Notation

Bevor wir mit dem eigentlichen Inhalt beginnen, ist es notwendig, uns auf einige grundlegende Notationen zu einigen, damit der Textfluss nicht unnötig unterbrochen wird und die einzelnen Kapitel auch unabhängig voneinander betrachtet werden können.

- $d \in \mathbb{N}$ beschreibt in allen Fällen die Dimension des betrachteten Raumes
- Die Indikatorfunktion ist
$$\mathbb{1}_A(x) = \mathbb{1}_{x \in A} = \begin{cases} 1 & \text{falls } x \in A \\ 0 & \text{sonst} \end{cases}$$
- $\lambda = \lambda^d$ ist das Lebesgue-Maß, die Dimension d werden wir in der Notation meist unterdrücken
- $L_p = L_p\left(\mathbb{R}^d\right)$ mit $p \geq 1$, die Räume der p-fach λ-integrierbaren Funktionen auf $\mathbb{R}^d$. Wir werden den zugrundeliegenden Raum in der Notation meist unterdrücken
- Wenn wir über den gesamten Raum $\mathbb{R}$ oder $\mathbb{R}^d$ integrieren werden wir dies in der Regel nur notieren, wenn es aus dem Kontext heraus nicht absolut klar ist. Außerdem lassen wir der besseren Übersicht halber an manchen Stellen die Integrationsvariable weg:
$$\int_{\mathbb{R}^d} f(x)dx = \int f(x)dx = \int f$$
- $\partial_x = \frac{\partial}{\partial x}$ ist die partielle Ableitung in x-Richtung, ∇ ist der Gradienten-Operator und Δ der Laplace-Operator
- $S_{a,b} = \left\{x \in \mathbb{R}^d | \|a - x\| \leq b\right\}$ ist die abgeschlossene Sphäre um a mit Radius b
- $c_d = \lambda(S_{0,1})$ ist das Volumen der Einheitssphäre in Dimension d
- $supp(f) = \left\{x \in \mathbb{R}^d | f(x) \neq 0\right\}$ ist der Träger der Funktion f
- Nullmengen und "fast alle" beziehen sich – wenn nicht näher spezifiziert – immer auf das Lebesgue-Maß, "fast sicher" immer auf das gerade betrachtete Wahrscheinlichkeitsmaß
- $\mathfrak{B}(\mathbb{R}^d)$ ist die Menge aller Borelmengen auf $\mathbb{R}^d$

- ϕ_{μ,σ^2} ist die Dichte einer $N(\mu, \sigma^2)$-Verteilung, $\phi = \phi_{0,1}$ (außer im Kontext der Fourier-Transformierten)
- Die Faltung zweier Funktionen $f, g \in L_1$ ist definiert durch:

$$f * g(x) = \int f(x-y)g(y)dy = \int f(y)g(x-y)dy$$

1 Einführung in die Kerndichteschätzung

Wir betrachten ein klassisches Problem der Statistik: Wir haben n unabhängige, identisch verteilte Datenpunkte oder Beobachtungen $X_1, \ldots, X_n$ in $\mathbb{R}^d$, $d \geq 1$ einer uns unbekannten Verteilung $\mathbb{P}^X$ gegeben und wollen nun die Wahrscheinlichkeitsdichte f bezüglich des Lebesgue-Maßes bestimmen – diese Notation dient als Argumentationsgrundlage für den vorliegenden Abschnitt. Dabei setzen wir natürlich voraus, dass f überhaupt existiert.

Dem parametrischen Ansatz folgend wäre nun $\mathbb{P}^X$ aufgrund von Vorwissen oder Schätzungen einer bestimmten Klasse von Verteilungen zuzuordnen – etwa den Normalverteilungen. Daran anschließend würde man die zugehörigen Parameter – in diesem Fall also den Erwartungswert und die Varianz – schätzen. Wie wir weiter unten sehen werden, birgt dieser Ansatz einige Nachteile, weshalb hier der nichtparametrische Fall untersucht wird.

Zur Herleitung werden wir zuerst nur den Fall $d = 1$ betrachten und dann auf höhere Dimensionen verallgemeinern. Um die Dichte zu schätzen betrachten wir zuerst die zugehörige Verteilungsfunktion F. Der natürliche Schätzer für diese Größe ist die empirische Verteilungsfunktion $F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty,x]}(X_i)$, die für jedes feste $x \in \mathbb{R}$ auch erwartungstreu ist: $\mathbb{E}(F_n(x)) = \mathbb{E}(\mathbb{1}_{(-\infty,x]}(X_1)) = F(x)$.

Da $f = F'$ liegt es zunächst nahe, auch die empirische Verteilungsfunktion abzuleiten, was allerdings nicht weit führt; F_n ist nur fast überall differenzierbar und die resultierende schwache Ableitung ist fast überall gleich 0, wir kommen also auf diesem Weg zu keiner Dichtefunktion. Stattdessen können wir jedoch die Ableitung durch den zweiseitigen Differenzenquotienten approximieren:

$$\begin{aligned} f(x) &\approx \frac{1}{h}\left(F\left(x+\frac{h}{2}\right) - F\left(x-\frac{h}{2}\right)\right) \\ &\approx \frac{1}{h}\left(F_n\left(x+\frac{h}{2}\right) - F_n\left(x-\frac{h}{2}\right)\right) \\ &= \frac{1}{nh}\sum_{i=1}^{n} \mathbb{1}_{\left(x-\frac{h}{2}, x+\frac{h}{2}\right]}(X_i) \\ &= \frac{1}{nh}\sum_{i=1}^{n} \mathbb{1}_{\left(-\frac{1}{2}, \frac{1}{2}\right]}\left(\frac{x-X_i}{h}\right), \end{aligned}$$

wobei $h > 0$ ein sogenannter Glättungsfaktor ist, meist auch Bandweite genannt; dazu später mehr. Das Ergebnis ist eine stückweise konstante

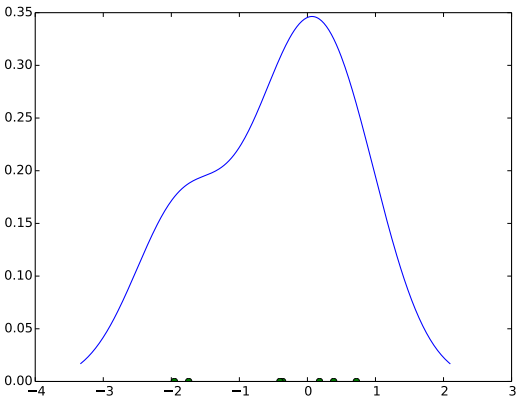

Abbildung 1: Ein Kerndichteschätzer für $n = 7$ und $f \sim N(0,1)$.

Funktion, die aber offensichtlich eine Dichte ergibt: Sie ist nichtnegativ und $\frac{1}{h} \int \mathbb{1}_{\left(-\frac{1}{2},\frac{1}{2}\right]} \left(\frac{x-X_1}{h}\right) dx = 1$. Um das Ergebnis weiter zu 'verstetigen' und zu 'glätten', können wir die Funktion $\mathbb{1}_{\left(-\frac{1}{2},\frac{1}{2}\right]}(\cdot)$, die die Dichte einer $Unif\left(\left(-\frac{1}{2},\frac{1}{2}\right]\right)$-Verteilung darstellt, durch einen sogenannten Kern K ersetzen. Damit können wir den eindimensionalen Kerndichteschätzer definieren:

Definition. Kerndichteschätzer (univariat)

Eine Funktion $f_n : \mathbb{R} \to \mathbb{R}$ heißt (eindimensionaler) Kerndichteschätzer, abgekürzt auch KDE (für Kernel Density Estimator), falls

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

mit $h > 0$ und $K : \mathbb{R} \to \mathbb{R}$ einer Funktion mit $\int K = 1$.

Meist wird es sich bei dem Kern aber auch um eine W-Dichte handeln, die außerdem noch symmetrisch um den Ursprung ist. Ein typischer Kern ist die Dichte der Standardnormalverteilung, der Gaußkern (siehe Abschnitt 1.3.1). Abbildung 1 zeigt einen solchen Kerndichteschätzer für 7 Datenpunkte einer Standardnormalverteilung, normalerweise sind allerdings deutlich mehr Beobachtungen für ein aussagekräfiges Ergebnis notwendig.

Der Schätzer ist im Allgemeinen auch von h abhängig, was in der Notation jedoch unterdrückt wird. Offensichtlich vererben sich die Stetigkeits-,

Integrierbarkeits- und Differenzierbarkeitseigenschaften des Kerns auch auf f_n, womit unter anderem folgt, dass auch f_n eine W-Dichte ist.

Der mehrdimensionale Fall wird hieraus abgeleitet:

Definition. Kerndichteschätzer (multivariat), nach [11]

Eine Funktion $f_n : \mathbb{R}^d \to \mathbb{R}$ heißt (mehrdimensionaler) Kerndichteschätzer, falls

$$f_n(x) := \frac{1}{nh_1 \cdots h_d} \sum_{i=1}^{n} \prod_{l=1}^{d} K_l \left(\frac{x_l - X_i^l}{h_l} \right),$$

mit $h_l > 0$ und $K_l : \mathbb{R} \to \mathbb{R}$, so dass

$$\begin{aligned} 0 \leq K_l \leq c < \infty, \\ K_l(y) = K_l(-y) \text{ für alle } y \in \mathbb{R} \\ \int K_l(y) y^2 dy < \infty \\ \int K_l(y) dy = 1. \end{aligned} \tag{1}$$

für alle $l = 1, \ldots, d$.

Tatsächlich gibt es eine Reihe von Variationen für die die Anforderungen an die K_l abgemildert werden können oder für die der Kern anders aussieht, aber der hier definierte Fall ist für unsere Zwecke allgemein genug. Ein häufiger Spezialfall tritt auf, wenn

$$\begin{aligned} h := h_1 = \ldots = h_d \\ \kappa := K_1 = \ldots = K_d \\ f_n(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K \left(\frac{x - X_i}{h} \right) \end{aligned} \tag{2}$$

mit $K(x) := \prod_{l=1}^{d} \kappa(x_l)$

Die verschiedenen Bandweiten in den verschiedenen Dimensionen sind insbesondere in der Praxis relevant, etwa wenn die Dimensionen verschieden skaliert sind, für theoretische Resultate jedoch meist unerheblich. Stattdessen kann man den Datensatz auf Kosten der Anschaulichkeit auch auf $(-1, 1)^d$ zentrieren und skalieren.

1.1 Verschiedene Fehlermaße

Um die Güte eines Dichteschätzers bewerten zu können brauchen wir zuerst ein passendes Fehlermaß. Die richtige Wahl hierfür kann zu unterschiedlichen

Resultaten führen und sollte nicht unterschätzt werden. Das natürlichste Maß für die Entfernung zweier Dichten g und h ist der L_1-Fehler, $\int |g-h|$. Den Gedanken kann man weiterführen und für beliebiges $p \in [1, \infty]$ den L_p-Fehler betrachten – der dann jedoch nicht mehr notwendigerweise endlich ist. Insbesondere die Fälle $p = 2$ und $p = \infty$ liegen intuitiv nah, wobei sich gerade der erste Fall in der Literatur stark durchgesetzt hat. Anstatt des L_2-Fehlers werden wir im Folgenden der Einfachheit halber ausschließlich dessen Quadrat verwenden, an dem Verhalten ändert sich nichts.

Der Vorteil des L_1-Fehlers liegt vor allem darin, dass er immer wohldefiniert ist – insbesondere gilt $\int |g-h| \leq \int g + \int h = 2$ – und dass er die tatsächliche, anschauliche Entfernung zwischen zwei Dichten beschreibt. Andererseits lässt es sich in praktischen Anwendungen schwerer mit dem L_1-Fehler rechnen. Für den L_2-Fehler gibt es dagegen einige gute Rechentechniken wie wir später sehen werden. Es ist jedoch klar, dass er die Regionen unterbetont, in denen $|g-h|$ tendenziell kleiner ist und die anderen dafür überbetont. Gerade in höheren Dimensionen kann das zu Problemen führen.

Der L_p-Fehler betrachtet immer nur den Fehler für zwei bestimmte Dichten. Wenn wir die Eigenschaften von auf zufälligen Daten beruhenden Kerndichteschätzern untersuchen, müssen wir dagegen den Erwartungswert betrachten:

Definition. MIAE und MISE

Sei f eine Wahrscheinlichkeitsdichte und f_n ein Schätzer für diese Dichte, dann gelten die Bezeichnungen:

$$\text{MIAE} = \mathbb{E}\left(\int |f - f_n|\right) \text{ (\textbf{M}ean \textbf{I}ntegrated \textbf{A}bsolute \textbf{E}rror)}$$
$$\text{MISE} = \mathbb{E}\left(\int (f - f_n)^2\right) \text{ (\textbf{M}ean \textbf{I}ntegrated \textbf{S}quare \textbf{E}rror)}$$

In diesem Teil werden wir nur den MISE betrachten, was für praktische Zwecke ausreichend ist, im zweiten Teil dagegen werden wir ein Konsistenzresultat für den MIAE zeigen.

