# Introduction to Nonparametric Estimation of Distribution, Density and Regression

by

Carlos Brunet Martins-Filho
Department of Economics
University of Colorado at Boulder
256 UCB
Boulder, CO 80309-0256
USA
email: carlos.martins@colorado.edu

# 1 Univariate density estimation over $\mathbb{R}$

## 1.1 The Rosenblatt density estimator and basic asymptotic results

Let $X : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be a random variable. Then, $P_X := P \circ X^{-1}$ is a probability on $\mathcal{B}(\mathbb{R})$. Let $F_X : \mathbb{R} \to [0,1]$ with $F_X(x) := P_X((-\infty, x]) : \mathbb{R} \to [0,1]$ be a proper distribution function.

**Definition 1** (Absolute continuity)**.** *A function $F : [a,b] \to \mathbb{R}$ is* absolutely continuous on $[a,b]$ *if for every $\epsilon > 0$ there exists $\delta > 0$ such that for every finite family $\{(a_1, b_1), \ldots, (a_n, b_n)\}$ of pairwise disjoint open subintervals of $[a,b]$,*

$$\sum_{k=1}^{n} (b_k - a_k) < \delta \quad \Longrightarrow \quad \sum_{k=1}^{n} \big|F(b_k) - F(a_k)\big| < \epsilon.$$

*A function $F : \mathbb{R} \to \mathbb{R}$ that is absolutely continuous on $[a,b]$ for every $a < b$ is said to be* locally absolutely continuous.

Note that if $F$ is absolutely continuous on $[a,b]$ and if $\{(a_k, b_k)\}_{k \geq 1}$ is a countable family of pairwise disjoint open subintervals of $[a,b]$ with $\sum_{k \geq 1}(b_k - a_k) < \delta$, then

$$\sum_{k \geq 1} \big|F(b_k) - F(a_k)\big| \ \leq \ \epsilon.$$

This follows since $\sum_{k \leq n}(b_k - a_k) \leq \sum_{k \geq 1}(b_k - a_k) < \delta$. Hence $\sum_{k \leq n} |F(b_k) - F(a_k)| < \epsilon$ for every $n$. The partial sums of a series of nonnegative terms are nondecreasing, so the series converges and its sum, being the supremum of the partial sums, is $\leq \epsilon$.

**Definition 2** (Absolute continuity of measures)**.** *Let $\mu, \nu$ be measures on a measurable space $(E, \mathcal{E})$. We write $\mu \ll \nu$ ($\mu$ is* absolutely continuous with respect to *$\nu$) if*

$$A \in \mathcal{E}, \ \nu(A) = 0 \quad \Longrightarrow \quad \mu(A) = 0.$$

Recall that every nonempty open $U \subseteq \mathbb{R}$ is a countable union of pairwise disjoint open intervals, $U = \bigsqcup_{k \in K} (a_k, b_k)$ with $K$ countable, and the Lebesgue measure $\lambda(U) = \sum_{k \in K}(b_k - a_k)$. If moreover $\lambda(U) < \infty$, every component is bounded.

**Lemma 1** (Outer approximation of null sets)**.** *Let $A \in \mathcal{B}(\mathbb{R})$ with $\lambda(A) = 0$ and let $\delta > 0$. Then there is an open set $U \supseteq A$ with $\lambda(U) < \delta$. If in addition $A \subseteq [-m, m]$, then $U$ may be chosen with $U \subseteq (-m-1, m+1)$.*

*Proof.* By the definition of Lebesgue outer measure as an infimum over countable covers by open intervals, there are open intervals $(c_j, d_j)$, $j \geq 1$, with $A \subseteq \bigcup_j (c_j, d_j)$ and $\sum_j (d_j - c_j) < \lambda^*(A) + \delta = \delta$, where the last equality follows from the fact that for measurable sets the Lebesgue outer measure coincides with the Lebesgue measure ($\lambda(A) = \lambda^*(A) = 0$). Put $U := \bigcup_j (c_j, d_j)$, which is open, contains $A$, and satisfies $\lambda(U) \leq \sum_j (d_j - c_j) < \delta$ by subadditivity. For the last statement replace $U$ by $U' := U \cap (-m-1, m+1)$, which is still open, still contains $A$, and $\lambda(U') \leq \lambda(U) < \delta$. □

**Theorem 1.** *Let $F_X$ be the distribution function of $X$. If $F_X$ is locally absolutely continuous on $\mathbb{R}$, then $P_X \ll \lambda$.*

*Proof.* Let $A \in \mathcal{B}(\mathbb{R})$ with $\lambda(A) = 0$. We will show that $P_X(A) = 0$. Fix $\epsilon > 0$ and let $\delta > 0$ be as in Definition 1 for this $\epsilon$.

By Lemma 1 choose an open $U \supseteq A$ with $\lambda(U) < \delta$. Write $U = \bigsqcup_{k \geq 1} (a_k, b_k)$ as a countable disjoint union of bounded open intervals with

$$\sum_{k \geq 1} (b_k - a_k) = \lambda(U) < \delta.$$

(If $U = \emptyset$ then $A = \emptyset$ and there is nothing to prove; if the family is finite, the argument below is unchanged.) The fact that $F_X$ is nondecreasing lets us drop the absolute values and

$$\sum_{k \geq 1} \big(F_X(b_k) - F_X(a_k)\big) \;\leq\; \epsilon. \tag{1}$$

On the other hand, for each $k$, monotonicity of $P_X$ gives

$$P_X\big((a_k, b_k)\big) \;\leq\; P_X\big((a_k, b_k]\big) \;=\; F_X(b_k) - F_X(a_k).$$

Hence, by monotonicity and countable additivity of $P_X$ together with (1),

$$P_X(A) \;\leq\; P_X(U) \;=\; \sum_{k \geq 1} P_X\big((a_k, b_k)\big) \;\leq\; \sum_{k \geq 1} \big(F_X(b_k) - F_X(a_k)\big) \;\leq\; \epsilon.$$

Since $\epsilon > 0$ was arbitrary and $P_X(A)$ does not depend on $\epsilon$, we conclude $P_X(A) = 0$. □

We now state the Radon–Nikodým theorem without proof.[1]

**Theorem 2** (Radon–Nikodým)**.** *Let $\mu$ and $\nu$ be measures on the measurable space $(E, \mathcal{E})$. If $\nu$ is $\sigma$-finite, then the following statements are equivalent:*

[1] See Schilling, R. (2017) Measures, Integrals and Martingales, p. 230 for a proof.

1. $\mu(A) = \int_A f d\nu$ *for some* $\nu$*-almost everywhere unique* $\mathcal{E}$*-measurable* $f \geq 0$

2. $\mu \ll \nu$.

