

# Consistencies of the kernel density estimator

Dominik Wied\* & Rafael Weißbach

Institut für Wirtschafts- und Sozialstatistik, Universität Dortmund, Dortmund, Germany

## Abstract

Proofs for the consistency of the kernel density estimator have historically developed. Four important milestones are the pointwise consistency, the almost sure uniform convergence, the rate of convergence on a bounded interval and the rate of convergence on  $\mathbb{R}$ . The underlying concepts of total variation, oscillation modulus and generalized empirical process are explained.

Keywords: Kernel estimation; Pointwise consistency; Strong uniform consistency; Empirical process; Rate of convergence.

AMS classification: 60-02, 62-02.

## 1 Introduction

For more than 50 years, mathematical statistics has been dealing with the problem to estimate efficiently the probability density function of continuous random variables from a random sample. In 1956, Rosenblatt proposed a kernel-based estimator with the basic idea to look at the difference ratio of the distribution function. This idea is used until today; it was and is a topic of scientific research, see e.g. Härdle (1991), Hall and Marron (1995) or Wand and Jones (1995). Over the years, the principle of kernel-based estimation has carried over for example to regression estimation (e.g. Dette (2002)) or survival analysis (e.g. Marron et al. (1996)).

In this survey article, we compare different proofs for (different types of) consistency in the historical development. Four important articles have had great influence on this field

---

\*address for correspondence: Dominik Wied, Institut für Wirtschafts- und Sozialstatistik, Fakultät Statistik, Technische Universität Dortmund, 44221 Dortmund, Germany, email: dominik.wied@tu-dortmund.de, Fon: +49/231/7555735, Fax: +49/231/7555284.

of research. The first one is an article of Parzen from 1962 proving pointwise consistency for the first time. In 1965, Nadaraja showed almost sure uniform convergence; an important technique for this is the partial integration for Lebesgue-Stieltjes integrals. About two decades later (1982), Stute achieved results on convergence rates of the estimator depending on the sample size, the kernel and the true density. He used an empirical process approach. However, he restricted his results to bounded intervals. Einmahl and Mason in 2005 extended the rates to the whole  $\mathbb{R}$ . With (generalized) empirical processes and mathematical techniques of other fields (e.g. topology) they proved results about almost sure convergence in the situation when the bandwidth is not fixed but may vary in a small interval. In the articles of Parzen, Nadaraja and Stute the bandwidth depends on the sample size and is fixed as a function of the sample size. Using the variable bandwidth approach, Einmahl and Mason achieved convergence rates for the whole  $\mathbb{R}$  that are the same as Stute's for bounded intervals, even extended to multivariate densities.

## 2 Consistency

### 2.1 Weak consistency

We consider a probability space  $(\Omega, \mathfrak{A}, \mathbb{P})$  and i.i.d. random variables  $X_i : \Omega \rightarrow \mathbb{R}^d, 1 \leq i \leq n, d \geq 1$ , distributed as  $X$ , with Lebesgue-density  $f$  and distribution function  $F$ . Let  $d = 1$  in the subsections 2.1, 2.2 and 3.1 if not stated otherwise.

The empirical distribution function  $F_n : \mathbb{R} \times \Omega \rightarrow [0, 1]$  is defined by

$$F_n(x, \omega) := \frac{1}{n} \sum_{i=1}^n I_{(-\infty, x]}(X_i(\omega)).$$

Because of the strong law of large numbers,  $F_n$  converges almost surely, i.e. with probability 1 to  $F$ . With the theorem of Gliwenko and Cantelli the convergence is uniform. Following Parzen, a heuristic approach to estimate the density is the consideration of the difference ratio of  $F_n$  and to use it as an estimator  $f_n$  for the density and sufficient small  $h_n$ :

$$f_n(x, \omega) := \frac{F_n\left(x + \frac{h_n}{2}, \omega\right) - F_n\left(x - \frac{h_n}{2}, \omega\right)}{h_n}. \quad (1)$$

Here and in the whole paper,  $(h_n)_{n \in \mathbb{N}}$  is a zero sequence. In order to generalize the approach, let

$$K(y) := I_{[-\frac{1}{2}, \frac{1}{2}]}(y),$$

then the estimator in (1) can be written as a Lebesgue-Stieltjes-integral of  $\frac{1}{h_n}K\left(\frac{\cdot}{h_n}\right)$  with the measure generating function  $F_n$ :

$$f_n(x, \omega) = \frac{1}{h_n} \int_{-\infty}^{\infty} K\left(\frac{x-y}{h_n}\right) dF_n(y, \omega) = \frac{1}{nh_n} \sum_{j=1}^n K\left(\frac{x - X_j(\omega)}{h_n}\right). \quad (2)$$

The second equal sign follows from the definition of the Lebesgue-Stieltjes integral.

It is useful to consider other nonnegative kernel functions  $K$  in (2), especially continuous kernel functions. This would guarantee that  $f_n$  is nonnegative and continuous as a sum of nonnegative and continuous functions. In the present paper, we consider general kernel functions with assumptions that depend on the situation. In general, however, the choice of  $K$  is rather irrelevant for the estimations. (see Wand and Jones (1995), page 31)

Parzen looks at the expected value and variance of the estimator with fixed point  $x \in \mathbb{R}$ . Since  $f$  is a Lebesgue-density, we have  $\int_{\mathbb{R}} |f(x)| dx = 1 < \infty$ , i.e.  $f \in L^1(\mathbb{R})$ . Parzen makes the assumption that  $K$  is a real-valued, Borel-measurable function with  $\sup_{y \in \mathbb{R}} |K(y)| =: \|K\|_{\infty} < \infty$ , i.e.  $K \in L^{\infty}(\mathbb{R})$ . In addition,  $K \in L^1(\mathbb{R})$ . We assume  $\lim_{y \rightarrow \infty} |yK(y)| = \lim_{y \rightarrow \infty} |yK^2(y)| = 0$ .

With these assumptions, at every point of continuity  $x$  of  $f$

$$\lim_{n \rightarrow \infty} \mathbb{E}(f_n(x)) = f(x) \int_{-\infty}^{\infty} K(y) dy.$$

$\mathbb{E}(f_n(x))$  is the convolution of  $f$  and  $\frac{1}{h_n}K\left(\frac{\cdot}{h_n}\right)$ , i.e.

$$\begin{aligned} \mathbb{E}(f_n(x)) &= \mathbb{E}\left(\frac{1}{h_n}K\left(\frac{x-X}{h_n}\right)\right) \\ &= \frac{1}{h_n} \int_{-\infty}^{\infty} K\left(\frac{x-y}{h_n}\right) dF(y) \\ &= \frac{1}{h_n} \int_{-\infty}^{\infty} K\left(\frac{x-y}{h_n}\right) f(y) dy =: f * \frac{1}{h_n}K\left(\frac{\cdot}{h_n}\right)(x). \end{aligned}$$

In order to get asymptotic unbiasedness, the integral of the kernel function over  $y$  has to be 1. This assumption holds for the whole paper.

It holds  $K \in L^2(\mathbb{R})$ .  $K$  is bounded and so,  $K$  is in  $L^2(\mathbb{R})$  if and only if  $\frac{K}{\sup_{x \in \mathbb{R}} |K(x)|}$  in  $L^2(\mathbb{R})$ . Since the absolute values are smaller or equal to 1,  $K \in L^2(\mathbb{R})$  follows from  $K \in L^1(\mathbb{R})$ .

Based on this, Parzen calculates the asymptotic variance of the density estimator:

At every point of continuity  $x$  of  $f$ ,

$$\lim_{n \rightarrow \infty} nh_n \text{Var}(f_n(x)) = f(x) \int_{-\infty}^{\infty} K^2(y) dy < \infty. \quad (3)$$

The proofs for the asymptotic expected value and for the asymptotic variance are similar; we give a sketch of the proof for the asymptotic variance:

$$\begin{aligned} \text{Var}(f_n(x)) &= \frac{1}{n} \text{Var} \left( \frac{1}{h_n} K \left( \frac{x - X}{h_n} \right) \right) \\ &= \frac{1}{n} \mathbb{E} \left( \frac{1}{h_n^2} K^2 \left( \frac{x - X}{h_n} \right) \right) - \frac{1}{n} (\mathbb{E} f_n(x))^2 \end{aligned}$$

and hence,  $nh_n \text{Var}(f_n(x)) = \mathbb{E} \left( \frac{1}{h_n} K^2 \left( \frac{x - X}{h_n} \right) \right) - h_n (\mathbb{E} f_n(x))^2$ .

With some techniques from the integration theory such as the theorem of the dominated convergence, one can show that the second summand tends to 0 and that the first summand tends to the desired term.

We introduce the assumption

$$(A1) \quad \lim_{n \rightarrow \infty} nh_n = \infty.$$

**Theorem 2.1** (Weak consistency) *Let the assumption (A1) hold. Then at every point of continuity  $x$  of  $f$  the estimator  $f_n(x)$  is weakly consistent, i.e. for every  $\epsilon > 0$*

$$\lim_{n \rightarrow \infty} \mathbb{P}(|f_n(x) - f(x)| > \epsilon) = 0.$$

*Proof.* We consider the mean square error of  $f_n(x)$ ,  $MSE(f_n(x)) = \text{Var}(f_n(x)) + (\text{Bias}(f_n(x)))^2$ .  $f_n(x)$  is asymptotically unbiased and so,  $\lim_{n \rightarrow \infty} (\text{Bias}(f_n(x)))^2 = 0$ .  $\text{Var}(f_n(x))$  is surely nonnegative. Under the assumption  $\lim_{n \rightarrow \infty} \text{Var}(f_n(x)) > 0$  we would have  $\lim_{n \rightarrow \infty} nh_n \text{Var}(f_n(x)) = \infty$ . This is not possible because of (3), therefore  $\lim_{n \rightarrow \infty} \text{Var}(f_n(x)) = 0$ . So,  $f_n(x)$  is consistent in the quadratic mean and hence weakly consistent.  $\square$

The theorem shows that the sequence  $(h_n)_{n \in \mathbb{N}}$  may not become too large nor too small. For controlling the bias, it is necessary that the sequence is a zero sequence, for controlling the asymptotic variance, the sequence may not tend to 0 too fast.

