|^2 - |\mathcal{F}K((h_n+\delta)u)|^2)|\varphi(u)|^2 du \Big|.$$

Benutzen wir nun die gleichmäßige Lipschitzstetigkeit von $\mathcal{F}K$, so ist das Integral von der Ordnung

$$\int (1 \wedge |\delta u|)|\varphi(u)|^2 du \leqslant \int |\delta u|^{2s\wedge 1}|\varphi(u)|^2 du \leqslant |\delta|^{2s\wedge 1}\|f\|_{H^s}^2.$$

Also ist der Biasanteil im Diskretisierungsfehler ebenfalls von kleinerer Ordnung als $n^{-2s/(2s+d)}$, die Ordnung von $MISE_n(h_n^*)$. Wir haben insgesamt gezeigt, dass der Diskretisierungsfehler vernachlässigbar ist:

$$\lim_{n\to\infty} \frac{MISE_n(h_n) - MISE_n(h_n^*)}{MISE_n(h_n^*)} = 0.$$

Dieses deterministische Konvergenzresultat ergibt zusammen mit (5.13) die Behauptung. □

5.4 Thresholding und Wavelets

In diesem Abschnitt betrachten wir wiederum ein Signal in weißem Rauschen

$$dY_t = f(t)\,dt + \frac{\sigma}{\sqrt{n}}\,dW_t, \quad t \in [0,1]$$

und das äquivalente Folgenraummodell bezüglich einer Orthonormalbasis $(\varphi_k)_{k\geqslant 1}$

$$y_k := \int_0^1 \varphi_k(t)\,dY_t = f_k + \frac{\sigma}{\sqrt{n}}\zeta_k, \quad k \geqslant 1,$$

mit $f_k = \langle f, \varphi_k\rangle_{L^2}$ und $\zeta_k \sim N(0,1)$ i.i.d. Bei Wahl einer geeigneten Basis (φ_k) können wir hoffen, dass sich f bereits durch wenige Koeffizienten f_k gut darstellen lässt, das heißt die meisten Koeffizienten sehr klein sind. Die beobachteten empirischen Koeffizienten y_k streuen um die wahren Koeffizienten f_k jeweils in der Grösenordnung $\frac{\sigma}{\sqrt{n}}$. Die Idee des *Hard-Thresholding* ist nun, jeden Koeffizienten f_k direkt durch y_k zu schätzen, sofern $|y_k| \gg \frac{\sigma}{\sqrt{n}}$ gilt, andernfalls jedoch durch Null zu schätzen (*keep or kill*). Wir setzen für $K \in \mathbb{N}$ und einen Schwellwert $\kappa \geqslant 1$

$$\hat{f}^{(hard)} := \sum_{k=1}^{K} \hat{f}_k^{(hard)}\varphi_k, \quad \hat{f}_k^{(hard)} := y_k \mathbf{1}(|y_k| \geqslant \kappa\tfrac{\sigma}{\sqrt{n}}).$$

Falls $|y_k|$ nicht viel größer als das Rauschniveau ist, so ist der stochastische Fehler im Allgemeinen nicht kleiner als das Signal, und es zahlt sich aus,

auf Kosten eines Bias die Varianz auf Null zu reduzieren. Das Verfahren kann auch als multiples Testproblem interpretiert werden, wo jeweils die Hypothese $H_k : f_k = 0$ getestet wird, oder aber als Modellwahlproblem, wo unter allen Teilmengen der Indexmenge ausgewählt werden kann.

Erwünschter Nebeneffekt dieser Methode ist eine starke Datenkompression ; $\hat{f}^{(hard)}$ wird im Allgemeinen wenige Koeffizienten ungleich Null besitzen, zumindest wenn f nur wenige signifikante Koeffizienten besitzt, das heißt solche vom Betrag größer als das Rauschniveau. Diese sparsame Basisdarstellung (*sparse representation, sparsity*) ist ein wichtiges Paradigma moderner Statistik und Bestandteil vieler Methoden (z.B. LASSO).

Für die Analyse des Hard-Thresholding betrachten wir zunächst die Orakelwahl $\hat{f}_k^{(Orakel)} \in \{0, y_k\}$, $1 \leqslant k \leqslant K$. Wegen

$$\mathbb{E}_f[(\hat{f}_k^{(Orakel)} - f_k)^2] = \begin{cases} \sigma^2/n, & \hat{f}_k^{(Orakel)} = 0, \\ f_k^2, & \hat{f}_k^{(Orakel)} = y_k \end{cases}$$

ist $\hat{f}_k^{(Orakel)} = y_k \mathbf{1}(|f_k| \geqslant \frac{\sigma}{\sqrt{n}})$ die Orakelwahl, und der Orakelschätzer besitzt das Orakelrisiko

$$\mathbb{E}_f[\|\hat{f}^{(Orakel)} - f\|_{L^2}^2] = \sum_{k=1}^{K} (f_k^2 \wedge \frac{\sigma^2}{n}) + \sum_{k>K} f_k^2.$$

Zum Vergleich betrachte einen Projektionsschätzer $\hat{f}_M$ auf die ersten M Koeffizienten mit MISE

$$\mathbb{E}_f[\|\hat{f}_M - f\|_{L^2}^2] = \sum_{k=1}^{M} \frac{\sigma^2}{n} + \sum_{k>M} f_k^2.$$

Wir sehen also, dass für $M \leqslant K$ der Orakel-Thresholding-Schätzer stets mindestens genauso gut ist wie ein Projektionsschätzer. Insbesondere erhalten wir daher für die Fourierbasis und Funktionen f aus $H^s_{per}([0,1])$, $s > 0$, die optimalen Minimaxraten, sofern nur $K \geqslant (2s\frac{n}{\sigma^2}\|f\|_s^2)^{1/(2s+1)}$ gewählt ist. Eine kanonische Wahl von K ist im Regressionsproblem die Stichprobengröße n, weil die n Daten maximal zu n linear unabhängigen empirischen Koeffizienten führen können (für $K > n$ bricht die Analogie zum Signal im weißen Rauschen zusammen). Asymptotisch ist ein solcher Orakel-Schätzer dann raten-optimal in der Skala aller Sobolevräume der Ordnung $s > 0$. Die Hoffnung, dass der Thresholding-Schätzer damit ein adaptiver raten-optimaler Schätzer ist, erfüllt sich fast, das heißt bis auf einen logarithmischen Faktor.

5.24 Satz. *Betrachte im Folgenraummodell den Hard-Thresholding-Schätzer*

$$\hat{f}^{(hard)} := \sum_{k=1}^{K} \hat{f}_k^{(hard)} \varphi_k, \quad \hat{f}_k^{(hard)} := y_k \mathbf{1}(|y_k| \geqslant \kappa \tfrac{\sigma}{\sqrt{n}})$$

mit $\kappa \geqslant 2$ und $K \in \mathbb{N}$ sowie den entsprechenden Orakel-Thresholding-Schätzer $\hat{f}^{(Orakel)}$. Dann gilt folgende Orakelungleichung für beliebige $f \in L^2([0,1])$

$$\mathbb{E}_f[\|\hat{f}^{(hard)} - f\|_{L^2}^2] \leqslant (4\kappa^2+1)\,\mathbb{E}[\|\hat{f}^{(Orakel)} - f\|_{L^2}^2] + \big(K\tfrac{\sigma^2}{n} + \|f\|_{L^2}^2\big)\tfrac{\kappa}{\sqrt{2\pi}} e^{-(\kappa-1)^2/2}.$$

Bei Wahl von $\kappa = \sqrt{2\log K}$ ($\sqrt{2\log K}$ heißt universal threshold) gilt dann insbesondere

$$\mathbb{E}_f[\|\hat{f}^{(hard)} - f\|_{L^2}^2] \leqslant (8\log K+1)\,\mathbb{E}[\|\hat{f}^{(Orakel)} - f\|_{L^2}^2] + \big(\tfrac{\sigma^2}{n} + \tfrac{\|f\|_{L^2}^2}{K}\big)\sqrt{\log K} K^{1/\sqrt{\log K}}.$$

Beweis. Die Idee ist es, den Fehler in jedem Koeffizienten in vier Fälle aufzuspalten, die sich durch Thresholding / kein Thresholding im Fall großer / kleiner Koeffizienten ergeben ($k \leqslant K$):

$$\begin{aligned}
\mathbb{E}_f[(\hat{f}_k^{(hard)} - f_k)^2] &= \mathbb{E}_f[(y_k \mathbf{1}(|y_k| \geqslant \kappa\tfrac{\sigma}{\sqrt{n}}) - f_k)^2] \\
&= \mathbb{E}_f[(y_k - f_k)^2 \mathbf{1}(|y_k| \geqslant \kappa\tfrac{\sigma}{\sqrt{n}}, |f_k| \leqslant \tfrac{\sigma}{\sqrt{n}})] \\
&\quad + \mathbb{E}_f[(y_k - f_k)^2 \mathbf{1}(|y_k| \geqslant \kappa\tfrac{\sigma}{\sqrt{n}}, |f_k| > \tfrac{\sigma}{\sqrt{n}})] \\
&\quad + \mathbb{E}_f[f_k^2 \mathbf{1}(|y_k| < \kappa\tfrac{\sigma}{\sqrt{n}}, |f_k| > 2\kappa\tfrac{\sigma}{\sqrt{n}})] \\
&\quad + \mathbb{E}_f[f_k^2 \mathbf{1}(|y_k| < \kappa\tfrac{\sigma}{\sqrt{n}}, |f_k| \leqslant 2\kappa\tfrac{\sigma}{\sqrt{n}})] \\
&=: T_1 + T_2 + T_3 + T_4.
\end{aligned}$$

Die Terme T_1 und T_3 werden mittels *großer Abweichungen* für normalverteilte Zufallsvariablen abgeschätzt, während T_2 und T_4 bis auf den Schwellenwert κ den entsprechenden Fehlern beim Orakelschätzer entsprechen; wir erhalten

$$T_2 \leqslant \mathbb{E}[\tfrac{\sigma^2}{n}\zeta_k^2 \mathbf{1}(|f_k| > \tfrac{\sigma}{\sqrt{n}})] = \tfrac{\sigma^2}{n}\mathbf{1}(|f_k| > \tfrac{\sigma}{\sqrt{n}})$$

sowie $T_4 \leqslant f_k^2 \mathbf{1}(|f_k| \leqslant 2\kappa\tfrac{\sigma}{\sqrt{n}})$ und somit

$$T_2 + T_4 \leqslant (1 + 4\kappa^2)(f_k^2 \wedge \tfrac{\sigma^2}{n}).$$

Für T_1 verwenden wir das Integral

$$\frac{1}{\sqrt{2\pi}}\int_A^\infty x(xe^{-x^2/2})dx = \frac{1}{\sqrt{2\pi}}\Big(-xe^{-x^2/2}|_A^\infty + \int_A^\infty e^{-x^2/2}dx\Big) = \frac{A}{\sqrt{2\pi}}e^{-A^2/2} + 1 - \Phi(A)$$

sowie die Abschätzung $1 - \Phi(A) \leqslant \frac{1}{\sqrt{2\pi}A}e^{-A^2/2}$ für $A > 1$, so dass wegen $\kappa \geqslant 2$

$$T_1 \leqslant \frac{\sigma^2}{n}\,\mathbb{E}_f[\zeta_k^2 \mathbf{1}(|\zeta_k| \geqslant \kappa - 1)] \leqslant \frac{\sigma^2}{n}\frac{\kappa - 1 + (\kappa-1)^{-1}}{\sqrt{2\pi}} e^{-(\kappa-1)^2/2} \leqslant \frac{\sigma^2}{n}\frac{\kappa}{\sqrt{2\pi}} e^{-(\kappa-1)^2/2}$$

gilt. Wir schließen in ähnlicher Weise

$$T_3 \leqslant f_k^2 P(|\zeta_k| > \kappa\tfrac{\sigma}{\sqrt{n}}) \leqslant f_k^2 \frac{\kappa^{-1}}{\sqrt{2\pi}} e^{-\kappa^2/2}.$$

In Summe ergibt sich also

$$\mathbb{E}_f[(\hat{f}_k^{(hard)} - f_k)^2] \leqslant (1+4\kappa^2)(f_k^2 \wedge \tfrac{\sigma^2}{n}) + \big(\tfrac{\sigma^2}{n} + f_k^2\big)\tfrac{\kappa}{\sqrt{2\pi}} e^{-(\kappa-1)^2/2}.$$

Für den Orakelschätzer gilt ja $\mathbb{E}_f[(\hat{f}_k^{(Orakel)} - f_k)^2] = f_k^2 \wedge \frac{\sigma^2}{n}$, so dass Summation über $k \in \{1, \dots, K\}$ ergibt

$$\mathbb{E}_f[\|\hat{f}^{(hard)} - f\|_{L^2}^2] \leqslant (1+4\kappa^2)\,\mathbb{E}[\|\hat{f}^{(Orakel)} - f\|_{L^2}^2] + \big(K\tfrac{\sigma^2}{n} + \|f\|_{L^2}^2\big)\tfrac{\kappa}{\sqrt{2\pi}} e^{-(\kappa-1)^2/2},$$

wie behauptet. Einsetzen liefert das zweite Resultat. □

Die Wahl der *universal threshold* κ garantiert insbesondere, dass der Restterm in der Orakelungleichung in n und K von kleinerer Ordnung als $o(\frac{\sigma^2}{n} K^p)$ für alle $p > 0$ ist, so dass im wesentlichen die Größenordnung des parametrischen Fehlers $\frac{\sigma^2}{n}$ gilt, der für nichtparametrische Probleme vernachlässigbar ist. Beachte, dass im asymptotisch äquivalenten Regressionsmodell n Beobachtungen vorliegen und daher $K = n$ eine kanonische Wahl ist. Im folgenden werden wir daher sehen, dass der Hard-Thresholding-Schätzer Minimaxraten bis auf einen logarithmischen Faktor erreicht.