**Lemma 2.** *$P_X$ is a finite measure and $\lambda$ is $\sigma$-finite on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.*

*Proof.* $P_X(\mathbb{R}) = \mathbb{P}(X^{-1}(\mathbb{R})) = \mathbb{P}(\Omega) = 1 < \infty$. And $\mathbb{R} = \bigcup_{m\in\mathbb{N}}[-m, m]$ with $\lambda([-m, m]) = 2m < \infty$. □

**Theorem 3.** *Let $X$ be a random variable whose distribution function $F_X$ is locally absolutely continuous. Then there exists a Borel measurable function $f_X : \mathbb{R} \to [0, \infty)$ such that*

$$P_X(B) = \int_B f_X \, d\lambda \quad \textit{for all } B \in \mathcal{B}(\mathbb{R}), \qquad F_X(x) = \int_{-\infty}^{x} f_X \, d\lambda \quad \textit{for all } x \in \mathbb{R}, \tag{2}$$

*and $\int_{\mathbb{R}} f_X \, d\lambda = 1$. The function $f_X$ is unique up to $\lambda$-null sets.*

*Proof.* By Theorem 1 we have $P_X \ll \lambda$, and by Lemma 2 the hypotheses of Theorem 2 hold with $\mu = P_X$, $\nu = \lambda$, $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Then, there exists a Borel measurable function $g \geq 0$, real-valued $\lambda$-a.e. such that

$$P_X(B) = \int_B g \, d\lambda \qquad \text{for all } B \in \mathcal{B}(\mathbb{R}). \tag{3}$$

Taking $B = \mathbb{R}$ gives $\int_{\mathbb{R}} g \, d\lambda = 1$ and $g < \infty$ $\lambda$-a.e.

Now set

$$f_X := g \cdot \mathbf{1}_{\{0 \leq g < \infty\}},$$

which is Borel measurable, takes values in $[0, \infty)$ *at every point*, and differs from $g$ only on the $\lambda$-null set $N \cup \{|g| = \infty\}$. Since altering the integrand on a $\lambda$-null set changes no integral, (3) holds with $f_X$ in place of $g$, which is the first identity in (2).

Now, taking $B = (-\infty, x]$ in (2) gives

$$F_X(x) = P_X\big((-\infty, x]\big) = \int_{(-\infty, x]} f_X \, d\lambda = \int_{-\infty}^{x} f_X \, d\lambda,$$

the endpoint being immaterial as $\lambda(\{x\}) = 0$. Taking $B = \mathbb{R}$ gives $\int_{\mathbb{R}} f_X \, d\lambda = P_X(\mathbb{R}) = 1$; in particular $f_X \in L^1(\mathbb{R}, \lambda)$.

Lastly, if $f$ and $f'$ both satisfy (2) then $\int_B (f - f') \, d\lambda = 0$ for every $B \in \mathcal{B}(\mathbb{R})$; both are integrable, so taking $B = \{f > f'\}$ and $B = \{f < f'\}$ forces $\lambda(B) = 0$ in each case, hence $f = f'$ $\lambda$-a.e. □

**Definition 3.** *A measurable function $f : \mathbb{R} \to \mathbb{R}$ is said to be locally integrable with respect to Lebesgue measure $\lambda$ if $\int_K |f| d\lambda < \infty$ for every compact set $K \subset \mathbb{R}$. The space of such functions is denoted by $L^1_{\text{loc}}(\mathbb{R}, \lambda)$.*

**Theorem 4.** *Let $L_f := \left\{ x : \lim_{r \to 0} \frac{1}{2r} \int_{(x-r, x+r)} |f - f(x)| d\lambda = 0 \right\}$ be the Lebesgue set of $f$. If $f \in L^1_{\text{loc}}(\mathbb{R}, \lambda)$ then $\lambda(L_f^c) = 0$.*

**Theorem 5.** *Let $f \in L^1_{\text{loc}}(\mathbb{R}, \lambda)$ and define*

$$F(x) \;:=\; \int_{-\infty}^{x} f \, d\lambda, \qquad x \in \mathbb{R},$$

*assuming the integral is finite for every $x$. Then for $\lambda$-almost every $x \in \mathbb{R}$ the two-sided limit*

$$F'(x) \;=\; \lim_{h \to 0} \frac{F(x+h) - F(x)}{h}$$

*exists in $\mathbb{R}$ and satisfies*

$$F'(x) \;=\; f(x).$$

*Proof.* By Theorem 4 there is a Borel set $E \subset \mathbb{R}$ with $\lambda(E) = 0$ such that every $x \in \mathbb{R} - E$ is a Lebesgue point of $f$ with $|f(x)| < \infty$, i.e.

$$\lim_{r \downarrow 0} \frac{1}{2r} \int_{x-r}^{x+r} \big| f(t) - f(x) \big| \, dt \;=\; 0. \tag{4}$$

Fix such an $x$ and let $h \neq 0$. Additivity of the integral over disjoint Borel sets gives $F(x+h) - F(x) = \int_x^{x+h} f \, d\lambda$, and since $\int_x^{x+h} f(x) \, dt = h \, f(x)$ we obtain

$$\frac{F(x+h) - F(x)}{h} - f(x) \;=\; \frac{1}{h} \int_x^{x+h} \big( f(t) - f(x) \big) \, dt.$$

The interval with endpoints $x$ and $x+h$ is contained in $[x - |h|,\, x + |h|]$ for either sign of $h$, hence

$$\left| \frac{F(x+h) - F(x)}{h} - f(x) \right| \;\le\; \frac{1}{|h|} \int_{x-|h|}^{x+|h|} \big| f(t) - f(x) \big| \, dt \;=\; 2 \cdot \frac{1}{2|h|} \int_{x-|h|}^{x+|h|} \big| f(t) - f(x) \big| \, dt.$$

By (4) applied with $r = |h|$, the right-hand side tends to 0 as $h \to 0$. Therefore $F'(x)$ exists and equals $f(x)$ for every $x \in \mathbb{R} - E$, and $\lambda(E) = 0$. □

Suppose we observe a random sample of size $n \in \mathbb{N}$ denoted by $\{X_i\}_{i=1}^n$. Our first objective is to use these observations to estimate the value of the density $f$ at some $x \in \mathbb{R}$, i.e., $f(x)$.

