Noname manuscript No.
(will be inserted by the editor)

Consistency of the kernel density estimator - a survey

Dominik Wied and Rafael Weißbach

Version 31-05-10 - accepted by Statistical Papers

Abstract Various consistency proofs for the kernel density estimator have been developed over the last few decades. Important milestones are the pointwise consistency and almost sure uniform convergence with a fixed bandwidth on the one hand and the rate of convergence with a fixed or even a variable bandwidth on the other hand. While considering global properties of the empirical distribution functions is sufficient for strong consistency, proofs of exact convergence rates use deeper information about the underlying empirical processes. A unifying character, however, is that earlier and more recent proofs use bounds on the probability that a sum of random variables deviates from its mean.

Keywords Kernel estimation; Pointwise consistency; Strong uniform consistency; Empirical process; Rate of convergence, Variable bandwidth

Mathematics Subject Classification (2000) 60-02, 62-02

1 Introduction

For more than 50 years, mathematical statistics has been dealing with the problem of efficiently estimating the probability density function of continuous random variables from a random sample. In 1956, M. Rosenblatt proposed a kernel-based estimator with the basic idea of looking at the difference quotient of the empirical distribution function. This idea is still in use; 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 been transferred, for example, to regression estimation (e.g. Nadaraja 1964 or Watson 1964), survival analysis (e.g. Diehl and Stute 1988 or Marron et al. 1996) or the theory of jump processes (e.g. Schäbe and Tiedge 1995).

In this survey article, we compare different proofs of (different types of) consistency in terms of their historical development. Several important articles have exerted a major influence on this field of research. The first is an article by E. Parzen from 1962, proving pointwise consistency for the first time. In 1965, E.A. Nadaraja demonstrated almost sure uniform convergence, for which he needed the

Dominik Wied
Institut für Wirtschafts- und Sozialstatistik, Technische Universität Dortmund, 44221 Dortmund, Germany
Phone: +49 231 755 3869
Fax: +49 231 755 5284
E-mail: dominik.wied@tu-dortmund.de

Rafael Weißbach
Lehrstuhl für Statistik, Institut für Volkswirtschaftslehre, Universität Rostock, 18051 Rostock, Germany

concept of bounded variation. About two decades later, W. Stute obtained in 1982 some valuable results on convergence rates of the estimator, depending on the sample size, kernel and true density. He did not just look at global properties of the empirical distribution function as did the other authors before, but considered local properties of empirical processes instead. Another generalized empirical process approach was proposed in 2005 by U. Einmahl and D.M. Mason. They proved almost sure convergence in a situation in which the bandwidth is not fixed, but may randomly vary within a small interval. They used sophisticated mathematical techniques from other fields such as topology basing on papers by M. Talagrand, and considered aspects like measurability.

2 Consistency

2.1 Weak consistency

We consider a probability space $(\Omega, \mathfrak{A}, \mathbb{P})$ and i.i.d. random variables $X_i : \Omega \to \mathbb{R}^d, 1 \leq i \leq n, d \geq 1$, 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 $F_n : \mathbb{R} \times \Omega \to [0, 1]$ 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 Glivenko and Cantelli, the convergence is uniform.

Following Parzen, a heuristic approach to estimating the density entails considering the difference quotient of F_n and using it as an estimator f_n of the density for a sufficiently small h_n:

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

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

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

The estimator in (1) can then be written as an integral of $\frac{1}{h_n} K\left(\frac{\cdot}{h_n}\right)$ with respect to F_n which is an average of Dirac point measures δ_{X_i}. Thus, integration with respect to F_n is obtained by evaluating the integrand at X_i:

$$f_n(x, \omega) = \frac{1}{h_n} \int_{-\infty}^{\infty} K\left(\frac{x - y}{h_n}\right) dF_n(y, \omega) = \frac{1}{n h_n} \sum_{i=1}^{n} K\left(\frac{x - X_i(\omega)}{h_n}\right). \tag{2}$$

It is useful to consider other nonnegative kernel functions K in (2), especially continuous kernel functions. This guarantees that f_n is nonnegative and continuous as a finite sum of nonnegative and continuous functions. In the present paper, we let the assumptions on the kernel functions depend on the situation. In practice, however, the choice of K is of minor interest for the estimation, see Wand and Jones (1995, p. 31).

Parzen looks at the expected value and variance of the estimator at a 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 assumes that K is a real-valued, Borel-measurable function with $\sup_{y \in \mathbb{R}} |K(x)| =: ||K||_\infty < \infty$, i.e. $K \in L^\infty(\mathbb{R})$. In addition, $K \in L^1(\mathbb{R})$

and hence, also $K \in L^2(\mathbb{R})$. We assume $\lim_{x\to\infty} |xK(x)| = \lim_{x\to\infty} |xK^2(x)| = 0$.
With these assumptions, at each point of continuity x of f

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

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

For finite n, $\mathbb{E}(f_n(x))$ is the convolution of f and $\frac{1}{h_n}K(\frac{\cdot}{h_n})$, or, by substitution $z = \frac{x-y}{h_n}$, $\int_{-\infty}^{\infty} K(z)f(x-zh_n)dz$.
Parzen also calculates the asymptotic variance of the density estimator. At each point of continuity x of f,

$$\lim_{n\to\infty} nh_n Var(f_n(x)) = f(x) \int_{-\infty}^{\infty} K^2(y)dy =: k < \infty. \tag{3}$$

This means that the asymptotic variance is proportional to $f(x)$. One needs more restrictive assumptions on $(h_n)_{n\in\mathbb{N}}$ for consistency than for asymptotic unbiasedness: With

(A1) $\lim_{n\to\infty} nh_n = \infty$

we ensure that the variance tends to 0 for $n \to \infty$. Thus, there is a trade-off between controlling the bias and controlling the asymptotic variance.