Der Hard-Thresholding-Schätzer ergibt sich äquivalent auch durch das Minimierungsproblem ▶ Übung

$$\hat{f}^{(hard)} = \operatorname{argmin}_{g \in \operatorname{span}(\varphi_1,\dots,\varphi_k)} \Big(\underbrace{\sum_{k=1}^{K} (y_k - g_k)^2}_{\text{empirisches Risiko}} + \underbrace{\kappa^2 \tfrac{\sigma^2}{n} |\{k \in \{1,\dots,K\} \mid g_k \neq 0\}|}_{\ell^0\text{-Penalisierung}} \Big).$$

Diese Idee der Minimerung des empirischen Risikos plus eines Strafterms, der eine spärliche Darstellung bewirkt, wird auch für viele komplexe Schätzprobleme, verwendet. Allerdings führt ℓ^0-Penalisierung nicht auf ein konvexes Minimierungsproblem in den Koeffizienten und ist daher oft nicht einfach berechenbar (NP-hartes Problem). Als sogenannte *konvexe Relaxation* erhält man die ℓ^1-Penalisierung

$$\hat{f}^{(soft)} = \operatorname{argmin}_{g \in \operatorname{span}(\varphi_1,\dots,\varphi_k)} \Big(\underbrace{\sum_{k=1}^{K} (y_k - g_k)^2}_{\text{empirisches Risiko}} + \underbrace{\kappa \tfrac{\sigma}{\sqrt{n}} \sum_{k=1}^{K} |g_k|}_{\ell^1\text{-Penalisierung}} \Big).$$

Hier lassen sich die Koeffizienten der Lösung wiederum auch explizit angeben▶ Übung :

$$\hat{f}_k^{(soft)} = (y_k - \kappa\tfrac{\sigma}{\sqrt{n}})_+ - (y_k + \kappa\tfrac{\sigma}{\sqrt{n}})_-.$$

Der Schätzer heißt *Soft-Thresholding-Schätzer*. Auch unter statistischen Gesichtspunkten ergeben sich Vorteile des Soft-Thresholding. Insbesondere ist der Schätzer stetig in den Daten und im Schwellwert κ, so dass geringe Schwankungen in den Daten oder in κ zu ähnlichen Ergebnissen führen, was beim Hard-Thresholding nicht notwendigerweise der Fall ist. Die statistische Theorie beider Thresholding-Verfahren ist sehr ähnlich, auch beim Soft-Thresholding ergibt sich eine entsprechende Orakelungleichung mit logarithmischem Faktor unter der *universal threshold* $\kappa = \sqrt{2 \log K}$.

Für die sparsame (*sparse*) Darstellung von Funktionen sind Wavelet-Basen besonders gut geeignet. Wir führen die Theorie anhand der einfachen Haarwavelets vor, die Resultate lassen sich aber alle auf allgemeine Waveletbasen (ψ_{jk}) verallgemeinern, siehe z.B. Härdle, Kerkyacharian, Picard, and Tsybakov (1998). Für eine anwendungsorientierte Übersicht und Diskussion von Waveletmethoden in der Statistik sei (Nason 2008) empfohlen.

5.25 Definition. Setze $\varphi(x) := \mathbf{1}_{[0,1]}(x)$, $\psi(x) = \mathbf{1}_{[0,1/2]}(x) - \mathbf{1}_{[1/2,1]}(x)$ sowie für $j \in \mathbb{N}_0$, $k = 0, \dots, 2^j - 1$

$$\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k) = 2^{j/2}(\mathbf{1}_{[k2^{-j},(k+1/2)2^{-j}]}(x) - \mathbf{1}_{[(k+1/2)2^{-j},(k+1)2^{-j}]}).$$

Man nennt (ψ_{jk}) *Haar-Wavelets*, ψ *Mutter-Wavelet* und φ *Skalierungsfunktion*. Mit der Notation $\psi_{-1,k} := \varphi$ bildet $(\psi_{jk})_{j \geqslant -1,k}$ eine Orthonormalbasis in $L^2([0,1])$, die *Haar-Wavelet-Basis*.

Weiterhin betrachten wir die Approximationsräume $V_J := \operatorname{span}(\psi_{j,k}, -1 \leqslant j < J, k = 0, \dots, 2^j - 1)$, $J \in \mathbb{N}$, sowie die Orthogonalprojektionen $\Pi_J : L^2([0,1]) \to V_J$.

Die Approximationsräume V_J lassen sich durch

$$V_J = \{f \in L^2([0,1]) \,|\, f \text{ ist f.ü. konstant auf } [k2^{-J}, (k+1)^{-J}],\ k = 0, \dots, 2^J - 1\}$$

beschreiben und besitzen die Dimension 2^J. Anhand einfacher Beispiele wollen wir die Approximationseigenschaften der (Haar-)Wavelets ergründen.

5.26 Beispiele.

(a) Betrachte $f \in C^\alpha([0,1])$ mit $\alpha \in (0,1]$ und Hölderkonstanten $L > 0$. Wir erhalten:

$$|\langle f, \psi_{jk}\rangle_{L^2}| = 2^{j/2}\left|\int_{k2^{-j}}^{(k+1/2)2^{-j}} (f(x) - f(x + 2^{-j-1}))\, dx\right| \leqslant L 2^{-j(\alpha+1/2)}.$$

Damit folgt als (linearer) Approximationsfehler

$$\|f - \Pi_J f\|_{L^2}^2 = \sum_{j \geqslant J,k} \langle f, \psi_{jk}\rangle^2 \leqslant L^2 \sum_{j \geqslant J} 2^j 2^{-j(2\alpha+1)} = 2L^2 2^{-2\alpha J}.$$

Bezüglich der Dimension von V_J ergibt sich die Schranke $2L^2 \dim(V_J)^{-2\alpha}$. Dies hatten wir auf anderem Wege bereits in Beispiel 3.22(a) gesehen.

(b) Betrachte die einfache Sprungfunktion $f(x) = \mathbf{1}_{[0,s]}(x)$ mit $s \in (0,1)$. Dann gilt offenbar $\langle f, \psi_{jk}\rangle = 0$ im Fall $s \notin (k2^{-j}, (k+1)2^{-j})$ sowie $|\langle f, \psi_{jk}\rangle| \leqslant \|f\|_\infty \|\psi_{jk}\|_{L^1} = 2^{-j/2}$ im Fall $s \in (k2^{-j}, (k+1)2^{-j})$. Mit $k_j := \lfloor s2^j \rfloor$ ergibt sich also kurz $|\langle f, \psi_{jk}\rangle| \leqslant 2^{-j/2}\mathbf{1}(k = k_j)$.

Als linearen Approximationsfehler erhalten wir

$$\|f - \Pi_J f\|_{L^2}^2 \leqslant \sum_{j \geqslant J,k} \langle f, \psi_{jk}\rangle^2 \leqslant \sum_{j \geqslant J} 2^{-j} = 2^{-J+1}.$$

Dies ist dieselbe Rate wie für $f \in C^\alpha([0,1])$ mit $\alpha = 1/2$. In der Tat kann man zeigen, dass die Sprungfunktion jede Glattheit $s < 1/2$ im L^2-Sobolev-Sinn besitzt (bestimme z.B. die Fourierkoeffizienten), allerdings ist die Größe der Wavelet-Koeffizienten grundverschieden vom Hölderfall, weil auf jedem Niveau j nur ein Koeffizient die Ordnung $2^{-j/2}$ besitzt, anstatt dass alle Koeffizienten die maximale Ordnung 2^{-j} besitzen. Die Sprungfunktion hat also bedeutend weniger signifikante Wavelet-Koeffizienten. Falls wir diese Koeffizienten kennen würden, so könnten wir mit entsprechenden N Basisfunktionen einen sehr viel kleineren (nichtlinearen) Approximationsfehler erreichen (*beste N-Term-Approximation*):

$$\Big\| f - \sum_{-1 \leqslant j \leqslant N-2} \langle f, \psi_{j,k_j}\rangle \psi_{j,k_j} \Big\|_{L^2}^2 = \sum_{j \geqslant N-1} \langle f, \psi_{j,k_j}\rangle^2 = 2^{-N+2}.$$

Bei analoger Wahl $N = \dim(V_J) = 2^J$ wie im linearen Projektionsfall wäre der Fehler also äußerst klein (hyperexponentiell in J).

(c) Betrachte nun eine Funktion f, die α-Hölder-stetig mit $\alpha \geqslant 1/2$ ist, jenseits von einer Sprungstelle $s \in (0,1)$. Dann können wir f als Linearkombination der Funktionen aus (a) und (b) schreiben. Die Größenordnung der Approximationsfehler ergibt sich als Summe aus (a) und (b), d.h $\|f - \Pi_J f\|_{L^2}^2 = O(2^{-J})$, aber $\|f - \tilde{f}_{2^J}\|_{L^2}^2 = O(2^{-2\alpha J})$ für die beste N-Term-Approximation $\tilde{f}_N$. Die lokale Nichtregularität von f bei s ist also unwesentlich für den Approximationsfehler. Wie wir sehen werden, können Funktionen mit lokalen Irregularitäten (wie Sprüngen oder Spitzen) in L^p-Normen bzw. in ℓ^p-Normen für ihre Koeffizienten mit $p < 2$ recht gut beschrieben werden.

5.27 Definition. Eine Funktion $f \in L^p([0,1])$ liegt im *Besovraum* $B^s_{pq}([0,1])$ mit $s > 0$, $p, q \in [1, \infty]$, falls für den L^p-Stetigkeitsmodul

$$\omega_n(f,h)_p := \sup_{0 \leqslant y \leqslant h} \|\Delta_y^n f\|_{L^p([0,1-h])}, \quad \Delta_y f(x) := f(x+y) - f(x),$$

mit $n = \lfloor s \rfloor + 1$ und $\Delta_y^n f := \Delta_y^{n-1}(\Delta_y f)$ gilt $(2^{sj}\omega_n(f, 2^{-j})_p)_{j \geqslant 0} \in \ell^q$. Der Besovraum wird normiert durch

$$\|f\|_{s,p,q} = \|f\|_{L^p} + \Big\| (2^{sj}\omega_n(f, 2^{-j})_p)_{j \geqslant 0} \Big\|_{\ell^q}.$$

5.28 Bemerkung. Man kann zeigen, dass $B^s_{pq}([0,1])$ ein Banachraum ist. Für $p = q = 2$ und $s \notin \mathbb{N}$ ergibt sich der L^2-Sobolevraum $B^s_{2,2}([0,1]) = H^s([0,1])$ (definiert z.B. durch Restriktion von $H^s(\mathbb{R})$) sowie für $p = q = \infty$ und $s \notin \mathbb{N}$ der entsprechende Hölderraum $B^s_{\infty,\infty}([0,1]) = C^s([0,1])$. Mit der Besovnorm wird also die Glattheit einer Funktion im L^p-Sinn gemessen, der Parameter q ist eher unwichtig und liefert nur noch ein *Fine-tuning*.

