Aspects of Nonparametric Density Estimation

Bert van Es

Aspects of Nonparametric Density Estimation

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam,
op gezag van de Rector Magnificus
prof. dr. S. K. Thoden van Velzen,
in het openbaar te verdedigen in de Aula der Universiteit
(Oude Lutherse Kerk, ingang Singel 411, hoek Spui),
op woensdag 2 november 1988 te 13.30 uur.

door

Albertus Jacob van Es

geboren te Amsterdam

Promotor	:	Prof. Dr. P. Groeneboom.
Co-promotor	:	Dr. P. L. J. Janssen.

Faculteit der Wiskunde en Informatica.

To my mother and the memory of my father.

Preface.

Afterwards one of the interesting things of preparing a thesis is that you go through stages of different ways of doing scientific research. The first part of the thesis, concerning bandwidth selection for kernel estimators, took place at a fair distance of the experts in this particular part of density estimation. One of the consequences is that at times you worry whether *they* haven't already written up your results. On the other hand the research on deconvolution and the Wicksell problem received so much attention from colleagues nearby, working on related problems, that one of your main worries is why *you* haven't already written up your results. For a young statisticians selfconfidence in my view the second situation is preferable.

I started the research for a thesis while I was working at the Centre for Mathematics and Computer Science (CWI) in Amsterdam. Later I worked at the Mathematical Institute of the University of Amsterdam. I thank both institutes for allowing me to do this research and for the facilities they offered.

Some people I want to mention specially. I want to thank my promotor Piet Groeneboom for a long period of pleasant cooperation. Although the circumstances have not always been optimal I feel confident that in the near future we can prolong the interesting research related to the problems in the last chapter of this thesis.

I am grateful to my co-promotor Paul Janssen for the effort he put into reading the manuscript in the limited time there was. It has benefitted a lot from his remarks.

With Peter de Jong I had many discussions on some theoretical aspects of this research, i.e. the asymptotic distribution of statistics appearing in the chapters about kernel estimation.

I thank Richard Gill and Adriaan Hoogendoorn for suggesting the Wicksell problem, which inspired the research on the deconvolution problem.

Without Te Yung Fu writing (typing) this thesis would have required even more effort, since then I could not have avoided reading all those manuals. I thank him for his technical guidance.

I am grateful to Dick Zwarst and his team for fitting the printing of this monograph in their tight schedule.

Finally I thank Marian for putting up with someone showing increasing signs of obsession. Hopefully in the near future these are not replaced by signs of increase by another obsession.

Contents:

1. INTRODUCTION.

If $X_1,...,X_n$ are independent observations from a distribution with density f, then one of the oldest nonparametric estimators of the density is the *Parzen-Rosenblatt kernel estimator* (Parzen (1962), Rosenblatt (1956))

$$(1.1) \qquad f_{nh}(x) := \frac{1}{nh}\sum_{i=1}^{n} K((x-X_i)/h),$$

where h is a positive real number called the *window* or the *bandwidth* and K is a probability density function called the *kernel*. This estimator is studied in the first two chapters. Of course many other nonparametric density estimators have been proposed, for instance the well known histogram estimator and several refinements of the kernel estimator. For reviews of density estimation we refer to Prakasa Rao (1983), Devroye & Györfi (1985), Silverman (1986) and Devroye (1987).

To compute a kernel estimate we have to choose a kernel and a bandwidth. It is generally recognized that for most loss functions the choice of the bandwidth is more important than the choice of the kernel. In chapter 3 we consider so called cross-validation methods to determine a good bandwidth for a kernel estimator. *Least squares cross-validation* was introduced and studied by Rudemo (1982) and Bowman (1984) and has since received considerable attention. Stone (1984) established an important optimality result with respect to the integrated squared error. It states that a kernel estimator with a bandwidth computed by least squares cross-validation asymptotically performs as well as a kernel estimator with the best possible non-random bandwidth. This optimality holds for all bounded densities. For smooth densities, i.e. essentially densities with a continuous second derivative, the asymptotic distribution of the computed bandwidths and the corresponding integrated squared error was derived by Hall & Marron (1987b). *Likelihood cross-validation* was introduced earlier by Habbema, Hermans & Van de Broek (1974) and Duin (1976). We establish the almost sure rates of convergence to zero of bandwidths computed by this method and the asymptotic distribution theory. It turns out that the asymptotic behavior is similar to the asymptotic behavior of least squares cross-validation, provided we use a modification proposed by Marron (1985), and provided we *exclude densities with jumps*. We show that likelihood cross-validation does not give asymptotically optimal bandwidths for densities with jumps. For densities without jumps likelihood cross-validation gives bandwidths which are asymptotically optimal with respect to a weighted integrated squared error, where the weight is equal to 1/f (Marron (1985)). For a detailed introduction to cross-validation methods we refer to section 3.1.

Following Van Eeden (1985) and Cline & Hart (1986) we consider not only estimation of smooth densities, but also of densities with discontinuity points. We also allow discontinuity points in the first or second derivative. At those points we require the densities to have left and right Taylor expansions. Thus we also consider densities with *jumps* and *kinks*. In section 2.1 we state the precise conditions on f. For the moment we suffice with giving two examples.

Example 1.1. Let the density f be given by

$$(1.2) \qquad f(x) := \begin{cases} 0 & \text{if } x<0 \\ (2-x/2)/\alpha & \text{if } 0\leq x<2 \\ (2-(x-3)^2)/\alpha & \text{if } 2\leq x<3+\sqrt{2} \\ 0 & \text{if } x\geq 3+\sqrt{2} \end{cases},$$

where $\alpha := \int_0^2 (2-\frac{1}{2}x)dx + \int_2^{3+\sqrt{2}} (2-(x-3)^2)dx \approx 6.5523$. Then f has a jump in the point 0, a kink in the point 2 and a kink in the point $3+\sqrt{2}$.

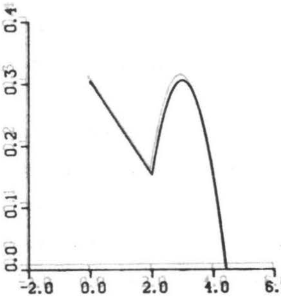

Figure 1.1. A non-smooth density.

We use this density repeatedly as an example of a typical non-smooth density.

Example 1.2. A situation where a jump and a kink appear naturally is given by *Wicksell's corpuscle problem*. Let $X_1, \dots, X_n$ denote n radii of spheres of different random size in an opaque medium, such as drops of oil in a piece of rock. Suppose that we can not observe these spheres directly. Instead we can observe the radii of the circular profiles of the spheres obtained by taking a slice of the medium. Denote the radii of the circular profiles by $Y_1, \dots, Y_n$, which we assume to be independent. Defining f as the sphere radius density and g as the circle radius density, Wicksell (1925) showed that under suitable regularity conditions the next relations between f and g hold,

$$(1.3) \qquad g(y) = \frac{1}{\mu} \int_y^\infty \frac{y}{\sqrt{r^2 - y^2}} f(r)\, dr,\ 0 < y < \infty$$

and

$$(1.4) \qquad f(r) = -\frac{2\mu}{\pi} \frac{d}{dr} \int_y^\infty \frac{r}{\sqrt{y^2 - r^2}} g(y)\, dy,\ r\geq 0,$$

where μ equals the expectation of the sphere radii. Several parametric and nonparametric methods have been proposed for estimating the density f or its distribution function. For reviews of the

Wicksell problem and related methods see Ripley (1981) and Stoyan, Kendall & Mecke (1987). Estimators of the density f related to the kernel estimator were proposed by Taylor (1983), Hall & Smith (1988) and Van Es & Hoogendoorn (1988). All these estimators suffer from a large bias close to zero, which can be explained from the fact that , no matter how smooth the density f is, the density g has a kink in zero. This is immediate from relation (1.3). Moreover, Hall & Smith propose an estimator based on the squared circle radii. It is readily seen that since the density of the squared circle radii equals $g_1(r) := (2r^{1/2})^{-1}g(r^{1/2})$ it has a jump in zero.

Kernel estimation of non-smooth densities is studied extensively in chapter 2. In our opinion kernel estimators can be used for estimating such densities, even though they have a larger error and thus require larger sample sizes. Moreover, these densities might occur without the statistician being aware of it. For this reason we have also studied likelihood cross-validation for such non smooth densities. In fact, in an important special case, treated in corollary 3.6, the density f has jumps. Some of the results derived in this chapter are also used frequently in chapter 3.

In the last chapter we leave density estimation and consider *deconvolution*, i.e. estimation of an unknown distribution function in a situation where we have a sample from a distribution which is the convolution of the unknown distribution and a known one. Since the Wicksell problem, properly transformed, also has a convolution structure, estimation of the distribution function of the sphere radii is one of the examples. We present a minimax theorem which shows that even for estimating a distribution function at one fixed point different rates of convergence can appear, a phenomenon well known in density estimation. Also, for three examples, we derive the nonparametric maximum likelihood estimator of the distribution function. An extended version of this chapter will appear separately as a joint report with P. Groeneboom.

2. KERNEL ESTIMATION IN NON-SMOOTH CASES.

We examine the performance of the kernel estimator (1.1) with the emphasis on its properties if $X_1, ..., X_n$ is a sample from a distribution with a density f which does not satisfy the usual smoothness conditions. Under these conditions f is essentially required to have two continuous derivatives. While the results for the smooth case date back to Rosenblatt (1956), studies on the behavior in non-smooth cases, allowing discontinuities in f and its derivative, are fairly recent, see for example Van Eeden (1985) and Cline & Hart (1986).

The conditions we impose on f and K are given in section 2.1. In section 2.2 we discuss the basic properties of the kernel estimator f_{nh}, evaluated at a fixed point x of the real line. The results presented in this section are needed to derive global properties of kernel estimators in later sections. They also have independent interest. The global behavior with respect to the integrated squared error and the supremum distance is treated in sections 2.3 and 2.4. For the properties of kernel estimators with respect to the L_1 norm we refer to Devroye & Györfi (1985) and Devroye (1987). The last section of this chapter contains technical (parts of) proofs of results in the preceding sections.

2.1. Assumptions.

We consider densities satisfying the following conditions. Essentially we allow the densities to have jumps and kinks. A typical example is given in figure 1.1 in the introduction.

Condition F:

(F.1) *The first and second derivatives of* f, *denoted as* f' *and* f ", *exist at every point of the real line, except at a countable set of points which we denote as* D. *In these points we give* f' *and* f " *arbitrary values . We assume that* $\inf\{|d_1-d_2| : d_1,d_2 \in D\}$ *is positive, i.e. the points in D are separated.*

(F.2) *The functions* f, f' *and* f " *have finite left and right limits at the points in* D.

(F.3) *The function* f *has finite left and right first and second derivatives at the points in* D.

(F.4) *The second derivative* f" *is continuous on the open intervals between the points in* D.

The elements of the set D are called *singular points*. For the density of example 1.1 the set D is equal to $\{0,2,3+\sqrt{2}\}$. The jumpsizes of f, f ' and f " in the singular points are denoted by $\delta^{(0)}(d)$, $\delta^{(1)}(d)$ and $\delta^{(2)}(d)$, so we have

$$\delta^{(0)}(d) := f(d+) - f(d-),$$

$$\delta^{(1)}(d) := f'(d+) - f'(d-),$$

$$\delta^{(2)}(d) := f''(d+) - f''(d-).$$

Condition (F.3) needs some further explanation. By the existence of a finite right derivative of f at $d \in D$ we mean that the limit

$$\lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon}(f(d+\varepsilon) - f(d+))$$

exists and is finite. By Taylor's theorem and (F.2) this limit equals f '(d+). By the existence of the second right derivative of f at $d \in D$ we mean that the limit

$$\lim_{\varepsilon \downarrow 0} \frac{2}{\varepsilon^2}(f(d+\varepsilon) - f(d+) - \varepsilon f'(d+))$$

exists and is finite. This limit then equals f "(d+). Hence

$$\lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon^2}(f(d+\varepsilon) - f(d+) - \varepsilon f'(d+) - \tfrac{1}{2}\varepsilon^2 f''(d+)) = 0.$$

The left derivatives are defined similarly. With left limits replacing the right limits the relation above also holds for f(d-ε). This means that we can use left and right Taylor expansions in the singular points.

Given the fact that we use a probability density, the choice of kernel is relatively unimportant. Hence we feel free to consider bounded support kernels only. This is further motivated in section 2.3.1. We assume that the kernels satisfy the next condition.

Condition K:

(K.1) K *is a probability density function.*

(K.2) K *has support* [-1,1].

(K.3) K *is bounded.*

(K.4) K *is symmetric.*

With respect to (K.2) note that

$$\frac{1}{nh}\sum_{i=1}^{n} K((x-X_i)/h) = \frac{1}{nh_c}\sum_{i=1}^{n} K_c((x-X_i)/h_c), \tag{2.1}$$

where $h_c = c^{-1}h$ and $K_c(x) := cK(cx)$, for all x. This implies that to study the case of kernels with bounded support we can restrict attention to the support [-1,1], without loss of generality.

2.2. Basic properties of the kernel estimator.

Since $f_{nh}(x)$ is an average of i.i.d. random variables its expectation is given by

$$\text{(2.2)} \qquad E\, f_{nh}(x) = E\,\frac{1}{h}\,K((x-X_1)/h)\ .$$

To compute the variance note that a straightforward computation gives

$$E\, f_{nh}^2(x) = \frac{1}{nh}\,E\,\frac{1}{h}\,K^2((x-X_1)/h)) + \frac{n-1}{n}\,(E\,\frac{1}{h}\,K((x-X_1)/h))^2\ ,$$

and therefore

$$\text{(2.3)} \qquad \text{var}\,(f_{nh}(x)) = \frac{1}{nh}\,E\,\frac{1}{h}\,K^2((x-X_1)/h)) - \frac{1}{n}(E\,\frac{1}{h}\,K((x-X_1)/h))^2\ .$$

Also note that the expectation (2.2) depends on the bandwidth but not on the sample size . The variance depends on both the sample size and the bandwidth. A further observation is that for $Ef_{nh}(x)$ to converge to f(x) we have to assume that h tends to zero.

The expectations appearing in (2.2) and (2.3) are of the same form. They can be written as

$$g(x,h) := E\, G_h(x-X_1),$$

where G is a measurable function, not necessarily a density, and G_h is defined by

$$G_h(x) := \frac{1}{h}G(\frac{x}{h}).$$

In (2.3) we take G equal to K^2 for the first term and equal to K for the second term. This shows the necessity of expansions of such quantities for bandwidths h tending to zero. The next lemma consists of two parts. Suppose that f satisfies condition F and recall that that D denotes the set of singular points of f. The first part gives an expansion for g(x,h) in terms of the bandwidth with x a fixed point in D_h, where

$$\text{(2.4)} \qquad D_h := \{x: |x-d|>h \text{ for all } d\in D\},$$

i.e. the set of all points of the real line which are at least at distance h of the singular points of f. In example 1.1 the set D is equal to $\{0,2,3+\sqrt{2}\}$. The set D_h is equal to the following union of intervals,

$$D_h = (-\infty,-h)\cup(h,2-h)\cup(2+h,3+\sqrt{2}-h)\cup(3+\sqrt{2}+h,\infty).$$

For technical reasons we also establish the uniformity of the expansion over the sets $D_h \cap[-M,M]$ for arbitrary positive integers M. The second part of the lemma gives an expansion of g(x,h) for x in a shrinking neighborhood of some fixed point x_0. Here we consider points $x=x_0+th$ and we let h tend to zero. The expansion holds uniformly on bounded t-intervals. Furthermore we prove uniformity of these expansions for the bandwidths h in intervals $(0,h_n']$, where (h_n') is a fixed sequence of real numbers satisfying

$$h_n'>0 \text{ for } n = 1,2,\ldots \text{ and } \lim_{n\to\infty} h_n' = 0.$$

Lemma 2.1. *Let* G *denote a bounded measurable function with support* [-1,1] *and let* X *denote a random variable having a distribution with density* f. *Suppose that* f *satisfies condition* F.
(a) Then

$$g(x,h) = E\, G_h(x-X) =$$

$$f(x)\int_{-1}^{1} G(u)du - hf'(x)\int_{-1}^{1} uG(u)du + \tfrac{1}{2}h^2f''(x)\int_{-1}^{1} u^2G(u)du + r_1(x,h), \tag{2.5}$$

where the remainder r_1 *satisfies*

$$\lim_{n\to\infty}\ \sup_{0<h\le h_n'}\ \sup_{x\in D_h\cap[-M,M]} h^{-2}r_1(x,h) = 0, \tag{2.6}$$

for every positive M.
(b) For x_0 *a fixed point we have*

$$g(x_0+th,h) = E\, G_h(x_0+th-X) =$$

$$f(x_0-)\int_{-\infty}^{0} G(t-u)du + f(x_0+)\int_{0}^{\infty} G(t-u)du +$$

$$h(f'(x_0-)\int_{-\infty}^{0} uG(t-u)du + f'(x_0+)\int_{0}^{\infty} uG(t-u)du) + \tag{2.7}$$

$$\tfrac{1}{2}h^2(f''(x_0-)\int_{-\infty}^{0} u^2G(t-u)du + f''(x_0+)\int_{0}^{\infty} u^2G(t-u)du) + r_2(t,h),$$

where the remainder r_2 *satisfies*

$$\lim_{n\to\infty}\ \sup_{0<h\le h_n'}\ \sup_{-M\le t\le M} h^{-2}r_2(t,h) = 0, \tag{2.8}$$

for every positive M. □

Proof. By a substitution we obtain

$$g(x+th,h) = E\, G_h(x+th-X) =$$

$$(2.9) \qquad \int_{-\infty}^{\infty} \frac{1}{h} G(\frac{x+th-v}{h}) f(v)dv = y$$

$$\int_{-\infty}^{\infty} G(t-u)f(x+hu)du \; .$$

To show (a) we take t equal to zero and we assume that x lies in the set D_h. Relation (2.9) then becomes

$$g(x,h) = \int_{-1}^{1} G(-u)f(x+hu)du \; .$$

Since $x \in D_h$ the interval [x-h,x+h] contains no points of D and hence by condition F the function f allows a three term Taylor expansion for f(x+hu) around the point x. We get

$$g(x,h) = \int_{-1}^{1} G(-u)\{f(x) + huf'(x) + \tfrac{1}{2}h^2u^2f''(x))\}du + r_1(x,h),$$

where r_1 equals

$$(2.10) \qquad r_1(x,h) = \tfrac{1}{2}h^2 \int_{-1}^{1} u^2G(-u)\{f''(\xi(x,hu)) - f''(x))\}du$$

and $\xi(x,hu)$ is the point between x and x+hu appearing in Lagranges version of the remainder term in the Taylor expansion of f(x+hu). In order to complete the proof of part (a) it remains to show (2.6). Let (h_n) be an arbitrary sequence of bandwidths satisfying $0<h_n \leq h_n'$ for all n, and let (x_n) be an arbitrary sequence of points in $D_h \cap [-M,M]$, where M is an arbitrary positive number. It suffices to show

$$(2.11) \qquad \lim_{n\to\infty} h_n^{-2}\, r_1(x_n,h_n) = 0.$$

Under condition F the interval [-M,M] contains a finite number of singular points, $-M \leq d_1 \leq d_2 \leq \ldots \leq d_m \leq M$, say. The second derivative f '' is uniformly continuous on the intervals $[-M,d_1)$, $(d_m,M]$, and (d_i,d_{i+1}), i=1,...,m-1. Since for -1<u<1 the points x_n and $\xi(x_n,h_nu)$ belong to the same interval we have

$$\lim_{n\to\infty} f''(\xi(x_n,h_nu)) - f''(x_n) = 0,$$

so the integrand in (2.10) converges pointwise to zero. By the dominated convergence theorem we then obtain (2.11) and the proof of part (a) is finished.

The proof of part (b) is similar, except that since x_0 is allowed to belong to D we have to use left and right Taylor expansions of $f(x_0+th)$. In fact this is an important special case. The details of the proof of part (b) are given in section (2.5). □

By (2.2) we can now expand $Ef_{nh}(x)$. Since condition K implies that the integral of uK(u) vanishes, taking G equal to K, part (a) of the lemma gives

$$
\begin{aligned}
&E\,f_{nh}(x) = E\,K_h(x-X_1) = \\
&f(x)\int_{-1}^{1}K(u)du - hf'(x)\int_{-1}^{1}uK(u)du + \tfrac{1}{2}h^2f''(x)\int_{-1}^{1}u^2K(u)du + r_1(x,h) = \\
&f(x) + \tfrac{1}{2}h^2f''(x)\int_{-1}^{1}u^2K(u)du + r_1(x,h).
\end{aligned}
$$

By (2.6) this expansion is only meaningful for x a non-singular point of f. On the other hand for x a singular point we can apply part (b) with t=0. We get

$$
\begin{aligned}
&E\,f_{nh}(x) = \\
&\tfrac{1}{2}(f(x-) + f(x+)) + h\delta^{(1)}(x)\int_{0}^{1}uK(u)du + \tfrac{1}{2}h^2(f''(x-) + f''(x+))\int_{0}^{1}u^2K(u)du + r_2(0,h),
\end{aligned}
$$

where r_2 satisfies

$$\lim_{n\to\infty}\ \sup_{0<h\le h_n'} h^{-2}r_2(0,h) = 0.$$

The next two theorems give expansions of the bias $b(x,h):=Ef_{nh}(x)-f(x)$ and the variance of $f_{nh}(x)$. Note that the bias, just as the expectation, is independent of the sample size. It only depends on the bandwidth. Similar to part (b) of lemma 2.1 we give an expansion of $b(x_0+th,h)=Ef_{nh}(x_0+th)-f(x_0+th)$, i.e. for values $x=x_0+th$ close to a point x_0. However, since by condition F the value of f in jumping points is arbitrary, we have to exclude t=0. We first introduce some functions which appear in the expansion of $b(x_0+th,h)$.

Definition 2.2. *The functions* b_0, b_1 *and* b_2 *are defined as*

$$
b_m(t) := \begin{cases} \displaystyle\int_{-\infty}^{t}(t-u)^mK(u)du & \textit{if } t<0 \\[2ex] \displaystyle -\int_{t}^{\infty}(t-u)^mK(u)du & \textit{if } t\ge 0 \end{cases},
$$

for m = 0,1,2.

The next pictures show the graphs of b_0, b_1 and b_2. We have used the kernel

$$K(x) = \tfrac{35}{32}(1-x^2)^3I_{[-1,1]}(x),$$

a symmetric bounded support kernel with support [-1,1].

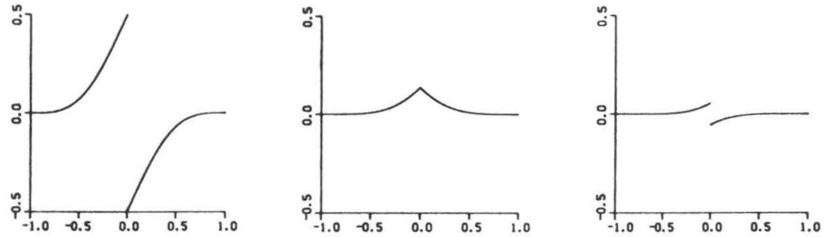

Figure 2.1. The functions b_0,b_1 and b_2.

Theorem 2.3. *Assume that the kernel* K *satisfies condition* K *and that the density* f *satisfies condition* F.
(a) Then

$$b(x,h) = \tfrac{1}{2}h^2 f''(x) \int_{-1}^{1} u^2K(u)du + r_3(x,h) \tag{2.12}$$

where the remainder r_3 *satisfies*

$$\lim_{n\to\infty} \sup_{0<h\leq h_n'} \sup_{x\in D_h\cap[-M,M]} h^{-2}r_3(x,h) = 0,$$

for every positive M.
(b) For x_0 *a fixed point we have*

$$\begin{aligned}b(x_0+th,h) = \; & b_0(t)\delta^{(0)}(x_0) + hb_1(t)\delta^{(1)}(x_0) + \tfrac{1}{2}h^2b_2(t)\delta^{(2)}(x_0) + \\ & \tfrac{1}{2}h^2 \int_{-1}^{1} u^2K(u)du\{f''(x_0-)I_{(-\infty,0)}(t) + f''(x_0+)I_{(0,\infty)}(t)\} + \\ & r_4(t,h),\end{aligned} \tag{2.13}$$

where the remainder r_4 *satisfies*

$$\lim_{n\to\infty} \sup_{0<h\leq h_n'} \sup_{-M\leq t\leq M, t\neq 0} h^{-2}r_4(t,h) = 0,$$

for every positive M. □

Proof. Notice that by (2.2) the expansion in part (a) is a direct consequence of the expansion in part (a) of lemma 2.1 if we choose G equal to K. By the symmetry of K we have

$$\int_{-1}^{1} uG(u)du = \int_{-1}^{1} uK(u)du = 0, \tag{2.14}$$

and so, since K integrates to one, the remainder r_3 is equal to the remainder r_1.
To prove part (b) notice that $b(x_0+th,h)$ equals $Ef_{nh}(x_0+th)-f(x_0+th)$. Again by relation (2.2) the expectation can be expanded using part (b) of lemma 2.1. Together with left and right Taylor expansions of $f(x_0+th)$ around x_0 the result can be derived. The details are left to section (2.5). □

Theorem 2.4. *Assume that the kernel* K *is a bounded probability density with support equal to* [-1,1] *and that the density* f *satisfies condition* F.
(a) Then

$$\mathrm{var}(f_{nh}(x)) = \frac{1}{nh} f(x) \int_{-1}^{1} K^2(u)du + r_5(x,h),$$

where the remainder term r_5 *satisfies*

$$\sup_{0<h\leq h_n'} \ \sup_{x\in D_h\cap[-M,M]} r_5(x,h) = O(\frac{1}{n}), \text{ for } n \to \infty,$$

for every positive M.
(b) For x_0 *a fixed point we have*

$$\mathrm{var}(f_{nh}(x_0+th)) = \frac{1}{nh}(f(x_0-)\int_{t}^{1} K^2(u)du + f(x_0+) \int_{-1}^{t} K^2(u)du)+ r_6(t,h),$$

where the remainder r_6 *satisfies*

$$\sup_{0<h\leq h_n'} \ \sup_{-M\leq t\leq M} r_6(t,h) = O(\frac{1}{n}), \text{ for } n \to \infty,$$

for every positive M. □

Proof. Recall that by (2.3) we have

$$\mathrm{var}\ (f_{nh}(x)) = \frac{1}{nh} E \frac{1}{h} K^2((x-X_1)/h)) - \frac{1}{n} (E \frac{1}{h} K((x-X_1)/h))^2 .$$

Both terms can be expanded by lemma 2.1, taking the function G equal to K^2 to deal with the first term and equal to K to deal with the second term. It turns out that the second term is negligible. The leading terms in the expansions of the first term give the leading terms in the expansions of the variance. □

Remark 2.5. To get the bias of order h^2 on the set D_h in (2.12) we have explicitly used that the integral of uK(u) is equal to zero. Assuming more smoothness of f, a bias of order h^m, with $m>2$, can be obtained using kernels satisfying

$$\int_{-1}^{1} u^i K(u)du = 0 \quad \text{for } i = 1,2,...,m-1,$$

and

$$\int_{-1}^{1} u^m K(u)du \neq 0.$$

Such kernels, called *higher order kernels*, clearly take on negative values. As a consequence they produce density estimates which can be negative. We don't consider higher order kernels here. We only mention that cross-validation, a technique discussed in the next chapter, can be used to select an appropriate order for a kernel, see Hall & Marron (1988).

Example 2.6. To illustrate the bias expansions we have computed the bias of a kernel estimator of the density f in example 1.1. The kernel we have used is

(2.15) $$K(x) = \tfrac{35}{32}(1-x^2)^3 I_{[-1,1]}(x).$$

Figure 2.2.1 shows a graph of f and a graph of $Ef_{nh}(x)$ where we have taken the bandwidth h equal to $\frac{1}{2}$. Figure 2.2.2 shows a graph of the bias $b(x,\frac{1}{2})$ of $f_{nh}(x)$.

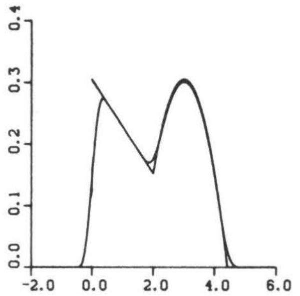

Figure 2.2.1. The density f and Ef_{nh} for $h=\frac{1}{2}$.

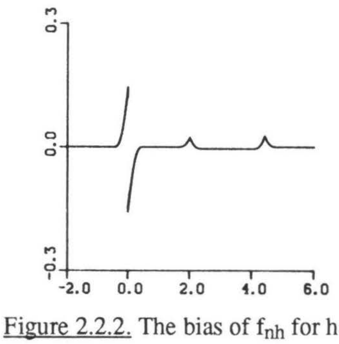

Figure 2.2.2. The bias of f_{nh} for $h=\frac{1}{2}$.

Recall that for this density the set D is equal to $\{0,2,3+\sqrt{2}\}$ and the set D_h is equal to the following union of intervals,

$$D_h = (-\infty,-h)\cup(h,2-h)\cup(2+h,3+\sqrt{2}-h)\cup(3+\sqrt{2}+h,\infty).$$

If (h_n) is a sequence of bandwidths converging to zero then for every fixed $x \notin D$ we have

$$b(x,h_n) = \tfrac{1}{2}h_n^2 f''(x) \int_{-1}^{1} u^2K(u)du + o(h_n^2), \quad \text{for } n\to\infty,$$

since $x\in D_{h_n}$ for n large enough. In fact this expansion holds uniformly on D_{h_n} ,that is in all points of at least a distance h_n to the singular points of f. In the picture we see that the bias is much larger close to these points.

Next let us consider the jumping point x=0. Then theorem 2.3 gives the following expansion

$$b(th_n,h_n) = b_0(t)\delta^{(0)}(0) + O(h_n) \text{ , for } n\to\infty \text{ and } t\neq 0.$$

This approximation holds uniformly for t in [-1,1], so on $[-h_n,h_n]$ the bias asymptotically resembles the function $b_0(t)$ times the jump size of f in zero. Notice that if K is a symmetric kernel then the function b_0 is an uneven function.The bias will not converge to zero close to a jumping point if the distance to the jumping point is measured in terms of h.

For the point x=2 we have the expansion

$$b(2+th_n,h_n) = h_nb_1(t)\delta^{(1)}(2) + O(h_n^2), \text{ for } n\to\infty \text{ and } t\neq 0,$$

again uniformly for t in [-1,1]. This expansion shows that on the interval $[2-h_n,2+h_n]$ the bias asymptotically resembles h_n times the function b_1 times the jump size of f' in 2. By the symmetry of K the function b_1 is an even function. Close to a kink the bias does converge to zero but it is not of the same order h_n^2 as it would have been in the smooth points in D_{h_n}. Here the bias is of order h_n! The point $x=3+\sqrt{2}$ can be treated similarly since f has a kink in this point too.

The consequences of theorem 2.3 for the bias close to a point where f and f' are continuous and f'' has a jump is left to the reader.

All the previous considerations about the bias suggest that very small bandwidths give good density estimates. This is far from true. Using theorem 2.4 we obtain the next expansion of the variance of f_{nh} in a point x which does not belong to D

$$\text{var}(f_{nh}(x)) = \frac{1}{nh} f(x) \int_{-1}^{1} K^2(u)du + O(\tfrac{1}{n}) \text{ , for } n\to\infty \text{ and } h\downarrow 0.$$

This expansion shows that very small bandwidths cause large variances of the kernel estimator. Part (b) of the theorem implies that this is also true if x belongs to D. It follows that we should require $nh_n\to\infty$, for $n\to\infty$, otherwise the variance does not vanish asymptotically. For optimal choices of the bandwidth these two effects have to be balanced. Of course what we mean by optimal should be made precise. Two global optimality concepts are discussed in the next sections.

Here let us briefly discuss estimation of f in a fixed point x. A common loss function when estimating a real valued parameter is the mean squared error. The mean squared error of $f_{nh}(x)$ is defined by

$$\mathrm{MSE}_n(x,h) := E(f_{nh}(x) - f(x))^2.$$

A simple computation shows

$$\mathrm{MSE}_n(x,h) = b(x,h)^2 + \mathrm{var}(f_{nh}(x)).$$

Let (h_n) denote a sequence of bandwidths converging to zero. By theorems 2.3 and 2.4 for a point $x \notin D$ the mean squared error can be expanded as follows

$$\mathrm{MSE}_n(x,h_n) = \tfrac{1}{4}h_n^4 f''(x)^2\Big(\int_{-1}^{1} u^2K(u)du\Big)^2 + \frac{1}{nh_n}f(x)\int_{-1}^{1} K^2(u)du \;+ o(h_n^4+\frac{1}{nh_n}).$$

Minimizing the leading term in this expansion we obtain the asymptotically optimal bandwidth

$$h_n^{opt} = \Big(f(x)\int_{-1}^{1} K^2(u)du \,/\, (f''(x)\int_{-1}^{1} u^2K(u)du)^2\Big)^{1/5} n^{-1/5}.$$

This choice results in a mean squared error of order $n^{-4/5}$. Since the expansion of the bias in a kink is different we also have a different expansion of the mean squared error. If x is a point where f has a kink then we have

$$\mathrm{MSE}_n(x,h_n) = h_n^2 b_1^2(t)\delta^{(1)}(x)^2 + \frac{1}{nh_n}f(x)\int_{-1}^{1} K^2(u)du \;+ o(h_n^2+\frac{1}{nh_n}),$$

which leads to an optimal bandwidth of order $n^{-1/3}$ and a mean squared error of order $n^{-2/3}$. It is not clear what the value of f should be in a jumping point so we don't consider estimation of f in such a point.

There is one more unexpected lesson to be learned from example 2.6. Careful examination of figure 2.2.2 on the interval (1/2,3/2) suggests that the bias is identically equal to zero on this interval. The next remark shows that this is no coincidence.

Remark 2.7. If a density f is linear on an interval [a,b] then the bias of a kernel estimator is equal to zero on the set [a+h,b-h], provided h is smaller than b-a. The proof is left to the reader. Now suppose that we want to estimate f at a point x inside [a,b]. In that case bandwidths h_n, which asymptotically minimize the mean squared error of $f_{nh_n}(x)$, don't converge to zero. This is immediate from the fact that the mean squared error for vanishing sequences of bandwidths can always be decreased by taking larger bandwidths. This follows since the estimator $f_{nh_n}(x)$ is unbiased and has a variance of order

$1/(nh_n)$. On the other hand the bandwidths h_n can not converge to infinity either, since then, if the kernel is bounded by $K^*>0$, by

$$f_{nh}(x) \le \frac{K^*}{h_n}, \text{ for all x,}$$

the estimate would converge to zero at every point of the real line. The conclusion is that in this case good choices for the bandwidth are asymptotically bounded away from zero and infinity.

At this point it should be noted that the merit of theorems 2.3 and 2.4 lies not only in the pointwise properties just discussed, but also in the fact that these theorems give uniform approximations of the bias and the variance on any bounded interval of the real line. This can be achieved by considering D_{h_n} and the h_n intervals of around the points of D separately. Thus we can also expand integrals involving the bias and the variance, provided we integrate over bounded areas.

2.3. The integrated squared error criterion.

In the remainder of this chapter we approach density estimation from a global point of view. Suppose that we want to estimate the density "well" on some subset E of the real line instead of in a fixed point. What we mean by "well" could be quantified for instance by requiring that the estimate minimizes the integrated squared error loss

$$\text{(2.16)} \qquad ISE_n(h) := \int_E (f_{nh}(x)-f(x))^2 w(x)dx,$$

where w is a nonnegative measurable weight function. Incorporating the indicator function of the set E in the weight function we can rewrite (2.16) in the more convenient form

$$ISE_n(h) = \int_{-\infty}^{\infty} (f_{nh}(x)-f(x))^2 w(x)dx.$$

Since the integrated squared error measures the discrepancy between the random function f_{nh} and the true density f, it is a random variable itself. The mean integrated squared error, defined as the expectation of the integrated squared error,

$$MISE_n(h) := E\ ISE_n(h) = E \int_{-\infty}^{\infty} (f_{nh}(x)-f(x))^2 w(x)dx,$$

is a deterministic loss function. We discuss the asymptotic behavior of the mean integrated squared error in the following section. We also derive the asymptotic distribution of the integrated squared error about its mean and discuss the relation between the two loss criteria.

2.3.1. The mean integrated squared error.

The mean integrated squared error can be rewritten as follows,

$$MISE_n(h) = \int_{-\infty}^{\infty} E(f_{nh}(x)-f(x))^2 w(x)dx =$$

$$(2.17) \qquad \int_{-\infty}^{\infty} \{b(x,h)^2 + \mathrm{var}(f_{nh}(x))\} w(x)dx.$$

This shows that $\mathrm{MISE}_n(h)$ is a weighted average of the mean squared error of $f_{nh}(x)$, the estimate at the point x. We can use the expansions of the bias and the variance in the previous section to derive an expansion of the mean intgerated squared error.

Assume that f satisfies condition F. The set D of singular points of f contains at most countably many points $d_1, d_2, \ldots$. Recall that $\delta^{(0)}(d_i)$, $\delta^{(1)}(d_i)$ and $\delta^{(2)}(d_i)$ denote the jump sizes of f, f ' and f " at the point d_i. We have to impose some extra conditions on the weight function w. We assume that w has a bounded support, which we denote by supp(w), and we assume that w has finite left and right limits in the singular points of f. We further assume that these limits are not both equal to zero in those singular points of f which also belong to supp(w). Define

$$\Delta_w^{(0)} := \sum_{i=1}^{\infty} (w(d_i-) + w(d_i+))\delta^{(0)}(d_i)^2$$

and

$$\Delta_w^{(1)} := \sum_{i=1}^{\infty} (w(d_i-) + w(d_i+))\delta^{(1)}(d_i)^2.$$

It follows from condition (F.1) and the fact that w has bounded support that these sums exist of only finitely many nonvanishing terms, since there are only finitely many elements of D contained in the support of w. So $\Delta_w^{(0)}$ and $\Delta_w^{(1)}$ are finite nonnegative real numbers. It turns out that the mean integrated squared error has a different asymptotic expansion in the following three cases:

case I	:	$\Delta_w^{(0)} > 0$,
case II	:	$\Delta_w^{(0)} = 0$ and $\Delta_w^{(1)} > 0$,
case III	:	$\Delta_w^{(0)} = \Delta_w^{(1)} = 0$.

The meaning of these cases will become clear after we have proved the next theorem.

Theorem 2.8. *Suppose that the density* f *satisfies condition* F *and that the kernel* K *satisfies condition* K. *Let* w *be a bounded measurable nonnegative weight function with bounded support and finite left and right limits in the singular points of* f. *We assume that these limits are not both equal to zero for singular points in* supp(w). *Then for any sequence of bandwidths* (h_n) *converging to zero and for* n *tending to infinity we have*

$$(2.18) \qquad \mathrm{MISE}_n(h_n) = \frac{1}{nh_n} \int_{-1}^{1} K^2(u)du \int_{-\infty}^{\infty} f(x)w(x)dx + O(\frac{1}{n}) +$$

$$\begin{cases} h_n \Delta_w^{(0)} \int_0^1 b_0^2(t)dt + o(h_n) & \textit{in case I} \\ h_n^3 \Delta_w^{(1)} \int_0^1 b_1^2(t)dt + o(h_n^3) & \textit{in case II} \\ \frac{1}{4} h_n^4 \left(\int_{-1}^{1} u^2 K(u)du \right)^2 \int_{-\infty}^{\infty} f''(x)^2 w(x)dx + o(h_n^4) & \textit{in case III.} \end{cases} \qquad \square$$

Proof. By (2.17) the mean integrated weighted squared error can be written as the integrated weighted squared bias plus the integrated weighted squared variance. The basic idea of the proof is that we split up the integration area in the set D_{h_n} and its complement, which in its turn is a countable union of h_n neighborhoods of the elements of D. Starting with the integrated weighted variance we write

$$(2.19) \qquad \int_{-\infty}^{\infty} \text{var}(f_{nh}(x))w(x)dx = \int_{D_{h_n}} \text{var}(f_{nh}(x))w(x)dx + \sum_{i=1}^{\infty} \int_{d_i-h_n}^{d_i+h_n} \text{var}(f_{nh}(x))w(x)dx.$$

The same decomposition is used for the integrated weighted squared bias. Since w has bounded support theorems 2.3 and 2.4 provide us with asymptotic expansions of the integrands over the integration areas. Thus part (a) of theorem 2.4 implies that we have,

$$(2.20) \qquad \int_{D_{h_n}} \text{var}(f_{nh}(x))w(x)dx = \frac{1}{nh_n} \int_{-1}^{1} K^2(u)du \int_{D_{h_n}} f(x)w(x)dx + \int_{D_{h_n}} r_5(x,h_n)w(x)dx,$$

where r_5 satisfies

$$\sup_{x \in D_{h_n} \cap \text{supp}(w)} r_5(x,h_n) = O(\frac{1}{n}), \text{ for } n \to \infty,$$

since the support of w is bounded. This implies that the second term of (2.20) is of order $O(\frac{1}{n})$. To deal with the first term notice that the Lebesgue measure of $\text{supp}(w) \backslash D_{h_n}$ is of order $O(h_n)$, so replacing the integral over D_{h_n} by an integral over supp(w) the difference is of order $O(\frac{1}{n})$ and therefore (2.19) equals the first term in the expansion (2.18).

By part (b) of theorem 2.4 it follows that for each d_i belonging to D we have

$$\int_{d_i-h_n}^{d_i+h_n} \text{var}(f_{nh}(x))w(x)dx = O(\frac{1}{n}),$$

and since there are only finitely many of such integrals which give a non zero contribution to (2.19) the sum of these terms is also of order $O(\frac{1}{n})$. This deals with the integrated weighted variance term.

Next we concentrate on the integrated weighted squared bias term. By part (a) of theorem 2.3 we have

$$\int_{D_{h_n}} b(x,h_n)^2 w(x)dx =$$

$$(2.21) \qquad \int_{D_{h_n}} (\tfrac{1}{2} h_n^2 f''(x) \int_{-1}^{1} u^2 K(u)du + r_3(x,h_n))^2 w(x)dx =$$

$$\tfrac{1}{4}h_n^4 \Big(\int_{-1}^{1} u^2K(u)du\Big)^2 \int_{-\infty}^{\infty} f\,''(x)^2w(x)dx + o(h_n^4),$$

which follows from similar arguments as above. We proceed with observing that for each d_i belonging to D we have

$$\int_{d_i-h_n}^{d_i+h_n} b(x,h_n)^2w(x)dx =$$

$$h_n \int_{-1}^{1} b(d_i+th_n,h_n)^2w(d_i+th_n)dt = \tag{2.22}$$

$$h_n \int_{-1}^{0} b(d_i+th_n,h_n)^2w(d_i+th_n)dt + h_n \int_{0}^{1} b(d_i+th_n,h_n)^2w(d_i+th_n)dt\ .$$

Since cases II and III are similar to case I we only treat case I . By part (b) of theorem 2.3 and the dominated convergence theorem (2.22) is asymptotically equivalent to,

$$h_n \int_{-1}^{0} b_0^2(t)\delta^{(0)}(d_i)^2w(d_i-)dt + h_n \int_{0}^{1} b_0^2(t)\delta^{(0)}(d_i)^2w(d_i+)dt + o(h_n),$$

which can be rewritten as

$$h_n \int_{-1}^{0} b_0^2(t)dt\ \delta^{(0)}(d_i)^2(w(d_i-) + w(d_i+)) + o(h_n).$$

The proof is completed by selecting the leading terms and adding them up. It should be noted that in cases I and II the terms (2.22) dominate over the term (2.21) while in case III it is the other way around. We need the condition that for singular points d in supp(w) either w(d-) or w(d+) is positive to ensure that $\Delta_w^{(0)} = 0$ implies that all the jump sizes $\delta^{(0)}(d)$ for points d in supp(w) are equal to zero, and similarly that $\Delta_w^{(1)} = 0$ implies that the jump sizes $\delta^{(1)}(d)$ are equal to zero for points d in supp(w).