Theorem 1 (Weak consistency) *Let the assumption (A1) hold. Then, at each point of continuity x of f, the estimator $f_n(x)$ is weakly consistent, i.e. for each $\epsilon > 0$*

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

Proof We consider the mean square error of $f_n(x)$, $MSE(f_n(x)) = Var(f_n(x)) + (Bias(f_n(x)))^2$. The estimator $f_n(x)$ is asymptotically unbiased and so, $\lim_{n\to\infty}(Bias(f_n(x)))^2 = 0$. With (3), we have $\lim_{n\to\infty} Var(f_n(x)) = \lim_{n\to\infty} \frac{k}{nh_n} = 0$. Hence, $f_n(x)$ is consistent in the quadratic mean and hence weakly consistent. □

We will see later on, that for proving strong consistency or for proving convergence rates even more assumptions are necessary.

For assertion (3), one can even relax the assumption of continuity; with results from differentiation theory (see Wheeden and Zygmund 1977, p. 100-109) this holds for Lebesgue-almost any x, as long as $\int_{-\infty}^{\infty} K^2(x)dx$ is finite.

Note that we have proved pointwise, but not uniform consistency in probability. Using Fourier analysis, Parzen showed uniform consistency in probability under the assumption $\lim_{n\to\infty} nh_n^2 = \infty$.

2.2 Strong consistency

To the best of our knowledge, in 1965, Nadaraja was the first researcher to formulate a theorem dealing with the almost sure uniform convergence of the kernel density estimator. This constituted progress over Parzen, who did not deal with almost sure convergence. We need the additional assumption of bounded variation for the kernels.

Definition 1 (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

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

where the supremum is taken over the set $\mathfrak{P} = \{P = \{x_0, \ldots, x_{n_P}\} : \text{P is partition of } [a,b]\}$. 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 for the entire $\mathbb{R}$, we define

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

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

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

holds with probability 1. Hence,

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

With the theorem of Glivenko 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, including the concept of bounded variation:

(B1) The kernel K is a right-continuous function.
(B2) K is of bounded variation.
(B3) $\lim_{|x|\to\infty} K(x) = 0$.
(B4) f is a uniformly continuous density function.
(B5) $\sum_{n=1}^{\infty} \exp(-\gamma n h_n^2) < \infty$ for every $\gamma > 0$.

The concept of bounded variation is needed to ensure that the kernel does not fluctuate excessively. For assumption (B5), the assumption $\lim_{n\to\infty} nh_n^2 = \infty$ that Parzen needed for uniform consistency in probability is not sufficient: If you choose h_n so that

$$nh_n^2 = \log\log n,$$

then $\exp(-\gamma nh_n^2) = (\log n)^{-\gamma}$ and the series summing this does not converge.

The assumption $\lim_{|x|\to\infty} K(x) = 0$ is new, because it does not strictly follow from $K \in L^1(\mathbb{R})$. In fact, it is possible to construct a continuous and bounded function K with $\int_{\mathbb{R}} K(x)dx = 1$ so that $\limsup_{|x|\to\infty} K(x) = 1$. K is a jagged function the serrations of which have a height of 1 and which, for $|x| \to \infty$, become so small that the area below is 1. We can think of serrations with a length of $\left(\frac{1}{2}\right)^{|x|}$ which lie around the numbers $x \in \mathbb{Z}, |x| \geq 1$. Otherwise, the function equals 0. The integral of the function is 1, i.e. the value of the geometric series $\sum_{n\geq 1} \left(\frac{1}{2}\right)^n$.

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

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

The basic idea behind proving the almost sure uniform convergence is the consideration of the (suitably scaled) sequence of random variables $(||f_n(x) - f(x)||_\infty)_{n\in\mathbb{N}}$. Firstly, we apply the triangle inequality for norms, i.e.

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

It is sufficient to prove that both summands in the last expression tend to 0 for $n \to \infty$. The first summand is stochastic, the second one deterministic. The convergence proof of the deterministic component (the bias) is much easier, but we need assumptions about f, namely uniform continuity, which are not necessary for the convergence of the stochastic component. On the other hand, the first summand requires assumptions about the kernel, namely bounded variation, which are not necessary for the bias convergence.

We now prove Theorem 2, concentrating on the stochastic part: For the convergence of the stochastic component, we need the Borel-Cantelli lemma. Assume that we wish to show that, for $\mathbb{P}$-almost any $\omega \in \Omega$, the limit superior of the sequence of random variables $B_n := (||f_n(x) - \mathbb{E}(f_n(x))||_\infty)_{n\in\mathbb{N}}$ is bounded from above by a number. With Borel-Cantelli, we have to show that, for any $\epsilon > 0$,

$$\sum_{n\geq 1} \mathbb{P}(B_n \geq \epsilon) < \infty.$$

In the following, we show how to get a suitable inequality for $\mathbb{P}(B_n \geq \epsilon)$ to achieve this. Write

$$B_n = \sup_{-\infty<x<\infty} \left| \frac{1}{h_n} \int_{-\infty}^{\infty} K\left(\frac{x-y}{h_n}\right) dF_n(y) - \frac{1}{h_n} \int_{-\infty}^{\infty} K\left(\frac{x-y}{h_n}\right) dF(y) \right|.$$

