University of Tübingen

— Seminar —
Statistical Learning

The strong universal consistency of the kernel estimate and the k-NN estimate

inspired by the book
"A distribution-free theory of nonparametric Regression"
by László Györfi

Authors
Valentin Pfisterer
Tobias Winkler

Supervisors
Prof. Dr. A. Prohl
Dr. A. Chaudhary

May 24, 2023

Contents

1. Quick reminder

To begin with, there is a graphic (Figure 1) presented as a visual aid to illustrate our current content position. The topics of weak universal consistency and strong universal consistency of partitioning estimates have already been addressed, as indicated by the checkmarks. However, the two question marks imply that these topics have not yet been discussed, which we will now proceed to do.

	Partitioning	**Kernel**	**k-NN**
weak universal consistency	✓	✓	✓
strong (universal) consistency	✓	**?**	**?**

Figure 1: Current position regarding to our content.

Definition 1.1 (kernel estimate)**.** Let $K : \mathbb{R}^d \to \mathbb{R}_+$ be a function called kernel function, and let $h > 0$ be a bandwidth. The kernel estimate is defined by

$$m_n(x) = \frac{\sum_{i=1}^n K(\frac{x-X_i}{h})Y_i}{\sum_{i=1}^n K(\frac{x-X_i}{h})}.$$

Definition 1.2 (k-NN estimate)**.** For $x \in \mathbb{R}^d$ let

$$(X_{(1)}(x), Y_{(1)}(x)), ..., (X_{(n)}(x), Y_{(n)}(x))$$

be a permutation of

$$(X_1, Y_2), ..., (X_n, Y_n)$$

such that

$$\|x - X_{(1)}(x)\| \leqslant ... \leqslant \|x - X_{(n)}(x)\|.$$

The k-NN estimate is defined by

$$m_n(x) = \frac{1}{k}\sum_{i=1}^k Y_{(i)}(x).$$

Definition 1.3 (strong universal consistency)**.** A sequence of regression function estimates $\{m_n\}_n$ is called strongly consistent for a certain distribution of (X, Y), if

$$\lim_{n\to\infty} \int |m_n(x) - m(x)|^2 \mu(\mathrm{d}x) = 0 \quad \text{with probability one.}$$

This sequence is called strongly universal consistent if it is strongly consistent for all distributions of (X, Y) with $\mathbb{E}(Y^2) < \infty$.

2. Kernel Estimates

In this section we are going to proof the strong consistency for the kernel estimate. Therefore, we consider a rather general class of kernels.

Definition 2.1 (regular kernel)**.** A kernel is called regular if it is non-negative, and there exists a ball $B_r(0)$ with radius $r > 0$ and a constant $b > 0$ such that

$$1 \geqslant K(x) \geqslant b\mathbb{1}_{\{x \in B_r(0)\}}$$

and

$$\int \sup_{u \in x + B_r(0)} K(u)\,\mathrm{d}x < \infty. \tag{1}$$

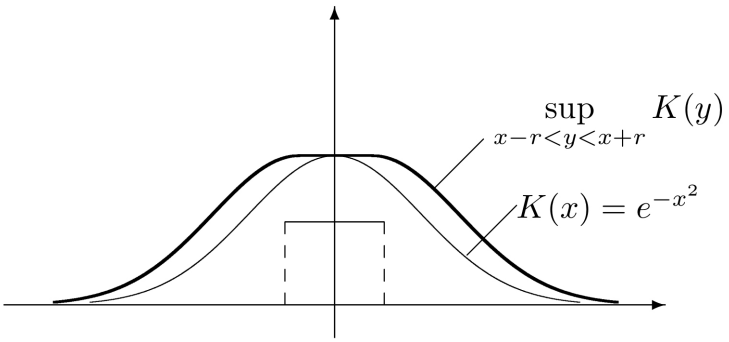

Figure 2: Regular kernel.

Put $K_h(x) := K(\frac{x}{h})$.

The following organizational chart (Figure 3) is intended to provide an overview of the structure of the proof of theorem 2.5, in order to better understand the organization and connections of the individual proof components.

Lemma 2.1 (Covering Lemma)**.** *If the kernel is regular then there exists a constant $\varrho \equiv \varrho(K) < \infty$ such that for any $u \in \mathbb{R}^d$, $h > 0$ and every probability measure μ*

$$\int \frac{K_h(x-u)}{\int K_h(x-z)\mu(\mathrm{d}z)}\mu(\mathrm{d}x) \leqslant \varrho.$$

Moreover, for any $\delta > 0$

$$\lim_{n\to\infty} \sup_{u \in \mathbb{R}^d} \int \frac{K_h(x-u)\mathbb{1}_{\{\|x-u\|>\delta\}}}{\int K_h(x-z)\mu(\mathrm{d}z)}\mu(\mathrm{d}x) = 0.$$

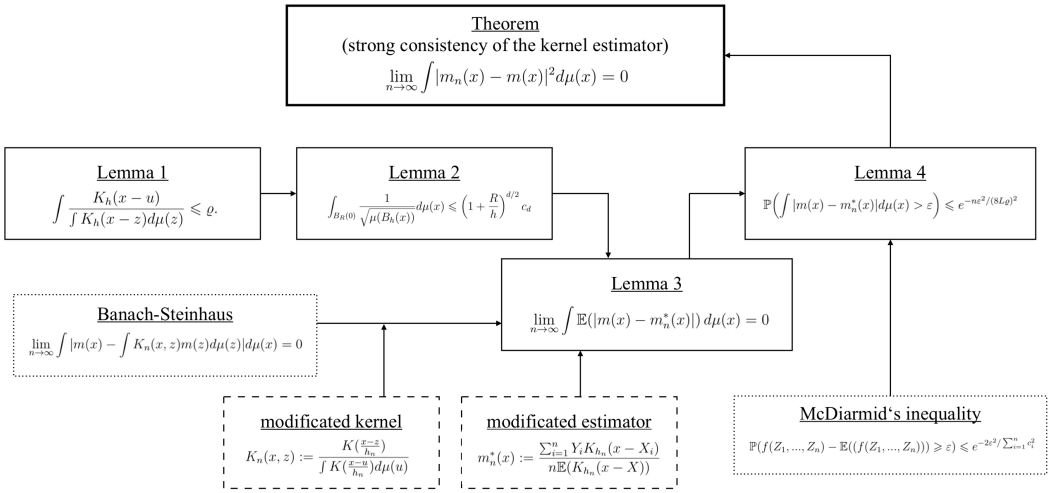

Figure 3: Organizational chart of the proof of theorem 2.5.

Proof. The kernel is regular which especially implies the existence of $r > 0$ and $b > 0$ such that

$$1 \geqslant K(x) \geqslant b\mathbb{1}_{\{x\in B_r(0)\}}.$$

Hence, there exists a bounded covering of $\mathbb{R}^d$ of balls with radius $\frac{r}{2}$ and centers x_i $(i = 1, 2, ...)$, i.e.

$$\mathbb{R}^d = \bigcup_{i\in\mathbb{N}} B_{r/2}(x_i).$$

This cover has an infinite number of member balls, but every x gets covered at most k_1 times (see Figure 4), where k_1 depends upon d only, which implies

$$\sum_{i=1}^{\infty} \mathbb{1}_{\{x\in B_{r/2}(x_i)\}} \leqslant k_1. \tag{2}$$

The integral condition (1) on K implies that

$$\begin{aligned}
(3)\quad & \sum_{i=1}^{\infty} \sup_{z\in B_{r/2}(x_i)} K(z) \\
&= \sum_{i=1}^{\infty} \sup_{z\in B_{r/2}(x_i)} K(z) \cdot \frac{\int_{\{x\in B_{r/2}(x_i)\}} \mathrm{d}x}{\int_{\{x\in B_{r/2}(0)\}} \mathrm{d}x} \quad \text{(by translation invariance)} \\
&= \sum_{i=1}^{\infty} \frac{1}{\int_{B_{r/2}(0)} \mathrm{d}x} \int_{\{x\in B_{r/2}(x_i)\}} \sup_{z\in B_{r/2}(x_i)} K(z)\,\mathrm{d}x \\
&= \frac{1}{\int_{B_{r/2}(0)} \mathrm{d}x} \int \sum_{i=1}^{\infty} \mathbb{1}_{\{x\in B_{r/2}(x_i)\}} \sup_{z\in B_{r/2}(x_i)} K(z)\,\mathrm{d}x \quad \text{(by Fubini's theorem)}
\end{aligned}$$

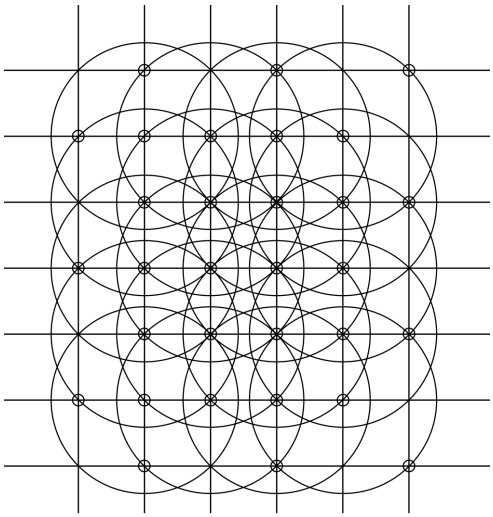

Figure 4: An example of a bounded overlap of $\mathbb{R}^2$.