**Definition 4.** *Let $\{h_n\}_{n\in\mathbb{N}}$ be a sequence of nonstochastic positive scalars, let $x \in \mathbb{R}$ and let $K : \mathbb{R} \to \mathbb{R}$ be a measurable function. Then, given a set of observations $\{X_i\}_{i=1,2,\cdots,n}$ we define*

$$\hat{f}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h_n}\right).$$

**Remark 1.** *1. We call $\hat{f}(x)$ the Rosenblatt kernel density estimator of $f(x)$. In this form it was first proposed by Rosenblatt (1956). $h_n$ is called a bandwidth and $K$ is called a kernel function.*

*2. The estimator can be motivated using the empirical distribution function and Theorem 3. Given a random sample $\{X_i\}_{i=1}^n$, the empirical distribution function is given by*

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

*where $I_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A \\ 0, & \text{if } \omega \notin A \end{cases}$ is the indicator function for the set $A$. Since*

$$E\left(I_{\{\omega: X_i(\omega) \le x\}}(\omega)\right) = F(x),$$

*by the measurability of $I_A$ and its boudedness, we have by Khinchin's Law of Large Numbers, $F_n(x) \xrightarrow{p} F(x)$ for all $x \in \mathbb{R}$. From Calculus we have that $\lim_{h\to 0} \frac{F(x+h)-F(x-h)}{2h} = f(x)$ for every point $x$ at which $f$ is continuous. Hence, for $h$ sufficiently small we have $f(x) \approx \frac{F(x+h)-F(x-h)}{2h}$. Replacing $F$ with $F_n$ we have $f_s(x) = \frac{1}{nh_n} \sum_{i=1}^{n} \frac{1}{2} I_{\{\omega: x-h_n < X_i(\omega) \le x+h_n\}}(\omega) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h_n}\right)$ where $K(u) = \frac{1}{2} I(-1 < u \le 1)$. The estimator in Definition 4 allows $K$ to be a generic measurable function. Restrictions on $K$ that go beyond measurability are imposed to obtain properties of $\hat{f}$.*

*3. Commonly used kernels with $\int_{\mathbb{R}} K(u)du = 1$ are:*

1. *$K(u) = \frac{1}{2} I(|u| \le 1)$ (rectangular)*
2. *$K(u) = (1 - |u|) I(|u| \le 1)$ (triangular)*
3. *$K(u) = \frac{3}{4}(1 - u^2) I(|u| \le 1)$ (Epanechnikov/parabolic)*
4. *$K(u) = \frac{15}{16}(1 - u^2)^2 I(|u| \le 1)$ (biweight)*
5. *$K(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)$ (Gaussian)*

Several questions regarding this estimator arise:

- How to choose $K$?
- How to choose $\{h_n\}_{n\in\mathbb{N}}$?
- Does $\hat{f}$ converge to $f$? If it does, then in what sense?
- How does the estimator perform when the support of the density is a finite interval?

Of course, a related problem is the estimation of the distribution function $F(x) = P(X \leq x)$. Because of the link $F(x) = \int_{-\infty}^{x} f(t)\,dt$ a natural idea is to estimate $F(x)$ by integrating $\hat{f}$:

$$\hat{F}(x) = \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{x} K\left(\frac{X_i - t}{h}\right)dt. \tag{5}$$

This definition gives rise to the same questions as for $\hat{f}$.

**Theorem 6.** *Assume that: a)* $\{X_i\}_{i=1}^n$ *is a random sample of observations on* $X$ *which has a distribution* $F$ *that is absolutely continuous with respect to* $\lambda$*, with density* $f$*; b.1)* $K : \mathbb{R} \to \mathbb{R}$ *is measurable; b.2)* $E\left|K\left(\frac{X-x}{h_n}\right)\right| < \infty$*; b.3)* $E\left(K\left(\frac{X-x}{h_n}\right)^2\right) < \infty$*. Then,*

$$E(\hat{f}(x)) = \int K(u)f(x + h_n u)du$$

*and*

$$V(\hat{f}(x)) = \frac{1}{nh_n}\int K^2(v)f(x + h_n v)dv - \frac{1}{n}\left(\int K(v)f(x + h_n v)dv\right)^2.$$

*Proof.* Fix $n$. Since $\{X_i\}_{i=1}^n$ is an independent and identically distributed (IID) sequence with the distribution of $X$ and $K$ is measurable, $\left\{\frac{1}{h_n}K\left(\frac{X_i-x}{h_n}\right)\right\}_{i=1}^n$ is IID sequence of random variables. Thus, given b.2)

$$E(\hat{f}(x)) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}E\left(K\left(\frac{X_i - x}{h_n}\right)\right) = \frac{1}{h_n}E\left(K\left(\frac{X - x}{h_n}\right)\right) = h_n^{-1}\int K\left(\frac{u-x}{h_n}\right)f(u)du.$$

Letting $v = \frac{u-x}{h_n}$, we have $E(\hat{f}(x)) = h_n^{-1}\int K(v)f(x + h_n v)h_n dv = \int K(v)f(x + h_n v)dv$. Given b.3) $V(\hat{f}(x)) = \frac{1}{nh_n^2}V\left(K\left(\frac{X-x}{h_n}\right)\right)$. Therefore,

$$\begin{aligned} V(\hat{f}(x)) &= \frac{1}{nh_n^2}\left(\int K^2\left(\frac{u-x}{h_n}\right)f(u)du - \left(\int K\left(\frac{u-x}{h_n}\right)f(u)du\right)^2\right) \\ &= \frac{1}{nh_n}\int K^2(v)f(x + h_n v)dv - \frac{1}{n}\left(\int K(v)f(x + h_n v)dv\right)^2. \end{aligned}$$

□

**Remark 2.** *1. Note that by the Cauchy-Schwarz Inequality, condition b.2) is implied by b.3).*

*2. In general, $E(\hat{f}(x)) \neq f(x)$ indicating that $\hat{f}(x)$ is a biased estimator for $f(x)$. The behavior of the bias and the variance as $n \to \infty$ can be investigated by using the following theorem.*

**Theorem 7.** *Assume that: a) $\int |K(x)|dx < \infty$; b) $|x||K(x)| \to 0$ as $|x| \to \infty$; c) $h_n \to 0$ as $n \to \infty$. Let $f : \mathbb{R} \to \mathbb{R}$ such that d) $\int |f(x)|dx < \infty$. Then, for every point of continuity $x$ of $f$,*

$$\frac{1}{h_n} E\left(K\left(\frac{X_i - x}{h_n}\right)\right) \to f(x) \int K(u)du \text{ as } n \to \infty. \tag{6}$$