5.29 Lemma. *Die Haar-Wavelet-Koeffizienten einer Funktion $f \in B^s_{p,q}([0,1])$ erfüllen im Fall $s \in (0,1)$*

$$\left\| \left(2^{j(s+1/p-1/2)} \left\| (\langle f, \psi_{jk} \rangle)_k \right\|_{\ell^p} \right)_{j \geqslant 0} \right\|_{\ell^q} \leqslant \|f\|_{s,p,q}.$$

Insbesondere gilt also

$$\sum_{j \geqslant 0} 2^{j(s+1/p-1/2)p} \sum_{k=0}^{2^j-1} |\langle f, \psi_{jk} \rangle|^p \leqslant \|f\|^p_{s,p,p}.$$

Beweis. Wegen $s < 1$ betrachten wir $n = 1$ und erhalten für jedes $j \geqslant 0$ mit der Jensenschen Ungleichung

$$\begin{aligned}
\sum_{k=0}^{2^j-1} |\langle f, \psi_{jk} \rangle|^p &= \sum_{k=0}^{2^j-1} \left| 2^{j/2} \int_{k2^{-j}}^{(k+1/2)2^{-j}} (f(x) - f(x + 2^{-j-1}))\, dx \right|^p \\
&\leqslant \sum_{k=0}^{2^j-1} 2^{-jp/2} 2^j \int_{k2^{-j}}^{(k+1/2)2^{-j}} |f(x) - f(x + 2^{-j-1})|^p\, dx \\
&\leqslant 2^{j(1-p/2)} \omega_1(f, 2^{-j-1})_p.
\end{aligned}$$

Dies impliziert $2^{j(1/2-1/p)} (\sum_{k=0}^{2^j-1} |\langle f, \psi_{jk} \rangle|^p)^{1/p} \leqslant \omega_1(f, 2^{-j+1})_p$, und Einsetzen in die Definitionen liefert das Ergebnis. □

5.30 Bemerkung. Auch mit einem geeigneten Faktor multipliziert, kann die Ungleichung nicht immer auch in die andere Richtung gelten, da eine Funktion mit nur endlich vielen Haar-Koeffizienten ungleich Null eine Sprungfunktion ist und nicht in allen B^s_{pq} mit $s \in (0,1)$ liegt. Sogenannte reguläre Wavelet-Basen erlauben jedoch gerade eine Normäquivalenz zwischen den Besovnormen und den wie oben gewichteten Folgenraumnormen in den Koeffizienten, vergleiche z.B. (Wojtaszczyk 1997, Cor. 9.10). Im Zusammenhang mit dem Folgenraummodell erlaubt dies eine gewichtete ℓ^p-Analyse der Schätzprobleme, wobei der Fall $p < 2$ gerade *sparsity* codiert.

5.31 Beispiele.

(a) Betrachte die Sprungfunktion $f(x) = \mathbf{1}_{[0,a]}(x)$, $a \in (0,1)$. Dann gilt $\omega_1(f,h)^p_p = \int_{a-h}^a 1^p dx = h$ für $h \in (0,a]$ und daher $2^{js}\omega_1(f, 2^{-j}) \in \ell^\infty$

genau für $s \leqslant 1/p$. Wir schließen $f \in B^{1/p}_{p,\infty}([0,1])$ (für $p=1$ benötigen wir dazu allerdings die ebenfalls korrekte Abschätzung für $\omega_2(f,h)_p$) sowie $f \in B^s_{p,q}([0,1])$ für $s < 1/p$ im Fall $q < \infty$.

(b) Betrachte $f(x) = x^\alpha$ mit $\alpha \in (0,1]$. Eine klassische Analyse beruht dann auf der Beobachtung, dass $f'(x) = \alpha x^{\alpha-1}$ in $L^p([0,1])$ für $p < (1-\alpha)^{-1}$ liegt. Im Sinne der L^p-Sobolevräume $W^{m,p}$ der Regularität $m \in \mathbb{N}$ gilt also $f \in W^{1,p}$ für alle $p < (1-\alpha)^{-1}$, insbesondere $f \in W^{1,2} = H^1$ für $\alpha > 1/2$. In der Skala der Besovräume mit $s < 1$ schätzen wir den L^p-Stetigkeitsmodul mittels $\Delta_y f(x) = \int_x^{x+y} f'$ ab:

$$\begin{aligned}\omega_1(f,h)_p^p &\leqslant \sup_{0\leqslant y\leqslant h} \int_0^{1-h} \Big(\int_x^{x+y} \alpha z^{\alpha-1} dz \Big)^p dx \\ &\leqslant \int_0^h h^{\alpha p} dx + \int_h^1 (\alpha x^{\alpha-1} h)^p \, dx \leqslant h^{1+\alpha p} + \tfrac{\alpha^p}{(1-\alpha)p} h^{1+\alpha p}.\end{aligned}$$

Somit gilt $2^{js}\omega_1(f,2^{-j})_p \leqslant C 2^{j(s-\alpha-1/p)}$ mit einer Konstanten $C > 0$. Wir schließen also $2^{js}\omega_1(f,2^{-j})_p \in \ell^\infty$ für $s \leqslant \alpha + 1/p$ sowie $2^{js}\omega_1(f,2^{-j})_p \in \ell^q$ mit $q < \infty$ für $s < \alpha + 1/p$. Dasselbe Resultat folgt analog auch für $\omega_n(f,2^{-j})_p$ mit $n \geqslant 2$, so dass $f \in B^{\alpha+1/p}_{p,\infty}$ für alle $p \in [1,\infty]$ gilt. Im Fall $p = \infty$ erhalten wir gerade die Glattheit α als Hölder-Regularität, im Fall $p = 2$ (L^2-Sobolev-Klasse) gewinnen wir jedoch eine halbe Glattheitsordnung, im Fall $p = 1$ sogar eine ganze Glattheitsordnung. Für $q < \infty$ erreichen wir mit $s < \alpha + 1/p$ jeweils nicht ganz diese Glattheitsordnung, vergleiche auch die ▶ Übung entsprechenden Einbettungen der Besovräume. Wie wir oben gesehen haben, spielt für unseren statistischen Schätzfehler im MISE bei der Projektionsmethode gerade die L^2-Sobolevklasse die entscheidende Rolle. Durch Verwendung von Wavelet-Thresholding-Schätzern können wir sogar L^p-Glattheiten mit $p < 2$ ausnutzen.

5.32 Satz. *Betrachte den Orakel-Thresholding-Schätzer im Folgenraummodell bezüglich der Haar-Wavelet-Basis und mit $K = 2^J \geqslant n/\sigma^2$. Dann gilt für $f \in B^s_{pp}([0,1])$ mit $p \in [1,2]$ und $s - \frac{1}{p} \geqslant \frac{s}{2s+1} - \frac{1}{2}$*

$$\mathbb{E}_f[\|\hat f^{(Orakel)} - f\|_{L^2}^2] \leqslant C_1 \max\big(\|f\|_{s,p,p}^2, \|f\|_{s,p,p}^{2/(2s+1)}\big)\Big(\frac{\sigma^2}{n}\Big)^{2s/(2s+1)}$$

mit einer Konstanten $C_1 > 0$. Für den entsprechenden Hard-Thresholding-Schätzer mit der universal threshold κ und der zusätzlichen Eigenschaft $K = 2^J \leqslant c_0 n/\sigma^2$ für ein $c_0 \in [1,\infty)$ folgt daraus

$$\mathbb{E}_f[\|\hat f^{(hard)} - f\|_{L^2}^2] \leqslant C_2 \max\big(\|f\|_{s,p,p}^2, \|f\|_{s,p,p}^{2/(2s+1)}\big)\Big(\frac{\sigma^2 \log(n/\sigma^2)}{n}\Big)^{2s/(2s+1)}$$

mit einer Konstanten $C_2 > 0$.

Beweis. Nach Lemma 5.29 gilt für die Waveletkoeffizienten $f_{jk} = \langle f, \psi_{jk}\rangle$

$$\sum_{j\geqslant 0} 2^{s(j+1/p-1/2)p} \sum_{k=0}^{2^j-1} |f_{jk}|^p \leqslant \|f\|_{s,p,p}^p.$$

Wähle nun den Index $J_s \in \mathbb{N}$ mit $J_s \leqslant J$ so, dass 2^{J_s} von der Größenordnung $(n\|f\|_{s,p,p}^2/\sigma^2)^{1/(2s+1)}$ ist, was der raten-optimalen Wahl bei der Projektionsmethode entspricht. Dann ergeben sich mittels obiger ℓ^p-Abschätzung drei verschiedene Zonen in j bei der Risikoabschätzung:

$$\begin{aligned}
&\mathbb{E}_f[\|\hat{f}^{(Orakel)} - f\|_{L^2}^2] = \sum_{j<J,k} f_{jk}^2 \wedge \tfrac{\sigma^2}{n} + \sum_{j\geqslant J,k} f_{jk}^2 \\
&\leqslant 2^{J_s}\tfrac{\sigma^2}{n} + \sum_{J_s\leqslant j<J,k} |f_{jk}|^p \tfrac{\sigma^{2-p}}{n^{1-p/2}} + \Big(\sum_{j\geqslant J,k} |f_{jk}|^p\Big)^{2/p} \\
&\leqslant c_1\|f\|_{s,p,p}^{2/(2s+1)}\big(\tfrac{\sigma^2}{n}\big)^{2s/(2s+1)} + \tfrac{\sigma^{2-p}}{n^{1-p/2}}\|f\|_{s,p,p}^p 2^{-2J_s(s+1/2-1/p)p} + \|f\|_{s,p,p}^2 2^{-2J(s+1/2-1/p)} \\
&\leqslant c_2\Big(\|f\|_{s,p,p}^{2/(2s+1)}\big(\tfrac{\sigma^2}{n}\big)^{2s/(2s+1)} + \|f\|_{s,p,p}^{2/(2s+1)}\big(\tfrac{\sigma^2}{n}\big)^{2s/(2s+1)} + \|f\|_{s,p,p}^2\big(\tfrac{\sigma^2}{n}\big)^{2(s+1/2-1/p)}\Big)
\end{aligned}$$

mit Konstanten $c_1, c_2 > 0$. Wegen $s + 1/2 - 1/p \geqslant s/(2s+1)$ folgt daher die behauptete Abschätzung für den Orakelschätzer. Mit der Orakelungleichung aus Satz 5.24 impliziert dies sofort die Abschätzung für den Hard-Thresholding-Schätzer. □

5.33 Beispiel. Eine Lipschitz-stetige Funktion f mit endlich vielen Sprungstellen liegt in $B_{p,p}^s([0,1])$ für $s < 1/p$, weil Sprungfunktionen in diesem Raum liegen, ebenso Lipschitz-Funktionen wegen ► Übung $B_{\infty,\infty}^1([0,1]) \subseteq B_{p,p}^s([0,1])$, $s < 1$, und f als Linearkombination dieser Funktionen im selben Besovraum liegt. Nach obigem Satz hat der Wavelet-Hard-Thresholding-Schätzer demnach die Rate $(n/\log n)^{-2s/(2s+1)}$ für alle $s < 1/p$, sofern $s - \frac{1}{p} \geqslant \frac{s}{2s+1} - \frac{1}{2}$ erfüllt ist. Für $p = 1$ können wir jedoch gerade $s \in (1-\varepsilon, 1)$ mit $\varepsilon > 0$ hinreichend klein wählen, so dass die letzte Bedingung erfüllt ist. Diese Argumentation über Besov-Glattheiten zeigt, dass wir die Rate $(n/\log n)^{-2/3+\varepsilon'}$ für jedes $\varepsilon' > 0$ erreichen, was fast der Rate für Lipschitz-Funktionen ohne Sprünge entspricht. Ein lineares Verfahren wie Projektionsschätzung hätte demgegenüber nur die L^2-Sobolev-Glattheit $s < 1/2$ ausgenutzt, was zu einer Rate der Ordnung $n^{-1/2}$ im MISE führt. Für solche Funktionen ergibt Wavelet-Thresholding also bedeutende Effizienzgewinne. Andere nichtlineare Schätzverfahren wie lokale Bandweitenwahl bei Kernschätzern oder adaptive Knotenwahl bei Spline-Schätzern erlauben ähnliche Resultate, jedoch mit aufwändigeren Beweisen. Es sei hier betont, dass sich die Einführung der Besov-Glattheitsklassen als ein gutes allgemeines Konzept bewährt hat, diese Konvergenzraten abzuschätzen ; unsere detaillierte Einzelanalyse für Waveletkoeffizienten von Sprungfunktionen zeigt nur noch, dass wir sogar genau die Rate $(n/\log n)^{-2/3}$ erreichen.

5.5 Lepski-Methode

Ein sequentielles Testverfahren ist das Herz der Lepski-Methode. Sie findet als adaptives Schätzverfahren von Funktionalen Anwendung und wurde ursprünglich zur lokalen Bandweitenwahl für Kernschätzer entwickelt. Die fundamentale Idee ist, aus vorliegenden reellwertigen Schätzern $\tilde{\vartheta}_0, \tilde{\vartheta}_1, \ldots, \tilde{\vartheta}_K$, deren Varianz mit wachsendem Index fällt und deren Bias im Schnitt wächst, sequentiell die Hypothesen $H_k : \tilde{\vartheta}_0 = \tilde{\vartheta}_1 = \cdots = \tilde{\vartheta}_k$ zu testen. Sind $H_0, H_1, \ldots, H_k$ akzeptiert, weicht jedoch $\tilde{\vartheta}_{k+1}$ signifikant von $\tilde{\vartheta}_0, \tilde{\vartheta}_1, \ldots, \tilde{\vartheta}_k$ ab, so lehne H_{k+1} ab und wähle $\hat{k} = k$. Der Schätzer $\hat{\vartheta} := \tilde{\vartheta}_{\hat{k}}$ sollte dann unter den Schätzern $\tilde{\vartheta}_k$, die einen quadrierten Bias von kleinerer Größenordnung als die Varianz besitzen, die kleinste Varianz besitzen. Natürlich ist die mathematische Analyse von $\hat{\vartheta}$ komplexer, und es bedarf einer genauen Festlegung der Tests und ihrer kritischen Werte.