Notice that we proved pointwise, but not uniform consistency in probability. Using Fourier analysis, Parzen showed uniform consistency in probability under the assumption  $\lim_{n \rightarrow \infty} nh_n^2 = \infty$ . We do not consider questions of measurability in detail, we just remark that  $f_n(x)$  is rightcontinuous if  $K$  is rightcontinuous. Consequently,  $\sup_{-\infty < x < \infty} |f_n(x) - f(x)| = \sup_{x \in \mathbb{Q}} |f_n(x) - f(x)|$  because  $\mathbb{Q}$  is dense in  $\mathbb{R}$ . Each  $|f_n(x) - f(x)|$  is measurable as a composition of measurable functions, then also the supremum of the (countable) series  $(|f_n(x) - f(x)|)_{x \in \mathbb{Q}}$  is measurable (see Amann and Escher (2001), page 73).

## 2.2 Strong consistency

As far as we know, in 1965 Nadaraja was the first author who formulated a theorem dealing with the almost sure uniform convergence of the kernel density estimator. This

was a progress to Parzen who did not deal with almost sure convergence. In this subsection we assume that the kernel is of bounded variation; we will see soon what this means.

**Definition 2.2** (Almost sure uniform convergence) *The kernel density estimator (2) converges almost sure uniformly to  $f$  on  $\mathbb{R}$  with bandwidth  $h_n$  if*

$$\lim_{n \rightarrow \infty} \sup_{-\infty < x < \infty} |f_n(x) - f(x)| =: \lim_{n \rightarrow \infty} \|f_n - f\|_\infty = 0$$

*holds with probability 1. Hence,*

$$\mathbb{P}(\omega \in \Omega \mid \lim_{n \rightarrow \infty} \|f_n - f\|_\infty \neq 0) = 0.$$

**Remark** The uniform convergence of the series of estimators  $f_n$  is similar to the uniform convergence of a deterministic series of functions. Almost sure convergence of a random variable implies convergence in probability.

With the theorem of Gliwenko and Cantelli,  $F_n$  always converges uniformly to  $F$ , for the uniform convergence of the kernel density estimator  $f_n$  to  $f$ , we need more assumptions, e.g. the concept of bounded variation. To understand this, we give a short introduction into the theory of Lebesgue-Stieltjes (LS) integrals. Nadaraja and Stute use this technique frequently.

Like the common Lebesgue-integral, the Lebesgue-Stieltjes integral for a nonnegative function  $f$  is defined as the limit of the integrals of a monotonous increasing series of step functions which converge against  $f$ . The integrals of the step functions are measure integrals. The connected measure  $\mu$  on  $(\mathbb{R}, \mathfrak{B})$  ( $\mathfrak{B}$  is the Borel  $\sigma$ -field on  $\mathbb{R}$ ) is defined with a generalized distribution function  $G$ , a real-valued monotonously increasing, right-continuous function on  $\mathbb{R}$ . It holds  $\mu((a, b]) = G(b) - G(a)$  for all  $-\infty \leq a \leq b < \infty$ , this measure can be extended uniquely on  $\mathfrak{B}$ .

The kernel density estimator  $f_n$  can be interpreted as a LS integral with  $F_n(x, \omega)$  as weighting function. For fixed  $\omega$ ,  $F_n(x, \omega)$  is a deterministic monotonous, rightcontinuous function, i.e. we integrate along the path of the stochastic process given by  $F_n(x, \omega)$ . Assuming that  $K$  is LS-integrable regarding  $F_n$ , the representation as a sum in (2) follows with the definition of the LS integral because  $F_n$  only jumps at the points  $X_i(\omega)$ , is constant otherwise and because the high of the jumps is always  $\frac{1}{n}$ .

The LS integral can also be suitably defined when  $G$  can be written as the difference of two monotonously increasing functions, i.e.  $G = G_1 - G_2$  and  $\int_a^b f dG = \int_a^b f dG_1 - \int_a^b f dG_2$ , if this integral exists. This fact is related to the total variation of a function.

**Definition 2.3** (Total variation) *The total variation of a real-valued function  $u$  on a closed interval  $[a, b]$ ,  $a, b \in \mathbb{R}$ , is defined as*

$$V_a^b(u) := \sup_{P \in \mathfrak{P}} \sum_{i=0}^{n_P-1} |u(y_{i+1}) - u(y_i)|,$$

where the supremum is taken over the set  $\mathfrak{P} = \{P = \{y_0, \dots, y_{n_P}\} : P \text{ is partition of } [a, b]\}$ . It is a parameter for the local oscillation behavior of a function. If  $V_a^b(u) < \infty$ ,  $u$  is of bounded variation. If  $u$  is defined on the whole  $\mathbb{R}$ , we define

$$V_{-\infty}^{\infty} := \sup_{a \leq b} V_a^b.$$

Then,  $u$  is of bounded variation if  $u$  is of bounded variation on each compact interval.

**Remark** If  $u$  is continuously differentiable, we have  $V_a^b(u) = \int_a^b |u'(y)| dy$ . Here we can see that e.g. the Gaussian kernel  $K(y) = \frac{1}{\sqrt{2\pi}} \exp(-y^2)$  (but also many other kernels) is of bounded variation.

The following lemma allows for the definition of the LS integral for non-monotonous integrators.

**Lemma 2.4** (Bounded variation and monotony) *A real-valued function  $u$  of bounded variation can be written as the difference of two monotonous functions  $u_1, u_2$ . If  $u$  is right-continuous, then also  $u_1$  and  $u_2$  are right-continuous.*

**Remark** The opposite direction is wrong in general, see e.g. the monotonous function  $u(y) = y^3$  on the whole  $\mathbb{R}$ .

Finally, we cite a lemma about partial integration which is used frequently in proofs, see Sirjaev (1984), page 204.

**Lemma 2.5** (Partial integration for LS integrals) *Let  $g_1, g_2$  be two real-valued functions on the closed interval  $[a, b]$ , where the bounds  $\pm\infty$  are possible. If  $g_1$  and  $g_2$  are generalized distribution functions or right-continuous functions of bounded variation, where the left-hand limit of  $g_2$  always exists,*

$$\int_a^b g_1(y) dg_2(y) = g_1(b)g_2(b) - g_1(a)g_2(a) - \int_a^b g_2(y-) dg_1(y).$$

with  $g(y-)$  the left-hand limit of  $g$  in  $y$ .

We require that the term  $\infty - \infty$  does not appear in this expression.

Now, we turn to the convergence of the kernel density estimator with the following assumptions:

(B1) The kernel  $K$  is a right-continuous function.

(B2)  $K$  is of bounded variation.

(B3)  $\lim_{|x| \rightarrow \infty} K(x) = 0$ .

(B4)  $f$  is a uniformly continuous density function.

(B5)  $\sum_{n=1}^{\infty} \exp(-\gamma n h_n^2) < \infty$  for every  $\gamma > 0$ .

**Theorem 2.6** (Almost sure uniform convergence with fixed bandwidth) *Under the assumptions (B1) - (B5) the kernel density estimator (2) uniformly converges almost sure on  $\mathbb{R}$ .*

**Remarks** A function  $f$  is uniformly continuous on  $\mathbb{R}$  if and only if

$$\forall \epsilon > 0 \exists \delta > 0 \forall x, y \in \mathbb{R} : |x - y| < \delta \Rightarrow |f(x) - f(y)| < \epsilon.$$

A sufficient condition for the convergence of the series is that  $\lim_{n \rightarrow \infty} n h_n^2 = \infty$ .

The idea for proving the almost sure uniform convergence is the consideration of the (suitably scaled) series of random variables  $(\|f_n(x) - f(x)\|_{\infty})_{n \in \mathbb{N}}$ . The first step is the appreciation with the triangle inequality for norms

$$\|f_n - f\|_{\infty} = \|f_n - \mathbb{E}(f_n) + \mathbb{E}(f_n) - f\|_{\infty} \leq \|f_n - \mathbb{E}(f_n)\|_{\infty} + \|\mathbb{E}(f_n) - f\|_{\infty}.$$

It is sufficient to prove that both summands in the last expression turn to 0 for  $n \rightarrow \infty$ . The first summand is stochastic, the second one deterministic. The convergence proof of the deterministic part (the bias) is much easier, but we need assumptions on  $f$  (uniformly continuous) which are not necessary for the convergence of the stochastic part. On the other hand, the last one needs assumptions on the kernel (bounded variation) which are not needed for the bias convergence.

For the convergence of the stochastic part, we need the Borel-Cantelli lemma. Assume that we want to show that for  $\mathbb{P}$ -almost every  $\omega \in \Omega$  the limit superior of the series of random variables  $(\|f_n(x) - \mathbb{E}(f_n(x))\|_{\infty})_{n \in \mathbb{N}}$  is bounded above by a number. Then one considers the series of sets describing the  $\omega \in \Omega$  for which the series is larger than this number and shows that the series of probabilities over them converges. With the Borel-Cantelli lemma one concludes that with probability 1 only finite many elements of the series are larger than this number. Hence, with probability 1 the limit superior of the series is finite.