*Proof.* We need to show that for all $\epsilon > 0$ there exists $N_{\epsilon,x}$ such that whenever $n > N_{\epsilon,x}$ we have

$$\left| \int (f(x + h_n u) - f(x))K(u)du \right| < \epsilon.$$

By the triangle inequality

$$\left| \int (f(x + h_n u) - f(x))K(u)du \right| \leq \int |f(x + h_n u) - f(x)| \, |K(u)| \, du$$

and since $x$ is a point of continuity of $f$, for all $\epsilon > 0$ there exists some $\delta_{\epsilon,x} > 0$ such that $|f(x + h_n u) - f(x)| < \epsilon$ provided that $|h_n u| < \delta_{\epsilon,x}$ or $|u| < \delta_{\epsilon,x}/h_n$. Hence, by partitioning the range of integration and using the triangle inequality we have

$$\begin{aligned} \left| \int (f(x + h_n u) - f(x))K(u)du \right| &\leq \epsilon \int_{|u| < \frac{\delta_{\epsilon,x}}{h_n}} |K(u)| \, du + \int_{|u| \geq \frac{\delta_{\epsilon,x}}{h_n}} |f(x + h_n u)| \, |K(u)| \, du \\ &+ |f(x)| \int_{|u| \geq \frac{\delta_{\epsilon,x}}{h_n}} |K(u)| \, du. \end{aligned}$$

By a) and c), $\int_{|\gamma| \geq \delta_{\epsilon,x}/h_n} |K(\gamma)| \, d\gamma \to 0$ as $n \to \infty$ since $h_n \to 0$. Furthermore, since $|f(x)|$ is $\mathbb{R}$-valued the third term after the inequality can be made as small as desired. Also, by a) $\int_{|\gamma| < \delta_{\epsilon,x}/h_n} |K(\gamma)| \, d\gamma < C$. Thus, the first and the third terms on the right hand side of the inequality can be made arbitrarily small. Now, by a change of variable ($h_n u = v$) in integration

$$\begin{aligned} \int_{|u| \geq \delta_{\epsilon,x}/h_n} |f(x + h_n u)| \, |K(u)| \, du &= \int_{|v| \geq \delta_{\epsilon,x}} \frac{1}{|v|} |f(x + v)| \left|\frac{v}{h_n}\right| \left|K\left(\frac{v}{h_n}\right)\right| dv \\ &\leq \frac{1}{\delta_{\epsilon,x}} \int_{|v| \geq \delta_{\epsilon,x}} |f(x + v)| \left|\frac{v}{h_n}\right| \left|K\left(\frac{v}{h_n}\right)\right| dv \\ &\leq \frac{1}{\delta_{\epsilon,x}} \sup_{|v| \geq \delta_{\epsilon,x}} \left|\frac{v}{h_n}\right| \left|K\left(\frac{v}{h_n}\right)\right| \int_{|v| \geq \delta_{\epsilon,x}} |f(x + v)| dv. \end{aligned}$$

By d) $\int_{|v|\geq\delta_{\epsilon,x}} |f(x+v)|dv < C$, hence we need $\sup_{|v|\geq\delta_{\epsilon,x}} \left|\frac{v}{h_n}\right| \left|K\left(\frac{v}{h_n}\right)\right| \to 0$ as $n \to \infty$. Since $h_n \to 0$ as $n \to \infty$ it suffices to have $|x||K(x)| \to 0$ as $|x| \to \infty$, which is assumed in b). □

**Remark 3.** *1. Note that for $1 < r < \infty$, if $|K(u)| < C$, $|K(u)|^r$ satisfies assumptions a)-b) in Theorem 7. Thus,*

$$h_n^{-1} E\left(\left|K^r\left(\frac{X_i - x}{h_n}\right)\right|\right) = \int |K(u)|^r f(x + h_n u)du \to f(x)\int |K(u)|^r du \text{ as } n \to \infty.$$

*2. Under the assumptions of Theorems 6 and 7 we have that $V(\hat{f}(x)) = \frac{f(x)}{nh_n}\int K^2(u)du + o((nh_n)^{-1})$.*

The next theorem is a direct consequence of Theorem 7.

**Theorem 8.** *Let $\{X_i\}_{i=1}^n$ be a sequence of IID random variables with density $f$. Assume that a) $|K(u)| \leq C$ for all $u \in \mathbb{R}$; b) $\int |K(u)|du < \infty$ and $\int K(u)du = 1$; c) $|u||K(u)| \to 0$ as $|u| \to \infty$; d) $h_n > 0$ for all $n$ and $h_n \to 0$ as $n \to \infty$. Then,*

1. *$E(\hat{f}(x)) - f(x) \to 0$ as $n \to \infty$, for every $x$ which is a point of continuity of $f$.*
2. *if $nh_n \to \infty$ as $n \to \infty$, $V(\hat{f}(x)) \to 0$ for every $x$ which is a point of continuity of $f$.*

**Remark 4.**
1. *If $h_n = Cn^\delta$, then it must be that $\delta < 0$ since $h_n \to 0$ as $n \to \infty$. Furthermore, in this case $nh_n = Cn^{1+\delta}$ and if $nh_n \to \infty$ as $n \to \infty$, then it must be that $1 + \delta > 0$. Combining these two inequalities we conclude that $-1 < \delta < 0$.*
2. *The speed of convergence in item 1 of Theorem 8 improves when $f$ is smooth. There are two main ways to characterize smoothness: using derivatives and finite differences.*

**Theorem 9.** *Let $\{X_i\}_{i=1}^n$ be a sequence of IID random variables with density $f$. Assume that a) $|K(u)| \leq C$ for all $u \in \mathbb{R}$; b) $\int |K(u)|du < \infty$ and $\int K(u)du = 1$; c) $|u||K(u)| \to 0$ as $|u| \to \infty$. If $h_n > 0$ for all $n$, $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$ then $E\left((\hat{f}(x) - f(x))^2\right) \to 0$ for every $x$ which is a point of continuity of $f$.*

*Proof.* Note that

$$\begin{aligned} E\left((\hat{f}(x) - f(x))^2\right) &= E\left((\hat{f}(x) - E(\hat{f}(x)) + E(\hat{f}(x)) - f(x))^2\right) = E\left((\hat{f}(x) - E(\hat{f}(x))^2\right) \\ &+ E\left((E(\hat{f}(x)) - f(x))^2\right) + 2\left(E(\hat{f}(x) - E(\hat{f}(x)))(E(\hat{f}(x)) - f(x))\right) \\ &= V(\hat{f}(x)) + (E(\hat{f}(x)) - f(x))^2. \end{aligned}$$

By Theorem 8 $(E(\hat{f}(x)) - f(x))^2 \to 0$ and $V(\hat{f}(x)) \to 0$ as $n \to \infty$. □

Since convergence in quadratic mean implies convergence in probability, a direct consequence of Theorem 9 is that $\hat{f}(x) - f(x) \xrightarrow{p} 0$ for points of continuity of $f$.

The next theorem establishes that a suitably centered and scaled version of the Rosenblatt density estimator converges in distribution to a standard normal random variable.