5.34 Definition. Als konkretes Modell betrachten wir das Regressionsproblem

$$Y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

mit (ε_i) i.i.d., $\mathbb{E}[\varepsilon_i] = 0$, $\mathbb{E}[\varepsilon_i^2] < \infty$ und Design $(x_i) \subseteq \mathbb{R}^d$. Die Verteilung von ε_i sei bekannt. Als Schätzer von f an der Stelle $x_0 \in \mathbb{R}^d$ verwende für Bandweiten $h_0 < h_1 < \cdots < h_K$ die Nadaraja-Watson-Schätzer

$$\tilde{\vartheta}_k := \frac{\sum_{i=1}^n Y_i K_{h_k}(x_0 - x_i)}{\sum_{i=1}^n K_{h_k}(x_0 - x_i)},$$

wobei eine Kernfunktion $K : \mathbb{R}^d \to [0, \infty)$ mit Träger in $\{x \in \mathbb{R}^d \mid |x| \leqslant 1\}$ gewählt sei. Setze $s_j^2 := \mathrm{Var}(\tilde{\vartheta}_j)$, $s_{kj}^2 := \mathrm{Var}(\tilde{\vartheta}_k - \tilde{\vartheta}_j)$ und nimm $\infty > s_0 > s_1 > \cdots > s_K$ an.

Für positive kritische Werte z_k, $k = 0, \ldots, K-1$, $z_K := 1$ bestimme den Schätzer $\hat{\vartheta} := \tilde{\vartheta}_{\hat{k}}$ gemäß der *Lepski-Methode* als

$$\hat{k} := \inf \left\{ k = 0, \ldots, K \,\middle|\, \exists j \leqslant k : \ |\tilde{\vartheta}_{k+1} - \tilde{\vartheta}_j| > z_j s_{k+1,j} + z_{k+1} s_{k+1} \right\} \wedge K.$$

Folgender Algorithmus beschreibt die Lepski-Methode iterativ:

- `initialisiere` $k := 0$;
- `wiederhole`
  `falls` $\forall j = 0, \ldots, k : \ |\tilde{\vartheta}_{k+1} - \tilde{\vartheta}_j| \leqslant z_j s_{k+1,j} + z_{k+1} s_{k+1}$,
  `erhöhe` k `um eins`,
  `sonst stoppe`;
  `solange, bis` $k = K$;
- `Setze` $\hat{k} := k$.

Zur Analyse der Lepski-Methode vergleichen wir $\hat{\vartheta}$ mit einem Orakel-Typ-Schätzer $\tilde{\vartheta}_{k^*}$. Der Fehler durch zu spätes Stoppen $\{\hat{k} > k^*\}$ wird direkt durch die Methode beschränkt, während zu frühes Stoppen $\{\hat{k} < k^*\}$ durch eine geeignete Wahl der kritischen Werte kontrolliert wird.

5.35 Lemma. *Für jedes $k^* = 0, 1, \dots, K-1$ gilt*

$$|\hat{\vartheta} - \tilde{\vartheta}_{k^*}| \mathbf{1}(\hat{k} > k^*) \leqslant \max_{k^* \leqslant j \leqslant K-1} (z_{k^*} s_{j+1,k^*} + z_{j+1} s_{j+1}).$$

Beweis. Die Definition der Lepski-Methode impliziert, dass aus $\hat{k} > k^*$ insbesondere

$$|\hat{\vartheta} - \tilde{\vartheta}_{k^*}| \leqslant z_{k^*} s_{\hat{k},k^*} + z_{\hat{k}} s_{\hat{k}}$$

folgt. Die rechte Seite wird durch das Maximum wegen $\hat{k} \in \{k^*+1, \dots, K\}$ abgeschätzt. □

5.36 Bemerkung. Die Ungleichung ist für alle Realisierungen von (Y_i) per Konstruktion wahr, nicht nur fast sicher. Wir sehen, dass der Fehler durch spätes Stoppen umso kleiner ist, je kleiner die kritischen Werte (z_k) gewählt sind.

5.37 Definition. Wir sagen, dass die kritischen Werte *unter der Null α-kalibriert* sind, falls für einen Konfidenzparameter $\alpha > 0$ gilt

$$\sum_{j=0}^{K-1} \mathbb{E}_0 \left[\tilde{\vartheta}_j^2 \mathbf{1}\big(\exists l \leqslant j : |\tilde{\vartheta}_{j+1} - \tilde{\vartheta}_l| > z_l s_{j+1,l}\big) \right] \leqslant \alpha s_K^2.$$

Eine kanonische Wahl zur Kalibrierung unter der Null erfolgt iterativ. Wähle zunächst $z_0 > 0$ so, dass

$$\sum_{j=0}^{K-1} \mathbb{E}_0 \left[\tilde{\vartheta}_j^2 \mathbf{1}\big(|\tilde{\vartheta}_{j+1} - \tilde{\vartheta}_0| > z_0 s_{j+1,0}\big) \right] \leqslant \frac{\alpha}{K} s_K^2$$

gilt. Für gegebene $z_0, \dots, z_{k-1}$ wähle dann $z_k > 0$ gemäß

$$\sum_{j=k}^{K-1} \mathbb{E}_0 \left[\tilde{\vartheta}_j^2 \mathbf{1}\big(|\tilde{\vartheta}_{j+1} - \tilde{\vartheta}_k| > z_k s_{j+1,k}, \forall l < k : |\tilde{\vartheta}_{j+1} - \tilde{\vartheta}_l| \leqslant z_l s_{j+1,l}\big) \right] \leqslant \frac{\alpha}{K} s_K^2.$$

Eine Summation dieser Ungleichungen ergibt genau die geforderte Kalibrierungsbedingung. In der Praxis bestimmt man die kritischen Werte, indem man die Lepski-Methode auf den Fall reinen Rauschens ($f = 0$) anwendet, jedoch den Term $z_{j+1} s_{j+1}$ in den Tests jeweils weglässt. Der Orakel-Index ist natürlich $k^* = K$, und z_k wird so gewählt, dass der Fehler durch Stoppen wegen einer Abweichung von $\tilde{\vartheta}_k$ nur das Vielfache $\frac{\alpha}{K}$ des Orakelfehlers s_K^2 beträgt. Später leiten wir eine asymptotische Größenordnung der (z_k) her.

5.38 Definition. Für die Regressionsfunktion f und den Punkt $x_0 \in \mathbb{R}^d$ definiere die *lokale Variation*

$$V_k(f) := \sup_{y_1,y_2 \in \{y:|y-x_0| \leqslant h_k\}} |f(y_1) - f(y_2)|$$

und betrachte den *Orakel-Typ-Index*

$$k^* := \inf\{k = 0, 1, \ldots, K-1 \,|\, V_{k+1}(f) > z_{k+1} s_{k+1}\} \wedge K.$$

Wir interpretieren k^* als Orakel, weil der durch V_k abgeschätzte Bias in $\tilde{\vartheta}_k$ dort gerade noch kleiner ist als der stochastische Fehler multipliziert mit dem kritischen Wert. Die Wahl $z_K = 1$ balanciert gerade Bias und stochastischen Fehler beim Index $k^* = K-1$, vorher wird z_k im Allgemeinen größer als Eins sein.

5.39 Satz. *Für den Orakel-Typ-Index k^* gilt bei α-Kalibrierung unter der Null*

$$\mathbb{E}_f[(\hat{\vartheta} - \tilde{\vartheta}_{k^*})^2 \mathbf{1}(\hat{k} < k^*)] \leqslant 2(z_{k^*}^2 + \alpha) s_{k^*}^2.$$

Beweis. Wir werden $\hat{k}(f), \tilde{\vartheta}_k(f)$ etc. schreiben, um die Abhängigkeit von der Regressionsfunktion zu kennzeichnen. Setzen wir $w_i^{(k)} := \frac{K_{h_k}(x-X_i)}{\sum_j K_{h_k}(x-X_j)}$, so gilt $w_i^{(k)} \geqslant 0$ (wegen $K \geqslant 0$), $\sum_{i=1}^n w_i^{(k)} = 1$ sowie $w_i^{(k)} = 0$ für $|x_i - x_0| > h_k$ (wegen $\operatorname{supp} K \subseteq \{x \in \mathbb{R}^d \,|\, |x| \leqslant 1\}$). Wir schließen für $j < k$

$$\begin{aligned}
|\tilde{\vartheta}_j(f) - \tilde{\vartheta}_k(f)| &\leqslant |\tilde{\vartheta}_j(0) - \tilde{\vartheta}_k(0)| + \Big|\sum_{i=1}^n (w_i^{(j)} - w_i^{(k)}) f(x_i)\Big| \\
&\leqslant |\tilde{\vartheta}_j(0) - \tilde{\vartheta}_k(0)| + \max_{i,i':|x_i-x_0|,|x_{i'}-x_0| \leqslant h_k} |f(x_i) - f(x_{i'})| \\
&\leqslant |\tilde{\vartheta}_j(0) - \tilde{\vartheta}_k(0)| + V_k(f).
\end{aligned}$$

Daher erhalten wir mit $(A+B)^2 \leqslant 2A^2 + 2B^2$ die Abschätzung

$$\begin{aligned}
&\mathbb{E}[(\hat{\vartheta}(f) - \tilde{\vartheta}_{k^*}(f))^2 \mathbf{1}(\hat{k}(f) < k^*)] \\
&\leqslant 2V_{k^*}(f) + 2\,\mathbb{E}[(\hat{\vartheta}(0) - \tilde{\vartheta}_{k^*}(0))^2 \mathbf{1}(\hat{k}(f) < k^*)] \\
&\leqslant 2V_{k^*}(f) + \\
&\quad 2\sum_{j=0}^{k^*-1} \mathbb{E}\Big[(\tilde{\vartheta}_j(0) - \tilde{\vartheta}_{k^*}(0))^2 \mathbf{1}\big(\exists l \leqslant j : |\tilde{\vartheta}_{j+1}(f) - \tilde{\vartheta}_l(f)| > z_l s_{j+1,l} + z_{j+1} s_{j+1}\big)\Big] \\
&\leqslant 2z_{k^*}^2 s_{k^*}^2 + 2\sum_{j=0}^{k^*-1} \mathbb{E}\Big[(\tilde{\vartheta}_j(0) - \tilde{\vartheta}_{k^*}(0))^2 \mathbf{1}\big(\exists l \leqslant j : |\tilde{\vartheta}_{j+1}(0) - \tilde{\vartheta}_l(0)| > z_l s_{j+1,l}\big)\Big] \\
&\leqslant 2z_{k^*}^2 s_{k^*}^2 + 2\alpha s_K^2,
\end{aligned}$$

wobei wir in der vorletzten Zeile die Definition von k^* und in der letzten Zeile die α-Kalibrierung unter der Null verwendet haben. Die Behauptung folgt nun aus $s_{k^*} \geqslant s_K$. □

5.40 Korollar. *Die kritischen Werte (z_k) seien α-kalibriert unter der Null und so gewählt, dass $(z_k s_k)$ monoton fallend in k ist. Dann folgt mit dem Orakel-Typ-Index k^**

$$\mathbb{E}_f[(\hat{\vartheta} - \tilde{\vartheta}_{k^*})^2] \leqslant \big(11 z_{k^*}^2 + 2\alpha\big) s_{k^*}^2.$$

Beweis. Aus Lemma 5.35 folgt

$$\mathbb{E}_f[(\hat{\vartheta} - \tilde{\vartheta}_{k^*})^2 \mathbf{1}(\hat{k} > k^*)] \leqslant \max_{k^* \leqslant j \leqslant K-1} (z_{k^*} s_{j+1,k^*} + z_{j+1} s_{j+1})^2.$$

Es gilt nun stets $s_{jk}^2 \leqslant 2s_k^2 + 2s_j^2 \leqslant 4s_k^2$ für $j > k$, so dass wir unter Benutzung der Monotonieannahme die rechte Seite mit $(3 z_{k^*} s_{k^*})^2$ weiter abschätzen können. Dies impliziert das Ergebnis durch Addition mit der Abschätzung aus Satz 5.39. □

5.41 Bemerkung. Die (z_k) werden im Allgemeinen bereits selbst konstant oder monoton fallend in k sein, so dass die Bedingung des Korollars in der Regel erfüllt ist. Umso stärker ist die Aussage dieses Korollars: der zusätzliche Fehler durch adaptive Bandweitenwahl ist beschränkt durch ein Vielfaches des Orakel-Typ-Fehlers, wobei α und der α-kalibrierte kritische Wert den Faktor bestimmen. Bei einer konkreten Stichprobe kann dieser Faktor noch in α optimiert werden; für ein allgemeines Verständnis ist eine asymptotische Analyse hilfreicher.

Wir stellen nun beispielhaft ein typisches asymptotisches Resultat vor, weitreichende Verallgemeinerungen, beispielsweise für höhere Dimensionen, irreguläre Designs oder nicht-Gaußsche Fehler, sind möglich.