$$
\begin{aligned}
&\leqslant \frac{1}{\int_{B_{r/2}(0)} \mathrm{d}x} \int \sum_{i=1}^{\infty} \mathbb{1}_{\{x \in B_{r/2}(x_i)\}} \sup_{z \in B_r(x)} K(z)\, \mathrm{d}x \\
&\qquad\qquad \text{(because } x \in B_{r/2}(x_i) \Rightarrow B_{r/2}(x_i) \subseteq B_r(x)\text{)} \\
&\leqslant \frac{k_1}{\int_{B_{r/2}(0)} \mathrm{d}x} \int \sup_{z \in B_r(x)} K(z)\, \mathrm{d}x \quad \text{(because of (2))} \\
&\leqslant k_2 \quad \text{(because of (1))}
\end{aligned}
$$

for another finite constant k_2. Furthermore,

$$
\begin{aligned}
K_h(x-u) = K\Big(\frac{x-u}{h}\Big) &\quad \text{(by definition)} \\
&\leqslant \sum_{i=1}^{\infty} \sup_{\frac{x-u}{h} \in B_{r/2}(x_i)} K\Big(\frac{x-u}{h}\Big) \\
&= \sum_{i=1}^{\infty} \sup_{x \in u + hB_{r/2}(x_i)} K_h(x-u)
\end{aligned}
$$

and, for $x \in u + hB_{r/2}(x_i)$,

$$
\begin{aligned}
(4) \quad \int K_h(x-z)\mu(\mathrm{d}z) &= \int K\Big(\frac{x-z}{h}\Big)\mu(\mathrm{d}z) \\
&\geqslant \int b \cdot \mathbb{1}_{\{\frac{x-z}{h} \in B_r(0)\}} \mu(\mathrm{d}z) \\
&\qquad \text{(by definition of regular kernels)}
\end{aligned}
$$

$$
\begin{aligned}
&= b \cdot \int \mathbb{1}_{\{z \in x + hB_r(0)\}} \mu(\mathrm{d}z) \\
&= b \cdot \mu(x + hB_r(0)) \\
&\geqslant b \cdot \mu(u + hB_{r/2}(x_i)), \quad \text{(because } x \in u + hB_{r/2}(x_i)\text{)}
\end{aligned}
$$

from which we conclude

$$
\begin{aligned}
&\int \frac{K_h(x-u)}{\int K_h(x-z)\mu(\mathrm{d}z)} \mu(\mathrm{d}x) \\
&\leqslant \sum_{i=1}^{\infty} \int_{\{x \in u + hB_{r/2}(x_i)\}} \frac{K_h(x-u)}{\int K_h(x-z)\mu(\mathrm{d}z)} \mu(\mathrm{d}x) \quad \text{(because of the covering)} \\
&\leqslant \sum_{i=1}^{\infty} \int_{\{x \in u + hB_{r/2}(x_i)\}} \frac{\sup_{z \in hB_{r/2}(x_i)} K_h(z)}{\int K_h(x-z)\mu(\mathrm{d}z)} \mu(\mathrm{d}x) \quad \text{(because of the choice of x)} \\
&\leqslant \sum_{i=1}^{\infty} \int_{\{x \in u + hB_{r/2}(x_i)\}} \frac{\sup_{z \in hB_{r/2}(x_i)} K_h(z)}{b \cdot \mu(u + hB_{r/2}(x_i))} \mu(\mathrm{d}x) \quad \text{(because of (4))} \\
&= \sum_{i=1}^{\infty} \frac{\sup_{z \in hB_{r/2}(x_i)} K_h(z)}{b \cdot \mu(u + hB_{r/2}(x_i))} \cdot \mu(u + hB_{r/2}(x_i)) \\
&= \frac{1}{b} \sum_{i=1}^{\infty} \sup_{z \in hB_{r/2}(x_i)} K_h(z) \\
&\leqslant \frac{k_2}{b} \quad \text{(by (3))},
\end{aligned}
$$

where k_2 depends on K and d only. Define $\varrho := \frac{k_2}{b}$ and get the first statement. To obtain the second statement in the lemma, substitute $K_h(z)$ above by $K_h(z)\mathbb{1}_{\{\|z\|>\delta\}}$ for $\delta > 0$, which implies

$$
\begin{aligned}
&\sup_{u \in \mathbb{R}^d} \int \frac{K_h(x-u)\mathbb{1}_{\{\|x-u\|>\delta\}}}{\int K_h(x-z)\mu(\mathrm{d}z)} d\mu(x) \\
&\leqslant \sup_{u \in \mathbb{R}^d} \frac{1}{b} \sum_{i=1}^{\infty} \sup_{z \in hB_{r/2}(x_i)} K_h(z)\mathbb{1}_{\{\|z\|>\delta\}} \quad \text{(by above)} \\
&= \sup_{u \in \mathbb{R}^d} \frac{1}{b} \sum_{i=1}^{\infty} \sup_{z \in B_{r/2}(x_i)} K(z)\mathbb{1}_{\{\|z\|>\delta/h\}} \\
&\longrightarrow 0
\end{aligned}
$$

as $h \to 0$ by dominated convergence. ■

Lemma 2.2. *Let $1 \leqslant R < \infty, 0 < h \leqslant R$ and let $B_R(0) \subseteq \mathbb{R}^d$ be a ball of*

radius R. Then, for every probability measure μ,

$$\int_{B_R(0)} \frac{1}{\sqrt{\mu(B_h(x))}} \mu(\mathrm{d}x) \leqslant \left(1 + \frac{R}{h}\right)^{d/2} c_d,$$

where c_d depends upon the dimension d only.

Proof. Let $1 \leqslant R < \infty, 0 < h \leqslant R$. Hence, there exists $M \in \mathbb{N}$ with

$$M \leqslant \frac{c_d'}{h^d} \leqslant \frac{c_d'}{\left(\frac{h}{h+R}\right)^d} = c_d' \left(\frac{h+R}{h}\right)^d$$

for a constant c_d' which depends on the dimension d only, such that

$$B_h(0) \subseteq \bigcup_{j=1}^{M} B_{h/2}(z_j)$$

with centers $z_1, ..., z_M$. Therefore, by the Cauchy-Schwarz inequality

$$\begin{aligned}
\int_{B_R(0)} \frac{1}{\sqrt{\mu(B_h(x))}} \mu(\mathrm{d}x) &\leqslant \left(\int_{B_R(0)} \frac{1}{\mu(B_h(x))} \mu(\mathrm{d}x)\right)^{\frac{1}{2}} \\
&\leqslant \left(\sum_{j=1}^{M} \int \frac{\mathbb{1}_{\{x \in B_{h/2}(z_j)\}}}{\mu(B_h(x))} \mu(\mathrm{d}x)\right)^{\frac{1}{2}} \\
&\leqslant \left(\sum_{j=1}^{M} \int \frac{\mathbb{1}_{\{x \in B_{h/2}(z_j)\}}}{\mu(B_{h/2}(z_j))} \mu(\mathrm{d}x)\right)^{\frac{1}{2}} \\
&\leqslant \left(\sum_{j=1}^{M} 1\right)^{\frac{1}{2}} \\
&= M^{\frac{1}{2}} \\
&\leqslant \left(c_d' \left(\frac{h+R}{h}\right)^d\right)^{\frac{1}{2}} \\
&= \left(1 + \frac{R}{h}\right)^{d/2} \underbrace{c_d}_{:=\sqrt{c_d'}}.
\end{aligned}$$

■

Define

$$m_n^*(x) := \frac{\sum_{i=1}^n Y_i K_{h_n}(x - X_i)}{n\mathbb{E}(K_{h_n}(x - X))}.$$

Lemma 2.3. *Let m_n be the kernel estimate of the regression function m with a regular kernel K. Assume that there is an $L < \infty$ such that $P(|Y| \leqslant L) = 1$. If*

$$h_n \to 0 \quad \textit{and} \quad nh_n^d \to \infty,$$

then we have

$$\lim_{n\to\infty} \int \mathbb{E}(|m(x) - m_n^*(x)|)\, \mu(\mathrm{d}x) = 0.$$

Proof. By the triangle inequality

$$\begin{aligned}
&\int \mathbb{E}(|m(x) - m_n^*(x)|)\, \mu(\mathrm{d}x) \\
&= \int \mathbb{E}(|m(x) - \mathbb{E}(m_n^*(x)) + \mathbb{E}(m_n^*(x)) - m_n^*(x)|)\, \mu(\mathrm{d}x) \\
&\leqslant \int \mathbb{E}(|m(x) - \mathbb{E}(m_n^*(x))\,|)\, \mu(\mathrm{d}x) + \int \mathbb{E}(|\mathbb{E}(m_n^*(x)) - m_n^*(x)|)\, \mu(\mathrm{d}x) \\
&=: I_n + J_n.
\end{aligned}$$

Concerning the term I_n verify the conditions of the Theorem of Banach-Steinhaus (Appendix, Theorem A.1) for

$$K_n(x, z) := \frac{K(\frac{x-z}{h_n})}{\int K(\frac{x-u}{h_n})\mu(\mathrm{d}u)}.$$

By this definition, we have

$$\begin{aligned}
(5) \quad \mathbb{E}(m_n^*(x)) &= \mathbb{E}\left(\frac{\sum_{i=1}^n Y_i K_{h_n}(x - X_i)}{nE(K_{h_n}(x - X))}\right) \quad \text{(by definition)} \\
&= \frac{1}{n\mathbb{E}(K_{h_n}(x - X))} n\mathbb{E}(Y K_{h_n}(x - X)) \quad (X, X_1, X_2, \ldots \text{ iid.}) \\
&= \frac{\mathbb{E}(Y K_{h_n}(x - X))}{\mathbb{E}(K_{h_n}(x - X))} \\
&= \frac{\mathbb{E}(\mathbb{E}(Y K_{h_n}(x - X)|X))}{\mathbb{E}(K_{h_n}(x - X))} \quad \text{(by tower property)} \\
&= \frac{\mathbb{E}(K_{h_n}(x - X)\mathbb{E}(Y|X))}{\mathbb{E}(K_{h_n}(x - X))}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\mathbb{E}(K_{h_n}(x-X)m(X))}{\mathbb{E}(K_{h_n}(x-X))} \quad \text{(by the definition of } m\text{)} \\
&= \int \frac{K(\frac{x-z}{h_n})}{\int K(\frac{x-u}{h_n}\mu(\mathrm{d}u))} m(z)\mu(\mathrm{d}z) \\
&= \int K_n(x,z)m(z)\mu(\mathrm{d}z). \quad \text{(by definition of } K_n\text{)}
\end{aligned}
$$