□

Remark 2.9. The expansion for the mean integrated squared error holds uniformly in interval $(0,h_n']$, where (h_n') is a fixed sequence of bandwidths converging to zero. This follows from the proof above using the fact that the orderbounds on the remainders in theorems 2.3 and 2.4 also hold uniformly on such intervals.

Theorem 2.8 supplements the results of Van Eeden (1985) and Cline & Hart (1986) for $w \equiv 1$ in the sense that we allow weight functions. In those papers however the kernels are not required to have a bounded support. Cline & Hart also consider the higher order kernels mentioned in remark (2.5).

In theorem 2.8 the weight function $w \equiv 1$ is not allowed because of its unbounded support. Let us instead consider the weight function $w(x) := I_E(x)$, $-\infty<x<\infty$, where E is a bounded interval [a,b], $-\infty<a<b<\infty$. With this weight function the mean integrated squared error equals

$$\mathrm{MISE}_n(h) = E\int_E (f_{nh}(x) - f(x))^2dx,$$

and the constants $\Delta_w^{(0)}$ and $\Delta_w^{(1)}$ are equal to

$$(2.23)\qquad \begin{aligned} \Delta_w^{(0)} &= 2\sum_{i=1}^{m} \delta^{(0)}(d_i)^2 + \delta^{(0)}(a)^2 + \delta^{(0)}(b)^2 \\ \Delta_w^{(1)} &= 2\sum_{i=1}^{m} \delta^{(1)}(d_i)^2 + \delta^{(1)}(a)^2 + \delta^{(1)}(b)^2, \end{aligned}$$

where $d_1,\dots,d_m$ denote the finitely many points of D inside (a,b). Notice that the contribution of the endpoints a and b is different than that of the points $d_1,\dots,d_m$ strictly inside E since $w(a-) = w(b+) = 0$ and $w(a+) = w(b-) = 1$. For the points d inside E both w(d-) and w(d+) are one. It follows that the cases we have distinguished in theorem 2.8 correspond to the fact whether f has jumping points in [a,b], case I, whether f has kinks in [a,b], but no jumps, case II, and whether there are neither kinks nor jumps in [a,b], which corresponds to case III. Thus the conclusion to be drawn from theorem 2.8 is that the presence of jumps and kinks in E causes a larger mean integrated squared error than in the smooth case III. Jumps increase the error most since in that case we are estimating a discontinuous function with a continuous one.

Similar conlusions hold for the error if the weight function is equal to $w(x) = f^{-1}(x)I_E(x)$, where E is an interval as above and the density f is assumed to be bounded away from zero on E. If f=0 we also set w=0. The mean integrated squared error criterion thus obtained, i.e.

$$\mathrm{MISE}_n(h) = E\int_E (f_{nh}(x) - f(x))^2 f^{-1}(x)dx,$$

plays an important role in the next chapter. Notice that this mean integrated squared error is the squared L_2 norm over the set E of the random function

$$\frac{f_{nh}(x) - f(x)}{f^{1/2}(x)},$$

which for each fixed point which is not a jumping point by theorem 2.4 has an asymptotic variance independent of x. In this case the values of $\Delta_w^{(0)}$ and $\Delta_w^{(1)}$ are given by

$$(2.24)\qquad \begin{aligned} \Delta_w^{(0)} &= \sum_{i=1}^{m} (f(d_i-)^{-1}+f(d_i+)^{-1})\delta^{(0)}(d_i)^2 + f(a+)^{-1}\delta^{(0)}(a)^2 + f(b-)^{-1}\delta^{(0)}(b)^2 \\ \Delta_w^{(1)} &= \sum_{i=1}^{m} (f(d_i-)^{-1}+f(d_i+)^{-1})\delta^{(1)}(d_i)^2 + f(a+)^{-1}\delta^{(1)}(a)^2 + f(b-)^{-1}\delta^{(1)}(b)^2. \end{aligned}$$

This shows that, apart from the constants, jumps and kinks have the same influence on the asymptotic behavior of the mean integrated squared error as in the case of the previously considered weight function.

Returning to the expansion of theorem 2.8 we see again that small bandwidths cause a large integrated variance term and that large bandwidths cause a large integrated squared bias term. Balancing these effects by minimizing the leading term in the expansion leads us to the following optimal bandwidths,

$$(2.25) \qquad h_n^{\text{opt}} = \begin{cases} \alpha_I(f,w)^{1/2}\beta_I(K)^{1/2}\, n^{-1/2} & \text{in case I} \\ \alpha_{II}(f,w)^{1/4}\beta_{II}(K)^{1/4}\, n^{-1/4} & \text{in case II} \\ \alpha_{III}(f,w)^{1/5}\beta_{III}(K)^{1/5}\, n^{-1/5} & \text{in case III} \end{cases} ,$$

where the constants α, depending on the density f and the weight function w, are given by

$$\alpha_I(f,w) \;= (\Delta_w^{(0)})^{-1} \int_{-\infty}^{\infty} f(x)w(x)dx)$$

$$\alpha_{II}(f,w) \;= (3\Delta_w^{(1)})^{-1} \int_{-\infty}^{\infty} f(x)w(x)dx,$$

$$\alpha_{III}(f,w) = \Big(\int_{-\infty}^{\infty} f\,''(x)^2 w(x)dx\Big)^{-1} \int_{-\infty}^{\infty} f(x)w(x)dx,$$

and the constants β, depending only on the kernel K, by

$$\beta_I(K) \;= \int_{-1}^{1} K^2(u)du \,\Big(\int_0^1 b_0^2(t)dt\Big)^{-1}$$

$$\beta_{II}(K) \;= \int_{-1}^{1} K^2(u)du \,\Big(\int_0^1 b_1^2(t)dt\Big)^{-1}$$

$$\beta_{III}(K) = \int_{-1}^{1} K^2(u)du \,\Big(\int_{-1}^{1} u^2K(u)du\Big)^{-2}.$$

Remark 2.10. It is no surprise that theorem 2.8 shows that the presence of jumps of f in the interval E, case I !, causes a large mean integrated squared error. Even if we use an asymptotically optimal bandwidth for case I, the mean integrated squared error is still of order $n^{-1/2}$, while in case II and in case III it would have been of orders $n^{-3/4}$ and $n^{-4/5}$ respectively. If we don't know where the jumping points are then this large error is unavoidable if we use a kernel estimator. However if a jumping point is known then the influence of this jumping point can be substantially reduced. For densities with support [c,d], [c,∞) or (-∞,d], $-\infty \le c < d \le \infty$, with jumps at the points c or d which are known points, Schuster (1985) shows that the kernel estimator can be improved by a symmetrization device. The symmetrization has the effect that the error caused by the jump is reduced to an error caused by a kink. The special case of kernel estimation of decreasing densities on $[0,\infty)$ with a jump at zero is also treated in Devroye (1987) section 8.4. Cline & Hart (1986) generalize this symmetrization device to be able to deal with known jumping points which are not necessarily endpoints of the support of the density f.

Until now we have only considered the choice of an optimal bandwidth. For all three cases there is also an optimal kernel.
First we consider case III. If we substitute the optimal bandwidth for this case in the expansion (2.18) we get

$$\lim_{n\to\infty} n^{4/5}\, \mathrm{MISE}_n(h_n^{\mathrm{opt}}) = \tfrac{5}{4}\Big(\int_{-\infty}^{\infty} f(x)w(x)dx\Big)^{4/5}\Big(\int_{-\infty}^{\infty} f''(x)^2 w(x)dx\Big)^{1/5}$$

$$\Big(\int_{-1}^{1} K^2(u)du\Big)^{4/5}\Big(\int_{-1}^{1} u^2K(u)du\Big)^{2/5}.$$

Under certain regularity conditions this expansion also holds for kernels with unbounded support. It is shown in Epanechnikov (1969) that the kernel which minimizes this expression over the class of symmetric kernels is

$$K(x) = \tfrac{3}{4}(1-x^2)I_{[-1,1]}(x), \tag{2.26}$$

which is the well known classical optimal kernel. It is less well known that the same procedure can be carried out in the non-smooth cases. The optimal kernel in case I was derived by Van Eeden (1985), it equals the Laplace density function

$$K(x) = \tfrac{1}{2}e^{-|x|}, \quad -\infty<x<\infty. \tag{2.27}$$

The optimal kernel in case III, derived by Cline & Hart (1986) and simultaniously by Swanepoel (1987), is a bounded support density given by

$$K(x) = (2\sinh(\tfrac{\pi}{2}))^{-1}(\cos(|x|)\cosh(\tfrac{\pi}{2}-|x|) + \sin(|x|)\sinh(\tfrac{\pi}{2}-|x|))\ I_{[-\pi/2,\pi/2]}(x), \tag{2.28}$$

The last two kernels don't have support [-1,1]. By the scaling property (2.1) we can transfer the kernel (2.28) to a kernel with support [-1,1], without disturbing the optimality property. This can of course not be done with the Laplace kernel because of its unbounded support. However, Swanepoel (1987) gives bounded support kernels which approach the Laplace kernel (2.27) arbitrarily closely in the sense that the constants in the expansion of the mean squared error for the optimal bandwidth in case II become close to the optimal constants in this case, attained by the Laplace kernel. This shows that our use of bounded support kernels is not restrictive from the point of view of the mean squared error criterion.

2.3.2. Asymptotic normality of the integrated squared error.

The previous section dealt with the expectation of the integrated squared error and the optimal bandwidths which we derived asymptotically minimize this expectation. However it is more natural to aim for minimizing the integrated squared error itself. From this point of view it is important that the variation of the integrated squared error around its mean does not dominate the leading terms in the asymptotic expansion of the mean integrated squared error, and thus disturb the optimality property of the optimal bandwidths derived in the previous section. To quantify this variation we give a central limit theorem for the integrated squared error. We consider all densities satisfying condition F.

Theorem 2.11. *Assume that* f *satisfies condition* F *and that the kernel* K *satisfies condition* K. *Furthermore assume that* w *is a bounded almost everywhere continuous nonnegative weight function with a bounded support and finite left and right limits in the singular points of* f. *Further we assume that these left and right limits are not both equal to zero. Let* (h_n), *a sequence of nonnegative bandwidths, satisfy* $h_n \to 0$ *and* $nh_n \to \infty$. *Then*

$$d_n\,(\mathrm{ISE}_n(h_n) - \mathrm{MISE}_n(h_n)) \overset{\mathcal{D}}{\rightarrow}$$

$$\begin{cases} N(0,2\sigma^2) & \textit{if } d_n = nh_n^{1/2} \quad \textit{and if } nh_n^2 \to 0 \\ N(0,2\sigma^2+\lambda\sigma_I^2) & \textit{if } d_n = nh_n^{1/2} \quad \textit{and if } nh_n^2 \to \lambda \\ N(0,\sigma_I^2) & \textit{if } d_n = n^{1/2}h_n^{-1/2} \quad \textit{and if } nh_n^2 \to \infty \end{cases} \quad , \textit{ in case I},$$

$$\begin{cases} N(0,2\sigma^2) & \textit{if } d_n = nh_n^{1/2} \quad \textit{and if } nh_n^4 \to 0 \\ N(0,2\sigma^2+\lambda\sigma_{II}^2) & \textit{if } d_n = nh_n^{1/2} \quad \textit{and if } nh_n^4 \to \lambda \\ N(0,\sigma_{II}^2) & \textit{if } d_n = n^{1/2}h_n^{-3/2} \quad \textit{and if } nh_n^4 \to \infty \end{cases} \quad , \textit{ in case II},$$

$$\begin{cases} N(0,2\sigma^2) & \textit{if } d_n = nh_n^{1/2} \quad \textit{and if } nh_n^5 \to 0 \\ N(0,2\sigma^2+\lambda\sigma_{III}^2) & \textit{if } d_n = nh_n^{1/2} \quad \textit{and if } nh_n^5 \to \lambda \\ N(0,\sigma_{III}^2) & \textit{if } d_n = n^{1/2}h_n^{-2} \quad \textit{and if } nh_n^5 \to \infty \end{cases} \quad , \textit{ in case III},$$

where the variances σ^2, σ_I^2, σ_{II}^2 *and* σ_{III}^2 *are given by*

$$\sigma^2 := \int_{-\infty}^{\infty} \Big(\int_{-1}^{1} K(v)K(v+z)dv \Big)^2 dz \int_{-\infty}^{\infty} w^2(u)f^2(u)du,$$

$$\sigma_I^2 := 4 \sum_{i=1}^{\infty} \delta^{(0)}(d_i)^2 \Big(f(d_i-) \int_{-\infty}^{0} \Big(w(d_i-) \int_{-1}^{0} K(t+v)b_0(t)dt \;+ w(d_i+)\int_{0}^{1} K(t+v)b_0(t)dt \Big)^2 dv$$

$$+ f(d_i+)\int_0^\infty \Big(w(d_i-)\int_{-1}^0 K(t+v)b_0(t)dt \ + w(d_i+)\int_0^1 K(t+v)b_0(t)dt \ \Big)^2 dv\Big),$$

$$\sigma_{II}^2 := 4 \sum_{i=1}^\infty f(d_i)\delta^{(1)}(d_i)^2 \int_{-\infty}^\infty \Big(w(d_i-)\int_{-1}^0 K(t+v)b_1(t)dt + w(d_i+)\int_0^1 K(t+v)b_1(t)dt\Big)^2 dv,$$

$$\sigma_{III}^2 := \Big(\int_{-1}^1 v^2K(v)dv \ \Big)^2 \Big(\int_{-\infty}^\infty f''(x)^2w^2(x)f(x)dx - \Big(\int_{-\infty}^\infty f''(x)w(x)f(x)dx\Big)^2\Big). \qquad \Box$$

Before we prove this theorem we first discuss its implications. Firstly the theorem shows that the asymptotically optimal bandwidths not only asymptotically minimize the mean integrated squared error but that they also minimize the order of the variance of the integrated squared error. Secondly it is readily checked that in all the cases considered we have for n tending to infinity

$$\text{(2.29)} \qquad \frac{ISE_n(h_n) - MISE_n(h_n)}{MISE_n(h_n)} \to 0, \text{ in probability,}$$

which implies

$$\frac{ISE_n(h_n)}{MISE_n(h_n)} \to 1, \text{ in probability.}$$

This theorem shows that the bandwidths which asymptotically minimize the mean integrated squared error also, in probability, asymptotically minimize the integrated squared error. Property (2.29) is shown for smooth densities by Hall (1982b). Under regularity conditions Marron & Härdle (1986) show that (2.29) holds almost surely uniformly in the bandwidths h_n. Furthermore they don't require smoothness of the density f. The uniformity in the bandwidths is useful for studying kernel estimation techniques with random bandwidths, in particular the cross-validation techniques discussed in the next chapter.

Central limit theorems for the integrated squared error for smooth densities have been derived by Bickel & Rosenblatt (1973) and by Hall (1984). Both theorems correspond to our case III. Bickel & Rosenblatt consider small bandwidths which satisfy $h_n = O(n^{-2/9})$, so their theorem is covered by the first line of the case III part of the theorem above. It should be noted that doing so they excluded the optimal bandwidths in that case which are of order $n^{-1/5}$. This was recognized by Hall who proved a central limit theorem for the integrated squared error of multivariate kernel estimators which is similar to our case III part in the one dimensional case.

Remark 2.12. A nice consequence of our theorem is that the asymptotic variance σ_{II}^2 of the integrated squared error in case II is equal to zero if the value of the density in all the singular points where f has a kink is equal to zero. This shows that for sequences of bandwidths with $nh_n^4 \to \infty$ the influence of a kink in the density on the variation of the integrated squared error is of smaller order if this kink is in a point where the density is zero.

Proof of theorem 2.11. We rewrite the integrated squared error $ISE_n(h)$ as follows,

$$ISE_n(h) =$$

$$\int_{-\infty}^{\infty} (f_{nh}(x) - f(x))^2 w(x)dx =$$

$$\int_{-\infty}^{\infty} \Big(\frac{1}{nh} \sum_{i=1}^{n} K((x-X_i)/h) - f(x)\Big)^2 w(x)dx =$$

$$\frac{1}{n^2h^2} \sum_{i \neq j} \int_{-\infty}^{\infty} K\Big(\frac{x-X_i}{h}\Big) K\Big(\frac{x-X_j}{h}\Big) w(x)dx +$$

$$- \frac{2}{nh} \sum_{i=1}^{n} \int_{-\infty}^{\infty} K\Big(\frac{x-X_i}{h}\Big) f(x)w(x)dx +$$

$$\frac{1}{n^2h^2} \sum_{i=1}^{n} \int_{-\infty}^{\infty} K^2\Big(\frac{x-X_i}{h}\Big) w(x)dx +$$

$$\int_{-\infty}^{\infty} f^2(x)w(x)dx.$$

It follows that the integrated squared error is equal to a quadratic form plus a linear term. Statistics of this type are treated in appendix C, where special attention is paid to this specific case in theorem C.2 and remark C.3. If $b(u,h)$ denotes the bias function of a kernel estimator, i.e. $b(u,h)=Ef_{nh}(u)-f(u)$, theorem C.2 states that if we assume

$$(2.30) \qquad 4nh_n^{-1} \mathrm{var}\Big(\int_{-\infty}^{\infty} K\Big(\frac{u-X_1}{h_n}\Big) b(u,h_n)w(u)du \Big) \rightarrow \alpha^2, \; 0 \leq \alpha^2 < \infty$$

then the integrated squared error is asymptotically normal and we have

$$(2.31) \qquad nh_n^{1/2}(ISE_n(h_n) - MISE_n(h_n)) \xrightarrow{\mathcal{D}} N(0, 2\sigma^2 + \alpha^2),$$

where σ^2 is defined above. Remark C.3 says that if (2.30) converges to infinity we also have asymptotic normality because then the linear terms in the proof of theorem C.2 dominate over the quadratic term. In that case we have by (C.13)

$$(2.32) \qquad \tfrac{1}{2} n^{1/2} h_n \Big(\mathrm{var}\Big(\int_{-\infty}^{\infty} K\Big(\frac{u-X_1}{h_n}\Big) b(u,h_n)w(u)du \Big)\Big)^{-1/2} (ISE_n(h_n) - MISE_n(h_n)) \xrightarrow{\mathcal{D}} N(0,1).$$

For a fixed sequence of bandwidths (h_n), whether we are actually dealing with situation (2.31) or with situation (2.32) depends on whether (2.30) converges to a finite number or to infinity. This means that we have to expand the variance in (2.30). It turns out that the presence of singular points of f in the support of w influences the order of magnitude of this variance. Distinguishing the cases I, II and III, introduced in the previous section, we have

$$\mathrm{var}\Big(\int_{-\infty}^{\infty} K\Big(\frac{u-X_1}{h_n}\Big)b(u,h_n)w(u)du\Big) \sim \tfrac{1}{4}h_n^3\sigma_I^2 \qquad \text{in case I,}$$

$$\mathrm{var}\Big(\int_{-\infty}^{\infty} K\Big(\frac{u-X_1}{h_n}\Big)b(u,h_n)w(u)du\Big) \sim \tfrac{1}{4}h_n^5\sigma_{II}^2 \qquad \text{in case II,}$$

$$\mathrm{var}\Big(\int_{-\infty}^{\infty} K\Big(\frac{u-X_1}{h_n}\Big)b(u,h_n)w(u)du\Big) \sim \tfrac{1}{4}h_n^6\sigma_{III}^2 \qquad \text{in case III,}$$

Since the proof of these expansions is rather technical it is postponed to section 2.5.
In case I we have

$$4nh_n^{-1}\mathrm{var}\Big(\int_{-\infty}^{\infty} K\Big(\frac{u-X_1}{h_n}\Big)b(u,h_n)w(u)du\Big) \to \begin{cases} 0 & \text{if } nh_n^2 \to 0 \\ \lambda\sigma_I^2 & \text{if } nh_n^2 \to \lambda \\ \infty & \text{if } nh_n^2 \to \infty \end{cases},$$

and

$$\tfrac{1}{2}n^{1/2}h_n\Big(\mathrm{var}\Big(\int_{-\infty}^{\infty} K\Big(\frac{u-X_1}{h_n}\Big)b(u,h_n)w(u)du\Big)\Big)^{-1/2} \sim n^{1/2}h_n^{-1/2}\sigma_I^{-1},$$

which proves the theorem for case I by (2.31) and (2.32). The other cases are obtained in a similar way. □

2.4. Properties with respect to supremum distances.

Let E be a closed bounded interval on the real line. An alternative for the integrated squared error criterion and the mean integrated squared error criterion is the weighted supremum distance

$$\text{(2.33)} \qquad \sup_{x\in E} |f_{nh}(x) - f(x)|\, w(x),$$

where w is a weight function with $w(x)>0$ for $x\in E$. Since f_{nh} is a continuous function the supremum distance between f_{nh} and f is always larger than some positive constant if the density f has a jump in E. Therefore density estimation from the point of view of supremum distance loss functions is only meaningful for densities f which are continuous on some ε neighborhood of E. Consequently only such densities are considered.

We discuss two aspects of kernel estimation from the point of view of supremum distances. In section 2.4.1, using the bias expansions of theorem 2.3, we supplement results of Stute (1982b) on the almost sure asymptotically optimal bandwidths for the specific supremum loss function

$$\sup_{x\in E} |f_{nh}(x) - f(x)|\, f^{-1/2}(x). \tag{2.34}$$

Note that here w equals $f^{-1/2}$. In section 2.4.2 we derive an almost sure order bound for

$$\sup_{h\in I_n} \left(\frac{nh}{\log n}\right)^{1/2} \sup_{x\in E} |f_{nh}(x) - Ef_{nh}(x)| \tag{2.35}$$

Such bounds on the supremum distance between $f_{nh}(x)$ and $Ef_{nh}(x)$, uniformly in the bandwidth, are important tools in studying bandwidth selection methods. Consequently the presented bound is frequently used in chapter 3.

The first results on strong uniform consistency were obtained by Nadaraya (1965). Other relevant references are Révész (1978), Silverman (1978b), Kolcinskii (1980), Serfling (1982) and Stute (1982b).

2.4.1. Almost sure asymptotically optimal bandwidths.

We consider the loss function (2.34). Note that by theorem 2.4 the pointwise asymptotic variance of $(f_{nh}(x)-f(x))f^{-1/2}(x)$ is independent of x. We derive asymptotically good bandwidths for the supremum distance (2.34) by studying the two terms in the right hand side of in the inequality

$$\sup_{x\in E} |f_{nh}(x) - f(x)|\, f^{-1/2}(x) \le \sup_{x\in E} |f_{nh}(x) - Ef_{nh}(x)|\, f^{-1/2}(x) + \sup_{x\in E} |Ef_{nh}(x) - f(x)|\, f^{-1/2}(x). \tag{2.36}$$

The next theorem of Stute (1982b) gives the exact almost sure rate of the supremum norm of the error part in (2.36).

Theorem 2.13. (Stute 1982). *Let* (h_n) *be a sequence of positive bandwidths with* $h_n\to 0$, $nh_n\to\infty$, $\log(h_n^{-1})=o(nh_n)$ *and* $\log(h_n^{-1})/\log(\log n)\to\infty$. *Assume that* f *is continuous on* $E=[a,b]$, *with* $-\infty<a<b<\infty$, *and assume* $0<m\le f(x)\le M<\infty$, *for all* $x\in E$. *Furthermore let K be any kernel function of bounded variation with* $K(x) = 0$ *outside some finite interval* $[r,s)$. *With probability one we have*

$$\lim_{n\to\infty} \left(\frac{nh_n}{2\log(h_n^{-1})}\right)^{1/2} \sup_{x\in E_\varepsilon} |f_{nh_n}(x) - Ef_{nh_n}(x)|\, f^{-1/2}(x) = \Big(\int_r^s K^2(u)du\Big)^{1/2}, \tag{2.37}$$

where E_ε *denotes the interval* $(a+\varepsilon,b-\varepsilon)$ *for some* $\varepsilon>0$. □

Since theorem 2.3 gives us uniform expansions of the bias function $Ef_{nh}(x)-f(x)$ the next lemma with expansions of the bias part in the right hand side of (2.36) readily follows (recall $b_1(0)=\int_0^1 uK(u)du$). The proof of this lemma is omitted. Just as in the previous section the presence of singular points in the set E plays an important role. We only consider densities satisfying condition F in section 2.1 which are continuous on the interval E = [a,b], $-\infty<a<b<\infty$. Let $d_1,\dots, d_m$ denote the singular points of f in the interval E. Define two special cases,

case II : all jump sizes $\delta^{(0)}(d_i)$, $i=1,\dots,m$, are equal to zero and at least one of the jump sizes $\delta^{(1)}(d_i)$, $i=1,\dots,m$, is unequal to zero,

case III : all jump sizes $\delta^{(0)}(d_i)$ and $\delta^{(1)}(d_i)$, $i=1,\dots,m$, are equal to zero.

Lemma 2.14. *Let* f, *a density satisfying condition* F, *be bounded away from zero on the interval* E=[a,b], $-\infty<a<b<\infty$. *Let* $d_1,\dots, d_m$ *denote the singular points of* f *in the interval* E *and let the kernel* K *satisfy condition* K. *We have in case II*

$$\lim_{n\to\infty} h_n^{-1} \sup_{x\in E} |Ef_{nh_n}(x) - f(x)| f^{-1/2}(x) = \Big(\int_0^1 uK(u)du\Big) \max_{i=1,\dots,m} f(d_i)^{-1/2}|\delta^{(1)}(d_i)| \tag{2.38}$$

and in case III

$$\lim_{n\to\infty} h_n^{-2} \sup_{x\in E} |Ef_{nh_n}(x) - f(x)| f^{-1/2}(x) = \frac{1}{2}\Big(\int_{-1}^1 u^2K(u)du\Big) \sup_{x\in E} |f''(x)| f^{-1/2}(x). \qquad \square \tag{2.39}$$

By balancing the error and bias term in the right hand side of (2.36) we can now derive the asymptotically optimal bandwidths in the two cases described above. Stute (1982b) accomplished this for densities with a continuous third derivative. In exactly the same manner the following asymptotically optimal bandwidths can be derived, we omit the proof. Notice that Stutes result is covered by case III.

Theorem 2.15. *Suppose that for some* $\varepsilon>0$ f, *a density satisfying condition* F, *is uniformly continuous on* $[a-\varepsilon,b+\varepsilon]$, $-\infty<a<b<\infty$, *with* $0<m\le f(x)\le M<\infty$ *for all* $x\in[a-\varepsilon,b+\varepsilon]$. *Let* K *be a kernel which satisfies condition* K *and let* h_n^{opt} *denote the bandwidth which minimizes the right hand side of* (2.36), *then*

$$h_n^{opt} \sim \left(\frac{\int_{-1}^{1} K^2(u)du}{6\,\Big(\int_0^1 uK(u)du\Big)^2 \max_{i=1,\dots,m} f(d_i)^{-1}\delta^{(1)}(d_i)^2} \;\frac{\log n}{n} \right)^{1/3} \qquad \textit{in case II}$$

and

$$h_n^{opt} \sim \left(\frac{\int_{-1}^{1} K^2(u)du}{10\ (\int_{-1}^{1} u^2K(u)du)^2 \sup_{x\in E} f\ ''(x)^2 f^{-1}(x)} \ \frac{\log n}{n} \right)^{1/5} \qquad \textit{in case III.}$$

□

The corresponding orders of the supremum loss are almost surely $O((\log n/n)^{1/3})$ in case II and $O((\log n/n)^{2/5})$ in case III.

2.4.2. Uniformity in the bandwidths.

Let $I_n=[h_n', h_n'']$ be an interval of bandwidths with $h_n'=n^{-1+\delta}$ and $h_n''=n^{-\delta}$ for some δ with $0<\delta<1/2$. Note that for all bandwidths h_n in I_n we have $h_n \to 0$ and $nh_n \to \infty$. We derive a uniform orderbound for (2.35). This bound on the supremum over the set of bandwidths I_n is needed for proofs in later sections. There we also consider kernel functions which are not probability densities. For a related result see lemma 1 in Härdle & Marron (1985).

Theorem 2.16. *Let* f *be a bounded density and let* E *be a bounded interval. Suppose that the kernel* K *is a symmetric function with support* [-1,1], *not necessarily a density, and that* K *has a bounded derivative, then*

$$\limsup_{n\to\infty} \sup_{h\in I_n} \left(\frac{nh}{\log n}\right)^{1/2} \sup_{x\in E} |f_{nh}(x) - Ef_{nh}(x)| \le C, \textit{ almost surely,} \tag{2.40}$$

for some constant C>0. □

Remark 2.17. Notice that for $h\in I_n$ we have $\delta \log n \le \log(h^{-1}) \le (1-\delta)\log n$ so the norming constant in (2.40) is of the same order as the constant in (2.37). By the conditions on f in theorem 2.13 the factor $f^{-1}(x)$ in (2.37) is bounded on E.

As a step in the proof of theorem 2.16 we need a bound on the oscillations of $f_{nh}(x)-Ef_{nh}(x)$ as a function of both h and x. Define for nonnegative real numbers α and β the random variable

$$\Omega_n(\alpha,\beta) := \sup_{(h_1,h_2)\in A_n(\alpha)} \ \sup_{(x_1,x_2)\in B(\beta)} (h_1 \vee h_2)\ |f_{nh_1}(x_1)-Ef_{nh_1}(x_1)-f_{nh_2}(x_2)+Ef_{nh_2}(x_2)|,$$

where $h_1 \vee h_2$ denotes the maximum of h_1 and h_2,

$$A_n(\alpha) := \{(h_1,h_2) : h_1,h_2\in I_n,\ |h_1 - h_2| \le \alpha\},$$

and

$$B(\beta) := \{(x_1,x_2) : x_1,x_2\in E,\ |x_1 - x_2| \le \beta\}.$$

Proposition 2.18. *Assume that the conditions of theorem 2.16 hold. Let* (α_n) *and* (β_n) *be two sequences of real numbers such that*

$$\alpha_n = o(h_n'),\ \beta_n = O(1) \text{ for } n\to\infty,$$

and

$$n(\alpha_n + \beta_n)\,(\log n)^{-2} \to\infty,$$

then

(2.41) $$\Omega_n(\alpha_n,\beta_n) = o(\alpha_n + \beta_n),\ \textit{almost surely}. \qquad \square\ .$$

Proof. With $V_n(x) = n^{1/2}(F_n(x) - F(x))$, the empirical proces of the sample $X_1,\ldots,X_n$, we have by partial integration

$$f_{nh}(x) - Ef_{nh}(x) =$$

$$\frac{1}{h}\int_{-\infty}^{\infty} K((x-u)/h)d(F_n-F)(u) =$$

$$\frac{1}{h^2}\int_{-\infty}^{\infty} (F_n-F)(u)K'((x-u)/h)du \ =$$

$$\frac{1}{h}\int_{-1}^{1} (F_n-F)(x+hv)K'(v)dv$$

$$n^{-1/2}h^{-1}\int_{-1}^{1} V_n(x+hv)K'(v)dv.$$

Therefore for all $(h_1,h_2)\in A_n(\alpha_n)$ and all $(x_1,x_2)\in B(\beta_n)$ we have

$$|\, f_{nh_1}(x_1) - Ef_{nh_1}(x_1) - f_{nh_2}(x_2) + Ef_{nh_2}(x_2)\,| =$$

$$n^{-1/2}\,|h_1^{-1}\int_{-1}^{1} V_n(x_1+h_1v)K'(v)dv \ - \ h_2^{-1}\int_{-1}^{1} V_n(x_2+h_2v)K'(v)dv| \le$$

$$n^{-1/2}h_2^{-1}|\int_{-1}^{1} (V_n(x_1+h_1v) - V_n(x_2+h_2v))K'(v)dv|+$$

$$n^{-1/2}|h_1^{-1} - h_2^{-1}|\,|\int_{-1}^{1} V_n(x_1+h_1v)\ K'(v)dv|.$$

Assuming that $|K'|$ is bounded by the constant $c>0$ by $|\,x_1 + h_1v - x_2 - h_2v\,| \le \alpha_n + \beta_n$ the first term is bounded by

$$cn^{-1/2}h_2^{-1}\omega_n(\alpha_n + \beta_n),$$

with

$$(2.42) \qquad \omega_n(t) := \sup_{x \in E_{h_n''}, x+s \in E_{h_n''}, 0 \le s \le t} | V_n(x+s) - V_n(x) |,$$

the oscillation modulus of the empirical process V_n on the interval $E_{h_n''}$ (the h_n'' neighborhood of E). Since $\int_{-1}^{1} K'(v)dv = 0$ the second term can be rewritten as

$$n^{-1/2}(h_1^{-1} - h_2^{-1}) \Big|\Big(\int_{-1}^{1} (V_n(x_1+h_1 v) - V_n(x_1))K'(v)dv\Big|,$$

which is bounded by

$$cn^{-1/2}\alpha_n(h_1 h_2)^{-1}\omega_n(h_1).$$

Thus we have for n sufficiently large

$$\Omega_n(\alpha_n,\beta_n) =$$

$$\sup_{(h_1,h_2)\in A_n(\alpha_n)} \ \sup_{(x_1,x_2)\in B(\beta_n)} (h_1 \vee h_2) | f_{nh_1}(x_1) - Ef_{nh_1}(x_1) - f_{nh_2}(x_2) + Ef_{nh_2}(x_2) | \le$$

$$cn^{-1/2}\frac{(h_1 \vee h_2)}{h_2}\omega_n(\alpha_n + \beta_n) + cn^{-1/2}\alpha_n \frac{(h_1 \vee h_2)}{h_1 h_2}\omega_n(h_1) \le$$

$$c(1 + \alpha_n/h_2)\Big\{ (n(\alpha_n + \beta_n))^{-1/2}(\alpha_n + \beta_n)^{-1/2}\omega_n(\alpha_n + \beta_n)$$

$$+ (nh_n')^{-1/2} \sup_{h_1 \in I_n} h_1^{-1/2}\omega_n(h_1)\Big\} (\alpha_n + \beta_n) \le$$

$$o(1) (\log n)^{-1}\Big\{ (\alpha_n + \beta_n)^{-1/2}\omega_n(\alpha_n + \beta_n) + \sup_{h_1 \in I_n} h_1^{-1/2}\omega_n(h_1)\Big\}(\alpha_n + \beta_n).$$

The proof of proposition 2.18 is finished by an application of the next lemma about the oscillation modulus ω_n.

Lemma 2.19. *Let the oscillation modulus* ω_n *be defined by* (2.42) *then for any* $\varepsilon>0$, *any sequence of nonnegative real numbers* (t_n), *with* $t_n \to 0$ *and* $nt_n \to \infty$, *and any constant* $T>0$, *we have*

$$\limsup_{n\to\infty} \frac{1}{\log n} \sup_{t_n \le t < T} t^{-1/2}\omega_n(t) \le C, \textit{ almost surely,}$$

for some constant $C>0$. □

The proof of this lemma is given in section 2.5.

Proof of theorem 2.16. Let (ε_n) denote a sequence of nonnegative real numbers converging to zero. Since $Eh^{-1}|K((x-X_1)/h)| = \int_{-1}^{1}|K(u)|f(x+hu)du$ we have for some constant $c''>0$

$$\sup_{h\in I_n}\sup_{x\in E} E\, h^{-1}|K((x-X_1)/h)| \leq c''.$$

Consequently the exponential bound (A.4) in appendix A implies for n large enough, for all $x\in E$ and all $h\in I_n$

$$P(h^{1/2}\,|\, f_{nh}(x) - Ef_{nh}(x)\,|\geq \varepsilon_n) \leq$$

$$2\exp(-n\varepsilon_n^2/(2K^*\, E\, h^{-1}|K((x-X_1)/h)| + h^{-1/2}\varepsilon_n)) \leq$$

$$2\exp(-c'n\varepsilon_n^2),$$

for some constant $c'>0$. Here K^* is a constant bounding K, i.e. $|K(x)| \leq K^*$ for all x.

Next define subsets $\mathcal{I}_n$ and $\mathcal{E}_n$ of I_n and E where $\mathcal{I}_n$ consists of n^2 equidistant points, the endpoints included, and, similarly, $\mathcal{E}_n$ consists of $n^{1-\delta/4}$ equidistant points including the endpoints. Then for n large enough

$$P(\sup_{h\in \mathcal{I}_n}\sup_{x\in \mathcal{E}_n} h^{1/2}|f_{nh}(x) - Ef_{nh}(x)| \geq \varepsilon_n) \leq$$

$$\sum_{h\in \mathcal{I}_n}\sum_{x\in \mathcal{E}_n} P(h^{1/2}|f_{nh}(x) - Ef_{nh}(x)| \geq \varepsilon_n) \leq$$

$$2n^2n^{1-\delta/4}\exp(-c'n\varepsilon_n^2) =$$

$$2\exp(-c'n\varepsilon_n^2 + (3-\delta/4)\log n),$$

which is summable if we choose $\varepsilon_n = 5c'^{-1}(\log n/n)^{1/2}$. Thus by the Borel-Cantelli theorem we get

$$\limsup_{n\to\infty}\sup_{h\in \mathcal{I}_n}\sup_{x\in \mathcal{E}_n} \varepsilon_n^{-1}h^{1/2}|f_{nh}(x) - Ef_{nh}(x)| \leq 1, \text{ almost surely,}$$

which in its turn implies (2.40) with the sets I_n and E replaced by $\mathcal{I}_n$ and $\mathcal{E}_n$.

We finish the proof by showing that the difference between the supremum over the finite sets and the supremum over the continuous sets vanishes almost surely. Let (h_1,x_1) be a point in $I_n{\times}E$, and let (h_2,x_2) be the nearest point in $\mathcal{I}_n{\times}\mathcal{E}_n$. Then we have $|h_1-h_2|<n^{-2}$ and $|x_1-x_2|<cn^{-1+\delta/4}$ for some constant $c>0$. It suffices to show that

$$(2.43)\qquad |(nh_1)^{1/2}(\log n)^{-1/2}(f_{nh_1}(x_1) - Ef_{nh_1}(x_1)) - (nh_2)^{1/2}(\log\ n)^{-1/2}(f_{nh_2}(x_2) - Ef_{nh_2}(x_2))|$$

converges to zero, uniformly for all points (h_1,x_1) and (h_2,x_2) as described above. Here proposition 2.18 will be instrumental.

Writing $\alpha_n = n^{-2}$ and $\beta_n = cn^{-1+\delta/4}$ we see that (2.43) is bounded by

$$n^{1/2}(h_1\vee h_2)^{1/2}(\log\ n)^{-1/2}|f_{nh_1}(x_1) - Ef_{nh_1}(x_1) - f_{nh_2}(x_2) + Ef_{nh_2}(x_2)| +$$

$$n^{1/2}((h_1 \vee h_2)^{1/2} - (h_1 \wedge h_2)^{1/2})(\log n)^{-1/2} 2K^*(h_1 \wedge h_2)^{-1} =$$

$$n^{1/2}(h_1 \vee h_2)^{-1/2}(\log n)^{-1/2}(h_1 \vee h_2)|f_{nh_1}(x_1) - Ef_{nh_1}(x_1) - f_{nh_2}(x_2) + Ef_{nh_2}(x_2)| + \tag{2.44}$$

$$n^{1/2}((h_1 \vee h_2)^{1/2} - (h_1 \wedge h_2)^{1/2})(\log n)^{-1/2} 2K^*(h_1 \wedge h_2)^{-1},$$

where we have used

$$|f_{nh}(x) - Ef_{nh}(x)| \leq 2K^*/h,$$

for all x and all h>0.

The first term in (2.44) is bounded by

$$n^{1/2}(h_1 \vee h_2)^{-1/2}(\log n)^{-1/2}(\alpha_n + \beta_n)\,(\alpha_n + \beta_n)^{-1}\Omega_n(\alpha_n, \beta_n) \leq$$

$$n^{1/2} n^{1/2-\delta/2}(n^{-2} + cn^{-1+\delta/4})\,(\alpha_n + \beta_n)^{-1}\Omega_n(\alpha_n, \beta_n) \leq$$

$$2cn^{-\delta/4}\,(\alpha_n + \beta_n)^{-1}\Omega_n(\alpha_n, \beta_n) = o(1), \text{ almost surely,}$$

by proposition 2.18. In order to treat the second term in (2.44) notice

$$(h_1 \vee h_2)^{1/2} =$$

$$(h_1 \wedge h_2)^{1/2}(1 + ((h_1 \vee h_2) - (h_1 \wedge h_2))/(h_1 \wedge h_2))^{1/2} \leq$$

$$(h_1 \wedge h_2)^{1/2}(1 + \alpha_n/(h_1 \wedge h_2))^{1/2} \leq$$

$$(h_1 \wedge h_2)^{1/2}(1 + \alpha_n/(h_1 \wedge h_2)) =$$

$$(h_1 \wedge h_2)^{1/2} + \alpha_n (h_1 \wedge h_2)^{-1/2}.$$

Therefore

$$n^{1/2}((h_1 \vee h_2)^{1/2} - (h_1 \wedge h_2)^{1/2})(\log n)^{-1/2} 2K^*(h_1 \wedge h_2)^{-1} \leq$$

$$n^{1/2}\alpha_n (h_1 \wedge h_2)^{-1/2}(\log n)^{-1/2} 2K^*(h_1 \wedge h_2)^{-1} \leq$$

$$2K^*(n(h_1 \wedge h_2))^{-3/2}(\log n)^{-1/2} \leq$$

$$2K^* n^{-3\delta/2}(\log n)^{-1/2} = o(1).$$

Since these bounds don't depend on the h's or the x's we have indeed shown that (2.43) vanishes uniformly and the proof is completed. □

2.5. Proofs.

Proof of lemma 2.1 part (b). Let x_0 be a fixed point and let M be an arbitrary positive number. Recall that by (2.9) we have

$$g(x_0+th,h) = \int_{-\infty}^{\infty} G(t-u)f(x_0+hu)du \ .$$

We omit the proof for $x_0 \notin D$ since then the same Taylor expansion argument as for part (a) can be used. So we assume $x_0 \in D$ and write

$$g(x_0+th,h) = \int_{-\infty}^{0} G(t-u)f(x_0+hu)du + \int_{0}^{\infty} G(t-u)f(x_0+hu)du \ .$$

Next define $r_2^l(t,h)$ by

$$\text{(2.45)} \qquad r_2^l(t,h) := \int_{-\infty}^{0} G(t-u)\{f(x_0+hu) - f(x_0-) - huf'(x_0-) - \tfrac{1}{2}h^2u^2f''(x_0-)\}du$$

and similarly $r_2^r(t,h)$ by

$$\text{(2.46)} \qquad r_2^r(t,h) := \int_{0}^{\infty} G(t-u)\{f(x_0+hu) - f(x_0+) - huf'(x_0+) - \tfrac{1}{2}h^2u^2f''(x_0+)\}du.$$

Then $r_2(t,h) = r_2^l(t,h) + r_2^r(t,h)$. Use (F.3) and the dominated convergence theorem, which can be applied since the integrals in the definition of r_2^r is in fact an integral over a bounded area, to obtain

$$\lim_{n\to\infty} h_n^{-2}r_2^l(t_n,h_n) = 0,$$

for all sequences (h_n) with $0<h_n\leq h'_n$ for all n, and for all sequences (t_n), with $-M\leq t_n\leq M$ for all n. A similar result holds for r_2^r and therefore for r_2. □

Proof of theorem 2.3 part (b). By (2.2) we have

$$b(x_0+th,h) = E\frac{1}{h}K((x_0+th-X_1)/h) - f(x_0+th).$$

Part (b) of lemma 2.1 gives us an expansion of the first term in this expression. The second term can be expanded as follows

$$\begin{aligned} f(x_0+th) \quad = \quad & (f(x_0-) + thf'(x_0-) + \tfrac{1}{2}t^2h^2f''(x_0-))I_{(-\infty,0)}(t) + \\ & (f(x_0+) + thf'(x_0+) + \tfrac{1}{2}t^2h^2f''(x_0+))I_{(0,\infty)}(t) + \\ & r(t,h), \end{aligned}$$

where the remainder term r has the property we have to prove for the remainder r_4 in the theorem. If x_0 is not a singular point of f this follows from a Taylor expansion argument and if x_0 is a singular point it follows from condition (F.3), just as in the preceding proof. Combining these expansions we get

$$
\begin{aligned}
b(x_0+th,h) = {} & f(x_0-)\int_{-\infty}^{0} K(t-u)du + f(x_0+)\int_{0}^{\infty} K(t-u)du \\
& - [f(x_0-)I_{(-\infty,0)}(t) + f(x_0+)I_{(0,\infty)}(t)] + \\
& hf'(x_0-)\int_{-\infty}^{0} uK(t-u)du + hf'(x_0+)\int_{0}^{\infty} uK(t-u)du \\
& - ht[f'(x_0-)I_{(-\infty,0)}(t) + f'(x_0+)I_{(0,\infty)}(t)] + \\
& \tfrac{1}{2}t^2h^2f''(x_0-)\int_{-\infty}^{0} u^2K(t-u)du + \tfrac{1}{2}t^2h^2f''(x_0+)\int_{0}^{\infty} u^2K(t-u)du \\
& - \tfrac{1}{2}h^2t^2\,[f''(x_0-)I_{(-\infty,0)}(t) + f''(x_0+)I_{(0,\infty)}(t)] + \\
& r_4(t,h),
\end{aligned}
$$

where the remainder r_4 has the property claimed in the theorem.