**Theorem 10.** *Assume that: a)* $\{X_i\}_{i=1}^n$ *forms an IID sequence of random variables with density* $f$*; b)* $K$ *is measurable with* $|K(u)| \leq M$ *for all* $|u| \in \mathbb{R}$*; c)* $\int |K(u)|du < \infty$ *and* $K \neq 0$ *on a set of positive Lebesgue measure; d)* $|u||K(u)| \to 0$ *as* $|u| \to \infty$*; e)* $h_n > 0$ *for all* $n$*,* $h_n \to 0$ *and* $nh_n \to \infty$ *as* $n \to \infty$*. Then,*

$$\frac{\hat{f}(x) - E(\hat{f}(x))}{\sqrt{V(\hat{f}(x))}} \xrightarrow{d} N(0,1),$$

*where* $x$ *is a point of continuity of* $f$ *and* $f(x) > 0$*.*

*Proof.* $\hat{f}(x) - E(\hat{f}(x)) = \sum_{i=1}^n \left( \frac{1}{nh_n} K\left(\frac{X_i - x}{h_n}\right) - \frac{1}{nh_n} E\left(K\left(\frac{X_i - x}{h_n}\right)\right)\right)$. Let $Z_{ni} = \frac{1}{nh_n} K(\frac{X_i - x}{h_n})$, $E(Z_{ni}) = \mu_n$ and $s_n^2 = \sum_{i=1}^n E(Z_{ni} - \mu_n)^2$. Note that

$$\begin{aligned} s_n^2 &= \sum_{i=1}^n \frac{1}{n^2 h_n^2} E\left(K\left(\frac{X_i - x}{h_n}\right) - E\left(K\left(\frac{X_i - x}{h_n}\right)\right)\right)^2 = \frac{1}{nh_n^2} V\left(K\left(\frac{X_i - x}{h_n}\right)\right) \\ &= V\left(\hat{f}(x)\right). \end{aligned}$$

Hence, $\frac{\hat{f}(x) - E(\hat{f}(x))}{\sqrt{V(\hat{f}(x))}} = \sum_{i=1}^n \left(\frac{Z_{ni} - \mu_n}{s_n}\right) = \sum_{i=1}^n X_{ni}$ with $E(X_{ni}) = 0$, $E(X_{ni}^2) = \frac{1}{s_n^2} E(Z_{ni} - \mu_n)^2$ and $\sum_{i=1}^n E(X_{ni}^2) = 1$. Consequently, by Lyapunov's Central Limit Theorem (CLT) (see Davidson, 1994, p. 372)

$$\frac{\hat{f}(x) - E(\hat{f}(x))}{\sqrt{V(\hat{f}(x))}} \xrightarrow{d} N(0,1)$$

provided that $\lim_{n\to\infty} \sum_{i=1}^n E|X_{ni}|^{2+\delta} = 0$ for some $\delta > 0$. We now verify that this limit is in fact zero.

$$\begin{aligned} \sum_{i=1}^n E|X_{ni}|^{2+\delta} &= \sum_{i=1}^n E\left(\left|\frac{Z_{ni} - \mu_n}{s_n}\right|^{2+\delta}\right) = \sum_{i=1}^n V(\hat{f}(x))^{-1-\frac{\delta}{2}} E\left(|Z_{ni} - \mu_n|^{2+\delta}\right) \\ &= V(\hat{f}(x))^{-(1+\frac{\delta}{2})} n E\left(|Z_{ni} - \mu_n|^{2+\delta}\right) \end{aligned}$$

By the $c_r$ inequality we have

$$E\left(|Z_{ni} - \mu_n|^{2+\delta}\right) \leq 2^{1+\delta} E\left(|Z_{ni}|^{2+\delta} + |\mu_n|^{2+\delta}\right),$$

and by Jensen's Inequality $|\mu_n|^{2+\delta} = |E(Z_{ni})|^{2+\delta} \leq E\left(|Z_{ni}|^{2+\delta}\right)$. Therefore,

$$\begin{aligned}
\sum_{i=1}^{n} E|X_{ni}|^{2+\delta} &\leq V(\hat{f}(x))^{-(1+\frac{\delta}{2})} n 2^{2+\delta} E\left(|Z_{ni}|^{2+\delta}\right) \\
&= V(\hat{f}(x))^{-(1+\frac{\delta}{2})} 2^{2+\delta} n^{-1-\delta} E\left(\left|\frac{1}{h_n} K\left(\frac{X_i - x}{h_n}\right)\right|^{2+\delta}\right) \\
&= \frac{2^{1+\delta}}{(nh_n)^{1+\delta}} \left(V(\hat{f}(x))\right)^{-1-\frac{\delta}{2}} 2^{2+\delta} \frac{1}{h_n} E\left(\left|K\left(\frac{X_i - x}{h_n}\right)\right|^{2+\delta}\right) \\
&= \frac{1}{(nh_n V(\hat{f}(x)))^{1+\frac{\delta}{2}}} \frac{2^{2+\delta}}{(nh_n)^{\delta/2}} \frac{1}{h_n} E\left(\left|K\left(\frac{X_i - x}{h_n}\right)\right|^{2+\delta}\right)
\end{aligned}$$

By Theorem 7, as $n \to \infty$, $nh_n V(\hat{f}(x)) \to f(x) \int K^2(v)dv$ and

$$\frac{1}{h_n} E\left(\left|K\left(\frac{X_i - x}{h_n}\right)\right|^{2+\delta}\right) \to f(x) \int |K(u)|^{2+\delta} du.$$

Also, by assumption $\frac{1}{(nh_n)^{\delta/2}} \to 0$. Therefore, $\sum_{i=1}^{n} E|X_{ni}|^{2+\delta} \to 0$ as $n \to \infty$. □

**Remark 5.** *1. Since $nh_n V(\hat{f}(x)) \to f(x) \int K^2(u)du$ we can write*

$$\frac{(nh_n)^{\frac{1}{2}}(\hat{f}(x) - E(\hat{f}(x)))}{(f(x) \int K^2(u)du)^{\frac{1}{2}}} \xrightarrow{d} N(0,1) \text{ or } \sqrt{nh_n}(\hat{f}(x) - E(\hat{f}(x))) \xrightarrow{d} N\left(0, f(x) \int K^2(u)du\right).$$

*2. Note that the normalization $\sqrt{nh_n}$ is different from the standard parametric normalization $\sqrt{n}$. Since $h_n \to 0$ as $n \to \infty$, $\sqrt{nh_n}$ diverges more slowly than $\sqrt{n}$. Thus, $\hat{f}(x)$ converges at a rate slower than that of parametric estimators.*

*3. Centering is around $E(\hat{f}(x))$ instead of $f(x)$. Since $\hat{f}$ is biased, $\hat{f}(x) - E(\hat{f}(x)) \neq \hat{f}(x) - f(x)$. We will need to derive the order of decay of the bias to obtain asymptotic normality of $\sqrt{nh_n}(\hat{f}(x) - f(x))$.*

## 1.2 Bias approximation

It is desirable to provide approximations for the bias of $\hat{f}(x)$. In the next definition, for $0 < \beta \in \mathbb{R}$, $\lfloor \beta \rfloor$ denotes the largest integer strictly smaller than $\beta$.

