Nichtparametrische Statistik
Skript zur Vorlesung im Wintersemester 2012/13

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

VORLÄUFIGE FASSUNG: 18. Oktober 2012

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

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

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.

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.