First consider the constant term in this expansion. Since K integrates to one this term is for $t<0$ equal to

$$
\begin{aligned}
& f(x_0-)\int_{t}^{\infty} K(u)du + f(x_0+)\int_{-\infty}^{t} K(u)du - f(x_0-) = \\
& (f(x_0+) - f(x_0-))\int_{-\infty}^{t} K(u)du \;= \\
& b_0(t)\delta^{(0)}(x_0).
\end{aligned}
$$

Next consider the coefficient of h. Using the fact that the integral of $uK(u)$ is equal to zero we see that for $t<0$ this term equals

$$
\begin{aligned}
& hf'(x_0-)\int_{t}^{\infty} (t-u)K(u)du + hf'(x_0+)\int_{-\infty}^{t} (t-u)K(u)du - htf'(x_0-) = \\
& h(f'(x_0+) - f'(x_0-))\int_{-\infty}^{t} (t-u)K(u)du + hf'(x_0-)\int_{-\infty}^{\infty} (t-u)K(u)du - htf'(x_0-)= \\
& hb_1(t)\delta^{(1)}(x_0).
\end{aligned}
$$

The coefficient of h^2 is for $t<0$ equal to

$$\tfrac{1}{2}h^2f''(x_0-)\int_t^{\infty}(t-u)^2K(u)du + \tfrac{1}{2}h^2f''(x_0+)\int_{-\infty}^{t}(t-u)^2K(u)du - \tfrac{1}{2}h^2t^2f''(x_0-) =$$

$$\tfrac{1}{2}h^2(f''(x_0+) - f''(x_0-))\int_{-\infty}^{t}(t-u)^2K(u)du + \tfrac{1}{2}h^2f''(x_0-)\int_{-\infty}^{\infty}(t-u)^2K(u)du - \tfrac{1}{2}h^2t^2f''(x_0-)=$$

$$\tfrac{1}{2}h^2b_2(t)\delta^{(2)}(x_0) + \tfrac{1}{2}h^2f''(x_0-)\int_{-\infty}^{\infty}u^2K(u)du.$$

For t>0 a similar discussion holds. □

Proof of theorem 2.11. The proof is completed by checking condition (C.7) of theorem C.2 in appendix C which means that we have to expand the variance

$$\mathrm{var}\Big(\int_{-\infty}^{\infty}K\big(\tfrac{u-X_1}{h_n}\big)b(u,h_n)w(u)du\Big)$$

for n tending to infinity. If the set D = $\{d_1,d_2,...\}$ denotes the set of singular points of f and the set D_h, defined in (2.4), denotes the set of points at least at distance h of D, then we write,

$$E\Big(\int_{-\infty}^{\infty}K\big(\tfrac{u-X_1}{h_n}\big)b(u,h_n)w(u)du\Big)^2 =$$

$$E\Big(\int_{D_{h_n}}K\big(\tfrac{u-X_1}{h_n}\big)b(u,h_n)w(u)du + \sum_{i=1}^{\infty}\int_{d_i-h_n}^{d_i+h_n}K\big(\tfrac{u-X_1}{h_n}\big)b(u,h_n)w(u)du\Big)^2 =$$

$$E\Big(\int_{D_{h_n}}K\big(\tfrac{u-X_1}{h_n}\big)b(u,h_n)w(u)du\Big)^2 +$$

$$\sum_{i=1}^{\infty}E\Big(\int_{d_i-h_n}^{d_i+h_n}K\big(\tfrac{u-X_1}{h_n}\big)b(u,h_n)w(u)du\Big)^2 +$$

$$2\sum_{i=1}^{\infty}E\Big(\int_{D_{h_n}}K\big(\tfrac{u-X_1}{h_n}\big)b(u,h_n)w(u)du\Big)\Big(\int_{d_i-h_n}^{d_i+h_n}K\big(\tfrac{u-X_1}{h_n}\big)b(u,h_n)w(u)du\Big) + \tag{2.47}$$

$$\sum_{i\neq j}E\Big(\int_{d_i-h_n}^{d_i+h_n}K\big(\tfrac{u-X_1}{h_n}\big)b(u,h_n)w(u)du\Big)\Big(\int_{d_j-h_n}^{d_j+h_n}K\big(\tfrac{u-X_1}{h_n}\big)b(u,h_n)w(u)du\Big).$$

Notice that since w has a bounded support the conditions on f imply that there are only finitely many singular points $d_1,...,d_m$ say, which are in the support of w. These points are the only singular points of f which can give a nonzero contribution to the sums above. All the singular points outside the support of w are at a positive distance from this support which means that their contributions are exactly equal to zero if n is larger than some fixed n_0. Using the fact that K has support [-1,1] and using the expansions of $b(u,h_n)$ given by theorem 2.3 we derive the following bounds,

$$\int_{d-h_n}^{d+h_n} K\left(\frac{u-x}{h_n}\right) b(u,h_n) w(u)\, du =$$

$$h_n \int_{-1}^{1} K\left(t+\frac{d-x}{h_n}\right) b(d+th_n,h_n) w(d+th_n)\, dt = \tag{2.48}$$

$$I_{[d-2h_n,d+2h_n]}(x)\, O(h_n \delta^{(0)}(d) + h_n^2 \delta^{(1)}(d) + h_n^3),$$

for each $d \in D$, and

$$\int_{D_{h_n}} K\left(\frac{u-x}{h_n}\right) b(u,h_n) w(u)\, du =$$

$$\int_{D_{h_n} \cap [x-h_n,x+h_n]} K\left(\frac{u-x}{h_n}\right) b(u,h_n) w(u)\, du = O(h_n^3). \tag{2.49}$$

From (2.48) it is immediately clear that the fourth term of (2.47) vanishes for n large enough. For the third term (2.48) and (2.49) give

$$E\left(\int_{D_{h_n}} K\left(\frac{u-X_1}{h_n}\right) b(u,h_n) w(u)\, du\right)\left(\int_{d_i-h_n}^{d_i+h_n} K\left(\frac{u-X_1}{h_n}\right) b(u,h_n) w(u)\, du\right) =$$

$$\int_{d_i-2h_n}^{d_i+2h_n} \left(\int_{D_{h_n}} K\left(\frac{u-x}{h_n}\right) b(u,h_n) w(u)\, du\right)\left(\int_{d_i-h_n}^{d_i+h_n} K\left(\frac{u-x}{h_n}\right) b(u,h_n) w(u)\, du\right) f(x)\, dx =$$

$$O(4h_n h_n^3 (h_n \delta^{(0)}(d_i) + h_n^2 \delta^{(1)}(d_i) + h_n^3)) =$$

$$O(h_n^5 (\delta^{(0)}(d_i) + h_n \delta^{(1)}(d_i) + h_n^2))$$

It turns out that this term is also asymptotically negligible compared to the first two terms in (2.47). Using the expansion of $b(u,h_n)$ given by theorem 2.3 we can expand the first term as follows,

$$E\left(\int_{D_{h_n}} K\left(\frac{u-X_1}{h_n}\right) b(u,h_n) w(u)\, du\right)^2 =$$

$$\int_{-\infty}^{\infty} \left(\int_{D_{h_n}} K\left(\frac{u-x}{h_n}\right) b(u,h_n) w(u)\, du\right)^2 f(x)\, dx \sim$$

$$\tfrac{1}{4} h_n^4 \left(\int_{-1}^{1} v^2 K(v)\, dv\right)^2 \int_{-\infty}^{\infty} \left(\int_{D_{h_n}} K\left(\frac{u-x}{h_n}\right) f''(u) w(u)\, du\right)^2 f(x)\, dx = \tag{2.50}$$

$$\tfrac{1}{4}h_n^6\Big(\int_{-1}^{1}v^2K(v)dv\Big)^2\int_{-\infty}^{\infty}\Big(\int_{(D_{h_n}-x)/h_n}K(v)f''(x+h_nv)w(x+h_nv)dv\Big)^2f(x)dx\sim$$

$$\tfrac{1}{4}h_n^6\Big(\int_{-1}^{1}v^2K(v)dv\Big)^2\int_{-\infty}^{\infty}f''(x)^2w^2(x)f(x)dx.$$

The last equivalence holds since for each $x\notin D$ the set $(D_{h_n}-x)/h_n$ converges to $(-\infty,\infty)$ which, since f'' is continuous outside D and since w is almost surely continuous, implies that for each fixed $x\notin D$ we have

$$\int_{(D_{h_n}-x)/h_n}K(v)f''(x+h_nv)w(x+h_nv)dv\rightarrow\int_{-\infty}^{\infty}K(v)dv\ f''(x)w(x)=f''(x)w(x),$$

almost surely as a function of x.

Concerning the second term in (2.47), just as in (2.48), for each $d\in D$ again by theorem 2.3 we get,

$$E\Big(\int_{d-h_n}^{d+h_n}K\big(\tfrac{u-X_1}{h_n}\big)b(u,h_n)w(u)du\Big)^2=$$

$$h_n^2\int_{-\infty}^{\infty}\Big(\int_{-1}^{1}K\big(t+\tfrac{d-x}{h_n}\big)b(d+th_n,h_n)w(d+th_n)dt\Big)^2f(x)dx\sim$$

$$h_n^3\int_{-\infty}^{\infty}\Big(\int_{-1}^{1}K(t+v)b(d+th_n,h_n)w(d+th_n)dt\Big)^2f(d+vh_n)dv\sim$$

$$h_n^3\int_{-\infty}^{\infty}\Big(\int_{-1}^{1}K(t+v)\delta^{(0)}(d)b_0(t)w(d+th_n)dt\Big)^2f(d+vh_n)dv\sim$$

$$h_n^3\,\delta^{(0)}(d)^2\Big(\quad f(d-)\int_{-\infty}^{0}\Big(w(d-)\int_{-1}^{0}K(t+v)b_0(t)dt\ +w(d+)\int_{0}^{1}K(t+v)b_0(t)dt\Big)^2dv\ +$$

$$f(d+)\int_{0}^{\infty}\Big(w(d-)\int_{-1}^{0}K(t+v)b_0(t)dt\ +w(d+)\int_{0}^{1}K(t+v)b_0(t)dt\Big)^2dv\Big).$$

If $\delta^{(0)}(d)$ is equal to zero a similar argument gives

$$E\Big(\int_{d-h_n}^{d+h_n}K\big(\tfrac{u-X_1}{h_n}\big)b(u,h_n)w(u)du\Big)^2\sim$$

$$h_n^5\,f(d)\delta^{(1)}(d)^2\int_{-\infty}^{\infty}\Big(w(d-)\int_{-1}^{0}K(t+v)b_1(t)dt\ +\ w(d+)\int_{0}^{1}K(t+v)b_1(t)dt\Big)^2dv,$$

and if both $\delta^{(0)}(d)$ and $\delta^{(1)}(d)$ are equal to zero then we have

$$E\Big(\int_{d-h_n}^{d+h_n} K\Big(\frac{u-X_1}{h_n}\Big)b(u,h_n)w(u)du\Big)^2 = O(h_n^7).$$

In order to derive expansions for the variance let us successively consider the three cases introduced in section 2.3.1. Recall the definition of the quantities $\Delta_w^{(0)}$ and $\Delta_w^{(1)}$,

$$\Delta_w^{(0)} := \sum_{i=1}^{\infty}(w(d_i-) + w(d_i+))\delta^{(0)}(d_i)^2$$

and

$$\Delta_w^{(1)} := \sum_{i=1}^{\infty}(w(d_i-) + w(d_i+))\delta^{(1)}(d_i)^2.$$

In case I we have $\Delta_w^{(0)}>0$ which means that there is at least one singular point d_i with $(w(d_i-)+w(d_i+))\delta^{(0)}(d_i)^2>0$. Since this implies that for such a point either $w(d_i-)\delta^{(0)}(d_i)^2$ or $w(d_i+)\delta^{(0)}(d_i)^2$ is positive we get

$$E\Big(\int_{-\infty}^{\infty} K\Big(\frac{u-X_1}{h_n}\Big)b(u,h_n)w(u)du\Big)^2 \sim$$

$$h_n^3 \sum_{i=1}^{\infty} \delta^{(0)}(d_i)^2 \Big(f(d-) \int_{-\infty}^{0} \Big(w(d_i-) \int_{-1}^{0} K(t+v)b_0(t)dt \; + w(d_i+)\int_{0}^{1} K(t+v)b_0(t)dt\Big)^2 dv +$$

$$f(d_i+)\int_{0}^{\infty} \Big(w(d_i-) \int_{-1}^{0} K(t+v)b_0(t)dt \; + w(d_i+)\int_{0}^{1} K(t+v)b_0(t)dt\Big)^2 dv\Big).$$

The bounds (2.48) and (2.49) imply that the squared expectation,

$$\Big(E\int_{-\infty}^{\infty} K\Big(\frac{u-X_1}{h_n}\Big)b(u,h_n)w(u)du\Big)^2,$$

is asymptotically negligible in this case. Thus we get

$$\mathrm{var}\Big(\int_{-\infty}^{\infty} K\Big(\frac{u-X_1}{h_n}\Big)b(u,h_n)w(u)du\Big) \sim \tfrac{1}{4}h_n^3\sigma_I{}^2.$$

In case II the situation is similar. Here we have $\Delta_w^{(0)}=0$ and $\Delta_w^{(1)}>0$. By our condition that for points d_i in the support of w either $w(d_i-)$ or $w(d_i+)$ is positive the fact that $\Delta_w^{(0)}=0$ implies that all the jumps $\delta^{(0)}(d_i)$ for points in the support of w are equal to zero. Since $\Delta_w^{(1)}$ is positive there is at least one point d_i such that $(w(d_i-)+w(d_i+))\delta^{(1)}(d_i)^2$ is positive. We then have

$$E\Big(\int_{-\infty}^{\infty} K\Big(\frac{u-X_1}{h_n}\Big)b(u,h_n)w(u)du\Big)^2 \sim$$

$$h_n^5 \sum_{i=1}^{\infty} f(d_i)\delta^{(1)}(d_i)^2 \int_{-\infty}^{\infty} \Big(w(d_i-) \int_{-1}^{0} K(t+v)b_1(t)dt + w(d_i+)\int_{0}^{1} K(t+v)b_1(t)dt\Big)^2 dv$$

and since in this case the squared expectation is also negligible we arrive at

$$\operatorname{var}\Big(\int_{-\infty}^{\infty} K\big(\tfrac{u-X_1}{h_n}\big)b(u,h_n)w(u)du\Big) \sim \tfrac{1}{4}h_n^5\sigma_{II}{}^2.$$

In case III the situation is different because the squared expectation is no longer negligible. Here both $\Delta_w^{(0)}$ and $\Delta_w^{(1)}$ are equal to zero. Therefore all $\delta^{(0)}(d_i)$ and $\delta^{(1)}(d_i)$ for points d_i in the support of w are equal to zero. This leads to

$$E\Big(\int_{-\infty}^{\infty} K\big(\tfrac{u-X_1}{h_n}\big)b(u,h_n)w(u)du\Big)^2 \sim$$

$$\tfrac{1}{4}h_n^6\Big(\int_{-1}^{1} v^2K(v)dv\Big)^2 \int_{-\infty}^{\infty} f''(x)^2w(x)^2f(x)dx.$$

Since we have

$$E\Big(\int_{D_{h_n}} K\big(\tfrac{u-X_1}{h_n}\big)b(u,h_n)w(u)du\Big) =$$

$$\int_{-\infty}^{\infty}\Big(\int_{D_{h_n}} K\big(\tfrac{u-x}{h_n}\big)b(u,h_n)w(u)du\Big)f(x)dx =$$

$$h_n\int_{D_{h_n}}\Big(\frac{1}{h_n}\int_{-\infty}^{\infty} K\big(\tfrac{u-x}{h_n}\big)f(x)dx\Big)b(u,h_n)w(u)du \sim$$

$$h_n\int_{D_{h_n}}\big(f(u) + b(u,h_n)\big)b(u,h_n)w(u)du \sim$$

$$h_n\int_{D_{h_n}} f(u)b(u,h_n)w(u)du \sim$$

$$\tfrac{1}{2}h_n^3\int_{-1}^{1} v^2K(v)dv \int_{-\infty}^{\infty} f''(u)w(u)f(u)du,$$

we find

$$\operatorname{var}\Big(\int_{-\infty}^{\infty} K\big(\tfrac{u-X_1}{h_n}\big)b(u,h_n)w(u)du\Big) \sim$$

$$\tfrac{1}{4}h_n^6\Big(\int_{-1}^{1} v^2K(v)dv\Big)^2 \int_{-\infty}^{\infty} f''(x)^2w^2(x)f(x)dx\ -$$

$$\Big(\tfrac{1}{2}h_n^3\int_{-1}^{1} v^2K(v)dv \int_{-\infty}^{\infty} f''(u)w(u)f(u)du\Big)^2 =$$

$$\tfrac{1}{4}h_n^6\sigma_{III}{}^2.$$

This completes the proof of theorem 2.11. □

Proof of lemma 2.19. First we use the Bernstein inequality for the binomial distribution, i.e. inequality (A.3) in appendix A, to derive the following exponential bound. For all $x\in E_{h_n''}$, all $t_n<t<T$ and all $0\leq s\leq t$ such that $x+s\in E_{h_n''}$ we have for any sequence (ε_n) tending to infinity and n sufficiently large

$$P(\, t^{-1/2}\,|\, V_n(x+s) - V_n(x)\,|\geq \varepsilon_n\,) \leq$$

$$P(\,|\, F_n(x+s) - F_n(x) - (\, F(x+s) - F(x))\,|\geq n^{-1/2}t^{1/2}\varepsilon_n\,) \leq$$

$$P\big(\tfrac{1}{n}\,\big|\sum_{i=1}^{n} I_{(x,x+s]}(X_i) - P(x<X_i\leq x+s)\big|\geq n^{-1/2}t^{1/2}\varepsilon_n\,\big) \leq$$

$$2\exp\big(\,-\,\varepsilon_n{}^2/2(c'+(nt)^{-1/2}\varepsilon_n)\big) \leq$$

$$2\exp(-\,\varepsilon_n),$$

where c' is a constant bounding f, i.e. $0\leq f(x)\leq c'$ for all x. We have used $0\leq P(x<X_i\leq x+s)\leq c's\leq c't$, and $(nt)^{-1/2}\leq(nt_n)^{-1/2}\rightarrow 0$.

Next let J_n denote the interval $[t_n,T]$. Since E is bounded and since h_n'' converges to zero the intervals $E_{h_n''}$ are uniformly bounded. Hence there exists a positive constant M such that the interval [-M,M] covers both $E_{h_n''}$ and J_n for all n. Let G_n denote the grid of $2n^3$ points of [-M,M], given by $g_i=iMn^{-3}$, $i=-n^3+1,\ldots,\, n^3$. Notice that consecutive points have a distance equal to Mn^{-3}. It follows for n sufficiently large

$$P\Big(\sup_{t\in J_n\cap G_n}\;\sup_{x\in E_{h_n''}\cap G_n}\;\sup_{s\in[0,t]\cap G_n}\; t^{-1/2}\,|\, V_n(x+s) - V_n(x)\,|\geq \varepsilon_n\Big) \leq$$

$$\sum_{t\in J_n\cap G_n}\;\sum_{x\in E_{h_n''}\cap G_n}\;\sum_{s\in[0,t]\cap G_n} P\big(t^{-1/2}\,|\, V_n(x+s) - V_n(x)\,|\geq \varepsilon_n\big) \leq$$

$$(2n^3)^3\exp(-\varepsilon_n) =$$

$$8\exp(\,-\varepsilon_n + 9\log n\,),$$

which is summable if we take ε_n equal to $11\log n$, which we assume from now on.

Let S_n denote the supremum over the discrete sets,

$$S_n := \varepsilon_n{}^{-1}\sup_{t\in J_n\cap G_n}\;\sup_{x\in E_{h_n''}\cap G_n}\;\sup_{s\in[0,t]\cap G_n}\; t^{-1/2}\,|\, V_n(x+s) - V_n(x)\,|\,,$$

and S_n the supremum over the continuous sets,

$$\mathrm{S}_n := \varepsilon_n^{-1} \sup_{t\in J_n} \sup_{x\in E_{h_n''}} \sup_{s\in[0,t]} t^{-1/2} \mid V_n(x+s) - V_n(x) \mid .$$

By the Borel-Cantelli theorem we have for S_n ,

$$\limsup_{n\to\infty} S_n \leq 1, \text{ almost surely.}$$

It remains to show that the difference between S_n and S_n vanishes almost surely, since then

$$\limsup_{n\to\infty} \mathrm{S}_n \leq 1, \text{ almost surely.}$$

From this result lemma 2.19 is immediate since

$$\frac{1}{\log n} \sup_{t_n\leq t<T} t^{-1/2}\omega_n(t) \leq 11\mathrm{S}_n,$$

we would have established the result of lemma 2.19. To show that S_n - S_n vanishes almost surely define the set A

$$A := \{d_n \leq Mn^{-3}, \text{ infinitely often}\},$$

where d_n denotes the smallest spacing of the sample $X_1,...,X_n$. It follows from a result of Devroye (1982) that the probability of A is zero. Actually Devroye condiders uniform spacings but since the density f is bounded it also follows for the spacings of the sample $X_1,...,X_n$. On the complement of A for $t\in J_n$, $x\in E_{h_n''}$and $s\in[0,t]$ the value of

$$\varepsilon_n^{-1}t^{-1/2} \mid V_n(x+s) - V_n(x) \mid =$$

$$\varepsilon_n^{-1}t^{-1/2}n^{1/2} \mid F_n(x+s) - F_n(x) - (F(x+s) - F(x)) \mid$$

changes for n larger than a certain random index N_0 at most

$$\varepsilon_n^{-1}t_n^{-1/2}n^{1/2}(3n^{-1} - 3c'Mn^{-3}) =$$

$$3\varepsilon_n^{-1}(nt_n)^{-1/2}(1 + c'Mn^{-2}) = o(1),$$

if we replace t,x, and s by their nearest points in the interval J_n, $E_{h_n''}$ and [0,t] which also lie on the grid G_n. Hence on the complement of A we have

$$S_n \leq \mathrm{S}_n \leq S_n + o(1), n\to\infty,$$

and since the complement of A has probability one we have indeed shown

$$\mathrm{S}_n - S_n = o(1), \text{ almost surely,}$$

which completes the proof of lemma 2.19. □

3. BANDWIDTH SELECTION BY LIKELIHOOD CROSS-VALIDATION.

3.1. Introduction and results.

Results in the previous chapter show that optimal bandwidths for kernel estimators depend on the unknown density f. One way to avoid this problem is to design procedures which compute a bandwidth, $H_n = H_n(X_1,...,X_n)$, from the sample $X_1,...,X_n$. For large sample sizes these bandwidths should be close to the optimal ones. Kernel estimators using these bandwidths are called *automatic* or *data adaptive*. Two such data adaptive bandwidth selection methods are *likelihood cross-validation*, which originates from a likelihood approach to the problem, and *least squares cross-validation*. Least squares cross-validation is briefly discussed in section 3.1.1. Likelihood cross-validation has a history of trial and error. A review of its development is given in section 3.1.2. For a comparison of cross-validation techniques in a more general setting see Marron (1987). Next, in section 3.1.3, we give a heuristic derivation of our results on the rates of convergence to zero and the asymptotic distribution of the bandwidths selected by likelihood cross-validation. These results are proved in the further sections of this chapter.

3.1.1. Least squares cross-validation.

Let us first consider the least squares cross-validation. Suppose that our aim is to find bandwidths and corresponding density estimates with a small integrated squared error. In order to do so write

$$\begin{aligned}&\mathrm{MISE}_n(h) = \\ &E\int_{-\infty}^{\infty} (f_{nh}(u) - f(u))^2 w(u)du = \\ &E\int_{-\infty}^{\infty} f_{nh}^2(u)w(u)du - 2\,E\int_{-\infty}^{\infty} f_{nh}(u)f(u)w(u)du + \int_{-\infty}^{\infty} f^2(u)w(u)du,\end{aligned}$$

where w is a nonnegative weight function. Since the third term is independent of h it suffices to find a bandwidth which minimizes

$$E\int_{-\infty}^{\infty} f_{nh}^2(u)w(u)du - 2\,E\int_{-\infty}^{\infty} f_{nh}(u)f(u)w(u)du, \tag{3.1}$$

an expression depending on the unknown density f. The least squares cross-validation method results in an unbiased estimator of (3.1). So we can estimate (3.1) as a function of h and compute the value of h which minimizes the estimate.

Define the "leave one out estimator" based on the sample $X_1,...,X_n$ with X_i left out by

$$f_{nh}^{(i)}(x) := \frac{1}{(n-1)h}\sum_{j=1,j\neq i}^{n} K((x-X_j)/h)), \ -\infty < x < \infty. \tag{3.2}$$

Then

$$E\, f_{nh}^{(i)}(x) = \frac{1}{h} E\, K((x - X_1)/h) = E\, f_{nh}(x),$$

and the independence of $f_{nh}^{(i)}(x)$ and X_i implies

$$E \frac{2}{n} \sum_{i=1}^{n} f_{nh}^{(i)}(X_i)w(X_i) =$$

$$\frac{2}{n} \sum_{i=1}^{n} E\, f_{nh}^{(i)}(X_i)w(X_i) =$$

$$2 \int_{-\infty}^{\infty} E\, f_{nh}^{(1)}(u)w(u)f(u)du =$$

$$2 \int_{-\infty}^{\infty} E\, f_{nh}(u)w(u)f(u)du =$$

$$2\, E \int_{-\infty}^{\infty} f_{nh}(u)f(u)w(u)du.$$

Therefore

$$\text{(3.3)} \qquad LS_n(h) := \int_{-\infty}^{\infty} f_{nh}^{2}(u)w(u)du - \frac{2}{n} \sum_{i=1}^{n} f_{nh}^{(i)}(X_i)w(X_i),\ h > 0,$$

is an unbiased estimator of (3.1). For $w \equiv 1$ this reduces to

$$LS_n(h) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{h} K^{(2)}(X_i - X_j)/h) - \frac{2}{n(n-1)} \sum_{i \neq j} \frac{1}{h} K(X_i - X_j)/h),$$

where $K^{(2)}$ denotes the convolution of K with itself. The factor 2/((n(n-1)) is often replaced by $2/n^2$.

This method is introduced and studied by Rudemo (1982) and Bowman (1984). Further relevant references are Hall (1983a, 1983b), Stone (1984), Scott (1985), Burman (1985), Hall & Marron (1987a, 1987b) and Scott & Terrell (1987). Silverman (1986) also considered computational aspects of the method.

Hall (1983a) obtained the first asymptotic optimality result for densities f with a finite second moment and a continuous square integrable second derivative. Generalizing this result Stone (1984) showed that the optimality property holds for all bounded densities f. In the univariate case and for kernels satisfying condition K in section 2.1 the theorem states the following.

Theorem 3.1. (Stone 1984). *If* K *is Lipschitz of order β, i. e. for some positive constants* β *and* c

$$|K(y) - K(x)| \leq c|y - x|^{\beta}, \textit{for all real } x \textit{ and } y,$$

then we have for all bounded densities f *on the real line*

$$(3.4) \qquad \lim_{n\to\infty} \frac{\mathrm{ISE}_n(H_n)}{\inf\limits_h \mathrm{ISE}_n(h)} = 1, \textit{ almost surely,}$$

where H_n *is the bandwidth obtained by least squares cross-validation, i.e. the bandwidth which minimizes* (3.3) *and* $\mathrm{ISE}_n(h)$ *equals*

$$(3.5) \qquad \int_{-\infty}^{\infty} (f_{nh}(u) - f(u))^2 du,$$

the integrated squared error of the kernel estimate f_{nh}.

This theorem shows that asymptotically the bandwidths obtained by least squares cross-validation perform as well as the best possible deterministic ones.

The rate of convergence in (3.4) was investigated by Hall & Marron (1987a, 1987b). If H_n^* denotes the random bandwidth which minimizes the integrated squared error (3.5) then under some smoothness conditions on K and f, essentially our smooth case III in chapter 2, they show

$$(3.6) \qquad \frac{H_n - H_n^*}{H_n^*} = O_p(n^{-1/10}),$$

and

$$\frac{\mathrm{ISE}_n(H_n) - \mathrm{ISE}_n(H_n^*)}{\mathrm{ISE}_n(H_n^*)} = O_p(n^{-1/5}).$$

In spite of the nice asymptotic optimality result (3.4), the convergence is very slow. However, Hall and Marron show that no data adaptive bandwidth selection method can have a faster convergence.

Before we discuss likelihood cross-validation we briefly mention other data adaptive methods. Silverman (1978a, 1986) gives a graphical method to compute bandwidths with good properties with respect to the supremum distance loss function, the so called test graph method. Scott & Factor (1981) and Bowman (1985) compare several other data adaptive methods by means of simulation studies.

3.1.2. The likelihood approach to bandwidth selection.

Again we consider the problem of selecting a bandwidth for a kernel estimator with a kernel satisfying condition K. Now we argue as follows. A "good" bandwidth h will give a large value to the "likelihood" L_n, defined by

$$L_n(h) := \prod_{i=1}^{n} f_{nh}(X_i).$$

This suggests that we should use the value of h which maximizes L_n over $[0,\infty)$. However, we can't use this value because it is always equal to zero. This can be seen from the inequalities

$$(3.7) \qquad f_{nh}(X_i)) = \frac{1}{nh}\sum_{j=1}^{n} K((X_i-X_j)/h)) \geq \frac{K(0)}{nh}$$

and

$$L_n(h) \geq (K(0)/(nh))^n,$$

which show that $L_n(h)$ tends to infinity if h decreases to zero. Recognizing this phenomenon Habbema, Hermans & Van de Broek (1974) and Duin (1976) proposed to replace L_n by

$$\text{(3.8)} \qquad LCV_n(h) := \prod_{i=1}^{n} f_{nh}^{(i)}(X_i),$$

where $f_{nh}^{(i)}$ is the "leave one out" kernel estimator defined by (3.2) in the previous section. The value of h which maximizes LCV_n is always finite, since for n fixed we have

$$0 < h < \max_{i=1,\dots,n} \min_{j \neq i} |X_i - X_j| \Rightarrow LCV_n(h) = 0$$

and

$$0 \leq f_{nh}^{(i)}(X_i) \leq K^*/h \rightarrow 0 \text{ for } h \rightarrow \infty,$$

where we assume that the kernel is bounded by $K^*>0$. In this way we lose the i-th term in (3.7), which is exactly the term which made it converge to infinity for h tending to zero. This technique is called likelihood cross-validation or Kullback-Leibler cross-validation.

The first undesirable property was reported by Schuster & Gregory (1981). Let H_n denote the positive value of h which maximizes $LCV_n(h)$, then, since the kernel K has a support equal to [-1,1], the next inequality holds. We have

$$H_n \geq X_{n:n} - X_{n-1:n} \, ,$$

where $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$ denotes the ordered sample. This inequality follows from the fact that $LCV_n(h)$ is equal to zero for all bandwidths h with $0 < h \leq X_{n:n} - X_{n-1:n}$, since for these bandwidths the term in the product (3.8) corresponding to the largest sample point $X_{n:n}$ is equal to zero. It follows that the computed bandwidth is always at least equal to the difference between the largest sample point and the second to largest. For certain densities f however this difference converges almost surely to infinity. Moreover, these densities are by no means pathalogical. It turns out that densities with an exponential tail form the border line. For densities with heavier tails the bandwidths H_n converge almost surely to infinity and therefore produce inconsistent estimates.

One possibility to avoid the problem discussed above is to restrict attention to densities with a compact support. If we know that f does not have a compact support we can always disregard all observations outside some bounded interval E, next estimate the probability of this interval and use likelihood cross-validation to compute a bandwidth for a kernel estimate of the density, conditional on being in E. Chow, Geman & Wu (1983) and Devroye & Györfi (1985) prove some positive results concerning the estimation of bounded support densities. An alternative to avoid the tail problems is to maximize the product

(3.9) $$\mathrm{LCV}_n(h) := \prod_{i:X_i\in E} f_{nh}^{(i)}(X_i),$$

where E is a bounded interval on the real line. Note that this definition coincides with (3.8) if we can take E equal to the support of f. Here we evaluate the leave one out estimators $f_{nh}^{(i)}$ only in the points X_i in the interval E, instead of in all the points as we did in the original definition of LCV_n. It is important to notice that here $f_{nh}^{(i)}$ is still based on the whole sample $X_1,\dots,X_n$ minus X_i, contrary to above where it was based on observations in E only! By maximizing (3.9) we aim at finding a good bandwidth for estimating f on the interval E rather than on the whole real line as in the original definition (3.8). Accepting this restriction indeed avoids the tail problems discussed above, but instead we are faced with the next property reported in Hall (1982). The theorem is reformulated to hold for kernels satisfying condition K.

Theorem 3.2. (Hall 1982). *Let* $E=[a,b]$, $-\infty < a < b < \infty$. *Assume that* f *is twice continuously differentiable on* $(a-\varepsilon,b+\varepsilon)$ *for some positive* ε. *Furthermore assume that* f *is bounded and that* f *is bounded away from zero on* E. *Then the bandwidths computed by maximizing* LCV_n, *as defined by* (3.9), *are of order* $n^{-1/3}$ *if* $f'(b) < f'(a)$, *and they are much larger if* $f'(b) > f'(a)$. *In the last case we might even have inconsistency.*

Notice that in neither case the order is $n^{-1/5}$ which is the optimal one for the integrated squared error criterion (see (2.25)). Also the dependence on the derivatives in the endpoints of E is very undesirable. However, Marron (1985) showed that if we maximize a modification $\mathrm{LCV}_n^c(h)$, instead of $\mathrm{LCV}_n(h)$, this behavior can be avoided. Then we even achieve asymptotic optimality with respect to a weighted integrated squared error with respect to the weight function $f^{-1}I_E$. We obtain $\mathrm{LCV}_n^c(h)$ by multiplying LCV_n by a correction factor,

$$\mathrm{LCV}_n^c(h) \;:=\; \mathrm{LCV}_n(h)\exp\Big(-n\int_E f_{nh}(u)du\Big) = \mathrm{LCV}_n(h)\exp\Big(-\sum_{i=1}^n \frac{1}{h}\int_E K((u-X_i)/h)du\Big).$$

A heuristic motivation for this correction factor is given in section 3.1.3. The corrected method has the following optimality property. The theorem is reformulated to hold for kernels satisfying condition K.

Theorem 3.3. (Marron 1985). *Let* $E=[a,b]$, $-\infty<a<b<\infty$. *Suppose f is bounded away from zero on* E *and suppose that* f *satisfies a Lipschitz condition,*

$$|f(x) - f(y)| \le M|x - y|^{\gamma}, \textit{for all } x,y,$$

for some positive constants M *and* γ. *If* H_n^c *denotes the value of* h *which maximizes* $\mathrm{LCV}_n^c(h)$ *over the set* $I_n=[h_n',h_n'']$, *where* $h_n'=n^{-1+\sigma}$ *and* $h_n''=n^{-\sigma}$ *for some* $\sigma>0$, *then*

(3.10) $$\frac{\mathrm{ISE}_n(H_n^c)}{\inf\limits_{h\in[h_n',h_n'']} \mathrm{ISE}_n(h)} \to 1 \qquad \textit{, almost surely,}$$

and similarly

$$(3.11)\qquad \frac{\mathrm{MISE}_n(H_n^c)}{\inf\limits_{h\in[h_n',h_n'']} \mathrm{MISE}_n(h)} \to 1 \quad \textit{, almost surely.}$$

Here the integrated squared error is defined by

$$\mathrm{ISE}_n(h) = \int_E (f_{nh}(x)-f(x))^2 f^{-1}(x)dx,$$

and $\mathrm{MISE}_n(h)$ *as its expectation.*

This theorem was the first asymptotic optimality result for the likelihood cross-validation method. Just like theorem 3.1 it says that the random bandwidths computed by cross-validation asymptotically perform just as well as the best deterministic ones except that here we are dealing with a weighted integrated squared error. In section 3.1.3 we give an heuristic explanation for the appearance of this particular weighted integrated squared error.

The method studied by Chow, Geman & Wu and Devroye & Györfi differs from the one studied by Hall and Marron in one important aspect. Apart of course from the correction factor in Marron's modification, Hall and Marron assume that the interval E = [a,b] is strictly contained in the support of f in the sense that both the endpoints a and b are strictly inside the support. Chow, Geman & Wu and Devroye & Györfi study the case where E is equal to the support of f. The results described in the next section show that this causes a quite different behavior.

Theorems 3.1 and 3.3 show that the two cross-validation methods have optimality properties with respect to appropriate (mean) integrated squared error loss functions. For these loss functions the choice of the kernel is relatively unimportant. Consequently we consider bounded support kernels only. It should be noted however that things change considerably if instead we want to minimize the Kullback-Leibler distance between our estimate and the true density, which can be desirable for instance in problems of discrimination. Likelihood cross-validation is studied from this point of view in Hall (1987a,1987b). Actually in Hall (1987a) it is shown that in this context the choice of the kernel is important and that it is unwise to use kernels with a compact support.

3.1.3. Likelihood cross-validation: heuristics and results.

The original likelihood cross-validation method prescribes that we maximize the function $\mathrm{LCV}_n(h)$ given by (3.9). The main ingredients in the proofs in Hall (1982) and Marron (1985) are expansions of the logarithm of this function. Using such expansions they prove theorem 3.2, Hall's surprising theorem about the original method, and theorem 3.3, Marron's optimality result. In this section by heuristics in the same spirit we present our results concerning likelihood cross-validation, both uncorrected and corrected. Later sections contain rigorous proofs of these results. The basic theme of these proofs is the analysis of the derivative of the logarithm of $\mathrm{LCV}_n(h)$. We assume that the density f satisfies condition F, so we also consider non-smooth densities. We don't impose the restriction required by Hall and Marron that the set E is strictly contained in the support of f. On the other hand we do also have to require that f is bounded away from zero on E.

Let I_n denote the interval $[h_n', h_n'']$, where $h_n' = n^{-1+\sigma}$ and $h_n'' = n^{-\sigma}$ for some $\sigma > 0$. From now on we assume $n \to \infty$, $h \to 0$ and $nh \to \infty$. Following Hall and Marron we write

$$\frac{1}{n}\log(LCV_n(h)) =$$

$$\frac{1}{n}\sum_{i=1}^{n} \log(f_{nh}^{(i)}(X_i))I_E(X_i) =$$

$$\frac{1}{n}\sum_{i=1}^{n} \log(f(X_i))I_E(X_i) + \frac{1}{n}\sum_{i=1}^{n} \log\Big(1 + \frac{f_{nh}^{(i)}(X_i) - f(X_i)}{f(X_i)}\Big)I_E(X_i).$$

Since the first term is independent of h the problem is to maximize

$$(3.12) \qquad \frac{1}{n}\sum_{i=1}^{n} \log(1 + \Delta_{ni}(X_i,h))I_E(X_i),$$

where $\Delta_{ni}(x,h)$ is defined by

$$\Delta_{ni}(x,h) := \frac{f_{nh}^{(i)}(x) - f(x)}{f(x)}, \quad i = 1,\ldots,n.$$

Defining

$$(3.13) \qquad g(x) := \log(1 + x) - x + \frac{1}{2}x^2,$$

we can rewrite (3.12) as

$$(3.14) \qquad \frac{1}{n}\sum_{i=1}^{n} \Delta_{ni}(X_i,h)I_E(X_i) - \frac{1}{2}\frac{1}{n}\sum_{i=1}^{n} \Delta_{ni}^2(X_i,h)I_E(X_i) + \frac{1}{n}\sum_{i=1}^{n} g(\Delta_{ni}(X_i,h))I_E(X_i).$$

Now assume that the variation in (3.14) is asymptotically negligible compared to the expectation, in the sense that, asymptotically, by maximizing $LCV_n(h)$ we are maximizing the expectation of (3.14). We don't give a proof of this assumption. However, proofs of the results coming from this heuristic approach are given in sections 3.2 to 3.5.

The expectations of the first two terms in (3.14) are easily computed. Since

$$\frac{1}{n}\sum_{i=1}^{n} E(\Delta_{ni}(X_i,h)I_E(X_i) \mid X_j, j=1,\ldots,n, j \neq i) =$$

$$\frac{1}{n}\sum_{i=1}^{n} \int_E (f_{nh}^{(i)}(u) - f(u))f^{-1}(u)f(u)du =$$

$$\frac{1}{n}\sum_{i=1}^{n} \int_E \Big(\frac{1}{(n-1)h}\sum_{j=1,j\neq i}^{n} K((u-X_j)/h)\Big)du - \int_E f(u)du =$$

$$\begin{aligned}
&\frac{1}{n(n-1)h}\sum_{j=1}^{n}\sum_{i=1,i\neq j}^{n}\int_E K((u-X_j)/h)\,du - \int_E f(u)du = \\
&\frac{1}{nh}\sum_{j=1}^{n}\int_E K((u-X_j)/h)\,du - \int_E f(u)du = \\
&\int_E f_{nh}(u)du - \int_E f(u)du,
\end{aligned}\tag{3.15}$$

the expectation of the first term in (3.14) is

$$E\int_E (f_{nh}(u)du - f(u))du = \int_E b(u,h)du, \tag{3.16}$$

where $b(u,h)$ is the bias of the kernel estimator f_{nh} at the point u. Since theorem 2.3 gives uniform expansions of the bias function we can also derive expansions for (3.16). To obtain the expectation of the second term in (3.14) note that

$$\begin{aligned}
&-\frac{1}{2}\frac{1}{n}\sum_{i=1}^{n} E(\Delta_{ni}^2(X_i,h)I_E(X_i) \mid X_j,\ j=1,\ldots,n,\ j\neq i) = \\
&-\frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}\int_E (f_{nh}^{(i)}(u) - f(u))^2 f^{-2}(u)f(u)du.
\end{aligned}$$

Therefore the expectation is given by

$$\begin{aligned}
&-\frac{1}{2}\frac{1}{n}\sum_{i=1}^{n} E\int_E (f_{nh}^{(i)}(u) - f(u))^2 f^{-1}(u)du = \\
&-\frac{1}{2}E\int_E (f_{nh}^{(1)}(u) - f(u))^2 f^{-1}(u)du \approx \qquad (3.17)\\
&-\frac{1}{2}E\int_E (f_{nh}(u) - f(u))^2 f^{-1}(u)du.
\end{aligned}$$

Expansions of this mean integrated squared error are given by theorem 2.8. From (3.16) and (3.17) we conclude that by maximizing $\log(LCV_n(h))$ is asymptotically equivalent to maximizing

$$\int_E b(u,h)du - \frac{1}{2}E\int_E (f_{nh}(u) - f(u))^2 f^{-1}(u)du + E\, g(\Delta_{n1}(X_1,h)I_E(X_1)). \tag{3.18}$$

So if the second term dominates the other two terms then we are asymptotically minimizing the weighted mean integrated suared error

$$E\int_E (f_{nh}(u) - f(u))^2 f^{-1}(u)du. \tag{3.19}$$

However it turns out that this not always true.