1.2 Nichtparametrische oder parametrische Schätzer

Es stellt sich die Frage, weshalb überhaupt nichtparametrische Schätzer benötigt werden, denn tatsächlich werden in der Praxis meist parametrische Schätzer verwendet. Die Vorteile der parametrischen Schätzer liegen zum einen darin, dass man viele bekannte Eigenschaften von bekannten

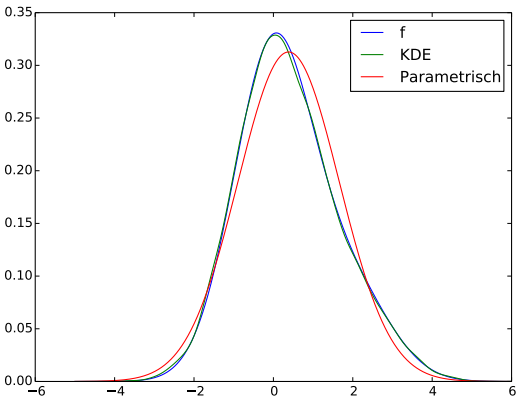

Abbildung 2: Vergleich von parametrischem und nichtparametrischem Schätzer bei leichten Unsymmetrien; $f = 0.8\phi_{0,1} + 0.2\phi_{2,1}$.

Verteilungsfamilien verwenden kann und viele theoretische Resultate über diese Verteilungen vorliegen. Zum anderen ist die Berechnung einer kleinen Anzahl von Parametern auch bei großen Datensätzen meist ohne Probleme möglich und danach ist eine geschlossene, analytische Form der Dichte vorhanden.

Bei Kerndichteschätzern hingegen sind sämtliche Datenpunkte für die Form des Schätzers relevant und müssen daher immer zugreifbar sein. Obwohl beispielsweise Kerne mit kompaktem Träger den Rechen- und Speicheraufwand erheblich minimieren können, ist dieser Mehraufwand nicht unerheblich und muss bei großen Datensätzen beachtet werden. Ein möglicher Lösungsansatz für den Rechenaufwand ist es, den Kerndichteschätzer durch ein Polynom zu interpolieren; mit modernen Algorithmen und ausreichend hohem Polynomgrad genügt die resultierende Funktion in der Praxis durchaus den meisten Anforderungen und ist leichter zu berechnen.

Für die nichtparametrischen Schätzer spricht vor allem ihre gute Anpassungsfähigkeit. Es müssen keine beschränkenden Annahmen über die Verteilung getroffen werden und viele Details der tatsächlich vorliegenden Verteilung fallen nicht unter den Tisch. Ein Beispiel hierfür sind leichte Unsymmetrien um den Mittelwert herum bei sonst scheinbar normalverteilten Zufallsvariablen – Abbildung 2 veranschaulicht das Problem. Insbesondere für ungeübte Praktiker können sich Kerndichteschätzer daher als vorteilhaft erweisen.

Aus theoretischer Sicht und bei sehr großen Datensätzen ist zu bedenken, dass parametrische Schätzer natürlich nur für die vorher spezifizierten Verteilungen auch konsistent sind – wir werden zeigen, dass Kerndichteschätzer dagegen schon bei minimalen Anforderungen an die Bandweite für alle Dichten konsistent sind; siehe Kapitel 2.

1.3 Der MISE und die Wahl von Kern und Bandweite

Wir stellen eine kleine Vorüberlegung an: Kern und Bandweite sollten möglichst so gewählt werden, dass der Fehler minimiert wird. Wie bereits erwähnt betrachten wir hier der Einfachheit halber den MISE, der minimiert werden soll. Da dieser selbst aber auch noch zu unhandlich ist, werden wir nur eine Approximation verwenden, die jedoch für ausreichend große Stichproben angemessen ist. Wir betrachten nur den bereits erwähnten Spezialfall in Gleichung (2) mit gleichem Kern und gleicher Bandweite in allen Dimensionen.

Es gilt:

$$\begin{aligned}\text{MISE} &= \int \mathbb{E}\left((f(x) - f_n(x))^2\right) dx \\ &= \int Var(f(x) - f_n(x))dx + \int \left(\mathbb{E}(f_n(x)) - f(x)\right)^2 dx \\ &= \int Var(f_n(x))dx + \int bias_h(x)^2 dx,\end{aligned}$$

wobei $bias_h(x) = \mathbb{E}\left(f_n(x)\right) - f(x)$ also ein Maß für die tendenzielle Abweichung des Schätzers in x angibt und $Var(f_n(x))$ die zugehörige Streuung beschreibt.

Wir approximieren nun beide Terme einzeln und definieren eine Approximation an den MISE:

Satz und Definition 1.1 (Epanechnikov [11], Silverman [23])**.** *Sei f eine mindestens zweimal stetig differenzierbare Wahrscheinlichkeitsdichte, f_n ein Kerndichteschätzer wie in (2) und $h = h(n)$ eine monoton fallende Nullfolge von Bandweiten. Mit der abkürzenden Notation*

$$k_1 := k_1(K) := \int_{\mathbb{R}} \kappa(y) y^2 dy \text{ und } k_2 := k_2(K) := \int_{\mathbb{R}^d} K(t)^2 dt$$

gilt:

$$\int bias_h(x)^2 dx = \frac{1}{4} h^4 k_1^2 \int \left(\Delta f(x)\right)^2 + O(h^5) \tag{3}$$

$$\int Var(f_n(x))dx = n^{-1} h^{-d} k_2 + O(n^{-1} h^{-d+2}). \tag{4}$$

Daher definieren wir den asymptotischen (oder approximierten) MISE:

$$AMISE = n^{-1}h^{-d}k_2 + \frac{1}{4}h^4k_1^2\int (\Delta f(x))^2$$

Beweis. Wir zeigen zuerst Gleichung (3).

$$\begin{aligned} bias_h(x) &= \mathbb{E}(f_n(x)) - f(x) \\ &= n^{-1}\sum_{i=1}^{n}\mathbb{E}\left(h^{-d}K\left(\frac{x-X_i}{h}\right)\right) - f(x) \\ &= \int h^{-d}K\left(\frac{x-y}{h}\right)f(y)dy - f(x) \\ &= \int K(t)f(x-ht)dt - f(x)\int K(t)dt \quad \text{für } y = x-ht,\ t\in\mathbb{R}^d \\ &= \int K(t)(f(x-ht)-f(x))dt. \end{aligned} \tag{5}$$

Für $f(x-ht)$ führen wir eine mehrdimensionale Taylor-Approximation zweiten Grades durch:

$$f(x-ht) = f(x) - h\langle\nabla f(x), t\rangle + \frac{1}{2}h^2\sum_{j=1}^{d}\sum_{k=1}^{d}\frac{\partial^2 f(x)}{\partial x_j\partial x_k}t_jt_k + O(h^3),$$

wobei $\langle\cdot,\cdot\rangle$ das Skalarprodukt im $\mathbb{R}^d$ ist. Wir können Gleichung (5) also weiterführen:

$$bias_h(x) = \int K(t)\left(-h\langle\nabla f(x), t\rangle + \frac{1}{2}h^2\sum_{j=1}^{d}\sum_{k=1}^{d}\frac{\partial^2 f(x)}{\partial x_j\partial x_k}t_jt_k\right)dt + O(h^3).$$

Der erste Teil davon verschwindet, denn:

$$\begin{aligned} \int_{\mathbb{R}^d} hK(t)\langle\nabla f(x), t\rangle dt &= \sum_{j=1}^{d} h\partial_{x_j}f(x)\int_{\mathbb{R}^d}K(t)t_jdt \\ &= \sum_{j=1}^{d} h\partial_{x_j}f(x)\int_{\mathbb{R}}\kappa(t_d)\cdots\int_{\mathbb{R}}\kappa(t_1)t_jdt_1\cdots dt_d \\ &= \sum_{j=1}^{d} h\partial_{x_j}f(x)\int_{\mathbb{R}}\kappa(t_j)t_jdt_j \\ &= 0, \end{aligned}$$

wobei wir zweimal Fubini und im letzten Schritt die Symmetrie von κ benutzt haben.

Den hinteren Teil können wir auch vereinfachen:

$$\begin{aligned}
\int_{\mathbb{R}^d} \frac{1}{2}h^2 K(t) \sum_{j=1}^{d}\sum_{k=1}^{d} \frac{\partial^2 f(x)}{\partial x_j \partial x_k} t_j t_k dt &= \frac{1}{2}h^2 \sum_{j=1}^{d}\sum_{k=1}^{d} \frac{\partial^2 f(x)}{\partial x_j \partial x_k} \int_{\mathbb{R}^d} K(t) t_j t_k dt \\
&= \frac{1}{2}h^2 \sum_{i=1}^{d} \frac{\partial^2 f(x)}{\partial x_i^2} \int_{\mathbb{R}} \kappa(t_i) t_i^2 dt_i \\
&+ h^2 \sum_{j=1}^{d}\sum_{k<j} \frac{\partial^2 f(x)}{\partial x_j \partial x_k} \int_{\mathbb{R}} \kappa(t_j) t_i dt_j \int_{\mathbb{R}} \kappa(t_k) t_k dt_k \\
&= \frac{1}{2}h^2 \sum_{i=1}^{d} \frac{\partial^2 f(x)}{\partial x_i^2} \int_{\mathbb{R}} \kappa(t_i) t_i^2 dt_i, \\
&= \frac{1}{2}h^2 \Delta f(x) k_1,
\end{aligned}$$

wobei wir auch hier wieder Fubini und die Symmetrie von κ verwenden. Insgesamt haben wir

$$bias_h(x) = \frac{1}{2}h^2 \Delta f(x) k_1 + O(h^3).$$

Quadrieren und integrieren liefert Gleichung (3)

Um Gleichung (4) zu zeigen gehen wir ähnlich vor. Es gilt

$$\begin{aligned}
Var(f_n(x)) &= n^{-1} Var\left(h^{-d} K\left(\frac{x - X_1}{h}\right)\right) \\
&= n^{-1}\mathbb{E}\left(h^{-2d} K\left(\frac{x - X_1}{h}\right)^2\right) - n^{-1}(\mathbb{E}\left(h^{-d} K\left(\frac{x - X_1}{h}\right)\right)^2 \\
&= n^{-1}\int h^{-2d} K\left(\frac{x - y}{h}\right)^2 f(y) dy - n^{-1}(f(x) + bias_h(x))^2 \\
&= n^{-1}h^{-d}\int K(t)^2 f(x - ht) dt + O(n^{-1}).
\end{aligned}$$

Der hintere Term ist $O(n^{-1})$, da $f(x)+bias_h(x)$ in n monoton fällt. Wieder ersetzen wir $f(x - ht)$ durch eine Taylor-Approximation, diesmal aber erster Ordnung:

$$f(x - ht) = f(x) - h\langle \nabla f(x), t\rangle + O(h^2),$$

und damit:

$$
\begin{aligned}
n^{-1}h^{-d}\int K(t)^2 f(x-ht)dt &= n^{-1}h^{-d}\int K(t)^2(f(x)-h\langle\nabla f(x),t\rangle+O(h^2))dt\\
&= n^{-1}h^{-d}f(x)\int K(t)^2dt\\
&\quad - n^{-1}h^{-d+1}\sum_{i=1}^{d}\partial_{x_i}f(x)\int_{\mathbb{R}^d}K(t)^2t_i dt + O(n^{-1}h^{-d+2})\\
&= n^{-1}h^{-d}f(x)k_2\\
&\quad - n^{-1}h^{-d+1}\sum_{i=1}^{d}\partial_{x_i}f(x)\int_{\mathbb{R}}\kappa(t_i)^2t_i dt_i + O(n^{-1}h^{-d+2})\\
&= n^{-1}h^{-d}f(x)k_2 + O(n^{-1}h^{-d+2}),
\end{aligned}
$$

wobei im letzten Schritt der hintere Term wieder aufgrund der Symmetrie verschwindet. Es gilt $O(n^{-1}h^{-d+2}) + O(n^{-1}) = O(n^{-1}h^{-d+2})$ und für den zu minimierenden MISE folgt somit:

$$
\begin{aligned}
&\int Var(f_n(x))dx + \int bias_h(x)^2dx\\
&= \int n^{-1}h^{-d}f(x)k_2dx + O(n^{-1}h^{-d+2})\\
&+ \int \frac{1}{4}h^4\left(\Delta f(x)\right)^2k_1^2dx + O(h^5)\\
&= n^{-1}h^{-d}k_2 + \frac{1}{4}h^4k_1^2\int(\Delta f(x))^2dx + O(n^{-1}h^{-d+2}+h^5).
\end{aligned}
$$

□

Satz 1.1 zeigt ein grundlegendes Problem bei der Wahl der Bandweite auf: Je kleiner die Bandweite gewählt ist, desto geringer ist der Bias, aber dafür wird die Varianz größer. Diese Schwierigkeit ist auch unter dem Begriff "Bias-Variance-Tradeoff" bekannt.