5.42 Satz. *Betrachte den Fall äquidistanten Designs $x_i = i/n$ auf dem Einheitsintervall, Gaußsche Fehler $\varepsilon_i \sim N(0, \sigma^2)$ sowie $x_0 \in (0,1)$. Weiterhin seien die Bandweiten $h_k = h_0 q^k$, $k = 0, \dots, K$, geometrisch wachsend mit $h_0 = 1/n$, $q > 1$ und $K = \lfloor \log_q(n) \rfloor$ gewählt. Dann sind die kritischen Werte α-kalibriert unter der Null für alle $\alpha > 0$, falls nur $z_k = \zeta\sqrt{\log n}$, $k = 0, 1, \dots, K-1$, mit $\zeta > 0$ hinreichend groß gewählt sind (asymptotisch reicht $\zeta > 2$). Es folgt für alle Hölderklassen $H_{[0,1]}(\beta, L)$ mit $\beta \in (0,1]$, $L > 0$*

$$\sup_{f \in H_{[0,1]}(\beta,L)} \mathbb{E}_f[(\hat{\vartheta} - f(x_0))^2] = O\big(L^{2/(2\beta+1)} (n/\log n)^{-2\beta/(2\beta+1)}\big).$$

Beweis. Da die Fehler normalverteilt sind, erhalten wir die Gaußsche Konzentrationsungleichung

$$P_0(|\tilde{\vartheta}_{j+1} - \tilde{\vartheta}_k| > z_k s_{j+1,k}) \leqslant 2e^{-z_k^2/2}.$$

Außerdem erhalten wir aus der Analyse von $\hat{f}_{n,h}(x_0)$ für die Varianz $s_j^2 \in [c_1\sigma^2(nh_j)^{-1}, C_1\sigma^2(nh_j)^{-1}]$ mit Konstanten $C_1 \geqslant c_1 > 0$. Eine Anwendung

der Cauchy-Schwarz-Ungleichung und die Identität $nh_j = q^j$ ergeben

$$\begin{aligned}\mathbb{E}_0[\tilde\vartheta_j^2 \mathbf{1}(|\tilde\vartheta_{j+1} - \tilde\vartheta_k| > z_k)] &\leqslant \mathbb{E}_0[\tilde\vartheta_j^4]^{1/2} P_0(|\tilde\vartheta_{j+1} - \tilde\vartheta_k| > z_k)^{1/2} \\ &\leqslant \sqrt{6} s_j^2 e^{-z_k^2/4} \leqslant \sqrt{6} C_1^2 \sigma^2 q^{-2j} n^{-\zeta^2/4}.\end{aligned}$$

Summation über $j = 0, \dots, K-1$ und eine triviale Abschätzung des Indikators zeigen daher

$$\sum_{j=0}^{K-1} \mathbb{E}_0 \left[\tilde\vartheta_j^2 \mathbf{1}\big(\exists l \leqslant j : |\tilde\vartheta_{j+1} - \tilde\vartheta_l| \leqslant z_l s_{j+1,l}\big)\right] \leqslant C_2 \sigma^2 n^{-\zeta^2/4}$$

mit einer Konstanten $C_2 > 0$. Dies ist kleiner als $\alpha c_1 n^{-1}\sigma^2 \leqslant \alpha s_K^2$ für $\zeta^2 \geqslant 4(1 + \log(\alpha c_1/C_2)/(\log n)) = 4 + o(1)$. Die erste Behauptung folgt.

Nach Korollar 5.40 folgt mit der angegebenen Wahl der kritischen Werte

$$\mathbb{E}_f[(\hat\vartheta - \tilde\vartheta_{k^*})^2] = O(\log(n)(nh_{k^*})^{-1}).$$

Für $f \in H_{[0,1]}(\beta, L)$ ist $V_k(f) \leqslant L(2h_k)^\beta$ erfüllt, und somit erhalten wir für den Orakel-Typ-Index mit Konstanten $c_3, c_4 > 0$

$$Lh_{k^*}^\beta \geqslant c_3 \sqrt{\log n}(nh_{k^*})^{-1/2} \quad \Rightarrow \quad h_{k^*} \geqslant c_4 \big(L^2 n/\log n\big)^{-1/(2\beta+1)}.$$

Dies impliziert $\mathbb{E}_f[(\hat\vartheta - \tilde\vartheta_{k^*})^2] = O(L^{2/(2\beta+1)}(n/\log n)^{-2\beta/(2\beta+1)})$. Beachte auch, dass dies zeigt, dass k^* mindestens wie $\log n$ wächst, der Fall $k^* = 0$ also für große n nie vorkommt.

Andererseits gilt nach Definition für $k^* > 0$ auch $V_{k^*}(f) \leqslant z_{k^*} s_{k^*}$ und für den Biasterm im Schätzer $\tilde\vartheta_{k^*}$ folgt

$$\Big|\frac{\sum_{i=1}^n (f(i/n) - f(x_0)) K_{h_{k^*}}(x_0 - i/n)}{\sum_{i=1}^n K_{h_{k^*}}(x_0 - i/n)}\Big| \leqslant V_{k^*}(f) = O(\sqrt{\log n}(nh_{k^*})^{-1/2}).$$

Damit sind quadrierter Bias und Varianz von $\tilde\vartheta_{k^*}$ beide von der Ordnung $O(L^{2/(2\beta+1)}(n/\log n)^{-2\beta/(2\beta+1)})$, und die Behauptung folgt mit der Ungleichung $(\hat\vartheta - f(x_0))^2 \leqslant 2(\hat\vartheta - \tilde\vartheta_{k^*})^2 + 2(\tilde\vartheta_{k^*} - f(x_0))^2$. □

In Lepski (1990) wird bewiesen (für den Fall eines Signals in weißem Rauschen), dass die Rate $(n/\log n)^{-2\beta/(2\beta+1)}$ für punktweises Risiko minimax ist bei adaptiver Schätzung über die Familie von Hölderklassen $(H_{[0,1]}(\beta, L))_{\beta \in (0,1]}$. Bei punktweisem Risiko müssen wir also in der Konvergenzrate verlieren, wenn wir adaptiv schätzen wollen, im scharfen Kontrast zum MISE, wo ja sogar asymptotisch exakte Orakelungleichungen gelten. Es sei erwähnt, dass die Lepski-Methode angewendet auf lokale Polynome oder Projektionsschätzer dieselben Ergebnisse auch für Hölderklassen mit $\beta > 1$ zeigt. Nur die Definition von V_k, die den maximalen Bias misst, muss entsprechend angepasst werden. Unsere Idee zur Kalibrierung der kritischen Werte unter der Null ist inspiriert durch Spokoiny and Vial (2009).

6 Nichtparametrische Konfidenz

7 Klassifikation und Lerntheorie

7.1 Klassifikation und Bayes-Klassifizierer

Häufig müssen in der Statistik Entscheidungen zwischen zwei oder mehr Alternativen aufgrund komplexer Daten getroffen werden. Dies führt auf sogenannte Klassifikationsprobleme. Beispiele sind Spam-Filter (Klassifikationen zwischen 'mit Sicherheit Spam', 'mit Sicherheit kein Spam' und Klassen dazwischen), medizinische Datenanalyse (wie EKG weist auf Krankheit x hin oder nicht) oder Schrifterkennung (ASCII-Code auf der Basis von Pixelmuster). Hier werden wir nur binäre Klassifikation mit Klassen (*Labels*) '-1', '+1' betrachten.

Für die Modellierung betrachte auf dem Stichprobenraum $\mathcal{X}$ (in der Regel $\mathbb{R}^d$) und der zugehörigen σ-Algebra Wahrscheinlichkeitsverteilungen P^- und P^+. Bei Vorliegen von Klasse '-1' beobachten wir eine Zufallsvariable $X \sim P^-$, bei Klasse '+1' jedoch $X \sim P^+$. Die a-priori-Wahrscheinlichkeit für Klasse '+1' sei $\pi \in [0,1]$, das heißt die Klasse Y werde gemäß π gezogen und danach $X \sim P^{\pm}$ für $Y = \pm 1$. Nach der Bayesformel folgt mit entsprechenden Dichten p^+, p^- bezüglich einem dominierenden Maß Q:

$$\eta(x) := P(Y = +1 \,|\, X = x) = \frac{\pi p^+(x)}{(1-\pi)p^-(x) + \pi p^+(x)}.$$

Ziel der Klassifikation ist es, bei einer Realisierung von X die Klasse Y so gut wie möglich vorherzusagen.

7.1 Definition. Jede messbare Funktion $h : \mathcal{X} \to \{-1, +1\}$ heißt *Klassifizierer*. Ihr *Klassifikationsfehler* ist $R(h) := P(Y \neq h(X))$. Durch

$$h^*(x) = \begin{cases} +1, & \text{falls } \eta(x) > 1/2, \\ -1, & \text{falls } \eta(x) \leqslant 1/2 \end{cases} = \begin{cases} +1, & \text{falls } \pi p^+(x) > (1-\pi)p^-(x), \\ -1, & \text{falls } \pi p^+(x) \leqslant (1-\pi)p^-(x) \end{cases}$$

wird der *Bayes-Klassifizierer* festgelegt. $R^* := R(h^*)$ heißt *Bayesrisiko* des Klassifikationsproblems.

Beachte, dass ein binäres Klassifikationsproblem gerade einem Test entspricht und der Bayes-Klassifizierer dem Bayestest, wenn die beiden möglichen Testfehler gleichen Verlust besitzen.

7.2 Lemma. *Für jeden Klassifizierer h gilt*

$$R(h) - R^* = \mathbb{E}[|2\eta(X) - 1| \mathbf{1}(h(X) \neq h^*(X))].$$

Insbesondere besitzt der Bayes-Klassifizierer minimales Risiko. Darüberhinaus gilt $R^* = \mathbb{E}[\eta(X) \wedge (1 - \eta(X))]$.

Beweis. Durch Bedingen und Einsetzen von $\eta(x)$ erhalten wir

$$\begin{aligned}
&P(Y \neq h(X) \,|\, X = x) \\
&= P(Y = 1 \,|\, X = x)\mathbf{1}(h(x) = -1) + P(Y = -1 \,|\, X = x)\mathbf{1}(h(x) = 1) \\
&= 1 - \eta(x) + (2\eta(x) - 1)\mathbf{1}(h(x) = -1).
\end{aligned}$$

Nun gilt $\mathbf{1}(h^*(x) = -1) = \mathbf{1}(2\eta(x) - 1 \leqslant 0)$, so dass insbesondere $P(Y \neq h^*(X) \,|\, X = x) = \eta(x) \wedge (1 - \eta(x))$ und damit $R^* = \mathbb{E}[\eta(X) \wedge (1 - \eta(X))]$ gilt. Weiterhin schließen wir

$$\begin{aligned}
&P(Y \neq h(X) \,|\, X = x) - P(Y \neq h^*(X) \,|\, X = x) \\
&= (2\eta(x) - 1)(\mathbf{1}(h(x) = -1) - \mathbf{1}(h^*(x) = -1)) \\
&= |2\eta(x) - 1|\mathbf{1}(h(x) \neq h^*(x)).
\end{aligned}$$

Integration bezüglich der Verteilung von X ergibt

$$R(h) - R^* = P(Y \neq h(X)) - P(Y \neq h^*(X)) = \mathbb{E}[|2\eta(X) - 1|\mathbf{1}(h(X) \neq h^*(X))].$$

Der letzte Ausdruck ist offensichtlich nicht-negativ. □

Wir können also niemals besser als mit dem Bayesrisiko R^* klassifizieren, weshalb für Klassifizierer h auch gerne nur das *Exzessrisiko* $\mathcal{E}(h) := R(h) - R^*$ betrachtet wird.

7.2 Minimierung des empirischen Risikos

In der Praxis sind P^+, P^-, π und $\eta(x)$ nicht bekannt, der Bayes-Klassifizierer dient nur als Orakel. Im folgenden wird von der Beobachtung einer mathematischen Stichprobe $(X_i, Y_i)_{i=1,\ldots,n}$ mit $X_i \in \mathcal{X}$, $Y_i \in \{-1, +1\}$ und $(X_i, Y_i) \sim P$ i.i.d. ausgegangen, der *Trainingsmenge*. Ziel ist es, aus dieser Beobachtung einen Klassifizierer zu konstruieren, der dem Bayesrisiko (zur Klassenvorhersage bei einer neuen Beobachtung $(X, Y) \sim P$) möglichst nahe kommt. Dafür ist ein naheliegender Ansatz $\eta(x)$ über einen Regressionsansatz nichtparametrisch zu schätzen und den Schätzer dann in den Bayes-Klassifizierer einzusetzen. Dafür muss allerdings das viel komplexere Problem der Funktionsschätzung von $\eta(x)$ anstelle des einfacheren Entscheidungsproblems $\{\eta(x) > 1/2\}$ gelöst werden. Dies ist numerisch häufig sehr aufwändig und kann auch mathematisch suboptimal sein. Ein direkter Ansatz ist das Konzept der *empirischen Risiko-Minimierung (ERM)*.