Before we proceed with considering separate cases note that at this stage we can also show the intuition behind Marron's correction term. From the definition of $\mathrm{LCV}_n^c(h)$ we have

$$\frac{1}{n}\log(\mathrm{LCV}_n^c(h)) = \frac{1}{n}\log(\mathrm{LCV}_n(h)) - \int_E f_{nh}(u)du.$$

By (3.16) and (3.18) maximizing $\mathrm{LCV}_n^c(h)$ is asymptotically equivalent to maximizing

$$-\int_E f(u)du - \tfrac{1}{2}E\int_E (f_{nh}(u) - f(u))^2 f^{-1}(u)du + E\ g(\Delta_{n1}(X_1,h)I_E(X_1)).$$

Note that the first term is independent of h. This means that in those cases where the third term is negligible compared to the second term, we are minimizing the weighted mean integrated squared error (3.19).

Let E be a bounded interval [a,b], $-\infty < a < b < \infty$, and let us again consider the three cases introduced in section 2.3.1. If we take the weight function w equal to $f^{-1}I_E$ then these cases were defined by

case I : $\Delta^{(0)} > 0$,

case II : $\Delta^{(0)} = 0$ and $\Delta^{(1)} > 0$,

case III : $\Delta^{(0)} = \Delta^{(1)} = 0$,

where $\Delta^{(0)}$ and $\Delta^{(1)}$ are given by

$$\Delta^{(0)} = \sum_{i=1}^{m} (f(d_i-)^{-1}+f(d_i+)^{-1})\delta^{(0)}(d_i)^2 + f(a+)^{-1}\delta^{(0)}(a)^2 + f(b-)^{-1}\delta^{(0)}(b)^2$$

$$\Delta^{(1)} = \sum_{i=1}^{m} (f(d_i-)^{-1}+f(d_i+)^{-1})\delta^{(1)}(d_i)^2 + f(a+)^{-1}\delta^{(1)}(a)^2 + f(b-)^{-1}\delta^{(1)}(b)^2.$$

Here $d_1,\dots,d_m$ denote the singular points of f in the open interval (a,b). Further we assume without proof that for the cases II and III we have

$$(3.20) \qquad \sup_{i=1,\dots,n}\ \sup_{x\in E} |\Delta_{ni}(x,h)| \to 0, \text{ almost surely.}$$

By $|g(x)|\le|x|^3$, for x small enough, this implies that the third term in (3.18) is negligible. Since cases II and III correspond to densities which are smooth on E, having at most kinks, condition (3.20) is not an unreasonable assumption. This condition is not satisfied for case I. In that case there are two possibilities. If there is at least one jumping point d in (a,b) then

$$\sup_{i=1,\dots,n}\ \sup_{x\in E} |\Delta_{ni}(x,h)| \ge \tfrac{1}{2}\delta^{(0)}(d)\,(\inf_{x\in E} f)^{-1} > 0,$$

and if one of the endpoints of E is a jumping point then (3.20) also can't be valid.

In case III using the expansions given by theorem 2.3 and theorem 2.8 we see that (3.18) is asymptotically equivalent to

$$\tfrac{1}{2}h^2 \int_{-1}^{1} u^2K(u)du \int_E f''(u)du +$$

$$-\tfrac{1}{2}\left\{\tfrac{1}{4}h^4 \left(\int_{-1}^{1} u^2K(u)du\right)^2 \int_E f''(u)^2f^{-1}(u)du + \frac{b-a}{nh} \int_{-1}^{1} K^2(u)du\right\} \approx$$

$$\tfrac{1}{2}h^2(f'(b) - f'(a)) \int_{-1}^{1} u^2K(u)du - \frac{b-a}{2nh} \int_{-1}^{1} K^2(u)du. \tag{3.21}$$

This is exactly n times the expansion derived by Hall (1982) to prove theorem 3.2. Clearly if $f'(b) - f'(a) > 0$ then (3.21) is an increasing function of h which does not have a maximum . It does have a maximum if $f'(b) - f'(a) <0$. The point h where the maximum is attained is of order $n^{-1/3}$.

Next we consider case II. Let $d_1,\ldots,d_m$ denote the points in (a,b) where f has a kink and recall that in case II there are no jumping points of f in E. Let D_h denote the set of points which are at least at a distance h of the singular points of f. Then theorem 2.3 gives the following expansion for the first term in (3.18),

$$\int_E b(u,h)du = \int_a^b b(u,h)du =$$

$$\int_a^{a+h} b(u,h)du + \int_{b-h}^{b} b(u,h)du + \sum_{i=1}^{m} \int_{d_i-h}^{d_i+h} b(u,h)du + \int_{D_h\cap[a,b]} b(u,h)du =$$

$$h \int_0^1 b(a+th,h)dt + h \int_{-1}^{0} b(b+th,h)dt + h \sum_{i=1}^{m} \int_{-1}^{1} b(d_i+th,h)dt + \int_{D_h\cap[a,b]} b(u,h)du \approx \tag{3.22}$$

$$h^2\delta^{(1)}(a)\int_0^1 b_1(t)dt + h^2\delta^{(1)}(b) \int_{-1}^{0} b_1(t)dt + h^2\sum_{i=1}^{m}\delta^{(1)}(d_i) \int_{-1}^{1} b_1(t)dt +$$

$$\tfrac{1}{2}h^2 \int_{-1}^{1} u^2K(u)du \int_a^b f''(u)du =$$

$$h^2 \left(\delta^{(1)}(a) + \delta^{(1)}(b) + 2\sum_{i=1}^{m}\delta^{(1)}(d_i)\right)\int_0^1 b_1(t)dt + \tfrac{1}{2}h^2 \int_{-1}^{1} u^2K(u)du \int_a^b f''(u)du.$$

The terms of order h^2 in (2.13) don't appear in this expansion because they are integrated over intervals of length 2h. We have also used

$$\int_{D_h\cap[a,b]} f''(u)du \approx \int_a^b f''(u)du.$$

Because f' can be discontinuous on (a,b) this integral is not necessarily equal to f '(b) - f '(a). By theorem 2.8 it now follows that in this case we are asymptotically maximizing

$$h^2\left\{\left(\delta^{(1)}(a)+\delta^{(1)}(b)+2\sum_{i=1}^{m}\delta^{(1)}(d_i)\right)\int_0^1 b_1(t)dt+\tfrac{1}{2}h^2\int_{-1}^{1}u^2K(u)du\int_a^b f\,''(u)du\right\}+$$

$$-\tfrac{1}{2}\left\{h^3\Delta^{(1)}\int_0^1 b_1^2(t)dt+\frac{b-a}{nh}\int_{-1}^{1}K^2(u)du\right\}\approx$$

$$h^2\left\{\left(\delta^{(1)}(a)+\delta^{(1)}(b)+2\sum_{i=1}^{m}\delta^{(1)}(d_i)\right)\int_0^1 b_1(t)dt+\tfrac{1}{2}h^2\int_{-1}^{1}u^2K(u)du\int_a^b f\,''(u)du\right\}+$$

$$-\frac{b-a}{2nh}\int_{-1}^{1}K^2(u)du,$$

which leads for the uncorrected method to the same type of behavior as in case III above.

Since in cases II and III the third term in (3.18) is negligible the corrected method indeed asymptotically minimizes the weighted mean integrated squared error (3.19). This corresponds to Marron's optimality result given in theorem 3.3. Notice that in cases II and III the density f satisfies a Lipschitz condition on an ε-neighborhood of [a,b] for some ε small enough. If the kernel K has a bounded support then this property can replace the condition in theorem 3.3 that f should be Lipschitz on the whole real line.

Finally we consider case I. Suppose that $d_1,\ldots, d_m$ denote the jumping points of f in (a,b) then we have similarly to (3.22)

$$\int_E b(u,h)du \approx$$

$$h\delta^{(0)}(a)\int_0^1 b_0(t)dt+h\delta^{(0)}(b)\int_{-1}^{0} b_0(t)dt+h\sum_{i=1}^{m}\delta^{(0)}(d_i)\int_{-1}^{1} b_0(t)dt=$$

$$h\,(\delta^{(0)}(a)-\delta^{(0)}(b))\int_0^1 b_0(t)dt,$$

since b_0 is odd. Notice that since b_0 is negative on [0,1] the integral above is also negative. For the third term in (3.18) we have for h small enough

$$E\,g(\Delta_{n1}(X_1,h))I_E(X_1)=$$

$$E\,g((f_{nh}^{(1)}(X_1)-f(X_1))f^{-1}(X_1))I_E(X_1)=$$

$$\int_E E\;g((f_{nh}^{(1)}(u)-f(u))f^{-1}(u))f(u)du=$$

$$\Big(\int_{a}^{a+h} + \int_{b-h}^{b} + \sum_{i=1}^{m}\int_{d_i-h}^{d_i+h}\Big)\, E\, g((f_{nh}^{(1)}(u) - f(u))f^{-1}(u))f(u)du \approx$$

$$\Big(\int_{a}^{a+h} + \int_{b-h}^{b} + \sum_{i=1}^{m}\int_{d_i-h}^{d_i+h}\Big)\, g((Ef_{nh}^{(1)}(u) - f(u))f^{-1}(u))f(u)du =$$

$$\Big(\int_{a}^{a+h} + \int_{b-h}^{b} + \sum_{i=1}^{m}\int_{d_i-h}^{d_i+h}\Big)\, g(b(u,h)f^{-1}(u))f(u)du \approx$$

$$h\gamma(f,K),$$

where

$$\gamma(f,K) := f(a+)\int_0^1 g\big(f(a+)^{-1}\delta^{(0)}(a)b_0(t)\big)dt +$$

$$f(b-)\int_{-1}^{0} g\big(f(b-)^{-1}\delta^{(0)}(b)b_0(t)\big)dt + \tag{3.23}$$

$$\sum_{i=1}^{m}\Big(f(d_i+)\int_0^1 g\big(f(d_i+)^{-1}\delta^{(0)}(d_i)b_0(t)\big)dt + f(d_i-)\int_{-1}^{0} g\big(f(d_i-)^{-1}\delta^{(0)}(d_i)b_0(t)\big)dt\Big).$$

So in case I this term is *not negligible*. Again by theorem 2.8 we see that in case I we are asymptotically maximizing

$$h\,(\delta^{(0)}(a) - \delta^{(0)}(b))\int_0^1 b_0(t)dt - \tfrac{1}{2}\Big\{h\,\Delta^{(0)}\int_0^1 b_0^2(t)dt + \frac{b-a}{nh}\int_{-1}^{1}K^2(u)du\Big\} + h\gamma(f,K) =$$

$$h\Big\{(\delta^{(0)}(a) - \delta^{(0)}(b))\int_0^1 b_0(t)dt - \tfrac{1}{2}\Delta^{(0)}\int_0^1 b_0^2(t)dt + \gamma(f,K)\Big\} - \frac{b-a}{2nh}\int_{-1}^{1}K^2(u)du. \tag{3.24}$$

However using the corrected method we are asymptotically maximizing

$$h\Big\{-\tfrac{1}{2}\Delta^{(0)}\int_0^1 b_0^2(t)dt + \gamma(f,K)\Big\} - \frac{b-a}{2nh}\int_{-1}^{1}K^2(u)du.$$

Since in case I situations the third term in (3.18) is not negligible neither the uncorrected nor the corrected method asymptotically minimizes the integrated squared error (3.19). So Marron's optimality result does not hold for densities with jumps in the interval [a,b]. Notice that if neither a nor b is a jumping point of f then there is no difference in the asymptotic behavior of the uncorrected and the corrected method because the first term in (3.24) vanishes.

These heuristics lead to the next theorem which gives the rates of convergence of the bandwidths obtained by likelihood cross-validation.

Theorem 3.4. *Suppose that* E *is a bounded interval* [a,b], $-\infty<a<b<\infty$, *and that the density* f *satisfies condition* F *and is bounded away from zero on* E. *Let* $d_1,\ldots,d_m$ *denote the singular points of* f *in* (a,b). *Further assume that the kernel* K *satisfies condition* K *and has a bounded second derivative. For some* $\sigma>0$ *let* I_n *denote the interval* $[h_n', h_n'']$ *with* $h_n'=n^{-1+\sigma}$ *and* $h_n''=n^{-\sigma}$. *Let* H_n *denote the value of* h *which maximizes* $LCV_n(h)$ *over* I_n *and let* H_n^c *denote the value of* h *which maximizes* $LCV_n^c(h)$ *over* I_n. *The next statements hold almost surely.*

(a) Case I: If $H_n=C_n n^{-1/2}$ *then*

$$(3.25)\qquad \lim_{n\to\infty} C_n = \left\{ \frac{\frac{1}{2}(b-a)\int_{-1}^{1} K^2(u)du}{(\delta^{(0)}(b)-\delta^{(0)}(a))\int_0^1 b_0(t)dt + \frac{1}{2}\Delta^{(0)}\int_0^1 b_0^2(t)dt - \gamma(f,K)} \right\}^{1/2},$$

provided

$$(\delta^{(0)}(b)-\delta^{(0)}(a))\int_0^1 b_0(t)dt + \tfrac{1}{2}\Delta^{(0)}\int_0^1 b_0^2(t)dt - \gamma(f,K) > 0.$$

If $H_n^c=C_n^c n^{-1/2}$ *then*

$$(3.26)\qquad \lim_{n\to\infty} C_n^c = \left\{ \frac{\frac{1}{2}(b-a)\int_{-1}^{1} K^2(u)du}{\frac{1}{2}\Delta^{(0)}\int_0^1 b_0^2(t)dt - \gamma(f,K)} \right\}^{1/2},$$

provided

$$\tfrac{1}{2}\Delta^{(0)}\int_0^1 b_0^2(t)dt - \gamma(f,K) > 0.$$

(b) Case II : If $H_n=C_n n^{-1/3}$ *then*

$$(3.27)\qquad \begin{aligned} &\liminf_{n\to\infty}\,(\log n)^{1/2+\varepsilon}\,C_n \geq 1 \\ &\limsup_{n\to\infty}\,\frac{1}{(\log n)^{1+\varepsilon}}\,C_n \leq 1, \end{aligned}$$

provided

$$\Big(\delta^{(1)}(a) + \delta^{(1)}(b) + 2\sum_{i=1}^{m}\delta^{(1)}(d_i)\Big)\int_0^1 b_1(t)dt + \tfrac{1}{2}h^2\int_{-1}^{1} u^2K(u)du \int_a^b f''(u)du < 0.$$

If $H_n^c=C_n^c n^{-1/4}$ *then*

$$(3.28)\qquad \lim_{n\to\infty} C_n^c = \alpha_{II}(f,w)^{1/4}\beta_{II}(K)^{1/4}.$$

(c) Case III : If $H_n=C_n\, n^{-1/3}$ *then*

$$(3.29)\qquad \begin{aligned} &\liminf_{n\to\infty} (\log n)^{1/2+\varepsilon}\, C_n \geq 1 \\ &\limsup_{n\to\infty} \frac{1}{(\log n)^{1+\varepsilon}}\, C_n \leq 1, \end{aligned}$$

provided f'(b) < f'(a).

If $H_n^c = C_n^c n^{-1/5}$ *then*

$$(3.30)\qquad \lim_{n\to\infty} C_n^c = \alpha_{III}(f,w)^{1/5}\beta_{III}(K)^{1/5}. \qquad \square$$

Here the constants $\alpha_{II}(f,w)$, $\alpha_{III}(f,w)$, $\beta_{II}(K)$ and $\beta_{III}(K)$ are the factors in the optimal bandwidths given in (2.25), where the weight function should be taken equal to $f^{-1}I_E$, and $\gamma(f,K)$ is defined in (3.23). So the limits (3.28) and (3.30) are the optimal constants in cases II and III respectively.

Remark 3.5. Since by remark 2.9 the expansions of the mean integrated squared error hold uniformly for h in $(0,h_n'']$ for any sequence of positive h_n'' converging to zero, for the corrected method in cases II and III the theorem above implies (3.11) of theorem 3.3. By an argument based on a result of Marron & Härdle (1986), similar to the one Cline & Hart (1986) use to prove their theorem 6, it can be shown that (3.10) also holds.

Next we consider the type of densities studied by Chow, Geman & Wu (1983) and Devroye & Györfi (1985), i.e. we assume that f has bounded support [c,d] and E=[c,d]. This means that we compute the product (3.9) over all the data points X_i. Also assume f continuous and bounded away from zero on E. This is a *case I situation* with

$$\begin{aligned} \gamma(f,K) \;=\; & f(c+)\int_0^1 g(f(c+)^{-1}f(c+)b_0(t))dt \;+ f(d-)\int_0^1 g(-f(d-)^{-1}f(d-)b_0(t))dt = \\ & (f(c+) + f(d-))\int_0^1 g(b_0(t))dt < 0. \end{aligned}$$

This constant is negative because g is an increasing function on $(-1,\infty)$ with $g(0) = 0$. So g is negative on (-1,0) and since b_0 is negative on (0,1) the function $g(b_0)$ is also negative. We also have by partial integration

$$\begin{aligned} & (\delta^{(0)}(c)-\delta^{(0)}(d))\int_0^1 b_0(t)dt = \\ & -\,(f(c+) + f(d-))\int_0^1 \Big(\int_t^1 K(u)du\Big)dt = \\ & -\,(f(c+) + f(d-))\int_0^1 uK(u)du \;<\; 0. \end{aligned}$$

Using the equality $\Delta^{(0)}=f(c+)+f(d-)$ the next result now follows from theorem 3.4.

Corollary 3.6. *Let* f *satisfy condition* F *and have bounded support* [c,d]. *Let* E=[c,d] *and let* f *be continuous and bounded away from zero on* E. *If* $H_n=C_n n^{-1/2}$ *and* $H_n^c=C_n^c n^{-1/2}$ *then under the conditions of theorem* 3.4 *we have almost surely*

$$(3.31)\qquad \lim_{n\to\infty} C_n = \left\{ \frac{\frac{1}{2}(d-c)\int_{-1}^{1} K^2(u)du}{(f(c+)+f(d-))(\int_0^1 uK(u)du + \frac{1}{2}\int_0^1 b_0^2(t)dt - \int_0^1 g(b_0(t))dt)} \right\}^{1/2}$$

and

$$(3.32)\qquad \lim_{n\to\infty} C_n^c = \left\{ \frac{\frac{1}{2}(d-c)\int_{-1}^{1} K^2(u)du}{(f(c+)+f(d-))(\frac{1}{2}\int_0^1 b_0^2(t)dt - \int_0^1 g(b_0(t))dt)} \right\}^{1/2}.$$

□

Remark 3.7. The asymptotically optimal constant for the weight function $f^{-1}I_E$ in the case I situation of this corollary is given by (2.25). It equals

$$c_{opt} = \left\{ \frac{\frac{1}{2}(d-c)\int_{-1}^{1} K^2(u)du}{(f(c+)+f(d-))\int_0^1 b_0^2(t)dt} \right\}^{1/2}.$$

The corresponding optimal bandwidth h_n^{opt} is equal to $c_{opt}n^{-1/2}$. Note that the quotients of the limits in (3.31) and (3.32) and c_{opt} depend only on the kernel function K and not on the density f. This means that we can obtain almost sure convergence to the asymptotical optimal constant c_{opt} by multiplying the computed bandwiths H_n and H_n^c by a known constant. However, even using the optimal bandwidths, unavoidably we have a large error since we are dealing with a case I situation. It would be better to use the symmetrization device described by Schuster (1985) combined with cross-validation to determine a good bandwidth. Cline & Hart (1986) discuss this approach for least squares cross-validation.

The two previous theorems show that in the cases II and III, i.e. if the density f has no jumps in the interval [a,b], the bandwidths H_n^c are asymptotically almost surely equivalent to the deterministic asymptotically optimal bandwidths with respect to the weighted mean integrated squared error $MISE_n(h)$, where

$$MISE_n(h) = E\ ISE_n(h) = E\int_a^b (f_{nh}(x) - f(x))f^{-1}(x)dx.$$

Let H_n^* denote the positive value of h which minimizes the integrated squared error $ISE_n(h)$. Since H_n^* is the random bandwidth which we would like to approximate we derive the asymptotic distribution of $H_n^c - H_n^*$. The next theorem establishes the asymptotic normality of $H_n^c - H_n^*$ in the cases II and III. The proof is given in section 3.4.

Theorem 3.8. *Suppose that the conditions of theorem* 3.4 *are satisfied. With* $L(u):=K(u)+uK'(u)$ *we define the constants* σ^2, σ_{II}^2 *and* σ_{III}^2 *by*

$$\sigma^2 := 4(b-a)\int_{-1}^{1} L^2(u)du,$$

$$\sigma_{II}^2 := \Delta^{(1)}\int_0^1 \Big(\int_{-\infty}^{t} (t-u)L(u)du\Big)^2 dt,$$

$$\sigma_{III}^2 := \frac{1}{4}\Big(\int_{-1}^{1} u^2K(u)du\Big)^2\Big(\int_a^b f''(x)^2 f^{-1}(x)dx - (f'(b) - f'(a))^2\Big),$$

and the constants α_0, α_1 *and* α_2 *by*

$$\alpha_0 := \frac{b-a}{2}\int_{-1}^{1} K^2(u)du,$$

$$\alpha_1 := \frac{1}{2}\Big(\int_{-1}^{1} u^2K(u)du\Big)^2 \int_a^b f''(x)^2 f^{-1}(x)dx$$

$$\alpha_2 := \frac{3}{2}\Delta^{(1)}\int_0^1 \Big(\int_{-\infty}^{t} (t-u)K(u)dt\Big)^2 dt.$$

Then we have in case II

$$n^{3/8}(H_n^c - H_n^*) \xrightarrow{\mathcal{D}} N(0,\tfrac{1}{16}(2\alpha_0^{-5/4}\alpha_2^{-3/4}\sigma^2 + \alpha_0^{-1/4}\alpha_2^{-7/4}\sigma_{III}^2))$$

and in case III

$$n^{3/10}(H_n^c - H_n^*) \xrightarrow{\mathcal{D}} N(0,\tfrac{1}{25}(2\alpha_0^{-7/5}\alpha_1^{-3/5}\sigma^2 + \alpha_0^{-2/5}\alpha_1^{-8/5}\sigma_{II}^2)).$$ □

The second statement of this theorem is similar to theorem 2.1 in Hall & Marron (1987a), the asymptotic normality result for the bandwidths computed by least squares cross-validation, the only difference is in the asymptotic variance. It shows that for smooth densities we also have the slow convergence demonstrated by (3.6) for least squares cross-validation. Though formally it doesn't apply here since we use a different weight function, this result is coherent with theorem 2.1 in Hall & Marron (1987b), which states that we can not expect a faster rate of convergence. This theorem assumes the densities to be twice differentiable, essentially our case III situation. It is a nice surprise that the first statement of our theorem shows a faster rate of convergence. In that case we have

$$\frac{H_n^c - H_n^*}{H_n^*} = O_p\left(\frac{n^{-3/8}}{n^{-1/4}}\right) = O_p(n^{-1/8}),$$

which is of smaller order than the bound $O_p(n^{-1/10})$ which holds in the smooth case. Proceeding as in Hall & Marron (1987a) we would also obtain

$$\frac{ISE_n(H_n^c) - ISE_n(H_n^*)}{ISE_n(H_n^*)} = O_p(n^{-1/4}), \text{ in case II}$$

and

$$\frac{ISE_n(H_n^c) - ISE_n(H_n^*)}{ISE_n(H_n^*)} = O_p(n^{-1/5}), \text{ in case III,}$$

which shows that the minimal integrated squared error is also better approximated in case II. However we should keep in mind that if f has kinks in [a,b] this integrated squared error is of a larger order than it is for smooth densities.

3.2. The derivative of $\log(LCV_n(h))$.

The proofs of theorem 3.4 and theorem 3.8 in the previous section are based on expansions of the derivative of the function $\log(LCV_n(h))$. Before we can derive these expansions we give two successive decompositions of this derivative in sections 3.2.1 and 3.2.2 . In section 3.3 we then obtain the expansions which prove theorem 3.4. The proof of theorem 3.8, stating the asymptotic normality of $H_n^c - H_n^*$, is given next in section 3.4.

3.2.1. A decomposition.

We first consider the derivative of the kernel estimator f_{nh} with respect to the bandwidth h. For K differentiable we have

$$\frac{d}{dh} f_{nh}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{d}{dh}\frac{1}{h} K((x-X_i)/h) =$$

(3.33) $$\frac{1}{n}\sum_{i=1}^{n} \left(-\frac{1}{h^2} K((x-X_i)/h) - \frac{(x-X_i)}{h^3} K'((x-X_i)/h)\right) =$$

$$-\frac{1}{nh^2}\sum_{i=1}^{n} L((x-X_i)/h)$$

with

(3.34) $$L(x) := K(x) + xK'(x), -\infty < x < \infty.$$

This function plays an important role in the sequel. For kernels K, satisfying condition K and having a bounded derivative, L has the following properties, (L.1) L has support [-1,1],

(L.2) L is bounded,

(L.3) L is symmetric,

(L.4) $\int_{-1}^{1} L(u)du = 0.$

The first three properties are immediate and property (L.4) follows by partial integration.
The next figure shows the graph of the function L for the kernel K, given by

$$K(x) = \frac{35}{32}(1-x^2)^3 I_{[-1,1]}(x).$$

Notice that K' continuous implies L continuous, and that L has a bounded derivative if K has a bounded second derivative. This last property is required if we want to apply theorem 2.16 to the derivative (3.33).

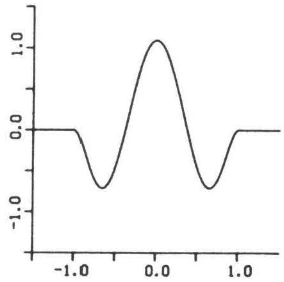

Figure 3.1. The function L.

Now consider $\log(LCV_n(h))$. Since

$$\frac{1}{n}\log(LCV_n(h)) = \frac{1}{n}\sum_{i:X_i\in E} \log(f_{nh}^{(i)}(X_i)).$$

and

$$f_{nh}^{(i)}(X_i) = \frac{n}{n-1} f_{nh}(X_i) - \frac{1}{(n-1)h} K(0)$$

we have

$$\frac{1}{n}\log(LCV_n(h)) =$$

$$\frac{1}{n}\sum_{i:X_i\in E} \log(\frac{n}{n-1} f_{nh}(X_i) - \frac{1}{(n-1)h} K(0))$$

$$\frac{1}{n}\sum_{i:X_i\in E} \log(f_{nh}(X_i) - \frac{1}{nh} K(0)) + \log(\frac{n}{n-1}).$$

Next use (3.33) to obtain

$$\frac{1}{n}\frac{d}{dh}\log(LCV_n(h)) =$$

$$-\frac{1}{nh}\sum_{i:X_i\in E}\frac{\frac{1}{nh}\sum_{j=1}^{n}L((X_i-X_j)/h)-\frac{1}{nh}K(0)}{f_{nh}(X_i)-\frac{1}{nh}K(0)} = \tag{3.35}$$

$$-\frac{1}{nh}\sum_{i:X_i\in E}\frac{\frac{1}{nh}\sum_{j=1}^{n}L((X_i-X_j)/h)-\frac{1}{nh}K(0)}{\frac{1}{nh}\sum_{j=1}^{n}K((X_i-X_j)/h)-\frac{1}{nh}K(0)}.$$

The following decomposition of this derivative is the key tool in our analysis of the behavior of likelihood cross-validation.

Proposition 3.9. *If the kernel* K *is differentiable and satisfies condition* K, *and if we define the function* L *by* (3.34), *then*

$$\frac{1}{n}\frac{d}{dh}\log(LCV_n(h)) =$$

$$U_n(h) + V_n(h) + W_n(h) + Y_n(h) + R_n(h).$$

where

$$U_n(h) := \frac{1}{n^2}\sum_{i\neq j}U_{ij}(h),\ V_n(h) := \frac{1}{n^3}\sum_{i\neq j}V_{ij}(h) \textit{ and } W_n(h) := \frac{1}{n^3}\sum_{i\neq j\neq k}W_{ijk}(h),$$

with

$$U_{ij}(h) \ := \ -\frac{2}{h^2}L((X_i-X_j)/h)\ f(X_i)^{-1}\ I_E(X_i),$$

$$V_{ij}(h) \ := \ \frac{1}{h^3}K((X_i-X_j)/h)\ L((X_i-X_j)/h)\ f(X_i)^{-2}\ I_E(X_i),$$

$$W_{ijk}(h) := \ \frac{1}{h^3}K((X_i-X_j)/h)\ L((X_i-X_k)/h)\ f(X_i)^{-2}\ I_E(X_i),$$

and where

$$Y_n(h) := -\frac{1}{nh}\sum_{i=1}^{n}\{\frac{1}{nh}\sum_{j=1}^{n}L((X_i-X_j)/h)-\frac{1}{nh}K(0)\}$$

$$\{\ \frac{1}{nh}\sum_{j=1}^{n}K((X_i-X_j)/h)-f(X_i)-\frac{1}{nh}K(0)\}^2 \tag{3.36}$$

$$\{\ \frac{1}{nh}\sum_{j=1}^{n}K((X_i-X_j)/h)-\frac{1}{nh}K(0)\}^{-1}\ f(X_i)^{-2}\ I_E(X_i),$$

and

$$R_n(h) := \frac{2K(0)^2}{(nh)^3} \sum_{i=1}^{n} f(X_i)^{-2}\, I_E(X_i).$$ □

Proof. Write the denominator in (3.35) as

$$f_{nh}(X_i) - \frac{1}{nh}K(0) = f(X_i) + f_{nh}(X_i) - f(X_i) - \frac{1}{nh}K(0) = f(X_i) + \Delta_{ni}(h),$$

thus defining $\Delta_{ni}(h)$. Next we introduce the function g by

$$g(x,s) := \frac{1}{x+s} - \frac{1}{x} + s\frac{1}{x^2} = \frac{s^2}{x^2(x+s)}.$$

This gives

$$\frac{1}{f_{nh}(X_i) - \frac{1}{nh}K(0)} = \frac{1}{f(X_i)} - \frac{\Delta_{ni}(h)}{f(X_i)^2} + g(f(X_i),\Delta_{ni}(h)),$$

and therefore by (3.35) we have

$$\frac{1}{n}\frac{d}{dh}\log(LCV_n(h)) =$$

$$-\frac{1}{nh}\sum_{i:X_i\in E}\left\{\frac{1}{nh}\sum_{j=1}^{n} L((X_i-X_j)/h) - \frac{1}{nh}K(0)\right\}\left\{\frac{1}{f(X_i)} - \frac{\Delta_{ni}(h)}{f(X_i)^2} + g(f(X_i),\Delta_{ni}(h))\right\} =$$

$$-\frac{1}{nh}\sum_{i=1}^{n} I_E(X_i)\left\{\frac{1}{nh}\sum_{j=1}^{n} L((X_i-X_j)/h) - \frac{1}{nh}K(0)\right\}$$

$$\left\{\frac{2}{f(X_i)} - \frac{1}{f(X_i)^2}\Big(\frac{1}{nh}\sum_{j=1}^{n} K((X_i-X_j)/h) - \frac{1}{nh}K(0)\Big) + g(f(X_i),\Delta_{ni}(h))\right\}.$$

This can be rewritten as

$$\sum_{i=1}^{6} Z_{ni}(h) + Y_n(h), \tag{3.37}$$

with

$$Z_{n1}(h) := -\frac{2}{(nh)^2}\sum_{i=1}^{n}\sum_{j=1}^{n} L((X_i-X_j)/h)\, f(X_i)^{-1} I_E(X_i),$$

$$Z_{n2}(h) := \frac{1}{(nh)^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} L((X_i-X_j)/h)K((X_i-X_k)/h)\, f(X_i)^{-2} I_E(X_i),$$

$$Z_{n3}(h) := -\frac{1}{(nh)^3}K(0)\sum_{i=1}^{n}\sum_{j=1}^{n} L((X_i-X_j)/h)\, f(X_i)^{-2} I_E(X_i),$$

$$Z_{n4}(h) := \frac{2}{(nh)^2} K(0) \sum_{i=1}^{n} f(X_i)^{-1} I_E(X_i),$$

$$Z_{n5}(h) := - \frac{1}{(nh)^3} K(0) \sum_{i=1}^{n} \sum_{j=1}^{n} K((X_i - X_j)/h)\ f(X_i)^{-2} I_E(X_i),$$

$$Z_{n6}(h) := \frac{K(0)^2}{(nh)^3} \sum_{i=1}^{n} f(X_i)^{-2}\ I_E(X_i).$$

and

$$Y_n(h) := - \frac{1}{nh} \sum_{i=1}^{n} \left\{ \frac{1}{nh} \sum_{j=1}^{n} L((X_i - X_j)/h) - \frac{1}{nh} K(0) \right\} g(f(X_i), \Delta_{ni}(h))\ I_E(X_i).$$

Now note that

$$Z_{n1}(h) = -Z_{n4}(h\ ) + U_n(h),$$

$$Z_{n2}(h) = -Z_{n3}(h) - Z_{n5}(h) + Z_{n6}(h) + V_n(h) + W_n(h),$$

$$Z_{n6}(h) = \tfrac{1}{2} R_n(h),$$

which shows that (3.37) equals

$$U_n(h) + V_n(h) + W_n(h) + Y_n(h) + R_n(h).$$

This completes the proof of the proposition. □

3.2.2. The relation to U-statistic theory: a second decomposition.

The statistics $U_n(h)$, $V_n(h)$ and $W_n(h)$ in the preceding section are U-statistics. If φ is a symmetric real valued function defined on the m dimensional Euclidean space then a U-statistic of degree m with kernel φ is defined as

$$\binom{n}{m}^{-1} \sum_{(i_1,\ldots,i_m) \in C_{n,m}} \varphi(X_{i_1},\ldots,X_{i_m}),$$

where $C_{n,m}$ is the set of all ordered m-tuples $(i_1,\ldots,i_m)$ of different indices from the set $\{1,2,\ldots,n\}$. Note that, with

$$\varphi_h^U(x,y) := - \frac{2}{h^2} L((x-y)/h)(f(x)^{-1} I_E(x) + f(y)^{-1} I_E(y)),$$

$$\varphi_h^V(x,y) := \frac{1}{h^3} K((x-y)/h) L(x-y)/h)(f(x)^{-2} I_E(x) + f(y)^{-2} I_E(y)),$$

we have

$$U_n(h) = \frac{1}{n^2} \sum_{i<j} \varphi_h^U(X_i, X_j),$$

$$V_n(h) = \frac{1}{n^3} \sum_{i<j} \varphi_h^V(X_i,X_j),$$

so up to normalizing factors $U_n(h)$ and $V_n(h)$ are indeed U-statistics. Also notice

$$\varphi_h^U(X_i,X_j) = U_{ij}(h)+U_{ji}(h)$$

and

$$\varphi_h^V(X_i,X_j) = V_{ij}(h)+V_{ji}(h).$$

Similarly we can write $W_n(h)$ as

$$W_n(h) = \frac{1}{n^3} \sum_{i<j<k} \varphi_h^W(X_i,X_j,X_k),$$

where

$$\varphi_h^W(X_i,X_j,X_k) = W_{ijk}(h)+W_{ikj}(h)+W_{kji}(h)+W_{jik}(h)+W_{jki}(h)+W_{kij}(h),$$

i.e. the sum over all permutations of the indices. So $W_n(h)$ is a U-statistic of order 3.

The kernel functions φ_h^U, φ_h^V and φ_h^W depend on the bandwidth. To derive the asymptotic distribution theory for likelihood cross-validation we need the asymptotic distribution of $U_{h_n}(h_n)$ for a sequence of bandwidths (h_n) (The terms $V_{h_n}(h_n)$ and $W_{h_n}(h_n)$ are negligible). In that case we are dealing with a statistic of the form

$$U_n(h) = \frac{1}{n^2} \sum_{i<j} \varphi_{h_n}^U(X_i,X_j),$$

a U-statistic with a kernel depending on the sample size. The asymptotic distribution theory for this type of statistics is studied by Hall (1984), Jammalamadaka & Jansson (1986), De Jong (1987, 1988), Nolan & Pollard (1987, 1988).

Although we can not use the theory for U-statistics with fixed kernels, we can employ Hoeffding's projection technique to derive a decomposition of a U-statistic (Hoeffding (1948), Serfling (1980)). This results in the following decompositions

$$U_n(h) = \frac{n-1}{n} EU_{ij}(h) +$$

$$\frac{n-1}{n}\frac{1}{n}\sum_{i=1}^{n} (E(U_{ij}(h)|X_i) - EU_{ij}(h)) + \frac{n-1}{n}\frac{1}{n}\sum_{j=1}^{n} (E(U_{ij}(h)|X_j) - EU_{ij}(h)) +$$

$$\frac{1}{n^2}\sum_{i\neq j} (U_{ij}(h) - E(U_{ij}(h)|X_i) - E(U_{ij}(h)|X_j) + EU_{ij}(h)),$$

which we rewrite as

$$U_n(h) = \frac{n-1}{n} EU_{ij}(h) +$$

$$\frac{n-1}{n}\frac{1}{n}\sum_{i=1}^{n} (E(U_{ij}(h)|X_i) - EU_{ij}(h)) + \frac{n-1}{n}\frac{1}{n}\sum_{j=1}^{n} (E(U_{ij}(h)|X_j) - EU_{ij}(h)) +$$

$$\hat{U}_n(h),$$

where

$$\hat{U}_n(h) := \frac{1}{n^2}\sum_{i\neq j} \hat{U}_{ij}(h),$$

and

$$\hat{U}_{ij}(h) := U_{ij}(h) - E(U_{ij}(h)|X_i) - E(U_{ij}(h)|X_j) + EU_{ij}(h).$$

Similarly we decompose $V_n(h)$ as

$$V_n(h) = \frac{n-1}{n^2} EV_{ij}(h) +$$

$$\frac{n-1}{n^2}\frac{1}{n}\sum_{i=1}^{n} (E(V_{ij}(h)|X_i) - EV_{ij}(h)) + \frac{n-1}{n^2}\frac{1}{n}\sum_{j=1}^{n} (E(V_{ij}(h)|X_j) - EV_{ij}(h)) +$$

$$\hat{V}_n(h),$$

where

$$\hat{V}_n(h) := \frac{1}{n^3}\sum_{i\neq j} \hat{V}_{ij}(h),$$

and

$$\hat{V}_{ij}(h) := V_{ij}(h) - E(V_{ij}(h)|X_i) - E(V_{ij}(h)|X_j) + EV_{ij}(h).$$

Finally we also decompose the statistic $W_n(h)$. We get

$$W_n(h) = \frac{(n-1)(n-2)}{n^2} EW_{ijk}(h) +$$

$$\frac{(n-1)(n-2)}{n^2}\frac{1}{n}\sum_{i=1}^{n} (E(W_{ijk}(h)|X_i) - EW_{ijk}(h)) +$$

$$\frac{(n-1)(n-2)}{n^2}\frac{1}{n}\sum_{j=1}^{n} (E(W_{ijk}(h)|X_j) - EW_{ijk}(h)) +$$

$$\frac{(n-1)(n-2)}{n^2}\frac{1}{n}\sum_{k=1}^{n} (E(W_{ijk}(h)|X_k) - EW_{ijk}(h)) +$$

$$\hat{W}_n(h),$$

where

$$\hat{W}_n(h) := \frac{1}{n^3}\sum_{i\neq j\neq k} \hat{W}_{ijk}(h)$$

and

$$\hat{W}_{ijk}(h) := W_{ijk}(h) - E(W_{ijk}(h)|X_i) - E(W_{ijk}(h)|X_j) - E(W_{ijk}(h)|X_k) + 2EW_{ijk}(h).$$

An important property of these decompositions is that the conditional expectations of the terms of $\hat{U}_n(h)$ given the X's vanish, i.e. for k=1, ..., n

$$E(\hat{U}_{ij}(h)|X_k) = E(\hat{U}_{ij}(h)|X_k) = 0.$$

This implies that $\hat{U}_n(h)$ and the linear terms in the decomposition of $U_n(h)$ are uncorrelated, an inherent property of the Hoeffding decomposition. The other two decompositions have a similar property.

We obtain a further decomposition of the derivative of $n^{-1}\log(LCV_n(h))$ by plugging in the previous ones of the statistics $U_n(h)$, $V_n(h)$ and $W_n(h)$ in the decomposition derived in the previous section. Then we compute the various conditional expectations appearing above. These conditional expectations can be expressed in terms of functions b^G with G equal to one of K, L, or KL, the product of the functions K and L. These functions b^G are generalizations of the bias function b in section 2.2. We also introduce generalizations of the functions b_0, b_1 and b_2 which appeared in the expansion of the bias.

Definition 3.10. *The functions* b^G, b^G_0, b^G_1 *and* b^G_2 *are defined by*

$$b^G(x,h) := E\,G_h(x-X_1) - f(x)\int_{-1}^{1} G(u)du = \frac{1}{h}\int_{-\infty}^{\infty} G((x-u)/h)f(u)du - f(x)\int_{-1}^{1} G(u)du$$

and

$$b^G_m(t) := \begin{cases} \int_{-\infty}^{t} (t-u)^m G(u)du & \textit{if } t<0 \\ -\int_{t}^{\infty} (t-u)^m G(u)du & \textit{if } t\geq 0 \end{cases},$$

for m=0,1,2.

The proof of the next lemma is a direct generalization of the proof of theorem 2.3 and is therefore omitted.

Lemma 3.11. *Assume that* G *is a bounded symmetric measurable function with support equal to* [-1,1] *and that the density* f *satisfies condition* F. *Let* (h'_n) *be a vanishing sequence of positive real numbers.*

(a) Then

$$b^G(x,h) = \tfrac{1}{2}h^2 f''(x)\int_{-1}^{1} u^2 G(u)du + r_3(x,h)$$

where the remainder r_3 *satisfies*

$$\lim_{n\to\infty} \sup_{0<h\leq h_n'} \sup_{x\in D_h\cap[-M,M]} h^{-2}r_3(x,h) = 0$$

for every positive M.

(b) For x_0 *a fixed point we have*

$$\begin{aligned} b^G(x_0+th,h) = \quad & b_0^G(t)\delta^{(0)}(x_0) + hb_1^G(t)\delta^{(1)}(x_0) + \tfrac{1}{2}h^2b_2^G(t)\delta^{(2)}(x_0) + \\ & \tfrac{1}{2}h^2 \int_{-1}^{1} u^2G(u)du\{f''(x_0-)I_{(-\infty,0)}(t) + f''(x_0+)I_{(0,\infty)}(t)\} + \\ & r_4(t,h), \end{aligned}$$

where the remainder r_4 *satisfies*

$$\lim_{n\to\infty} \sup_{0<h\leq h_n'} \sup_{-M\leq t\leq M, t\neq 0} h^{-2}r_4(t,h) = 0$$

for every positive M.

We now state the main proposition of this section.