We apply integration by parts to both integrals. Formally, the integrand is the function $K\left(\frac{x-\cdot}{h_n}\right)$. Because of the bijective linear map $y \to \frac{x-y}{h_n}$, $K\left(\frac{x-\cdot}{h_n}\right)$ has the same total variation as K. Especially, it is of bounded variation. The first summands vanish, because we have assumed that $\lim_{|x|\to\infty} K(x) = 0$. Hence,

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

The last equation follows from the fact that we consider the supremum over all $x \in \mathbb{R}$ and because of the right-continuity of F_n (see definition). Accordingly, we have reduced the contribution of the stochastic component to the maximum distance between the empirical and the theoretical distribution function. Using the inequality of Dvoretzky et al. (1956), using Massart's constant (Massart 1990), it holds for arbitrary $\lambda > 0$:

$$\mathbb{P}\left(||F_n(x) - F(x)||_\infty > \frac{\lambda}{\sqrt{n}} \right) \leq 2\exp(-2\lambda^2). \tag{4}$$

With the above upper limit of B_n we finally achieve, for any arbitrary $\epsilon > 0$,

$$\mathbb{P}(B_n > \epsilon) \leq \mathbb{P}\left(||F_n(x) - F(x)||_\infty > \epsilon h_n \frac{1}{V_{-\infty}^{\infty}(K)} \right) \leq 2\exp(-2\epsilon^2 (V_{-\infty}^{\infty}(K))^{-2} n h_n^2) = 2\exp(-\beta n h_n^2),$$

where $\beta := 2\epsilon^2 (V_{-\infty}^{\infty})^{-2}$ is finite and deterministic.

With the assumptions of the theorem, specifically (B5), 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 sequence of random variables $(B_n)_{n\in\mathbb{N}}$ is smaller than or equal to ϵ with probability 1. Since ϵ was arbitrary, the limit superior is 0. Since the sequence is nonnegative, the limit inferior and therefore the limit must also be 0. □

Remark The idea for proving the convergence of the bias is similar to Parzen's idea of showing the pointwise asymptotic unbiasedness and the variance of the kernel density estimator. For the convergence of the bias, we need no other assumption about the bandwidth than it is a null sequence.

3 Rates of convergence with advanced empirical process techniques

Comparable to Nadaraja, several authors provided similar results about the almost sure uniform convergence. However, a general problem with many approaches was that the theorems did not take the form of f and K into account. The researchers were able to obtain results on the convergence itself, but not on the rates of convergence. The reason is that the examination of the behavior of f_n was reduced to the global examination of F_n and F, and moreover to the supremum norm of their distance, so that information was lost.

3.1 Rates of convergence with local properties of the empirical process

Stute found an improvement in 1982, by describing exact rates using local properties of empirical processes. The major progress attributable to him was that he provided a local study of the empirical process (i.e. of global properties of empirical distributions) in order to study F locally in terms of f. With this, he achieved precise convergence rate results on a bounded interval $J := (a, b), a < b$. The reason for this restriction, as we will see, is that on the entire J, the density must be above a certain threshold $m > 0$: In his limit results, Stute uses a standardizing factor $\sqrt{f(x)}$ in the denominator with which he gets precise limit results. Stute considers kernels with bounded support and focuses on the stochastic part.

Definition 3 (Empirical process) For a set $\mathfrak{G}$ of Borel-measurable functions $g : \mathbb{R}^d \to \mathbb{R}$, the empirical process $\alpha_n(g)$ is defined by

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

Stute refines the empirical process to the distribution functions on $\mathbb{R}$, i.e. for $x \in \mathbb{R}$ he studies $\alpha_n(I_{(-\infty,x]}(\cdot)) = \sqrt{n}(F_n(x) - F(x)$. The first step is to explore the example of the uniform empirical process, i.e. of $\gamma_n(x) = \sqrt{n}(\bar{F}_n(x) - x)$ for $x \in [0, 1]$, where $\bar{F}_n$ is the empirical distribution function of random variables that are uniformly distributed on $[0, 1]$. For uniform distribution, we have $F(x) = x$. It is well known that the example can be generalized to distributions with a positive density. The positivity $f(x) > 0$ implies a strictly monotonously increasing, and hence invertible, F. Thus, with the fundamental theorem of statistics (see Büning and Trenkler 1994, p. 52), $F(X_i)$ is distributed uniformly on $[0, 1]$. We can conclude that $F_n(x) = \bar{F}_n(F(x))$ and that

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

Instead of studying the (total) variation of the kernel, as in the previous section, Stute studies the oscillation modulus ω_n of γ_n that describes the local oscillation behaviour of the uniform empirical process.

Definition 4 (Oscillation modulus) For an $h \in \mathbb{R}_0^+$, the oscillation modulus ω_n of the uniform empirical process γ_n is defined by

$$\omega_n(h) := \sup_{|y-x| \le h} |\gamma_n(y) - \gamma_n(x)|.$$

Stute makes slightly different assumptions on the bandwidth, compared to Parzen or Nadaraja. Specifically, for $n \to \infty$ Assumption (A1), $nh_n \to \infty$, is maintained. In addition, it is assumed that $\frac{-\log h_n}{nh_n} \to 0$ (A2) and $\frac{-\log h_n}{\log \log n} \to \infty$ (A3). Assumption (A2) states, similarly to (A1), that h_n does not converge to 0 too fast. Assumption (A3) urges h_n to converge sufficiently fast to 0. Interestingly, the former condition (B5), or $nh_n^2 \to \infty$, becomes obsolete.