7.3 Definition. Das *empirische Risiko* eines Klassifizierers h bei einer Trainingsmenge $(X_i, Y_i)_{i=1,\ldots,n}$ ist gegeben durch

$$R_n(h) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(Y_i \neq h(X_i)).$$

Ein Klassifizierer $\hat{h}_n^{ERM}$ aus einer Familie $\mathcal{H}$ von Klassifizierern heißt *ERM-Klassifizierer*, wenn $R_n(\hat{h}_n^{ERM}) = \min_{h \in \mathcal{H}} R_n(h)$ gilt.

Das Exzessrisiko des ERM-Klassifizierers erlaubt eine Aufspaltung analog zur Bias-Varianz-Zerlegung

$$R(\hat{h}_n^{ERM}) - R^* = \underbrace{R(\hat{h}_n^{ERM}) - \inf_{h\in\mathcal{H}} R(h)}_{\text{stochastischer Fehler}} + \underbrace{\inf_{h\in\mathcal{H}} R(h) - R^*}_{\text{Approximationsfehler}}.$$

Je größer die Familie $\mathcal{H}$ gewählt wird, desto kleiner ist der Approximationsfehler, desto größer ist jedoch im Allgemeinen der stochastische Fehler. Wählt man beispielsweise $\mathcal{H}$ als die Familie aller Klassifizierer, so ist jedes h mit $h(X_i) = Y_i$ und $h(x)$ beliebig für $x \notin \{X_1, \dots, X_n\}$ ein ERM-Klassifizierer, was zu beliebig schlechtem Risko von h führen kann.

7.4 Lemma. *Der stochastische Fehler lässt folgende Abschätzung zu:*

$$R(\hat{h}_n^{ERM}) - \inf_{h\in\mathcal{H}} R(h) \leqslant 2 \sup_{h\in\mathcal{H}} |R_n(h) - R(h)|.$$

Beweis. Nach Definition gilt für alle $h' \in \mathcal{H}$

$$\begin{aligned} R(\hat{h}_n^{ERM}) &\leqslant R_n(\hat{h}_n^{ERM}) + \sup_{h\in\mathcal{H}} |R_n(h) - R(h)| \\ &\leqslant R_n(h') + \sup_{h\in\mathcal{H}} |R_n(h) - R(h)| \\ &\leqslant R(h') + 2 \sup_{h\in\mathcal{H}} |R_n(h) - R(h)|. \end{aligned}$$

Subtrahiere nun $R(h')$ auf beiden Seiten und bilde das Supremum über $h' \in \mathcal{H}$. □

7.5 Satz. *Ist die Familie $\mathcal{H} = \{h_1, \dots, h_M\}$ endlich, so gilt für jedes $\delta \in (0,1)$ mit Wahrscheinlichkeit $1-\delta$ die Orakelungleichung*

$$R(\hat{h}_n^{ERM}) \leqslant \min_{1\leqslant m\leqslant M} R(h_m) + \sqrt{\tfrac{8}{n}\log(2M/\delta)}.$$

Beweis. Nach dem Lemma erhalten wir für $t > 0$

$$\begin{aligned} P\Big(R(\hat{h}_n^{ERM}) - \min_{1\leqslant m\leqslant M} R(h_m) > t\Big) &\leqslant P\Big(\max_{1\leqslant m\leqslant M} |R_n(h_m) - R(h_m)| > t/2\Big) \\ &\leqslant \sum_{m=1}^{M} P(|R_n(h_m) - R(h_m)| > t/2). \end{aligned}$$

Schreiben wir $R_n(h_m) - R(h_m) = \frac{1}{n}\sum_{i=1}^n Z_i$ mit $Z_i := \mathbf{1}(Y_i \neq h_m(X_i)) - \mathbb{E}[\mathbf{1}(Y_i \neq h_m(X_i))] \in [-1,1]$ i.i.d., so können wir die Hoeffding-Ungleichung anwenden und erhalten

$$\sum_{m=1}^{M} P(|R_n(h_m) - R(h_m)| > t/2) \leqslant 2Me^{-nt^2/8}.$$

Damit liefert die Wahl $t = \sqrt{\frac{8}{n}\log(2M/\delta)}$ das Ergebnis. □

7.6 Bemerkung. Eine schärfere Variante der Hoeffding-Ungleichung ergibt auf der rechten Seite $\sqrt{\frac{2}{n}\log(2M/\delta)}$. Unter direkter Verwendung der exponentiellen Momente der Binomialverteilung von $R_n(h)$ kann man auch eine Orakelungleichung im Erwartungswert erhalten:

$$\mathbb{E}[R(\hat{h}_n^{ERM})] \leqslant \min_{1\leqslant m\leqslant M} R(h_m) + \sqrt{\tfrac{\log(2M)}{2n}}.$$

Für allgemeine, unendliche Familien $\mathcal{H}$ kann man den stochastischen Fehler in natürlicher Weise mit der Theorie der empirischen Prozesse kontrollieren. Hierbei ist insbesondere die sogenannte *Vapnik-Chervonenkis-Dimension* der Familie $\mathcal{H}$ ein wichtiges Maß für die Komplexität von $\mathcal{H}$. Wir werden im folgenden nur eine spezielle Form der empirischen Risikominimierung betrachten, wo maßgeschneiderte Methoden zum Einsatz kommen.

7.3 Support Vector Machines

Aus numerischer Sicht ist die Minimierung des empirischen Risikos im Allgemeinen sehr schwierig, da $R_n(h)$ nicht-konvex in h ist (potentiell NP-vollständiges Problem). Die erste Idee, die auf sogenannte *Support Vector Machines* (*SVM, Stützvekor-Methoden*) führt, ist es daher, ein verwandtes konvexes Problem zu lösen. Dabei wird $\mathbf{1}(Y_i \neq h(X_i)) = \mathbf{1}(-Y_i h(X_i) > 0)$ durch $\varphi(-Y_i h(X_i))$ mit einer konvexen Funktion $\varphi : \mathbb{R} \to [0,\infty)$ ersetzt. Wir werden im folgenden die in der Praxis am meisten verwendete Funktion $\varphi(x) = (1+x)_+$ (*hinge loss*) betrachten.Außerdem gehen wir von $\{-1,+1\}$-wertigen zu reellwertigen Funktionen über.

7.7 Definition. Für eine konvexe Funktion $\varphi : \mathbb{R} \to [0,\infty)$ und eine Familie $\mathcal{F} \subseteq \{\{ : \mathcal{X} \to \mathbb{R} \text{ messbar}\}$ reellwertiger Funktionen setze für $f \in \mathcal{F}$

$$\begin{aligned}
R_\varphi(f) &:= \mathbb{E}[\varphi(-Yf(X))] \text{ (φ-Risiko)},\\
R_{n,\varphi}(f) &:= \frac{1}{n}\sum_{i=1}^n \varphi(-Y_i f(X_i)) \text{ (empirisches φ-Risiko)},\\
h_f(x) &:= \begin{cases} +1, & \text{falls } f(x) > 0,\\ -1, & \text{falls } f(x) \leqslant 0 \end{cases} \text{ (zu } f \text{ assoziierter Klassifizierer)},\\
\hat{f}_{n,\varphi} &:= \operatorname{argmin}_{f\in\mathcal{F}} R_{n,\varphi}(f) \text{ (verallg. φ-ERM-Klassifizierer)},\\
\hat{h}_{n,\varphi} &:= h_{\hat{f}_{n,\varphi}} \text{ (φ-ERM-Klassifizierer)}.
\end{aligned}$$

7.8 Lemma. *Für den Bayes-Klassifizierer h^* und für hinge loss $\varphi(x) = (1+x)_+$ gilt*

$$\min_{f:\mathcal{X}\to\mathbb{R} \text{ messbar}} R_\varphi(f) = R_\varphi(h^*) = 2R^*.$$

Beweis. Durch Bedingen erhalten wir

$$\begin{aligned}\mathbb{E}[\varphi(-Yf(X)) \,|\, X = x] &= (1 + f(x))_+(1 - \eta(x)) + (1 - f(x))_+\eta(x) \\ &\geqslant 2(\eta(x) \wedge (1 - \eta(x))),\end{aligned}$$

wobei wir $(1+A)_+ + (1-A)_+ \geqslant (1+A)+(1-A) = 2$ verwendet haben. Es gilt Gleichheit genau dann, wenn $f(x) = -1$ im Fall $\eta(x) < 1/2$ bzw. $f(x) = +1$ im Fall $\eta(x) > 1/2$ erfüllt ist. Also minimiert der Bayesklassifizierer $h^*(x)$ diese bedingte Erwartung, und es folgt $R_\varphi(f) \geqslant R_\varphi(h^*) = 2R^*$. □

Die zweite Idee, die SVMs attraktiv machen, ist ein allgemeiner Ansatz zur Wahl der Klassifizierermenge. Er beruht auf einer Transformation der Daten (X_i) in den sogenannten *Feature-Raum* mittels Kernen. Dazu sind mathematische Vorarbeiten nötig.

7.9 Definition. Eine Funktion $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ heißt *reproduzierender Kern* für einen Hilbertraum W von reellwertigen Funktionen auf $\mathcal{X}$, falls

(a) $\forall x \in \mathcal{X} :\ k(x, \bullet) \in W$;

(b) $\forall x \in \mathcal{X},\ f \in W :\ f(x) = \langle f, k(x, \bullet)\rangle_W$ (*Reproduktion*).

In diesem Fall heißt W auch *RKHS (reproducing kernel Hilbert space).*

7.10 Lemma. *Es gilt* $\langle k(x, \bullet), k(x', \bullet)\rangle_W = k(x, x')$ *sowie* $\sup_{x\in\mathcal{X}} |f(x)| \leqslant \|f\|_W \sup_{x\in\mathcal{X}} \sqrt{k(x, x)}$.

Beweis. Die erste Iedentität folgt aus der Reproduktion $\langle f, k(x', \bullet)\rangle_W = f(x')$ angewendet auf $f(y) = k(x, y)$. Daraus folgt die Abschätzung mit der Cauchy-Schwarz-Ungleichung:

$$\sup_{x\in\mathcal{X}} f(x)^2 = \sup_{x\in\mathcal{X}} \langle f, k(x, \bullet)\rangle_W^2 \leqslant \|f\|_W^2 \sup_{x\in\mathcal{X}} \|k(x, \bullet)\|_W^2 = \|f\|_W^2 \sup_{x\in\mathcal{X}} k(x, x).$$

□

7.11 Beispiele.

(a) Die (bezüglich Lebesguemaß) quadrat-integrierbaren Funktion $L^2(\mathcal{X})$ bilden keinen RKHS, weil die Punktauswertungen $f \mapsto f(x)$ nicht stetig bezüglich L^2 sind. Deshalb sind die Elemente von $L^2(\mathcal{X})$ natürlich auch nur f.ü. festgelegt.

(b) Bildet $\varphi_1, \ldots, \varphi_K$ ein Orthonormalsystem bezüglich $L^2(\mathcal{X})$, so ist

$$k(x, y) := \sum_{k=1}^{K} a_k \varphi_k(x)\varphi_k(y) \text{ für beliebige } a_k > 0$$

ein reproduzierender Kern von $W = \operatorname{span}(\varphi_1, \dots, \varphi_K)$ bezüglich

$$\langle f, g\rangle_W := \sum_{k=1}^{K} a_k^{-1} \langle f, \varphi_k\rangle_{L^2} \langle g, \varphi_k\rangle_{L^2}.$$

Dies folgt mit $f \in W$ sofort aus

$$\begin{aligned}\langle f, k(x, \bullet)\rangle_W &= \sum_{k=1}^{K} a_k^{-1} \langle f, \varphi_k\rangle_{L^2} \sum_{l=1}^{K} a_l \langle \varphi_l, \varphi_k\rangle_{L^2} \varphi_l(x) \\ &= \sum_{k=1}^{K} \langle f, \varphi_k\rangle_{L^2} \varphi_k(x) = f(x).\end{aligned}$$

Dies lässt sich auf unendliche Reihenentwicklungen ($K = \infty$) verallgemeinern, sofern $k(x, y)$ wohldefiniert ist. Ein wichtiges Beispiel dafür ist die Fourierbasis (φ_k) mit $(a_k) \in \ell^1$, insbesondere führt $a_k = (1 + k^2)^{-s}$ für $s > 1/2$ auf den RKHS $W = H^s_{per}([0, 1])$.

(c) Betrachte $W = \{f : [0, 1] \to \mathbb{R} \mid f(0) = 0, \int_0^1 f'(x)^2 dx < \infty\}$, wobei die Ableitung im schwachen Sinne existieren möge. Dies ist bezüglich $\langle f, g\rangle_W = \langle f', g'\rangle_{L^2}$ ein Hilbertraum. Mit $k(x, y) = x \wedge y$ gilt $k(x, \bullet) \in W$ sowie

$$\langle f, k(x, \bullet)\rangle_W = \int_0^1 f'(y) \mathbf{1}(y \leqslant x)\, dx = f(x).$$

Also ist k ein reproduzierender kern von W. Beachte nun, dass $k(x, y) = \mathbb{E}[B_x B_y]$ gerade die Kovarianzfunktion der Brownschen Bewegung B ist und W gerade der *Cameron-Martin-Raum* aus der stochastischen Analysis. Im Sinne der stochastischen Integration gilt $\mathbb{E}[\langle f, B\rangle_W \langle g, B\rangle_W] := \mathbb{E}[\int f' dB \int g' dB] = \langle f, g\rangle_W$. Allgemeiner führen zentrierte Gaußprozesse auf $\mathcal{X}$ über die Kovarianzfunktion als Kern zu RKHS, die auch in der Theorie der Gaußprozesse eine tragende Rolle spielen.