We now prove Theorem 2.6:

*Convergence of the bias* At first,

$$\begin{aligned} A_n &:= \sup_{-\infty < x < \infty} |\mathbb{E}(f_n(x)) - f(x)| \\ &= \sup_{-\infty < x < \infty} \left| \frac{1}{h_n} \int_{-\infty}^{\infty} K\left(\frac{x-y}{h_n}\right) f(y) dy - f(x) \int_{-\infty}^{\infty} K(y) dy \right|. \end{aligned}$$

In the first integral, we substitute  $x - y$  with  $y$ , in the second one  $y$  with  $\frac{y}{h_n}$  and we take the absolute value in the integral. In addition, we separate the area of integration in  $|y| \leq \delta$  and  $|y| > \delta$  for arbitrary  $\delta > 0$ . Then

$$\begin{aligned} A_n &\leq \sup_{-\infty < x < \infty} \int_{|y| \leq \delta} |f(x-y) - f(x)| \frac{1}{h_n} K\left(\frac{y}{h_n}\right) dy + \\ &\quad \sup_{-\infty < x < \infty} \int_{|y| > \delta} |f(x-y) - f(x)| \frac{1}{h_n} K\left(\frac{y}{h_n}\right) dy. \end{aligned}$$

We pull  $\sup_{|y| \leq \delta} |f(x-y) - f(x)|$  respectively  $\sup_{|y| > \delta} |f(x-y) - f(x)|$  in front of the two integrals. We can appreciate the first integral with 1 because the integral value of  $K$  is 1. In the second one, we resubstitute  $\frac{y}{h_n}$  with  $y$  and use the inequality  $|f(x-y) - f(x)| \leq 2 \max_{-\infty < x < \infty} f(x) =: 2M$ . With the theorem of the maximum and minimum for continuous functions, the maximum exists because  $\lim_{|x| \rightarrow \infty} f(x) = 0$  is a uniformly continuous density function. Hence,

$$A_n \leq \sup_{-\infty < x < \infty} \sup_{|y| \leq \delta} |f(x-y) - f(x)| + 2M \int_{|y| > \frac{\delta}{h_n}} K(y) dy.$$

Now, let  $\epsilon > 0$  arbitrary. For sufficient small  $\delta$  the first summand becomes smaller than  $\epsilon$  because of the uniform continuity of  $f$ . For each of this fixed  $\delta$  the integral on the right becomes smaller than  $\epsilon$  for sufficient large  $n$  by change of integration and limit (theorem of the dominated convergence) because of  $h_n \rightarrow 0$ . Consequently,  $\lim_{n \rightarrow \infty} A_n < 2\epsilon$ . Because  $\epsilon$  was arbitrary, the proof is completed.

*Convergence of the stochastic part* Let

$$\begin{aligned} B_n &:= \sup_{-\infty < x < \infty} |f_n(x) - \mathbb{E}(f_n(x))| \\ &= \sup_{-\infty < x < \infty} \left| \frac{1}{h_n} \int_{-\infty}^{\infty} K\left(\frac{x-y}{h_n}\right) dF_n(y) - \frac{1}{h_n} \int_{-\infty}^{\infty} K\left(\frac{x-y}{h_n}\right) dF(y) \right|. \end{aligned}$$

We consider both integrals as improper LS integrals and apply partial integration. Formally, the integrand is the function  $K\left(\frac{x-y}{h_n}\right)$ . Because of the bijective linear map  $y \rightarrow \frac{x-y}{h_n}$



$K\left(\frac{x-}{h_n}\right)$  has the same total variation as  $K$  and is especially of bounded variation. The boundary terms are dropped because we have assumed  $\lim_{|x| \rightarrow \infty} K(x) = 0$ . Hence,

$$\begin{aligned} B_n &= \sup_{-\infty < x < \infty} \left| \frac{1}{h_n} \int_{-\infty}^{\infty} (F(y) - F_n(y-)) dK\left(\frac{x-y}{h_n}\right) \right| \\ &\leq \sup_{-\infty < x < \infty} |F_n(x-) - F(x)| \frac{1}{h_n} V_{-\infty}^{\infty} \\ &= \sup_{-\infty < x < \infty} |F_n(x) - F(x)| \frac{1}{h_n} V_{-\infty}^{\infty}. \end{aligned}$$

The last equation follows from the fact that we consider the supremum over all  $x \in \mathbb{R}$  and because of the rightcontinuity of  $F_n$ . We want to make the last but one appreciation plausible for continuously differentiable  $K$ . Then, the LS integral becomes a common Lebesgue-integral

$$\left| \int_{-\infty}^{\infty} (F_n(y) - F(y)) K'(y) dy \right|.$$

But then,

$$\begin{aligned} \left| \int_{-\infty}^{\infty} (F_n(y) - F(y)) K'(y) dy \right| &\leq \int_{-\infty}^{\infty} |(F_n(y) - F(y))| |K'(y)| dy \\ &\leq \|(F_n(y) - F(y))\|_{\infty} \int_{-\infty}^{\infty} |K'(y)| dy \\ &= \|(F_n(y) - F(y))\|_{\infty} V_{-\infty}^{\infty}. \end{aligned}$$

So, we have reduced the appreciation of the stochastic part to an appreciation of the maximum distance between the empirical and the theoretical distribution function. Now define

$$\begin{aligned} D_n &:= \sup_{-\infty < x < \infty} |F_n(x) - F(x)|, \\ D_n^+ &:= \sup_{-\infty < x < \infty} (F_n(x) - F(x)), D_n^- := \sup_{-\infty < x < \infty} (F(x) - F_n(x)). \end{aligned}$$

If  $D_n$  is larger than an arbitrary number, either  $D_n^+$  or  $D_n^-$  is larger than this number. In addition,  $D_n^+$  or  $D_n^-$  are distributed equally for symmetric reasons, see Büning and Trenkler (1994), page 70. Hence it holds for arbitrary  $\lambda > 0$ :

$$\mathbb{P}\left(D_n > \frac{\lambda}{\sqrt{n}}\right) \leq \mathbb{P}\left(D_n^+ > \frac{\lambda}{\sqrt{n}}\right) + \mathbb{P}\left(D_n^- > \frac{\lambda}{\sqrt{n}}\right) = 2\left(1 - \mathbb{P}\left(D_n^+ \leq \frac{\lambda}{\sqrt{n}}\right)\right).$$

An inequality of Smirnov (see Nadaraja (1965), page 187) shows that

$$1 - \mathbb{P}\left(D_n^+ \leq \frac{\lambda}{\sqrt{n}}\right) \leq C \exp(-\alpha \lambda^2)$$

with a constant  $0 < C < \infty$  which is not defined more concrete and  $0 < \alpha \leq 2$ . Hence,

$$\mathbb{P}\left(D_n > \frac{\lambda}{\sqrt{n}}\right) \leq 2C \exp(-\alpha\lambda^2).$$

With the previous appreciation of  $B_n$  we finally achieve for arbitrary  $\epsilon > 0$

$$\mathbb{P}(B_n > \epsilon) \leq \mathbb{P}\left(D_n > \epsilon h_n \frac{1}{V_{-\infty}^\infty}\right) \leq 2C \exp(-\alpha\epsilon^2(V_{-\infty}^\infty)^{-2}nh_n^2) = 2C \exp(-\beta nh_n^2),$$

where  $\lambda = \frac{\sqrt{n}\epsilon h_n}{V_{-\infty}^\infty}$  and  $\beta := \alpha\epsilon^2(V_{-\infty}^\infty)^{-2}$  is a finite, deterministic number.

With the assumptions of the theorem, the series over  $\mathbb{P}(B_n > \epsilon)$  converges. Therefore, as described in the paragraph on the Borel-Cantelli lemma, the limit superior of the series of random variables  $(B_n)_{n \in \mathbb{N}}$  is smaller or equal to  $\epsilon$  with probability 1. Since  $\epsilon$  was arbitrary, the limit superior is 0. Since the series is nonnegative, also the limit inferior and therefore the limit has to be 0.  $\square$

**Remarks** The idea for proving the bias convergence is similar to the idea Parzen showed the pointwise asymptotic unbiasedness and the variance of the kernel density estimator with. For the bias convergence we do not need any special assumptions on the bandwidth.

The assumption  $\lim_{|x| \rightarrow \infty} K(x) = 0$  is necessary because it does not follow out of  $K \in L^1(\mathbb{R})$ . It is possible to construct a continuous and bounded function  $Z$  with  $\int_{\mathbb{R}} Z(y)dy = 1$  so that  $\limsup_{|z| \rightarrow \infty} K(z) = 1$ .  $Z$  is a jagged function which jags have height 1 and which for  $|z| \rightarrow \infty$  become so tiny that the area below is 1. We can think of jags with length  $(\frac{1}{2})^{|z|}$  which lie around the numbers  $z \in \mathbb{Z}, |z| \geq 1$ . Otherwise, the function equals to 0. The integral of the function is 1, i.e. the value of the geometric series  $\sum_{n \geq 1} (\frac{1}{2})^n$ .

Sufficient for the assumption  $\lim_{|x| \rightarrow \infty} K(x) = 0$  is the uniform continuity of  $K$ , see Einmahl and Mason (2005), page 1393.

### 3 Rates of convergence

Comparably to Nadaraja, several authors proved similar results about the almost sure uniform convergence. However, a general problem in many approaches was that the theorems did not consider the form of  $f$  and  $K$ . The authors were able to achieve results on the convergence itself but not on the rates of convergence. The reason lies in the fact that the examination of the behavior of  $f_n$  was reduced to the examination of  $F_n$  and  $F$  so that information was lost.