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

Lemma 1 (Convergence of the oscillation modulus) *Let J be a subinterval of $[0, 1]$ and bandwidth h_n fulfill Assumptions (A1)-(A3). Then, for arbitrary $0 < c_1 \leq c_2 < \infty$ it holds $\mathbb{P}$-almost surely*

$$\lim_{n\to\infty} \sup_{c_1 h_n \leq y-x \leq c_2 h_n; x,y\in J} \frac{|\gamma_n(y) - \gamma_n(x)|}{\sqrt{2(y-x)(-\log h_n)}} = 1. \tag{5}$$

In addition, $\mathbb{P}$-almost surely

$$\lim_{n\to\infty} \frac{\omega_n(h_n)}{\sqrt{2h_n(-\log h_n)}} = 1. \tag{6}$$

Remarks The proof of (5) consists of two parts: Firstly, one shows that the limit superior of the sequence 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 complete. The proof of the limit-inferior part uses the technique of poissonization. The basic idea is that $\sqrt{n}\gamma_n$, apart from the standardization and for fixed x, 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 compared with the normal distribution.

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

As in proof of (6), the limit superior and the limit inferior are analyzed separately. The result for the limit inferior follows from (5) by choosing $c_1 = c_2 = 1$. The result for the limit superior is proved similarly to that for the limit superior used in (5).

From (5) by substituting $y := F(y)$ and $x := F(x)$, with the mean value theorem for differentiation, an analogous result follows for the empirical process $\alpha_n(I_{(-\infty,x]}(\cdot))$. To this end, we introduce two new assumptions:

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

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

It is interesting to note that, later in the 1980s, H. Schäfer proved convergence rates for Lipschitz-continuous f, which is a stronger property than the uniform continuity (Schäfer 1986). However, Schäfer extended the scope from density estimation in the setup of independent and identically distributed observations to survival analysis and a variable bandwidth. Recently, Weißbach (2006) followed-up on this with a yet broader scope, but the same techniques, in order to increase the applicability to survival analysis. The variable bandwidth is the subject of the next section.

Lemma 2 (Convergence of the empirical process) *Let the assumptions (A1)-(A3),(C1) and (C2) hold. Let $\xi_{x,y}$ be an arbitrary point between x and y. Then, it holds for arbitrary $0 < c_1 \leq c_2 < \infty$ $\mathbb{P}$-almost surely*

$$\lim_{n\to\infty} \sup_{c_1 h_n \leq y-x \leq c_2 h_n; x,y\in J} \frac{|\alpha_n(I_{(-\infty,y]}(\cdot)) - \alpha_n(I_{(-\infty,x]}(\cdot))|}{\sqrt{2(y-x)f(\xi_{x,y})(-\log h_n)}} = 1.$$

Theorem 3 *With the assumptions (A1)-(A3),(C1) and (C2), the rectangular kernel satisfies $\mathbb{P}$-almost surely*

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

Proof With Lemma 2, we directly obtain a result about the convergence rate of the stochastic component of the rectangular kernel density estimator with kernel $K(x) = I_{[-\frac{1}{2},\frac{1}{2})}(x)$. This is simple, because the rectangular estimator (in contrast to estimators with more complicated kernels) works directly with the empirical distribution function. In this context, Lemma 2 gives a convergence result. At first, we choose $c_1 = c_2 = 1$ and $\xi_{x,y} = \frac{x+y}{2}$. With the definitions of f_n and $\mathbb{E}(f_n)$, we then obtain

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

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

The rectangular kernel is surely not an ideal estimator in practice, because $f_n(x)$ is not continuous in general, even if the true density is uniformly continuous. To overcome this, Stute obtains results for more general kernels, based on the results on the empirical process. He does this in several steps, at first by considering step functions like

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

for a finite m, nonnegative real numbers $a_i, 1 \le i \le m$ and $\infty < d_1 < d_2 < \ldots < d_{m+1}$. For such kernels, the convergence result is similar to the result for the rectangular kernel, and can be proved similarly.

Corollary 4 *With the assumptions (A1)-(A3),(C1) and (C2), and for a step-function kernel $\mathbb{P}$-almost surely*

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

The next step is the extension to kernels with bounded variation. Consider the following assumptions:

(C3) $K(x) = 0$ outside a bounded interval $[r, s)$, where $K(r) = 0$.
(C4) F is Lipschitz-continuous on an interval $J = (a, b), a < b$, with Lipschitz-constant $m^\star < \infty$.

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

Lemma 3 *With the assumptions (A1)-(A3),(B1), (C3) and (C4), for each $\epsilon > 0$*

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

where the constant C satisfies $C \le \sqrt{m^\star(s-r)}V_r^s(K)$ with the total variation $V_r^s(K) < \infty$.

Proof We again apply the technique of partial integration, however, now distinctly different as compared to Nadaraja. Since global properties of empirical processes do not apply here, we use local properties as presented in Lemma 1 and 2. At first, for a fixed $x \in J_\epsilon$, y must lie within 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

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

This is necessary to work with the standard empirical process $\alpha_n(I_{(-\infty,x]}(\cdot))$ later on. For a fixed $x \in J_\epsilon$, we get

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

Now, $F_n(y-) \leq F_n(y)$ and the distance between y and $x - rh_n$ 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

$$\sup_{|y-x|\leq(s-r)h_n, x,y\in J} |\alpha_n(I_{(-\infty,y]}(\cdot)) - \alpha_n(I_{(-\infty,x]}(\cdot))| = \sup_{|y-x|\leq(s-r)h_n, x,y\in J} |\gamma_n(F(y)) - \gamma_n(F(x))|.$$

As in Theorem 3, we need a small $\epsilon > 0$ as a buffer for x, so that $x - sh_n$ and $x - rh_n$ can potentially be in J.