Den Tiefpunkt für $h(z) = az^\alpha + bz^{-\beta}$, $z, a, b > 0$ und $\alpha, \beta \geq 1$, finden wir durch Ableiten:

$$
\begin{aligned}
h'(z) &= a\alpha z^{\alpha-1} - b\beta z^{-\beta-1} = 0\\
&\Leftrightarrow z = \left(\frac{\beta b}{\alpha a}\right)^{\frac{1}{\alpha+\beta}}\\
h''(z) &= \alpha(\alpha-1)az^{\alpha-2} + \beta(\beta+1)bz^{-\beta-2} > 0.
\end{aligned}
$$

Dementsprechend optimiert also

$$h_{opt} = \left(dk_2 k_1^{-2} \left(\int (\Delta f)^2 \right)^{-1} n^{-1} \right)^{\frac{1}{d+4}}$$

den AMISE. Auffällig ist, dass h_{opt} als Funktion von n bereits für kleines d nur sehr langsam gegen 0 konvergiert. Wie zu erwarten war ist h_{opt} auch von der unbekannten Dichte f abhängig, was die Wahl der Bandweite in der Praxis problematisch macht; k_1 und k_2 sind jedoch im Allgemeinen ohne Probleme bestimmbar, entweder durch analytische oder numerische Verfahren, da K bekannt ist. Eine einfache Methode zur Wahl von h in der Praxis stellen wir nach einer genaueren Betrachtung des Kerns dar.

1.3.1 Wahl des Kerns

Wir betrachten zuerst nur den eindimensionalen Fall und schränken uns zusätzlich ein, indem wir für einen Kern $\int K(x)xdx = 0$ voraussetzen (was insbesondere für symmetrische Kerne erfüllt ist) und wie für den Multivariaten Kerndichteschätzer $K \in L_2$ fordern. Für den approximierten Fehler können wir dann $h_{opt} = (k_2 k_1^{-2} n^{-1} (\int f''^2)^{-1})^{\frac{1}{5}}$ einsetzen und erhalten:

$$\text{AMISE}_1 = n^{-1}h^{-1}k_2 + \frac{1}{4}h^4 k_1^2 \int f''(x)^2 dx = \frac{5}{4} n^{-\frac{4}{5}} k_2^{\frac{4}{5}} k_1^{\frac{2}{5}} \left(\int f''^2 \right)^{\frac{1}{5}}.$$

Wir suchen also einen Kern K, der $C(K) := k_2^{\frac{4}{5}} k_1^{\frac{2}{5}}$ minimiert. Wir können ohne Einschränkungen so skalieren, dass $k_1 = 1$, indem wir ihn durch

$$K^s(t) := k_1^{\frac{1}{2}} K\left(k_1^{\frac{1}{2}} t \right) \tag{6}$$

ersetzen, denn

$$\begin{aligned} \int k_1^{\frac{1}{2}} K\left(k_1^{\frac{1}{2}} y \right) dy &= \frac{1}{k_1^{\frac{1}{2}}} \int K\left(k_1^{\frac{1}{2}} y \right) \left(k_1^{\frac{1}{2}} y \right)^2 dy \\ &= \frac{1}{k_1} \int K(z) z^2 dz \\ &= 1, \end{aligned}$$

außerdem bleibt $C(K)$ durch die Skalierung unverändert:

$$\begin{aligned} C(K^s) &= 1 \cdot \left(\int k_1 K \left(k_1^{\frac{1}{2}} t \right)^2 dt \right)^{\frac{4}{5}} \\ &= \left(\int k_1^{\frac{1}{2}} K(t)^2 dt \right)^{\frac{4}{5}} \\ &= k_1^{\frac{2}{5}} \left(\int K(t)^2 dt \right)^{\frac{4}{5}} \\ &= C(K), \end{aligned}$$

wobei wir natürlich $k_1 = k_1(K)$ haben. Mit anderen Worten heißt das, wir müssen nur diejenigen Kerne untersuchen, für die $k_1 = 1$ gilt, anstatt alle möglichen Kerne K zu betrachten; die Skalierung übernimmt die Bandweite. Damit ergibt sich ein Optimierungsproblem mit Nebenbedingungen ($k_1 = 1, \int K = 1, K \geq 0, \int K(x)xdx = 0$), das mit Lagrange-Multiplikatoren gelöst werden kann (siehe etwa [13], Gleichungen 1.10-1.13). Der daraus resultierende sogenannte Epanechnikov-Kern ist

$$K_e(x) = \frac{3}{4\sqrt{5}} \left(1 - \frac{1}{5} x^2 \right) \mathbb{1}_{[-\sqrt{5},\sqrt{5}]}(x).$$

Die Effizienz eines beliebigen anderen Kerns K_0 geben wir dazu im Verhältnis an, indem wir betrachten, wie viele Beobachtungen wir bei Benutzung von K_e benötigen, um den gleichen approximierten Fehler wie bei Benutzung von K_0 zu machen:

$$\begin{aligned} &\frac{5}{4} n_e^{-\frac{4}{5}} C(K_e) \left(\int f''^2 \right)^{\frac{1}{5}} = \frac{5}{4} n_0^{-\frac{4}{5}} C(K_0) \left(\int f''^2 \right)^{\frac{1}{5}} \\ &\Leftrightarrow n_e = \left(\frac{C(K_e)}{C(K_0)} \right)^{\frac{5}{4}} n_0, \end{aligned}$$

also

Definition. Effizienz eines Kerns

Für eine beliebige Funktion $K_0 : \mathbb{R} \to \mathbb{R}$, die die Bedingungen in (1) erfüllt, ist die Effizienz durch

$$\mathit{eff}(K_0) := \left(\frac{C(K_e)}{C(K_0)} \right)^{\frac{5}{4}}$$

definiert.

Name	Definition	Effizienz
Epanechnikov	$\frac{3}{4\sqrt{5}}\left(1-\frac{1}{5}x\right)\mathbb{1}_{\left[-\sqrt{5},\sqrt{5}\right]}(x)$	1
Gauß (Normalverteilung)	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$	0.9512
Dreieckskern	$\frac{1}{\sqrt{6}}\left(1-\frac{x}{\sqrt{6}}\right)\mathbb{1}_{\left[-\sqrt{6},\sqrt{6}\right]}(x)$	0.9859
Gleichverteilung	$\frac{1}{2\sqrt{3}}\mathbb{1}_{\left[-\sqrt{3},\sqrt{3}\right]}(x)$	0.9295

Tabelle 1: Effizienz einiger skalierter Kerne. Die Daten zur Effizienz stammen aus [23], Abschnitt 3.3 und sind auf 4 Nachkommastellen genau.

Die Definition und Effizienz einiger häufig benutzter Kerne werden in Tabelle 1 festgehalten, die zugehörigen Graphen sind in Abbildung 3 aufgeführt. Es fällt sofort auf, dass die Effizienz bei allen aufgeführten Kernen sehr nahe an 1 liegt. Entsprechend bietet es sich an, die Wahl des Kerns anhand anderer Kriterien zu treffen, etwa anhand von Stetigkeits- und Differenzierbarkeitseigenschaften – hier bietet sich vor allem der Gaußkern an –, aber auch Berechnungseffizienz kann ausschlaggebend sein – hier bieten sich etwa Kerne mit kompaktem Träger an.

Tatsächlich ist die ganze Approximation des Fehlers auf die beschränkende Annahme gegründet, dass der Kern notwendigerweise selbst eine Wahrscheinlichkeitsdichte ist. Wenn man auch Kerne mit negativen Werten zulässt, so lässt sich zeigen, dass man die Konvergenzrate von $n^{-\frac{4}{5}}$ auf $n^{-\frac{2k}{2k+1}}$ für $k \in \mathbb{N}$ verbessern kann, also beliebig nahe an n^{-1} herankommt – für Details siehe [26] (Abschnitt 2.8). Ein Nachteil hierbei ist, dass der resultierende Kerndichteschätzer nicht mehr automatisch eine W-Dichte ist.

Für den mehrdimensionalen Fall liegen deutlich weniger ausgefeilte Ergebnisse vor. In unserer Definition des Kerndichteschätzers haben wir die restriktive Annahme eines Produktkerns getroffen. In der Literatur werden verschiedene Ansätze gewählt, der allgemeinste Fall generalisiert die Bandweite zu einer positiv definiten Matrix $H \in \mathbb{R}^{d\times d}$ und definiert $f_n(x) = n^{-1}\sum_{i=1}^{n} K_H(x - X_i)$ mit $K_H(x) = |H|^{-\frac{1}{2}} K\left(H^{-\frac{1}{2}}x\right)$ und K einer Funktion auf $\mathbb{R}^d$, die zu 1 integriert.

Ein typischer Kern, der mit jeder zugrundeliegenden Definition herleitbar ist, ist die multivariate Standardnormalverteilung.

Wir werden nicht hier nicht weiter auf die Eigenschaften von multivariaten Kernen eingehen und verweisen für mehr Details auf [26], Kapitel 4 und [23], Kapitel 4.

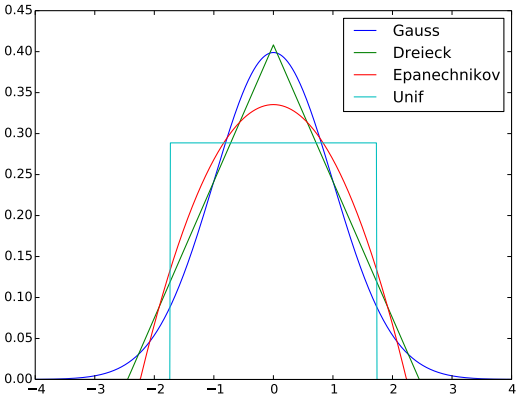

Abbildung 3: Vier weit verbreitete Kerne, skaliert wie in Gleichung (6).

1.3.2 Wahl der Bandweite

Die Wahl der Bandweite ist in der Literatur über Kerndichteschätzer das wohl am meisten untersuchte und komplexeste Thema. Im Rahmen dieser Arbeit ist es leider nicht möglich, eine ausführliche Betrachtung durchzuführen und wir werden nur eine einfache Methode für die Praxis angeben, die in [23] (Abschnitt 3.4.2) vorgeschlagen wurde.

Wieder betrachten wir den Fall $d = 1$ gesondert. Theoretisch haben wir mit h_{opt} bereits eine ausreichend gute Lösung, in der Praxis ist f'' allerdings nicht bekannt. Ein weiterer Nachteil ist, dass h_{opt} auch nur den approximierten Fehler optimiert, die Approximierung greift aber nur für große n und entsprechend kleine Größenordnungen von h. Wir gehen nicht weiter auf dieses Problem ein und versuchen eine Näherung für h_{opt} zu finden, indem wir $\int f''^2$ versuchen anzunähern.

Da wir nur an einem tendenziellen Wert interessiert sind, ist es naheliegend, f'' mit einer Standardverteilung parametrisch zu schätzen, etwa mit der Normalverteilung. Der Erwartungswert dieser Verteilung ist irrelevant, da wir ohnehin das Integral über ganz $\mathbb{R}$ betrachten. Wenn also $\hat{f}$ unser parametrischer Schätzer für $f - \int xf(x)dx$ ist, dann gilt $\hat{f}(x) = \frac{1}{\sigma}\phi(\frac{x}{\sigma})$, $\sigma > 0$ und $\hat{f}''(x) = \sigma^{-3}\phi''(\frac{x}{\sigma})$. Für das Integral gilt also unter Beachtung

von $\phi''(x) = (x^2 - 1)\phi(x)$:

$$\begin{aligned}
\int \hat{f}''(x)^2 dx &= \frac{1}{\sigma^6} \int \phi'' \left(\frac{x}{\sigma}\right)^2 dx \\
&= \frac{1}{\sigma^5} \int \phi''(x)^2 dx \\
&= \frac{1}{\sigma^5} \int \left(x^2 - 1\right)^2 \frac{1}{2\pi} \exp\left(-x^2\right) dx \\
&= \frac{1}{2\sigma^5\pi} \left(\int x^4 \exp\left(-x^2\right) dx - 2 \int x^2 \exp\left(-x^2\right) dx + \int \exp\left(-x^2\right) dx \right) \\
&= \frac{1}{2\sigma^5\pi} \left(\frac{3}{4}\sqrt{\pi} - \sqrt{\pi} + \sqrt{\pi} \right) \\
&= \frac{3}{8\sqrt{\pi}\sigma^5},
\end{aligned}$$

wobei wir im vorletzten Schritt die bekannten zweiten und vierten Momente einer $N\left(0, \frac{1}{2}\right)$-Verteilung benutzt haben, $\frac{1}{2}$ und $\frac{3}{4}$. σ kann jetzt aus den $X_1, \ldots X_n$ geschätzt werden und wir erhalten durch Einsetzen eine Näherung an h_{opt}, hier $\hat{h}_{opt}$ genannt.