Part (i) follows from the covering lemma with $c = \varrho$: Let $z \in \mathbb{R}^d$. We get

$$
\begin{aligned}
\int |K_n(x,z)|\mu(\mathrm{d}x) &= \int \frac{K(\frac{x-z}{h_n})}{\int K(\frac{x-u}{h_n})\mu(\mathrm{d}u)}\mu(\mathrm{d}x) \quad \text{(by definition of } K_n\text{)} \\
&= \int \frac{K_{h_n}(x-z)}{\int K_{h_n}(x-u)\mu(\mathrm{d}u))}\mu(\mathrm{d}x) \quad \text{(by } K_h(x) := K(x/h)\text{)} \\
&\leqslant \varrho. \quad \text{(by Lemma 2.1)}
\end{aligned}
$$

Part (ii) follows from the covering lemma with $d = \varrho$: Let $x \in \mathbb{R}^d$. Hence,

$$
\begin{aligned}
\int |K_n(x,z)|\mu(\mathrm{d}z) &= \int \frac{K(\frac{x-z}{h_n})}{\int K(\frac{z-u}{h_n})\mu(\mathrm{d}u)}\mu(\mathrm{d}z) \quad \text{(by definition of } K_n\text{)} \\
&= \int \frac{K_{h_n}(x-z)}{\int K_{h_n}(z-u)\mu(\mathrm{d}u))}\mu(\mathrm{d}z) \quad \text{(by } K_h(x) := K(x/h)\text{)} \\
&\leqslant \varrho. \quad \text{(by Lemma 2.1)}
\end{aligned}
$$

Let $a > 0$. We get part (iii) by

$$
\begin{aligned}
&\int\int K_n(x,z)\mathbb{1}_{\{\|x-z\|>a\}}\mu(\mathrm{d}z)\mu(\mathrm{d}x) \\
&= \int\int \frac{K(\frac{x-z}{h_n})}{\int K(\frac{x-u}{h_n}\mu(\mathrm{d}u))}\mathbb{1}_{\{\|x-z\|>a\}}\mu(\mathrm{d}z)\mu(\mathrm{d}x) \quad \text{(by definition of } K_n\text{)} \\
&= \int \frac{\int K(\frac{x-z}{h_n})\mathbb{1}_{\{\|x-z\|>a\}}\mu(\mathrm{d}z)}{\int K(\frac{x-u}{h_n})\mu(\mathrm{d}u)}\mu(\mathrm{d}x) \\
&= \int \frac{\int K_{h_n}(x-z)\mathbb{1}_{\{\|x-z\|>a\}}\mu(\mathrm{d}z)}{\int K_{h_n}(x-u)\mu(\mathrm{d}u)}\mu(\mathrm{d}x) \\
&\longrightarrow 0
\end{aligned}
$$

by the second statement of the covering lemma (2.1) by using a instead of δ. Part (iv) is obvious since

$$
ess\sup_{x\in\mathbb{R}^d} \int K_n(x,z)\mu(\mathrm{d}z) = ess\sup_{x\in\mathbb{R}^d} \frac{\int K(\frac{x-z}{h_n})\mu(\mathrm{d}z)}{\int K(\frac{x-u}{h_n})\mu(\mathrm{d}u)} = ess\sup_{x\in\mathbb{R}^d} 1 = 1.
$$

By applying the theorem of Banach-Steinhaus (Appendix, A.1) we get

$$\begin{aligned}
&\int \mathbb{E}(|m(x) - \mathbb{E}(m_n^*(x))\,|)\,\mu(\mathrm{d}x) \\
&= \int \mathbb{E}\Big(|m(x) - \int K_n(x,z)m(z)\mu(\mathrm{d}z)|\Big)\,\mu(\mathrm{d}x) \quad \text{(by (5))} \\
&\longrightarrow 0.
\end{aligned}$$

For the second term J_n we have with $h = h_n$ for convenience,

$$\begin{aligned}
&\mathbb{E}(|m_n^*(x) - E(m_n^*(x))|) \\
&= \sqrt{\mathbb{E}(|m_n^*(x) - \mathbb{E}(m_n^*(x))\,|^2)} \quad \text{(by Cauchy-Schwarz)} \\
&= \sqrt{\mathbb{E}\left(\left|\frac{\sum_{i=1}^n Y_i K_h(x-X_i)}{n\mathbb{E}(K_n(x-X_i))} - \mathbb{E}\left(\frac{\sum_{i=1}^n Y_i K_h(x-X_i)}{n\mathbb{E}(K_n(x-X_i))}\right)\right|^2\right)} \quad \text{(by definition)} \\
&= \sqrt{\mathbb{E}\left(\left|\frac{nYK_h(x-X) - n\mathbb{E}(YK_h(x-X))}{n\mathbb{E}(K_h(x-X))}\right|^2\right)} \quad (X, X_1, X_2, \ldots \text{ iid}) \\
&= \sqrt{\frac{n\mathbb{E}\Big(|YK_h(x-X) - \mathbb{E}(YK_h(x-X)|\Big)^2}{n^2\mathbb{E}(K_h(x-X)))^2}} \\
&= \sqrt{\frac{\mathbb{E}(|YK_h(x-X) - \mathbb{E}(YK_h(x-X)|^2))}{n\mathbb{E}(K_h(x-X))^2}} \\
&\leqslant \sqrt{\frac{\mathbb{E}(|YK_h(x-X)|^2)}{n\mathbb{E}(K_h(x-X))^2}} \quad \text{(by definition of the variance: } \mathbb{E}(YK_h(x-X)) \geqslant 0) \\
&\leqslant L\sqrt{\frac{\mathbb{E}(|K_h(x-X)|^2)}{n\mathbb{E}(K_h(x-X))^2}} \quad \text{(since } |Y| \leqslant L) \\
&\leqslant L\sqrt{\frac{\mathbb{E}\Big(|K(\frac{x-X}{h})|^2\Big)}{n\mathbb{E}\Big(K(\frac{x-X}{h})\Big)^2}} \quad \text{(by } K_h(x) := K(x/h)) \\
&\leqslant L\sqrt{K_{max}}\sqrt{\frac{\mathbb{E}\Big(K(\frac{x-X}{h})\Big)}{n\mathbb{E}\Big(K(\frac{x-X}{h})\Big)^2}} \\
&\leqslant L\sqrt{\frac{K_{max}}{b}}\sqrt{\frac{1}{n\mu(B_h(x))}}, \quad \text{(by (4))}
\end{aligned}$$

where K_{max} is a upper bound of K, i.e. $K(x) \leqslant K_{max}\ \forall x \in \mathbb{R}$ which is possible reasoned by the definition of regular kernels.
Let $R > 0$. Hence, we can split the integral such that

$$\begin{aligned}
&\int_{\mathbb{R}^d} \mathbb{E}(|m_n^*(x) - \mathbb{E}(m_n^*(x))\,|)\,\mu(\mathrm{d}x) \\
&= \int_{B_R(0)} \mathbb{E}(|m_n^*(x) - \mathbb{E}(m_n^*(x))\,|)\,\mu(\mathrm{d}x) + \int_{B_R(0)^c} \mathbb{E}(|m_n^*(x) - \mathbb{E}(m_n^*(x))\,|)\,\mu(\mathrm{d}x).
\end{aligned}$$

For the integral outside the ball we have

$$\begin{aligned}
&\int_{B_R(0)^c} \mathbb{E}(|m_n^*(x) - \mathbb{E}(m_n^*(x))\,|)\,\mu(\mathrm{d}x) \\
&\leqslant 2\int_{B_R(0)^c} \mathbb{E}(|m_n^*(x)|)\,\mu(\mathrm{d}x) \quad \text{(by triangle inequality)} \\
&\leqslant 2L\mu(B_R(0)^c) \quad \text{(by assumption } |Y| \leqslant L \text{ and by definition of } m_n^*) \\
&\longrightarrow 0
\end{aligned}$$

since $R \to \infty$.
To bound the integral over $B_R(0)$ we employ Lemma 2.2:

$$\begin{aligned}
&\int_{B_R(0)} \mathbb{E}(|m_n^*(x) - \mathbb{E}(m_n^*(x))\,|)\,\mu(\mathrm{d}x) \\
&\leqslant \int_{B_R(0)} L\sqrt{\frac{K_{max}}{b}}\sqrt{\frac{1}{n\mu(B_h(x))}}\mu(\mathrm{d}x) \quad \text{(by the inequality obtained above)} \\
&= L\sqrt{\frac{K_{max}}{b}}\frac{1}{\sqrt{n}}\int_{B_R(0)} \sqrt{\frac{1}{\mu(B_h(x))}}\mu(\mathrm{d}x) \\
&\leqslant L\sqrt{\frac{K_{max}}{b}}\frac{1}{\sqrt{n}}\left(1+\frac{R}{h}\right)^{d/2} c_d \quad \text{(by Lemma 2.2)} \\
&= L\sqrt{\frac{K_{max}}{b}}\left(1+\frac{R}{hn^{1/d}}\right)^{d/2} c_d \\
&\longrightarrow 0 \quad \text{(by assumption } nh_n^d \to \infty \text{ and continuity).}
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\int \mathbb{E}(|m(x) - m_n^*(x)|)\,\mu(\mathrm{d}x) \\
&\leqslant \int \mathbb{E}(|m(x) - \mathbb{E}(m_n^*(x))\,|)\,\mu(\mathrm{d}x) + \int \mathbb{E}(\mathbb{E}(m_n^*(x)) - m_n^*(x)|)\,\mu(\mathrm{d}x) \\
&\longrightarrow 0.
\end{aligned}$$