**Definition 5.** *Let $\mathcal{I}$ be an interval in $\mathbb{R}$ and let $\beta, C > 0$. The Hölder class $\Sigma(\beta, C)$ on $\mathcal{I}$ is defined as the set of $\lfloor \beta \rfloor$-times differentiable functions $f : \mathcal{I} \to \mathbb{R}$ whose $\lfloor \beta \rfloor$-th derivative $f^{(\lfloor \beta \rfloor)}$ satisfies*

$$|f^{(\lfloor \beta \rfloor)}(x) - f^{(\lfloor \beta \rfloor)}(x')| \leq C|x - x'|^{\beta - \lfloor \beta \rfloor}, \, \forall \, x, x' \in \mathcal{I}, \, x \neq x'.$$

**Remark 6.** *1. Note that $0 < \beta - \lfloor\beta\rfloor \leq 1$. Since $\beta - \lfloor\beta\rfloor > 0$, $f^{(\lfloor\beta\rfloor)}(x)$ is continuous at every $x \in \mathcal{I}$.*

*2. If $f \in \Sigma(\beta, C)$ then all of its derivatives of order $\leq \lfloor\beta\rfloor$ exist for all $x \in \mathcal{I}$.*

*3. If $\beta - \lfloor\beta\rfloor = 1$, then $f^{(\lfloor\beta\rfloor)}$ satisfies a Lipschitz condition of order 1 in $\mathcal{I}$.*

**Definition 6.** *We say that $K : \mathbb{R} \to \mathbb{R}$ is a kernel of order $m \in \mathbb{N}$ if*

*1. $\int K(u)du = 1$*

*2. $\int u^j K(u)du = 0$ for $j = 1, \cdots, m$.*

**Remark 7.** *In much of the nonparametric statistical literature a kernel is said to be of order $m$ if $\int K(u)du = 1$, $\int u^j K(u)du = 0$ for $j = 1, \ldots, m-1$, and $\int |u|^m |K(u)|du < \infty$. In particular, when $\beta = 2$ and $\lfloor\beta\rfloor = 1$, the following Theorem 11 calls for a kernel satisfying $\int K = 1$ and $\int uK(u)du = 0$, which the usual convention would describe as a kernel of order 2.*

**Theorem 11.** *Assume that $f \in \Sigma(\beta, C)$ for $\beta \geq 2$ and $K$ is a kernel of order $\lfloor\beta\rfloor$. If, in addition, $\int |K(u)||u|^\beta du < C$ then*

$$E(\hat{f}(x)) - f(x) = O(h_n^\beta). \tag{7}$$

*Proof.* Since $f \in \Sigma(\beta, C)$, by Taylor's Theorem, for some $\lambda \in (0, 1)$

$$f(x + h_n u) - f(x) = \sum_{j=1}^{\lfloor\beta\rfloor - 1} \frac{1}{j!} f^{(j)}(x)(h_n u)^j + \frac{1}{\lfloor\beta\rfloor!} f^{(\lfloor\beta\rfloor)}(x + h_n u\lambda)(h_n u)^{\lfloor\beta\rfloor}.$$

Since $K$ is of order $\lfloor\beta\rfloor$ and $f \in \Sigma(\beta, C)$,

$$\begin{aligned}
E(\hat{f}(x)) - f(x) &= \int K(u)(f(x + h_n u) - f(x))du \\
&= \int K(u) \left( \sum_{j=1}^{\lfloor\beta\rfloor - 1} \frac{1}{j!} f^{(j)}(x)(h_n u)^j + \frac{1}{\lfloor\beta\rfloor!} f^{(\lfloor\beta\rfloor)}(x + h_n u\lambda)(h_n u)^{\lfloor\beta\rfloor} \right) du \\
&= \int K(u) \frac{1}{\lfloor\beta\rfloor!} f^{(\lfloor\beta\rfloor)}(x + h_n u\lambda)(h_n u)^{\lfloor\beta\rfloor} du \\
&= \frac{1}{\lfloor\beta\rfloor!} \int K(u) \left( f^{(\lfloor\beta\rfloor)}(x + h_n u\lambda) - f^{(\lfloor\beta\rfloor)}(x) \right) (h_n u)^{\lfloor\beta\rfloor} du
\end{aligned}$$

By the Triangle Inequality, $f \in \Sigma(\beta, C)$ and the fact that $\lambda \in (0,1)$, we have

$$\begin{aligned}|E(\hat{f}(x)) - f(x)| &\leq \frac{C}{\lfloor\beta\rfloor!}\int |K(u)||h_n u\lambda|^{\beta-\lfloor\beta\rfloor}|h_n u|^{\lfloor\beta\rfloor}du \\ &\leq C\int |K(u)||h_n u|^{\beta-\lfloor\beta\rfloor}|h_n u|^{\lfloor\beta\rfloor}du \leq Ch_n^{\beta}\end{aligned}$$

where the last inequality follows from the assumption that $\int |K(u)||u|^{\beta}du < C$ and $h_n > 0$. □

**Remark 8.** *1. It is common to assume that $\beta = 2$. In this case $E(\hat{f}(x) - f(x)) = O(h_n^2)$, provided $K$ is of order 1 and $\int u^2|K(u)|du < C$. If $K(u) \geq 0$, then the last condition becomes $\int u^2K(u)du < C$.*

*2. Given that $(nh_n)^{1/2}(\hat{f}(x) - f(x)) = (nh_n)^{1/2}(\hat{f}(x) - E(\hat{f}(x)) + (nh_n)^{1/2}(E(\hat{f}(x)) - f(x))$, we have, under the conditions of Theorem 11 that*

$$(nh_n)^{1/2}(\hat{f}(x) - f(x)) = (nh_n)^{1/2}(\hat{f}(x) - E(\hat{f}(x)) + (nh_n^{2\beta+1})^{1/2}O(1).$$

*Hence, provided $nh_n^{2\beta+1} \to 0$ we can write,*

$$(nh_n)^{1/2}(\hat{f}(x) - f(x)) = (nh_n)^{1/2}(\hat{f}(x) - E(\hat{f}(x)) + o(1).$$

Using Theorems 10, 11 and Remark 8.2 we have,

**Theorem 12.** *Under the assumptions of Theorems 10 and 11, if $nh_n^{2\beta+1} \to 0$ as $n \to \infty$, then*

$$\sqrt{nh_n}(\hat{f}(x) - f(x)) \xrightarrow{d} N\left(0, f(x)\int K^2(u)du\right).$$