On J, we have $|F(a) - F(b)| \leq m^\star|a - b|$ and thus, the integrand is smaller than or equal to $\omega_n(m^\star(s - r)h_n)$. In summary, we have the inequality

$$\sqrt{nh_n}\sup_{x\in J_\epsilon} |f_n(x) - \mathbb{E}(f_n(x))| \leq (h_n)^{-\frac{1}{2}}\omega_n(m^\star(s - r)h_n)V_r^s(K). \tag{7}$$

With (6), it follows that the expression

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

is $\mathbb{P}$-almost surely finite, because of the appropriately scaled $\omega_n(m^\star(s - r)h_n)$ (even $\leq V_r^s(K)$). The theorem follows from the fact that $-\log(m^\star(s - r)h_n)$ and $-\log(h_n)$ have the same limiting behavior. □

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

The finiteness of the Lipschitz-constant of F is necessary for proving the finiteness of the limit superior of the sequence. The condition is fulfilled if the density is bounded.

In the lemma, the supremum is taken over $x \in J_\epsilon$, but the lemma also holds for $x \in \mathbb{R}$, if one assumes that F is Lipschitz-continuous on $\mathbb{R}$. However, for the following corollary, the restriction on J_ϵ is necessary.

If $F' = f$ is uniformly continuous on J, we can calculate C explicitly. The basic idea is to split the kernel into the sum of two functions, where one is a step function. By combining Corollary 4 and Lemma 3, we obtain

Corollary 5 *With the assumptions of Corollary 4 and Lemma 3, it holds $\mathbb{P}$-almost surely*

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

Note that in contrast to Lemma 3, this corollary gives a precise limit result.

The results cannot be extended to the entire $\mathbb{R}$, because of the value $f(x)$ in the denominator. In the tails, the kernel estimator is in the domain of attraction of a Poisson and not of a normal law. If one excludes $f(x)$, however, the limit results would be imprecise.

In the course of our discussion, the bandwidth was identified as the main obstacle to applications. One attempt to reduce the impact of the bandwidth on performance was to allow the bandwidth to depend on the point of estimation x. For example, in the 1970s Wagner used a nearest-neighbor bandwidth (Wagner 1975). However, even with the fixed bandwidth, the need to select the bandwidth h_n for a fixed sample size n even impelled Parzen (1962) to propose a rule. The idea was to let the bandwidth depend on the data. Weißbach et al. (2008) combined both concepts to achieve good performance in (medical) practice. However, the property of the bandwidth to vary now with respect to more than the sample size raises the question of how such variable bandwidths affect the consistency of the kernel estimate.

As the above results can be extended to the multivariate case, similarly, see e.g. Stute (1984), the development has turned to a simultaneous analysis, which is taken up in the next section.

3.2 Rates of convergence with an inequality by Talagrand

From a mathematical point of view, the usefulness of a bandwidth that depends not only on n, but also on the point x, and furthermore on the data, was found by Stute while studying the mean square error (Stute 1982a):

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

For a larger $f(x)$, h_n should become larger, so as to reduce the variance and hence the error. There is a wide range of data-adaptive bandwidth selectors with the aim of minimizing the mean integrated square error. For instance, Silverman (1986, p.45) assumed a Gaussian distribution, Hall (1978) started using cross-validation and Hall et al. (1991) concluded that plug-in estimation is optimal. Einmahl and Mason (2005), however, do not look at a specific method, but aim at obtaining a general result about the almost sure uniform convergence of the kernel density estimator, if the bandwidth varies within an interval $[a_n, b_n]$. In this section, we consider, as did Einmahl and Mason, the multivariate estimation on $\mathbb{R}^d$ with the Borel σ-field $\mathfrak{B}^d$.

The univariate Definition 2 for the fixed bandwidth has to be extended to a multivariate definition, with a variable bandwidth playing the role of an additional variable, almost like x, that is subject to the supremum norm.

Definition 5 (Almost sure convergence with variable bandwidth) 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

$$\lim_{n\to\infty} \sup_{a_n \le h_n \le b_n} \sup_{x\in\mathbb{R}^d} |f_n(x) - f(x)| =: \lim_{n\to\infty} \sup_{a_n \le h_n \le b_n} ||f_n - f||_\infty = 0$$

with probability 1.

Even though the definition allows for variability of h_n, it has to be admitted that the definition is quite restrictive. For given n, the bandwidth must be in the interval $[a_n, b_n]$. This can be achieved for a bandwidth that varies in x. In contrast, data-adaptive bandwidths depending on $X_1, \ldots, X_n$ rarely remain in a fixed interval. The lesson learnt from studying the mean square error, that the bandwidth should be adapted to $f(x)$, requires an estimation of the density, a priori, with a pilot estimate in order to obtain a reasonable bandwidth. For instance, the nearest-neighbor bandwidth relies on estimating the density with the dirac estimator, i.e. with the empirical distribution function, see Dette and Gefeller (1995). Ideally, Definition 5 should have covered bandwidth sequences that lie only asymptotically in the desired interval, i.e.

$$\lim_{n\to\infty} \mathbb{P}(a_n \le h_n \le b_n) = 1.$$

Some techniques have even been developed to account for such bandwidth sequences, such as those with plug-in-estimators from Deheuvels and Mason (2004), or for the nearest-neighbor bandwidth from Schäfer (1986) and Weißbach (2006). The nearest-neighbor techniques used to prove almost sure uniform convergence, however, must restrict x to a bounded support. Those techniques are similar to Theorem 2, making use only of the Bennett-Hoeffding inequality, instead of the Smirnov-inequality.

In the following, we restrict ourselves to Definition 5, in order to obtain consistency on the unbounded support of f.

One can obtain strong consistency with a variable bandwidth also with the methods from W. Stute, but in the following, we will present the result from 2005 by U. Einmahl and D.M. Mason. We will shortly come back to Stute in the end of the section.