■

Lemma 2.4. *For n large enough*

$$\mathbb{P}\Big(\int |m(x) - m_n^*(x)|\mu(\mathrm{d}x) > \varepsilon\Big) \leqslant e^{-n\varepsilon^2/(8L\varrho)^2}.$$

Proof. We use the decomposition

$$\begin{aligned}
&\int |m(x) - m_n^*(x)|\mu(\mathrm{d}x) \\
&= \int |m(x) - m_n^*(x)| - \mathbb{E}(|m(x) - m_n^*(x)|) + \mathbb{E}(|m(x) - m_n^*(x)|)\,\mu(\mathrm{d}x) \\
&= \int \mathbb{E}(|m(x) - m_n^*(x)|)\,\mu(\mathrm{d}x) + \int |m(x) - m_n^*(x)| - \mathbb{E}(|m(x) - m_n^*(x)|)\,\mu(\mathrm{d}x).
\end{aligned}$$

The first term on the right-hand side tends to 0 by Lemma 2.3. It remains to show that the second term on the right-hand side is small with large probability. To do this, we use McDiarmid's inequality (Appendix, A.2) for

$$\int |m(x) - m_n^*(x)|\mu(\mathrm{d}x) - \mathbb{E}\Big(\int |m(x) - m_n^*(x)|\mu(\mathrm{d}x)\Big).$$

Fix the training data at $((x_1, y_1), ..., (x_n, y_n))$ and replace the ith pair (x_i, y_i) by $(\hat{x}_i, \hat{y}_i)$, changing the value of $m_n^*(x)$ to $m_{ni}^*(x)$. In detail, we consider the transformation T_i for $i = 1, ..., n$ defined as

$$((x_1, y_1), ..., (x_n, y_n)) \mapsto ((x_1, y_1), ..., (x_{i-1}, y_{i-1}), (\hat{x}_i, \hat{y}_i), (x_{i+1}, y_{i+1}), ..., (x_n, y_n))$$

where $(\hat{x}_i, \hat{y}_i) \in \mathbb{R}^d \times \mathbb{R}$. Hence,

$$m_{ni}^*(x, D_n(\omega)) = m_n^*(x, T_i(D_n(\omega))).$$

Clearly, by the covering lemma (Lemma 2.1), we get for fixed training data, i.e. $\omega \in \Omega$ fixed,

$$\begin{aligned}
&\int |m(x) - m_n^*(x)|\mu(\mathrm{d}x) - \int |m(x) - m_{ni}^*(x)|\mu(\mathrm{d}x) \\
&= \int |m(x) - m_n^*(x)| - |m(x) - m_{ni}^*(x)|\mu(\mathrm{d}x) \\
&\leqslant \int |m(x) - m_n^*(x) - m(x) - m_{ni}^*(x)|\mu(\mathrm{d}x) \quad \text{(by reverse triangle inequality)} \\
&= \int |m_n^*(x) - m_{ni}^*(x)|\mu(\mathrm{d}x) \\
&= \int \left| \frac{\sum_{j=1}^n y_j K_{h_n}(x - x_j)}{n\mathbb{E}(K_{h_n}(x - X))} - \frac{\hat{y}_i K_{h_n}(x - \hat{x}_i) + \sum_{j \neq i}^n y_j K_{h_n}(x - x_j)}{n\mathbb{E}(K_{h_n}(x - X))} \right| \mu(\mathrm{d}x) \quad \text{(by definition)}
\end{aligned}$$

$$
\begin{aligned}
&= \int \left| \frac{y_i K_{h_n}(x - x_i) - \hat{y}_i K_{h_n}(x - \hat{x}_i)}{n\mathbb{E}(K_{h_n}(x - X))} \right| \mu(\mathrm{d}x) \quad \text{(by definition)} \\
&\leqslant \sup_{y \in \mathbb{R}^d} \int \frac{2L K_{h_n}(x - y)}{n\mathbb{E}(K_{h_n}(x - X))} \mu(\mathrm{d}x) \\
&= \frac{2L}{n} \sup_{y \in \mathbb{R}^d} \int \frac{K_{h_n}(x - y)}{\int K_{h_n}(x - z)\mu(\mathrm{d}z)} \mu(\mathrm{d}x) \\
&\leqslant \frac{2L\varrho}{n}. \quad \text{(by Lemma 2.1)}
\end{aligned}
$$

So by Theorem A.2, for n large enough we get

$$
I_n := \mathbb{E}\left(\int |m(x) - m_n^*(x)| \mu(\mathrm{d}x) \right) < \frac{\varepsilon}{2} \quad \text{(by Lemma 2.3)}.
$$

We notice that for random Variables A, B, C, we have that

$$
\mathbb{P}(B > C) \leqslant \mathbb{P}(A > C) \quad \text{if } A \geqslant B, \tag{6}
$$

and thus

$$
\begin{aligned}
&\mathbb{P}\left(\int |m(x) - m_n^*(x)| \mu(\mathrm{d}x) > \varepsilon \right) \\
&= \mathbb{P}\left(\int |m(x) - m_n^*(x)| \mu(\mathrm{d}x) - I_n > \varepsilon - I_n \right) \\
&\leqslant \mathbb{P}\left(\int |m(x) - m_n^*(x)| \mu(\mathrm{d}x) - \mathbb{E}\left(\int |m(x) - m_n^*(x)| \mu(\mathrm{d}x) \right) > \frac{\varepsilon}{2} \right) \\
&\leqslant e^{-2\frac{(\varepsilon/2)^2}{4L^2\varrho^2}} \\
&= e^{-n\varepsilon^2/(8L^2\varrho^2)}.
\end{aligned}
$$

The proof is now complete. ■

Theorem 2.5. *Let m_n be the kernel estimate of the regression function m with a regular kernel K. Assume that there is an $L < \infty$ such that $P(|Y| \leqslant L) = 1$. If*

$$
h_n \to 0 \quad \textit{and} \quad n h_n^d \to \infty,
$$

then the kernel estimate is strongly consistent.

Proof. (proof of Theorem 2.5) It suffices to show that

$$
\lim_{n \to \infty} \int |m_n(x) - m(x)| \mu(\mathrm{d}x) = 0, \quad \mathbb{P}\text{-a.s.}
$$

due to

$$|m_n(x) - m(x)| \leqslant |m_n(x)| + |m(x)| \leqslant 2L$$

and thus

$$|m_n(x) - m(x)|^2 \leqslant 2L|m_n(x) - m(x)|.$$

With the triangle inequality we get the decomposition

$$\int |m_n(x) - m(x)|\mu(\mathrm{d}x) \leqslant \int |m_n(x) - m_n^*(x)|\mu(\mathrm{d}x) + \int |m_n^*(x) - m(x)|\mu(\mathrm{d}x)$$

and, according to Lemma 2.4,

$$\int |m_n^*(x) - m(x)|\mu(\mathrm{d}x) \to 0$$

with probability 1. On the other hand,

$$\begin{aligned}
&|m_n^*(x) - m_n(x)| \\
&= \left| \frac{\sum_{i=1}^n Y_i K_{h_n}(x - X_i)}{n\mathbb{E}(K_{h_n}(x - X))} - \frac{\sum_{i=1}^n Y_i K_{h_n}(x - X_i)}{\sum_{i=1}^n K_{h_n}(x - X_i)} \right| \\
&= \left| \sum_{i=1}^n Y_i K_{h_n}(x - X_i) \right| \left| \frac{1}{n\mathbb{E}(K_{h_n}(x - X))} - \frac{1}{\sum_{i=1}^n K_{h_n}(x - X_i)} \right| \\
&\leqslant L \left| \sum_{i=1}^n K_{h_n}(x - X_i) \right| \left| \frac{1}{n\mathbb{E}(K_{h_n}(x - X))} - \frac{1}{\sum_{i=1}^n K_{h_n}(x - X_i)} \right| \\
&= L \left| \frac{\sum_{i=1}^n Y_i K_{h_n}(x - X_i)}{n\mathbb{E}(K_{h_n}(x - X))} - 1 \right| \\
&= L|M_n^*(x) - 1|,
\end{aligned}$$

where M_n^* is a special form of $m_n^*(x)$ for $(X, 1)$.
In this case, $M(x) = \mathbb{E}(Y|X = x) = \mathbb{E}(1|X = x) = 1 \ \forall x \in \mathbb{R}^d$ with $Y \equiv 1$. Therefore,

$$\int |m_n^*(x) - m_n(x)|\mu(\mathrm{d}x) \leqslant L \int |M_n^*(x) - 1|\mu(\mathrm{d}x) \to 0 \quad \mathbb{P}\text{-a.s.}$$

which completes the proof. ■

The subsequent lemmas help us to prove the following theorem which considers the strong universal consistency of the kernel estimate. These both lemmas will not be proved here. To review their proofs, please refer to section 23.1 of Györfi's book.

Lemma 2.6. *Let m_n be a local averaging regression function estimate with subprobability weights $\{\alpha_{n,i}(x)\}$ that is strongly consistent for all distributions of (X,Y) such that Y is bounded with probability one. Assume that there is a constant c such that for all Y with $\mathbb{E}(Y^2) < \infty$,*

$$\limsup_{n\to\infty} \sum_{i=1}^{n} Y_i^2 \int \alpha_{n,i}(x)\mu(\mathrm{d}x) \leqslant c\mathbb{E}\left(Y^2\right) \quad \text{with probability one.}$$

Then m_n is strongly universally consistent.

Lemma 2.7. *Let $K_n\colon \mathbb{R}^d \times \mathbb{R}^d \to \{0,1\}$ be a measurable function. Assume that a constant $\varrho > 0$ exists with*

$$\int \frac{K_n(x,z)}{\int K_n(x,s)\mu(\mathrm{d}s)}\mu(\mathrm{d}x) \leqslant \varrho$$

for all n, all z, and all distributions μ. If $K_{n-1} \neq K_n$ at most for the indices $n = n_1, n_2, \ldots$, where $n_{k+1} \geqslant Dn_k$ for some fixed $D > 1$, then

$$\limsup_{n\to\infty} \int \sum_{i=1}^{n} \frac{Y_i K_n(x,X_i)}{1+\sum_{j\in\{1,\ldots,n\}\setminus\{i\}} K_n(x,X_j)}\mu(\mathrm{d}x) \leqslant 2\varrho\mathbb{E}(Y) \quad \mathbb{P}\text{-}a.s.,$$

for each integrable $Y \geqslant 0$.