*for every point of continuity $x$ such that $f(x) > 0$.*

**Remark 9.** *1. Theorem 10 requires that $h_n \to 0$ and $nh_n \to \infty$ and the condition in Remark 8.2 requires $nh_n^{2\beta+1} \to 0$. If $h_n = Cn^{\delta}$ for some $C > 0$, then it is easy to see that any $-1 < \delta < -\frac{1}{2\beta+1}$ produces a bandwidth sequence that satisfies all speed requirements. If $\beta$ is large, i.e., if the Hölder class includes functions with a large degree of smoothness, the upper bound on the range of values for $\delta$ is close to zero, thereby relaxing the constraint imposed on the choice of $h_n$.*

*2. If $\beta = 2$ in Theorem 11, then for any given $n$ the mean squared error in Theorem 9 is bounded above as $E\left((\hat{f}(x) - f(x))^2\right) \leq h_n^4C_1 + \frac{1}{nh_n}C_2$, if $\sup_x f(x) < \infty$. The first*

*term is a bound on the square of the bias and the second is a bound on the variance. The bias bound increases as $h_n$ increases whereas the variance bound diminishes as $h_n$ increases. This is generally referred to as the bias-variance tradeoff associated with the choice of $h_n$ (for fixed n). The bound on the right-hand side of the inequality is minimized at $h_n^* = \left(\frac{C_2}{4C_1}\right)^{1/5} n^{-1/5}$. Hence, substituting $h_n^*$ in the expression for the bound we have $E\left((\hat{f}(x) - f(x))^2\right) = O(n^{-4/5})$. Note that in this case the squared bias and the variance decay at the same rate.*

## 1.3 Uniform convergence in probability

In many contexts it is desirable to obtain uniform convergence of $\hat{f}$ over a specified subset of the range of $X_i$. The next theorem gives conditions under which $\hat{f}$ converges to $f$ uniformly in probability over a compact subset of $\mathbb{R}$.

**Theorem 13.** *Assume that: a)* $|K(u)| \leq C$ *for all* $u \in \mathbb{R}$*; b)* $\int |K(u)|du < \infty$*; c)* $|u||K(u)| \to 0$ *as* $|u| \to \infty$*; d) for any* $u, u' \in \mathbb{R}$ *with* $u \neq u'$ *we have* $|K(u') - K(u)| < C|u' - u|$ *for some arbitrary constant* $C < \infty$ *(Lipschitz condition of order 1 in* $\mathbb{R}$*); e)* $h_n \to 0$*,* $nh_n^3 \to \infty$ *and* $\frac{nh_n}{\log n} \to \infty$ *as* $n \to \infty$*. Then, if* $f$ *is continuous in* $G$*, a compact subset of* $\mathbb{R}$

$$\sup_{x \in G} |\hat{f}(x) - E(\hat{f}(x))| = O_p\left(\left(\frac{\log n}{nh_n}\right)^{1/2}\right).$$

*Proof.* Since $G$ is a compact subset of $\mathbb{R}$, it is closed and bounded. Therefore, there exists $x_0 \in \mathbb{R}$ and some $r > 0$ such that $G \subseteq B(x_0, r) = \{x \in \mathbb{R} : |x - x_0| < r\}$. In addition, for any two $x, x' \in G$ we have $|x - x'| < 2r$. For every $x \in G$ consider $B(x, r_x)$ and note that the collection $\{B(x, r_x) :$ for $x \in G$ and $r_x > 0\}$ is an open covering for $G$. By the Heine-Borel Theorem every open cover of $G$ contains a finite subcover, which we denote by $\{B(x_k, r_{x_k}) :$ for $x_k \in G$, $r_{x_k} > 0$ and $k = 1, \cdots, m$, $m \in \mathbb{N}.\}$. Now, by the triangle inequality

$$|\hat{f}(x) - E(\hat{f}(x))| \leq |\hat{f}(x_k) - E(\hat{f}(x_k))| + |\hat{f}(x) - \hat{f}(x_k)| + |E(\hat{f}(x)) - E(\hat{f}(x_k))|.$$

Given d), for $x \in B(x_k, r_{x_k})$ we have

$$|\hat{f}(x) - \hat{f}(x_k)| \leq \frac{C}{h_n^2}|x_k - x| < C\frac{r_{x_k}}{h_n^2} \text{ and } \left|E(\hat{f}(x)) - E(\hat{f}(x_k))\right| \leq \frac{C}{h_n^2}|x_k - x| < C\frac{r_{x_k}}{h_n^2}.$$

Hence,

$$|\hat{f}(x) - E(\hat{f}(x))| \leq |\hat{f}(x_k) - E(\hat{f}(x_k))| + 2C\frac{r_{x_k}}{h_n^2}.$$

Since for every $x \in G$ there exists $B(x_k, r_{x_k})$ that contains $x$,

$$\sup_{x \in G} |\hat{f}(x) - E(\hat{f}(x))| \leq \max_{1 \leq k \leq m} |\hat{f}(x_k) - E(\hat{f}(x_k))| + 2C\frac{r_{x_k}}{h_n^2}$$

Now, for every $n$, let $m := l_n$ and $r_{x_k} := r_n$ with $G \subset \cup_{k=1}^{l_n} B(x_k, r_n)$ with $r > l_n r_n$ or $l_n < r/r_n$. Letting $r_n = \frac{h_n^{3/2}}{n^{1/2}}$ we have

$$\frac{(nh_n)^{1/2}}{log(n)^{1/2}}\sup_{x \in G}|\hat{f}(x) - E(\hat{f}(x))| \leq \frac{(nh_n)^{1/2}}{log(n)^{1/2}}\max_{1 \leq k \leq l_n}|\hat{f}(x_k) - E(\hat{f}(x_k))| + \frac{2C}{log(n)^{1/2}}$$

and to prove the theorem it suffices to show that for all $\epsilon > 0$ there exists $\Delta_\epsilon > 0$ such that

$$P\left(\frac{(nh_n)^{1/2}}{log(n)^{1/2}} \max_{1\le k\le l_n} |\hat{f}(x_k) - E(\hat{f}(x_k))| > \Delta_\epsilon\right) < \epsilon \text{ for all } n.$$

By Lemma 11 in the Appendix, we have

$$P\left(\frac{(nh_n)^{1/2}}{log(n)^{1/2}} \max_{1\le k\le l_n} |\hat{f}(x_k) - E(\hat{f}(x_k))| > \Delta_\epsilon\right) \le \sum_{k=1}^{l_n} P\left(\frac{(nh_n)^{1/2}}{log(n)^{1/2}} |\hat{f}(x_k) - E(\hat{f}(x_k))| > \Delta_\epsilon\right).$$