Proposition 3.12. *If the kernel* K *is differentiable and satisfies condition* K, *and if we define the function* L *by* (3.34), *then we have the following decomposition,*

$$\begin{aligned} & \frac{1}{n}\frac{d}{dh}\log(LCV_n(h)) = \\ & \frac{n-1}{n}EU_{ij}(h) + \frac{n-1}{n^2}EV_{ij}(h) + \frac{(n-1)(n-2)}{n^2}EW_{ijk}(h) + \\ & \frac{n-1}{n}\frac{1}{n}\sum_{i=1}^{n}(u_1(X_i,h) - Eu_1(X_i,h)) + \\ & \frac{n-1}{n}\frac{1}{n}\sum_{i=1}^{n}(u_2(X_i,h) - Eu_2(X_i,h)) + \\ & \frac{n-1}{n}\frac{1}{n^2}\sum_{i=1}^{n}(v_1(X_i,h) - Ev_1(X_i,h)) + \\ & \frac{n-1}{n}\frac{1}{n^2}\sum_{i=1}^{n}(v_2(X_i,h) - Ev_2(X_i,h)) + \\ & \frac{(n-1)(n-2)}{n^2}\frac{1}{n}\sum_{i=1}^{n}(w_1(X_i,h) - Ew_1(X_i,h)) + \\ & \frac{(n-1)(n-2)}{n^2}\frac{1}{n}\sum_{i=1}^{n}(w_2(X_i,h) - Ew_2(X_i,h)) + \end{aligned}$$

$$\frac{(n-1)(n-2)}{n^2}\frac{1}{n}\sum_{i=1}^{n}(w_3(X_i,h) - Ew_3(X_i,h)) +$$

$$\hat{U}_n(h) + \hat{V}_n(h) + \hat{W}_n(h) +$$

$$Y_n(h) + R_n(h),$$

where the functions $u_1, u_2, v_1, v_2, w_1, w_2, w_3$ *are defined by*

$$\begin{aligned}
u_1(x,h) &:= -\frac{1}{h}\, b^L(x,h)\, f(x)^{-1} I_E(x), \\
u_2(x,h) &:= -\frac{1}{h^2}\int_E L((u-x)/h)du\,, \\
v_1(x,h) &:= \frac{1}{h^2}\,(f(x)\int_{-1}^{1} KL(u)du \;+ b^{KL}(x,h))f(x)^{-2} I_E(x), \\
v_2(x,h) &:= \frac{1}{h^3}\int_E KL((u-x)/h)f(u)^{-1}du\,, \\
w_1(x,h) &:= \frac{1}{h}\, b^K(x,h) b^L(x,h) f(x)^{-2} I_E(x)\,, \\
w_2(x,h) &:= \frac{1}{h^2}\int_E K((u-x)/h)f(u)^{-1} b^L(u,h)du\,, \\
w_3(x,h) &:= \frac{1}{h^2}\int_E L((u-x)/h)f(u)^{-1} b^K(u,h)du\,.
\end{aligned} \tag{3.38}$$

To prove this result we only have to compute the conditional expectations in the decompositions of $U_n(h)$, $V_n(h)$ and $W_n(h)$. These conditional expectations are given by the next lemma.

Lemma 3.13. *The conditional expectations of* $U_{ij}(h)$, $V_{ij}(h)$ *and* $W_{ijk}(h)$ *are given by*

(a) $$E(U_{ij}(h) \mid X_i) = -\frac{2}{h}\, b^L(X_i,h)\, f(X_i)^{-1} I_E(X_i) = 2u_1(X_i,h),$$

$$E(U_{ij}(h) \mid X_j) = -\frac{2}{h^2}\int_E L((u-X_j)/h)du = 2u_2(X_j,h),$$

(b) $$E(V_{ij}(h) \mid X_i) = \frac{1}{h^2}(f(X_i)\int_{-1}^{1} KL(u)du \;+ b^{KL}(X_i,h))f(X_i)^{-2} I_E(X_i) = v_1(X_i,h),$$

$$E(V_{ij}(h) \mid X_j) = \frac{1}{h^3}\int_E KL((u-X_j)/h)f(u)^{-1}du = v_2(X_j,h),$$

(c) $$E(W_{ijk}(h) \mid X_i) = -\frac{1}{2}E\,(U_{ik}(h) \mid X_i) + \frac{1}{h} b^K(X_i,h) b^L(X_i,h) f(X_i)^{-2} I_E(X_i) =$$

$$- u_1(X_i,h) + w_1(X_i,h),$$

$$E(W_{ijk}(h) \mid X_j) = \frac{1}{h^2}\int_E K((u-X_j)/h)f(u)^{-1}b^L(u,h)du = w_2(X_j,h),$$

$$E(W_{ijk}(h) \mid X_k) = -\tfrac{1}{2}E\,(U_{ik}(h) \mid X_k) + \frac{1}{h^2}\int_E L((u-X_k)/h)f(u)^{-1}b^K(u,h)du =$$

$$- u_2(X_k,h) + w_3(X_k,h).$$

Proof. We only derive the expressions for $E(W_{ijk}(h)|X_i)$ and $E(W_{ijk}(h)|X_k)$. The other expressions are obtained similarly. We get the conditional expextation of $W_{ijk}(h)$ given X_i and X_k by integrating out X_j,

$$E(W_{ijk}(h)|X_i,X_k) =$$

$$\int_{-\infty}^{\infty} \frac{1}{h^3}K((X_i-v)/h)L((X_i-X_k)/h)f(X_i)^{-2}I_E(X_i)f(v)dv =$$

$$\frac{1}{h^2}L((X_i-X_k)/h)f(X_i)^{-2}I_E(X_i)\{f(X_i) + \frac{1}{h}\int_{-\infty}^{\infty} K((X_i-v)/h)f(v)dv - f(X_i)\} =$$

$$-\tfrac{1}{2}U_{ik}(h) + \frac{1}{h^2}L((X_i-X_k)/h)f(X_i)^{-2}I_E(X_i)b^K(X_i,h).$$

Next we obtain $E(W_{ijk}(h)|X_i) = E(E(W_{ijk}(h)|X_i,X_k)|X_i)$ by integrating out X. This gives

$$E(W_{ijk}(h)|X_i) =$$

$$-\tfrac{1}{2}E(U_{ik}(h)|X_i) + \int_{-\infty}^{\infty} \frac{1}{h^2}L((X_i-w)/h)f(X_i)^{-2}I_E(X_i)b^K(X_i,h)f(w)dw =$$

$$-\tfrac{1}{2}E(U_{ik}(h)|X_i) + \frac{1}{h}b^K(X_i,h)b^L(X_i,h)f(X_i)^{-2}I_E(X_i).$$

Similarly we compute $E(W_{ijk}(h)|X_k)$ by integrating out X_i. This gives

$$E(W_{ijk}(h)|X_k) =$$

$$-\tfrac{1}{2}E(U_{ik}(h)|X_k) + \int_{-\infty}^{\infty} \frac{1}{h^2}L((u-X_k)/h)f(u)^{-2}I_E(u)b^K(u,h)f(u)du =$$

$$-\tfrac{1}{2}E(U_{ik}(h)|X_k) + \frac{1}{h^2}\int_E L((u-X_k)/h)f(u)^{-1}b^K(u,h)f(u)du\ ,$$

which is the correct expression. □

Proposition 3.12 gives us a basis for deriving both the rates of convergence to zero as well as the asymptotic distributions of the bandwidths computed by likelihood cross-validation. In the following section we obtain results on the rates of convergence to zero. Distributional properties are studied in section 3.4.

3.3. Rates of convergence: proof of theorem 3.4.

Recall that the random bandwidth H_n computed by the uncorrected likelihood cross-validation method is equal to the value of h which maximizes the random function LCV_n, defined in (3.9), over the interval $I_n=[h_n',h_n'']$, where $h_n'=n^{-1+\sigma}$ and $h_n''=n^{-\sigma}$ for some $\sigma>0$. The random bandwidth H_n^c computed by the corrected likelihood cross-validation method is equal to the value of h which maximizes the random function LCV_n^c over the interval I_n. The function LCV_n^c is obtained from LCV_n by

$$(3.39) \qquad LCV_n^c(h) = LCV_n(h) \exp\Big(-\sum_{i=1}^{n} \frac{1}{h} \int_E K((u-X_i)/h)du\Big).$$

In this section we prove theorem 3.4 concerning the rates of convergence to zero of the random bandwidths H_n and H_n^c. We consider the root and the sign of the derivative of the random functions $\log(LCV_n(.))$ and $\log(LCV_n^c(.))$.

Throughout this section we assume that the conditions of theorem 3.4 are satisfied, i.e. we assume that E=[a,b] and that the density f satisfies condition F and is bounded away from zero on E. Further we assume that K satisfies condition K and has a bounded second derivative. By $d_1,..., d_m$ we denote the singular points of f in the open interval (a,b). We treat the points a and b separately.

The decomposition given in proposition 3.12 gives the next expansion of the derivative of $\log(LCV_n(.))$. The proof of this expansion is given at the end of the section.

Proposition 3.14. *If we write*

$$\frac{1}{n}\frac{d}{dh}\log(LCV_n(h)) =$$

$$-\frac{1}{n}\sum_{j=1}^{n}\frac{1}{h^2}\int_E L((u-X_j)/h)du \ +$$

$$\frac{b-a}{2nh^2}\int_{-1}^{1} K^2(u)du \ +$$

$$\frac{1}{h}\int_E b^L(u,h)b^K(u,h)f(u)^{-1}du \ +$$

$$Y_n(h) +$$

$$R_{n1}(h),$$

where $Y_n(h)$ *is defined by* (3.36), *then the remainder term* R_{n1} *satisfies*

$$\sup_{h\in I_n} \alpha_n(h)\, R_{n1}(h) = o(1), \textit{ almost surely,}$$

with $\alpha_n(h)$ *equal to*

$$\alpha_n(h) = \begin{cases} \left(\frac{1}{nh^2} + 1\right)^{-1} & \textit{in case I} \\ \left(\frac{1}{nh^2} + h^2\right)^{-1} & \textit{in case II} \\ \left(\frac{1}{nh^2} + h^3\right)^{-1} & \textit{in case III.} \end{cases}$$ □

Now we also automatically obtain an expansion for the derivative of $\log(LCV_n^c(.))$ since by (3.39) and (3.33) we have

$$\frac{1}{n}\frac{d}{dh}\log(LCV_n^c(h)) =$$

$$\frac{1}{n}\frac{d}{dh}\log(LCV(h)) - \frac{1}{n}\frac{d}{dh}\Big(\sum_{i=1}^{n}\frac{1}{h}\int_E K((u-X_i)/h)du\Big) = \tag{3.40}$$

$$\frac{1}{n}\frac{d}{dh}\log(LCV(h)) + \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^2}\int_E L((u-X_i)/h)du \ .$$

It follows that the correction factor removes the first term in the expansion of $\frac{1}{n}\frac{d}{dh}\log(LCV_n(h))$ given by proposition 3.14. Using the expansions of the bias functions b^K and b^L, provided by lemma 3.11, next we expand the deterministic third term.

Lemma 3.15. *We have*

$$\frac{1}{h}\int_E b^L(u,h)b^K(u,h)f(u)^{-1}du =$$

$$\begin{cases} -\frac{1}{2}\Delta^{(0)}\int_0^1 b_0^K(t)^2dt + r_1(h) & \textit{in case I} \\ -\frac{3}{2}h^2\Delta^{(1)}\int_0^1 b_1^K(t)^2dt + h^2r_2(h) & \textit{in case II} \\ -\frac{1}{2}h^3\Big(\int_{-1}^1 u^2K(u)du\Big)^2\int_E f\ ''(u)^2f^{-1}(u)du + h^3r_3(h) & \textit{in case III} \end{cases}, \tag{3.41}$$

where $r_1(h)$, $r_2(h)$ *and* $r_3(h)$ *converge to zero uniformly for* $h\in I_n$. □

Proof. Let D_h denote the set of points on the real line which are at least at a distance h from the singular points of f. For n large enough and $h \in I_n$ write

$$\frac{1}{h}\int_E b^L(u,h)b^K(u,h)f^{-1}(u)du =$$

$$\frac{1}{h}\Big(\int_{E \cap D_h} + \int_a^{a+h} + \int_{b-h}^{b} + \sum_{i=1}^{m} \int_{d_i-h}^{d_i+h}\Big) b^L(u,h)b^K(u,h)f^{-1}(u)du.$$

First consider case III. Then the interval E = [a,b] contains no singular points. By lemma 3.11 we have for n large enough

$$\frac{1}{h}\int_E b^L(u,h)b^K(u,h)f^{-1}(u)du =$$

$$\frac{1}{h}\int_E \Big(\tfrac{1}{2}h^2 f''(u)\Big(\int_{-1}^{1} v^2L(v)dv\Big)\tfrac{1}{2}h^2 f''(u)\Big(\int_{-1}^{1} v^2K(v)dv\Big)f^{-1}(u)du\Big) + h^3 r_1(h),$$

where the remainder term $r_1(h)$ vanishes uniformly for $h \in I_n$ for n tending to infinity. Since by partial integration we have

$$\int_{-1}^{1} v^2L(v)dv = -2\int_{-1}^{1} v^2K(v)dv$$

this proves (3.41) for case III. In the cases I and II this term is asymptotically negligible and the term

$$\frac{1}{h}\sum_{i=1}^{m} \int_{d_i-h}^{d_i+h} b^L(u,h)b^K(u,h)f^{-1}(u)du \tag{3.42}$$

dominates. Let d_i be a singular point of f in the open interval (a,b). The term corresponding to d_i in the sum (3.42) is equal to

$$\frac{1}{h}\int_{d_i-h}^{d_i+h} b^L(u,h)b^K(u,h)f^{-1}(u)du. \tag{3.43}$$

By lemma 3.11 this term is equal to

$$\int_{-1}^{1} \delta^{(0)}(d_i)b_0^L(t)\delta^{(0)}(d_i)b_0^K(t)f^{-1}(d_i+th)dt + o(1) =$$

$$\delta^{(0)}(d_i)^2 \int_{-1}^{1} b_0^L(t)b_0^K(t)f^{-1}(d_i+th)dt + o(1) =$$

$$\delta^{(0)}(d_i)^2\Big(f^{-1}(d_i+)\int_0^1 b_0^L(t)b_0^K(t)dt + f^{-1}(d_i-)\int_{-1}^{0} b_0^L(t)b_0^K(t)dt\Big) + o(1) =$$

$$\delta^{(0)}(d_i)^2(f^{-1}(d_i+) + f^{-1}(d_i-))\int_0^1 b_0^L(t)b_0^K(t)dt \; + o(1),$$

since b_0^K and b_0^L are odd functions. If $\delta^{(0)}(d_i)$ is equal to zero then the expansion of (3.43) becomes

$$\int_{-1}^{1} \delta^{(1)}(d_i)b_1^L(t)\delta^{(1)}(d_i)b_1^K(t)f^{-1}(d_i+th)dt + o(h^2) =$$

$$h^2\delta^{(1)}(d_i)^2(f^{-1}(d_i+) + f^{-1}(d_i-))\int_0^1 b_1^L(t)b_1^K(t)dt \; + o(h^2).$$

Similar expansions hold for the points a and b. The uniformity of these expansions is readily verified so it remains to show the equalities

$$\int_0^1 b_0^L(t)b_0^K(t)dt = -\tfrac{1}{2}\int_0^1 b_0^K(t)^2dt \tag{3.44}$$

and

$$\int_0^1 b_1^L(t)b_1^K(t)dt = -\tfrac{3}{2}\int_0^1 b_1^K(t)^2dt. \tag{3.45}$$

The proof of these equalities is postponed to the end of section 3.5. □

We also need a bound on the term $Y_n(h)$. For cases II and III it is given by lemma 3.16. Lemma 3.17 provides information on $Y_n(h)$ for case I. Both lemmas are proved in section 3.5.

Lemma 3.16. *For some constant* c > 0

$$\limsup_{n\to\infty} \sup_{h\in I_n} \alpha_n(h)\, |Y_n(h)| < c, \textit{ almost surely,}$$

where

$$\alpha_n(h) = \begin{cases} \left(\frac{1}{h}\left(\frac{\log n}{nh}\right)^{3/2} + h^3\right)^{-1} & \textit{in case II} \\ \left(\frac{1}{h}\left(\frac{\log n}{nh}\right)^{3/2} + h^5\right)^{-1} & \textit{in case III} \end{cases}.$$ □

First consider the corrected method in the smooth case III. Proposition 3.14, (3.40), lemma 3.15 and lemma 3.16 imply

$$\frac{1}{n}\frac{d}{dh}\log(LCV_n^c(h)) = \alpha_0\frac{1}{nh^2} - \alpha_1 h^3 + R_{n2}(h)\left(\frac{1}{nh^2} + h^3\right), \tag{3.46}$$

where for some sequence of almost surely vanishing random variables S_n we have for all $h \in I_n$

$$|R_{n2}(h)| \le S_n, \tag{3.47}$$

and the constants α_0 and α_1 are defined in theorem 3.8. Substituting $h = cn^{-1/5}$ in relation (3.46) and multiplying by $n^{3/5}$ we get

(3.48) $$\alpha_0\frac{1}{c^2} - \alpha_1 c^3 + R_{n2}(cn^{-1/5})\left(\frac{1}{c^2} + c^3\right).$$

Let c_0 be equal to $(\alpha_0/\alpha_1)^{1/5}$ then c_0 is equal to the optimal constant for the mean integrated squared error in case III for the weight function $w = f^{-1}I_E$ (see (2.25)). Next we rewrite (3.48) as

(3.49) $$\left(\alpha_0\frac{1}{c^2} - \alpha_1 c^3\right)\left(1 + R_{n2}(cn^{-1/5})\,\frac{\frac{1}{c^2} + c^3}{\alpha_0\frac{1}{c^2} - \alpha_1 c^3}\right).$$

We see by (3.47) that for any $0 < \varepsilon < c_0$ and for all n larger than a random integer $N(\varepsilon)$ the expression (3.49) is positive for all c in $(0,c_0-\varepsilon)\cap n^{1/5}I_n$ and negative for all c in $(c_0+\varepsilon,\infty)\cap n^{1/5}I_n$. So if we write $H_n^c = C_n^c\, n^{-1/5}$ then for all ε in $(0,c_0)$

$$c_0-\varepsilon \le C_n^c \le c_0+\varepsilon, \text{ for all } n \ge N(\varepsilon).$$

Thus we have shown

$$\lim_{n\to\infty} C_n^c = c_0, \text{ almost surely,}$$

which proves (3.30), i.e. the almost sure convergence to the optimal constant for case III.

The proof of statement (3.28) of theorem 3.4 for the case II is exactly the same except that the second term in (3.46) is of order h^2 instead of h^3. In case II we have

(3.50) $$\frac{1}{n}\frac{d}{dh}\log(LCV_n^c(h)) = \alpha_0\frac{1}{nh^2} - \alpha_2 h^2 + R_{n3}(h)\left(\frac{1}{nh^2} + h^2\right),$$

where α_2 is given in theorem 3.8, and $R_{n3}(h)$ satisfies a condition similar to (3.47).

In case I the situation is different since then the term $Y_n(h)$ in the expansion of proposition 3.14 is no longer negligible. The next lemma deals with this term. The proof is given in section 3.5.

Lemma 3.17. *Let* $d_1,\ldots, d_m$ *denote the jumping points of* f *in* (a,b). *Then*

$$Y_n(h) = \gamma(f,K) + R_{n4}(h),$$

where $\gamma(f,K)$ *is defined in* (3.23) *and* $R_{n4}(h)$ *satisfies*

$$\sup_{h\in I_n}\left(\frac{1}{nh^2} + 1\right)^{-1} R_{n4}(h) = o(1), \textit{ almost surely.}$$ □

We now get the following expansion of the derivative of $\log(LCV_n^c(.))$,

(3.51) $$\frac{1}{n}\frac{d}{dh}\log(LCV_n^c(h)) = \alpha_0\frac{1}{nh^2} - \alpha_3 + \gamma(f,K) + R_{n5}(h)\left(\frac{1}{nh^2} + 1\right),$$

where for some almost surely vanishing sequence of random variables S'_n we have for all $h \in I_n$

$$|R_{n5}(h)| \leq S'_n.$$

Here α_0 is the same as above and α_3 is given by

$$\alpha_3 := \tfrac{1}{2}\Delta^{(0)} \int_0^1 b_0^K(t)^2 dt.$$

Contrary to the previous cases here the leading term of (3.51) does not always have a root in $(0,\infty)$. If $\alpha_3 - \gamma(f,K) < 0$ then by the same argument as above the derivative is positive for all $h \in I_n$ for n larger than some random integer N. This means that we get find large values of H_n^c. On the other hand if $\alpha_3 - \gamma(f,K) > 0$, defining C_n^c by $H_n^c = C_n^c n^{-1/5}$, we get (3.26) of theorem 3.4.

Having dealt with the part of theorem 3.4 about the corrected method we proceed with proving the results concerning the uncorrected method. The next lemma gives expansions of the expectation of the correction term

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^2} \int_E L((u-X_i)/h)du \; .$$

Since the proof is a straigthforward application of lemma 3.11 it is omitted.

Lemma 3.18. *We have*

$$E\frac{1}{h^2}\int_a^b L((u-X_i)/h)du =$$

$$\begin{cases} (\delta^{(0)}(a) - \delta^{(0)}(b)) \int_0^1 b_0^L(t)dt + r_1(h) & \textit{in case I} \\ h(\delta^{(1)}(a) + \delta^{(1)}(b) + 2\sum_{i=1}^{m} \delta^{(1)}(d_i)) \int_0^1 b_1^L(t)dt + & \\ \qquad \tfrac{1}{2}h(f'(b) - f'(a)) \int_{-1}^1 u^2 L(u)du + hr_2(h) & \textit{in case II} \\ \tfrac{1}{2}h(f'(b) - f'(a)) \int_{-1}^1 u^2 L(u)du + hr_3(h) & \textit{in case III,} \end{cases}$$

where the functions $r_1(h)$, $r_2(h)$ *and* $r_3(h)$ *converge to zero uniformly for* $h \in I_n$. □

By the statement concerning u_2 in lemma 3.19 below we also have almost surely

$$\frac{1}{n\log n} \sup_{h \in I_n} (nh)^{1/2} \Big|\sum_{i=1}^{n} \Big(\frac{1}{h^2}\int_a^b L((u-X_i)/h)du - E\frac{1}{h^2}\int_a^b L((u-X_i)/h)du\Big)\Big| = o(1),$$

so the correction term is equal to

$$E\frac{1}{h^2}\int_a^b L((u-X_i)/h)du + R_{n6}(h)\,\frac{\log n}{(nh)^{1/2}}, \tag{3.52}$$

where $R_{n6}(h)$ for all $h\in I_n$ satisfies

$$|R_{n6}(h)| \leq S''_n$$

for some almost surely vanishing sequence of random variables S''_n. It is not hard to show that the asymptotic standard deviation of the correction term is of the order $(nh)^{-1/2}$, so apart from the factor $\log n$ the bound in (3.52) is sharp.

First consider the uncorrected method in case III. Substracting expansions (3.46) and (3.52) we get

$$\frac{1}{n}\frac{d}{dh}\log(LCV_n(h)) =$$

$$\frac{1}{n}\frac{d}{dh}\log(LCV_n^c(h)) - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^2}\int_E L((u-X_i)/h)du =$$

$$\alpha_0\frac{1}{nh^2} - \alpha_1 h^3 + R_{n2}(h)\left(\frac{1}{nh^2} + h^3\right) - E\frac{1}{h^2}\int_a^b L((u-X_i)/h)du - R_{n6}(h)\,\frac{\log n}{(nh)^{1/2}},$$

which by lemma 3.18 equals

$$\alpha_0\frac{1}{nh^2} - \alpha_1 h^3 + \alpha_4 h + hr_3(h) + R_{n2}(h)\left(\frac{1}{nh^2} + h^3\right) - R_{n6}(h)\,\frac{\log n}{(nh)^{1/2}} =$$

$$\alpha_0\frac{1}{nh^2} + \alpha_4 h + hr_3(h) + R_{n2}(h)\left(\frac{1}{nh^2} + h^3\right) - R_{n6}(h)\,\frac{\log n}{(nh)^{1/2}}, \tag{3.53}$$

where α_4 is given by

$$\alpha_4 = -\frac{1}{2}(f'(b) - f'(a))\int_{-1}^{1} u^2 L(u)du = (f'(b) - f'(a))\int_{-1}^{1} u^2 K(u)du$$

and $r_4(h)$ converges to zero uniformly for $h\in I_n$. Next notice

$$\frac{1}{(nh)^{1/2}} = \begin{cases} (nh^3)^{1/2}\,\dfrac{1}{nh^2} \\ \dfrac{1}{(nh^3)^{1/2}}\,h \end{cases},$$

so if nh^3 converges to zero or infinity fast enough the term $(nh)^{-1/2}\log n$ is negligible compared to the leading terms in (3.53), uniformly for $h\in I_n$. However if nh^3 remains bounded away from zero and infinity then this term is not negligible. Writing $h = cn^{-1/3}$ and multiplying (3.53) by $n^{1/3}$ we rewrite (3.53) as

$$\alpha_0\frac{1}{c^2} + \alpha_4 c - R_{n6}(cn^{-1/3}) \log n + cr_3(cn^{-1/3}) + R_{n2}(cn^{-1/3}) \left(\frac{1}{c^2} + c^3 n^{-2/3}\right).$$

Now assume that α_4 is negative then $\alpha_0 c^{-2} + \alpha_4 c$ has a root in $(0,\infty)$. By a similar argument as we used for the corrected method we then find that for n larger than a random integer N we have for all $\varepsilon>0$

$$\frac{1}{(\log n)^{1/2+\varepsilon}} \le C_n \le (\log n)^{1+\varepsilon}$$

which gives

$$\liminf_{n\to\infty} (\log n)^{1/2+\varepsilon}\, C_n \ge 1, \qquad \text{almost surely,}$$

and

$$\limsup_{n\to\infty} \frac{1}{(\log n)^{1+\varepsilon}} C_n \le 1, \qquad \text{almost surely,}$$

thus proving (3.29).

For case II the proof of (3.27) is exactly the same except that the constant α_4 is different since lemma 3.18 gives a different constant. Here α_4 is equal to

$$- h\, (\delta^{(1)}(a) + \delta^{(1)}(b) + 2\sum_{i=1}^{m} \delta^{(1)}(d_i)) \int_0^1 b_1^L(t)dt - \frac{1}{2}h(f'(b) - f'(a)) \int_{-1}^{1} u^2 L(u)du =$$

$$2h\, (\delta^{(1)}(a) + \delta^{(1)}(b) + 2\sum_{i=1}^{m} \delta^{(1)}(d_i)) \int_0^1 b_1^K(t)dt + h(f'(b) - f'(a)) \int_{-1}^{1} u^2 K(u)du \;,$$

where we use (see the end of section 3.5 for the proof)

(3.54) $$\int_0^1 b_1^L(t)dt = - 2\int_0^1 b_1^K(t)dt.$$

Statement (3.25) for case I can be proved in the same way as we proved (3.26) because in this case the variation of the correction term is negligible since uniformly for all $h \in I_n$ we have $\log n\ (nh)^{-1/2} = o(1)$. Here we need the relation

(3.55) $$\int_0^1 b_0^L(t)dt = - \int_0^1 b_0^K(t)dt,$$

the proof of which is also postponed to section 3.5.

To complete the proof of theorem 3.4 we now prove proposition 3.14.

Proof of proposition 3.14. The proof is based on the decomposition given by proposition 3.12. Combining some of the terms of the decomposition we write

$$\frac{n-1}{n} EU_{ij}(h) + \frac{n-1}{n^2} EV_{ij}(h) + \frac{(n-1)(n-2)}{n^2} EW_{ijk}(h) + \frac{n-1}{n}\frac{1}{n}\sum_{i=1}^{n} (u_2(X_i,h) - Eu_2(X_i,h)) =$$

$$\frac{n-1}{n}(E(EU_{ij}(h)|X_j) - Eu_2(X_i,h)) + \frac{n-1}{n^2}E(EV_{ij}(h)|X_i) + \frac{(n-1)(n-2)}{n^2}E(EW_{ijk}(h)|X_j) +$$

$$- \frac{n-1}{n}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^2}\int_a^b L((u-X_i)/h)du.$$

By lemma 3.13 this is equal to

$$\frac{n-1}{n}(2Eu_2(X_1,h) - Eu_2(X_1,h)) + \frac{n-1}{n^2}Ev_1(X_1,h) +$$

$$\frac{(n-1)(n-2)}{n^2}(-Eu_1(X_1,h) + Ew_1(X_1,h)) - \frac{n-1}{n}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^2}\int_a^b L((u-X_i)/h)du =$$

$$\frac{n-1}{n}Eu_2(X_1,h) - \frac{(n-1)(n-2)}{n^2}Eu_1(X_1,h) +$$

$$\frac{n-1}{n^2}Ev_1(X_1,h) + \frac{(n-1)(n-2)}{n^2}Ew_1(X_1,h) +$$

$$- \frac{n-1}{n}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^2}\int_a^b L((u-X_i)/h)du.$$

Notice that by lemma 3.11 we have

$$\frac{1}{n}Ev_1(X_1,h) = \frac{1}{nh^2}\Big(\int_{-1}^{1}KL(u)du\int_{-\infty}^{\infty}f^{-1}(x)I_E(x)f(x)dx + hr_4(h)\Big) =$$

(3.56) $$\frac{b-a}{2nh^2}\int_{-1}^{1}K^2(u)du + \frac{h}{nh^2}r_4(h),$$

where $r_4(h)$ converges to zero uniformly for $h \in I_n$. Here we use the equality

$$\int_{-1}^{1}KL(u)du = \int_{-1}^{1}(K^2(u) + uK'(u)K(u))du = \frac{1}{2}\int_{-1}^{1}K^2(u)du.$$

Furthermore we have

$$Ew_1(X_1,h) = \frac{1}{h}\int_E b^L(u,h)b^K(u,h)f(u)^{-1}du$$

and since $Eu_1(X_1,h) = Eu_2(X_1,h) = EU_{ij}(h)$ we also get

$$\frac{n-1}{n}Eu_2(X_1,h) - \frac{(n-1)(n-2)}{n^2}Eu_1(X_1,h) =$$

(3.57) $$2\frac{n-1}{n^2}Eu_2(X_1,h).$$

It follows from lemma 3.18 that (3.57) is asymptotically negligible and from lemma 3.15, lemma 3.18, (3.52) and (3.56) that the factors $(n-1)/n$ and $(n-1)(n-2)/n^2$ can be replaced by one.
Since the term $R_n(h)$ is readily dealt with it remains to show that the linear terms corresponding to the functions u_1, v_1, v_2, w_1, w_2 and w_3 and the quadratic terms $\hat{U}_n(h)$, $\hat{V}_n(h)$ and $\hat{W}_n(h)$ are asymptotically negligible. This is achieved by the next two lemmas which are proved in section 3.5.

Lemma 3.19. *Let φ be one of the functions u_1, u_2, v_1, v_2, w_1, w_2 and w_3 and let $(\alpha_n(.))$ be sequence of positive functions on $(0,\infty)$. The statement*

$$\text{(3.58)} \qquad \frac{1}{n\log n} \sup_{h\in I_n} \alpha_n(h) \left|\sum_{i=1}^{n} (\varphi(X_i,h) - E\varphi(X_i,h))\right| = o(1), \textit{ almost surely,}$$

is valid for $\varphi = u_1$, $\varphi = w_2$ and $\varphi = w_3$ if

$$\alpha_n(h) = \begin{cases} n^{1/2}h^{1/2} & \textit{in case I} \\ n^{1/2}h^{-1/2} & \textit{in case II} \\ n^{1/2}h^{-1} & \textit{in case III} \end{cases} .$$

It is valid for $\varphi = u_2$ if we take $\alpha_n(h)$ equal to $n^{1/2}h^{1/2}$, for $\varphi = v_1$ and $\varphi = v_2$ if we take $\alpha_n(h)$ equal to $n^{1/2}h^2$ and for $\varphi = w_1$ if

$$\alpha_n(h) = \begin{cases} n^{1/2}h^{1/2} & \textit{in case I} \\ n^{1/2}h^{-3/2} & \textit{in case II} \\ n^{1/2}h^{-3} & \textit{in case III} \end{cases} .$$

□

Lemma 3.20. *For any $\alpha>0$ we have*

$$\text{(3.59)} \qquad \begin{aligned} \sup_{h\in I_n} n^{-\alpha}(nh^{3/2})\,|\hat{U}_n(h)| &= o(1), \textit{ almost surely,} \\ \sup_{h\in I_n} n^{-\alpha}(n^2h^{5/2})\,|\hat{V}_n(h)| &= o(1), \textit{ almost surely,} \\ \sup_{h\in I_n} n^{-\alpha}(n^{3/2}h^2)\,|\hat{W}_n(h)| &= o(1), \textit{ almost surely.} \end{aligned}$$

□

The proof of proposition 3.14 is completed by checking that the bounds provided by these two lemmas are small enough. For instance for the term corresponding to the function u_1 we use

$$n^{-1/2}h^{-1/2} = \begin{cases} \dfrac{1}{nh^2}(nh^2)^{1/2}\, h^{1/2} \\ 1 \; \dfrac{1}{(nh^2)^{1/2}}\, h^{1/2} \end{cases} ,$$

which shows by distinguishing the cases $nh^2 \geq 1$ and $nh^2<1$ that we have

$$n^{-1/2}h^{-1/2} < \left(\frac{1}{nh^2} + 1\right),$$

for n large enough. So in case I the linear term corresponding to the function u_1 is indeed small enough. The other linear terms can be treated similarly. For the quadratic term $\hat{U}_n(h)$ we write

$$|\hat{U}_n(h)| = \frac{1}{nh^2} h^{1/2} nh^{3/2} |\hat{U}_n(h)| \leq \frac{1}{nh^2} (n^{-\sigma/2} nh^{3/2} |\hat{U}_n(h)|),$$

so by lemma 3.20 that this term is also asymptotically negligible. By similar bounds the other two quadratic terms are also negligible and the proof of proposition 3.14 is completed. □

3.4. Asymptotic distribution theory: proof of theorem 3.8.

Before we study the asymptotic distribution of the bandwidths obtained by likelihood cross-validation we first derive some properties of H_n^*, the value of h in the interval I_n which minimizes the integrated squared error $ISE_n(h)$, given by

$$(3.60) \qquad ISE_n(h) = \int_E (f_{nh}(x) - f(x))^2 f^{-1}(x)dx.$$

In the proof of theorem 2.11 we already noticed that

$$ISE_n(h) =$$

$$\frac{1}{n^2h^2} \sum_{i \neq j} \int_E K(\frac{x-X_i}{h}) K(\frac{x-X_j}{h}) f^{-1}(x)dx +$$

$$- \frac{2}{nh} \sum_{i=1}^{n} \int_E K(\frac{x-X_i}{h})dx +$$

$$\frac{1}{n^2h^2} \sum_{i=1}^{n} \int_E K^2(\frac{x-X_i}{h}) f^{-1}(x)dx +$$

$$\int_E f(x)dx.$$

Since with L as in (3.34),

$$\frac{d}{dh} \frac{1}{h} K(\frac{x}{h}) = - \frac{1}{h^2} L(\frac{x}{h})$$

we have

$$\frac{d}{dh} ISE_n(h) =$$

$$- \frac{2}{n^2h^3} \sum_{i \neq j} \int_E K(\frac{x-X_i}{h}) L(\frac{x-X_j}{h}) f^{-1}(x)dx +$$

$$\frac{2}{nh^2} \sum_{i=1}^{n} \int_E L(\frac{x-X_i}{h})dx +$$

$$-\frac{2}{n^2h^3}\sum_{i=1}^{n}\int_E K\left(\frac{x-X_i}{h}\right)L\left(\frac{x-X_i}{h}\right)f^{-1}(x)dx.$$

Therefore

$$\frac{d}{dh}\mathrm{ISE}_n(h) =$$

$$-\frac{2}{n^2}\sum_{j\neq k} E(W_{ijk}(h)|X_j,X_k) +$$

$$-\frac{1}{n}\sum_{j=1}^{n} E(U_{ij}(h)|X_j) +$$

$$-\frac{2}{n^2}\sum_{j=1}^{n} E(V_{ij}(h)|X_j),$$

with $U_{ij}(h)$, $V_{ij}(h)$ and $W_{ijk}(h)$ as in proposition 3.9. Just as in section 3.2.2. we use the Hoeffding projection technique and lemma 3.13 to obtain the decomposition

$$\frac{d}{dh}\mathrm{ISE}_n(h) =$$

$$-\left(\frac{n-1}{n}+\frac{1}{2n^2}\right)EU_{ij}(h) - \frac{2}{n}EV_{ij}(h) - 2\frac{n-1}{n}EW_{ijk}(h) +$$

$$-2\frac{n-1}{n^2}\sum_{i=1}^{n}(w_2(X_i,h) - Ew_2(X_i,h)) +$$

$$-2\frac{n-1}{n^2}\sum_{i=1}^{n}(w_3(X_i,h) - Ew_3(X_i,h)) + \tag{3.61}$$

$$-2\frac{1}{n^2}\sum_{i=1}^{n}(v_2(X_i,h) - Ev_2(X_i,h)) +$$

$$-\frac{1}{n^2}\sum_{i=1}^{n}(u_2(X_i,h) - Eu_2(X_i,h)) +$$

$$-2\,\tilde{W}_n(h),$$

where the functions w_2, w_3 and v_2 are defined by (3.38) and

$$\tilde{W}_n(h) := \frac{1}{n^2}\sum_{j\neq k}(E(W_{ijk}(h)|X_j,X_k) - E(W_{ijk}(h)|X_j) - E(W_{ijk}(h)|X_k) + EW_{ijk}(h)).$$

By the same arguments we used to prove the bound on $\hat{U}_n(h)$ in lemma 3.20 we have for any $\alpha>0$

$$\sup_{h\in I_n} n^{-\alpha}(nh^{3/2})\,\tilde{W}_n(h) = o(1), \text{ almost surely.} \tag{3.62}$$

Next notice the similarity of (3.61) and the decomposition given by proposition 3.12, and also notice that the linear term corresponding to the function u_2, which dominated the behavior in case of the uncorrected likelihood cross-validation method, is of lower order in (3.61). Proceeding in the same way as in section 3.3 we obtain the next result which states that in the three cases I, II and III the random bandwidths H_n^* are asymptotically almost surely equivalent to the deterministic optimal bandwidths for the mean integrated suared error, which is no surprise since we are directly minimizing $ISE_n(h)$. The proof of the theorem is omitted.

Proposition 3.21. *Suppose that* E *is a bounded interval* $[a,b]$, $-\infty<a<b<\infty$, *and that the density* f *satisfies condition* F *and is bounded away from zero on* E. *Let* $d_1, ..., d_m$ *denote the singular points of* f *in* (a,b). *Further assume that the kernel* K *satisfies condition* K *and has a bounded first derivative. For some* $\sigma>0$ *let* I_n *denote the interval* $[h_n', h_n'']$, *with* $h_n'=n^{-1+\sigma}$ *and* $h_n''=n^{-\sigma}$. *Let* H_n^* *denote the value of* h *which minimizes* $ISE_n(h)$, *given by* (3.60), *over* I_n. *The next statements hold almost surely.*

a) Case I: If $H_n^*=C_n^*n^{-1/2}$ *then*

$$\lim_{n\to\infty} C_n^* = \alpha_I(f,w)^{1/2}\beta_I(K)^{1/2}.$$

(a) Case II: If $H_n^*=C_n^*n^{-1/4}$ *then*

$$\lim_{n\to\infty} C_n^* = \alpha_{II}(f,w)^{1/4}\beta_{II}(K)^{1/4}.$$

(b) Case III: If $H_n^*=C_n^*n^{-1/5}$ *then*

$$\lim_{n\to\infty} C_n^* = \alpha_{III}(f,w)^{1/5}\beta_{III}(K)^{1/5}.$$

Here the factors α_I, α_{II}, α_{III}, β_I, β_{II} and β_{III} are the factors in the optimal bandwidths for the mean integrated squared error given in (2.25).

Another important property of H_n^* which we need is the fact that since the derivative of ISE_n is equal to zero in the point H_n^* we have by (3.61)

$$\begin{aligned}
&\{(\tfrac{n-1}{n}+\tfrac{1}{2n^2})EU_{ij}(h)+\tfrac{2}{n}EV_{ij}(h)+2\tfrac{n-1}{n}EW_{ijk}(h)+\\
&2\tfrac{n-1}{n^2}\sum_{i=1}^{n}(w_2(X_i,h)-Ew_2(X_i,h))+\\
(3.63)\qquad &2\tfrac{n-1}{n^2}\sum_{i=1}^{n}(w_3(X_i,h)-Ew_3(X_i,h))+\\
&2\tfrac{1}{n^2}\sum_{i=1}^{n}(v_2(X_i,h)-Ev_2(X_i,h))+
\end{aligned}$$

$$\frac{1}{n^2}\sum_{i=1}^{n} (u_2(X_i,h) - Eu_2(X_i,h)) \}\Big|_{h=H_n^*} =$$

$$- 2\, \tilde{W}_n(H_n^*).$$

In order to derive the asymptotic distribution of H_n^c - H_n^* we assume that the conditions of theorem 3.4 are satisfied. Define the two random functions $D_n^{(1)}(.)$ and $D_n^{(2)}(.)$ by

$$D_n^{(1)}(h) := \frac{d}{dh}\log(LCV_n^c(h)),\ h>0$$

and

$$D_n^{(2)}(h) := \frac{d^2}{dh^2}\log(LCV_n^c(h)),\ h>0.$$

By the mean value theorem we have

$$D_n^{(1)}(H_n^c) - D_n^{(1)}(H_n^*) = - D_n^{(1)}(H_n^*) = D_n^{(2)}(\tilde{H}_n)\,(H_n^c - H_n^*),$$

for some random variable $\tilde{H}_n$ between H_n^c and H_n^*. Thus we have the equality

$$(3.64) \qquad (H_n^c - H_n^*) = - \frac{D_n^{(1)}(H_n^*)}{D_n^{(2)}(\tilde{H}_n)}.$$

First consider the denominator. In cases II and III it follows from theorem 3.4 and proposition 3.21 that H_n^c and H_n^* are asymptotically almost surely equivalent to the deterministic optimal bandwidths given by (2.25). The same is clearly true for $\tilde{H}_n$. By examining the derivative of the decomposition given by proposition 3.9, using the same techniques which led to (3.50) in case II, and to (3.46) in case III, it can be shown that we have

$$\frac{1}{n}D_n^{(2)}(h) = \begin{cases} -2\alpha_0\frac{1}{nh^3} - 2\alpha_2 h + R_{n7}(h)\left(\frac{1}{nh^3} + h^2\right) & \text{in case II} \\ -2\alpha_0\frac{1}{nh^3} - 3\alpha_1 h + R_{n8}(h)\left(\frac{1}{nh^3} + h^2\right) & \text{in case III} \end{cases},$$

with α_0, α_1 and α_2 as in theorem 3.8 and where we have almost surely

$$\sup_{h\in I_n} |R_{n7}(h)| = o(1) \text{ and } \sup_{h\in I_n} |R_{n8}(h)| = o(1).$$

Since the optimal constants in the cases II and III are $(\alpha_0/\alpha_2)^{1/4}$ and $(\alpha_0/\alpha_1)^{1/5}$, respectively, we have almost surely

$$(3.65) \qquad \frac{1}{n}D_n^{(2)}(\tilde{H}_n) \sim \begin{cases} -4\alpha_0^{1/4}\alpha_2^{3/4}n^{-1/4} & \text{in case II} \\ -5\alpha_0^{2/5}\alpha_1^{3/5}n^{-2/5} & \text{in case III} \end{cases}.$$

Next we examine the nominator of (3.64). Recall that by proposition 3.12 and (3.74) we have

$$\frac{1}{n}D_n^{(1)}(h) = \frac{d}{dh}\log(LCV_n(h)) - \frac{1}{n}\sum_{i=1}^{n} u_2(X_i,h).$$

Use proposition 3.12, (3.40) and (3.63) to show that

$$\frac{1}{n}D_n^{(1)}(H_n^*) = T_n^{(1)}(H_n^*) + T_n^{(2)}(H_n^*), \tag{3.66}$$

where

$$T_n^{(1)}(h) := \hat{U}_n(h) + \frac{n-1}{n}\frac{1}{n}\sum_{i=1}^{n}(u_1(X_i,h) - Eu_1(X_i,h))$$

and

$$\begin{aligned} T_n^{(2)}(h) := {} & \frac{n-1}{n}\frac{1}{n^2}\sum_{i=1}^{n}(v_1(X_i,h) - Ev_1(X_i,h)) + \\ & \frac{(n-1)(n-2)}{n^2}\frac{1}{n}\sum_{i=1}^{n}(w_1(X_i,h) - Ew_1(X_i,h)) + \\ & \hat{V}_n(h) + \hat{W}_n(h) - \tilde{W}_n(h) + Y_n(h) + R_n(h) + \\ & -\frac{3}{2n^2}\sum_{i=1}^{n}(u_2(X_i,h) - Eu_2(X_i,h)) + \\ & -\frac{2n-1}{4n^2}EU_{ij}(h) - \frac{1}{n^2}EV_{ij}(h) - 2\frac{n-1}{n^2}EW_{ijk}(h). \end{aligned}$$

It turns out that $T_n^{(2)}(H_n^*)$ is negligible compared to $T_n^{(1)}(H_n^*)$. By the next two lemma's we derive the asymptotic normality of

$$T_n^{(1)}(H_n^*) = \hat{U}_n(H_n^*) + \left\{\frac{n-1}{n}\frac{1}{n}\sum_{i=1}^{n}(u_1(X_i,h) - Eu_1(X_i,h))\right\}\Big|_{h=H_n^*}. \tag{3.67}$$

Lemma 3.22.
(a) In case II we have

$$n^{5/8}\,(T_n^{(1)}(H_n^*) - T_n^{(1)}(c_{opt}n^{-1/4})) \xrightarrow{\mathcal{P}} 0,$$

where c_{opt} *denotes the asymptotically optimal constant for case II given by* (2.25).