A key element of the approach is to construct a set of functions, generated by a kernel K and indexed in x and h_n.

$$\mathfrak{K} := \left\{ K\left(\frac{x - \cdot}{h_n^{\frac{1}{d}}}\right) : h_n > 0, x \in \mathbb{R}^d \right\}. \tag{8}$$

The research by Einmahl and Mason was only possible because of earlier research by M. Talagrand dealing with bounds for the supremum of empirical processes. He chose a more general approach than Stute, who restricted himself to the empirical process for the set of functions $g_x(\cdot) = I_{(-\infty,x]}(\cdot)$ (indexed in x). In contrast, Talagrand (1994) considers a general set of functions $\mathfrak{G}$ and the generalization of $||(F_n(x) - F(x))||_\infty$

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

The inequality of interest, now with a different set of functions as compared to $g_x(\cdot) = I_{(-\infty,x]}(\cdot)$ in (4), is

$$r_{\mathfrak{G}}(\lambda) := \mathbb{P}(||\alpha_n||_{\mathfrak{G}} \ge \lambda).$$

This probability has been studied in various articles. Einmahl and Mason use a result for the subset of $\mathfrak{K}$, where the bandwidth h lies in a specific interval.

The motivation for 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, with random variables as indices. The expected value of the kernel, indexed at the random variables, is $\mathbb{E}(f_n(x))$. Consequently, if we divide $||\alpha_n||_{\mathfrak{G}}$ by $\sqrt{n}$, for a general set of kernels $\mathfrak{G}$, through this inequality we achieve an approximation of the stochastic component over all kernels of this set. If, for example, the point x and the bandwidth varies in a set of kernels, we obtain a bound for the supremum of all points and all bandwidths. It is then possible to let the bandwidth vary, as Einmahl and Mason prove.

We now define a random variable centering the $g(X_i)$:

Definition 6 (Rademacher-variables) A Rademacher-variable is a discrete random variable ϵ_i with $P(\epsilon_i = -1) = P(\epsilon_i = 1) = \frac{1}{2}$.

We wish to introduce a sequence $\epsilon_1, \dots, \epsilon_n$ of independent Rademacher-variables which are independent of the $X_1, \dots, X_n$. We consider the expression $\sup_{g \in \mathfrak{G}} |\sum_{i=1}^n \epsilon_i g(X_i)|$ which is similar to $||\sqrt{n}\alpha_n||_{\mathfrak{G}}$. Both processes are centered, if we do not use the absolute value. However, the advantage of the symmetrization is that we can study this process more efficiently, as we will see below.

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 its expected value is well defined. The pointwise measurability of $\mathfrak{G}$ is sufficient for this purpose. 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 is a sequence of functions g_m of $\mathfrak{G}_0$ with

$$\lim_{m \to \infty} g_m(x) = g(x) \; \forall x \in \mathbb{R}^d.$$

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. Parzen, Nadaraja, or Stute did not consider this issue.

Einmahl and Mason use a central theorem about the bound of the probability that $||\sqrt{n}\alpha_n||_{\mathfrak{G}}$ differs considerably from the expected value $\mathbb{E}(\sup_{g \in \mathfrak{G}} |\sum_{i=1}^n \epsilon_i g(X_i)|)$. The theorem is basically due to Talagrand (1994, p. 45). We can imagine this as a generalization of Tchebychev's inequality. $||\alpha_n||_{\mathfrak{G}}$ is measurable, if $\mathfrak{G}$ is measurable pointwisely.

Lemma 4 (Talagrand) *Let $\mathfrak{G}$ be measurable pointwisely with $\sup_{g \in \mathfrak{G}} ||g||_\infty \leq M < \infty$, $\sigma_{\mathfrak{G}}^2 := \sup_{g \in \mathfrak{G}} Var(g(X)) < \infty$ and let A_1, A_2 be some constants. Then, for each $t > 0$*

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

Remarks The formal similarity to Tchebychev's inequality is that we obtain an upper bound (in a generalized sense) for the probability that a random variable differs strongly 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 requires the second moment to be finite. Talagrand achives this result for the empirical processes, based on similar results for Gaussian processes. We will point out the connection later.

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

Lemma 5 (Bound for the expected value for general empirical processes) *Let $\mathfrak{G}$ be a pointwise measurable set of bounded functions with the assumptions*

1. *$\exists G : \mathbb{R}^d \to \mathbb{R}$ with $G(x) \geq \sup_{g \in \mathfrak{G}} |g(x)| \; \forall x \in \mathbb{R}^d$*

2. $\mathbb{E}((G(X))^2) \leq \zeta^2$
3. $N(\epsilon, \mathfrak{G}) \leq C\epsilon^{-\nu}$ *for all* $0 < \epsilon < 1$
4. $\sigma_0^2 := \sup_{g \in \mathfrak{G}}((g(X))^2) \leq \sigma^2$
5. $\sup_{g \in \mathfrak{G}} ||g||_\infty \leq M$

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

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

Remark The third condition was not necessary in Talagrand's inequality for the probability of the distance from the expected value. For the proof of this moment inequality, it is necessary however. Being a topological entropy-condition about the kernel set, it must not be too rich.

The conditions discussed in this section must be true only for some subsets of $\mathfrak{K}$. Yet, we assume that they are true for the entire 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 ϕ is a right-continuous function with bounded variation. For $d = 1$, the measurability condition is true whenever K is right-continuous. 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 about the kernels that Nadaraja and Stute needed, too.