Diese Methode ist nicht für jede Verteilung geeignet. $\int f''^2$ kann als ein Maß dafür angesehen werden, wie gleichmäßig f ist, wobei niedrige Werte mit gleichmäßigeren Dichten einhergehen. Die Normalverteilungen sind im Vergleich mit anderen Dichten allerdings häufig zu optimistisch und entsprechend $\int \hat{f}''^2$ zu klein und $\hat{h}_{opt}$ zu groß; der Schätzer wird "überglättet". Eine Möglichkeit damit umzugehen ist, anstatt einer Normalverteilung andere Verteilungsklassen zu betrachten, die den tatsächlichen Daten näher liegen. Ein anderer Ansatz benutzt $\hat{h}_{opt}$ lediglich als Startpunkt; damit plottet man dann den Kerndichteschätzer und versucht, h so anzupassen, dass er angemessen erscheint – dieser Vorgang ist natürlich subjektiv und minimiert im Allgemeinen auch nicht den tatsächlichen Fehler, kann aber nichtsdestotrotz in vielen Situationen völlig ausreichend sein.

Ein Beispiel findet sich in Abbildung 4: Für zu kleine Werte von h werden zufällige Schwankungen überbewertet und führen zu einer Verrauschung des Graphen. Zu große Werte dagegen führen dazu, dass wichtige Details übergangen werden; so ist etwa der Hochpunkt in -2 bei einer Bandweite von 1.69 nicht als solcher zu erkennen.

Gehen wir im mehrdimensionalen Fall wie oben auch wieder von gleicher Bandweite in allen Dimensionen aus, so haben wir im Prinzip das gleiche Problem wie im eindimensionalen Fall. h_{opt} hängt wieder von der tatsächlichen Verteilung ab, diesmal durch $\int (\Delta f)^2$. Auch dieses Funktional kann durch eine vorher festgelegte Verteilung geschätzt werden, etwa der multivariaten Normalverteilung. Da die Ergebnisse nicht vollständig exakt zu sein

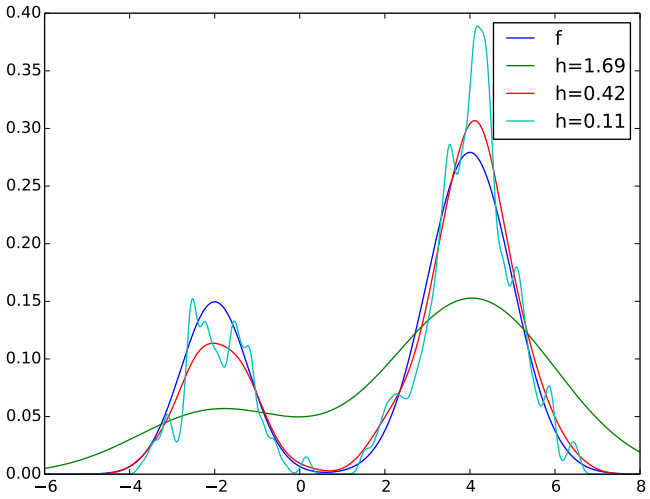

Abbildung 4: Kerndichteschätzer mit Gaußkern und verschiedenen Bandweiten für $f = 0.3\phi_{-2,0.8} + 0.7\phi_{4,1}$ und $n = 500$.

brauchen und $\int (\Delta f)^2$ auch nur einmal berechnet werden muss, bietet es sich gerade hier an, numerische Verfahren zur Lösung heranzuziehen.

Eine andere, weniger genaue Möglichkeit, die dafür jedoch direkt für jede einzelne Dimension passende Werte angibt, besteht darin, den im Eindimensionalen hergeleiteten Wert für $\hat{h}_{opt}$ mit einem Schätzer für die Standardabweichung entlang der einzelnen Achsen anzugeben, das heißt $h = (h_1, \ldots, h_d)$ mit $h_j = \frac{3}{8\sqrt{\pi}\hat{\sigma}_j^5}$ und $\hat{\sigma}_j^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i^j - \overline{X}_j)^2$, wobei $\overline{X} = n^{-1}\sum_{i=1}^n X_i$.

1.4 Varianten des Kerndichteschätzers

Während die hier definierten Kerndichteschätzer im Eindimensionalen in der Regel gute Ergebnisse liefern, werden im Höherdimensionalen vor allem die Ausläufer meist unterschätzt. Für große d verhält sich $\mathbb{R}^d$ kontraintuitiv; so beträgt das Verhältnis zwischen Einheitssphäre und dem umschließenden Quader mit Seitenlänge 2 für $d = 1$ und $d = 2$ beispielsweise 1 und $\frac{\pi}{4} \approx 0.79$, für $d = 10$ dagegen $\frac{\pi^5}{120 \cdot 2^{10}} \approx 0.0025$. Angewandt auf eine $Unif([-1,1]^{10})$-verteilte Zufallsvariable Y bedeutet das, dass $\mathbb{P}(Y \in S_{0,1}) = 0.0025$, während der überragende Anteil in den Ausläufern liegt – für andere Verteilungen liegen ähnliche Zahlen vor. Da jedoch in den Flanken bei den meisten

Verteilungen naturgemäß weniger Datenpunkte vorliegen, ist es schwierig, dort die Wahrscheinlichkeitsdichte vernünftig zu schätzen, ein Phänomen, das auch als "Curse of dimensionality" bekannt ist. Für mehr Details siehe etwa [22], Kapitel 7.

Um diesen Problemen entgegenzuwirken, bietet es sich an, den Glättungsfaktor h an die Region anzupassen, in der man $f(x)$ schätzen will. In dichteren Regionen wird h also kleiner gewählt um Details nicht zu "verschmieren", in den anderen Regionen wird h dagegen größer gewählt, um mehr Datenpunkte mit einfließen lassen zu können. Die zwei resultierenden Möglichkeiten sind, entweder $h = h(x)$ oder $h = h(X_i)$ zu wählen (wir betrachten wieder nur den Fall, dass $h_1 = \ldots = h_d = h$).

Im Folgenden werden wir zwei wichtige Abwandlungen des Kerndichteschätzers betrachten, die jeweils einen der beiden Ansätze verfolgen. Wir werden diese Methoden nur kurz einführen und deren Eigenschaften und genaueres Verhalten nicht detaillierter betrachten.

1.4.1 Nächste-Nachbarn-Verfahren

Die Nächste-Nachbarn-Methode verallgemeinert den Kerndichteschätzer auf die erste der oben beschriebenen Weisen, indem wir $h = h(x)$ wählen. Wir benötigen einen weiteren Parameter $k = k(n) \in \mathbb{N}$, der nur von n abhängig ist. Damit definieren wir $r_k(x) = h$ als den Abstand von x zum k-nächsten Nachbarn bezüglich der Metrik auf $\mathbb{R}^d$, das heißt in $S_{x,r_k(x)}$ befinden sich mindestens k Punkte aus $X_1, \ldots, X_n$. Da der Rand von $S_{x,r_k(x)}$ jedoch eine λ-Nullmenge ist, können wir den Fall, dass mehr als k Punkte in $S_{x,r_k(x)}$ liegen vorerst ignorieren. Das heißt, dass sich für fast alle $x \in \mathbb{R}^d$ genau k Punkte in der definierten Umgebung befinden. Wir wählen als Kern $K(x) = c_d^{-1} \mathbb{1}_{\|x\| \leq 1}$. Es ergibt sich damit:

Definition. k-Nächste-Nachbarn-Schätzer, nach [23]

Seien k, r_k und K definiert wie oben, dann heißt eine Funktion $f_{kNN} : \mathbb{R}^d \to \mathbb{R}$

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

der k-Nächste-Nachbarn-Schätzer für f.

Die Bezeichnung ergibt mehr Sinn, wenn man bedenkt, dass

$$\begin{aligned} f_{kNN}(x) &= n^{-1} r_k(x)^{-d} c_d^{-1} \sum_{i=1}^{n} \mathbb{1}_{\|x - X_i\| \leq r_k(x)} \\ &= \frac{k}{n c_d r_k(x)^d} \end{aligned}$$

für fast alle $x \in \mathbb{R}^d$.

Offensichtlich ist r_k als Funktion von x stetig, entsprechend also auch f_{kNN}, jedoch nur fast überall differenzierbar. Die wichtigste Eigenschaft liegt jedoch in der Integrierbarkeit: Da $r_k(x)$ für x im Betrag ausreichend groß linear verläuft, gilt $\int f_{kNN} = \infty$. Zum Schätzen der gesamten Dichte erweist sich f_{kNN} also als denkbar ungeeignet. Man kann jedoch zeigen, dass $f_{kNN}(x) \to_{n\to\infty} f(x)$ in Wahrscheinlichkeit für fast alle $x \in \mathbb{R}^d$, vorrausgesetzt dass $k(n) \to_{n\to\infty} 0$ und $\frac{k(n)}{n} \to_{n\to\infty} 0$, siehe etwa [17].

Anstatt des hier vorgeschlagenen $Unif(S_{0,1})$-verteilten Kerns kann man wie beim normalen Kerndichteschätzer auch wieder andere Wahrscheinlichkeitsdichten wählen, um glattere Ergebnisse zu erzielen. Diese Flexibilität wird allerdings durch ineffizientere Berechnung und komplexere (allerdings stärkere) theoretische Resultate erkauft.

Auf das Nächste-Nachbarn-Verfahren werden wir im dritten Kapitel als Grundlage für ein Klassifizierungsverfahren wieder zurückkommen.

1.4.2 Adaptive Kerndichteschätzer

Auch das Konzept vom adaptiven Kerndichteschätzer entfernt sich von einer statisch festgelegten Bandweite als einzigem Glättungsfaktor. Im Gegensatz zum Nächste-Nachbarn-Verfahren jedoch hängt der Faktor, den wir neu einführen, nicht mehr von dem Punkt an dem wir f schätzen ab, sondern von den $X_1, \ldots, X_n$. Wir definieren:

Definition. Adaptiver Kerndichteschätzer, nach [23]
Sei $h > 0$, K wie in (1),

$$\lambda_i = \left(\frac{\prod_{i=1}^n \hat{f}(X_i)^{\frac{1}{n}}}{\hat{f}(X_i)} \right)^{\alpha},$$

mit $\alpha \in [0,1]$ und $\hat{f}$ einem beliebigen Dichteschätzer für den $\hat{f}(X_i) > 0$, $i = 1, \ldots n$ gilt, dann heißt eine Funktion $f_{adapt} : \mathbb{R}^d \to \mathbb{R}$ adaptiver Kerndichteschätzer, falls

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

In der Definition erfüllen der Glättungsfaktor h und der Kern K die gleiche Funktion wie bei dem normalen Kerndichteschätzer und werden genauso ausgewählt. λ_i erhalten wir mithilfe eines zuvor berechneten "Pilotschätzers" $\hat{f}$. In der Regel wird $\hat{f}$ entweder ein normaler Kerndichteschätzer mit der gleichen Bandweite h wie auch f_{adapt} oder ein Nächste-Nachbarn-Schätzer sein.

Die Differenzierbarkeit des Pilotschätzers spielt hier natürlich keine Rolle mehr, da er nur an den Punkten $X_1, \ldots, X_n$ ausgewertet wird, weswegen sich etwa der Epanechnikov-Kern anbietet.

Für α gibt es verschiedene Ansätze. $\alpha = 0$ führt natürlich wieder zu dem normalen Kerndichteschätzer mit fester Bandweite; Standardwerte sind $\alpha = \frac{1}{d}$ und $\alpha = \frac{1}{2}$, siehe [23], Abschnitt 5.3.

Der offensichtliche Vorteil des adaptiven Kerndichteschätzers ist, dass er einerseits so wie f_n auch eine Wahrscheinlichkeitsdichte ist, andererseits aber auch in Regionen mit weniger Datenpunkten verhältnissmäßig gute Ergebnisse liefert; die Nachteile liegen vor allem darin, dass durch α ein weiterer Parameter ausgewählt werden muss und bei großem n auch der Rechenaufwand für die λ_i nicht zu vernachlässigen ist.

2 Konsistenz in L_1

In diesem Abschnitt beweisen wir ein starkes Konsistenzresultat in L_1: Wenn der L_1-Fehler des Kerndichteschätzers in Wahrscheinlichkeit für eine Dichte f gegen 0 konvergiert, so konvergiert er schon für alle Dichten und das sogar fast sicher und mit exponentieller Geschwindigkeit. Dafür sind nur sehr geringe Anforderung an den Kern K, n und die Bandweite h notwendig. Wir betrachten der Einfachheit halber nur den Fall, dass $h := h_1 = \ldots = h_d > 0$. Der Übergang zu verschiedenen Bandweiten in verschiedenen Dimensionen ist dann möglich, erfordert aber gesonderte Voraussetzungen.

Der Beweis orientiert sich in weiten Teilen an [9], Kapitel 2 und 3, und [10], Kapitel 2 und 3, mit Änderungen wo sie angebracht erschienen.

Zur speziellen Notation in diesem Kapitel:

Weiterhin seien $X_1, \ldots, X_n$ unabhängig und identisch verteilt, außerdem ist $\mu := \mathbb{P}^{X_1}$ mit zugehöriger Dichte f und $\mu_n(\cdot) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(\cdot)}(X_i)$ das entsprechende empirische Maß. Wir verwenden $K_h(x) := h^{-d}K(\frac{x}{h})$ und daher $f_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_h\left(x - X_i\right)$ für den Kerndichteschätzer. Den Erwartungswert von f_n werden wir häufig brauchen und definieren deshalb:

$$g_h(x) := \mathbb{E}(f_n(x)) = \int K_h(x-y)f(y)dy = f * K_h(x). \tag{7}$$

Außerdem können wir f_n weiter umformen zu

$$f_n(x) = \int K_h(x-y)\mu_n(dy). \tag{8}$$

Der Beweis des folgenden Satzes wird das gesamte Kapitel in Anspruch nehmen.