### 3.1 Rates of convergence on a bounded interval

In this case, Stute achieved an improvement in 1982 by describing exact rates using empirical processes. He could achieve results on a bounded interval  $J := (a, b)$ ,  $a < b$ , yet not on the whole  $\mathbb{R}$  as Nadaraja. Stute considers kernels with bounded support and focusses on an appreciation of the stochastic part.

**Definition 3.1** (Empirical process) *For a Borel-measurable function  $g : \mathbb{R}^d \rightarrow \mathbb{R}$ , the empirical process  $\alpha_n^*(g)$  is defined by*

$$\alpha_n^*(g) := \frac{1}{\sqrt{n}} \left( \sum_{i=1}^n g(X_i) - n\mathbb{E}g(X_1) \right).$$

In addition,  $\alpha_n(g) := \sqrt{n}\alpha_n^*(g)$ .

Stute only defines the empirical process for distribution functions on  $\mathbb{R}$ . Here, for  $x \in \mathbb{R}$  we have  $\beta_n(x) := \alpha_n^*(g_x)$  with  $g_x(\cdot) = I_{(-\infty, x]}(\cdot)$  respectively  $\beta_n(x) = \sqrt{n}(F_n(x) - F(x))$ . The first step is the consideration of the uniform empirical process, i.e.  $\gamma_n(x) = \sqrt{n}(\bar{F}_n(x) - x)$  for  $x \in [0, 1]$ , where  $\bar{F}_n$  is the empirical distribution function of random variables uniformly distributed on  $[0, 1]$ . For uniform distribution, it holds  $F(x) = x$ .

This makes sense because we point out convergence results on a bounded interval. There, we assume that  $f(x) > 0$  and hence that  $F$  is strictly monotonously increasing and invertible. Thus, with the central theorem of statistics (see Büning and Trenkler (1994), page 52),  $F(X_i)$  is uniformly distributed on  $[0, 1]$ . We can conclude that  $F_n(x) = \bar{F}_n(F(x))$  and that

$$\beta_n(x) = \sqrt{n}(F_n(x) - F(x)) = \sqrt{n}(\bar{F}_n(F(x)) - F(x)) = \gamma_n(F(x)).$$

Comparable with the total variation or with the modulus of continuity of a function, the oscillation modulus  $\omega_n$  of  $\gamma_n$  describes the local oscillation behaviour of the uniform empirical process.

**Definition 3.2** (Oscillation modulus) *The oscillation modulus  $\omega_n$  of the uniform empirical process  $\gamma_n$  is for some  $a \in \mathbb{R}$  defined by*

$$\omega_n(a) := \sup_{|t-s| \leq a} |\gamma_n(t) - \gamma_n(s)|.$$

Stute partly makes different assumptions on the bandwidth than Parzen or Nadaraja. Here, for  $n \rightarrow \infty$  it also holds  $nh_n \rightarrow \infty$  (assumption 1), but in addition  $\frac{-\log h_n}{nh_n} \rightarrow 0$  (assumption 2) and  $\frac{-\log h_n}{\log \log n} \rightarrow \infty$  (assumption 3). Assumption 2 says similarly to

assumption 1 that  $h_n$  does not converge to 0 too fast, assumption 3 permits that  $h_n$  becomes too large. The condition  $nh_n^2 \rightarrow \infty$  is not necessary.

An important result is a lemma about uniform convergence of the oscillation modulus and a corollary. This is an important step to appreciate the local behavior of  $F$  and thus the form of  $f$ .

**Lemma 3.3** (Convergence of the oscillation modulus) *Let  $J$  be a subinterval of  $[0, 1]$  and  $h_n$  bandwidth. Then, for arbitrary  $0 < c_1 \leq c_2 < \infty$  it holds  $\mathbb{P}$ -almost sure*

$$\lim_{n \rightarrow \infty} \sup_{c_1 h_n \leq t-u \leq c_2 h_n; t, u \in J} \frac{|\gamma_n(t) - \gamma_n(u)|}{\sqrt{2(t-u)(-\log h_n)}} = 1. \quad (4)$$

*In addition,  $\mathbb{P}$ -almost sure*

$$\lim_{n \rightarrow \infty} \frac{\omega_n(h_n)}{\sqrt{2h_n(-\log h_n)}} = 1. \quad (5)$$

**Remarks** The proof of (4) consists of two parts: At first, one shows that the limit superior of the series of random variables is smaller than or equal to 1, then one shows that the limit inferior is larger than or equal to 1. Since the limit superior is always larger than or equal to the limit inferior, the proof is completed. The proof of the limit inferior-part uses the technique of poissonization. The basic idea is that  $\sqrt{n}\gamma_n$  is a centered Poisson process under the condition that there are  $n$  observations at time 1. The Poisson probabilities appearing in the proof are then appreciated against the normal distribution.

The proof of the limit superior-part uses an exponential inequality for the probability that  $\frac{\omega_n(a)}{\sqrt{a}}$  becomes large. Both proofs use the Borel-Cantelli lemma.

Also in the proof of (5), the limit superior and the limit inferior are analyzed seperately. The result for the limit inferior follows from (4) by choosing  $c_1 = c_2 = 1$ , the result for the limit superior is proved similarly to the result for the limit superior used in (4).

From (4) it follows by substituting  $t := F(t)$  and  $u := F(u)$  with the mean value theorem for differentiation an analog result for the empirical process  $\beta_n$ . We introduce the assumptions

(C1)  $f$  is uniformly continuous on  $J = (a, b) \subset \mathbb{R}, a < b$ .

(C2)  $0 < m \leq f(x) \leq M < \infty$  for all  $x \in J$ .

**Lemma 3.4** (Convergence of the empirical process) *Let the assumptions (C1) and (C2) hold. Let  $\xi_{u,t}$  be an arbitrary point between  $u$  and  $t$ . Then it holds for arbitrary  $0 < c_1 \leq$*

$c_2 < \infty$   $\mathbb{P}$ -almost sure

$$\lim_{n \rightarrow \infty} \sup_{c_1 h_n \leq t-u \leq c_2 h_n; t, u \in J} \frac{|\beta_n(t) - \beta_n(u)|}{\sqrt{2(t-u)f(\xi_{u,t})(-\log h_n)}} = 1.$$

**Theorem 3.5** *With the assumptions (C1) and (C2), it holds  $\mathbb{P}$ -almost sure for the naive kernel*

$$\lim_{n \rightarrow \infty} \sup_{x \in J_\epsilon} \frac{\sqrt{nh_n}}{\sqrt{2(-\log h_n)}} \frac{|f_n(x) - \mathbb{E}(f_n(x))|}{\sqrt{f(x)}} = 1.$$

*Proof.* With Lemma 3.4 we directly achieve a result about the convergence rate of the stochastic part of the naive kernel density estimator with kernel  $K(y) = I_{[-\frac{1}{2}, \frac{1}{2}]}(y)$ . This is that simple because the naive estimator (in contrast to estimators with more complicated kernels) directly works with the empirical distribution function. For this, Lemma 3.4 gives a convergence result. At first, we choose  $c_1 = c_2 = 1$  and  $\xi_{u,t} = \frac{u+t}{2}$ . Then we get with the definitions of  $f_n$  and  $\mathbb{E}(f_n)$

$$\begin{aligned} 1 &= \lim_{n \rightarrow \infty} \sup_{t-u=h_n; t, u \in J} \frac{|\beta_n(t) - \beta_n(u)|}{\sqrt{2(t-u)f\left(\frac{u+t}{2}\right)(-\log h_n)}} \\ &= \lim_{n \rightarrow \infty} \sup_{t-u=h_n; t, u \in J} \frac{\sqrt{n}|F_n(t) - F(t) - F_n(u) + F(u)|}{\sqrt{2h_n f\left(u + \frac{h_n}{2}\right)(-\log h_n)}} \\ &= \lim_{n \rightarrow \infty} \sup_{u \in J, u+h_n \in J} \frac{\sqrt{n}}{\sqrt{2h_n(-\log h_n)}} \frac{|F_n(u+h_n) - F_n(u) - (F(u+h_n) - F(u))|}{\sqrt{f\left(u + \frac{h_n}{2}\right)}} \\ &= \lim_{n \rightarrow \infty} \sup_{x \in J_\epsilon} \frac{\sqrt{nh_n}}{\sqrt{2(-\log h_n)}} \frac{|f_n(x) - \mathbb{E}(f_n(x))|}{\sqrt{f(x)}}. \end{aligned}$$

$J_\epsilon$  is for an arbitrary  $\epsilon > 0$  defined by  $J_\epsilon = (a + \epsilon, b - \epsilon)$ . We need such a small  $\epsilon$  because we consider  $u + h_n$  in the third step of the calculation. For finite  $n$ ,  $u$  or  $u + h_n$  (because of  $h_n > 0$ ) do not have to lie in  $J$ , just because  $u + \frac{h_n}{2}$  lies in  $J$ . On the other hand,  $\epsilon$  may become arbitrarily small for  $\lim_{n \rightarrow \infty} h_n = 0$ .  $\square$

Surely, the naive kernel is not a suitable estimator for practice because  $f_n(x)$  is not continuous in general, even if the true density is continuous. Hence, Stute develops results on more general kernels basing on the results on the empirical process. He does this in several parts, at first he considers step functions like

$$K(y) := \sum_{i=1}^m a_i I_{[d_i, d_{i+1})}(y)$$

for finite  $m$ , nonnegative real numbers  $a_i$ ,  $1 \leq i \leq m$  and  $\infty < d_1 < d_2 < \dots < d_{m+1}$ . For this, he achieves a convergence result that is similar to the result on the naive kernel and is proved similarly:

**Corollary 3.6** *With the assumptions (C1) and (C2), it holds  $\mathbb{P}$ -almost sure for the kernel of step functions*

$$\lim_{n \rightarrow \infty} \sup_{x \in J_\epsilon} \frac{\sqrt{nh_n}}{\sqrt{2(-\log h_n)}} \frac{|f_n(x) - \mathbb{E}(f_n(x))|}{\sqrt{f(x)}} = \sqrt{\sum_{i=1}^m (K(d_i))^2 (d_{i+1} - d_i)}.$$

The next step is the extension on kernels with bounded variation; we make the following assumptions:

(C3)  $K(y) = 0$  outside of a bounded interval  $[r, s]$  with  $K(r) = 0$ .