(b) In case III we have

$$n^{7/10}\,(T_n^{(1)}(H_n^*) - T_n^{(1)}(c_{opt}n^{-1/5})) \xrightarrow{\mathcal{P}} 0,$$

where c_{opt} *denotes the asymptotically optimal constant for case III given by* (2.25). □

Lemma 3.23.
(a) In case II we have for any $c>0$

$$n^{5/8}T_n^{(1)}(cn^{-1/4}) \xrightarrow{\mathcal{D}} N(0,c^{-3}(2\sigma^2+c^4\sigma_{II}^2)),$$

with

$$\sigma^2 := 4(b\text{-}a)\int_{-1}^{1} L^2(v)dv$$

and

$$\sigma_{II}^2 := \Delta^{(1)}\int_0^1 b_1^L(t)^2dt.$$

(b) In case III we have for any c>0

$$n^{7/10}T_n^{(1)}(cn^{-1/5}) \xrightarrow{\mathcal{D}} N(0,c^{-3}(2\sigma^2+c^5\sigma_{III}^2)),$$

with

$$\sigma_{III}^2 := \tfrac{1}{4}\Big(\int_{-1}^{1} u^2K(u)du\Big)^2\Big(\int_a^b f''(x)^2f^{-1}(x)dx - (f'(b) - f'(a))^2\Big). \qquad □$$

A sketch of the proof of lemma 3.22 and the proof of lemma 3.23 are postponed to section 3.5.

Lemma 3.22 shows that to obtain asymptotic normality of (3.67) H_n^* can be replaced by the deterministic asymptotically optimal bandwidth $(\alpha_0/\alpha_2)^{1/4}n^{-1/4}$ in case II, and by $(\alpha_0/\alpha_1)^{1/5}n^{-1/5}$ in case III. Provided the other terms in (3.66) are negligible by (3.64) we see from (3.64), (3.65) and the two previous lemmas that in case II

$$n^{3/8}(H_n^c - H_n^*) = -\, n^{3/8}\frac{\frac{1}{n}D_n^{(1)}(H_n^*)}{\frac{1}{n}D_n^{(2)}(\tilde{H}_n)}\ ,$$

is asymptotically normally distributed with zero mean and variance

$$\frac{(\alpha_0/\alpha_2)^{-3/4}(2\sigma^2 + (\alpha_0/\alpha_2)\sigma_{II}^2)}{16\alpha_0^{1/2}\alpha_2^{3/2}} = \tfrac{1}{16}(2\alpha_0^{-5/4}\alpha_2^{-3/4}\sigma^2 + \alpha_0^{-1/4}\alpha_2^{-7/4}\sigma_{III}^2).$$

In case III $n^{3/10}(H_n^c - H_n^*)$ is asymptotically normally distributed with zero mean and variance

$$\frac{(\alpha_0/\alpha_1)^{-3/5}(2\sigma^2 + (\alpha_0/\alpha_1)\sigma_{III}^2)}{25\alpha_0^{4/5}\alpha_1^{6/5}} = \tfrac{1}{25}(2\alpha_0^{-7/5}\alpha_1^{-3/5}\sigma^2 + \alpha_0^{-2/5}\alpha_1^{-8/5}\sigma_{II}^2).$$

To complete the proof of theorem 3.8 it remains to show that $T_n^{(2)}(H_n^*)$ is indeed negligible.
In order to deal with the term $\hat{W}_n(H_n^*) - \tilde{W}_n(H_n^*)$ we write

$$\hat{W}_n(h) - \tilde{W}_n(h) =$$

$$\frac{1}{n^3}\sum_{i\neq j\neq k}(W_{ijk}(h) - E(W_{ijk}(h)|X_i) - E(W_{ijk}(h)|X_j) - E(W_{ijk}(h)|X_k) + 2EW_{ijk}(h)) -$$

$$\frac{1}{n^2}\sum_{j\neq k}(E(W_{ijk}(h)|X_j,X_k) - E(W_{ijk}(h)|X_j) - E(W_{ijk}(h)|X_k) + EW_{ijk}(h)) =$$

$$\frac{1}{n^3}\sum_{i\neq j\neq k}(W_{ijk}(h) - E(W_{ijk}(h)|X_i) - E(W_{ijk}(h)|X_j,X_k) + EW_{ijk}(h)) -$$

$$\frac{2}{n^3}\sum_{j\neq k}(E(W_{ijk}(h)|X_j,X_k) - E(W_{ijk}(h)|X_j) - E(W_{ijk}(h)|X_k) + EW_{ijk}(h)) =$$

$$\hat{\hat{W}}_n(h) - \frac{2}{n}\tilde{W}_n(h),$$

with

(3.68) $$\hat{\hat{W}}_n(h) := \frac{1}{n^3}\sum_{i\neq j\neq k}(W_{ijk}(h) - E(W_{ijk}(h)|X_i) - E(W_{ijk}(h)|X_j,X_k) + EW_{ijk}(h)).$$

Similarly to the proof of the statement concerning $\hat{W}_n(h)$ in lemma 3.20 it can be shown that we have for any $\alpha>0$

(3.69) $$\sup_{h\in I_n} n^{-\alpha}(n^{3/2}h^2)\,|\hat{\hat{W}}_n(h)| = o(1), \text{ almost surely.}$$

In case III we now have

$$|n^{7/10}(\hat{W}_n(H_n^*) - \tilde{W}_n(H_n^*))| \leq$$

$$n^{7/10}n^{-3/2}(H_n^*)^{-2}\,n^{3/2}(H_n^*)^2|\hat{\hat{W}}_n(H_n^*)| +$$

$$|n^{7/10}\frac{2}{n}\frac{1}{n}(H_n^*)^{-3/2}\,n(H_n^*)^{3/2}|\tilde{W}_n(H_n^*)| \leq$$

$$(C_n^*)^{-2}n^{-4/10}\sup_{h\in I_n}(n^{3/2}h^2)\,|\hat{\hat{W}}_n(h)| +$$

$$2\,(C_n^*)^{-3/2}n^{-1}\sup_{h\in I_n}(nh^{3/2})\,|\tilde{W}_n(h)|,$$

which almost surely vanishes by statements (3.62) and (3.69) and theorem 3.21. Thus we have shown that the term $\hat{W}_n(H_n^*)$-$\tilde{W}_n(H_n^*)$ is indeed negligible in case III. Case II can be treated similarly. The terms

$$\{\frac{n-1}{n}\frac{1}{n^2}\sum_{i=1}^{n}(v_1(X_i,h) - Ev_1(X_i,h))\}\Big|_{h=H_n^*},$$

$$\{\frac{(n-1)(n-2)}{n^2}\frac{1}{n}\sum_{i=1}^{n}(w_1(X_i,h) - Ew_1(X_i,h))\}\Big|_{h=H_n^*},$$

and

$$\{-\frac{3}{2n^2}\sum_{i=1}^{n}(u_2(X_i,h) - Eu_2(X_i,h))\}\Big|_{h=H_n^*}$$

can be dealt with using lemma 3.19 and the term $\hat{V}_n(H_n^*)$ using lemma 3.20.

The three expectations appearing in (3.66) can be treated in the same way as in the proof of proposition 3.14. Lemmas 3.15 and 3.18 can then be used to show that they are also asymptotically negligible. The fact that the term $Y_n(H_n^*)$ is negligible follows from lemma 3.16. Since this is obvious for $R_n(H_n^*)$ the proof of theorem 3.8 is completed. □

3.5. Proofs.

Before we give the remaining proofs we derive the next bound on the number of points X_i in intervals of length h.

Lemma 3.24. *Let* f *be a bounded density then we have for any point* d *for some positive constant* c

$$\limsup_{n\to\infty} \sup_{h\in I_n} h^{-1} \Big| \int_{d-h}^{d+h} dF_n \Big| \le c, \textit{ almost surely.}$$ □

Proof. By a discretization argument and the Bernstein inequality for the binomial distribution, i.e. inequality (A.3) in appendix A, we can show for any $\varepsilon>0$

(3.70) $$\sup_{h\in I_n} \frac{n^{1/2}}{h^{1/2}(\log n)^{1/2+\varepsilon}} \Big| \int_{d-h}^{d+h} d(F_n-F) \Big| = o(1), \text{ almost surely.}$$

Since f is bounded we also have for some positive constant c'

$$\Big| \int_{d-h}^{d+h} dF \Big| \le c'h,$$

for all $h\in I_n$. Together these bounds complete the proof by the triangle inequality. □

Proof of lemma 3.16. First we introduce some notation. Define f_{nh}^K and f_{nh}^L by

$$f_{nh}^K(x) := \frac{1}{nh}\sum_{j=1}^{n} K((x - X_j)/h)$$

and

$$f_{nh}^L(x) := \frac{1}{nh}\sum_{j=1}^{n} L((x - X_j)/h),$$

so $f_{nh}^K(x)$ is the usual kernel estimator and $f_{nh}^L(x)$ is of the same form except for the fact that L is not a probability density. Specifically it integrates to zero instead of to one (see (L.1) - (L.4) in section 3.2.1). Next define the random variables S_n^K and S_n^L by

(3.71) $$S_n^K := \sup_{h\in I_n} \sup_{x\in E} \Big(\frac{nh}{\log n}\Big)^{1/2} |f_{nh}^K(x) - Ef_{nh}^K(x)|$$

and

(3.72) $$S_n^L := \sup_{h\in I_n} \sup_{x\in E} \Big(\frac{nh}{\log n}\Big)^{1/2} |f_{nh}^L(x) - Ef_{nh}^L(x)|.$$

Notice that by theorem 2.16 we have with probability one for some constant $c > 0$

$$(3.73)\qquad \limsup_{n\to\infty} S_n^K \le c$$

and

$$(3.74)\qquad \limsup_{n\to\infty} S_n^L \le c,$$

It follows that for any subset E' of E for all $x \in E'$ and all $h \in I_n$ we have

$$| f_{nh}^K(x) - f(x) | \le$$

$$(3.75)\qquad | f_{nh}^K(x) - Ef_{nh}^K(x) | + | Ef_{nh}^K(x) - f(x) | \le$$

$$\Big(\frac{\log n}{nh}\Big)^{1/2} S_n^K + \sup_{x\in E'} |b^K(x,h)|,$$

and similarly for f_{nh}^L,

$$(3.76)\qquad | f_{nh}^L(x) | \le$$

$$\Big(\frac{\log n}{nh}\Big)^{1/2} S_n^L + \sup_{x\in E'} |b^L(x,h)|.$$

Here the functions b^K and b^L are defined in definition 3.10.

Let $d_1, \ldots, d_m$ denote the singular points of f in the open interval (a,b) and let D_h as usual denote the set of points on the real line which are at least at a distance h of all the singular points of f. Notice that $Y_n(h)$, defined by (3.36), can be written as

$$Y_n(h) = -\frac{1}{nh}\sum_{j=1}^{n} \Big\{f_{nh}^L(X_i) - \frac{1}{nh}K(0)\Big\}\Big\{f_{nh}^K(X_i) - f(X_i) - \frac{1}{nh}K(0)\Big\}^2$$

$$\Big\{f_{nh}^K(X_i) - \frac{1}{nh}K(0)\Big\}^{-1} f(X_i)^{-2} I_E(X_i) =$$

$$Y_n^{(1)}(h) + Y_n^{(2)}(h),$$

with

$$Y_n^{(1)}(h) := -\frac{1}{nh}\sum_{j=1}^{n} \Big\{f_{nh}^L(X_i) - \frac{1}{nh}K(0)\Big\}\Big\{f_{nh}^K(X_i) - f(X_i) - \frac{1}{nh}K(0)\Big\}^2$$

$$\Big\{f_{nh}^K(X_i) - \frac{1}{nh}K(0)\Big\}^{-1} f(X_i)^{-2} I_{E\cap D_h}(X_i)$$

and

$$Y_n^{(2)}(h) := -\frac{1}{nh}\sum_{j=1}^{n} \Big\{f_{nh}^L(X_i) - \frac{1}{nh}K(0)\Big\}\Big\{f_{nh}^K(X_i) - f(X_i) - \frac{1}{nh}K(0)\Big\}^2$$

$$\Big\{f_{nh}^K(X_i) - \frac{1}{nh}K(0)\Big\}^{-1} f(X_i)^{-2}$$

$$\Big\{I_{E\cap[a,a+h]}(X_i) + I_{E\cap[b-h,b]}(X_i) + \sum_{k=1}^{m} I_{E\cap[d_k-h,d_k+h]}(X_i)\Big\}.$$

For $Y_n^{(1)}(h)$ we have by (3.75) and (3.76)

$$|Y_n^{(1)}(h)| \leq \frac{1}{h}\Big\{\Big(\frac{\log n}{nh}\Big)^{1/2} S_n^L + \sup_{x\in E\cap D_h} |b^L(x,h)| + \frac{1}{nh}K(0)\Big\}$$

$$\Big\{\Big(\frac{\log n}{nh}\Big)^{1/2} S_n^K + \sup_{x\in E\cap D_h} |b^K(x,h)| + \frac{1}{nh}K(0)\Big\}^2$$

$$\Big\{\inf_{x\in E} f(x) - \Big(\frac{\log n}{nh}\Big)^{1/2} S_n^K - \sup_{x\in E\cap D_h} |b^K(x,h)| - \frac{1}{nh}K(0)\Big\}^{-1}\Big\{\inf_{x\in E} f(x)\Big\}^{-2},$$

for all $h\in I_n$. By lemma 3.11 we have for some positive constant c' and for n large enough uniformly for $h\in I_n$

$$\sup_{x\in E\cap D_h} |b^K(x,h)| \leq c'h^2$$

and

$$\sup_{x\in E\cap D_h} |b^L(x,h)| \leq c'h^2,$$

Since for n large enough uniformly for $h\in I_n$ we have

$$\frac{1}{nh} < \Big(\frac{\log n}{nh}\Big)^{1/2} \to 0$$

the term $\frac{1}{nh}K(0)$ is asymptotically negligible. A combination of these bounds then gives

$$\limsup_{n\to\infty} \sup_{h\in I_n} h\,\Big(\Big(\frac{\log n}{nh}\Big)^{1/2} + h^2\Big)^{-3} |Y_n^{(1)}(h)| < c, \text{ almost surely.}$$

Using $(x + y)^3 \leq 2^3(x^3 + y^3)$ for all x,y>0 we obtain

$$\text{(3.77)} \qquad \limsup_{n\to\infty} \sup_{h\in I_n} \Big(\frac{1}{h}\Big(\frac{\log n}{nh}\Big)^{3/2} + h^5\Big)^{-1} |Y_n^{(1)}(h)| < 2^{-3}c, \text{ almost surely.}$$

Since in case III the term $Y_n^{(2)}(h)$ is equal to zero for all $h\in I_n$ for n large enough we also have (3.77) for $Y_n(h)$ which proves the case III part of the lemma. Next assume that we are dealing with a case II situaution and consider the term $Y_n^{(2)}(h)$. If $N_n(h)$ denotes the number of points X_i in the set

$$\text{(3.78)} \qquad E\backslash D_h = [a,a+h] \cup [b-h,b] \cup \bigcup_{i=1}^{m} [d_i-h,d_i+h],$$

then by lemma 3.24 we have for some positive constant c

$$\limsup_{n\to\infty} \sup_{h\in I_n} \frac{N_n(h)}{nh} \leq c, \text{ almost surely.}$$

By lemma 3.11 we have for any point d where f has a kink, for some positive constant c' and for n large enough uniformly for $h \in I_n$

$$\sup_{x \in E \cap [d-h,d+h]} |b^K(x,h)| \leq c'h,$$

and

$$\sup_{x \in E \cap [d-h,d+h]} |b^L(x,h)| \leq c'h.$$

By a similar argument as above, taking into account the number of points X_i in the set (3.78), we find for some positive constant c''

$$\limsup_{n \to \infty} \sup_{h \in I_n} \left(\left(\frac{\log n}{nh}\right)^{3/2} + h^3 \right)^{-1} |Y_n^{(2)}(h)| < c'', \text{ almost surely.} \tag{3.79}$$

From (3.77) and (3.79) the lemma follows for case II. □

Proof of lemma 3.17. We use the same notation as in the previous proof. Since

$$\frac{1}{h}\left(\frac{\log n}{nh}\right)^{3/2} + h^5 < \frac{1}{nh^2} + 1,$$

for all $h \in I_n$ for n large enough (3.77) implies

$$\sup_{h \in I_n} \left(\frac{1}{nh^2} + 1\right)^{-1} |Y_n^{(1)}(h)| = o(1), \text{ almost surely.}$$

Now consider $Y_n^{(2)}(h)$. We can write $Y_n^{(2)}(h)$ as the sum of m+2 terms

$$\begin{aligned} - \frac{1}{nh} \sum_{j=1}^{n} & \left\{ f_{nh}^L(X_i) - \frac{1}{nh} K(0) \right\} \left\{ f_{nh}^K(X_i) - f(X_i) - \frac{1}{nh} K(0) \right\}^2 \\ & \left\{ f_{nh}^K(X_i) - \frac{1}{nh} K(0) \right\}^{-1} f(X_i)^{-2} I_{E \cap [d-h,d+h]}(X_i), \end{aligned} \tag{3.80}$$

where d is one of the points a, b, $d_1, \ldots, d_m$. We can write (3.80) as

$$- \frac{1}{h} \int_{E \cap [d-h,d+h]} G_{nh}(x) dF_n(x), \tag{3.81}$$

where F_n is the empirical distribution function based on the sample $X_1, \ldots, X_n$ and

$$G_{nh}(x) := \frac{\left\{ f_{nh}^L(x) - \frac{1}{nh} K(0) \right\} \left\{ f_{nh}^K(x) - f(x) - \frac{1}{nh} K(0) \right\}^2}{\left\{ f_{nh}^K(x) - \frac{1}{nh} K(0) \right\} f(x)^2} =$$

$$\frac{\left\{b^{L}(x,h) + f_{nh}^{L}(x) - Ef_{nh}^{L}(x) - \frac{K(0)}{nh}\right\}\left\{b^{K}(x,h) + f_{nh}^{K}(x) - Ef_{nh}^{K}(x) - \frac{K(0)}{nh}\right\}^{2}}{\left\{f(x) + b^{K}(x,h) + f_{nh}^{K}(x) - Ef_{nh}^{K}(x) - \frac{K(0)}{nh}\right\} f(x)^{2}}$$

Recall that in case I lemma 3.11 implies that for some constant $c > 0$ and for n large enough we have for all $h \in I_n$

$$\sup_{x\in E} |b^{K}(x,h)| \leq c \quad \text{and} \quad \sup_{x\in E} |b^{L}(x,h)| \leq c.$$

Further, defining S_n^K and S_n^L as in (3.71) and (3.72), notice that we have

$$|f_{nh}^{K}(x) - Ef_{nh}^{K}(x)| \leq \left(\frac{\log n}{nh}\right)^{1/2} S_n^K$$

and

$$|f_{nh}^{L}(x) - Ef_{nh}^{L}(x)| \leq \left(\frac{\log n}{nh}\right)^{1/2} S_n^L.$$

By (3.73) and (3.74) and the fact that $\log n/nh$ vanishes uniformly on I_n we see that with probability one for n large enough

$$|G_{nh}(x) - g_{nh}(x)| \leq c'\left(\frac{\log n}{nh}\right)^{1/2},$$

for all $x \in E$ and $h \in I_n$, where we define $g_{nh}(x)$ by

$$g_{nh}(x) := -\frac{b^{L}(x,h)b^{K}(x,h)^{2}}{(f(x) + b^{K}(x,h))f(x)^{2}}.$$

Then, with

$$R_{n9}(h) := \frac{1}{h}\int_{E\cap[d-h,d+h]} [G_{nh}(x) - g_{nh}(x)]\, dF_n(x),$$

it follows that

$$|R_{n9}(h)| \leq c'\frac{1}{h}\left(\frac{\log n}{nh}\right)^{1/2} \left|\int_{E\cap[d-h,d+h]} dF_n(x)\right|.$$

Hence by lemma 3.24 we have now shown

$$\sup_{h\in I_n} |R_{n9}(h)| = o(1), \text{ almost surely.}$$

Next we consider the term

(3.82) $$\frac{1}{h} \int_{E\cap[d-h,d+h]} g_{nh}(x)\; dF_n(x).$$

By lemma 3.11 we have for n large enough for some constant c">0

$$\left| g_{nh}(x) - \Big(- \frac{\delta^{(0)}(d)b_0^L((x-d)/h)\delta^{(0)}(d)^2 b_0^K((x-d)/h)^2}{f(x) + \delta^{(0)}(d)b_0^K((x-d)/h)} f(x)^{-2}\Big) \right| I_{E\cap[d-h,d+h]}(x) \le c''h,$$

for all $x\in E$ and all $h\in I_n$. Then, with

$$R_{n10}(h) := \frac{1}{h} \int_{E\cap[d-h,d+h]} \Big(g_{nh}(x) + \frac{\delta^{(0)}(d)^3 b_0^L((x-d)/h)b_0^K((x-d)/h)^2}{f(x) + \delta^{(0)}(d)b_0^K((x-d)/h)} f(x)^{-2}\Big) dF_n(x),$$

we have

$$|R_{n10}(h)| \le \frac{c''}{h} h \,\Big| \int_{E\cap[d-h,d+h]} dF_n(x) \Big|,$$

and by lemma 3.24

$$\sup_{h\in I_n} |R_{n10}(h)| = o(1), \text{ almost surely.}$$

We continue with

(3.83) $$- \frac{1}{h} \delta^{(0)}(d)^3 \int_{E\cap[d-h,d+h]} \frac{b_0^L((x-d)/h)b_0^K((x-d)/h)^2}{f(x) + \delta^{(0)}(d)b_0^K((x-d)/h)} f(x)^{-2}\, dF_n(x).$$

Since the integrand in (3.83) is a bounded function, by a discretization argument and the Bernstein inequality, (A.2) in appendix A, with

$$R_{n11}(h) := - \frac{1}{h} \delta^{(0)}(d)^3 \int_{E\cap[d-h,d+h]} \frac{b_0^L((x-d)/h)b_0^K((x-d)/h)^2}{f(x) + \delta^{(0)}(d)b_0^K((x-d)/h)} f(x)^{-2}\, d(F_n\text{-}F)(x)$$

we have almost surely

$$\sup_{h\in I_n} |R_{n11}(h)| = o(1).$$

Finally, assuming that d is not equal to a or b, notice that

$$- \delta^{(0)}(d)^3 \frac{1}{h} \int_{E\cap[d-h,d+h]} \frac{b_0^L((x-d)/h)b_0^K((x-d)/h)^2}{f(x) + \delta^{(0)}(d)b_0^K((x-d)/h)} f(x)^{-2} dF(x) =$$

$$- \delta^{(0)}(d)^3 \int_{-1}^{1} \frac{b_0^L(t)b_0^K(t)^2}{f(d+th) + \delta^{(0)}(d)b_0^K(t)} f(d+th)^{-1} dt$$

converges uniformly for $h \in I_n$ to

$$(3.84) \qquad \begin{aligned} &- \delta^{(0)}(d)^3 f(d+)^{-1} \int_0^1 \frac{b_0^L(t) b_0^K(t)^2}{f(d+) + \delta^{(0)}(d) b_0^K(t)}\, dt\ + \\ &- \delta^{(0)}(d)^3 f(d-)^{-1} \int_{-1}^0 \frac{b_0^L(t) b_0^K(t)^2}{f(d-) + \delta^{(0)}(d) b_0^K(t)}\, dt\ . \end{aligned}$$

Using

$$b_0^L(t) = tK(t) = t \frac{d}{dt} b_0^K(t)$$

and

$$g'(x) = x^2/(1+x)$$

by partial integration (3.84) equals

$$\begin{aligned} &- f(d+) \int_0^1 t \frac{d}{dt} g(f(d+)^{-1} \delta^{(0)}(d) b_0^K(t)) dt\ + \\ &- f(d-) \int_0^1 t \frac{d}{dt} g(f(d-)^{-1} \delta^{(0)}(d) b_0^K(t)) dt = \\ &f(d+) \int_0^1 g(f(d+)^{-1} \delta^{(0)}(d) b_0^K(t)) dt\ + \\ &f(d-) \int_{-1}^0 g(f(d-)^{-1} \delta^{(0)}(d) b_0^K(t)) dt. \end{aligned}$$

If d equals a or b then one of the terms of (3.84) vanishes. By adding up these expansions for all the m+2 terms in (3.80) the lemma is proved. □

Proof of lemma 3.19. First notice that the conditions we have imposed on the kernel function K imply that K and L are Lipschitz functions. Using this property it can be shown that it suffices to prove the lemma for suprema over discrete subsets $\mathcal{I}_n$ of I_n with an at most algebraically fastly increasing number of elements, i.e. we assume $\#\mathcal{I}_n \leq n^a$, for some integer a. If $(\alpha_n(.))$ is a sequence of positive functions on $(0,\infty)$ then by the Bernstein inequality, i.e. inequality (A.2) in appendix A, we have for any $\varepsilon > 0$

$$P\Big(\Big|\sum_{i=1}^n (\varphi(X_i,h) - E\varphi(X_i,h))\Big| > n\log n\ \alpha_n^{-1}(h)\, \varepsilon \Big) \leq$$

$$2 \exp\left(\frac{-\ n\ (\log n\ \alpha_n^{-1}(h)\varepsilon)^2}{2\mathrm{var}(\varphi(X_1,h)) + \frac{2}{3} m(h) \log n\ \alpha_n^{-1}(h)\varepsilon} \right),$$

where m(h) is a constant such that $|\varphi(X_1,h) - E\varphi(X_1,h)| \le m(h)$ with probability one. For $\varepsilon < 1$ this bound is dominated by

$$2\exp\left(\frac{-\frac{1}{2}n\alpha_n^{-2}(h)(\log n)^2\varepsilon^2}{E\varphi^2(X_1,h) + m(h)\alpha_n^{-1}(h)\log n}\right).$$

Assume that the functions α_n can be chosen such that for some constant c>0 and for n large enough

$$\frac{n\alpha_n^{-2}(h)}{E\varphi^2(X_1,h) + m(h)\alpha_n^{-1}(h)\log n} \ge c > 0, \tag{3.85}$$

for all $h \in I_n$. If $\#I_n \le n^a$, for some integer a, then

$$P\Big(\frac{1}{n\log n}\sup_{h\in I_n}\alpha_n(h)\Big|\sum_{i=1}^{n}(\varphi(X_i,h) - E\varphi(X_i,h))\Big| > \varepsilon\Big) \le$$

$$\sum_{h\in I_n} P\Big(\Big|\sum_{i=1}^{n}(\varphi(X_i,h) - E\varphi(X_i,h))\Big| > n\log n\,\alpha_n^{-1}(h)\,\varepsilon\Big) \le$$

$$2\,\#I_n\exp(-\tfrac{1}{2}c\varepsilon^2(\log n)^2) =$$

$$2\exp(-\tfrac{1}{2}c\varepsilon^2(\log n)^2 + a\log n),$$

which is summable. Hence by the Borel-Cantelli theorem

$$\frac{1}{n\log n}\sup_{h\in I_n}\alpha_n(h)\Big|\sum_{i=1}^{n}(\varphi(X_i,h) - E\varphi(X_i,h))\Big| = o(1), \text{ almost surely.}$$

First we take φ equal to u_1. Recall that the set E is equal to the bounded interval [a,b]. We shall choose a suitable sequence of functions $(\alpha_n(.))$ and then check (3.85). Write

$$Eu_1^2(X_1,h) =$$

$$\frac{1}{h^2}\int_{-\infty}^{\infty}(b^L(x,h)f(x)^{-1}I_E(x))^2f(x)dx.$$

$$\frac{1}{h^2}\int_a^b b^L(x,h)^2f(x)dx.$$

Since the order of magnitude of b^L is different in the three cases I, II and III we also get three different bounds for this expectation. By lemma 3.11 we have for some constant c' > 0

$$(3.86) \qquad Eu_1^2(X_1,h) \le \begin{cases} c'\frac{1}{h} & \text{in case I} \\ c'h & \text{in case II} \\ c'h^2 & \text{in case III} \end{cases} ,$$

for n large enough uniformly for $h \in I_n$. Lemma 3.11 also provides us with suitable choices for m(h). We can use

$$(3.87) \qquad m(h) := \sup_x |u_1(x,h) - Eu_1(X_1,h)| \le \begin{cases} c''\frac{1}{h} & \text{in case I} \\ c'' & \text{in case II} \\ c''h & \text{in case III} \end{cases} ,$$

for n large enough uniformly for $h \in I_n$. Here c" is a positive constant. The inequalities (3.86) and (3.87) imply that the condition (3.85), i.e. for n large enough

$$\frac{n\alpha_n^{-2}}{Eu_1^2(X_1,h) + m(h)\alpha_n^{-1}(h)\log(n)} \ge c > 0, \text{ for all } h \in I_n,$$

is satisfied for the choices

$$\alpha_n(h) = \begin{cases} n^{1/2}h^{1/2} & \text{in case I} \\ n^{1/2}h^{-1/2} & \text{in case II} \\ n^{1/2}h^{-1} & \text{in case III} \end{cases} .$$

Notice that we have taken $\alpha_n(h)$ equal to $n^{1/2}$ times the inverse of the root of the bounds in (3.86). We have now shown that (3.58) is valid for the function u_1.

Next consider the function u_2. For this function we have for n large enough

$$Eu_2^2(X_1,h) =$$

$$\frac{1}{h^4}\int_{-\infty}^{\infty}\Big(\int_a^b L((u-x)/h)du\Big)^2 f(x)dx \le$$

$$\frac{1}{h^4}\int_{-\infty}^{\infty}\Big(\int_{[x-h,x+h]\cap[a,b]} L((u-x)/h)du\Big)^2 f(x)dx \le$$

$$\frac{1}{h^4}\Big(\int_a^{a+h} + \int_{b-h}^{b}\Big)\Big(\int_{[x-h,x+h]\cap[a,b]} L((u-x)/h)du\Big)^2 f(x)dx ,$$

since $\int_{-1}^{1} L(u)du$ is equal to zero. It follows that for some constant c' > 0 we have

$$Eu_2^2(X_1,h) \le c'\frac{1}{h^4}h\,h^2 = c'\,\frac{1}{h} ,$$

uniformly for $h \in I_n$ and n large enough. We also have for some constant $c'' > 0$

$$m(h) := \sup_x |u_2(x,h) - Eu_2(X_1,h)| \le c''\frac{1}{h^2}h = c''\frac{1}{h}.$$

It is readily verified that with the choice $\alpha_n(h) = n^{1/2}h^{1/2}$ condition (3.85) is satisfied which proves statement (3.58) for the function u_2. For the other functions this statement can be proved in the same manner. □

Proof of lemma 3.20. Using the fact that the functions K and L are Lipschitz functions it can be shown that it suffices to prove (3.59) for finite subsets I_n of I_n instead of for the intervals I_n themselves. We choose I_n such that the number of its points increases sufficiently rapidly but still at most algebraically fast in n. The Lipschitz property can be used to show that sufficiently small changes in h result in negligible changes in $\hat{U}_n(h)$, $\hat{V}_n(h)$ and $\hat{W}_n(h)$.
We start with $\hat{U}_n(h)$. Write

$$\sum_{n=1}^{\infty} P(\sup_{h\in I_n} nh^{3/2}n^{-\alpha}|\hat{U}_n(h)| > \varepsilon) \le$$

$$\sum_{n=1}^{\infty}\sum_{h\in I_n} P(nh^{3/2}n^{-\alpha}|\hat{U}_n(h)| > \varepsilon) \le$$

$$\sum_{n=1}^{\infty} \#I_n \sup_{h\in I_n} P(nh^{3/2}n^{-\alpha}|\hat{U}_n(h)| > \varepsilon) \le$$

$$\sum_{n=1}^{\infty} \#I_n \sup_{h\in I_n} (\varepsilon n^{-1}h^{-3/2}n^{\alpha})^{-p}\, E(\hat{U}_n(h))^p =$$

$$\varepsilon^{-p}\sum_{n=1}^{\infty} \#I_n\, n^{-\alpha p}\sup_{h\in I_n} (nh^{3/2})^p\, E(\hat{U}_n(h))^p,$$

for every even positive integer p. Here $\#I_n$ denotes the number of elements of I_n. In order to show that this sum is finite, which would enable us to apply the Borel-Cantelli theorem, we derive a bound for the p-th moment of $\hat{U}_n(h)$. Recall

$$\hat{U}_n(h) = n^{-2}\sum_{i\ne j} \hat{U}_{ij}(h),$$

where $\hat{U}_{ij}(h)$ is defined above. Since

$$E(\hat{U}_{ij}(h)|X_k) = E(\hat{U}_{ij}(h)|X_k) = 0\ , \text{ for } k=1,\dots,n,$$

any product of $\hat{U}_{ij}$'s, such as $\hat{U}_{i_1j_1}(h)\dots\hat{U}_{i_pj_p}(h)$, with at least one index i or j appearing only once in $i_1,j_1,\dots,i_p,j_p$, has zero expectation. Therefore

$$E(\hat{U}_n(h))^p = n^{-2p} \sum_{m=2}^{p} ES_m,$$

where S_m is the sum of all products $\hat{U}_{i_1j_1}(h)\ldots\hat{U}_{i_pj_p}(h)$, with $i_1,j_1,\ldots,i_p,j_p$ containing exactly m different indices, every index appearing at least twice. Since $X_1,\ldots,X_n$ are identically distributed we can rewrite ES_m as

$$ES_m = \binom{n}{m} E\tilde{S}_m,$$

with $\tilde{S}_m$ equal to the sum of all possible terms of S_m with indices in $\{1,2,\ldots,m\}$. Since $m \le p$ the number of such terms is bounded by a constant depending only on p, c_p, say. From corollary B.3 (appendix B) it follows that the expectation of the absolute value of each of the terms appearing in $\tilde{S}_m$ is bounded by a constant times $h^{m/2-2p}$. Therefore

$$E(\hat{U}_n(h))^p \le n^{-2p} \sum_{m=2}^{p} \binom{n}{m} c_p\, h^{m/2-2p} \le$$

$$\tilde{c}_p n^{-2p} h^{-2p} \sum_{m=2}^{p} n^m h^{m/2} \le$$

$$\tilde{c}_p (n^{-2}h^{-2})^p \sum_{m=2}^{p} (nh^{1/2})^m =$$

$$\tilde{c}_p (n^{-2}h^{-2})^p ((nh^{1/2})^{p+1} - (nh^{1/2})^2)(nh^{1/2}-1) \le$$

$$2\tilde{c}_p (nh^{3/2})^{-p},$$

for n large enough. Here $\tilde{c}_p$ denotes a constant independent of h and n. We have used that for large n and $h \in I_n$ we have $nh^{1/2} > nh \ge nh_n' \to \infty$.

Combining these bounds and assuming $\#I_n \le n^a$ for some positive integer a, we have for every positive integer p, even and large enough, and for every $\varepsilon > 0$

$$\sum_{n=1}^{\infty} P(\sup_{h \in I_n} nh^{3/2} n^{-\alpha} |\hat{U}_n(h)| > \varepsilon) \le$$

$$2\tilde{c}_p \varepsilon^{-p} \sum_{n=1}^{\infty} n^a n^{-\alpha p} \sup_{h \in I_n} (nh^{3/2})^p (nh^{3/2})^{-p} =$$

$$2\tilde{c}_p \varepsilon^{-p} \sum_{n=1}^{\infty} n^{a-\alpha p} < \infty,$$

which proves the statement of this lemma for $\hat{U}_n(h)$ by the Borel-Cantelli lemma.

The argument for $\hat{V}_n(h)$ is similar. Since by corollary B.3 the bound on the expectation of the absolute values of the terms appearing in $\tilde{S}_m$ is of order $h^{m/2-3p}$, here it leads to

$$E(\hat{V}_n(h))^p \leq 2\tilde{c}_p(n^{-3}h^{-3})^p(nh^{1/2})^p = 2\tilde{c}_p(n^{-2}h^{-5/2})^p,$$

for n large enough.

For $\hat{W}_n(h)$ the argument is similar too. In this case by corollary B.3 the bound on the expectation of the absolute values of the terms appearing in $\tilde{S}_m$ is of order $h^{2m/3-3p}$. Hence the bound for the p-th moment becomes

$$E(\hat{W}_n(h))^p \leq (n^{-3}h^{-3})^p \sum_{m=2}^{[3p/2]} \binom{n}{m} c_p^* h^{2m/3} \leq$$

$$2\,\tilde{c}_p^*\,(n^{-3}h^{-3})^p(nh^{2/3})^{3p/2} =$$

$$2\,\tilde{c}_p^*\,(n^{-3/2}h^{-2})^p,$$

for n large enough. Just as before c_p^* and $\tilde{c}_p^*$ are constants depending only on p. To derive these inequalities we have used $2m \leq 3p$ and that for large n and $h \in I_n$ we have $nh^{2/3} > nh \geq nh_n' \to \infty$. □

Proof of lemma 3.22. Since the proof of this lemma is tedious and very similar to the proofs of lemma 3.2 and lemma 3.3 in Hall & Marron (1987a) we only mention the basic steps. For case III it suffices to show that for some $\varepsilon_1>0$ we have

$$|H_n^* - c_{opt}n^{-1/5}| = O_p(n^{-1/5-\varepsilon_1}), \tag{3.88}$$

and that for all $\varepsilon_2>0$ we have

$$\begin{aligned} \sup_{|t - c_{opt}| \leq n^{-\varepsilon_2}} n^{7/10}\Big\{ \big|\hat{U}_n(tn^{-1/5}) - \hat{U}_n(c_{opt}n^{-1/5})\big| + \Big|\frac{n-1}{n^2}\sum_{i=1}^{n}(u_1(X_i,tn^{-1/5}) - Eu_1(X_i,tn^{-1/5})) \\ - \frac{n-1}{n^2}\sum_{i=1}^{n}(u_1(X_i,c_{opt}n^{-1/5}) - Eu_1(X_i,c_{opt}n^{-1/5}))\Big|\Big\} \overset{\mathcal{P}}{\to} 0, \end{aligned} \tag{3.89}$$

For case II we have to prove two similar properties, i.e. (3.88) with $n^{-1/5}$ replaced by $n^{-1/4}$, and (3.89) with $n^{-1/5}$ replaced by $n^{-1/4}$ and $n^{7/10}$ replaced by $n^{5/8}$. □

Proof of lemma 3.23. For any sequence of bandwidths (h_n) we define $T_n(h_n)$ by

$$T_n(h_n) := (nh_n)^2T_n^{(1)}(h) = (nh_n)^2\Big(\hat{U}_n(h_n) + \frac{n-1}{n}\frac{1}{n}\sum_{i=1}^{n} (u_1(X_i,h_n) - Eu_1(X_i,h_n))\Big).$$

Then the expectation of $T_n(h_n)$ is equal to zero and by the definition of $\hat{U}_n(h)$ we see that $T_n(h_n)$ equals

$$(nh_n)^2\Big(\frac{1}{n^2}\sum_{i\neq j} (U_{ij}(h_n) - E(U_{ij}(h_n)|X_i) - E(U_{ij}(h_n)|X_j) + EU_{ij}(h_n)) +$$

$$\frac{n-1}{n}\frac{1}{n}\sum_{i=1}^{n} (u_1(X_i,h_n) - Eu_1(X_i,h_n))\Big) =$$

$$\sum_{i\neq j} h_n^2 EU_{ij}(h_n) - (n-1)h_n^2 \sum_{i=1}^{n} Eu_1(X_i,h_n) + \sum_{i\neq j} h_n^2 U_{ij}(h_n) +$$

$$\sum_{i=1}^{n} \Big(-(n-1)h_n^2(E(U_{ij}(h_n)|X_i) + E(U_{ji}(h_n)|X_i)) + (n-1)h_n^2\, u_1(X_i,h_n)\Big).$$

Next we write

$$\sum_{i\neq j} h_n^2 EU_{ij}(h_n) = - \sum_{i\neq j} 2L((X_i-X_j)/h_n)f(X_i)^{-1}I_E(X_i) =$$

$$- \sum_{i\neq j} L((X_i-X_j)/h_n)\ (f(X_i)^{-1}I_E(X_i) + f(X_j)^{-1}I_E(X_j)) =$$

$$\sum_{i\neq j} G((X_i-X_j)/h_n)w(X_i,X_j),$$

where G is equal to -L and w is the function given by

$$w(x,y) := f(x)^{-1}I_E(x) + f(y)^{-1}I_E(y).$$

Thus $T_n(h_n)$ equals

$$\sum_{i\neq j} h_n^2 EU_{ij}(h_n) - (n-1)h_n^2 \sum_{i=1}^{n} Eu_1(X_i,h_n) +$$

$$\sum_{i\neq j} G((X_i-X_j)/h_n)w(X_i,X_j) + \sum_{i=1}^{n} g_n(X_i),$$

with $g_n(X_i)$ equal to $-(n-1)h_n^2(E(U_{ij}(h_n)|X_i) + E(U_{ji}(h_n)|X_i)) + (n-1)h_n^2\, u_1(X_i,h_n)$. Asymptotic normality of this type of statistic is treated in appendix C. In order to apply theorem C.1 notice that the function $g_n^*(x)$ which appears in the conditions of this theorem is here given by

$$g_n^*(x) = (n-1)h_n^2\, u_1(x,h) = - (n-1)h_n\, b^L(x,h)f^{-1}(x)I_E(x).$$

Condition (i) of theorem C.1 then requires

$$(nh_n^{1/2})^{-1} \sup_x |\, g_n^*(x) - Eg_n^*(X_1)\, | \to 0,$$

which is clearly satisfied here.To check condition (ii) we consider

$$(nh_n)^{-1}\mathrm{var}(g_n^*(X_1)) =$$

$$(3.90)\qquad (nh_n)^{-1}\Big((n-1)^2h_n^2\int_E b^L(x,h)^2f^{-2}(x)f(x)dx - (n-1)^2h_n^2\Big(\int_E b^L(x,h)dx\Big)^2\Big) =$$

$$\frac{(n-1)^2}{n}h_n\Big(\int_E b^L(x,h)^2f^{-1}(x)dx - \Big(\int_E b^L(x,h)dx\Big)^2\Big).$$

It follows from theorem C.1 that if this quantity converges to a constant α^2 then

$$\frac{1}{nh_n^{1/2}}(T_n(h_n) - ET_n(h_n)) = \frac{1}{nh_n^{1/2}}T_n(h_n) \overset{\mathcal{D}}{\to} N(0,2\sigma^2+\alpha^2),$$

with

$$\sigma^2 := \int_{-\infty}^{\infty} G^2(v)dv \int_{-\infty}^{\infty} w^2(x)f^2(x)dx.$$

In our case σ^2 equals

$$\int_{-1}^{1} L^2(v)dv \int_{-\infty}^{\infty}(f(x)^{-1}I_E(x) + f(x)^{-1}I_E(x))^2f^2(x)dx = 4(b-a)\int_{-1}^{1} L^2(v)dv.$$

So if (3.90) indeed converges to α^2 then we have shown

$$(3.91)\qquad nh_n^{3/2}\Big(\hat{U}_n(h_n) + \frac{n-1}{n}\frac{1}{n}\sum_{i=1}^{n}(u_1(X_i,h_n) - Eu_1(X_i,h_n))\Big) = \frac{1}{nh_n^{1/2}}T_n(h_n) \overset{\mathcal{D}}{\to} N(0,2\sigma^2+\alpha^2).$$

We proceed with computing α^2 in the two cases considered in this lemma. By lemma 3.11 in case II we have the following expansion for $h_n=cn^{-1/4}$

$$\frac{(n-1)^2}{n}h_n\Big(\int_E b^L(x,h)^2f^{-1}(x)dx - \Big(\int_E b^L(x,h)dx\Big)^2\Big) \sim$$

$$nh_nh_n^3\,\Delta^{(1)}\int_0^1 b_1^L(t)^2dt =$$

$$c^4\Delta^{(1)}\int_0^1 b_1^L(t)^2dt$$

and in case III for $h_n=cn^{-1/5}$ we have

$$\frac{(n-1)^2}{n}h_n\Big(\int_E b^L(x,h)^2f^{-1}(x)dx - \Big(\int_E b^L(x,h)dx\Big)^2\Big) \sim$$

$$nh_n\Big(\tfrac{1}{4}h_n^4\Big(\int_{-1}^{1} u^2L(u)du\Big)^2\int_E f''(x)^2f^{-1}(x)dx - \Big(\tfrac{1}{2}h_n^2\int_{-1}^{1} u^2L(u)du\int_E f''(x)dx\Big)^2\Big) =$$

$$\tfrac{1}{4}n^{-1}h_n^5\Big(\int_{-1}^{1} u^2L(u)du\Big)^2\Big(\int_a^b f''(x)^2f^{-1}(x)dx - (f'(b) - f'(a))^2\Big) =$$

$$c^5\tfrac{1}{4}\Big(\int_{-1}^{1} u^2L(u)du\Big)^2\Big(\int_a^b f''(x)^2f^{-1}(x)dx - (f'(b) - f'(a))^2\Big).$$

These two expansions can be derived by the same method we have used in the proof of theorem 2.8. The proof of the lemma is completed by observing that the norming factor $nh_n^{3/2}$ in (3.91) is equal to $c^{3/2}n^{5/8}$ if h_n is equal to $cn^{-1/4}$, and that it is equal to $c^{3/2}n^{7/10}$ if h_n is equal to $cn^{-1/5}$. □

Proof of relations (3.44), (3.45), (3.54) and (3.55). To prove the relations it suffices to consider any particular density f. Define f by

$$f(x) = \begin{cases} 0 & \text{if } x<0 \\ 1+x & \text{if } 0\leq x\leq\sqrt{3}-1 \\ 0 & \text{if } x\geq\sqrt{3}-1 \end{cases}.$$

Computing the bias of a kernel estimator of f at the point th for $0<t<1$ and $0<h<\frac{\sqrt{3}-1}{2}$ we get

$$b^K(th,h) = \frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{th-u}{h}\right)f(u)du - f(th) =$$

$$\int_{-\infty}^{\infty} K(t-v)f(vh)dv - f(th) =$$

$$\int_{0}^{\infty} K(t-v)(1+vh)dv - (1+th) =$$

$$b_0^K(t) + hb_1^K(t).$$

Similarly for $0<t<1$ and $0<h<\frac{\sqrt{3}-1}{2}$ the bias function for L is equal to

$$b^L(th,h) = b_0^L(t) + hb_1^L(t).$$

Next recall that by the definition of $b^K(x,h)$ and $b^L(x,h)$, and by (3.33) we have

$$\frac{d}{dh}b^K(x,h) = -\frac{1}{h}b^L(x,h).$$

In order to prove relations (3.54) and (3.55) consider the equation

$$\frac{d}{dh}\int_0^h b^K(x,h)dx = \int_0^h \frac{d}{dh}b^K(x,h)dx + b^K(h,h) = -\frac{1}{h}\int_0^h b^L(x,h)dx,$$

which follows from Leibnitz's theorem for differentiation of integrals, i.e. formula 3.3.7 in Abramowitz & Stegun (1965). By the substitution t=x/h we get

$$\frac{d}{dh}h\int_0^1 (b_0^K(t) + hb_1^K(t))dt = -\int_0^1 (b_0^L(t) + hb_1^L(t))dt,$$

which proves formulas (3.54) and (3.55) by comparing the constant term and the coefficient of h in the left and right hand side of this equality.