Satz 2.1 (Devroye, [9])**.** *Sei f_n wie weiter oben definiert, $K \geq 0$ eine Borelmessbare Funktion auf $\mathbb{R}^d$ mit $\int K(x)dx = 1$ und $h = h(n) > 0$ eine Folge, dann sind folgende Aussagen äquivalent:*

1. $\int |f_n - f| \to^{\mathbb{P}} 0$ *für mindestens eine Wahrscheinlichkeitdichte f*
2. $\int |f_n - f| \to^{\mathbb{P}} 0$ *für alle Wahrscheinlichkeitsdichten f*
3. $\int |f_n - f| \to 0$ *fast sicher für alle Wahrscheinlichkeitsdichten f*
4. $\int |f_n - f| \to 0$ *exponentiell für alle Wahrscheinlichkeitsdichten f, das heißt: Für alle $\epsilon > 0$ existieren $r, m > 0$ so dass $\mathbb{P}\left(\int |f_n - f| \geq \epsilon\right) \leq e^{-rn}$ für alle $n > m$, wobei r unabhängig von f wählbar ist*
5. $\lim_{n\to\infty} h = 0$, $\lim_{n\to\infty} nh^d = \infty$

Beweis. Die Hauptarbeit liegt im Beweis von *5* $\Rightarrow$ *4* und *1* $\Rightarrow$ *5*. Die Implikation *3* $\Rightarrow$ *2* $\Rightarrow$ *1* ist trivial und *4* $\Rightarrow$ *3* folgt sofort mit Satz 34.6 in [2], einem Korollar des Borel-Cantelli-Lemmas, denn $\sum_{n>m} e^{-rn} < \infty$.

Einige Hilfsaussagen wurden an das Ende des Kapitels verschoben, da sie entweder häufiger gebraucht werden oder ein zwischengeschobenes Lemma die Übersichtlichkeit des Beweises beeinträchtigt hätte.

5 $\Rightarrow$ 4

Es gelte also

$$\lim_{n\to\infty} h = 0 \text{ und } \lim_{n\to\infty} nh^d = \infty. \tag{9}$$

Im Folgenden werden wir die erste Bedingung durchgehend verwenden, die zweite wird erst im letzten Schritt benutzt.

Wir wollen zeigen, dass für festes $\epsilon_0 > 0$ r und $m > 0$ existieren, sodass $P(\int |f_n(x) - f(x)|dx > \epsilon_0) < e^{-rn}$, wobei r unabhängig von f sein soll. Der Beweis lässt sich in mehrere logische Abschnitte einteilen: Zuerst werden wir zeigen, dass wir uns für K auf Treppenfunktionen beschränken können und in Folge nur die exponentielle Konvergenz von $h^{-d}\int |\mu(x + hA) - \mu_n(x + hA)|dx$ für einen beliebigen halboffenen Quader $A \subset \mathbb{R}^d$ zu zeigen brauchen. Diesen Ausdruck werden wir im zweiten Schritt nach oben durch eine Summe beschränken, für deren einzelne Summanden wir im letzten Schritt die exponentielle Konvergenz zeigen.

Starten wir mit dem ersten Schritt: Wir können zuerst festhalten, dass

$$\int |f(x) - f_n(x)|dx \leq \int |f_n(x) - g_h(x)|dx + \int |f(x) - g_h(x)|dx.$$

Mit Hilfssatz 2.2 und Bedingung (9) können wir für n ausreichend groß $\int |f(x) - g_h(x)|dx < \delta_1 = \delta_1(n)$ für $0 < \delta_1 < \epsilon_0$ beliebig klein fordern. Daher folgt:

$$\mathbb{P}\left(\int |f(x) - f_n(x)|dx \geq \epsilon_0\right) \leq \mathbb{P}\left(\int |f_n(x) - g_h(x)|dx \geq \epsilon_0 - \delta_1(n)\right), \tag{10}$$

und die Behauptung folgt, falls wir die exponentielle Konvergenz von $\int |f_n(x) - g_h(x)|dx$ zeigen können.

Wie bereits erwähnt verschärfen wir jetzt die Bedingungen an den Kern. Da die Treppenfunktionen dicht in L_1 liegen und K als integrierbare Funktion f.s. endlich ist, finden wir zu festem $\epsilon > 0$ Konstanten $M, L, N, a_1, \ldots, a_N$ und disjunkte Quader $A_1, \ldots, A_N$ in $\mathbb{R}^d$, sodass die Funktion

$$K^*(x) := \sum_{i=1}^{N} a_i \mathbb{1}_{A_i}(x)$$

K in L_1 approximiert mit:

$$\begin{aligned} |K^*| &\leq M, \\ supp(K^*) &\subset [-L, L]^d \text{ und} \\ \int |K(x) - K^*(x)|dx &< \delta_2, \end{aligned} \tag{11}$$

mit $0 < \delta_2 < \frac{\epsilon}{2}$ beliebig klein. δ_2 wird später eine ähnliche Rolle spielen wie δ_1 in Gleichung (10).

Wir definieren nun f_n^* und g_h^* wie f_n und g_h durch ersetzen von K durch K^* und können mit (8) abschätzen:

$$\begin{aligned} \int |f_n(x) - f_n^*(x)|dx &= \int h^{-d} \left| \int K\left(\frac{x-y}{h}\right) - K^*\left(\frac{x-y}{h}\right) \mu_n(dy) \right| dx \\ &\leq \int h^{-d} \int \left| K\left(\frac{x-y}{h}\right) - K^*\left(\frac{x-y}{h}\right) \right| \mu_n(dy)dx \\ &= \int h^{-d} \int \left| K\left(\frac{x-y}{h}\right) - K^*\left(\frac{x-y}{h}\right) \right| dx\mu_n(dy) \\ &< \delta_2, \end{aligned}$$

und mit (7)

$$\begin{aligned} \int |g_h^*(x) - g_h(x)|dx &= \int |f * K_h^*(x) - f * K_h(x)|dx \\ &= \int h^{-d} \left| \int \left(K^*\left(\frac{x-y}{h}\right) - K\left(\frac{x-y}{h}\right) \right) f(y)dy \right| dx \\ &\leq \int h^{-d} \int \left| K^*\left(\frac{x-y}{h}\right) - K\left(\frac{x-y}{h}\right) \right| f(y)dydx \\ &= \int h^{-d} \int \left| K^*\left(\frac{x-y}{h}\right) - K\left(\frac{x-y}{h}\right) \right| dxf(y)dy \\ &< \delta_2. \end{aligned}$$

Damit folgt dann:

$$\begin{aligned} \int |f_n(x) - g_h(x)|dx &\leq \int |f_n(x) - f_n^*(x)|dx + \int |f_n^*(x) - g_h^*(x)|dx \\ &\quad + \int |g_h^*(x) - g_h^*(x)|dx \\ &\leq 2\delta_2 + \int |f_n^*(x) - g_h^*(x)|dx, \end{aligned}$$

wobei wir die Reihenfolge der Integration vertauscht und Bedingung (11) verwendet haben.

Den hinteren Term können wir weiter abschätzen:

$$
\begin{aligned}
&\int |f_n^*(x) - g_h^*(x)|dx \\
&= \int h^{-d} \left| \int \sum_{i=1}^N a_i \mathbb{1}_{A_i} \left(\frac{x-y}{h}\right) \mu_n(dy) - \int \sum_{i=1}^N a_i \mathbb{1}_{A_i} \left(\frac{x-y}{h}\right) f(y)dy \right| \\
&\leq \int h^{-d} \sum_{i=1}^N |a_i| \left| \int \mathbb{1}_{A_i} \left(\frac{x-y}{h}\right) \mu_n(dy) - \int \mathbb{1}_{A_i} \left(\frac{x-y}{h}\right) f(y)dy \right| dx \\
&= \int h^{-d} \sum_{i=1}^N |a_i| |\mu(x + hA_i) - \mu_n(x + hA_i)|dx \\
&\leq h^{-d} M \sum_{i=1}^N \int |\mu(x + hA_i) - \mu_n(x + hA_i)|dx \\
&\leq h^{-d} MN \max_{i=1}^N \int |\mu(x + hA_i) - \mu_n(x + hA_i)|dx
\end{aligned}
$$

nach Wahl von M und $a_1, \ldots, a_N$. Wir sind jetzt am Ende unseres ersten Schrittes angekommen, denn

$$
\begin{aligned}
&\mathbb{P}\left(\int |f_n - g_h| > \epsilon \right) \\
&\leq \mathbb{P}\left(2\delta_2 + h^{-d} MN \max_{i=1}^N \int |\mu(x + hA_i) - \mu_n(x + hA_i)|dx > \epsilon \right) \\
&= \mathbb{P}\left(h^{-d} \max_{i=1}^N \int |\mu(x + hA_i) - \mu_n(x + hA_i)|dx > (\epsilon - 2\delta_2) M^{-1} N^{-1} \right).
\end{aligned}
$$

Wir kommen nun zum zweiten Schritt, in dem wir den Ausdruck $h^{-d} \int |\mu(x + hA) - \mu_n(x + hA)|dx$ für einen beliebigen halboffenen Quader $A = \prod_{i=1}^d [x_i, x_i + b_i)$ nach oben beschränken. Zu diesem A und $\hat{\epsilon} = (\epsilon - 2\delta_2) M^{-1} N^{-1}$ partitionieren wir $\mathbb{R}^d$ jetzt in disjunkte halboffene Quader der Seitenlänge $\frac{h}{N_0}$, wobei wir N_0 erst später festlegen und jetzt nur verlangen, dass $b_i \geq \frac{2}{N_0}$, $i = 1, \ldots, d$. Diese Partition sei Ψ genannt und es gilt $\Psi = \left\{ \prod_{j=1}^d \left[\frac{(i_j - 1)h}{N_0}, \frac{i_j h}{N_0} \right) | i_1, \ldots, i_d \in \mathbb{Z} \right\}$

Zur besseren Übersicht definieren wir zusätzlich A^*, C_x und C_x^*:

$$
A^* := \prod_{i=1}^d \left[x_i + \frac{1}{N_0}, x_i + b_i - \frac{1}{N_0} \right),
$$

also A ohne einen "Rahmen" der Breite $\frac{1}{N_0}$,

$$C_x := (x + hA) \setminus \bigcup_{B \in \Psi, B \subset x+hA} B$$

und

$$C_x^* := x + h(A \setminus A^*).$$

Da aufgrund der respektiven Seitenlängen $hA^* \subset \bigcup_{B \in \Psi, B \subset hA} B$ gilt, gilt auch $C_x \subset C_x^*$. Wir können mit Maßadditivität für disjunkte Mengen und Dreiecksungleichung abschätzen:

$$\begin{aligned}
&\int |\mu_n(x + hA) - \mu(x + hA)|dx \\
&= \int |\mu_n(C_x) + \sum_{B \in \Psi, B \subset x+hA} \mu_n(B) - \mu(C_x) - \sum_{B \in \Psi, B \subset x+hA} \mu(B)|dx \\
&\leq \int \mu_n(C_x) + \mu(C_x) + |\sum_{B \in \Psi, B \subset x+hA} (\mu_n(B) - \mu(B))|dx \\
&\leq \int \sum_{B \in \Psi, B \subset x+hA} |\mu_n(B) - \mu(B)|dx + \int \mu_n(C_x^*) + \mu(C_x^*)dx.
\end{aligned} \tag{12}$$

Betrachten wir nun den ersten Teil des Terms. Zuerst fixieren wir eine aufsteigende Folge von endlichen Teilmengen von Ψ, $(\Psi_k)_{k \in \mathbb{N}} \uparrow \Psi$, das heißt

$$\begin{aligned}
&\Psi_k \subseteq \Psi_l \subset \Psi \text{ für } k \leq l, \\
&|\Psi_k| < \infty \text{ für alle } k \in \mathbb{N} \text{ und} \\
&\bigcup_{k \in \mathbb{N}} \Psi_k = \Psi.
\end{aligned}$$

Wie die Ψ_k konkret aussehen spielt dabei keine Rolle. Es gilt dann:

$$\psi(x) := \sum_{B \in \Psi, B \subseteq x+hA} |\mu_n(B) - \mu(B)| = \lim_{k \to \infty} \psi_k(x),$$

mit

$$\psi_k(x) := \sum_{B \in \Psi_k, B \subseteq x+hA} |\mu_n(B) - \mu(B)|,$$

wobei die ψ_k offensichtlich monoton steigend in k sind.