The following theorem concerns the strong universal consistency of the kernel estimate with naive kernel and a special sequence of bandwidths.

Theorem 2.8. *Let $K(x) = \mathbb{1}_{\{\|x\|\leqslant 1\}}$ and let h_n satisfy*

$$h_{n-1} \neq h_n \quad \text{at most for the indices } n = n_1, n_2, \ldots,$$

where $n_{k+1} \geqslant Dn_k$ for fixed $D > 1$. Additionally let

$$h_n \to 0 \quad \text{and} \quad nh_n^d \to \infty,$$

e.g., $h_n = ce^{-\gamma\lfloor q\log n\rfloor/q}$ with $c > 0$, $0 < \gamma d < 1$ and $q > 0$. Then m_n is strongly universal consistent.

Proof. The assertion holds when there exists a $L > 0$ with $|Y| \leqslant L$ (cf. Theorem 2.5). According to Lemma 2.6, it suffices to show, that for some constant $c > 0$,

$$c\mathbb{E}\left(Y^2\right) \geqslant \limsup_{n\to\infty} \sum_{i=1}^{n} Y_i^2 \int \alpha_{ni}(x)\mu(\mathrm{d}x)$$

$$= \limsup_{n\to\infty} \sum_{i=1}^{n} Y_i^2 \int \frac{K_{h_n}(x - X_i)}{\sum_{i=1}^{n} K_{h_n}(x - X_i)} \mu(\mathrm{d}x) \quad \text{(by definition)}$$
$$= \limsup_{n\to\infty} \int \sum_{i=1}^{n} Y_i^2 \frac{K_{h_n}(x - X_i)}{\sum_{i=1}^{n} K_{h_n}(x - X_i)} \mu(\mathrm{d}x), \quad \text{(by linearity of integrals)}$$

$\mathbb{P}$-a.s., which is equivalent to

$$\begin{aligned}
&\limsup_{n\to\infty} \int \frac{\sum_{i=1}^{n} |Y_i| K_{h_n}(x - X_i)}{\sum_{i=1}^{n} K_{h_n}(x - X_i)} \mu(\mathrm{d}x) \\
&= \limsup_{n\to\infty} \int \sum_{i=1}^{n} \frac{|Y_i| K_n(x, X_i)}{\sum_{i=1}^{n} K_n(x, X_i)} \mu(\mathrm{d}x) \\
&= \limsup_{n\to\infty} \int \sum_{i=1}^{n} \frac{|Y_i| K_n(x, X_i)}{1 + \sum_{j\in\{1,\dots,n\}\setminus\{i\}} K_n(x, X_j)} \mu(\mathrm{d}x) \\
&\leqslant c\mathbb{E}(|Y|) \quad \mathbb{P}\text{-a.s.}
\end{aligned}$$

with $K_n(x, z) := K_{h_n}(x - z)$, for every distribution of (X, Y) with $\mathbb{E}(|Y|) < \infty$. This can be justified by Lemma 2.7: Let $z \in \mathbb{R}^d$ and $n \in \mathbb{N}$. Hence,

$$\begin{aligned}
&\int \frac{K_n(x, z)}{\int K_n(x, s)\mu(\mathrm{d}s)} \mu(\mathrm{d}x) \\
&= \int \frac{K_{h_n}(x - z)}{\int K_{h_n}(x - u)\mu(\mathrm{d}u))} \mu(\mathrm{d}x) \quad \text{(by definition)} \\
&\leqslant \varrho. \quad \text{(by Lemma 2.1)}
\end{aligned}$$

With help of the assumption for the sequence of bandwidths of Theorem 2.8 we can apply Lemma 2.7 and get

$$\limsup_{n\to\infty} \int \sum_{i=1}^{n} \frac{Y_i K_n(x, X_i)}{1 + \sum_{j\in\{1,\dots,n\}\setminus\{i\}} K_n(x, X_j)} \mu(\mathrm{d}x) \leqslant 2\varrho\mathbb{E}(Y) \quad \mathbb{P}\text{-a.s..}$$

for each integrable $Y \geqslant 0$, which completes the proof. ■

3. k-NN Estimates

Similar to Chapter 2, we first of all proof the strong consistency of the k-NN estimates under certain assumptions and later on the strong universal consistency.

Theorem 3.1 (strong consistency of the k-NN estimate)**.** *Let $|Y| \leqslant L$ $\mathbb{P}$-a.s. for some $L < \infty$ and that for each $x \in \mathbb{R}^d$ the random variable $\|X - x\|$ is absolutely continuous, i.e. its cumulative distribution function (CDF) is absolutely continuous. If $k_n \to \infty$ and $k_n/n \to 0$, then the k_n-NN regression function estimate is strongly consistent.*

Analogous to the kernel estimate above, the following organizational chart (Figure 3) is intended to illustrate the basic structure of the proof of Theorem 3.1.

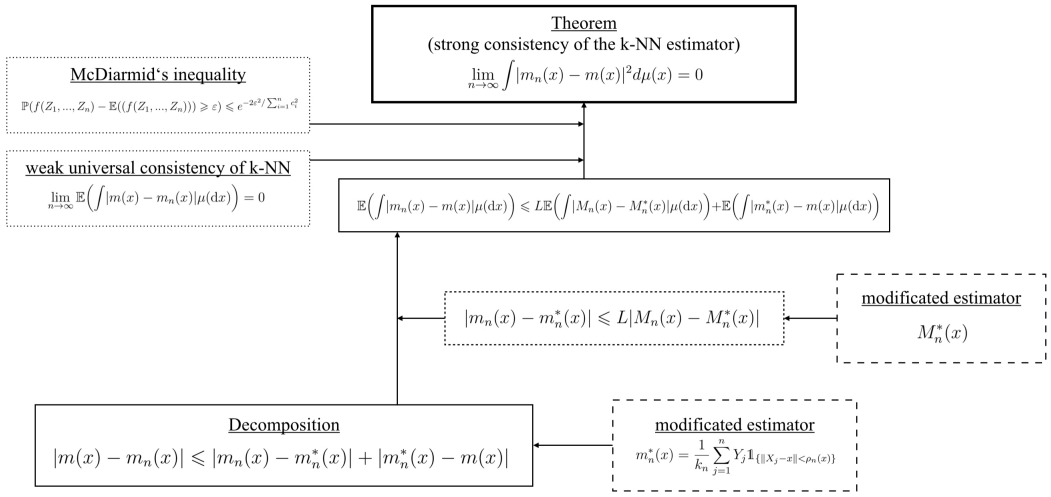

Figure 5: Organizational chart of the proof of theorem 3.1.

Proof. We first show that, for sufficiently large n

$$\mathbb{P}\left(\int |m(x) - m_n(x)|\mu(\mathrm{d}x) > \varepsilon\right) \leqslant 4e^{-n\varepsilon^2/(72L^2\gamma_d^2)}$$

where we defined γ_d as in Györfi's Lemma (see the Appendix). Due to $\|X - x\|$ being absolutely continuous. (note that this is the only part in

the proof where we will make use of this assumption), we know by Radon-Nikodým's Theorem A.4 and its corollary A.5, that $\|X - x\|$ has a probability density function (PDF) $f_{\|X-x\|}$. We now define $\rho_n(x)$ for every $x \in \mathbb{R}^d$ and $n \in \mathbb{N}$ as the solution to the equation

$$\frac{k_n}{n} = \mu(B_{\rho_n(x)}(x)) = P(\|X - x\| < \rho_n(x)) = \int_{-\infty}^{\rho_n(x)} f_{\|X-x\|}(t)\,\mathrm{d}t$$

and $\rho_n(x)$ is well-defined. Also define $m_n^*(x)$ by

$$m_n^*(x) = \frac{1}{k_n}\sum_{j=1}^{n} Y_j \mathbb{1}_{\{\|X_j - x\| < \rho_n(x)\}}.$$

For the proof we will rely on the following decomposition:

$$|m(x) - m_n(x)| \leqslant |m_n(x) - m_n^*(x)| + |m_n^*(x) - m(x)|.$$

By denoting $R_n(x) := \|X_{(k_n,n)}(x) - x\|$, we see that

$$\begin{aligned}
|m_n^*(x) - m_n(x)| &= \frac{1}{k_n}\left|\sum_{j=1}^{n} Y_j \mathbb{1}_{\{X_j \in B_{\rho_n(x)}(x)\}} - \sum_{j=1}^{n} Y_j \mathbb{1}_{\{X_j \in B_{R_n(x)}(x)\}}\right| \\
&\leqslant \frac{L}{k_n}\sum_{j=1}^{n}\left|\mathbb{1}_{\{X_j \in B_{\rho_n(x)}(x)\}} - \mathbb{1}_{\{X_j \in B_{R_n(x)}(x)\}}\right|. \quad (|Y| \leqslant L \text{ a.s.})
\end{aligned}$$