Relations (3.44) and (3.44) can be proved similarly by considering the equation

$$\frac{d}{dh}\int_0^h b^K(x,h)^2dx = \int_0^h \frac{d}{dh}b^K(x,h)^2dx + b^K(h,h)^2 = -\frac{2}{h}\int_0^h b^K(x,h)b^L(x,h)dx,$$

and comparing the constant terms and the coefficients of h^2. □

4. RECOVERING A DISTRIBUTION FUNCTION FROM A CONVOLUTION.

4.1. Introduction and results.

Suppose that we have a sample $X_1,...,X_n$ of observations with a distribution function G which is the convolution of two other distribution functions K and F, i.e. for all x we have

$$G(x) = \int_{-\infty}^{\infty} K(x-y)dF(y).$$

Assuming that the function K is known we consider the problem of estimating F at a fixed point x_0, in cases where the distribution function F is uniquely determined by G and K (if, for example, K is a distribution function with a characteristic function with compact support, this need not be true). Our main theorem gives lower bounds for a local minimax risk for estimation of $F(x_0)$ in two cases. It turns out that the rate of convergence to zero of the minimax risk depends on the smoothess properties of K. Theorem 4.1 states that if K has a density k with jumps then the rate of the lower bound is equal to $n^{-1/3}$, but, on the other hand, if the density k is smooth enough, then the lower bound has the larger rate $n^{-1/4}$! Supplementary to this result we show that two other minimax risks do not converge to zero.

We give three examples. In the first example we derive the nonparametric maximum likelihood estimator, NPMLE, of F in one particular case, where K is the uniform distribution function and the support of F is contained in the interval [0,1]. We show that this estimator converges with a rate $n^{-1/3}$. This suggests that the rate of the minimax lower bound in this case is sharp. However, a rigorous proof would require an additional uniformity argument. In the second example we propose an algorithm for computing the NPMLE of F for a different K. In the third example we use the convolution structure in the Wicksell problem, example 1.2, to derive the NPMLE of the distribution function of the sphere radii.

Let x_0 be a fixed point in the support of F and let (γ_n) be a vanishing sequence of positive numbers to be specified later. To define the local minimax risk for $n \in \mathbb{N}$ and $\theta \in (0,1)$, we introduce the functions $h_n(.)$ and $F_n(.;\theta)$, given by

$$h_n(u) := f(x_0)\{I_{(x_0-c\gamma_n,x_0)}(u) - I_{[x_0,x_0+c\gamma_n)}(u)\}$$

and

$$F_n(x;\theta) := F(x) + \theta\int_0^x h_n(u)du,$$

assuming that $f(x_0)$,the derivative of F at x_0, exists and is positive. Note that for n sufficiently large $F_n(.;\theta)$ is a distribution function.

We define the minimax risk $MR(n;0,\delta)$ by

$$MR(n;0,\delta) := \inf_{U_n} \max_{\theta\in\{0,\delta\}} E_\theta |U_n - F_n(x_0;\theta)|,$$

where the infimum is taken over the set of all possible estimators U_n of $F(x_0)$ based on the observations $X_1,...,X_n$ from the distribution $K*F$. Thus $MR(n;0,\delta)$ is the best possible maximal expected error for estimating the two values $F(x_0)=F_n(x_0;0)$ and $F_n(x_0;\delta)$.

We restrict ourselves to absolutely continuous distribution functions K with densities k. In theorem 4.1 we consider densities k which satisfy one of the next two conditions. Notice that by condition (A) we can use left and right one term Taylor expansions of k in the points $a_1,...,a_m$.

Conditions on k.

(A) The density k is differentiable except in m points $a_1,...,a_m$ where k has a jump. In these points the left and right limits of k exist and are finite, as well as the left and right derivatives of k. We further assume that, for $i = 1,...,m-1$, the restriction of k' to the interval (a_i,a_{i+1}) can be extended to a continuous function on $[a_i,a_{i+1}]$, such that the values at the endpoints coincide with the corresponding one-sided derivatives. We use similar assumptions on the intervals $(-\infty,a_1]$ and $[a_m,\infty)$ for the right and left endpoint, respectively.

(B) The density k is continuously differentiable on $\mathbb{R}$.

Let $*$ denote convolution, i.e. $K*F$ is the convolution of the distribution functions K and F, and $k*F$ denotes its density.

Theorem 4.1. *Assume that both* K *and* F *have a bounded support.*
(a) *If* k *satisfies condition* (A) *and* $\gamma_n=n^{-1/3}$ *then*

$$\sup_{\delta\in(0,1),c>0} \liminf_{n\to\infty} n^{1/3}\, MR(n;0,\delta) \geq \frac{3^{4/3}}{16} f(x_0)^{1/3} \Big(\sum_{i=1}^{m} \frac{(k(a_i+) - k(a_i-))^2}{(k*F)(x_0+a_i)}\Big)^{-1/3}.$$

(b) *If* k *satisfies condition* (B) *and* $\gamma_n=n^{-1/4}$ *then*

$$\sup_{\delta\in(0,1),c>0} \liminf_{n\to\infty} n^{1/4}\, MR(n;0,\delta) \geq 2^{3/4}5^{-5/4} f(x_0)^{1/2} \Big(\int_{-\infty}^{\infty} \frac{k'(x-x_0)^2}{(k*F)(x)} dx\Big)^{-1/4}. \qquad \square$$

The theorem suggests that the lower bounds for the minimax risk in the case of a nice smooth density k have a slower rate of convergence to zero than in the case of certain densities k with jump discontinuities. This is further illustrated in the next remark.

Remark 4.2. Suppose that F has a bounded density. If $K(x) = x^{\alpha} I_{[0,\infty)}$ for all $x \le \varepsilon$ for some $\varepsilon > 0$, if K satisfies the conditions of theorem 1.1(b) on the interval $[\frac{\varepsilon}{2},\infty)$ and if $\gamma_n = n^{-1/(2\alpha+1)}$, then we have for $0 \le \alpha < \frac{3}{2}$

$$\sup_{\delta\in(0,1),c>0} \liminf_{n\to\infty} n^{1/(2\alpha+1)} MR(n;0,\delta) \ \ge\ f(x_0)^{(2\alpha-1)/(2\alpha+1)} m(\alpha) > 0,$$

where $m(\alpha)$ depends on α only. We need the assumption that F has a bounded density to ensure that $k{*}F$ is continuous. This bound can be proved by the same arguments we have used for the proof of part (a) of theorem 4.1. An interesting feature of this bound is that for $\alpha=1/2$ the rate becomes $n^{1/2}$. However, it is still an open question whether this rate can actually be achieved by some particular estimator. Clearly for $0 \le \alpha < 1/2$ the rate is not sharp, since in that case it is smaller than $n^{1/2}$.

We assumed above that the derivative of the distribution function F at x_0 exists and is positive. The next theorem deals with a situation where this condition is not fulfilled. It states that the minimax risk for estimating the value of two degenerate distribution functions at a fixed point x_0 does not converge to zero.

Theorem 4.3. *Let , for* $\theta \ge 0$, *the distribution function* $F(.;\theta)$ *be defined by* $F(x;\theta) := I_{[x_0+\theta,\infty)}(x)$. *Then we have*

$$\liminf_{n\to\infty} \inf_{U_n} \max_{\theta\in\{0,\theta_n\}} E_\theta \,|U_n - F(x_0;\theta)| \ge \frac{1}{4},$$

provided θ_n *decreases to zero sufficiently fast.* □

Although our main interest in this chapter is estimation of the distribution function F we mention one result on estimation of the density f of F, assuming that F is absolutely continuous.

Theorem 4.4. *For any density* k *we have*

$$\inf_{f_n} \sup_{f\in\mathcal{F}} E_f \int_{-\infty}^{\infty} |f_n(x) - f(x)| dx \ge 1,$$

where $\mathcal{F}$ *denotes the class of all densities on the real line and the infimum is taken over the set of estimators* f_n *of the density* f *based on samples of size* n *from the distribution* $K{*}F$. □

In other words, the theorem states that for any $\varepsilon>0$ and any estimator f_n there exists a density f such that the expected L_1-distance between f_n and f is larger than $1-\varepsilon$. The proof of this result is based on a theorem in Devroye (1987) for estimation of a density within a convolution family. It should be noted that this proof requires that the supports of the densities $f \in \mathcal{F}$ are not uniformly bounded. It is not clear if the minimax risk converges to zero if we consider densities with uniformly bounded supports.

We close this section with some examples of situations where the nonparametric maximum likelihood estimator of F can actually be computed.

Example 4.5. For the special case that K is the uniform distribution function, the lower bound given by part (a) of theorem 1.1 becomes

$$\frac{3^{4/3}}{16} f(x_0)^{1/3} \Big(\frac{1}{(F(x_0) - F(x_0-1))} + \frac{1}{(F(x_0+1) - F(x_0))} \Big)^{-1/3},$$

and if we assume that F is concentrated on the interval [0,1] this bound reduces to

$$\frac{3^{4/3}}{16} f(x_0)^{1/3} \Big(\frac{1}{F(x_0)} + \frac{1}{(1 - F(x_0))} \Big)^{-1/3} =$$

$$\frac{3^{4/3}}{16} f(x_0)^{1/3} \Big(\frac{1}{F(x_0)(1 - F(x_0))} \Big)^{-1/3}.$$

Now, let $X_1,\ldots,X_n$ be a sample, generated by the density g, defined by

$$(4.1) \qquad g(x) = \int_{-\infty}^{\infty} k(x-y)dF(y),\ x \in \mathbb{R},$$

where k is the uniform density on [0,1]. We want to find the nonparametric maximum likelihood estimator (NPMLE) of F. Defining $\delta_i = 1_{\{X_i \le 1\}}$, the log likelihood, based on $X_1,\ldots,X_n$, can be written

$$\sum_{i=1}^{n} \log \int_{-\infty}^{\infty} k(X_i - y)dF(y) = \sum_{i=1}^{n} \log\{F(X_i) - F(X_i - 1)\} =$$

$$= \sum_{i=1}^{n} \{\delta_i \log F(X_i) + (1-\delta_i)\log\{1 - F(X_i - 1)\}\}.$$

Now let Y_i, $1 \le i \le n$, be defined by

$$Y_i := \begin{cases} X_i \ , \text{ if } X_i \le 1, \\ X_i - 1, \text{ if } X_i > 1. \end{cases}$$

Then $Y_1,\ldots,Y_n$ are distributed as a sample from a uniform distribution on [0,1]. Let $Z_1 \le\ldots\le Z_n$ denote the order statistics of the set $Y_1,\ldots,Y_n$, and let $\Delta_j = 1$, if the X_k, corresponding to Z_j, is ≤ 1, and let $\Delta_j = 0$, otherwise. Then the NPMLE $\hat{F}_n(Z_i)$ of F at Z_i is given by the left-continuous derivative at the point i of the convex minorant of the function H_n: $[0,n] \to \mathbb{R}$, defined by

$$H_n(i) := \sum_{j \le i} \Delta_j$$

at points i, and by linear interpolation at other points of [0,n] (see Barlow et al. (1972)). Moreover, we have the following result.

Theorem 4.5. *Let* t_0 *be such that* $0 < F(t_0) < 1$, *and let* F *be differentiable at* t_0, *with strictly positive derivative* $f(t_0)$. *Furthermore, let* $\hat{F}_n$ *be the* NPMLE *of* F, *based on the order statistics* $X_1,\ldots,X_n$ *of the sample, generated by the (convolution) density* g, *defined by* (4.1). *Then we have, as* $n \to \infty$,

$$n^{1/3}(\hat{F}_n(t_0) - F(t_0))/\{\tfrac{1}{2}F(t_0)(1 - F(t_0))f(t_0)\}^{1/3} \xrightarrow{\mathcal{D}} 2.Z, \tag{4.4}$$

where $\xrightarrow{\mathcal{D}}$ *denotes convergence in distribution, and* Z *is the last time that two-sided Brownian motion minus the parabola* $y(t) = t^2$ *reaches its maximum.* □

The proof of theorem 4.5 proceeds along the lines of the proof of theorem 1.1 in Groeneboom (1987) and is omitted here. The next three pictures show the NPMLE for the three distribution functions, $F(x) = x$, $F(x) = x^2$, and $F(x) = \sqrt{x}$, $0 \le x \le 1$, and simulated samples of size 1000, generated using the uniform random number generator from the IMSL library.

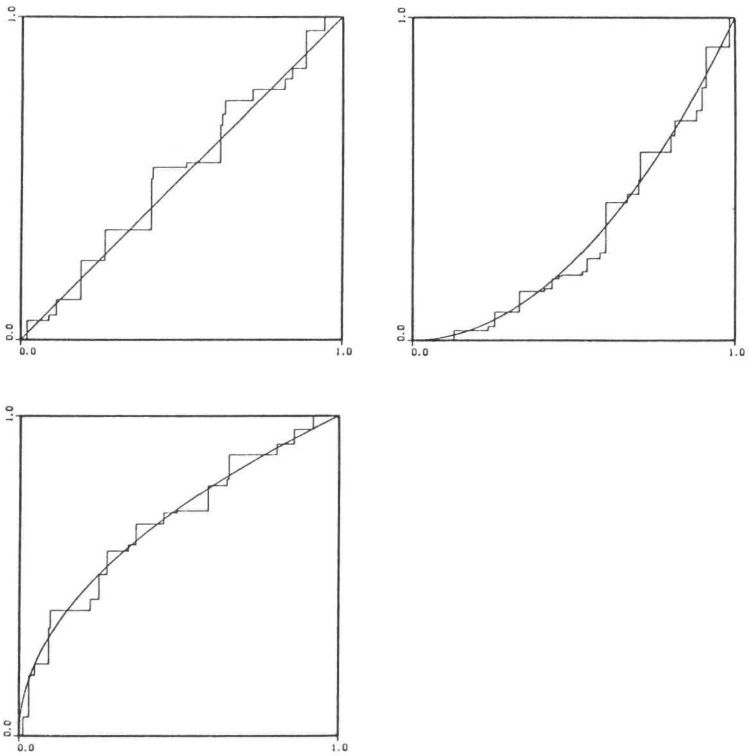

Figure 4.1. The NPMLE computed for samples of size 1000.

The next two examples are of the same type as the previous one. However, the algorithm for computing the NPMLE is much more complicated. For more details see Van Es & Groeneboom (1988).

Example 4.6. Let k be the probability density defined by

$$k(x) := \begin{cases} 2(1 - x), & x \in (0,1), \\ 0, & \text{elsewhere,} \end{cases}$$

and let $X_1,\dots,X_n$ be an ordered sample, generated by the density g, defined by

$$g(x) := \int_{-\infty}^{\infty} k(x - y)dF(y), \ x \in \mathbb{R}.$$

Notice that the density k satisfies condition (A), so in this case part (a) of theorem 4.1 holds and we have a minimax lower bound of order $n^{-1/3}$. The specific choice for the function 2(1-x) is not essential. The same method can be used for other decreasing functions as well.

The NPMLE is a discrete distribution function, with masses α_i at the points X_i, where the α_i maximize the function:

$$\text{(4.2)} \qquad \sum_{i=2}^{n} \log\Big(\sum_{j=1}^{i-1} k(X_i - X_j)\alpha_j\Big),$$

under the restrictions $\sum_{i=1}^{n} \alpha_i = 1$, $\alpha_i \geq 0$, $1 \leq i \leq n\text{-}1$ and $\alpha_n = 0$. We can write (4.2) in the form

$$\sum_{i=2}^{n} \log\Big(\sum_{j=1}^{i-1} w_{ij}\alpha_j\Big),$$

where $w_{ij} = k(X_i - X_j)$. Letting $\alpha = (\alpha_1,\dots,\alpha_n)$, for k = 1, ..., n-1 we define the derivative with respect to α_k by

$$\text{(4.3)} \qquad d_k(\alpha) := \sum_{i=k+1}^{n} w_{ik} \Big/ \Big(\sum_{j=1}^{i-1} w_{ij}\alpha_j\Big).$$

A maximum can be found using the gradient projection algorithm (Luenberger (1973)), an algorithm for maximizing a concave function subject to a number of linear constraints.

Example 4.7 (The Wicksell problem). Let $X_1 \leq ... \leq X_n$ be the order statistics of a sample of squared radii of sections of spheres. We assume that the support of the distribution of the radii of the spheres is a finite interval, which we take to be [0,1]. For a review of this estimation problem and related problems we refer to Stoyan, Kendall and Mecke (1987). As observed by Hall and Smith (1988), the distribution function of the squared section radii can be written as a convolution of the unknown distribution function of the squared sphere radii with a known function. Therefore a technique similar to the one in the previous example can be used.

The log likelihood $L(X_1,...,X_n)$ of the sample can be written in the following form:

$$L(X_1,...,X_n) = \sum_{i=1}^{n} \log \left\{\frac{1}{\gamma}\int_{(X_i,1]} \frac{1}{\sqrt{x - X_i}}\, dF(x)\right\}, \tag{4.4}$$

where F is the distribution function of the squared sphere radii, and γ is given by

$$\gamma = \int_0^1 dx \int_{(x,1]} \frac{1}{\sqrt{y - x}}\, dF(y).$$

The nonparametric maximum likelihood estimator of F is a discrete distribution function, with mass at the points $X_2,...,X_n$. So we can write:

$$L(X_1,...,X_n) = \sum_{i=1}^{n-1} \log \left\{\frac{1}{\gamma} \sum_{j>i} \frac{\alpha_j}{\sqrt{X_j - X_i}}\right\},$$

where $\alpha_2,...,\alpha_n$ are the masses of F at the points $X_2,...,X_n$, and where γ can be written

$$\gamma = \sum_{i=1}^{n} \sum_{j=i}^{n} 2\alpha_j\{\sqrt{X_j - X_{i-1}} - \sqrt{X_j - X_i}\} = 2\sum_{j=2}^{n} \alpha_j\sqrt{X_j},$$

defining $X_0 = 0$ and $\alpha_1=0$. Note that this example does not exactly fit into our previous set-up for two reasons: we look at the convolution with a function which is not a probability density and we have the extra parameter γ. It is possible to reformulate the problem in such a way that we would deal with the convolution with a *probability* density (looking at the *logarithms* of the observations), but we would not get rid of the extra parameter in this way. There does not seem to be a real advantage in this reformulation, so we keep to the above statement of the maximization problem.

Since $L(X_1,...,X_n)$ is not a concave function of $(\alpha_2,...,\alpha_n)$, using the gradient pojection algorithm as in the previous example we might find local maxima. However, for fixed γ the log likelihood $L(X_1,...,X_n)$ is concave. So by the gradient projection algorithm we can maximize $L(X_1,...,X_n)$ subject to $\alpha_i \geq 0$, for i = 2,...,n, and the *two* linear constraints

$$\sum_{i=2}^{n} \alpha_i = 1 \quad\text{and}\quad 2\sum_{i=2}^{n} \alpha_i\sqrt{X_i} = \gamma.$$

Next we can vary γ to find values of $\alpha_2, ..., \alpha_n$ and a corresponding γ which maximize the log likelihood $L(X_1,...,X_n)$. Notice that this procedure also yields a maximum likelihood estimate of γ. Theory for this NPMLE seems to be absent, but should be related to the theory for the estimator of example 4.5. In fact, because of the peakedness of the weight function $1 / \sqrt{x - X_i}$ in (4.4), we expect a faster rate of convergence of the NPMLE.

To illustrate this procedure we have simulated three samples of circle radii of size 100, for F equal to the three distribution functions in example 4.5. Since the computation for a fixed γ is already timeconsuming we have only computed the maximizing F for the three true values of γ, and for three estimated values of γ. We have used the estimator

$$\hat{\gamma}_n := \pi n / \sum_{i=1}^{n} \frac{1}{\sqrt{X_i}} ,$$

which is based on an estimator of $\mu = \gamma/2$ in example 1.2 (see Hall & Smith (1988)). The log likelihoods of the estimates are given in a table following the next figures.

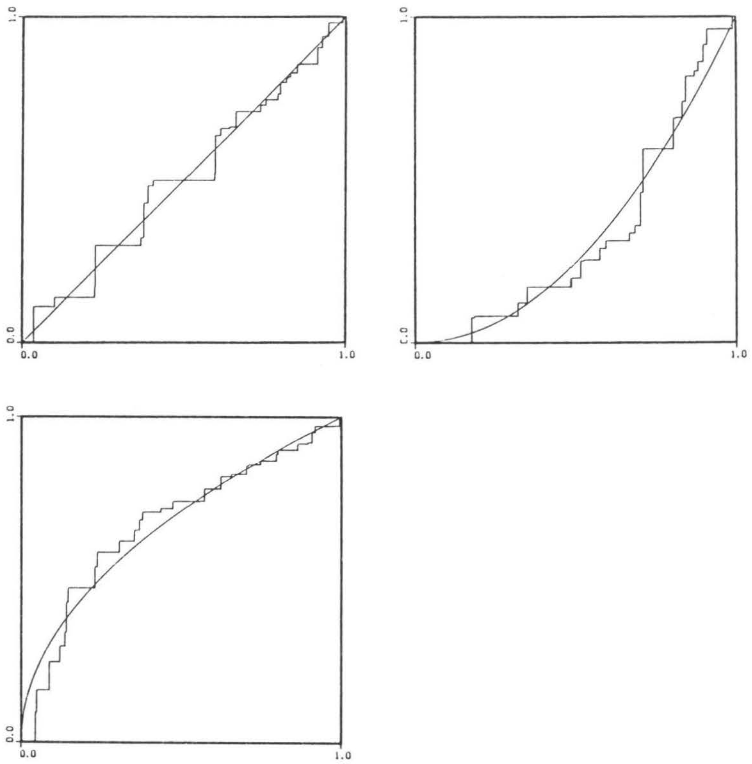

Figure 4.2. The maximizing F for the true values of γ for three samples of size 100.

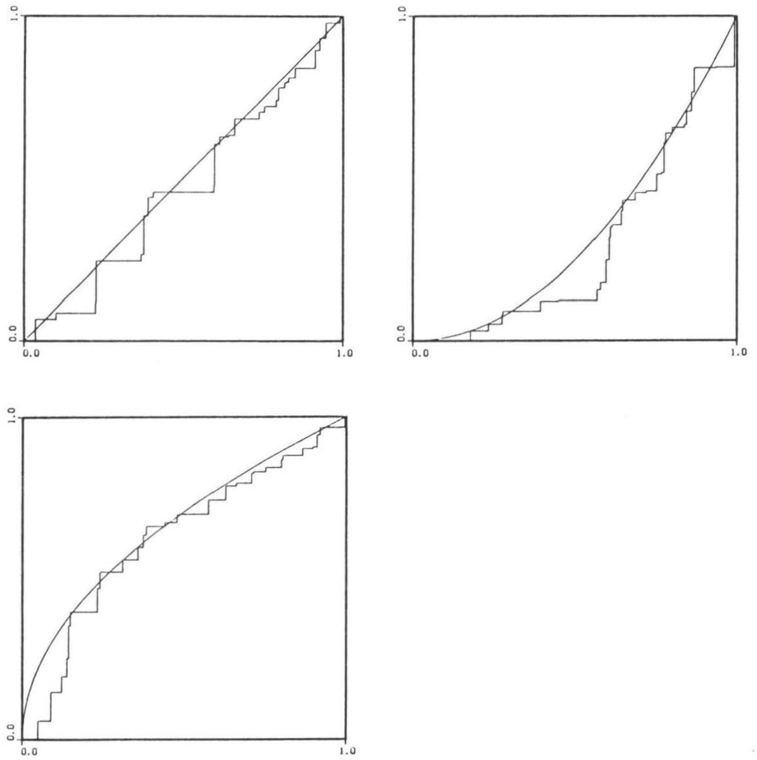

Figure 4.3. The maximizing F for the estimated values of γ for the same samples as in figure 4.3.

The next table gives the values of the true and estimated values of γ and the log likelihoods of the corresponding estimates of F.

F	γ	$\hat{\gamma}_n$	log lik. with true γ	log lik. with est. γ
x	1	1.07	29.71	29.57
x^2	4/3	1.41	21.95	20.57
$\sqrt{x}$	2/3	0.75	36.29	37.32

Roughly speaking the estimates in figure 4.2 for samples of size 100 have about the same error as the estimates in figure 4.1, which were based on samples of size 1000. This suggests that the NPMLE in the Wicksell problem indeed has a faster rate than $n^{-1/3}$, and perhaps even a rate close to $n^{-1/2}$. Because of limited computing time we have not been able to compute the true NPMLE, i.e. the estimate with a value of γ which maximizes $L(X_1,...,X_n)$. However we expect that for this value of γ the estimates would have a better fit than the estimates for the estimated γ in figure 4.3.

4.2. Proofs.

Proof of theorem 4.1. We use the same arguments as in the alternative proof of theorem 1.2 by Assouad's lemma in Groeneboom (1987). Notice that for n large enough we have

$$|F_n(x_0;\theta) - F_n(x_0;0)| \;=\; \int_{x_0-c\gamma_n}^{x_0} \theta f(x_0)du = \theta c\gamma_n f(x_0).$$

A combination of Assouad's lemma (Le Cam (1986), p.524) and Le Cam's inequality (Le Cam (1973)) now gives

$$\max_{\theta\in\{0,\delta\}} E_\theta\, |U_n - F_n(x_0;\theta)| \geq$$

$$\tfrac{1}{2}\delta c\gamma_n f(x_0)\|P_0^n \wedge P_\delta^n\| \geq$$

$$\tfrac{1}{4}\delta c\gamma_n f(x_0)\{1 - H^2(P_0^n,P_\delta^n)\} =$$

$$\tfrac{1}{4}\delta c\gamma_n f(x_0)\{1 - nH^2(P_0,P_\delta)\},$$

where P_θ is the measure corresponding to $(K{*}F_n)(.;\theta)$.

We proceed with examining the term $nH^2(P_0,P_\delta)$. By the definition of the Hellinger distance we have

$$nH^2(P_0,P_\delta) = n\,\tfrac{1}{2}\int_{-\infty}^{\infty} ((k{*}F)^{1/2}(x) - (k{*}F_n(.;\delta))^{1/2}(x))^2 dx.$$

The convolution of k and $F_n(.;\delta)$ can be written as

$$\int_{-\infty}^{\infty} k(x-y)dF(y) + \int_{-\infty}^{\infty} k(x-y)\delta f(x_0)h_n(y)dy =$$

$$(k{*}F)(x) + \delta f(x_0)\int_{-\infty}^{\infty} k(x-y)h_n(y)dy =$$

$$(k{*}F)(x) + \delta f(x_0)D_2^K(x-x_0,c\gamma_n),$$

where D_2^K denotes the second difference of the function K, i.e.

$$D_2^K(x,z) := K(x-z) - 2K(x) + K(x+z).$$

Since both K and F have bounded support the integrals are actually over a bounded area. By a Taylor expansion argument we then get for $\gamma_n \rightarrow 0$

$$nH^2(P_0,P_\delta) \sim \tfrac{1}{2} n\, \delta^2 f(x_0)^2 \int_{-\infty}^{\infty} \Big(\frac{D_2^K(x-x_0,c\gamma_n)}{2\sqrt{(k*F)(x)}}\Big)^2 dx = \tfrac{1}{8} n\, \delta^2 f(x_0)^2 \int_{-\infty}^{\infty} \frac{D_2^K(x-x_0,c\gamma_n)^2}{(k*F)(x)} dx.$$

To deal with the second difference D_2^K in case (a) we need the following lemma.

Lemma 4.8. *For any point* $a\in\mathbb{R}$ *we have*

$$\int_{a-z}^{a+z} D_2^K(x,z)^2 dx = \tfrac{2}{3} z^3 (k(a+) - k(a-))^2 + o(z^3),\ z\downarrow 0. \qquad □$$

Proof. By Taylor expansion and a substitution $t=(x-a)/z$ we get

$$\int_{a-z}^{a+z} D_2^K(x,z)^2 dx =$$

$$\int_{a}^{a+z} (K(x-z) - 2K(x) + K(x+z))^2 dx + \int_{a-z}^{a} (K(x-z) - 2K(x) + K(x+z))^2 dx =$$

$$\int_{a}^{a+z} (K(x-z) - K(a) - 2(K(x) - K(a)) + K(x+z) - K(a))^2 dx +$$

$$\int_{a-z}^{a} (K(x-z) - K(a) - 2(K(x) - K(a)) + K(x+z) - K(a))^2 dx =$$

$$\int_{a}^{a+z} ((x-z-a)k(a-) - 2(x-a)k(a+) + (x+z-a)k(a+))^2 dx +$$

$$\int_{a-z}^{a} ((x-z-a)k(a-) - 2(x-a)k(a-) + (x+z-a)k(a+))^2 dx + o(z^3) =$$

$$2z^3(k(a+) - k(a-))^2 \int_0^1 (1-t)^2 dt + o(z^3) =$$

$$\tfrac{2}{3} z^3 (k(a+) - k(a-))^2 + o(z^3)\ ,$$

which proves the lemma. □

Notice that the leading term is only nonzero if a is a jumppoint of k. By a Taylor expansion argument we see that for $z\to 0$ we have $D_2^K(x,z) \sim z^2 k'(x)$, uniformly for x at distance z of the jump points of k. Since under our conditions the density k is bounded the convolution k*F is continuous. This gives for case (a)

$$\int_{-\infty}^{\infty} \frac{D_2^K(x-x_0,c\gamma_n)^2}{(k*F)(x)}dx \sim c^3\gamma_n^3 \frac{2}{3}\sum_{i=1}^{m} \frac{(k(a_i+) - k(a_i-))^2}{(k*F)(x_0+a_i)} \ , n\rightarrow\infty.$$

We can now finish the proof of part (a) by observing that taking γ_n equal to $n^{-1/3}$ we get

$$\sup_{\delta\in(0,1),c>0} \ \liminf_{n\rightarrow\infty} n^{1/3}\, MR(n;0,\delta) \ \geq$$

$$\sup_{c>0} \frac{1}{4} cf(x_0)\Big\{1 - \frac{1}{12} f(x_0)^2 c^3 \sum_{i=1}^{m} \frac{(k(a_i+) - k(a_i-))^2}{(k*F)(x_0+a_i)}\Big\} \ =$$

$$\frac{3^{4/3}}{16} f(x_0)^{1/3} \Big(\sum_{i=1}^{m} \frac{(k(a_i+) - k(a_i-))^2}{(k*F)(x_0+a_i)}\Big)^{-1/3}.$$

For the last equality observe that the supremum is attained for c equal to

$$\Big(3/\Big(f(x_0)^2 \sum_{i=1}^{m} \frac{(k(a_i+) - k(a_i-))^2}{(k*F)(x_0+a_i)}\Big)\Big)^{1/3}.$$

For part (b) of the theorem the argument is similar, except that now we have

$$\int_{-\infty}^{\infty} \frac{D_2^K(x-x_0,c\gamma_n)^2}{(k*F)(x)}dx \sim c^4\gamma_n^4 \int_{-\infty}^{\infty} \frac{k'(x-x_0)^2}{(k*F)(x)}dx \ , n\rightarrow\infty.$$

The term γ_n^4 causes the $n^{-1/4}$ lower bound. □

Proof of theorem 4.3. Just as in the previous proof we use Assouad's lemma and Le Cam's inequality. Notice that for all $\theta>0$ we have $|F(x_0;\theta) - F(x_0;0)|=1$, so in this case we get

$$\max_{\theta\in\{0,\theta_n\}} E_\theta\, |U_n - F(x_0;\theta)| \geq \frac{1}{2} \|P_0^n \wedge P_{\theta_n}^n\| \geq \frac{1}{4}\{1 - nH^2(P_0,P_{\theta_n})\}.$$

Since the densities of the measures P_0 and P_{θ_n} are given by $k(x-x_0)$ and $k(x-x_0-\theta_n)$ respectively, we get

$$nH^2(P_0,P_{\theta_n}) = n\,\frac{1}{2}\int_{-\infty}^{\infty} (k^{1/2}(x-x_0) - k^{1/2}(x-x_0-\theta_n))^2 dx,$$

which converges to zero if θ_n decreases to zero fast enough and the proof is completed. □

Proof of theorem 4.4. Consider the family of densities $\mathcal{G}$ defined by

$$\mathcal{G} := k*\mathcal{F} := \{g : g(x) = \int_{-\infty}^{\infty} k(x-y)f(y)dy,\ f\in\mathcal{F}\}.$$

Such a family of densities is called a convolution family. Notice that, contrary to Devroye (1987, section 5.8), who considers convolutions of k with an arbitrary measure, we only allow convolutions of k with another density f. A minor adaptation of the proof of theorem 5.6 in Devroye (1987) gives the following minimax bound for estimating members of $\mathcal{G}$,

$$\inf_{g_n} \sup_{g\in\mathcal{G}} E_g \int_{-\infty}^{\infty} |g_n(x) - g(x)|dx \geq 1.$$

Next notice that if $g=k*f$ and $g_n=k*f_n$ then by $g_n-g=k*(f_n-f)$ and Young's inequality we have

$$\int_{-\infty}^{\infty} |g_n(x) - g(x)|dx \leq \int_{-\infty}^{\infty} |k(x)|dx \int_{-\infty}^{\infty} |f_n(x) - f(x)|dx = \int_{-\infty}^{\infty} |f_n(x) - f(x)|dx,$$

i.e. convolution with a probabilty density is a contraction operator for the L_1 norm. This gives

$$\inf_{f_n} \sup_{f\in\mathcal{F}} E_f \int_{-\infty}^{\infty} |f_n(x) - f(x)|dx \geq$$

$$\inf_{f_n} \sup_{f\in\mathcal{F}} E_f \int_{-\infty}^{\infty} |(k*f_n)(x) - (k*f)(x)|dx \geq$$

$$\inf_{g_n} \sup_{g\in\mathcal{G}} E_f \int_{-\infty}^{\infty} |g_n(x) - g(x)|dx \geq 1,$$

which shows that the minimax bound also holds for estimation of f. □

APPENDIX A. EXPONENTIAL BOUNDS.

In our proofs frequently we need an almost sure order bound for the supremum of some stochastic process. A standard way to derive such bounds is to consider finite subsets of the set where the supremum is taken over, and to derive a bound for the supremum over these finite subsets first. This is then usually followed by an argument showing that the difference between the supremum over the finite sets and the supremum over the original sets is asymptotically negligible. A useful tool to derive a bound for the supremum over a finite set is the next exponential inequality attributed to S.N.Bernstein. See Serfling (1980) who for the proof refers to Uspenski (1937). We omit the proof here.

Lemma A.1. *Let* $Y_1,\dots,Y_n$ *be independent random variables satisfying* $P(|Y_i - EY_i| \leq m) = 1$, *for each* i, *where* $m<\infty$. *Then for* $t>0$ *we have*

$$(A.1)\qquad P\Big(\ \Big|\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} EY_i\Big| \geq nt\ \Big) \leq 2\exp\Big(-n^2t^2/(2\textstyle\sum_{i=1}^{n}\mathrm{var}(Y_i) + \frac{2}{3}mnt)\Big),$$

for $n = 1,2,\dots$.

If we impose the extra condition that the random variables are identically distributed then the bound becomes

$$(A.2)\qquad P\Big(\ \Big|\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} EY_i\Big| \geq nt\ \Big) \leq 2\exp\Big(-nt^2/(2\mathrm{var}(Y_1) + \tfrac{2}{3}mt)\Big),$$

which gives the next bound in the even more special case that Y_i is binomial (1,p), and $\Sigma_{i=1}^{n}Y_i$ is consequently binomial (n,p), distributed,

$$(A.3)\qquad P\Big(\ \Big|\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} EY_i\Big| \geq nt\ \Big) \leq 2\exp\Big(-\tfrac{1}{2}nt^2/(p + t)\Big).$$

Recall that the kernel estimator is a sum of i.i.d. random variables. Lemma A.1 then gives us the next exponential bound, which is a minor adaptation of lemma 5 in chapter 6 of Devroye and Györfi (1985). We prove this bound for bounded measurable kernel functions K, so we don't require that K is a density function.

Theorem A.2. *Let* K *be a bounded measurable function then for arbitrary* $t>0$ *and* $h>0$ *we have for any point* x *on the real line*

$$(A.4)\qquad P(\ |f_{nh}(x) - Ef_{nh}(x)| \geq t\ ) \leq 2\exp\Big(-nht^2/(2K^*(Eh^{-1}|K((x-X_1)/h)| + t))\Big).$$

Here K *is bounded by* K^*, *i.e.* $|K(x)|\leq K^*$, *for all x.*

Proof. First we estimate the variance of $h^{-1}K((x-X_i)/h)$. We have

$$\mathrm{var}(h^{-1}K((x-X_i)/h)) \leq E(h^{-1}K((x-X_i)/h))^2 \leq$$

$$h^{-1}K^*Eh^{-1}|K((x-X_1)/h)|.$$

A direct application of lemma A.1 gives,

$$P(\ |f_{nh}(x) - Ef_{nh}(x)| \geq t\ ) =$$
$$P\Big(\ \Big|\sum_{i=1}^{n}\frac{1}{h}K((x-X_i)/h) - \sum_{i=1}^{n}E\frac{1}{h}K((x-X_i)/h)\Big| \geq nt\ \Big) \leq$$
$$2\exp\Big(-nt^2/(2h^{-1}K^*Eh^{-1}|K((x-X_1)/h)| + \tfrac{2}{3}h^{-1}K^*t)\Big) \leq$$
$$2\exp\Big(-nht^2/(2K^*(Eh^{-1}|K((x-X_1)/h)| + t))\Big),$$

which proves the theorem. □

APPENDIX B. MOMENT BOUNDS.

Investigating the performance of kernel estimators and cross-validation techniques the following type of statistic is often encountered. If $X_1,...,X_n$ is an i.i.d. sample from a distribution with a bounded density f then for h>0 we consider statistics $T_n(h)$,

$$\text{(B.1)} \qquad T_n(h) := \sum_{i \neq j} G\Big(\frac{X_i-X_j}{h}\Big)\, w(X_i,X_j) + \sum_{i=1}^{n} g_n(X_i).$$

Here G, w and g_n are bounded measurable functions for which we additionally require that G is symmetric around zero, that G is integrable and that w is symmetric in its two arguments. The first term of $T_n(h)$, $G_n(h)$ say, is a U-statistic of degree two. We have

$$G_n(h) = \sum_{i \neq j} \varphi_h(X_i,X_j),$$

with

$$\text{(B.2)} \qquad \varphi_h(x,y) := G\Big(\frac{x-y}{h}\Big)\, w(x,y).$$

$T_n(h)$ is the sum of a U-statistic of degree two and a sum of i.i.d. random variables. Examples of these statistics are $(nh)^2U_n(h)$ and $(nh)^3V_n(h)$, where $U_n(h)$ and $V_n(h)$ are defined in proposition 3.9. In the notation of chapter 3 we have

$$(nh)^2U_n(h) = -\sum_{i \neq j} L\Big(\frac{X_i-X_j}{h}\Big)(f(X_i)^{-1}I_E(X_i) + f(X_j)^{-1}I_E(X_j)).$$

and

$$(nh)^3V_n(h) = \sum_{i \neq j} KL\Big(\frac{X_i-X_j}{h}\Big)\tfrac{1}{2}(f(X_i)^{-2}I_E(X_i) + f(X_j)^{-2}I_E(X_j)).$$

In chapter 3 we need bounds on the moments of terms in the Hoeffding decomposition of these statistics. First consider the moments of the statistic $G_n(h)$. Writing $G_{ij}(h) := \varphi_h(X_i,X_j)$ we have

$$G_n(h) = \sum_{i \neq j} G_{ij}(h).$$

For any positive integer k we compute the k-th absolute moment of $G_n(h)$,

$$E\,|G_n(h)|^k = E\,\Big|\sum_{i \neq j} G_{ij}(h)\Big|^k =$$

$$\text{(B.3)} \qquad \sum_{(i_1,j_1)\in C_n,...,(i_k,j_k)\in C_n} E\,|G_{i_1j_1}(h)...\; G_{i_kj_k}(h)|,$$

where C_n denotes the set $\{(i,j) : i=1,...,n,\ j=1,...n,\ i\neq j\}$. Each of the terms $E\,|G_{i_1j_1}(h)...\; G_{i_kj_k}(h)|$ can be represented as a graph Γ with vertices corresponding to to the indices 1,2,...,n and with an undirected edge between two vertices i and j, i≠j, for each time the term $G_{ij}(h)$ appears in the product $G_{i_1j_1}(h)...\; G_{i_kj_k}(h)$. Let e_{ij} denote the number of edges between the vertices i and j and let $\nu(\Gamma)$ denote the number of vertices reached by at least one edge, which is equal to the number of different indices in $i_1,j_1,...,i_k,j_k$. For example the term $E\,|G_{12}(h)G_{23}(h)G_{24}(h)^3G_{56}(h)^2|$ is represented by the graph

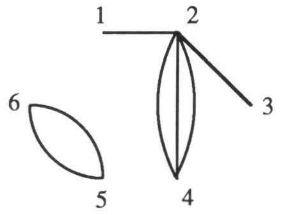

Figure B.1. The graph corresponding to E $|G_{12}(h)G_{23}(h)G_{24}(h)^3G_{56}(h)^2|$.