The conditions guarantee that the generalized empirical process α_n converges for $n \to \infty$ against the Brownian Bridge, a special Gaussian process. Accordingly, it behaves "normally" and thus, the maximal expected value can be bounded from above. The fact that many empirical processes converge to Gaussian processes was Talagrand's motivation.

Einmahl and Mason, in the first instance, consider only the stochastic component, just like Stute. We introduce the assumptions

(D1) $K : \mathbb{R}^d \to \mathbb{R}$, $K \in L^\infty(\mathbb{R}^d)$ with $\int_{\mathbb{R}^d} K(x)dx = 1$.
(D2) f is bounded.
(D3) The assumption of the polynomial covering number holds for (8).
(D4) The pointwise measurability is satisfied for (8).

Theorem 6 (Almost sure uniform consistency with variable bandwidth) *With (D1) - (D4) we have for every* $c > 0$ $\mathbb{P}$*-almost surely*

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

Proof We outline the proof. Let h_n be called h and we introduce two real sequences. For $j, k \geq 0$ and $c > 0$ define $n_k = 2^k$ and $h_{j,k} = \frac{c\ 2^j\ \log(n_k)}{n_k}$. The c is fixed, j and k will vary in the proof. Define, in addition, the kernel set which fulfills the conditions of Lemma 4 and 5:

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

We obtain two bounds upwards of 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 be bounded, so that the former bound can become small for vanishing 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 bound the expression

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

and obtain for a large k

$$\begin{aligned}\mathbb{E}\left(\sup_{g\in\mathfrak{K}_{j,k}}\left(\sum_{i=1}^{n_k}\epsilon_i g(X_i)\right)\right) &\leq D_3\sqrt{n_k h_{j,k}\log\frac{1}{D_2 h_{j,k}}} \\ &\leq D_3\sqrt{n_k h_{j,k}\max\left(\log\frac{1}{D_2 h_{j,k}}, \log\log n_k\right)} =: D_3 a_{j,k} \qquad (9)\end{aligned}$$

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

On $\mathfrak{K}_{j,k}$,we apply Talagrand's inequality with $M = \kappa$ and $\sup_{g\in\mathfrak{K}_{j,k}} Var(g(X)) \leq \sup_{g\in\mathfrak{K}_{j,k}} \mathbb{E}(g(X)^2) = \sigma_0^2 \leq D_0 h_{j,k}$. For each $t > 0$, we have

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

Minding this, we bound 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

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

and can show that, for a large k,

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

For this bound the term $\log\log n_k$ in (9) is very helpful.

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 $h_{j,k}$ runs 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 a small k and $n_{k-1} \leq n \leq n_k$, we can show

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

$\mathbb{P}$ is a probability measure, so that $\mathbb{P}(A_k(\rho)) \leq \mathbb{P}_k(\rho)$. The infinite sum over the right expression converges, so that the series over $\mathbb{P}(A_k(\rho))$ also converges. The Borel-Cantelli-Lemma now states that the probability of the limit superior of the set sequence $A_k(\rho)$ is 0. With probability 1, only a finite number of elements of the in $A_k(\rho)$ formulated sequence of random variables are larger than $A_1(D_3+\rho)$. Consequently, the limit superior of the sequence is finite with probability 1. The exact limit is unknown, but depends on c. □

Let, for the moment, $K^*(c) > 0$ and consider the case that we take the supremum over values h_n such that Stute's assumption (A3) holds. With it, $-\log h_n$ increases faster than $\log\log n$ and so

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

for sufficiently large n. Accordingly, Einmahl and Mason's theorem yields a convergence rate of $\sqrt{\frac{nh_n}{-\log h_n}}$ - this is Stute's convergence rate. However, the authors did not calculate an exact limit. To obtain this, it would probably be necessary to include a variance term in the denominator, as Stute did.

A result like Theorem 6 can also be proved with methods that Stute used in his papers. We show this at first for the rectangular kernel in one dimension, comparably to Theorem 3. For this let $a_n \leq h_n \leq b_n$ and F Lipschitz-continuous on $\mathbb{R}$ with Lipschitz-constant $m^* < \infty$. Uniformly in x, we have

$$\begin{aligned} |f_n(x) - \mathbb{E}(f_n(x))| &= h_n^{-1}n^{-\frac{1}{2}} \left| \alpha_n(I_{(-\infty,(x+\frac{h_n}{2})]}(\cdot)) - \alpha_n(I_{(-\infty,(x-\frac{h_n}{2})]}(\cdot)) \right| \\ &= h_n^{-1}n^{-\frac{1}{2}} \left| \gamma_n\left(F\left(x + \frac{h_n}{2}\right)\right) - \gamma_n\left(F\left(x - \frac{h_n}{2}\right)\right) \right| \\ &\leq h_n^{-1}n^{-\frac{1}{2}}\omega_n(m^* h_n). \end{aligned}$$

By monotonicity, the last term is further bounded from above by

$$a_n^{-1}n^{-\frac{1}{2}}\omega_n(m^* b_n).$$

If Assumptions (A1)-(A3) are fulfilled for b_n, we get with (6) that

$$\omega_n(m^* b_n) = O(\sqrt{b_n(\log b_n)})$$

and finally

$$||f_n(x) - \mathbb{E}(f_n(x))||_\infty = O(a_n^{-1}n^{-\frac{1}{2}}(\log b_n)^{\frac{1}{2}} b_n^{\frac{1}{2}}).$$

With suitable conditions on a_n and b_n the last term then tends to 0.
For general kernels, a result like this follows by adapting the proof of Lemma 3. With the assumptions of this lemma (with (A1)-(A3) holding for b_n), we get uniformly for $x \in \mathbb{R}$ (see the proof of (7))

$$|f_n(x) - \mathbb{E}(f_n(x))| \leq h_n^{-1}n^{-\frac{1}{2}}\omega_n(m^*(s-r)h_n)V_r^s(K).$$

With the same argumentation as for the naive kernel, we arrive at

$$||f_n(x) - \mathbb{E}(f_n(x))||_\infty = O(a_n^{-1} n^{-\frac{1}{2}} (\log b_n)^{\frac{1}{2}} b_n^{\frac{1}{2}}).$$

Note that an identical result like Theorem 6 cannot be proved with this method, anyway, because in Theorem 6 we have $b_n = 1$. This sequence is not a null sequence and thus does not fulfill the assumptions of Lemma 3.