(C4)  $F$  is Lipschitz-continuous on an interval  $J = (a, b)$ ,  $a < b$ , with Lipschitz-constant  $M < \infty$ , i.e.

$$M := \sup_{a, b \in J, a \neq b} \frac{|F(a) - F(b)|}{|a - b|}.$$

Assumption (C4) is less strong than the assumption of the uniform continuity of  $f$  because the uniform continuity of  $f$  implies the continuous differentiability of  $F$ . Stute's result is

**Lemma 3.7** *With the assumptions (B1), (C3) and (C4) for every  $\epsilon > 0$*

$$\limsup_{n \rightarrow \infty} \sqrt{\frac{nh_n}{-2 \log h_n}} \sup_{x \in J_\epsilon} |f_n(x) - \mathbb{E}(f_n(x))| = C,$$

where the constant  $C \leq \sqrt{M(s-r)} V_r^s(K)$  with the total variation  $V_r^s(K) < \infty$ .

*Proof.* We apply again the technique of the partial integration. At first, for a fixed  $x \in J_\epsilon$ ,  $y$  must lie in the interval  $(x - sh_n, x - rh_n]$  so that  $K\left(\frac{x-y}{h_n}\right)$  can be different from 0. In addition, for a fixed  $x \in J_\epsilon$  with the definition of the Lebesgue-Stieltjes integral:  $\int_{x-sh_n}^{x-rh_n} dK\left(\frac{x-y}{h_n}\right) = K(r) - K(s) = 0$  and thus

$$\int_{x-sh_n}^{x-rh_n} F_n(x - rh_n) - F(x - rh_n) dK\left(\frac{x-y}{h_n}\right) = 0.$$

We need this to be able to work with the  $\beta_n$  later on. Similarly to Nadaraja's proof it holds for a fixed  $x \in J_\epsilon$

$$\begin{aligned} f_n(x) - \mathbb{E}(f_n(x)) &= \frac{1}{h_n} \int_{\mathbb{R}} K\left(\frac{x-y}{h_n}\right) dF_n(y) - \frac{1}{h_n} \int_{\mathbb{R}} K\left(\frac{x-y}{h_n}\right) dF(y) \\ &= -\frac{1}{h_n} \int_{(x-sh_n, x-rh_n]} (F_n(y-) - F(y) - [F_n(x - rh_n) - F(x - rh_n)]) dK\left(\frac{x-y}{h_n}\right). \end{aligned}$$

Now,  $F_n(y-) \leq F_n(y)$  and the distance from  $y$  to  $x - rh_n$  in the interval  $(x - sh_n, x - rh_n]$  is smaller than or equal to  $(s - r)h_n$ . With the definition of the empirical process the integrand (multiplied by  $\sqrt{n}$ ) is smaller than or equal to

$$\sup_{|x-y| \leq (s-r)h_n, x, y \in J} |\beta_n(x) - \beta_n(y)| = \sup_{|x-y| \leq (s-r)h_n, x, y \in J} |\gamma_n(F(x)) - \gamma_n(F(y))|.$$

Like in Theorem 3.5 we need a small  $\epsilon > 0$  as a buffer for  $x$  so that  $x - sh_n$  and  $x - rh_n$  have the "possibility" to be in  $J$ .

On  $J$ ,  $|F(a) - F(b)| \leq M|a - b|$  and thus, the integrand is smaller than or equal to  $\omega_n(M(s - r)h_n)$ . Summed up, we have the appreciation

$$\sqrt{nh_n} \sup_{x \in J_\epsilon} |f_n(x) - \mathbb{E}(f_n(x))| \leq (h_n)^{-\frac{1}{2}} \omega_n(M(s - r)h_n) V_r^s(K).$$

With (5), it follows that the expression

$$\sqrt{\frac{nh_n}{2(-\log(M(s - r)h_n))}} \sup_{x \in J_\epsilon} |f_n(x) - \mathbb{E}(f_n(x))|$$

is  $\mathbb{P}$ -almost sure finite because of the properly scaled  $\omega_n(M(s - r)h_n)$  (even  $\leq V_r^s(K)$ ). The theorem follows because of the fact that  $-\log(M(s - r)h_n)$  and  $-\log(h_n)$  have the same limit behavior.  $\square$

**Remarks** The Borel-Cantelli lemma is used indirectly because it is used for proving the results on the oscillation modulus.

The finiteness of the Lipschitz-constant of  $F$  is necessary for the finiteness of the limit superior of the series. For this, it is sufficient that the density is bounded.

If  $F' = f$  is uniformly continuous on  $J$ , we can calculate  $C$  explicitly. The idea lies in separating the kernel into the sum of two functions, where one function is a step function. By combining Corollary 3.6 and Lemma 3.7 we achieve

**Corollary 3.8** *With the assumptions of (3.4) and (3.7) it holds  $\mathbb{P}$ -almost sure*

$$\lim_{n \rightarrow \infty} \sqrt{\frac{nh_n}{-2 \log h_n}} \sup_{x \in J_\epsilon} \frac{|f_n(x) - \mathbb{E}(f_n(x))|}{\sqrt{f(x)}} = \left( \int_r^s K^2(y) dy \right)^{\frac{1}{2}}.$$

## 3.2 Rates of convergence on $\mathbb{R}$

From the statistical point of view, it is in general useful if the bandwidth not just depends on  $n$  but also on the point  $x$  and especially from the data. We can see this in

the expression for  $MSE(f_n(x))$  from Stute (1982):

$$\begin{aligned} MSE(f_n(x)) &:= Var(f_n(x)) + (Bias(f_n(x)))^2 \\ &= \frac{f(x)}{nh_n} \int_{-\infty}^{\infty} K^2(y) dy + 2\pi h_n^4 (f''(x))^2 \left( \frac{1}{2\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 K^2(y) dy \right)^2. \end{aligned}$$

For larger  $f(x)$ ,  $h_n$  should become larger, too, to reduce the  $MSE$ . There are different approaches for a data-adaptive bandwidth, as Einmahl and Mason (2005) show. The two authors do not look at a special method, but get general results about the almost sure uniform convergence of the kernel density estimator if the bandwidth varies in a small interval  $[a_n, b_n]$ . In this chapter, we consider like the authors the  $\mathbb{R}^d$  with the Borel  $\sigma$ -field  $\mathfrak{B}^d$ .

Similarly to Definition 2.2, we say that the kernel density estimator

$$f_n(x) := \frac{1}{nh_n} \sum_{j=1}^n K\left(\frac{x - X_j}{h_n^{\frac{1}{d}}}\right)$$

converges on  $\mathbb{R}^d$  almost surely uniformly to  $f$  with variable bandwidth  $h_n \in [a_n, b_n]$  if

$$\lim_{n \rightarrow \infty} \sup_{a_n \leq h_n \leq b_n} \sup_{-\infty < x < \infty} |f_n(x) - f(x)| =: \lim_{n \rightarrow \infty} \sup_{a_n \leq h_n \leq b_n} \|f_n - f\|_{\infty} = 0$$

with probability 1.

There are two possibilities to realize such a data-adaptive bandwidth in practice. The first one is

$$\lim_{n \rightarrow \infty} \mathbb{P}(a_n \leq h_n \leq b_n) = 1,$$

i.e. the bandwidth lies at least asymptotically in the desired interval. This is not enough for almost sure uniform convergence and only gives us uniform convergence in probability. However, there have already been developed some techniques to get such a bandwidth, e.g. with plug-in-estimators analyzed in Deheuvels and Mason (2004).

It is also possible that (at least for large  $n$ )

$$\mathbb{P}(a_n \leq h_n \leq b_n) = 1$$

holds. Yet, this is probably difficult to realize in practice; just at the moment we do not know any application.

The general result about consistency for variable bandwidth is a great effort. We point out that the two authors simultaneously get a result about the (almost sure) rate of convergence for the case  $a_n = h_n = b_n$ , i.e.  $h_n$  is fixed. This is a special case. In contrast to Stute, this result does not only hold for a bounded interval  $(a, b)$ , but it is the same



rate.

Einmahl und Mason also proved similar results. They need more special mathematical techniques than Parzen, Nadaraja and Stute, e.g. from the field of topology and functional analysis.