(d) Gegeben sei eine symmetrische, positiv-definite Funktion $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, das heißt $k(x, y) = k(y, x)$ und $\sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j) \geqslant 0$ mit Gleichheit nur für $\alpha_1 = \dots = \alpha_n = 0$ für alle $x, y, x_i \in \mathcal{X}$, $\alpha_i \in \mathbb{R}$, $n \in \mathbb{N}$. Betrachte $W = \operatorname{span}(k(x_1, \bullet), \dots, k(x_n, \bullet))$ für paarweise verschiedene $x_1, \dots, x_n \in \mathcal{X}$. Für $f = \sum_{i=1}^n \alpha_i k(x_i, \bullet)$ gilt dann $f(x_j) = \sum_{i=1}^n \alpha_i k(x_j, x_i) = \langle f, k(x_j, \bullet)\rangle_W$, wenn man für $f = \sum_{i=1}^n \alpha_i k(x_i, \bullet)$, $g = \sum_{i=1}^n \beta_i k(x_i, \bullet)$ definiert $\langle f, g\rangle_W := \sum_{i=1}^n \alpha_i \beta_j k(x_i, x_j) = \alpha^\top \mathbf{K} \beta$ mit $\mathbf{K} = (k(x_i, x_j))_{i,j=1,\dots,n}$. Auf diese Weise kann aus k und Punkten $x_1, \dots, x_n$ ein zugehöriger RKHS W erzeugt werden, was in den Anwendungen von SVM der häufigste Zugang ist. Populär ist der Gaußsche Radialkern $k(x, y) = e^{-|x-y|^2/\gamma^2}$ mit $\gamma > 0$ geeignet, wo W gerade

aus Linearkombinationen d-dimensionaler Gaußscher Glockenfunktionen besteht. Der bilineare Kern $k(x,y) = \langle x,y\rangle_{\mathbb{R}^d}$ führt geometrisch auf den Raum W der Abstandsfunktionen zu den Hyperebenen mit Normalenvektoren $\sum_i \alpha_i x_i$. Eine unendlich-dimensionale Version ergibt sich aus dem Integraloperator mit Kern k, wobei der *Satz von Mercer* aus der Funktionalanalysis gerade über die Eigenfunktionen den Zusammenhang mit (b) herstellt.

7.12 Satz. *(Darsteller-Eigenschaft) Es seien W ein RKHS bezüglich $k : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$, $\Phi:\mathbb{R}\to\mathbb{R}$ streng monoton wachsend und $G:\mathbb{R}^n\to\mathbb{R}$ beliebig. Dann besitzt für $x_1,\dots,x_n\in\mathcal{X}$ jede Lösung des Minimierungsproblems*

$$G(f(x_1),\dots,f(x_n)) + \Phi(\|f\|_W) \longrightarrow \min_{f\in W}!$$

die Form $f(x) = \sum_{i=1}^n \alpha_i k(x_i,x)$ mit $\alpha_i\in\mathbb{R}$ geeignet.

Ist G konvex und nicht-negativ, so existiert für jedes $\lambda>0$ eine eindeutige Lösung des Minimierungsproblems

$$G(f(x_1),\dots,f(x_n)) + \lambda\|f\|_W^2 \longrightarrow \min_{f\in W}!$$

Beweis. Betrachte $V := \operatorname{span}(k(x_1,\bullet),\dots,k(x_n,\bullet))$ und $V^\perp := \{u\in W \mid \forall\, v\in V:\ \langle u,v\rangle_W = 0\}$. Dann gilt für $u\in V^\perp$ offenbar $u(x_i) = \langle u, k(x_i,\bullet)\rangle_W = 0$. Für jedes $f = u+v\in W$ mit $u\in V^\perp$, $v\in V$ gilt also

$$\forall\, i=1,\dots,n:\ f(x_i) = v(x_i),\quad \|f\|_W^2 = \|u\|_W^2 + \|v\|_W^2.$$

Ist daher solch ein $f=u+v$ Lösung des Minimierungsproblems, so muss $u=0$ gelten, weil andernfalls das Kriterium bei v kleiner ist als bei f. Dies zeigt $f\in V$ und die erste Behauptung.

Wenn G konvex und nicht-negativ ist, so trifft dies auch auf $f\mapsto K(f) := G(f(x_1),\dots,f(x_n)) + \lambda\|f\|_W^2$ als Komposition und Summe konvexer Funktionen zu. Darüberhinaus gilt für $f\in W$ mit $\lambda\|f\|_W^2 > G(0,\dots,0)$ natürlich $K(f) > K(0)$. Damit existieren $f_n\in W$ mit $\|f_n\|_W \leqslant \lambda^{-1/2} G(0,\dots,0)^{1/2}$ und $K(f_n)\to \min_{f\in W} K(f)$. Die Zerlegung $f_n = u_n + v_n$ mit $u_n\in V^\perp$, $v_n\in V$ zeigt wie oben $v_n\to\min_{f\in W}K(f)$. Nun liegt aber (v_n) in dem kompakten endlich-dimensionalen Ball $\{v\in V \mid \|v_n\|_W \leqslant \lambda^{-1/2}G(0,\dots,0)^{1/2}\}$, und jeder Häufungspunkt von (v_n) löst das Minimierungsproblem. Wären f_1,f_2 zwei verschiedene Lösungen des Minimierungsproblems, so würde für $\bar f = \frac12(f_1+f_2)$ nach der Parallelogramm-Identität gelten

$$\|\bar f\|_W^2 = \tfrac14(2\|f_1\|_W^2 + 2\|f_2\|_W^2 - \|f_1-f_2\|_W^2) < \tfrac12(\|f_1\|_W^2 + \|f_2\|_W^2).$$

Wegen der Konvexität von G folgte daraus $K(\bar f) < \frac12(K(f_1)+K(f_2))$ (K ist sogar strikt konvex), im Widerspruch zur Minimalität bei f_1,f_2. Also ist die Lösung eindeutig. □

7.13 Definition. Für einen reproduzierenden Kern $k\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ eines RKHS W und $\lambda > 0$ setze

$$\hat{f}_n^{SVM} := \operatorname{argmin}_{f \in W, \|f\|_W \leqslant \lambda} \Big(\frac{1}{n} \sum_{i=1}^{n} (1 - Y_i f(X_i))_+ \Big)$$

Dann ergibt $\hat{h}_n^{SVM} = h_{\hat{f}_n^{SVM}}$ den zugehörigen *SVM-Klassifizierer* (*support vector machine, Stützvektor-Klassifizierer*).

Für SVMs wählt man also einen Ball $\mathcal{F} = \{f \in W \,|\, \|f\|_W \leqslant \lambda\}$ in einem RKHS W auf $\mathcal{X}$ mit Radius $\lambda > 0$ als (verallgemeinerte) Klassifiziererfamilie. Nach der Lagrange-Theorie ergibt sich dann als φ-ERM-Klassifizierer für $\lambda' > 0$ geeignet

$$\hat{f}_n^{SVM} = \operatorname{argmin}_{f \in W} \Big(R_{n,\varphi(\lambda)}(f) + \lambda' \|f\|_W^2 \Big).$$

Wie zuvor beim Schätzen erhalten wir den Klassifizierer durch Minimierung eines penalisierten Datenfehlers, wobei λ bzw. λ' die Rolle eines Glättungsparameters spielt.

Nach dem Darsteller-Satz lässt sich wegen $R_{n,\varphi}(f) = G(f(X_1), \ldots, f(X_n))$ das zweite Minimierungsproblem nun als endlich-dimensionales Problem nur in dem Kern schreiben. Außerdem gilt nach dem Lemma $\|\sum_i \alpha_i k(x_i, \bullet)\|_W^2 = \sum_{i,j} \alpha_i \alpha_j K(x_i, x_j)$, so dass mit $\hat{f}_n^{SVM} = \sum_{i=1}^n \hat{\alpha}_i K(X_i, \bullet)$ für die Koeffizienten $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_n)$ gilt:

$$\hat{\alpha} = \operatorname{argmin}_{\alpha \in \mathbb{R}^n} \Big(\frac{1}{n} \sum_{i=1}^{n} \Big(1 - Y_i \sum_{j=1}^{n} \alpha_j K(X_i, X_j) \Big)_+ + \lambda' \sum_{i,j=1}^{n} \alpha_i, \alpha_j K(X_i, X_j) \Big)$$

Beachte, dass in dieser Charakterisierung der RKHS W nicht mehr erscheint und nur ein endlich-dimensionales Optimierungsproblem zu lösen ist.

Betrachte nun $I := \{i = 1, \ldots, n \,|\, Y_i \hat{f}_n^{SVM}(X_i) \leqslant 1\}$. Dann ist $\hat{f}_n^{SVM}$ auch Lösung des Minimierungsproblems

$$\hat{f}_n^{SVM} = \operatorname{argmin}_{f \in W} \Big(\frac{1}{n} \sum_{i \in I} (1 - Y_i f(X_i))_+ + \lambda' \|f\|_W^2 \Big).$$

Der Darstellersatz impliziert daher $\hat{f}_n^{SVM} = \sum_{i \in I} \hat{\alpha}_i K(X_i, \bullet)$, so dass die Koeffizienten $\hat{\alpha}_i$ gleich Null sind, die zu signifikant korrekt klassifizierten Beobachtungen (X_i, Y_i) im Sinne von $Y_i \hat{f}_n(X_i) > 1$ gehören. Es liegt also häufig eine sparsame (*sparse*) Darstellung von $\hat{f}_n^{SVM}$ vor. Man nennt $(X_i)_{i \in I}$ die *Stützvektoren* (*support vectors*).

Für die Interpretation ist es sinnvoll, dass Minimierungsproblem weiter äquivalent umzuschreiben in

$$(\hat{f}_{n,\varphi}, \hat{\xi}) := \operatorname{argmin}_{f \in W, \xi \in [0,\infty)^n, \forall i: Y_i f(X_i) \geqslant 1 - \xi_i} \Big(\lambda' \|f\|_W^2 + \frac{1}{n} \sum_{i=1}^{n} \xi_i \Big).$$

Beachte, dass die Nebenbedingungen an ξ sicherstellen, dass $\xi_i \geqslant (1 - Y_i f(X_i))_+$ gilt, so dass $\hat{\xi}_i = (1 - Y_i f(X_i))_+$ und damit die Äquivalenz mit dem ursprünglichen Problem folgt.

Werden alle (X_i, Y_i) signifikant korrekt klassifiziert im Sinne von $Y_i \hat{f}_n^{SVM}(X_i) \geqslant 1$, so folgt $\hat{\xi} = 0$ und $\hat{f}_n^{ERM}$ minimiert die Norm $\|f\|_W$ unter allen signifikant korrekten Klassifizierern $f \in W$. Das ist durch Umskalieren äquivalent dazu, bei gegebener Norm $\|f\|_W$ den sogenannten *margin* $\min_i(Y_i \hat{f}_n^{SVM}(X_i))$ zu maximieren. Diejenigen Beobachtungen (X_j) mit $Y_j \hat{f}_n^{SVM}(X_j) = \min_i(Y_i \hat{f}_n^{SVM}(X_i))$ sind gerade die Stützvektoren. Im Allgemeinen wird $\hat{f}_n^{SVM}$ nicht alle Punkte korrekt klassifizieren, was über den Term $\frac{1}{n}\sum_{i=1}^n \xi_i$ bestraft wird. Diese Penalisierung ist ein Maß für den Gesamtabstand der allgemeinen Stützvektoren, das heißt der Punkte auf der falschen Seite des durch die korrekt klassifizierten Punkte bestimmten *margin*. Im Vergleich zur linearen Diskriminanzanalyse, die auf empirischem Mittelwert und Kovarianzmatrix beruht, wird die Klassifikation mit SVMs nicht durch alle Beobachtungen, sondern nur durch die kritischen Stützvektoren in der Nähe der Entscheidungsgrenze bestimmt. Der *Kern-Trick* erlaubt darüberhinaus eine nichtlineare Entscheidungsgrenze in $\mathcal{X}$, die jedoch im *Featureraum* $\operatorname{span}(k(X_1, \bullet), \ldots, k(X_n, \bullet))$ linear ist.

Wir werden nun eine Orakelgleichung für SVM beweisen. Damit erhalten wir eine Kontrolle des stochastischen Fehlers. Der Approximationsfehler ist stark vom betrachteten Problem und der Wahl des Kerns abhängig und wird nicht weiter abgeschätzt. Für eine allgemeine Theorie und eine Analyse der Wahl des Parameters $\lambda > 0$ sei auf Steinwart and Christmann (2008) verwiesen.