We need some notions from graph theory which can be found for instance in Wilson (1975). A graph Γ is called *connected* if, going through consecutive edges, each vertex of Γ can be reached from any other vertex. If Γ is not connected then Γ is the union of finitely many disjoint connected subgraphs called the *components* of Γ. Let $\gamma(\Gamma)$ denote the number of such components. If for each pair of vertices of a graph there exists one and only one way to reach one vertex from the other then such a graph is called a *tree*. If Γ is an arbitrary connected graph and if Γ' is a subgraph of Γ with the same vertices, such that Γ' is a tree, then Γ' is called a *spanning tree* of Γ. The number of edges of any spanning tree of Γ is equal to $v(\Gamma)$-1.

Now consider E $|G_{i_1j_1}(h)... G_{i_kj_k}(h)|$ more closely. This expectation can be written as

$$\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} |G(\frac{x_{i_1}-x_{j_1}}{h})w(x_{i_1},x_{j_1})...G(\frac{x_{i_k}-x_{j_k}}{h})w(x_{i_k},x_{j_k})|\ f(x_1)...f(x_n)\ dx_1...dx_n.$$

If Γ is connected we have

$$\text{(B.4)}\qquad E\ |G_{i_1j_1}(h)... G_{i_kj_k}(h)| \le ch^{v(\Gamma)-1},$$

for some constant c>0 not depending on h. The fact that this inequality holds can be seen as follows. Let Γ' denote a spanning tree of Γ. We can rewrite the integral above by performing a series of substitutions which correspond to consecutive edges of Γ'. Each of these substitutions yields a factor h and the final integral is bounded because all integer powers of $|G|$ are integrable and because w and f are bounded. The argument is completed by the observation that the number of edges of any spanning tree is equal to $v(\Gamma)$-1. If Γ is not connected then it has $\gamma(\Gamma)>1$ disjoint connected components $C_1,...,C_{\gamma(\Gamma)}$. For each of these components the bound (B.4) holds. By the independence of the X's the expectation E $|G_{i_1j_1}(h)... G_{i_kj_k}(h)|$ is equal to a product of expectations, each concerning terms of one component only, so we have for general Γ,

$$\text{(B.5)}\qquad E\ |G_{i_1j_1}(h)... G_{i_kj_k}(h)| \le c^{\gamma(\Gamma)}h^{(v(C_1)-1)\ +...+\ (v(C_{\gamma(\Gamma)})-1)} = c^{\gamma(\Gamma)}h^{v(\Gamma)-\gamma(\Gamma)},$$

which gives the next result.

Lemma B.1. *Let* $X_1,...,X_n$ *denote a sample from a distribution with a bounded density. Under the conditions imposed on the functions* G *and* w *we have for any positive integer* k *and for any positive* h

$$\text{(B.6)} \qquad E\,|G_{i_1j_1}(h)...\,G_{i_kj_k}(h)| \le ch^{\nu(\Gamma)-\gamma(\Gamma)},$$

for some constant $c>0$ *independent of* h, *where* Γ *is the graph corresponding to the indices considered.* □

Next we decompose the statistic $G_n(h)$ by Hoeffding's projection method (Hoeffding (1984), Serfling (1980)). Writing the conditional expectations of $G_{ij}(h)$ as

$$E(G_{ij}(h)|X_i) = g_h^c(X_i),$$

$$E(G_{ij}(h)|X_j) = g_h^c(X_j),$$

with

$$g_h^c(x) = \int_{-\infty}^{\infty} \varphi_h(x,y)f(y)dy,$$

we define $\hat{G}_{ij}(h)$ and $\hat{G}_n(h)$ by

$$\hat{G}_{ij}(h) = G_{ij}(h) - g_h^c(X_i) - g_h^c(X_j) + EG_{ij}(h)$$

and

$$\hat{G}_n(h) := \sum_{i\neq j} \hat{G}_{ij}(h).$$

This gives the next decomposition of $G_n(h)$,

$$\text{(B.7)} \qquad G_n(h) = \hat{G}_n(h) + 2(n-1)\sum_{i=1}^{n} g_h^c(X_i) - EG_n(h).$$

Since $E(\hat{G}_{ij}(h)|X_i) = E(\hat{G}_{ij}(h)|X_j) = 0$ it follows that the terms are uncorrelated. Plugging (B.7) into (B.1) we get a similar decomposition for $T_n(h)$,

$$\text{(B.8)} \qquad T_n(h) = \hat{G}_n(h) + \sum_{i=1}^{n} (2(n-1)g_h^c(X_i) + g_n(X_i)) - EG_n(h).$$

Notice that the terms of this decomposition are also uncorrelated. It turns out that $\hat{G}_{ij}(h)$ also satisfies (B.6).

Lemma B.2. *Let* $X_1,...,X_n$ *denote a sample from a distribution with a bounded density. Under the conditions imposed on the functions* G *and* w *we have for any positive integer* k *and for any* $0<h<1$,

$$\text{(B.9)} \qquad E\,|\hat{G}_{i_1j_1}(h)...\,\hat{G}_{i_kj_k}(h)| \le ch^{\nu(\Gamma)-\gamma(\Gamma)},$$

for some constant $c>0$ *independent of* h, *where* Γ *is the graph corresponding to the indices considered.* □

Proof. By a simple substitution v = (y-x)/h we get

$$g_h^c(x) = \int_{-\infty}^{\infty} \varphi_h(x,y)f(y)dy =$$

$$\int_{-\infty}^{\infty} G\left(\frac{x-y}{h}\right)w(x,y)f(y)dy =$$

$$h\int_{-\infty}^{\infty} G(v)w(x,x+hv)f(x+hv)dv,$$

and by repeated integration

$$EG_{ij}(h) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi_h(x,y)f(x)f(y)dxdy =$$

$$\int_{-\infty}^{\infty} g_h^c(x)f(x)dx.$$

It follows that for some constant $\alpha>0$ we have

(B.10) $\quad |g_h^c(x)| \le \alpha h$, for all x,

and consequently

(B.11) $\quad |EG_{ij}(h)| \le \alpha h$.

Recall that $\hat{G}_{ij}(h)$ is equal to $Gij(h) - g_h^c(X_i) - g_h^c(X_j) + EG_{ij}(h)$. Returning to $E\,|\hat{G}_{i_1j_1}(h)\ldots\ \hat{G}_{i_kj_k}(h)|$ we see that this expectation is equal to the sum of 4^k terms of the form $E\,|\beta_{i_1j_1}(h)\ldots\ \beta_{i_kj_k}(h)|$ where $\beta_{ij}(h)$ equals either $G_{ij}(h)$, $g_h^c(X_i)$, $g_h^c(X_j)$ or $EG_{ij}(h)$. The proof is now completed by the same spanning tree argument as above for each of the terms $E\,|\beta_{i_1j_1}(h)\ldots\ \beta_{i_kj_k}(h)|$, with this exception that each edge of the spanning tree, between i and j say, now corresponds to a term

$$G_{ij}(h)^{e_{ij1}} g_h^c(X_i)^{e_{ij2}} g_h^c(X_j)^{e_{ij3}} EG_{ij}(h)^{e_{ij4}},$$

where $e_{ij1},\ldots,e_{ij4}$ are nonegative integers with $e_{ij1}+\ldots+e_{ij4} = e_{ij}$. If $e_{ij1}=e_{ij}$ this term yields a factor h by substitution just as above, and by (B10) and (B.11) since $0<h<1$ it yields a factor smaller than a constant times h otherwise. □

Corollary B.3. *Let* f *be bounded density which is bounded away from zero on the set* E. *For the statistics* $\hat{U}_n(h)$, $\hat{V}_n(h)$ *and* $\hat{W}_n(h)$ *defined in proposition 3.9 we have for some constant* $c>0$ *and for* $0<h<1$

$$E\,|\hat{U}_{i_1j_1}(h)\ldots\ \hat{U}_{i_pj_p}(h)| \le ch^{m/2-2p},$$

$$E\,|\hat{V}_{i_1j_1}(h)\ldots\ \hat{V}_{i_pj_p}(h)| \le ch^{m/2-3p},$$

provided there are exactly m *different numbers in the sequence* $i_1, j_1,\ldots, i_p, j_p$, *each index appearing at least twice.*

Similarly we have

$$E\,|\hat{W}_{i_1j_1k_1}(h)\ldots\ \hat{W}_{i_pj_pk_p}(h)| \le ch^{2m/3-3p},$$

provided there are exactly m *different numbers in the sequence* $i_1, j_1, k_1, \ldots, i_p, j_p, k_p$, *each index appearing at least twice.* □

Proof. Taking G equal to the function L defined by (3.34) and w equal to

$$w(x,y) = -\,(f(x)^{-1} I_E(x) + f(y)^{-1} I_E(y))$$

we see that $\hat{U}_{ij}(h)$ equals $h^{-2}\hat{G}_{ij}(h)$. By lemma B.2 we have

$$E\,|\hat{U}_{i_1 j_1}(h) \ldots \hat{U}_{i_p j_p}(h)| =$$

$$h^{-2p}\, E\,|\hat{G}_{i_1 j_1}(h) \ldots \hat{G}_{i_p j_p}(h)| \leq$$

$$ch^{\nu(\Gamma)-\gamma(\Gamma)-2p}.$$

The conditions of the lemma imply $\nu(\Gamma)=m$ and $\gamma(\Gamma) \leq m/2$. So by $0<h<1$ the bound above is smaller than $ch^{m-m/2-2p}$. This proves the first statement. The proof of the second statement is completely analogous, except that the factor h^{-2} should be replaced by h^{-3}. We cannot use lemma B.2 to derive the third statement . However, the expectation $E\,|\hat{W}_{i_1 j_1 k_1}(h) \ldots \hat{W}_{i_p j_p k_p}(h)|$ can also be represented as a graph Γ. In this case the conditions of the lemma imply that the number of components of Γ, $\gamma(\Gamma)$, does not exceed m/3. By the same method as above we can then derive a bound $ch^{m-m/3-3p}$. □

APPENDIX C. ASYMPTOTIC NORMALITY.

We consider the asymptotic distribution of the statistics $T_n(h_n)$ defined by (B.1) for sequences of positive bandwidths (h_n) tending to zero. In that case the kernel function of the U-statistic $G_n(h_n)$ depends on the sample size and we can not use standard U-statistic theory to derive asymptotic normality of $T_n(h_n)$. Instead we use a limit theorem of Jammalamadaka and Janson (1986). An alternative approach would be to use central limit theorems for degenerate U-statistics which can be found for example in Hall (1984), De Jong (1987, 1988), Nolan & Pollard (1987, 1988).

By decomposition (B.8) we have

$$T_n(h_n) = \hat{G}_n(h) + \sum_{i=1}^{n} g_n^*(X_i) - EG_n(h), \tag{C.1)}$$

with

$$g_n^*(x) := 2(n-1)g_{h_n}^c(x) + g_n(x). \tag{C.2}$$

Since the terms in this decomposition are uncorrelated and since $E\hat{G}_n(h_n)=0$ the variance of $T_n(h_n)$ equals

$$E(\hat{G}_n(h_n))^2 + n\ \mathrm{var}(g_n^*(X_1)),$$

Next we use the fact that $E(\hat{G}_{ij}(h)|X_k)$ and $E(\hat{G}_{ij}(h)|X_k)$ are both equal to zero for $k=1,\ldots,n$. We get

$$E(\hat{G}_n(h_n))^2 = E\big(2\sum_{i<j}\hat{G}_{ij}(h_n)\big)^2 = 2n(n-1)\ E(\hat{G}_{12}(h_n))^2,$$

and

$$\mathrm{var}(T_n(h_n)) = 2n(n-1)\ E(\hat{G}_{12}(h_n))^2 + n\ \mathrm{var}(g_n^*(X_1)).$$

Assume that w and f are almost everywhere continuous. Then by (B.10), (B.11) in appendix B and the dominated convergence theorem

$$E(\hat{G}_{12}(h_n))^2 =$$

$$E(G_{12}(h_n) - g_{h_n}^c(X_i) - g_{h_n}^c(X_j) + EG_{12}(h_n))^2 \sim$$

$$E(G_{12}(h_n))^2 =$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G^2\left(\frac{x_1-x_2}{h_n}\right)w^2(x_1,x_2)f(x_1)f(x_2)dx_1dx_2 =$$

$$h_n\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G^2(v)w^2(x,x+h_nv)f(x)f(x+h_nv)dxdv \sim \tag{C.3}$$

$$h_n\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G^2(v)w^2(x,x)f^2(x)dxdv =$$

$$h_n \int_{-\infty}^{\infty} G^2(v)dv \int_{-\infty}^{\infty} w^2(x,x)f^2(x)dx \ .$$

To examine the variance of the second term we compute $E(g^c_{h_n}(X_1))^2$. By the dominated convergence theorem we have

$$E(g^c_{h_n}(X_1))^2 =$$

$$\int_{-\infty}^{\infty} \Big(h_n \int_{-\infty}^{\infty} G(v)w(x,x+h_nv)f(x+h_nv)dv\Big)^2 f(x)dx \sim$$

$$h_n^2 \int_{-\infty}^{\infty} \Big(\int_{-\infty}^{\infty} G(v)w(x,x)f(x)dv\Big)^2 f(x)dx =$$

$$h_n^2 \Big(\int_{-\infty}^{\infty} G(v)dv\Big)^2 \Big(\int_{-\infty}^{\infty} w(x,x)f^3(x)dx\Big).$$

This implies that the variance of $2(n-1)g^c_{h_n}(X_1)$ is typically of order $(nh_n)^2$. Thus if g_n is identically equal to zero and if $nh_n \to \infty$, as in our applications, then $\hat{G}_n(h_n)$ is asymptotically negligible compared to the linear term. For an example of this situation see Veraverbeke (1985). Actually in the standard U-statistic theory where the kernels are fixed functions the linear term dominates too. In our applications however g_n is not identically equal to zero. It turns out that in those cases $g_n(X_i)$ compensates the terms $2(n-1)g^c_{h_n}(X_i)$ in such a way that the variances of both terms in (C.1) are of the same order, or that the variance of the second term is even of smaller order than the first term. We use a theorem of Jammalamadaka and Janson (1986) to prove the next theorem which establishes asymptotic normality of $T_n(h_n)$ in the case that $\hat{G}_n(h_n)$ is not asymptotically negligible.

Theorem C.1. *Let* f *be a bounded almost everywhere continuous density and let the functions* G, w *and* g_n *also be bounded. Further assume that* G *is symmetric and integrable, that* w *is symmetric in its two arguments and that* w *is almost everywhere continuous. Let the statistic* $T_n(h)$ *be defined by (B.1) and let* (h_n) *be a sequence of positive bandwidths converging to zero such that* $nh_n \to \infty$. *Let the function* g_n^* *be defined by (C.2) and suppose that this function satisfies*

(i) $(nh_n^{1/2})^{-1} \sup_x |g_n^*(x) - Eg_n^*(X_1)| \to 0,$

(ii) $(nh_n)^{-1} \mathrm{var}(g_n^*(X_1)) \to \alpha^2, \ 0 \leq \alpha^2 \leq \infty.$

Then

$$\text{(C.4)} \qquad \frac{1}{nh_n^{1/2}}(T_n(h_n) - ET_n(h_n)) \xrightarrow{\mathcal{D}} N(0, 2\sigma^2+\alpha^2),$$

with

(C.5) $$\sigma^2 := \int_{-\infty}^{\infty} G^2(v)^2 dv \int_{-\infty}^{\infty} w^2(x,x) f^2(x)dx.$$ □

Proof. To apply theorem 2.2 of Jammalamadaka and Janson (1986) we rewrite and renormalize $T_n(h_n)$ as follows,

$$T^*_n(h_n) := \frac{1}{nh_n^{1/2}}(T_n(h_n) - ET_n(h_n)) =$$

$$\sum_{i<j} 2\frac{1}{nh_n^{1/2}}\hat{G}_{ij}(h_n) + \sum_{i=1}^{n} \frac{1}{nh_n^{1/2}}(g^*_n(X_i) - Eg^*_n(X_i)) =$$

$$\sum_{i<j} 2\frac{1}{nh_n^{1/2}}\hat{\varphi}_{h_n}(X_i,X_j) + \sum_{i=1}^{n} \frac{1}{nh_n^{1/2}}(g^*_n(X_i) - Eg^*_n(X_i)),$$

with

(C.6) $$\hat{\varphi}_{h_n}(x,y) := \varphi_{h_n}(x,y) - g^c_{h_n}(x) - g^c_{h_n}(y) + E\varphi_{h_n}(X_1,X_2).$$

Suppose that we have checked the conditions. Then this theorem gives

$$\Big(\sum_{i<j} 2\frac{1}{nh_n^{1/2}}\hat{\varphi}_{h_n}(X_i,X_j)\,,\, \sum_{i=1}^{n} \frac{1}{nh_n^{1/2}}(g^*_n(X_i) - Eg^*_n(X_i))\Big) \xrightarrow{\mathcal{D}} N\left(0,0,\begin{pmatrix}2\sigma^2 & 0\\ 0 & \alpha^2\end{pmatrix}\right),$$

and consequently

$$\frac{1}{nh_n^{1/2}}(T_n(h_n) - ET_n(h_n)) = T^*_n(h_n) \xrightarrow{\mathcal{D}} N(0,2\sigma^2+\alpha^2),$$

which proves (C.4). All we have to do is to check

(i) $E\frac{1}{nh_n^{1/2}}(g^*_n(X_1) - Eg^*_n(X_1)) = E\,2\frac{1}{nh_n^{1/2}}\hat{\varphi}_{h_n}(x,X_2) = 0,$

(ii) $\sup_x \frac{1}{nh_n^{1/2}}|g^*_n(x) - Eg^*_n(X_1)| \to 0,$

(iii) $n\,E\,\big(\frac{1}{nh_n^{1/2}}(g^*_n(X_1) - Eg^*_n(X_1))\big)^2 \to \alpha^2,\ 0 \le \alpha^2 < \infty,$

(iv) $n^2\,E\,\big(\frac{1}{nh_n^{1/2}}\hat{\varphi}_{h_n}(X_1,X_2)\big)^2 \to \beta^2,\ 0 \le \beta^2 < \infty,$

(v) $\sup_{x,y} \big|\frac{1}{nh_n^{1/2}}\hat{\varphi}_{h_n}(x,y)\big| \to 0,$

(vi) $n\,\sup_x E\,\big|\frac{1}{nh_n^{1/2}}\hat{\varphi}_{h_n}(x,X_2)\big| \to 0.$

The first three conditions are clearly fulfilled by the fact that $\hat{G}_{12}(h_n) = \hat{\varphi}_{h_n}(X_1,X_2)$ has vanishing conditional expectations, and by conditions (i) and (ii) of our theorem. Condition (iv) with β^2 equal to $2\sigma^2$ follows from (C.3). In order to show (v) and (vi) notice that for n large enough we have $nh_n^{1/2} \ge nh_n \to \infty$, which together with (B.10) and (B.11) implies (v). Property (vi) follows by the same arguments as in the derivation of (B.10) and (B.11). □

This theorem is used in section 3.4 to prove the asymptotic normality of the statistic $U_n(h_n)$, thus serving as an important tool in the asymptotic distribution theory for likelihood cross-validation. Another place where it is used is in the derivation of the asymptotic distribution of the integrated squared error of kernel estimators in section 2.3.2. There the theorem can be directly applied only for $w \equiv 1$. However, for other weight functions, modifying the proof above we can also prove asymptotic normality.

Assume that the function K satisfies condition K and that w is a bounded nonnegative measurable weigth function with a bounded support. In section 2.3.2 we have shown that the integrated squared error of a kernel estimator f_{nh} can be written as

$$\mathrm{ISE}_n(h) =$$

$$\frac{1}{n^2h^2}\sum_{i\neq j}\int_{-\infty}^{\infty} K\left(\frac{u-X_i}{h}\right)K\left(\frac{u-X_j}{h}\right)w(u)du +$$

$$-\frac{2}{nh}\sum_{i=1}^{n}\int_{-\infty}^{\infty} K\left(\frac{u-X_i}{h}\right)f(u)w(u)du +$$

$$\frac{1}{n^2h^2}\sum_{i=1}^{n}\int_{-\infty}^{\infty} K^2\left(\frac{u-X_i}{h}\right)w(u)du +$$

$$\int_{-\infty}^{\infty} f^2(u)w(u)du.$$

For $w \equiv 1$ the first term equals

$$\frac{1}{n^2h}\sum_{i\neq j}\int_{-1}^{1} K(u)K\left(u+\frac{X_i-X_j}{h}\right)du.$$

The terms of this sum are symmetric functions of $(X_i-X_j)/h$ so we can directly apply the previous theorem. However, if w is not identically equal to one we get

$$\frac{1}{n^2h}\sum_{i\neq j}\int_{-1}^{1} K(u)K\left(u+\frac{X_i-X_j}{h}\right)w(X_i+hu)du,$$

which is not of the form considered above. A modification of the proof of theorem C.1 gives the next limit theorem for the integrated squared error.

Theorem C.2. *Let* f *be a bounded almost everywhere continuous density and let* w *be a bounded almost everywhere continuous weightfunction with a bounded support. Furthermore assume that the kernel* K *satisfies condition* K *in section* 2.1 *and that* (h_n) *is a sequence of nonnegative bandwidths converging to zero such that* $nh_n \to \infty$. *Let* $b(u,h)$ *denote the bias function* $Ef_{nh}(u) - f(u)$ *of the kernel* *estimator. If*

(C.7) $$4nh_n^{-1}\mathrm{var}\Big(\int_{-\infty}^{\infty} K\big(\tfrac{u-X_1}{h_n}\big)b(u,h_n)w(u)du\Big) \to \alpha^2,\ 0 \le \alpha^2 < \infty,$$

then

(C.8) $$nh_n^{1/2}\,(ISE_n(h_n) - MISE_n(h_n)) \xrightarrow{\mathcal{D}} N(0,2\sigma^2+\alpha^2),$$

with

(C.9) $$\sigma^2 := \int_{-\infty}^{\infty}\Big(\int_{-1}^{1} K(v)K(v+z)dv\Big)^2 dz \int_{-\infty}^{\infty} w^2(u)f^2(u)du.$$

Proof. We use the same notation as above. Define the statistic $T_n(h)$ by

$$T_n(h) := \sum_{i\neq j}\frac{1}{h}\int_{-\infty}^{\infty} K\big(\tfrac{u-X_i}{h}\big)K\big(\tfrac{u-X_j}{h}\big)w(u)du - 2n\sum_{i=1}^{n}\int_{-\infty}^{\infty} K\big(\tfrac{u-X_i}{h}\big)f(u)w(u)du\ .$$

We decompose this statistic using Hoeffding's projection technique. Write

$$\varphi_h(x,y) := \frac{1}{h}\int_{-\infty}^{\infty} K\big(\tfrac{u-x}{h}\big)K\big(\tfrac{u-y}{h}\big)w(u)du$$

and

$$g_h^c(x) = \int_{-\infty}^{\infty}\varphi_h(x,y)f(y)dy =$$

$$\int_{-\infty}^{\infty}\Big(\frac{1}{h}\int_{-\infty}^{\infty} K\big(\tfrac{u-x}{h}\big)K\big(\tfrac{u-y}{h}\big)w(u)du\Big)f(y)dy =$$

$$\int_{-\infty}^{\infty} K\big(\tfrac{u-x}{h}\big)\Big(\frac{1}{h}\int_{-\infty}^{\infty} K\big(\tfrac{u-y}{h}\big)f(y)dy\Big)w(u)du =$$

$$\int_{-\infty}^{\infty} K\big(\tfrac{u-x}{h}\big)f(u)w(u)du + \int_{-\infty}^{\infty} K\big(\tfrac{u-x}{h}\big)b(u,h)w(u)du.$$

We obtain the decomposition

$$T_n(h_n) = \sum_{i\neq j}\hat{\phi}_{h_n}(X_i,X_j) + \sum_{i=1}^{n} g_n^*(X_i) - n(n-1)E\varphi_{h_n}(X_1,X_2),$$

where $\hat{\phi}_h$ is defined by (C.6) and the function g_n^* is given by

$$g_n^*(x) := 2(n-1)g_{h_n}^c(x) - 2n\int_{-\infty}^{\infty} K\big(\tfrac{u-x}{h_n}\big)f(u)w(u)du =$$

$$2(n-1)\int_{-\infty}^{\infty} K\big(\tfrac{u-x}{h_n}\big)b(u,h_n)w(u)du - 2\int_{-\infty}^{\infty} K\big(\tfrac{u-x}{h_n}\big)f(u)w(u)du.$$

Using the fact that b(u,h) is bounded by a fixed constant for all real x and all positive h it is readily shown that $\hat{\phi}_{h_n}$ and g_n^* satisfy conditions (i),...,(vi) in the proof of the previous theorem. Therefore

(C.10) $$\frac{1}{nh_n^{1/2}}(T_n(h_n) - ET_n(h_n))\,) \xrightarrow{\mathcal{D}} N(0,2\sigma^2+\alpha^2).$$

Returning to the integrated squared error notice

$$\text{(C.11)}\qquad \begin{aligned} & n h_n^{1/2}(\mathrm{ISE}_n(h_n) - \mathrm{MISE}_n(h_n)) = \\ & \frac{1}{n h_n^{1/2}}(T_n(h_n) - ET_n(h_n)) + \frac{1}{n h_n^{3/2}} \sum_{i=1}^{n} \Big(\int_{-\infty}^{\infty} K^2\Big(\frac{u-X_i}{h_n}\Big) w(u)du - E \int_{-\infty}^{\infty} K^2\Big(\frac{u-X_i}{h_n}\Big) w(u)du \Big). \end{aligned}$$

The variance of the second term can be bounded as follows,

$$\mathrm{var}\Big(\frac{1}{n h_n^{3/2}} \sum_{i=1}^{n} \Big(\int_{-\infty}^{\infty} K^2\Big(\frac{u-X_i}{h_n}\Big) w(u)du - E \int_{-\infty}^{\infty} K^2\Big(\frac{u-X_i}{h_n}\Big) w(u)du \Big)\Big) \leq$$

$$\frac{1}{n^2 h_n^3}\, n\, E\Big(\int_{-\infty}^{\infty} K^2\Big(\frac{u-X_1}{h_n}\Big) w(u)du \Big)^2 =$$

$$\frac{1}{n h_n^3} \int_{-\infty}^{\infty} \Big(\int_{-\infty}^{\infty} K^2\Big(\frac{u-v}{h_n}\Big) w(u)du \Big)^2 f(v)dv =$$

$$\frac{1}{n h_n} \int_{-\infty}^{\infty} \Big(\int_{-1}^{1} K^2(w) w(v + h_n w)dw \Big)^2 f(v)dv = O\Big(\frac{1}{n h_n}\Big),$$

which shows that this term vanishes in probability. By (C.10) and (C.11) the proof is completed.

□

Remark C.3. If condition (C.7) of the previous theorem holds with α^2 equal to infinity then the linear term $\sum_{i=1}^{n} g_n^*(X_i)$ dominates over the quadratic term $\sum_{i \neq j} \hat{\phi}_{h_n}(X_i, X_j)$. Considering

$$\text{(C.12)}\qquad \sum_{i=1}^{n} (g_n^*(X_i) - Eg_n^*(X_i))/(n h_n^{1/2})$$

we recall

$$(n h_n^{1/2})^{-1} \sup_x |g_n^*(x) - Eg_n^*(X_1)| \to 0,$$

i.e. the terms of the sum (C.12) vanish uniformly in i for n tending to infinity. We also have

$$\mathrm{var}\Big(\sum_{i=1}^{n} (g_n^*(X_i) - Eg_n^*(X_i))/(n h_n^{1/2})\Big) \sim$$

$$n \frac{1}{n^2 h_n} 4\,(n-1)^2\, \mathrm{var}\Big(\int_{-\infty}^{\infty} K\Big(\frac{u-X_1}{h_n}\Big) b(u,h_n) w(u)du \Big) \sim$$

$$4 n h_n^{-1}\, \mathrm{var}\Big(\int_{-\infty}^{\infty} K\Big(\frac{u-X_1}{h_n}\Big) b(u,h_n) w(u)du \Big) \to \infty.$$

This implies asymptotic normality of the linear term by the Lindeberg Feller central limit theorem, so in case condition (C.7) is fulfilled with α^2 equal to infinity the integrated squared error is still asymptotically normal. The proof of theorem C.2 now implies

$$\text{(C.13)} \qquad \tfrac{1}{2}n^{1/2}h_n \Big(\mathrm{var}\Big(\int_{-\infty}^{\infty} K\Big(\frac{u-X_1}{h_n}\Big)b(u,h_n)w(u)du\Big)\Big)^{-1/2}(\mathrm{ISE}_n(h_n) - \mathrm{MISE}_n(h_n)) \overset{\mathcal{D}}{\rightarrow} N(0,1),$$

which gives the proper normalizing constant in this case.

REFERENCES.

Abramowitz, M. and A. Stegun (1965), *Handbook of Mathematical Functions*, Dover, New York.

Barlow, R.E., Bartholomew, D.J., Bremner, J.M. and H.D. Brunk (1972), *Statistical Inference under Order Restrictions*, Wiley, New York.

Bickel, P.J. and M. Rosenblatt (1973), On some global measures of the deviations of density function estimates, *Ann. Statist. 1*, 1071-1095.

Bowman, A.W. (1984), An alternative method of cross-validation for the smoothing of density estimates, *Biometrika 71*, 353-360.

Bowman, A.W. (1985), A comparative study of some kernel based nonparametric density estimators, *J. Statist. Comput. Simul. 21*, 131-327.

Burman, P. (1985), A data dependent approach to density estimation, *Z. Wahrsch. verw. Gebiete 69*, 609-628.

Chow, Y.S., Geman, S. & L.D. Wu. (1983), Consistent cross-validated density estimation, *Ann. Statist. 11*, 25-39.

Cline, D.B.H. and J.D. Hart. (1986), Kernel estimation of densities with discontinuities or discontinuous derivatives, *Unpublished manuscript.*

De Jong, P. (1987), A central limit theorem for generalized quadratic forms, *Probab. Th. Rel. Fields 75*, 261-277.

De Jong, P. (1988), *Central Limit Theorems for Generalized Multilinear Forms*, Dissertation, University of Amsterdam.

Devroye, L. (1982), Upper and lower class sequences for minimal uniform spacings, *Z. Wahrsch. verw. Geb. 61*, 237-254.

Devroye, L. (1987), *A Course in Density Estimation*, Birkhäuser, Boston.

Devroye, L. and L. Györfi. (1985), *Nonparametric Density Estimation, The L_1 View*, Wiley, New York.

Duin, R.P.W. (1976), On the choice of smoothing parameters for Parzen estimators of probability density functions, *IEEE Trans.Computers C 25*, 1175-1179.

Eeden Van , C. (1985), Mean integrated squared error of kernel estimators when the density and its derivatives are not necessarily continuous, *Ann. Inst. Statist. Math. 37,* Part A, 461-472.

Epanechnikov, V.A. (1969), Nonparametric estimation of a multivariate probability density, *Theory Prob. Appl. 14*, 153-158.

Es Van, A.J. and A.W. Hoogendoorn (1988), A kernel approach to estimation of the sphere radius density in Wicksell's corpuscle problem, *Report MS-R8809, Centrum voor Wiskunde en Informatica, Amsterdam.*

Es Van, A.J. and P. Groeneboom (1988), Recovering a distribution from a convolution, *in preparation.*

Groeneboom, P. (1987), Asymptotics for incomplete censored observations, *Report 87-18, University of Amsterdam.*

Habbema, J.D.F., Hermans, J.& K. Van de Broek. (1974), A stepwise discriminant analysis program using density estimation, *In Compstat 1974: Proceedings in Computational Statistics (G Bruckman ed.),* 101-110, Physica Verlag, Vienna.

Hall, P. (1982a), Cross-validation in density estimation, *Biometrika 69*, 383-390.

Hall, P. (1982b), Limit theorems for stochastic measures of the accuracy of density estimators, *Stochastic Processes and Applications 13,* 11-25.

Hall, P. (1983a), Large sample optimality of least squares cross-validation in density estimation, *Ann. Statist. 11*, 1156-1174.

Hall, P. (1983b), Asymptotic theory of minimum integrated squared error for multivariate density estimation, *Proc. Sixth Internat. Symp. Multivariate Anal. Pittsburg, 25 - 29 July 1983.*

Hall, P. (1984), Central limit theorem for integrated square error of multivariate nonparametric density estimators, *J. Multivar. Anal. 14,* 1-16.

Hall, P. (1985), Asymptotic theory of minimum integrated square error for multivariate density estimation, *Multivariate Analysis VI (P.R.Krishnaiah, ed.), 289-309*, North-Holland, Amsterdam.

Hall, P. (1987a), On the use of compactly supported density estimates in problems of discrimination, *J. Multivar. Anal. 23*, 131-159.

Hall, P. (1987b), On Kullback-Leibler loss and likelihood cross-validation, *Ann. Statist.15*, 1491-1520.

Hall, P. and J.S. Marron (1987a), Extent to which least squares cross-validation minimizes integrated square error in nonparametric density estimation, *Probab. Th. Rel. Fields 74*, 567-581.

Hall, P. and J.S. Marron (1987b), On the amount of noise inherent in bandwidth selection for a kernel density estimator, *Ann. Statist. 15*, 163-181.

Hall, P. and J.S. Marron (1988), Choice of kernel order in density estimation, *Ann. Statist.16*, 161-174.

Hall, P. and R.L. Smith (1988), The kernel method for unfolding sphere distributions, *J. Comp. Phys. 74*, 409-421.

Härdle, W. and J.S. Marron (1985), Optimal bandwidth selection in nonparametric regression estimation, *Ann. Statist. 13*, 1465-1481.

Hoeffding, W. (1948), A class of statistics with asymptotically normal distribution, *Ann. Math. Stat. 19*, 293-325.

Jammalamadaka, R.S. and S. Janson (1986), Limit theorems for a triangular scheme of U-statistics with applications to interpoint distances, *Ann. Probab. 14*, 1347-1358.

Kolcinskii, V. I. (1980), Some limit theorems for empirical measures, *Theor. Probability and Math. Statist. 21*, 79-86.

Le Cam, L.M. (1973), Convergence of estimates under dimensionality restrictions, *Ann. Statist. 1*, 38-53.

Le Cam, L.M. (1986), *Asymptotic Methods in Statistical Decision Theory*, Springer, Berlin.

Luenberger, D.G. (1973), *Introduction to Linear and Nonlinear Programming,* Addison-Wesley, London.

Marron, J.S. (1985), An asymptotically efficient solution to the bandwidth problem of kernel density estimation, *Ann. Statist. 13*, 1011-1023.

Marron, J.S. (1987), A comparison of cross-validation techniques in density estimation, *Ann. Statist. 15*, 152-162.

Marron, J.S. and W. Härdle (1986), Random approximations to some measures of accuracy in nonparametric curve estimation, *J. Multivariate Anal. 20*, 91-113.

Nadaraya, E. A. (1965), On nonparametric estimates of density functions and regression curves, *Theor. Probability Appl. 19*, 186-190.

Nolan, D. and D. Pollard (1987), U-processes: Rates of convergence, *Ann. Statist. 15*, 780-800.

Nolan, D. and D. Pollard (1988), Functional limit theorems for U-processes, *to appear in Ann. Probab.*.

Parzen, E. (1962), On the estimation of a probability density function and the mode, *Ann. Math. Stat. 33*, 1065-1076.

Prakasa Rao, B.L S. (1983), *Nonparametric Functional Estimation*, Academic Press, London.

Révész, P. (1978), A strong law for the empirical density function, *Periodica Math. Hung. 9*, 317-324.

Ripley, B.D. (1981), *Spatial Statistics*, Wiley, New York.

Rosenblatt, M. (1956), Remarks on some nonparametric estimates of a density function, *Ann. Math. Stat. 27*, 832-837.

Rudemo, M. (1982), Empirical choice of histogram and kernel density estimators, *Scand. J. Statist. 9*, 65-78.

Schuster, E. F. (1985), Incorporating support constraints into nonparametric estimators of densities, *Commun. Statist. - Theor. Meth. 14*, 1123-1136.

Schuster, E.F.and G.G. Gregory (1981), On the nonconsistency of maximum likelihood nonparametric density estimators, *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface. (W.F. Eddy, ed.),* 295-298, Springer Verlag, New York.

Scott, D.W. (1985), Averaged shifted histograms: Effective nonparametric density estimators in several dimensions, *Ann. Statist. 13*, 1024-1041.

Scott, D.W. and G.R. Terrell, (1987), Biased and unbiased cross-validation in density estimation, *JASA 82*, 1131-1146.

Scott, D.W. and L.E. Factor (1981), Monte Carlo study of three data based nonparametric probability density estimators, *JASA 76*, 9-15.

Serfling, R.J. (1980), Approximation Theorems in Mathematical Statistics, Wiley, New York.

Serfling, R.J. (1982), Properties and applications of metrics on nonparametric density estimators, *Proc. Int. Coll. on Nonparametric Statist. Inf., 859-873, (ed. B.V. Gnedenko, M.L. Puri, I. Vincze).*

Silverman, B.W. (1978a), Choosing the window width when estimating a density, *Biometrika 65*, 1-11.

Silverman, B.W. (1978b), Weak and strong uniform consistency of the kernel estimate of a density function and its derivatives, *Ann. Statist. 6*, 177-184 (Add. 8, 1175-1176 (1980)).

Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York.

Stone, C. J. (1983), Optimal uniform rates of convergence for nonparametric estimators of a density function or its derivatives, *Recent Advances in Statistics : Papers in Honor of Herman Chernoff on his Sixtieth Birthday, 393-406, M. H. Rizvi, J. S. Rustagi and D. Siegmund (eds)*, Academic Press, New York.

Stone, C.J. (1984), An asymptotically optimal window selection rule for kernel density estimates, *Ann. Statist. 12*, 1285-1297.

Stoyan, D., Kendall, W.S. and J. Mecke (1987), *Stochastic Geometry and Its Applications*, Akademie-Verlag, Berlin.

Stute, W. (1982a), The oscillation behavior of empirical processes, *Ann. Probab. 10*, 86-107.

Stute, W. (1982b), A law of the logarithm for kernel density estimators, *Ann. Probab. 10*, 414-422.

Swanepoel, J.H. (1987), Optimal kernels when estimating non-smooth densities, *Commun. Statist.- Theory Meth. 16*, 1835-1848.

Taylor, C.C. (1983), A new method for unfolding sphere size distributions, *J. Microsc. 132*, 57-66.

Uspensky, J.V. (1937), *Introduction to Mathematical Probability*, MacGraw-Hill, New York.

Wicksell, S.D. (1925), The corpuscle problem, Part I, *Biometrika 17*, 84-99.

Wilson, R.J. (1975), *Introduction to Graph Theory*, Longman, London.

SAMENVATTING.

Het grootste deel van dit proefschrift is gewijd aan de *Parzen-Rosenblatt kernschatter*, gedefinieerd door formule (1.1). Dit is een schatter van de kansdichtheidsfunktie, zeg f, van n onafhankelijke identiek verdeelde stochastische variabelen $X_1, ..., X_n$. In het laatste hoofstuk besteden we aandacht aan het *deconvolutie* probleem.

Na een korte introductie bestuderen we in hoofdstuk 2 eigenschappen van kernschatters met de nadruk op het gedrag in gevallen waarin de dichtheid f niet glad is, d.w.z. we laten toe dat f sprongen en knikken heeft. We beginnen met het afleiden van ontwikkelingen van de bias en variantie van een kernschatter in een vast punt. Daarna bestuderen we eigenschappen met betrekking tot een bekende verliesfunktie, de geintegreerde gekwadrateerde fout. We behandelen het asymptotische gedrag van de verwachting van deze verliesfunktie en het daarmee verbandhoudende probleem van optimale bandgrootten. Eveneens bewijzen we een centrale limietstelling voor deze verliesfunktie. In het laatste deel van het hoofdstuk geven we enige resultaten met betrekking tot de supremum afstand. Een aantal resultaten in dit hoofdstuk, met name de ontwikkelingen van de bias en een ordegrens voor het supremum van de fout van een kernschatter, zijn belangrijke technische hulpmiddelen in het volgende hoofdstuk.

Om een kernschatter te kunnen uitrekenen moeten we eerst een kernfunctie K en een bandgrootte h>0 kiezen. Het is bekend dat de keuze van de kernfunktie voor de meeste verliesfunkties minder belangrijk is dan de keuze van de bandgrootte h. Uit resultaten in hoofdstuk 2 blijkt dat asymptotisch optimale bandgrootten afhangen van de onbekende dichtheid f. We kunnen deze bandgrootten dus niet zomaar uitrekenen. Om die reden zijn er methoden voorgesteld om goede bandgrootten te schatten, d.w.z. om bandgrootten te berekenen op grond van de steekproef. We beperken ons tot de zogenaamde *cross-validation methoden.* Na een inleiding over *least squares cross-validation*, een methode waarover reeds vrij veel bekend is, richten we ons op *likelihood cross-validation.* We laten zien dat de bijna zekere orde van convergentie naar nul van de met deze methode berekende bandgrootten afhangt van de aanwezigheid van sprongen en knikken in de dichtheid f. Het blijkt dat, als f geen sprongen heeft, de berekende bandgrootten bijna zeker asymptotisch equivalent zijn met de deterministiche asymptotisch optimale bandgrootten met betrekking tot een speciale verwachtte geintegreerde gekwadrateerde fout. Als f echter wel sprongen heeft dan is dit niet meer het geval. Dan heeft de berekende bandgrootte wel bijna zeker de optimale orde van convergentie naar nul, n.l. $n^{-1/2}$, maar niet de juiste constante. Vervolgens veronderstellen we dat f geen sprongen heeft en bewijzen we asymptotische normaliteit van de berekende bandgrootten.

In het laatste hoofdstuk behandelen we *deconvolutie*, met als belangrijkste voorbeeld het *Wicksell probleem*. Bij het deconvolutie probleem hebben we de beschikking over een steekproef X_1, ..., X_n uit een verdeling die de convolutie is van een bekende verdeling en een onbekende verdeling die we willen schatten. We beperken ons tot het schatten van de onbekende verdelingsfunktie in een vast punt. Uit de resultaten blijkt dat, hoe gladder de bekende verdeling, des te moeilijker is het om de onbekende verdeling te schatten. We geven drie voorbeelden van problemen waarin de niet parametrische maximum likelihood schatter van de onbekende verdelingsfunktie uitgerekend kan worden. Voor het eerste voorbeeld en het Wicksell probleem zijn een aantal schattingen berekend op grond van gesimuleerde steekproeven.