In the following, we consider a general family of kernels, i.e. for a fixed kernel  $K$  the set

$$\mathfrak{K} := \left\{ K \left( \frac{x - \cdot}{h_n^{\frac{1}{d}}} \right) : h_n > 0, x \in \mathbb{R}^d \right\}. \quad (6)$$

The research of Einmahl und Mason was only possible because of former research of Talagrand who dealt with appreciations for the supremum of empirical processes. He chose a more general approach than Stute. The latter just defined the empirical process for the functions  $g_t(\cdot) = I_{(-\infty, t]}(\cdot)$ , so that the process is a description of the distance between the empirical and the theoretical distribution function. As in his article from 1994 shown, he considers a general function set  $\mathfrak{G}$  and considers the expression

$$\|\alpha_n\|_{\mathfrak{G}} := \sup_{g \in \mathfrak{G}} |\alpha_n(g)|.$$

It is of interest to find appreciations for the expression

$$r_{\mathfrak{G}}(M) := \mathbb{P}(\|\alpha_n\|_{\mathfrak{G}} \geq M\sqrt{n}).$$

This expression was described more precisely in different articles. Einmahl and Mason use the general results to get results about a subset of  $\mathfrak{K}$  (if the bandwidth  $h$  lies in a special interval).

The motivation in using Talagrand's approach for results about kernel density estimation lies in the fact that the kernel density estimator  $f_n(x)$  is the (divided by  $n$ ) sum of suitably scaled kernels, applied to the random variables. The expected value of the kernel applied on the random variables is  $\mathbb{E}(f_n(x))$ . Consequently, if we divide by  $n \|\alpha_n\|_{\mathfrak{G}}$  for a general set of kernels  $\mathfrak{G}$ , we get with this inequality an appreciation for the stochastic part of all kernels of this set. If for example the point  $x$  and the bandwidth varies in a set of kernels, we get an appreciation for the supremum of all points and all bandwidths. It is then possible to let the bandwidth vary, as Einmahl und Mason proof. Comparably to Nadaraja und Stute, Talagrand and Einmahl/Mason make bounding assumptions on the kernels. We consider a general set of functions  $\mathfrak{G}$  which only contains measurable functions and which is uniformly bounded, i.e.  $\exists G : \mathbb{R}^d \rightarrow \mathbb{R}$  with  $\infty > G(x) \geq \sup_{g \in \mathfrak{G}} |g(x)| \quad \forall x \in \mathbb{R}^d$ . In addition,  $\sup_{g \in \mathfrak{G}} \|g\|_{\infty} \leq M < \infty$  and  $\sigma_{\mathfrak{G}}^2 := \sup_{g \in \mathfrak{G}} \text{Var}(g(X)) < \infty$ , i.e., all variances are finite. Later,  $\mathfrak{G}$  is a subset of  $\mathfrak{K}$ .

Now we define a random variable symmetrizing the  $g(X_i)$ :

**Definition 3.9** (Rademacher-variables) *A Rademacher-variable is a discrete random variable  $\epsilon_i$  with  $P(\epsilon_i = -1) = P(\epsilon_i = 1) = \frac{1}{2}$ .*

We want to introduce a sequence  $\epsilon_1, \dots, \epsilon_n$  of independent Rademacher-variables which are independent from the  $X_1, \dots, X_n$  and look at the expression  $\sup_{g \in \mathfrak{G}} |\sum_{i=1}^n \epsilon_i g(X_i)|$ . It is similar to  $\|\alpha_n\|_{\mathfrak{G}}$ ; both processes are centered if we do not use the absolute value. However, the advantage of the symmetrization is that we can appreciate this process in a better way, as we seen soon.

It is important to assume the measurability of the expression  $\sup_{g \in \mathfrak{G}} |\sum_{i=1}^n \epsilon_i g(X_i)|$  so that we can calculate the expected value from it. The pointwise measurability of  $\mathfrak{G}$  is sufficient for this. Pointwise measurability means that we can find a countable subset  $\mathfrak{G}_0$  of  $\mathfrak{G}$  so that for each function  $g \in \mathfrak{G}$  there exists a sequence of functions  $g_n$  of  $\mathfrak{G}_0$  with

$$\lim_{n \rightarrow \infty} g_n(x) = g(x) \quad \forall x \in \mathbb{R}^d.$$

This suffices because the supremum of a (countable) sequence of measurable functions is measurable again. For the supremum of uncountable many functions, this is not true in general. The former articles (Parzen, Nadaraja, Stute) did not consider these questions. Talagrand shows a central theorem about the bound of the probability that  $\|\alpha_n\|_{\mathfrak{G}}$  differs much from the expected value  $\mathbb{E}(\sup_{g \in \mathfrak{G}} |\sum_{i=1}^n \epsilon_i g(X_i)|)$ . We can consider this as a generalization of Tchebychev's inequality.  $\|\alpha_n\|_{\mathfrak{G}}$  is measurable if  $\mathfrak{G}$  is pointwise measurable.

**Lemma 3.10** (Talagrand) *Let  $\mathfrak{G}$  be pointwise measurable with  $\sup_{g \in \mathfrak{G}} \|g\|_{\infty} \leq M < \infty$ ,  $\sigma_{\mathfrak{G}}^2 := \sup_{g \in \mathfrak{G}} \text{Var}(g(X)) < \infty$  and let  $A_1, A_2$  be some constants. Then we have for each  $t > 0$*

$$\begin{aligned} \mathbb{P} \left\{ \max_{1 \leq m \leq n} \|\alpha_m\|_{\mathfrak{G}} \geq A_1 \left( \mathbb{E} \left( \sup_{g \in \mathfrak{G}} \left| \sum_{i=1}^n \epsilon_i g(X_i) \right| \right) + t \right) \right\} \\ \leq 2 \left\{ \exp \left( -\frac{A_2 t^2}{n \sigma_{\mathfrak{G}}^2} \right) + \exp \left( -\frac{A_2 t}{M} \right) \right\}. \end{aligned}$$

**Remarks** The formal similarity to Tchebychev's inequality lies in the fact that we get an upper bound (in a generalized sense) for the probability that a random variable strongly differs from its expected value. The rate of distance is given by the parameter  $t$ ; the upper bound is strongly monotonous decreasing in  $t$ . Tchebychev's inequality also needs that the second moment is finite.

Talagrand gets results about empirical processes of this kind based on similar results about Gaussian processes. We will point out the coherence later.

For the examination of almost sure uniform convergence, it is very useful to appreciate  $\mathbb{E}(\sup_{g \in \mathfrak{G}} |\sum_{i=1}^n \epsilon_i g(X_i)|)$  itself. This appreciation is deterministic and not stochastic. However, we need more assumptions on the richness of the function set.

**Lemma 3.11** (Bound for the expected value for general empirical processes) *Let  $\mathfrak{G}$  be a pointwise measurable set of bounded functions with the assumptions*

1.  $\exists G : \mathbb{R}^d \rightarrow \mathbb{R}$  with  $G(x) \geq \sup_{g \in \mathfrak{G}} |g(x)| \quad \forall x \in \mathbb{R}^d$
2.  $\mathbb{E}((G(X))^2) \leq \beta^2$
3.  $N(\epsilon, \mathfrak{G}) \leq C\epsilon^{-\nu}$  for all  $0 < \epsilon < 1$  (consider the remark after this theorem)
4.  $\sigma_0^2 := \sup_{g \in \mathfrak{G}} ((g(X))^2) \leq \sigma^2$
5.  $\sup_{g \in \mathfrak{G}} \|g\|_\infty \leq U$

for constants  $C, \nu \geq 1, 0 < \sigma \leq \beta, \sigma_0 \leq U \leq C_2 \sqrt{n} \beta, C_2 = (4\sqrt{\nu \log C_1})^{-1}, C_1 = \max(C_1^{\frac{1}{\nu}}, e)$ . Let  $C_3 = \frac{C_1^2}{16\nu}$  and  $A$  a constant. Then we have:

$$\mathbb{E} \left( \sup_{g \in \mathfrak{G}} \left| \sum_{i=1}^n \epsilon_i g(X_i) \right| \right) \leq A \left( \sqrt{\nu n \sigma_0^2 \log \left( \frac{C_1 \beta}{\sigma_0} \right)} + 2\nu U \log \left( C_3 n \frac{\beta^2}{U^2} \right) \right).$$

**Remark** The third condition was not necessary for Talagrand's inequality for the probability of the distance from the expected value, for the proof of this moment inequality it is however necessary. It is a topological entropy-condition on the kernel set, i.e. is may not be too rich.

Let  $Q$  be any probability measure on the measurable space  $(\mathbb{R}^d, \mathfrak{B}^d)$ . This measure defines a measure integral which induces a metric to measure the distance of two functions  $g_1, g_2$ , i.e.

$$d_Q(g_1, g_2) := \left( \int_{\mathbb{R}^d} (g_1 - g_2)^2 dQ \right)^{\frac{1}{2}}.$$

The positive definiteness and the symmetry are evident, the triangle inequality follows from Minkowski's inequality which is proved with Hölder's inequality (see Amann and Escher (2001), page 116). Now, we can ask how to cover the kernel set with balls using this metric.

**Definition 3.12** (Balls) *A ball around a function  $g$  with radius  $\epsilon > 0$  is defined as*

$$B_\epsilon(g) := \{g^* : d_Q(g^*, g) < \epsilon\}$$

*and contains all functions that have a distance smaller than  $\epsilon$  to  $g$ .*

In  $\mathbb{R}$  all functions  $g^*$  with  $\int_{\mathbb{R}} (g^*)^2 dx = 1$  have the distance 1 to the constant function  $g(x) = 0$ .

Now,  $N(\epsilon, \mathfrak{G}, d_Q)$  is the minimal number of such balls with radius  $\epsilon$  that are needed to cover  $\mathfrak{G}$ , i.e. the minimal number so that every function of  $\mathfrak{G}$  lies in at least one ball. If  $\mathfrak{G}$  contains the zero function and in addition only kernels with our former assumptions, for every  $\epsilon > 0$  we have  $N(1 + \epsilon, \mathfrak{G}, d_Q) = 1$  for all kernels are contained in the ball around the zero function with radius  $1 + \epsilon$ .