For fixed x we have either $\rho_n(x) \leqslant R_n(x)$ or $\rho_n(x) \geqslant R_n(x)$, thus either $B_{\rho_n(x)}(x) \subseteq B_{R_n(x)}(x)$ or $B_{\rho_n(x)}(x) \supseteq B_{R_n(x)}(x)$ and therefore $\mathbb{1}_{\{X_j \in B_{\rho_n(x)}(x)\}} - \mathbb{1}_{\{X_j \in B_{R_n(x)}(x)\}}$ have the same sign for each j. It follows that

$$\begin{aligned}
|m_n^*(x) - m_n(x)| &\leqslant \frac{L}{k_n}\sum_{j=1}^{n}\left|\mathbb{1}_{\{X_j \in B_{\rho_n(x)}(x)\}} - \mathbb{1}_{\{X_j \in B_{R_n(x)}(x)\}}\right| \\
&= L\left|\frac{1}{k_n}\sum_{j=1}^{n}(\mathbb{1}_{\{X_j \in B_{\rho_n(x)}(x)\}} - \mathbb{1}_{\{X_j \in B_{R_n(x)}(x)\}})\right| \\
&= L\left|\frac{1}{k_n}\sum_{j=1}^{n}\mathbb{1}_{\{X_j \in B_{\rho_n(x)}(x)\}} - \frac{k_n}{k_n}\right| \quad \text{(by definition of } R_n(x)) \\
&= L\left|\frac{1}{k_n}\sum_{j=1}^{n}\mathbb{1}_{\{X_j \in B_{\rho_n(x)}(x)\}} - 1\right| \\
&= L|M_n^*(x) - M(x)|
\end{aligned}$$

where M_n^* is defined as m_n^* with Y replaced by the constant random variable $Y \equiv 1$, and $M \equiv 1$ is the corresponding regression function. We note that

$$\begin{aligned}
\mathbb{E}(M_n^*(x)) &= \mathbb{E}\left(\frac{1}{k_n}\sum_{j=1}^{n} \mathbb{1}_{\{X_j \in B_{\rho_n(x)}(x)\}}\right) \\
&= \frac{1}{k_n}\sum_{j=1}^{n} \mathbb{P}\Big(X_j \in B_{\rho_n(x)}(x)\Big) \\
&= \frac{1}{k_n}\sum_{j=1}^{n} \mu(B_{\rho_n(x)}(x)) \\
&= \frac{1}{k_n}\sum_{j=1}^{n} \frac{k_n}{n} \quad \text{(by the definition of } \rho(x)) \\
&= 1 \\
&= M(x)
\end{aligned}$$

So far, we've got

$$|m(x) - m_n(x)| \leqslant L|M_n^*(x) - M(x)| + |m_n^*(x) - m(x)|. \tag{7}$$

First we show that the expected values of the integrals of both terms on the right-hand side converge to zero. Then we use McDiarmid's inequality to prove that both terms are very close to their expected values with large probability. For the expected value of the first term on the right-hand side of (7), we have

$$\begin{aligned}
& L\mathbb{E}\Big(\int |M_n^*(x) - M(x)|\mu(\mathrm{d}x)\Big) \\
&= L\int \mathbb{E}(|M_n^*(x) - M(x)|)\,\mu(\mathrm{d}x) \quad \text{(Fubini)} \\
&\leqslant L\int \sqrt{\mathbb{E}(|M_n^*(x) - M(x)|^2)}\mu(\mathrm{d}x) \quad \text{(Cauchy-Schwarz inequality)} \\
&= L\int \sqrt{\mathbb{E}(|M_n^*(x) - \mathbb{E}(M_n^*(x))|^2)}\mu(\mathrm{d}x) \quad \text{(equation from above)} \\
&= L\int \sqrt{\mathrm{Var}(M_n^*(x))}\mu(\mathrm{d}x) \\
&= L\int \sqrt{\frac{1}{k_n^2} n \mathrm{Var}\Big(\mathbb{1}_{\{X \in B_{\rho_n(x)}(x)\}}\Big)}\mu(\mathrm{d}x) \quad (X, X_1, \ldots, X_n \text{ iid.}) \\
&\leqslant L\int \sqrt{\frac{1}{k_n^2} n \mathbb{E}\Big(\mathbb{1}^2_{\{X \in B_{\rho_n(x)}(x)\}}\Big)}\mu(\mathrm{d}x) \\
&= L\int \sqrt{\frac{1}{k_n^2} n \mathbb{P}\Big(X \in B_{\rho_n(x)}(x)\Big)}\mu(\mathrm{d}x)
\end{aligned}$$

$$
\begin{aligned}
&= L \int \sqrt{\frac{1}{k_n^2} n \mu(B_{\rho_n(x)}(x))} \mu(\mathrm{d}x) \\
&= L \int \sqrt{\frac{n}{k_n^2} \frac{k_n}{n}} \mu(\mathrm{d}x) \quad \text{(definition of } \rho_n(x)) \\
&= \frac{L}{\sqrt{k_n}},
\end{aligned}
$$

which converges to 0 for $n \to \infty$. For the expected value of the second term on the right-hand side of (7), we use the weak universal consistency of the k-NN Estimator:

$$
\lim_{n\to\infty} \mathbb{E}\Big(\int |m(x) - m_n(x)| \mu(\mathrm{d}x) \Big) = 0.
$$

Therefore,

$$
\begin{aligned}
(8) \qquad & \mathbb{E}\Big(\int |m_n^*(x) - m(x)| \mu(\mathrm{d}x) \Big) \\
& \leqslant \mathbb{E}\Big(\int |m_n^*(x) - m_n(x)| \mu(\mathrm{d}x) \Big) + \mathbb{E}\Big(\int |m(x) - m_n(x)| \mu(\mathrm{d}x) \Big) \\
& \leqslant L\mathbb{E}\Big(\int |M_n^*(x) - M_n(x)| \mu(\mathrm{d}x) \Big) + \mathbb{E}\Big(\int |m(x) - m_n(x)| \mu(\mathrm{d}x) \Big) \\
& \longrightarrow 0.
\end{aligned}
$$

We notice that for random Variables A, B, C,

$$
(9) \qquad \mathbb{P}(A + B > 2C) \leqslant \mathbb{P}(A > C) + \mathbb{P}(B > C)\,.
$$

Assume now by (8) that n is so large that

$$
I_n + J_n := L\mathbb{E}\Big(\int |M_n^*(x) - M(x)| \mu(\mathrm{d}x) \Big) + \mathbb{E}\Big(\int |m_n^*(x) - m(x)| \mu(\mathrm{d}x) \Big) < \frac{\varepsilon}{3}.
$$

Then, with (7), we have for all $\varepsilon > 0$

$$
\begin{aligned}
(10) \qquad & \mathbb{P}\Big(\int |m(x) - m_n(x)| \mu(\mathrm{d}x) > \varepsilon \Big) \\
& = \mathbb{P}\Big(\int |m(x) - m_n(x)| \mu(\mathrm{d}x) - I_n - J_n > \varepsilon - I_n - J_n \Big) \\
& \leqslant \mathbb{P}\Big(\int |m(x) - m_n(x)| \mu(\mathrm{d}x) - I_n - J_n > 2\frac{\varepsilon}{3} \Big) \quad \text{(by (6))} \\
& \leqslant \mathbb{P}\Big(\int L|M_n^*(x) - M(x)| \mu(\mathrm{d}x) + \int |m_n^*(x) - m(x)| \mu(\mathrm{d}x) - I_n - J_n > 2\frac{\varepsilon}{3} \Big) \quad \text{(by (6))}
\end{aligned}
$$

$$
\begin{aligned}
&\leqslant \mathbb{P}\bigg(\int |m_n^*(x) - m(x)|\mu(\mathrm{d}x) - I_n > \frac{\varepsilon}{3}\bigg) \\
&+ \mathbb{P}\bigg(L\int |M_n^*(x) - M(x)|\mu(\mathrm{d}x) - J_n > \frac{\varepsilon}{3}\bigg) \quad \text{(by (9))} \\
&= \mathbb{P}\bigg(\int |m_n^*(x) - m(x)|\mu(\mathrm{d}x) - \mathbb{E}\bigg(\int |m_n^*(x) - m(x)|\mu(\mathrm{d}x)\bigg) > \frac{\varepsilon}{3}\bigg) \\
&+ \mathbb{P}\bigg(L\int |M_n^*(x) - M(x)|\mu(\mathrm{d}x) - \mathbb{E}\bigg(L\int |M_n^*(x) - M(x)|\mu(\mathrm{d}x)\bigg) > \frac{\varepsilon}{3}\bigg).
\end{aligned}
$$

Next we get an exponential bound for the first probability on the right-hand side of (7) by McDiarmid's inequality. We fix an arbitrary realization of the data $D_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, and replace (x_i, y_i) by $(\hat{x}_i, \hat{y}_i)$, changing the value of $m_n^*(x)$ to $m_{ni}^*(x)$. Then by the inverse triangle inequality

$$
\begin{aligned}
&\left|\int |m_n^*(x) - m(x)|\mu(\mathrm{d}x) - \int |m_{ni}^*(x) - m(x)|\mu(\mathrm{d}x)\right| \\
&\leqslant \int |m_n^*(x) - m_{ni}^*(x)|\mu(\mathrm{d}x) \\
&= \int \left|\frac{1}{k_n}\left(\sum_{j=1}^n y_j \mathbb{1}_{\{\|x_j - x\| < \rho_n(x)\}} - \hat{y}_i \mathbb{1}_{\{\|\hat{x}_i - x\| < \rho_n(x)\}} - \sum_{j\neq 1}^n y_j \mathbb{1}_{\{\|x_j - x\| < \rho_n(x)\}}\right)\right| \mu(\mathrm{d}x) \\
&= \frac{1}{k_n}\int |y_i \mathbb{1}_{\{\|x_i - x\| < \rho_n(x)\}} - \hat{y}_i \mathbb{1}_{\{\|\hat{x}_i - x\| < \rho_n(x)\}}|\mu(\mathrm{d}x)
\end{aligned}
$$