Mit diesen Bezeichnungen gilt mit dem Satz von der monotonen Konvergenz im ersten Schritt und der Endlichkeit der Ψ_k im zweiten Schritt:

$$\begin{aligned}\int \psi(x)dx &= \lim_{k\to\infty} \int \psi_k(x)dx \\ &= \lim_{k\to\infty} \sum_{B\in\Psi_k} |\mu_n(B) - \mu(B)| \int_{B\subseteq x+hA} dx \\ &= \sum_{B\in\Psi} |\mu_n(B) - \mu(B)| \int_{B\subseteq x+hA} dx.\end{aligned}$$

Als Hilfsfakt brauchen wir, dass $\int_{B\subseteq x+hA} dx \leq \lambda(hA)$. Um dies einzusehen, reicht es aufgrund der Bewegungsinvarianz von λ, nur $B_0 = \left[0, \frac{h}{N_0}\right)^d$ und $A_0 = \prod_{i=1}^d [0, b_i)$ zu betrachten. Dafür gilt aber offensichtlich

$$\begin{aligned}\lambda\left(\left\{x \in \mathbb{R}^d | \left[0, \frac{h}{N_0}\right)^d \subseteq x + \prod_{i=1}^d [0, hb_i)\right\}\right) &= \lambda\left(\prod_{i=1}^d \left[0, hb_i - \frac{h}{N_0}\right)\right) \\ &\leq \lambda(hA_0).\end{aligned}$$

Wir führen nun eine neue Variable $R > 0$ ein, welche wir erst im nächsten Schritt fixieren werden.

$$\begin{aligned}&\int \psi(x)dx \\ &\leq h^d\lambda(A)\left(\sum_{B\in\Psi, B\cap S_{0,R}\neq\emptyset} |\mu_n(B) - \mu(B)| + \sum_{B\in\Psi, B\subseteq S_{0,R}^c} |\mu_n(B) - \mu(B)|\right) \\ &\leq h^d\lambda(A)\left(\sum_{B\in\Psi, B\cap S_{0,R}\neq\emptyset} |\mu_n(B) - \mu(B)| + \sum_{B\in\Psi, B\subseteq S_{0,R}^c} (\mu_n(B) + \mu(B))\right) \\ &\leq h^d\lambda(A)\left(\sum_{B\in\Psi, B\cap S_{0,R}\neq\emptyset} |\mu_n(B) - \mu(B)| + \mu_n\left(S_{0,R}^c\right) + \mu\left(S_{0,R}^c\right)\right) \\ &= h^d\lambda(A)\left(\sum_{B\in\Psi, B\cap S_{0,R}\neq\emptyset} |\mu_n(B) - \mu(B)| + \mu_n\left(S_{0,R}^c\right) - \mu\left(S_{0,R}^c\right) + 2\mu\left(S_{0,R}^c\right)\right)\end{aligned} \tag{13}$$

Für das dritte Ungleichheitszeichen haben wir dabei verwendet, dass die B disjunkt sind. Wir fixieren R ausreichend groß, sodass

$$2\lambda(A)\mu(S_{0,R}^c) < \delta_3 \tag{14}$$

mit $0 < \delta_3 < \frac{\hat{\epsilon}}{2}$ gilt; das δ_3 erfüllt die gleiche Funktion wie δ_1 und δ_2 zuvor.

Nachdem wir den ersten Teil aus Gleichung (12) abgeschätzt haben betrachten wir nun den zweiten.

$$\begin{aligned}
\int \mu_n(C_x^*) + \mu(C_x^*)dx &= \int\int f(y)\mathbb{1}_{C_x^*}(y)dydx + \int n^{-1}\sum_{i=1}^{n}\mathbb{1}_{C_x^*}(X_i)dx \\
&= \int f(y)\int \mathbb{1}_{C_x^*}(y)dxdy + n^{-1}\sum_{i=1}^{n}\int \mathbb{1}_{C_x^*}(X_i)dx \\
&= 2\lambda(C_x^*) \\
&= 2\lambda(h(A - A^*)) \\
&= 2h^d(\lambda(A) - \lambda(A^*)) \\
&= 2h^d\lambda(A)\left(1 - \prod_{i=1}^{d}\left(1 - \frac{2}{N_0 b_i}\right)\right) \\
&\leq 4h^d\lambda(A)\sum_{i=1}^{d}\frac{1}{N_0 b_i} \\
&\leq \delta_3 h^d,
\end{aligned} \tag{15}$$

für N_0 ausreichend groß. Im vorletzten Schritt haben wir dabei das folgende kleine Lemma benutzt:

Lemma 2.1.1. *Für $z_i \in (0, 1]$, $i = 1, \ldots, d$ gilt:*

$$1 - \prod_{i=1}^{d}(1 - z_i) \leq \sum_{i=1}^{d} z_i$$

Beweis. Per Induktion über d.

Sei $d = 1$, dann ist $1 - (1 - z_i) = z_i$.

Gelte die Behauptung bis zu d. Dann folgt

$$\begin{aligned}
1 - \prod_{i=1}^{d+1}(1 - z_i) &= 1 - \prod_{i=1}^{d}(1 - z_i) + z_{d+1}\prod_{i=1}^{d}(1 - z_i) \\
&\leq \sum_{i=1}^{d} z_i + z_{d+1}\prod_{i=1}^{d}(1 - z_i) \\
&\leq \sum_{i=1}^{d+1} z_i.
\end{aligned}$$

□

Wir haben jetzt durch Kombination von (13), (14) und (15)

$$h^{-d} \int |\mu_n(x + hA) - \mu(x + hA)|dx$$
$$\leq \lambda(A) \left(\sum_{B \in \Psi, B \cap S_{0,R} \neq \emptyset} |\mu_n(B) - \mu(B)| \right) + \lambda(A) \left(\mu_n \left(S_{0,R}^c \right) - \mu \left(S_{0,R}^c \right) \right) + 2\delta_3$$

und in dem verbleibenden dritten Schritt ist nur noch zu zeigen, dass

$$\mu_n \left(S_{0,R}^c \right) - \mu \left(S_{0,R}^c \right) \tag{16}$$

und

$$\sum_{B \in \Psi, B \cap S_{0,R} \neq \emptyset} |\mu_n(B) - \mu(B)| \tag{17}$$

jeweils exponentiell gegen 0 konvergieren, dann folgt die Behauptung.

Für den Term in (16) wollen wir die Hoeffding-Ungleichung anwenden:

Lemma 2.1.2 (Hoeffding-Ungleichung)**.** *Seien $Y_1, \ldots, Y_n$ unabhängig und $a_i \leq Y_i \leq b_i$ für $i = 1, \ldots, n$, dann gilt für $t > 0$*

$$\mathbb{P} \left(\frac{1}{n} \sum_{i=1}^{n} Y_i - \mathbb{E}(Y_i) \geq t \right) \leq \exp \left(- \frac{2n^2 t^2}{\sum_{i=1}^{n} b_i - a_i} \right)$$

Ein Beweis findet sich in [14] (Theorem 2).

Offensichtlich lässt sich $\mu_n(S_{0,R}^c)$ folgendermaßen als Zufallsvariable ausdrücken: $\mu_n \left(S_{0,R}^c \right) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{S_{0,R}^c}(X_i) = \frac{1}{n} \sum_{i=1}^{n} Z_i$, wobei die Z_i unabhängig, identisch $Bern \left(\mu \left(S_{0,R}^c \right) \right)$-verteilt sind. Daraus folgt dann durch Einsetzen in die Hoeffding-Ungleichung:

$$\mathbb{P} \left(\mu_n \left(S_{0,R}^c \right) - \mu \left(S_{0,R}^c \right) \geq \epsilon' \right) = \mathbb{P} \left(\sum_{i=1}^{n} Z_i - \mathbb{E}(Z_i) \geq n\epsilon' \right)$$
$$\leq \exp \left(-2n\epsilon'^2 \right)$$

mit $\epsilon' := \lambda(A)^{-1}(\hat{\epsilon} - 2\delta_3)$.

Für den Term (17) benötigen wir ein Lemma, für dessen Beweis wir auf [9] (Lemma 3.1) verweisen.

Lemma 2.1.3. *Sei $(Y_1, \ldots, Y_k)$ ein* Multinomial$(n, p_1, \ldots, p_k)$*-verteilter Zufallsvektor, $t \in (0, 1)$ und $\frac{k}{n} \leq \frac{t^2}{20}$, dann gilt:*

$$\mathbb{P} \left(\sum_{i=1}^{k} |Y_i - \mathbb{E}(Y_i)| > nt \right) \leq 3 \exp \left(- \frac{nt^2}{25} \right)$$

Um die Vorraussetzungen des Lemmas zu erfüllen, beachten wir zuerst, dass

$$k := |\{B \in \Psi | B \cap S_{0,R} \neq \emptyset\}| \leq 2^d \left(\frac{RN_0}{h} + 1\right)^d.$$

Um dies einzusehen, betrachte statt $S_{0,R}$ den um 0 zentrierten Würfel mit Seitenlänge $2R$, $Q_{0,R} \supset S_{0,R}$. Teile diesen Würfel entlang der Koordinatenachsen in 2^d Abschnitte und in jeden dieser Abschnitte passen aufgrund der jeweiligen Seitenlängen maximal $\frac{RN_0}{h} + 1$ der $B \in \Psi$.

Mit der Standard Groß-O-Notation gilt aber

$$2^d \left(\frac{RN_0}{h} + 1\right)^d = O\left(h^{-d}\right) = O(n), \tag{18}$$

das heißt $2^d \left(\frac{RN_0}{h} + 1\right)^d$ wächst nicht schneller als n. Dies gilt, da alle anderen Variablen bereits fixiert wurden und die Vorraussetzung $\lim_{n\to\infty} nh^d = \infty$ impliziert, dass n schneller steigt als h^{-d}. Daher können wir n ausreichend groß wählen, so dass $\frac{k}{n} \leq \frac{\epsilon^2}{20}$.

Da die $B \in \Psi, B \cap S_{0,R} \neq \emptyset$ paarweise disjunkt sind, ist der Vektor $(Y_1, \ldots, Y_k) := (n\mu_n(B))_{B\in\Psi, B\cap S_{0,R}\neq\emptyset}$ bereits multinomialverteilt, weshalb wir das Lemma darauf anwenden können:

$$\begin{aligned}\mathbb{P}\left(\sum_{B\in\Psi, B\cap S_{0,R}\neq\emptyset} |\mu_n(B) - \mu(B)| > \epsilon'\right) &= \mathbb{P}\left(\sum_{i=1}^{k} |Y_i - \mathbb{E}(Y_i)| > n\epsilon'\right) \\ &\leq 3\exp\left(-\frac{n\epsilon'^2}{25}\right),\end{aligned}$$

für n genügend groß. Die Behauptung folgt, denn:

$$\begin{aligned}&\mathbb{P}\left(h^{-d}\int |\mu_n(x + hA) - \mu(x + hA)|dx > \hat{\epsilon}\right) \\ &\leq \mathbb{P}\left(\sum_{B\in\Psi, B\cap S_{0,R}\neq\emptyset} |\mu_n(B) - \mu(B)| + \mu_n\left(S_{0,R}^c\right) - \mu\left(S_{0,R}^c\right) > \epsilon'\right) \\ &\leq \mathbb{P}\left(\sum_{B\in\Psi, B\cap S_{0,R}\neq\emptyset} |\mu_n(B) - \mu(B)| > \frac{\epsilon'}{2}\right) + \mathbb{P}\left(\mu_n\left(S_{0,R}^c\right) - \mu\left(S_{0,R}^c\right) > \frac{\epsilon'}{2}\right) \\ &\leq 3\exp\left(-\frac{n\epsilon'^2}{100}\right) + \exp\left(-\frac{n\epsilon'^2}{2}\right).\end{aligned}$$

1 $\Rightarrow$ 5

Wir setzen jetzt voraus, dass K und f beliebige Wahrscheinlichkeitsdichten sind und

$$\int |f(x) - f_n(x)| \to_{n\to\infty} 0 \text{ in Wahrscheinlichkeit.}$$

Wir wollen zeigen, dass dann sowohl

$$h = h(n) \to_{n\to\infty} 0, \tag{19}$$

als auch

$$nh(n)^d \to_{n\to\infty} \infty. \tag{20}$$

Die erste Behauptung folgt schnell: Da $\int |f(x) - f_n(x)|dx$ in Wahrscheinlichkeit konvergiert, konvergiert es auch in Verteilung, insbesondere gilt also $\mathbb{E}\left(\int |f(x) - f_n(x)|dx\right) \to 0$. Daher folgt mit der Jensenschen Ungleichung:

$$\begin{aligned}\mathbb{E}\left(\int |f_n(x) - f(x)|dx\right) &= \int \mathbb{E}(|f_n(x) - f(x)|)dx \\ &\geq \int |\mathbb{E}(f_n(x)) - f(x)|dx \\ &= \int |g_h(x) - f(x)|dx.\end{aligned} \tag{21}$$

Da der letzte Term immer noch nicht-negativ ist, konvergiert auch er gegen 0 und mit Hilfssatz 2.5 folgt sofort Behauptung (19).