With the former defined  $G$  in our case  $N(\epsilon, \mathfrak{G})$  is defined as the supremum of the  $N(\epsilon d_Q(G, 0), \mathfrak{G}, d_Q)$  of all probability measures  $Q$  with  $0 < d_Q(G, 0) < \infty$ , i.e.

$$N(\epsilon, \mathfrak{G}) := \sup_Q N(\epsilon d_Q(G, 0), \mathfrak{G}, d_Q).$$

Obviously,  $N(\epsilon, \mathfrak{G})$  is monotonously decreasing in  $\epsilon$ : The smaller the radii may be, the more balls are needed. The condition from the moment inequality now says that the necessary number of balls may increase polynomially but not exponentially. For example, for the set  $\mathfrak{F}$  of all indicator functions  $I_{(-\infty, t]}$  we can say that  $N(\epsilon, \mathfrak{F}) \leq \frac{2}{\epsilon^2}$ , see also van der Vaart and Wellner (1996), page 129.

The conditions discussed in this chapter must just be true for some subsets of  $\mathfrak{K}$ , yet we assume that they are true for the whole set  $\mathfrak{K}$ . For instance, the entropy condition is true if  $K(x) = \phi(p(x))$ , where  $p$  is a polynomial in  $d$  dimensions and  $\phi$  is a rightcontinuous function with bounded variation. For  $d = 1$ , the measurability condition is true whenever  $K$  is rightcontinuous. Because  $\mathbb{Q}$  is dense in  $\mathbb{R}$ , we can choose

$$\mathfrak{K}_0 := \left\{ K \left( \frac{x - \cdot}{h_n} \right) : h_n \in \mathbb{Q}^+, x \in \mathbb{Q} \right\}$$

as a subset of  $\mathfrak{K}$ . For  $d = 1$ , these are the assumptions on the kernels that Nadaraja and Stute need, too.

The conditions guarantee that the generalized empirical process  $\alpha_n^*$  converges for  $n \rightarrow \infty$  against the Brownian Bridge, a special Gaussian process. So, it behaves in a "normal" way and thus, the maximal expected value can be appreciated upwards. The fact that many empirical processes converge against Gaussian processes is the reason for Talagrand's motivation.

The authors Einmahl and Mason at first only consider the stochastic part, just like Stute. We introduce the assumptions

$$(D1) \quad K : \mathbb{R}^d \rightarrow \mathbb{R}, K \in L^\infty(\mathbb{R}^d) \text{ with } \int_{\mathbb{R}^d} K(y) dy = 1.$$

$$(D2) \quad f \text{ is bounded.}$$

(D3) The assumption of the polynomial covering number holds for (6).

(D4) The pointwise measurability is satisfied for (6).

**Theorem 3.13** (Almost sure uniform consistency with variable bandwidth) *With (D1) - (D4) we have for every  $c > 0$   $\mathbb{P}$ -almost sure*

$$\limsup_{n \rightarrow \infty} \sup_{c \frac{\log n}{n} \leq h_n \leq 1} \frac{\sqrt{nh_n} \|f_n - \mathbb{E}f_n\|_\infty}{\sqrt{\max(-\log h_n, \log \log n)}} = K(c) < \infty.$$

*Proof.* We give a sketch of the proof. Let  $h_n$  be called  $h$  and introduce two real sequences. For  $j, k \geq 0$  and  $c > 0$  define  $n_k = 2^k$  and  $h_{j,k} = \frac{c 2^j \log(n_k)}{n_k}$ .  $c$  is fixed,  $j$  and  $k$  will vary in the proof. Define in addition the kernel set which fulfills the conditions of chapter 5.2:

$$\mathfrak{K}_{j,k} := \left\{ K \left( \frac{x - \cdot}{h^{\frac{1}{d}}} \right) : h_{j,k} \leq h \leq h_{j+1,k}, x \in \mathbb{R} \right\}.$$

We get two appreciations upwards for the expected values of the square kernels. We have

$$\begin{aligned} \mathbb{E} \left( K^2 \left( \frac{x - X}{h^{\frac{1}{d}}} \right) \right) &= \int_{\mathbb{R}^d} K^2 \left( \frac{x - s}{h^{\frac{1}{d}}} \right) f(s) ds \\ &= h \int_{\mathbb{R}^d} K^2(u) f \left( x - uh^{\frac{1}{d}} \right) du \leq h \|f\|_\infty \|K\|_2^2. \end{aligned}$$

It is important that the density is bounded so that this bound can become small for small  $h$ . Furthermore, we get for  $h_{j,k} \leq h \leq h_{j+1,k}$

$$\begin{aligned} \mathbb{E} \left( K^2 \left( \frac{x - X}{h^{\frac{1}{d}}} \right) \right) &\leq \min(\kappa^2, h_{j+1,k}) \|f\|_\infty \|K\|_2^2 \\ &= \min(\kappa^2, 2 \|f\|_\infty \|K\|_2^2 h_{j,k}) =: \min(\kappa^2, D_0 h_{j,k}) =: \sigma_{j,k}^2. \end{aligned}$$

Applying the moment inequality with the Rademacher variables, we appreciate the expression

$$\mathbb{E} \left( \sup_{g \in \mathfrak{K}_{j,k}} \left( \sum_{i=1}^{n_k} \epsilon_i g(X_i) \right) \right)$$

and get for large  $k$

$$\begin{aligned} \mathbb{E} \left( \sup_{g \in \mathfrak{K}_{j,k}} \left( \sum_{i=1}^{n_k} \epsilon_i g(X_i) \right) \right) &\leq D_3 \sqrt{n_k h_{j,k} \log \frac{1}{D_2 h_{j,k}}} \\ &\leq D_3 \sqrt{n_k h_{j,k} \max \left( \log \frac{1}{D_2 h_{j,k}}, \log \log n_k \right)} =: D_3 a_{j,k} \end{aligned} \tag{7}$$

with a constant  $D_3$  and  $D_2 = \frac{D_0}{\beta^2}$ . The inequality is especially true if the sum begins with  $n_{k-1}$ .

On  $\mathfrak{K}_{j,k}$ , we apply Talagrand's inequality with  $M = \kappa$  and

$\sup_{g \in \mathfrak{K}_{j,k}} \text{Var}(g(X)) \leq \sup_{g \in \mathfrak{K}_{j,k}} \mathbb{E}(g(X)^2) = \sigma_0^2 \leq D_0 h_{j,k}$ . For every  $t > 0$  we have

$$\mathbb{P}\left\{\max_{n_{k-1} \leq n \leq n_k} \sup_{g \in \mathfrak{K}_{j,k}} |\alpha_n(g)| \geq A_1(D_3 a_{j,k} + t)\right\} \leq 2 \left( \exp\left(\frac{-A_2 t^2}{D_0 n_k h_{j,k}}\right) + \exp\left(\frac{-A_2 t}{\kappa}\right) \right).$$

With this, we appreciate the probability that the maximal distance between the estimated density and the expected value of the kernel density estimator is large. For any  $\rho > \max(1, 2\sqrt{\frac{D_0}{A_2}})$  and  $k \geq 1$  we define

$$p_{j,k}(\rho) := \mathbb{P}\left\{\max_{n_{k-1} \leq n \leq n_k} \sup_{g \in \mathfrak{K}_{j,k}} |\alpha_n(g)| \geq A_1(D_3 + \rho)a_{j,k}\right\}$$

and can show for large  $k$

$$p_{j,k}(\rho) \leq 4(\log n_k)^{-\frac{A_2}{D_0}\rho^2}.$$

For this appreciation the term  $\log \log n_k$  in (7) is very useful.

Let  $l_k := \max(j : h_{j,k} \leq 2)$ . Obviously,  $l_k \leq n_k = \frac{\log n_k}{\log 2}$  because for  $j = n_k$  the sequence  $h_{j,k}$  goes to infinity. It follows:

$$P_k(\rho) := \sum_{j=0}^{l_k-1} p_{j,k}(\rho) \leq \frac{4}{\log 2} (\log n_k)^{1-\frac{A_2}{D_0}\rho^2}.$$

The series over  $P_k(\rho)$  converges because the exponent of  $\log n_k$  is strictly smaller than  $-1$  and  $\log n_k$  has size  $k$ . This is essential for applying the Borel-Cantelli-Lemma.

For small  $k$  and  $n_{k-1} \leq n \leq n_k$  we can show

$$\begin{aligned} A_k(\rho) &:= \left\{ \max_{n_{k-1} \leq n \leq n_k} \sup_{\frac{c \log n}{n} \leq h \leq 1} \frac{\sqrt{nh} \|f_n - \mathbb{E}f_n\|_\infty}{\sqrt{\max(-\log h, \log \log n)}} > 2A_1(D_3 + \rho) \right\} \\ &\subset \left\{ \max_{n_{k-1} \leq n \leq n_k} \sup_{\frac{c \log n_k}{n_k} \leq h \leq h_{l_k,k}} \frac{\sqrt{nh} \|f_n - \mathbb{E}f_n\|_\infty}{\sqrt{\max(-\log h, \log \log n)}} > 2A_1(D_3 + \rho) \right\} \\ &\subset \bigcup_{j=0}^{l_k-1} \left\{ \max_{n_{k-1} \leq n \leq n_k} \sup_{g \in \mathfrak{K}_{j,k}} |\alpha_n(g)| \geq A_1(D_3 + \rho)a_{j,k} \right\}. \end{aligned}$$

$\mathbb{P}$  is a probability measure and so  $\mathbb{P}(A_k(\rho)) \leq P_k(\rho)$ . The series over the right expression converges and so, the series over  $\mathbb{P}(A_k(\rho))$  converges, too. The Borel-Cantelli-Lemma now says that the probability of the limit superior of the set sequence  $A_k(\rho)$  is 0. With

probability 1, only finitely many elements of the in  $A_k(\rho)$  formulated series of random variables are larger than  $A_1(D_3 + \rho)$ ; consequently the limit superior of the series is finite with probability 1. The exact limit is unknown but it depends on  $c$ .  $\square$

In the following, we consider a special series of intervals for  $h_n$ , i.e.  $h_n \in [a_n, b_n]$  with  $0 < a_n < b_n \leq 1$ . We assume

(D5)  $b_n \rightarrow 0$  for  $n \rightarrow \infty$  (this implies  $a_n \rightarrow 0$ ) and  $\frac{na_n}{\log n} \rightarrow \infty$ .