7.14 Satz. *Für den SVM-Klassifizierer $\hat{h}_n^{SVM}$ im RKHS W bezüglich einem Kern k und mit Radius $\lambda > 0$ gilt*

$$\mathbb{E}[\hat{h}_n^{SVM}] - R^* \leqslant \Big(\inf_{\|f\|_W \leqslant \lambda} R_\varphi(f) - R^* \Big) + 8\lambda\, \mathbb{E}[k(X,X)]^{1/2} n^{-1/2}.$$

Beweis.

(a) Zunächst bemerken wir mit $\varphi(x) = (1 + x)_+$

$$R(\hat{h}_n^{SVM}) = P(-Y \hat{f}_n^{SVM}(X) > 0) \leqslant \mathbb{E}[(1 - Y \hat{f}_n^{SVM}(X))_+] = R_\varphi(\hat{f}_n^{SVM}).$$

Wir erhalten also

$$R(\hat{h}_n^{SVM}) - R^* \leqslant R_\varphi(\hat{f}_n^{SVM}) - \inf_{\|f\|_W \leqslant \lambda} R_\varphi(f) + \inf_{\|f\|_W \leqslant \lambda} R_\varphi(f) - R^*.$$

Nun gilt wiederum (Beweis wie oben):

$$R_\varphi(\hat{f}_n^{SVM}) - \inf_{\|f\|_W \leqslant \lambda} R_\varphi(f) \leqslant 2 \sup_{\|f\|_W \leqslant \lambda} |R_{n,\varphi}(f) - R_\varphi(f)|.$$

Wir müssen also nur noch zeigen:

$$\mathbb{E}[\sup_{\|f\|_W \leqslant \lambda} |R_{n,\varphi}(f) - R_\varphi(f)|] \leqslant 4\lambda \sup_{x \in \mathcal{X}} \sqrt{k(x,x)} n^{-1/2}.$$

(b) Wir verwenden nun ein Symmetrisierungs- und Kontraktionsargument, um das Supremum abzuschätzen. Dazu sei $(X_i', Y_i')_{i=1,\ldots,n}$ eine *Phantom-Stichprobe* (*ghost sample*), das heißt, $(X_i', Y_i')_{i=1,\ldots,n}$ sei auf demselben Wahrscheinlichkeitsraum wie $(X_i, Y_i)_{i=1,\ldots,n}$ definiert, besitze die gleich Verteilung, sei aber unabhängig von $(X_i, Y_i)_{i=1,\ldots,n}$. Dann gilt nach Jensenscher Ungleichung und $\sup_t \mathbb{E}[Z_t] \leqslant \mathbb{E}[\sup_t Z_t]$

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{\|f\|_W \leqslant \lambda} \Big|\frac{1}{n}\sum_{i=1}^n (\varphi(-Y_i f(X_i)) - \mathbb{E}[\varphi(-Y_i' f(X_i'))])\Big|\Big] \\
&\leqslant \mathbb{E}\Big[\sup_{\|f\|_W \leqslant \lambda} \Big|\frac{1}{n}\sum_{i=1}^n (\varphi(-Y_i f(X_i)) - \varphi(-Y_i' f(X_i')))\Big|\Big]
\end{aligned}$$

Weiterhin existiere eine Rademacherfolge $(\sigma_i)_{i=1,\ldots,n}$, also $P(\sigma_i = \pm 1) = 1/2$ und (σ_i) unabhängig, die auch unabhängig von $(X_i, Y_i)_{i=1,\ldots,n}$ und $(X_i', Y_i')_{i=1,\ldots,n}$ ist. Nun sind $\pm(\varphi(-Y_i f(X_i)) - \varphi(-Y_i' f(X_i')))$ identisch verteilt und damit ebenso verteilt wie $\sigma_i(\varphi(-Y_i f(X_i)) - \varphi(-Y_i' f(X_i')))$. Dies zeigt

$$\begin{aligned}
&\mathbb{E}[\sup_{\|f\|_W \leqslant \lambda} |R_{n,\varphi}(f) - R_\varphi(f)|] \\
&\leqslant \mathbb{E}\Big[\sup_{\|f\|_W \leqslant \lambda} \Big|\frac{1}{n}\sum_{i=1}^n \sigma_i(\varphi(-Y_i f(X_i)) - 1 + 1 - \varphi(-Y_i' f(X_i')))\Big|\Big] \\
&\leqslant 2\,\mathbb{E}\Big[\sup_{\|f\|_W \leqslant \lambda} \Big|\frac{1}{n}\sum_{i=1}^n \sigma_i(\varphi(-Y_i f(X_i)) - 1)\Big|\Big].
\end{aligned}$$

Wir verwenden nun das Kontraktionsprinzip (Ledoux and Talagrand 1991, Thm. 4.12): Ist $\psi : [-1,1] \to \mathbb{R}$ eine Kontraktion, also $|\psi(x) - \psi(y)| \leqslant |x - y|$ und $\psi(0) = 0$, so gilt für jede Familie $\mathcal{G} \subseteq \{g : \mathcal{X} \times \{-1,+1\} \to [-1,1] \text{ messbar}\}$

$$\mathbb{E}\Big[\sup_{g \in \mathcal{G}} \Big|\frac{1}{n}\sum_{i=1}^n \sigma_i \psi(g(X_i, Y_i))\Big|\Big] \leqslant 2\,\mathbb{E}\Big[\sup_{g \in \mathcal{G}} \Big|\frac{1}{n}\sum_{i=1}^n \sigma_i g(X_i, Y_i)\Big|\Big].$$

Für $\psi(u) = (\varphi(\|f\|_\infty u) - 1)/\|f\|_\infty$ und $g(x,y) = -yf(x)/\|f\|_\infty \in [-1,1]$ erhalten wir so

$$\mathbb{E}\Big[\sup_{\|f\|_W \leqslant \lambda} \Big|\frac{1}{n}\sum_{i=1}^n \sigma_i \frac{\varphi(-Y_i f(X_i)) - 1}{\|f\|_\infty}\Big|\Big] \leqslant 2\,\mathbb{E}\Big[\sup_{\|f\|_W \leqslant \lambda} \Big|\frac{1}{n}\sum_{i=1}^n \sigma_i \frac{Y_i f(X_i)}{\|f\|_\infty}\Big|\Big],$$

und die Konstante $\|f\|_\infty$ fällt weg. Nun ist $\sigma_i Y_i$ wie σ_i verteilt, und wir haben insgesamt gezeigt

$$\mathbb{E}\Big[\sup_{\|f\|_W \leqslant \lambda} |R_{n,\varphi}(f) - R_\varphi(f)|\Big] \leqslant 4\,\mathbb{E}\Big[\sup_{\|f\|_W \leqslant \lambda} \Big|\frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)\Big|\Big].$$

(c) Schließlich profitieren wir noch von der Hilbertraumstruktur des RK-HS, indem wir mit der Cauchy-Schwarz-Ungleichung abschätzen:

$$\begin{aligned}\sup_{\|f\|_W \leqslant \lambda} \Big|\frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)\Big|^2 &= \sup_{\|f\|_W \leqslant \lambda} \Big|\frac{1}{n}\sum_{i=1}^n \langle f, \sigma_i k(X_i, \bullet)\rangle_W\Big|^2 \\ &\leqslant \sup_{\|f\|_W \leqslant \lambda} \|f\|_W^2 \Big\|\frac{1}{n}\sum_{i=1}^n \sigma_i k(X_i, \bullet)\Big\|_W^2 \\ &= \lambda^2 \frac{1}{n^2}\sum_{i,j=1}^n \sigma_i\sigma_j k(X_i, X_j).\end{aligned}$$

Dies impliziert wegen $\mathbb{E}[\sigma_i\sigma_j] = \delta_{i,j}$

$$\begin{aligned}\mathbb{E}\Big[\sup_{\|f\|_W \leqslant \lambda} \Big|\frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)\Big|\Big] &\leqslant \lambda\,\mathbb{E}\Big[\frac{1}{n^2}\sum_{i,j=1}^n \sigma_i\sigma_j k(X_i, X_j)\Big]^{1/2} \\ &= \lambda n^{-1/2}\sqrt{\mathbb{E}[k(X,X)]}.\end{aligned}$$

Zusammen mit den obigen Überlegungen ergibt dies die Behauptung.

□

7.15 Bemerkung. Der Restterm ist von der Ordnung $O(n^{-1/2})$ in n, die Komplexität der Klassifiziererfamilie wird durch $\lambda\,\mathbb{E}[k(X,X)]^{1/2}$ kontrolliert. Offensichtlich sollte der Radius λ so gewählt werden, dass Approximationsfehler und stochastischer Fehler sich in etwa austarieren. Eine adaptive Wahl von λ wird häufig über ein Kreuzvalidierungskriterium getroffen.

Literatur

Brown, L. D., and M. G. Low (1996): "Asymptotic equivalence of nonparametric regression and white noise.," *Ann. Stat.*, 24(6), 2384–2398.

De Boor, C. (2001): *A practical guide to splines. Rev. ed.* Applied Mathematical Sciences. 27. New York, NY: Springer.

Efromovich, S. (1999): *Nonparametric curve estimation. Methods, theory, and applications.* Springer Series in Statistics. New York.

Elstrodt, J. (2007): *Maß- und Integrationstheorie. 5. Auflage.* Springer-Lehrbuch. Berlin: Springer.

Fan, J., and I. Gijbels (1996): *Local polynomial modelling and its applications.* Monographs on Statistics and Applied Probability. 66. London: Chapman & Hall.

Györfi, L., M. Kohler, A. Krzyżak, and H. Walk (2002): *A distribution-free theory of nonparametric regression.* Springer Series in Statistics. New York, NY: Springer.

Hall, P., and J. S. Marron (1987): "Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation.," *Probab. Theory Relat. Fields*, 74, 567–581.

Härdle, W. (1991): *Applied nonparametric regression.* Econometric Society Monographs. 19. Cambridge: Cambridge University Press.

Härdle, W., G. Kerkyacharian, D. Picard, and A. Tsybakov (1998): *Wavelets, approximation, and statistical applications.* Springer, Berlin.

Hastie, T., R. Tibshirani, and J. Friedman (2001): *The elements of statistical learning. Data mining, inference, and prediction.* Springer Series in Statistics. New York, NY: Springer.

Houdré, C., and P. Reynaud-Bouret (2003): "Exponential inequalities, with constants, for U-statistics of order two.," Giné, Evariste (ed.) et al., Stochastic inequalities and applications. Selected papers presented at the Euroconference on "Stochastic inequalities and their applications", Barcelona, June 18–22, 2002. Basel: Birkhäuser. Prog. Probab. 56, 55-69 (2003).

Ledoux, M., and M. Talagrand (1991): *Probability in Banach spaces. Isoperimetry and processes.* Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge, 23. Berlin etc.: Springer-Verlag.

Lehmann, E., and G. Casella (1998): *Theory of point estimation. 2nd ed.* Springer Texts in Statistics. New York, NY: Springer.

Lepski, O. V. (1990): "One problem of adaptive estimation in Gaussian white noise.," *Theory Probab. Appl.*, 35, 459–470.

Massart, P. (2007): *Concentration inequalities and model selection. Ecole d'Eté de Probabilités de Saint-Flour XXXIII – 2003.* Lecture Notes in Mathematics 1896. Berlin: Springer.

Nason, G. P. (2008): *Wavelet methods in statistics with R.* Use R!. New York, NY: Springer.

Shiryaev, A. (1995): *Probability. 2nd ed.* Graduate Texts in Mathematics. 95. New York, Springer.

Silverman, B. (1986): *Density estimation for statistics and data analysis.* Monographs on Statistics and Applied Probability. London - New York: Chapman and Hall.

Spokoiny, V., and C. Vial (2009): "Parameter tuning in pointwise adaptation using a propagation approach.," *Ann. Stat.*, 37(5b), 2783–2807.

Steinwart, I., and A. Christmann (2008): *Support vector machines.* Information Science and Statistics. New York, NY: Springer.

Stone, C. J. (1984): "An asymptotically optimal window selection rule for kernel density estimates.," *Ann. Stat.*, 12, 1285–1297.

Tsybakov, A. B. (2004): *Introduction à l'estimation non-paramétrique.* Mathématiques & Applications (Paris). 41. Paris: Springer.

——— (2009): *Introduction to nonparametric estimation.* Springer Series in Statistics.

Wand, M., and M. Jones (1995): *Kernel smoothing.* Monographs on Statistics and Applied Probability. 60. London: Chapman & Hall.

Wasserman, L. (2006): *All of nonparametric statistics.* Springer Texts in Statistics. New York, Springer.

Werner, D. (2007): *Funktionalanalysis. 6. Auflage.* Springer-Lehrbuch. Berlin: Springer.

Wojtaszczyk, P. (1997): *A mathematical introduction to wavelets.* London Mathematical Society Student Texts. 37. Cambridge University Press.