Im Folgenden können wir also annehmen, dass (19) gilt. Außerdem wissen wir, dass auch $\mathbb{E}(\int |f_n(x) - g_h(x)|dx)$ gegen 0 konvergiert, denn mit Ungleichung (21) folgt auch:

$$\begin{aligned}0 &\leq \mathbb{E}\left(\int |f_n(x) - g_h(x)|dx\right) \\ &\leq \mathbb{E}\left(\int |f_n(x) - f(x)|dx\right) + \mathbb{E}\left(\int |f(x) - g_h(x)|dx\right) \\ &\leq 2\mathbb{E}\left(\int |f_n(x) - f(x)|dx\right) \to_{n\to\infty} 0.\end{aligned}$$

Wir wollen Behauptung (20) per Widerspruch zeigen. Nehmen wir an, es gebe eine Teilfolge von nh^d, die gegen einen reellen Limes konvergiert, also

$$\lim_{k\to\infty} n_k h_{n_k}^d =: s \in [0, \infty). \tag{22}$$

Im Folgenden approximieren wir K durch eine gestutzte und damit beschränkte Variante K^*. Dafür sei $M > 0$ beliebig und wird erst am Ende des Beweises festgelegt und

$$K^*(x) := K(x)\mathbb{1}_{K(x)\leq M}.$$

f_n^* und g_h^* sind wie f_n und g_h definiert mit K durch K^* ersetzt. Die L_1-Distanz zwischen diesen ist dann nur durch $\int K(x) - K^*(x)dx =: \delta$ beschränkt:

$$\begin{aligned}
\int |f_n(x) - f_n^*(x)|dx &= \int h^{-d}n^{-1}\sum_{i=1}^{n}\left(K\left(\frac{x-X_i}{h}\right) - K^*\left(\frac{x-X_i}{h}\right)\right)dx \\
&= n^{-1}\sum_{i=1}^{n}\int h^{-d}\left(K\left(\frac{x-X_i}{h}\right) - K^*\left(\frac{x-X_i}{h}\right)\right)dx \\
&= \int K(x) - K^*(x)dx \\
&= \delta
\end{aligned}$$

und

$$\begin{aligned}
\int |g_h(x) - g_h^*(x)|dx &= \int h^{-d}\int f(y)\left(K\left(\frac{x-y}{h}\right) - K^*\left(\frac{x-y}{h}\right)\right)dydx \\
&= \int h^{-d}f(y)\int K\left(\frac{x-y}{h}\right) - K^*\left(\frac{x-y}{h}\right)dxdy \\
&= \int K(x) - K^*(x)dx\int f(y)dy \\
&= \delta.
\end{aligned}$$

Damit und mit der inversen Dreiecksungleichung folgt dann:

$$\begin{aligned}
&\int |f_n(x) - g_h(x)|dx \\
&\geq \int |f_n^*(x) - g_h^*(x)|dx - \int |f_n(x) - f_n^*(x)|dx - \int |g_h(x) - g_h^*(x)|dx \quad (23) \\
&= \int |f_n^*(x) - g_h^*(x)|dx - 2\delta.
\end{aligned}$$

Sei $L > 0$ eine Konstante, die wir wie M erst am Ende des Beweises fixieren. Um K auf einen kompakten Träger einzuschränken, definieren wir

$$K' := \mathbb{1}_{S_{0,L}}K^*$$

und zusätzlich

$$K'' := \mathbb{1}_{S_{0,L}^c}K^*,$$

sodass also $K^* = K' + K''$ und $f_n^* = f_n' + f_n''$, wobei f_n' und f_n'' wieder durch ersetzen von K durch K' bzw. K'' definiert sind. Wenn wir

$$A_x := \bigcap_{i=i}^{n} \{X_i \notin S_{x,hL}\}$$

definieren, so gilt offensichtlich $\mathbb{1}_{A_x} f_n'(x) = 0$. Damit folgt dann

$$\begin{aligned} \mathbb{E}\left(\int |f_n^*(x) - g_h^*(x)|dx\right) &\geq \int \mathbb{E}\left(|g_h^*(x) - f_n^*(x)|\mathbb{1}_{A_x}\right) dx \\ &\geq \int \mathbb{E}(g_h^*(x)\mathbb{1}_{A_x})dx - \int \mathbb{E}(f_n^*(x)\mathbb{1}_{A_x})dx \qquad (24) \\ &= \int g_h^*(x)\mathbb{P}(A_x)dx - \int \mathbb{E}(f_n''(x)\mathbb{1}_{A_x})dx. \end{aligned}$$

Der Term auf der linken Seite der Ungleichungskette (24) konvergiert aber mithilfe von Ungleichung (23) gegen 0, da wir δ beliebig klein machen können. Daher reicht es, die rechte Seite weiter abzuschätzen und zu zeigen, dass sie nur gegen 0 konvergiert, wenn $s = \infty$.

Wir benötigen folgendes Lemma:

Lemma 2.1.4. *Mit der Notation wie zuvor gelten folgende drei Aussagen:*

(a) $g_h'(x) \to f(x) \int K'(z)dz$

(b) $\frac{\mu(S_{y+hz,hL})}{\lambda(S_{y-hz,hL})} \to f(y)$ *alle* $z \in \mathbb{R}^d$, λ*-fast alle* $y \in \mathbb{R}^d$.

(c) $1 - x \geq \exp\left(-\frac{x}{1-x}\right)$ *für* $0 \leq x \leq 1$

Beweis. (a) Es gilt:

$$\begin{aligned} g_h'(x) &= \int K_h'(x-y)f(y)dy \\ &= \int K'(z)dz \int \frac{K_h'(x-y)}{\int K'(z)dz} f(y)dy \\ &\to_{h\to 0} \int K'(z)dz f(x) \end{aligned}$$

mit Hilfssatz 2.3, da $\frac{K_h'(\cdot)}{\int K'(z)dz}$ beschränkt ist, kompakten Träger $S_{0,L}$ besitzt und zu 1 integriert.

(b) Wir benutzen das Lebesgue-Density-Theorem 2.2, wobei wir als Klasse $\mathcal{B} = \{S_{0,r} | r > 0\}$ wählen, für die die Voraussetzung des Satzes offensichtlich erfüllt ist. Damit gilt dann für beliebiges $z \in \mathbb{R}^d$ und fast alle $y \in \mathbb{R}^d$:

$$\begin{aligned}\frac{\mu(S_{y+hz,hL})}{\lambda(S_{y-hz,hL})} &= \frac{1}{\lambda(S_{0,hL})} \int_{S_{y+hz,hL}} f(y)dy \\ &\to f(y),\end{aligned}$$

da f λ-fast überall stetig ist.

(c) Standard-Ungleichung, siehe etwa [19] (Ungleichung 2.68).

□

Es gilt jetzt für den linken Term der rechten Seite von (24) unter Benutzung des Lemmas von Fatou, Lemma 2.1.4 (a), Lemma 2.1.4 (c) und schließlich der Annahme (22) in gleicher Reihenfolge:

$$\begin{aligned}
&\liminf_{k\to\infty} \int g_h^*(x)\mathbb{P}(A_x)dx \\
&\geq \liminf_{k\to\infty} \int g_h'(x)\mathbb{P}(A_x)dx \\
&\geq \int \liminf_{k\to\infty} g_h'(x) \liminf_{k\to\infty} \mathbb{P}(A_x)dx \\
&= \int f(x) \int K'(z)dz \liminf_{k\to\infty} (1-\mu(S_{x,hL}))^n dx \\
&\geq \int f(x) \liminf_{k\to\infty} \exp\left(-n_k \frac{\mu(S_{x,hL})}{1-\mu(S_{x,hL})}\right) dx \int K'(z)dz \\
&= \int f(x) \exp\left(-\limsup_{k\to\infty} \left(\left[n_k h^d\right]\left[\lambda(S_{0,1})L^d\right] \frac{\mu(S_{x,hL})}{\lambda(S_{x,hL})} \frac{1}{1-\mu(S_{x,hL})}\right)\right) dx \int_{S_{0,L}} K^*(z)dz \\
&= \int f(x) \exp\left(-s\lambda(S_{0,1})L^d f(x)\right) dx \int_{S_{0,L}} K^*(z)dz.
\end{aligned}$$

Andererseits gilt für den zweiten Term in (24) bei Beachtung der Unabhängigkeit der X_i, Vertauschung der Integrale und Verwendung von $z = \frac{x-y}{h}$:

$$
\begin{aligned}
\int \mathbb{E}\left(f_n''(x)\mathbb{1}_{A_x}\right)dx &= \int n^{-1}h^{-d}\sum_{i=1}^{n}\mathbb{E}\left(\mathbb{1}_{A_x}K''\left(\frac{x-X_i}{h}\right)\right)dx \\
&= \int h^{-d}\mathbb{E}\left(\mathbb{1}_{X_1\notin S_{x,hL}}K''\left(\frac{x-X_1}{h}\right)\right)\prod_{i=2}^{n}\mathbb{E}(\mathbb{1}_{X_i\notin S_{x,hL}})dx \\
&= \int h^{-d}(1-\mu(S_{x,hL}))^{n-1}\int_{y-x\notin S_{0,hL}}K''\left(\frac{x-y}{h}\right)f(y)dydx \\
&= \int h^{-d}f(y)\int_{x-y\notin S_{0,hL}}K''\left(\frac{x-y}{h}\right)(1-\mu(S_{x,hL}))^{n-1}dxdy \\
&= \int f(y)\int_{z\notin S_{0,L}}(1-\mu(S_{y+hz,L}))^{n-1}K''(z)dzdy \\
&\leq \int f(y)\int_{z\notin S_{0,L}}\exp(-(n-1)\mu(S_{y+hz,L}))K''(z)dzdy,
\end{aligned}
\tag{25}
$$

wobei die letzte Ungleichung gilt, weil für $1-0\leq e^0$ und $\partial_a(1-a) = -1 \leq \partial_a(e^{-a}) = -e^{-a} \in \left(-1,-\frac{1}{e}\right]$ und damit $(1-a)^{n-1}\leq e^{-(n-1)a}$ für alle $a\in(0,1]$.

Nehmen wir hiervon den Limes superior könnnen wir wieder das Lemma von Fatou anwenden, da $\exp(-(n-1)\mu(S_{y+hz,L}))K''(z)\leq K''(z)\in L_1$, also:

$$
\begin{aligned}
&\limsup_{k\to\infty}\int \mathbb{E}(f_{n_k}''(x)\mathbb{1}_{A_x})dx \\
&\leq \limsup_{k\to\infty}\int f(y)\int_{z\notin S_{0,L}}\exp(-(n_k-1)\mu(S_{y+hz,L}))K''(z)dzdy \\
&\leq \int f(y)\int_{z\notin S_{0,L}}\exp\left(-\liminf_{k\to\infty}\frac{n_k-1}{n_k}n_kh^d\frac{\mu(S_{y+hz,L})}{\lambda(S_{y+hz,L})}L^d\lambda(S_{0,1})\right)K''(z)dzdy \\
&= \int f(y)\int_{z\notin S_{0,L}}\exp\left(-sf(y)L^d\lambda(S_{0,1})\right)K''(z)dzdy \\
&= \int f(y)\exp\left(-sL^d\lambda(S_{0,1})f(y)\right)dy\int_{z\notin S_{0,L}}K^*(z)dz
\end{aligned}
\tag{26}
$$

Ungleichungen (23), (24), (25) und (26) können wir jetzt alle zusammenfassen:

$$
\begin{aligned}
&\liminf_{k\to\infty} \int \mathbb{E}(|f_{n_k}(x) - g_h(x)|)dx \\
&\geq \liminf_{k\to\infty} \mathbb{E}\left(\int |f^*_{n_k}(x) - g^*_h(x)|)dx\right) - 2\delta \\
&\geq \int f(y) \exp\left(-sL^d\lambda(S_{0,1})f(y)\right) dy \left(\int_{S_{0,L}} K^*(z)dz - \int_{S^c_{0,L}} K^*(z)dz\right) - 2\delta \\
&\geq \int f(y) \exp\left(-sL^d\lambda(S_{0,1})f(y)\right) dy \left(2\int_{S_{0,L}} K^*(z)dz - 1\right) - 2\delta \\
&\to_{M\to\infty} \int f(y) \exp\left(-sL^d\lambda(S_{0,1})f(y)\right) dy \left(2\int_{S_{0,L}} K(z)dz - 1\right),
\end{aligned}
$$

wobei der letzte Schritt gilt wegen des Satzes von der majorisierten Konvergenz ($|K^*| \leq K$ und $|K - K^*| \leq K$) und weil M bisher beliebig war. Außerdem können wir nun L so groß wählen, dass der hintere Faktor $2\int_{S_{0,L}} K(z)dz - 1 =: c > 0$ ist, und haben daher

$$
0 = \liminf_{k\to\infty} \int \mathbb{E}(|f_{n_k}(x) - g_h(x)|)dx \geq c \int f(y) \exp\left(-sL^d\lambda(S_{0,1})f(y)\right) dy \geq 0.
$$

Damit diese letzte Gleichung hält, muss aber $s = \infty$ gelten, also ein Widerspruch zur Annahme (22). Deshalb konvergiert keine Teilfolge von nh^d gegen einen rellen Limes und folglich gilt Behauptung (20). □

Hilfssätze

In diesem Abschnitt werden wir, wie zu Beginn des vorhergehenden Beweises bemerkt, einige Hilfssätze aufführen und beweisen. Auch diese stammen wieder aus [9] und [10]

Als Haupthilfsmittel benötigen wir das sogenannte Lebesgue-Density-Theorem, das ein Korollar des Lebesque-Differentiation-Theorems ist:

Satz 2.2 (Lebesgue-Density-Theorem)**.** *Sei* $\mathcal{B} \subset \mathfrak{B}(\mathbb{R}^d)$ *und* $\mathcal{Q}_0 = \left\{[-a,a]^d \subset \mathbb{R}^d | a > 0\right\}$ *und es gelte*

$$
\sup_{B\in\mathcal{B}} \left(\min_{Q\in\mathcal{Q}_0, B\subseteq Q} \frac{\lambda(Q)}{\lambda(B)} \right) < \infty.
$$

Außerdem sei f *eine beliebige Wahrscheinlichkeitsdichte und* $(B_k)_{k\in\mathbb{N}}$ *eine Folge in* $\mathcal{B}$*, für die* $\lambda(B_k) \to_{k\to\infty} 0$ *gilt, dann folgt:*

$$\lim_{k\to\infty} \frac{1}{\lambda(B_k)} \int_{x+B_k} |f(y) - f(x)|dy = 0 \text{ für f.a. } x \in \mathbb{R}^d$$

und insbesondere

$$\lim_{k\to\infty} \frac{1}{\lambda(B_k)} \int_{x+B_k} f(y)dy = f(x) \text{ für f.a. } x \in \mathbb{R}^d.$$

Für einen Beweis siehe beispielsweise [27] (Theorem 7.16).