The last condition is sufficient for the application of the theorem because then there is a  $n \in \mathbb{N}$  for every  $c > 0$  so that  $h_n \geq a_n \geq c \frac{\log n}{n}$ .

Now,

$$\frac{\sqrt{nh_n}}{\sqrt{\max(-\log h_n, \log \log n)}}$$

tends to  $\infty$  for  $n \rightarrow \infty$  for every  $h_n \in [a_n, b_n]$  and so,  $\sup_{a_n \leq h_n \leq b_n} \|f_n - \mathbb{E}f_n\|_\infty$  tends to 0 almost surely. We introduce the assumptions

(D6) For  $x \in \mathbb{R}^d$  define  $\psi_K(x) := \sup_{|y| \geq |x|} |K(y)|$  and let  $\psi_K \in L^1(\mathbb{R}^d)$ .

(D7)  $f$  is uniformly continuous on  $\mathbb{R}^d$ .

(D8)  $f$  is bounded on  $\mathbb{R}^d$ .

and get for the bias convergence

**Theorem 3.14** *With the assumptions (D6) - (D8) it holds  $\mathbb{P}$ -almost sure*

$$\lim_{n \rightarrow \infty} \sup_{a_n \leq h_n \leq b_n} \|\mathbb{E}f_n - f\|_\infty = 0.$$

*Proof.* Because of the continuity of  $f$ ,  $\sup_{a_n \leq h_n \leq b_n} \|\mathbb{E}f_n - f\|_\infty$  is realized for every  $n \in \mathbb{N}$  for some  $h_{nj}$  (theorem about maximum and minimum with continuous functions). The series  $(h_{nj})_{n \in \mathbb{N}}$  is then still a zero sequence. Consider in the following  $\mathbb{E}f_n$  in dependence of this series.

Let  $R > 0$  arbitrary. Then we can say

$$\begin{aligned} \|\mathbb{E}f_n - f\|_\infty &= \sup_{z \in \mathbb{R}^d} |\mathbb{E}f_n - f| \leq \sup_{|z| \leq R} |\mathbb{E}f_n - f| + \sup_{|z| > R} |\mathbb{E}f_n - f| \\ &\leq \sup_{|z| \leq R} |\mathbb{E}f_n - f| + \sup_{|z| > R} |\mathbb{E}f_n| + \sup_{|z| > R} |f| \\ &\leq \sup_{|z| \leq R} |\mathbb{E}f_n - f| + 2 \sup_{|z| > R} |f| \\ &=: A_n + B. \end{aligned}$$

The last inequality follows from  $\int_{\mathbb{R}^d} K(y)dy = 1$ .

Let  $\epsilon > 0$  arbitrary. Choose  $R$  so large, that  $B < \epsilon$ . This is possible because  $f$  is a uniformly continuous density. Choose  $n$  so large that  $A_n < \epsilon$ . For this, we use a theorem of (Stein (1970), page 65) because  $\{z \in \mathbb{R}^d \mid |z| \leq R\}$  is a compact set and  $(h_{nj})_{n \in \mathbb{N}}$  is a zero sequence.  $\square$

**Remark** For the application of Stein's theorem the assumptions on  $\psi_K$  and  $f$  are necessary. The assumption on  $\psi_K$  says that the kernel may not have too heavy tails. We have already noticed this phenomenon in Stute's articles.

Summarized, it follows

**Corollary 3.15** *Let  $h_n \in [a_n, b_n]$  with  $0 < a_n < b_n \leq 1$ . With the assumptions (D5) - (D8) we get  $\mathbb{P}$ -almost sure*

$$\lim_{n \rightarrow \infty} \sup_{a_n \leq h_n \leq b_n} \|f_n - f\|_\infty = 0.$$

We consider the special case  $a_n = b_n = h_n$ . In addition to the assumption  $\frac{nh_n}{\log n} \rightarrow \infty$  Stute's three assumptions are true. With these assumptions  $-\log h_n$  increases faster than  $\log \log n$  and so

$$\max(-\log h_n, \log \log n) = -\log h_n$$

for sufficiently large  $n$ . With this, Einmahl's and Mason's theorem gives a convergence rate of  $\sqrt{\frac{nh_n}{-\log h_n}}$ . Except of the factor 2, this is Stute's convergence rate but the result is stronger because it is true for the whole  $\mathbb{R}^d$ .

In contrast to Stute however, they do not calculate an exact limit and only say that the limit is finite.

Another approach to deal with variable bandwidth is described in Schäfer (1986) and Weißbach (2006).

## 4 Summary

In this survey article we compared different proofs for consistency of kernel density estimators and lay the focus on the almost sure convergence which is stronger than uniform convergence in probability. We saw a clear historical development, many techniques for the proofs were used frequently and former results were tightened in later articles.

Recurrent techniques are for example the lemma of Borel-Cantelli which is often useful for theorems about almost sure convergence, the partial integration for Lebesgue-Stieltjes-integrals and the theory of empirical processes. Especially here, we notice a development



over the years. In 1982, Stute just uses the empirical distribution function as a special case of an empirical process; the approach of Einmahl and Mason in 2005 is more general. With more complex mathematical techniques, they get stronger results. However, reduced to the case of a fixed bandwidth, both authors get the same convergence rates for the almost sure uniform convergence.

In general, the stochastic part of the error between  $f_n$  and  $f$  is a bigger problem than the deterministic one. The authors focus on the stochastic part and treat the bias secondarily. For appreciating the stochastic part, i.e. for finding almost sure bounds of the stochastic process, we especially need assumptions on the kernels; a recurrent assumption is right continuity and bounded variation. Often, we also need boundedness of  $f$ . The deterministic part needs assumptions on the smoothness of  $f$ , e.g. uniform continuity. Surely, there is more research needed in the theory of consistency. An approach would be the calculation of the precise constants in the consistency result of Einmahl and Mason. Here are parallels to Stute: At first, he achieved a convergence result with a constant not defined precisely (but finite) and later on, he calculated this constant in the case of uniformly continuous  $f$ .

In addition it would be interesting to think about cases in which the data-adaptive bandwidth lies almost sure in an interval  $[a_n, b_n]$  for sufficient large  $n$ . We have seen that Einmahl's and Mason's strong consistency result could be used in practice only then. With the described condition on the data-adaptive bandwidth that is less strong, the consistency result is also useful, but maybe, there is potential for improvement.

*Acknowledgement:* We would like to thank U. Einmahl for pointing out some references. The financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures") is gratefully acknowledged.

## References

- E. Amann and J. Escher. *Analysis III*, volume 1. Birkhäuser, Basel, 2001.
- H. Büning and G. Trenkler. *Nichtparametrische Statistische Methoden*, volume 2. de Gruyter, Berlin, 1994.
- P. Deheuvels and D.M. Mason. General asymptotic confidence bands based on kernel-type function estimators. *Stat. Inference Stoch. Process.*, 7:225–277, 2004.
- H. Dette. A consistent test for heteroscedasticity in nonparametric regression based on the kernel method. *Journal of Statistical Planning and Inference*, 103:311–330, 2002.

- U. Einmahl and D.M. Mason. Uniform in bandwidth consistency of kernel-type function estimators. *Annals of Statistics*, 33:1380–1403, 2005.
- P. Hall and J.S. Marron. Improved variable window kernel estimates of probability densities. *Annals of Statistics*, 23:1–10, 1995.
- W. Härdle. *Smoothing techniques*. Springer, New York, 1991.
- J.S. Marron, W. Gonzalez-Manteiga, and R. Cao. Bootstrap selection of the smoothing parameter in nonparametric hazard rate estimation. *Journal of the American Statistical Association*, 91:1130–1140, 1996.
- E.A. Nadaraja. On nonparametric estimates of density functions and regression curves. *Theory of Probability and Applications*, 10:186–190, 1965.
- E. Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.
- M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27:832–835, 1956.
- H. Schäfer. Local convergence of empirical measures in the random censorship situation with application to density and rate estimators. *Annals of Statistics*, 14:1240–1245, 1986.
- A.N. Sirjaev. *Probability*. Springer, New York, 1984.
- E.M. Stein. *Singular integrals and differentiability properties of functions*. Princeton University Press, New Jersey, 1970.
- W. Stute. The law of the logarithm for kernel density estimators. *Annals of Probability*, 10:414–422, 1982.
- A.W. van der Vaart and J.A. Wellner. *Weak convergence and empirical processes*. Springer, New York, 1996.
- M. Wand and M. Jones. *Kernel smoothing*. Chapman and Hall, London, 1995.
- R. Weißbach. A general kernel functional estimator with general bandwidth - strong consistency and applications. *Nonparametric Statistics*, 18:1–12, 2006.