But $|y_i \mathbb{1}_{\{\|x_i - x\| < \rho_n(x)\}} - \hat{y}_i \mathbb{1}_{\{\|\hat{x}_i - x\| < \rho_n(x)\}}|$ is bounded by $2L$ and can differ from zero only if $\|x - x_i\| < \rho_n(x)$ or $\|x - \hat{x}_i\| < \rho_n(x)$. Observe that $\|x - x_i\| < \rho_n(x)$ or $\|x - \hat{x}_i\| < \rho_n(x)$ if and only if $\mu(B_{\|x - x_i\|}(x)) < k_n/n$ due to the definition of $\rho_n(x)$. But the measure of such x's is bounded by $2 \cdot \gamma_d k_n/n$ by Györfi's Lemma. Therefore,

$$
\sup_{x_1, y_1, \ldots, x_n, y_n, \hat{x}_i, \hat{y}_i} \int |m_n^*(x) - m_{ni}^*(x)|\mu(\mathrm{d}x) \leqslant \frac{2L}{k_n}\frac{2 \cdot \gamma_d k_n}{n} = \frac{4L\gamma_d}{n}
$$

and by McDiarmid's inequality A.2

$$
\begin{aligned}
&\mathbb{P}\bigg(\int |m_n^*(x) - m(x)|\mu(\mathrm{d}x) - \mathbb{E}\bigg(\int |m_n^*(x) - m(x)|\mu(\mathrm{d}x)\bigg) > \frac{\varepsilon}{3}\bigg) \\
&\leqslant 2e^{-n\varepsilon^2/(72L^2\gamma_d^2)}.
\end{aligned}
$$

For the second term on the right-hand side of (7), we can do it in exactly the same way, yielding

$$
\mathbb{P}\bigg(L\int |M_n^*(x) - M(x)|\mu(\mathrm{d}x) - \mathbb{E}\bigg(L\int |M_n^*(x) - M(x)|\mu(\mathrm{d}x)\bigg) > \frac{\varepsilon}{3}\bigg)
$$

$\leqslant 2e^{-n\varepsilon^2/(72L^2\gamma_d^2)}$.

Putting that into (7) gives us

$$\mathbb{P}\Big(\int |m(x) - m_n(x)|\mu(\mathrm{d}x) > \varepsilon\Big) \leqslant 4e^{-n\varepsilon^2/(72L^2\gamma_d^2)}.$$

Thus by the Borel-Cantelli Lemma, we get for every $\varepsilon > 0$ that

$$\limsup_{n\to\infty} \int |m(x) - m_n(x)|\mu(\mathrm{d}x) < \varepsilon \text{ almost surely.}$$

Due to the k-NN estimate having probability weights, we get

$$|m(x) - m_n(x)| \leqslant |m(x)| + |m_n(x)| \leqslant 2L$$

almost surely. Therefore finally

$$\lim_{n\to\infty} \int |m(x) - m_n(x)|^2\mu(\mathrm{d}x) \leqslant \lim_{n\to\infty} 2L \int |m(x) - m_n(x)|\mu(\mathrm{d}x) = 0$$

almost surely because ε was chosen arbitrarily small.

■

Before we come to the formulation and proof of the theorem for strong universal consistency of the k-NN estimate, we first of all need to show two lemmata. We now define some useful sets that we will continue to use in the course of our proof.

Definition 3.1. Let A_i be the collection of all $x \in \mathbb{R}^d$ s.t. X_i is one of its k_n nearest neighbors of x in $\{X_1, \ldots, X_n\}$.

Let us define cones $x + C_1, \ldots, x + C_{\gamma_d}$ as we did in Lemma 6.2. where

$$\bigcup_{j=1}^{\gamma_d} C_j = \mathbb{R}^d.$$

Then obviously

$$\bigcup_{j=1}^{\gamma_d} \{x + C_j\} = \mathbb{R}^d \tag{11}$$

regardless of how we chose x. By the cone property, if $u, u' \in x + C_j$ and $\|x - u\| < \|x - u'\|$, then $\|u - u'\| < \|x - u'\|$. Define

$$C_{i,j} = X_i + C_j \quad 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant \gamma_d.$$

Definition 3.2. Let $B_{i,j}$ be the subset of $C_{i,j}$ consisting of all $x \in C_{i,j}$ that are among the k_n nearest neighbors of X_i in the set

$$\{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n, x\} \cap C_{i,j}.$$

(If $C_{i,j}$ contains fewer than $k_n - 1$ of the X_l points $i \neq l$, then $B_{i,j} = C_{i,j}$.) Equivalently $B_{i,j}$ is the subset of $C_{i,j}$ consisting of all x that are closer to X_i than the k_n-th nearest neighbor of X_i in $\{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\} \cap C_{i,j}$.

Lemma 3.2 (k-NN covering)**.** *Let $1 \leqslant i \leqslant n$. If $x \in A_i$, then $x \in \bigcup_{j=1}^{\gamma_d} B_{i,j}$, and thus*

$$\mu(A_i) \leqslant \sum_{j=1}^{\gamma_d} \mu(B_{i,j}).$$

where A_i and $B_{i,j}$ are defined as in Definition 3.1 and 3.2 respectively.

Proof. Take $x \in A_i$. Then locate a j for which $x \in C_{i,j}$ (this always exists because of (11)). We now have to show, that $x \in B_{i,j}$ and the claim is proven. Equivalently, x has to be one of the k_n nearest neighbors of X_i in the set

$$\{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n, x\} \cap C_{i,j}.$$

Take $X_l \in C_{i,j}$. If $\|X_l - X_i\| < \|x - X_i\|$, then by the cone property we get that $\|x - X_l\| < \|x - X_i\|$, and thus X_l is one of the $k_n - 1$ nearest neighbors of x in $\{X_1, \ldots, X_n\}$ because of $x \in A_i$ (by definition of A_i, X_l being the k_n-th nearest neighbor of x in $\{X_1, \ldots, X_n\}$ would already imply $l = i$). This shows that in $C_{i,j}$ there are most $k_n - 1$ points X_l closer to X_i than x. Thus x is one of the k_n nearest neighbors of X_i in the set

$$\{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n, x\} \cap C_{i,j}.$$

This concludes the proof of the claim. ■

Lemma 3.3 (k-NN upper bound for $\mu(B_{i,j})$ as $n \to \infty$)**.** *Assume that for each x the random variable $\|X - x\|$ is absolutely continuous. If $k_n / \log(n) \to \infty$ and $k_n/n \to 0$ then, for every $j \in \{1, \ldots, \gamma_d\}$,*

$$\limsup_{n\to\infty} \frac{n}{k_n} \max_{1 \leqslant i \leqslant n} \mu(B_{i,j}) \leqslant 2 \ a.s.,$$

where $B_{i,j}$ are defined as in Definition 3.2.

Proof. If we utilize the Borel-Cantelli Lemma, we need to show for every j, that

$$\sum_{n=1}^{\infty} \mathbb{P}\Big(\frac{n}{k_n} \max_{1\leqslant i\leqslant n} \mu(B_{i,j}) > 2\Big) < \infty$$

In order to do this we give a bound for

$$\mathbb{P}(\mu(B_{i,j}) > p \mid X_i)$$

for arbitrary $0 < p < 1$. If $\mu(C_{i,j}) \leqslant p$ then since $B_{i,j} \subseteq C_{i,j}$, we have $\mathbb{P}(\mu(B_{i,j}) > p \mid X_i) = 0$, therefore we can assume that $\mu(C_{i,j}) > p$. Fix X_i. Define

$$G_{i,p} := C_{i,j} \cap B_{R_n(X_i)}(X_i),$$

where $R_n(X_i) > 0$ is chosen such that $\mu(G_{i,p}) = p$ (Its existence is assured similar to the proof of Theorem 3.1). We notice that either $B_{i,j} \supseteq G_{i,p}$ or $B_{i,j} \subseteq G_{i,p}$, therefore we have the following relationship:

$$\begin{aligned}
&\mathbb{P}(\mu(B_{i,j}) > p \mid X_i) \\
&= \mathbb{P}(\mu(B_{i,j}) > \mu(G_{i,p}) \mid X_i) \\
&= \mathbb{P}(B_{i,j} \supset G_{i,p} \mid X_i) \\
&= \mathbb{P}(G_{i,p} \text{ captures } < k_n \text{ of the points } X_l \in C_{i,j}, l \neq i \mid X_i) \\
&= \mathbb{P}\left(\sum_{\substack{l=1 \\ l\neq i}}^{n} \mathbb{1}_{\{X_l \in G_{i,p}\}} < k_n \,\middle|\, X_i\right)
\end{aligned}$$

The number of points X_l $(l \neq i)$ captured by $G_{i,p}$ given X_i is binomially distributed with parameters $(n-1, p)$, so by Lemma A.1, with $p = 2k_n/(n-1)$ and $\varepsilon = p/2 = k_n/(n-1)$ (which we can both WLOG assume to be smaller than 1 due to $k_n/n \to 0$), we have that

$$\begin{aligned}
&\mathbb{P}\Big(\max_{1\leqslant i\leqslant n} \mu(B_{i,j}) > p\Big) \\
&= \mathbb{P}(\cup_{i=1}^{n} \{\mu(B_{i,j}) > p\}) \\
&\leqslant n\mathbb{P}(\mu(B_{1,j}) > p) \\
&= n\mathbb{E}(\mathbb{P}(\mu(B_{1,j}) > p \mid X_1)) \\
&= n\mathbb{E}(\mathbb{P}(G_{1,p} \text{ captures } < k_n \text{ of the points } X_l \in C_{1,j}, l \neq 1 \mid X_1)) \\
&= n\mathbb{E}\left(\mathbb{P}\left(\sum_{\substack{l=1 \\ l\neq 1}}^{n} \mathbb{1}_{\{X_l \in G_{1,p}\}} < k_n \,\middle|\, X_1\right)\right)
\end{aligned}$$

$$
\begin{aligned}
&= n\mathbb{E}\left(\mathbb{P}\left(\sum_{\substack{l=1\\ l\neq 1}}^{n} \mathbb{1}_{\{X_l\in G_{1,p}\}} < (n-1)\frac{k_n}{n-1}\,\middle|\, X_1\right)\right)\\
&= n\mathbb{E}\left(\mathbb{P}\left(\sum_{\substack{l=1\\ l\neq 1}}^{n} \mathbb{1}_{\{X_l\in G_{1,p}\}} < (n-1)\varepsilon\,\middle|\, X_1\right)\right)\\
&\leqslant ne^{-(n-1)[p-\varepsilon+\varepsilon\log(\varepsilon/p)]} \quad \text{(by Chernoff's inequality A.6)}\\
&= ne^{-2k_n+k_n+k_n\log 2}\\
&\leqslant ne^{-k_n(1-\log 2)},
\end{aligned}
$$