Hilfssatz 2.1 (Young'sche Ungleichung)**.** *Für beliebige $g, h \in L_1$ gilt:*

$$\int |f * g(x)|dx \leq \int |f(x)|dx \int |g(x)|dx$$

Beweis. Mit Fubini folgt sofort:

$$\begin{aligned}
\int |f * g(x)|dx &= \int |\int f(y)g(x-y)dy|dx \\
&\leq \int\int |f(y)||g(x-y)|dydx \\
&= \int |f(y)| \int g(x-y)|dxdy \\
&= \int |f(y)|dy \int |g(x)|dx.
\end{aligned}$$

□

Hilfssatz 2.2. *Sei f eine Wahrscheinlichkeitsdichte, $K \in L_1$ und $\int K = 1$, dann gilt:*

$$\lim_{h\downarrow 0} \int |f * K_h(x) - f(x)|dx = 0.$$

Beweis. Gelte die Behauptung anstatt für alle Wahrscheinlichkeitsdichten nur für eine dichte Teilmenge G. Dann gilt mit Hilfssatz 2.1 für $g \in G$ und f eine beliebige W-Dichte :

$$\begin{aligned}
\int |f * K_h(x) - f(x)|dx &\leq \int |(f-g) * K_h| + \int |f-g| + \int |g * K_h - g| \\
&\leq \int |(f-g)||K_h| + \int |f-g| + \int |g * K_h - g| \\
&= \Big(1 + \int |K|\Big) \int |f-g| + \int |g * K_h - g|,
\end{aligned}$$

wobei der erste Teil des letzten Terms aufgrund der Dichtheit beliebig klein gemacht werden kann und der zweite Teil nach Voraussetzung gegen

0 konvergiert. Es reicht also, die Behauptung für eine dichte Teilmenge zu zeigen.

Wir betrachten für G die stetigen Funtionen mit kompaktem Träger. Zu $M > 0$ und K definieren wir

$$K'(x) := K(x)\mathbb{1}_{\|x\| \leq M} \text{ und } K''(x) := K(x)\mathbb{1}_{\|x\| > M}$$

und es gilt für $g \in G$ wieder mit Hilfssatz 2.1:

$$\begin{aligned}
\int |g * K_h - g| &\leq \int \left| g * K'_h - g \int K'_h \right| + \int |g * K''_h| + \int \left| g \int K''_h \right| \\
&\leq \int \left| g * K'_h - g \int K'_h \right| + \int g \int |K''_h| + \int g \int |K''_h| \\
&= \int \left| g * K'_h - g \int K'_h \right| + 2 \int |K''_h|.
\end{aligned}$$

Der zweite Term kann durch Wahl von M beliebig klein gemacht werden, es reicht also zu zeigen, dass der erste Term gegen null konvergiert. Wir definieren

$$\omega(g, h) := \sup_{x,y:\|y\| \leq h} |g(x - y) - g(x)|.$$

Da g kompakten Träger hat existiert eine kompakte Menge C, sodass:

$$\begin{aligned}
\int \left| g * K'_h - g \int K'_h \right| &= \int_C \left| g * K'_h - g \int K'_h \right| \\
&= \int_C \left| \int (f(x - y) - f(x)) K'_h(y) dy \right| dx \\
&\leq \int_C \int |f(x - y) - f(x)||K'_h(y)| dy dx \\
&\leq \int_C \int \omega(y)|K'_h(y)| dy dx \\
&\leq \lambda(C) \int \omega(y)|K'_h(y)| dy \\
&\leq \lambda(C)\omega(hM) \int |K'(y)| dy,
\end{aligned}$$

wobei im letzten Schritt benutzt wurde, dass $K'_h(y) = 0$ für $\|y\| > hM$.

$\omega(g, hM)$ konvergiert bei festem M aber aufgrund der Stetigkeit von g für $h \to 0$ gegen 0, es folgt also die Behauptung. □

Hilfssatz 2.3. *Sei f eine Wahrscheinlichkeitsdichte und $K \in L_1$ beschränkt mit kompaktem Träger und $\int K = 1$, dann gilt:*

$$\lim_{h \downarrow 0} f * K_h(x) = f(x) \text{ für fast alle } x \in \mathbb{R}^d.$$

Beweis. Da K beschränkt ist und kompakten Träger hat, finden wir $C \in \mathbb{R}$ und $r > 0$, sodass

$$K \leq C \text{ und } supp(K) \subseteq S_{0,r} =: A.$$

Da auch K_h zu 1 integriert folgt daher:

$$\begin{aligned}
|f * K_h(x) - f(x)| &= \left| \int (f(x-y) - f(x)) K_h(y) dy \right| \\
&\leq \int |f(x-y) - f(x)||K_h(y)| dy \\
&= h^{-d} \int_{hA} |f(x-y) - f(x)||K_h\left(\frac{y}{h}\right)| dy \\
&\leq C\lambda(A)\lambda(hA)^{-1} \int_{hA} |f(x-y) - f(x)| dy \\
&= C\lambda(A) \frac{1}{\lambda(hA)} \int_{x+hA} |f(s) - f(x)| ds.
\end{aligned}$$

Der letzte Ausdruck konvergiert aber für $h \to 0$ und fast alle $x \in \mathbb{R}^d$ gegen 0 nach dem Lebesgue-Density-Theorem 2.2; daraus folgt die Behauptung. □

Hilfssatz 2.4. *Seien f und K beides Wahrscheinlichkeitsdichten, dann gilt für alle $h > 0$:*

$$\int |f * K_h(x) - f(x)| dx > 0.$$

Beweis. Für den Beweis benötigen wir die charakteristische Funktion – auch Fourier-Transformierte (F.T.) genannt – einer Verteilung $\mathbb{P}^X$ auf $\mathbb{R}^d$:

$$\phi_X(t) := \int \exp(itx) \mathbb{P}^X(dx), \text{ für } t \in \mathbb{R}^d.$$

Nach A.1 (siehe Anhang) ist eine solche Verteilung durch ihre Fourier-Transformierte bereits vollständig festgelegt. Bezeichnen ϕ, ψ und ν die charakteristischen Funktionen von f, K und $f * K_h$, so gilt für alle $t \in \mathbb{R}^d$:

$$\begin{aligned}\nu(t) &= \int f * K_h(x) e^{itx} dx \\ &= \int f(y) \int K_h(x-y) e^{itx} dx dy \\ &= \int f(y) e^{ity} \int h^{-d} K\left(\frac{x-y}{h}\right) e^{iht\frac{x-y}{h}} dx dy \\ &= \psi(ht)\phi(t).\end{aligned}$$

Nehmen wir an, es gelte $\int |f * K_h - f| = 0$ für ein $h > 0$, dann folgt schon $f * K_h(x) = f(x)$ für fast alle $x \in \mathbb{R}^d$ und somit aufgrund der Eindeutigkeit der charakteristischen Funktion auch $\psi(ht)\phi(t) = \nu(t) = \phi(t)$.

Offensichtlich nimmt jede F.T. in $t = 0$ den Wert 1 an, und nach A.2 ist sie auch stetig, daher gibt es eine Umgebung $S_{0,\epsilon}$ von 0 in der $\phi > 0$, also auch $\psi(ht) = 1$ für alle $t \in S_{0,\epsilon}$. In diesem Bereich gilt deshalb für die erste und zweite Ableitung: $\psi' = \psi'' = 0$ und somit nach A.3 insbesondere

$$\mathbb{E}(X_K^2) = -\psi''(0) = 0,$$

wobei $\mathbb{P}^{X_K}$ die Dichte K besitzt. Die Endlichkeit von $\mathbb{E}(X_K^2)$, die Voraussetzung für Gültigkeit der letzten Gleichung ist, folgt dabei aus der auch in Satz A.3 enthaltenen Umkehrung. Daraus folgt dann aber $X_K = 0$, ein Widerspruch, und somit die Behauptung.

□

Hilfssatz 2.5. *f und K seien Wahrscheinlichkeitsdichten, $h = h(n)$ eine Folge und*

$$\lim_{n\to\infty} \int |f * K_h(x) - f(x)| = 0. \tag{27}$$

Dann gilt schon

$$\lim_{n\to\infty} h = 0.$$

Beweis. Wir beweisen wieder über einen Widerspruch und nehmen zuerst an, dass

$$h \to c \in (0, \infty) \text{ entlang einer Teilfolge } n_k \text{ von } n.$$

Damit gilt

$$\int |f * K_h - f| \geq \int |f * K_c - f| - \int |f * K_c - f * K_h|.$$

Die linke Seite der Ungleichung konvergiert gegen 0 nach Voraussetzung und der erste Term der anderen Seite ist nach Hilfssatz 2.4 positiv. Wenn also der letzte Ausdruck gegen 0 konvergiert folgt der Widerspruch.

Sei dafür $K' : \mathbb{R}^d \to \mathbb{R}$ stetig mit kompaktem Träger, woraus insbesondere $K' \in L_1$ und $\max_{y\in\mathbb{R}^d} K'(y) < \infty$ folgt. Für eine ausreichend große, kompakte Menge A gilt also:

$$K'_h(x) = h^{-d}K'\left(\frac{x}{h}\right) \leq h^{-d}\mathbb{1}_A(x)\max_{y\in\mathbb{R}^d} K'(y) \leq (c-\delta)^{-d}\mathbb{1}_A(x)\max_{y\in\mathbb{R}^d} K'(y) \in \mathbb{R},$$

ab einem n_k groß genug (und damit $|c-h|$ klein genug) und $0 < \delta < c$ passend. K'_h wird also von einer L_1-Funktion majorisiert. Damit folgt dann mit Hilfssatz 2.1 und dem Satz von der majorisierten Konvergenz:

$$\begin{aligned}
\int |f * K_c - f * K_h| &\leq \int |K_c - K_h| \\
&\leq \int |K_c - K'_c| + \int |K'_c - K'_h| + \int |K_h - K'_h| \\
&= 2\int |K - K'| + \int |K'_c - K'_h| \\
&\to 2\int |K - K'| + \int \lim_{k\to\infty} |K'_c - K'_h| \\
&= 2\int |K - K'|
\end{aligned}$$

aufgrund der Stetigkeit von K'_h. Der letzte Term kann beliebig klein gemacht werden und damit folgt der erwünschte Widerspruch.

Es bleibt der Fall $\lim_{k\to\infty} h(n_k) = \infty$ entlang einer Teilfolge zu betrachten. Für die charakteristischen Funktionen (mit den gleichen Bezeichnungen wie im Beweis von Hilfssatz 2.4) wissen wir:

$$\begin{aligned}
\int |f * K_h(x) - f(x)|dx &= \int |e^{itx}||f * K_h(x) - f(x)|dx \\
&\geq |\int e^{itx}\left(f * K_h(x) - f(x)\right)dx| \\
&= |\psi(th)\phi(t) - \phi(t)|,
\end{aligned}$$

für $t \in \mathbb{R}^d$ beliebig und daraus folgt dann mit der Voraussetzung (27):

$$\lim_{k\to\infty} |\psi(th)\phi(t) - \phi(t)| \leq \lim_{k\to\infty} \int |f * K_h(x) - f(x)|dx = 0.$$

Da $\phi(t) > 0$ zumindest in einer Umgebung von $t = 0$, muss also $\lim_{h\to\infty} \psi(th) = 1$ gelten. Nach dem Riemann-Lebesgue-Lemma A.4 gilt aber $\lim_{\|t\|\to\infty} \phi(t) = 0$, also insbesondere auch $\lim_{h\to\infty} \psi(th) = 0 \neq 1$, ein Widerspruch. Die Behauptung folgt. □