4 Summary

In this survey article we compare different proofs for the consistency of kernel density estimators. We focus on the almost sure convergence, which is stronger than uniform convergence in probability. We see a clear historical development: While considering global properties of the empirical distribution functions is enough for strong consistency, proofs of exact convergence rates use deeper information about the underlying empirical processes. We give a survey on two different empirical process approaches, one by Stute and one by Einmahl/Mason, the latter being a generalization.

In general, the stochastic component of the error between f_n and f is a greater problem than the deterministic one. We focus on the stochastic component and treat the bias secondarily. For bounding the stochastic part, i.e. in order to find almost sure bounds of the stochastic process, we especially need assumptions about the kernels with a recurrent assumption being right-continuity and bounded variation. Often, we also need boundedness of f. The deterministic component requires assumptions about the smoothness of f, e.g. uniform continuity.

There is surely considerable potential for more research on the theory of consistency. It would be useful to consider cases in which the data-adaptive bandwidth lies almost surely in an interval $[a_n, b_n]$ for a sufficiently large n. The strong consistency result by Einmahl and Mason could be applied in practice only then. With the described condition of the data-adaptive bandwidth that is less strong, the consistency result is also useful, but there may still be potential for improvement.

Acknowledgement: We would like to thank U. Einmahl for pointing out some significant references and the editor and an anonymous referee for their valuable advice. Financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures" and SFB 823 "Statistical modelling of nonlinear dynamic processes") is gratefully acknowledged.

References

Büning, H. and Trenkler, G. (1994). *Nichtparametrische Statistische Methoden.* de Gruyter, Berlin, 2nd edition.

Deheuvels, P. and Mason, D. (2004). General asymptotic confidence bands based on kernel-type function estimators. *Statistical Inference for Stochastic Processes*, 7:225–277.

Dette, H. and Gefeller, O. (1995). Definitions of nearest neighbour distances for censored data on the nearest neighbour kernel estimators of the hazard rate. *Journal of Nonparametric Statistics*, 4:271–282.

Diehl, S. and Stute, W. (1988). Kernel density and hazard function estimation in the presence of censoring. *Journal of Multivariate Analysis*, 25:299–310.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Annals of Mathematical Statistics*, 27:642–669.

Einmahl, U. and Mason, D. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *Annals of Statistics*, 33:1380–1403.

Hall, P. (1978). Cross-validation in density estimation. *Biometrika*, 69:383–390.

Hall, P. and Marron, J. (1995). Improved variable window kernel estimates of probability densities. *Annals of Statistics*, 23:1–10.

Hall, P., Sheather, S., Jones, M., and Marron, J. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika*, 78:263–269.

Härdle, W. (1991). *Smoothing Techniques.* Springer, New York.

Marron, J., Gonzàlez-Manteiga, W., and Cao, R. (1996). Bootstrap selection of the smoothing parameter in nonparametric hazard rate estimation. *Journal of the American Statistical Association*, 91:1130–1140.

Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Annals of Probability*, 18:1269–1283.

Nadaraja, E. (1964). On estimating regression. *Theory of Probability and its Applications*, 9(1):141–142.

Nadaraja, E. (1965). On nonparametric estimates of density functions and regression curves. *Theory of Probability and Applications*, 10:186–190.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27:832–835.

Schäbe, H. and Tiedge, J. (1995). Kernel estimation for characteristics of pure jump processes. *Statistical Papers*, 36:131–144.

Schäfer, H. (1986). Local convergence of empirical measures in the random censorship situation with application to density and rate estimators. *Annals of Statistics*, 14:1240–1245.

Silverman, B. (1986). *Density Estimation.* Chapman & Hall, London.

Stute, W. (1982a). A law of the logarithm for kernel density estimators. *Annals of Probability*, 10:414–422.

Stute, W. (1982b). The oscillation behavior of empirical processes. *Annals of Probability*, 10:86–107.

Stute, W. (1984). The oscillation behavior of empirical processes: The multivariate case. *Annals of Probability*, 12:361–379.

Talagrand, M. (1994). Sharper bounds for gaussian and empirical processes. *Annals of Probability*, 22:28–76.

Wagner, T. (1975). Nonparametric estimates of probability densities. *IEEE Transactions on Information Theory*, 21:438–440.

Wand, M. and Jones, M. (1995). *Kernel Smoothing.* Chapman and Hall, London.

Watson, G. (1964). Smooth regression analysis. *Sankhya Series A*, 26:101–116.

Weißbach, R. (2006). A general kernel functional estimator with general bandwidth - strong consistency and applications. *Journal of Nonparametric Statistics*, 18:1–12.

Weißbach, R., Pfahlberg, A., and Gefeller, O. (2008). Double-smoothing in kernel hazard rate estimation. *Methods of Information in Medicine*, 47:167–173.

Wheeden, R. and Zygmund, A. (1977). *Measure and Integral - An Introduction to Real Analysis.* Marcel Dekker, New York.