which is summable because $k_n/\log n \to \infty$ implies there exists $N \in \mathbb{N}$ s.t. for all $n \geqslant N$, we have $k_n \geqslant \frac{3}{1-\log 2}\log n$, and thus

$$
\begin{aligned}
&\sum_{n=1}^{\infty} ne^{-k_n(1-\log 2)}\\
&= \sum_{n=1}^{N} ne^{-k_n(1-\log 2)} + \sum_{n=N+1}^{\infty} ne^{-k_n(1-\log 2)}\\
&\leqslant \sum_{n=1}^{N} ne^{-k_n(1-\log 2)} + \sum_{n=N+1}^{\infty} ne^{-\frac{3}{1-\log 2}\log(n)(1-\log 2)}\\
&= \sum_{n=1}^{N} ne^{-k_n(1-\log 2)} + \sum_{n=N+1}^{\infty} n\frac{1}{n^3}\\
&= \underbrace{\sum_{n=1}^{N} ne^{-k_n(1-\log 2)}}_{<\infty} + \underbrace{\sum_{n=N+1}^{\infty} \frac{1}{n^2}}_{<\infty} < \infty
\end{aligned}
$$

■

Theorem 3.4. *Assume that for each $x \in \mathbb{R}^d$ the random variable $\|X-x\|$ is absolutely continuous. If $k_n/\log n \to \infty$ and $k_n/n \to 0$ then the k_n-NN regression function estimate is strongly universally consistent, i.e. strongly consistent for all distributions (X,Y) with $\mathbb{E}(Y^2) < \infty$.*

Proof. By Lemma 2.6 and Theorem 3.1 it is enough to prove that there is a constant $c > 0$,

$$
\limsup_{n\to\infty} \sum_{i=1}^{n} \int W_{ni}(x)\mu(\mathrm{d}x)Y_i^2 \leqslant c\mathbb{E}\left(Y^2\right) \quad \text{a.s.}
$$

Observe that

$$\sum_{i=1}^{n} \int W_{ni}(x)\mu(\mathrm{d}x)Y_i^2 = \frac{1}{k_n}\sum_{i=1}^{n} Y_i^2\mu(A_i) \leqslant \left(\frac{n}{k_n}\max_i \mu(A_i)\right)\frac{1}{n}\sum_{i=1}^{n} Y_i^2.$$

If we can show, that

$$\limsup_{n\to\infty} \frac{n}{k_n}\max_i \mu(A_i) \leqslant c \quad \text{a.s.} \tag{12}$$

for some constant $c > 0$, then by the law of large numbers

$$\limsup_{n\to\infty} \left(\frac{n}{k_n}\max_i \mu(A_i)\right)\frac{1}{n}\sum_{i=1}^{n} Y_i^2 \leqslant \limsup_{n\to\infty} c\frac{1}{n}\sum_{i=1}^{n} Y_i^2 = c\mathbb{E}\left(Y^2\right) \quad \text{a.s.}$$

so we need to show (12). By the k-NN covering-Lemma 3.2, we have

$$\mu(A_i) \leqslant \sum_{j=1}^{\gamma_d} \mu(B_{i,j}) \quad \forall i \in \{1,\ldots,n\}$$

and thus

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{n}{k_n}\max_i \mu(A_i) \\
&\leqslant \limsup_{n\to\infty} \frac{n}{k_n}\max_i \sum_{j=1}^{\gamma_d} \mu(B_{i,j}) \\
&\leqslant \limsup_{n\to\infty} \frac{n}{k_n}\sum_{j=1}^{\gamma_d} \max_i \mu(B_{i,j}) \\
&\leqslant \sum_{j=1}^{\gamma_d} \limsup_{n\to\infty} \frac{n}{k_n}\max_i \mu(B_{i,j}) \\
&\leqslant \sum_{j=1}^{\gamma_d} 2 \quad \text{(by Lemma 3.3)} \\
&= 2\gamma_d
\end{aligned}$$

almost surely.

■

4. Comparison

Now, we have been able to fill in the remaining gaps concerning the weak and strong universal consistency of the three estimators. In order to conclude this topic, the two following charts (Figure 4, Figure 4) will provide an overview of the prerequisites required to achieve the different types of consistency. Starting for the kernel estimate:

KERNEL ESTIMATE		
boxed	regular	naive
$h_n \to 0$ $nh_n^d \to \infty$		
–	$\lvert Y\rvert \leqslant L$ in $\mathbb{P}$	$h_n \neq h_{n+1}$
$\Downarrow$	$\Downarrow$	$\Downarrow$
weak universal consistency	**strong** consistency	**strong universal** consistency

Figure 6: Consistencies of kernel estimate (overview).

Analogous for the k-NN estimate:

K-NN ESTIMATE		
$k_n \to \infty$		$k_n/log(n) \to \infty$
$k_n/n \to 0$		
$\mathbb{P}(ties) = 0$		
–	$\lvert Y\rvert \leqslant L$ a.s.	–
–	$\lVert X - x\rVert$ abs. continuous	
$\Downarrow$	$\Downarrow$	$\Downarrow$
weak universal consistency	**strong** consistency	**strong universal** consistency

Figure 7: Consistencies of k-NN estimate (overview).

A. Appendix

Theorem A.1 (Banach-Steinhaus). *Let $K_n(x,z)$ be functions on $\mathbb{R}^d \times \mathbb{R}^d$ satisfying the following conditions:*

(i) There is a constant $c > 0$ such that, for all n,

$$\int |K_n(x,z)|\mu(\mathrm{d}x) \leqslant c$$

for μ-almost all z.

(ii) There is a constant $D \geqslant 1$ such that, for all n,

$$\int |K_n(x,z)|\mu(\mathrm{d}z) \leqslant D$$

for all x.

(iii) For all $a > 0$,

$$\lim_{n\to\infty} \int\int |K_n(x,z)| \mathbb{1}_{\{\|x-z\|>a\}} \mu(\mathrm{d}z)\mu(\mathrm{d}x) = 0.$$

(iv)

$$\lim_{n\to\infty} ess\sup_x \left| \int K_n(x,z)\mu(\mathrm{d}z) - 1 \right| = 0.$$

Then, for all $m \in L_1(\mu)$,

$$\lim_{n\to\infty} \int \left| m(x) - \int K_n(x,z)m(z)\mu(\mathrm{d}z) \right| \mu(\mathrm{d}x) = 0.$$

Theorem A.2 (McDiarmid). *Let $Z_1, ..., Z_n$ be independent random variables taking values in a set A and assume that $f : A^n \to \mathbb{R}$ satisfies*

$$\sup_{z_1,...,z_n,\hat{z}_i \in A} |f(z_1, ..., z_n) - f(z_1, ..., z_{i-1}, \hat{z}_i, z_{i-1}, ..., z_n)| \leqslant c_i, \quad 1 \leqslant i \leqslant n.$$

Then, for all $\varepsilon > 0$,

$$\mathbb{P}(f(Z_1, ..., Z_n) - \mathbb{E}(f(Z_1, ..., Z_n)) \geqslant \varepsilon) \leqslant e^{-2\varepsilon^2 / \sum_{i=1}^n c_i^2}$$

and

$$\mathbb{P}(\mathbb{E}(f(Z_1, ..., Z_n)) - f(Z_1, ..., Z_n) \geqslant \varepsilon) \leqslant e^{-2\varepsilon^2 / \sum_{i=1}^n c_i^2}.$$

Lemma A.3 (Györfi's Lemma). *(Called Lemma 6.2 in Györfi's book) Let*

$$\mathcal{B}_a(x') := \{x \in \mathbb{R}^d \mid \mu(B_{\|x-x'\|}(x)) < a\}.$$

Then, for all $x' \in \mathbb{R}^d$,

$$\mu(\mathcal{B}_a(x')) \leqslant \gamma_d a,$$

where γ_d depends on the dimension d only.

Theorem A.4 (Radon-Nikodým). *Let μ and ν be measures on $(\Omega, \mathcal{F})$ and let μ be σ-finite, the following are equivalent*

a) ν is continuous with respect to μ.

b) There exists a density function $f \geqslant 0$, s.t. $\nu = f\mu$.

Corollary A.5. *Let X be an absolutely continuous random variable. Then X has a PDF.*

Proof. A measure μ is absolutely continuous if and only if its measure generating function F, i.e. $\mu((a,b]) = F(b) - F(a)$, is absolutely continuous. In the case of X, its measure generating function is its CDF and thus $\mu := \mathbb{P}_X$ is absolutely continuous and by A.4, X has a PDF. ■

Lemma A.6 (Chernoff). *Let B be a binomial random variable with parameters n and p. Then, for $1 > \varepsilon > p > 0$,*

$$\mathbb{P}(B > n\varepsilon) \leqslant \exp\left(-n\left[\varepsilon \log\frac{\varepsilon}{p} + (1-\varepsilon)\log\frac{1-\varepsilon}{1-p}\right]\right) \leqslant \exp(-n[p - \varepsilon + \varepsilon\log(\varepsilon/p)])$$

and, for $0 < \varepsilon < p < 1$,

$$\mathbb{P}(B < n\varepsilon) \leqslant \exp\left(-n\left[\varepsilon \log\frac{\varepsilon}{p} + (1-\varepsilon)\log\frac{1-\varepsilon}{1-p}\right]\right) \leqslant \exp(-n[p - \varepsilon + \varepsilon\log(\varepsilon/p)]).$$

References

[1] Györfi, L.: *A Distribution-Free Theory of Nonparametric Regression.* Springer (2002).