Now, using Lemma 12 in the Appendix (Bernstein's Inequality) with $\Delta = \left(\frac{\log n}{nh_n}\right)^{1/2} \frac{\Delta_\epsilon}{\sigma_n}$, $W_{in} = \frac{1}{nh_n} K\left(\frac{X_i - x_k}{h_n}\right) - E\left(\frac{1}{nh_n} K\left(\frac{X_i - x_k}{h_n}\right)\right)$, we have that if $|K(x)| < C$, $|W_{in}| \le \frac{2C}{nh_n}$ and

$$\begin{aligned} P\left(\frac{(nh_n)^{1/2}}{(log(n))^{1/2}} |\hat{f}(x_k) - E(\hat{f}(x_k))| > \Delta_\epsilon\right) &< 2\exp\left(-\frac{\frac{\log n}{nh_n}\frac{\Delta_\epsilon^2}{\sigma_n^2}}{2 + \frac{2}{3}\frac{2C}{nh_n}\frac{1}{\sigma_n}\frac{\Delta_\epsilon}{\sigma_n}\left(\frac{\log n}{nh_n}\right)^{1/2}}\right) \\ &= 2\exp\left(-\frac{\frac{\log n}{nh_n}\Delta_\epsilon^2}{2\sigma_n^2 + \frac{2}{3}\frac{2C}{nh_n}\Delta_\epsilon\left(\frac{\log n}{nh_n}\right)^{1/2}}\right) \\ &= 2\exp\left(-\frac{\log n\ \Delta_\epsilon^2}{2nh_n\sigma_n^2 + \frac{4}{3}C\Delta_\epsilon\left(\frac{\log n}{nh_n}\right)^{1/2}}\right) \end{aligned}$$

where $\sigma_n^2 = V(\hat{f}(x_k)) = \frac{1}{nh_n}\int K^2(v) f(x_k + h_n v)dv - \frac{1}{n}\left(\int K(v) f(x_k + h_n v)dv\right)^2$. Since $f$ is uniformly continuous on $G$ (continuity on a compact set implies uniform continuity), $nh_n\sigma_n^2 \to f(x_k)\int K^2(v)dv$ for all $x_k$ and, provided that $\frac{nh_n}{\log n} \to \infty$, we have $c_n := 2nh_n\sigma_n^2 + \frac{4}{3}C\Delta_\epsilon\left(\frac{\log n}{nh_n}\right)^{1/2} \to 2f(x_k)\int K^2(v)dv$. Now,

$$\begin{aligned} P\left(\frac{(nh_n)^{1/2}}{(log(n))^{1/2}} \max_{1\le k\le l_n} |\hat{f}(x_k) - E(\hat{f}(x_k))| > \Delta_\epsilon\right) &< l_n\, 2\exp\left(-\log n\ \frac{\Delta_\epsilon^2}{c_n}\right) \\ &< 2\frac{n^{1/2}}{h_n^{3/2}} r n^{-\Delta_\epsilon^2/c_n} \text{ since } l_n < \tfrac{r}{r_n} = \tfrac{n^{1/2}}{h_n^{3/2}} \\ &= 2\left(\frac{1}{nh_n^3}\right)^{1/2} n^{1-\frac{\Delta_\epsilon^2}{c_n}} r \to 0, \end{aligned}$$

provided $\Delta_\epsilon^2 > 2f(x_k)\int K^2(v)dv$ and $nh_n^3 \to \infty$. □

**Remark 10.** *1. From Theorem 13, we can conclude that*

$$\frac{1}{h_n}\sup_{x\in G}|\hat{f}(x) - E(\hat{f}(x))| = \left(\frac{\log n}{nh_n^3}\right)^{1/2} O_p(1).$$

*Consequently, if* $\frac{nh_n^3}{log(n)} \to \infty$ *then* $\frac{1}{h_n}\sup_{x\in G}|\hat{f}(x) - E(\hat{f}(x))| = o_p(1)$ *or* $\sup_{x\in G}|\hat{f}(x) - E(\hat{f}(x))| = o_p(h_n)$. *Note that if* $h_n \propto n^{\delta}$, *then* $nh_n^3 \to \infty$ *requires that* $\delta > -\frac{1}{3}$ *and in this case* $\frac{nh_n^3}{\log n} \to \infty$.

*2. If* $\frac{nh_n^3}{log(n)} \to \infty$, *then by the triangle inequality*

$$\begin{aligned}\frac{1}{h_n}\sup_{x\in G}|\hat{f}(x) - f(x)| &\leq \frac{1}{h_n}\sup_{x\in G}|\hat{f}(x) - E(\hat{f}(x))| + \frac{1}{h_n}\sup_{x\in G}|E(\hat{f}(x)) - f(x))| \\ &= o_p(1) + \frac{1}{h_n}\sup_{x\in G}|E(\hat{f}(x)) - f(x))|.\end{aligned}$$

*If* $f$ *satisfies a Lipschitz condition of order 1 in* $G$, *that is* $|f(x) - f(x')| < C|x - x'|$ *for all* $x, x' \in G$ *and* $\int K(u)du = 1$, *then* $|E(\hat{f}(x)) - f(x)| \leq \int |f(x + h_n u) - f(x)||K(u)|du \leq h_n C \int |u||K(u)|du$. *Hence, if* $\int |u||K(u)|du < \infty$, $\frac{1}{h_n}\sup_{x\in G}|E(\hat{f}(x)) - f(x))| = O(1)$ *and* $\sup_{x\in G}|\hat{f}(x) - f(x)| = h_n O_p(1)$.

*3. If* $f$ *satisfies the Lipschitz condition in remark 2 and* $\int uK(u)du < \infty$, *then for any given* $n$ *the mean squared error in Theorem 9 is bounded above as* $E\left((\hat{f}(x) - f(x))^2\right) \leq h_n^2 C_1 + \frac{1}{nh_n}C_2$. *The first term is a bound on the square of the bias and the second is a bound on the variance. The bias bound increases as* $h_n$ *increases whereas the variance bound diminishes as* $h_n$ *increases. This is the bias-variance tradeoff associated with the choice of* $h_n$ *(for fixed* $n$*). The bound on the right-hand side of the inequality is minimized at* $h_n^* = \left(\frac{C_2}{2C_1}\right)^{1/3} n^{-1/3}$. *Hence, substituting* $h_n^*$ *in the expression for the bound we have* $E\left((\hat{f}(x) - f(x))^2\right) = O(n^{-2/3})$.

*4. If* $h_n = Cn^{\delta}$ *with* $-1 < \delta < 0$ *then* $\frac{nh_n}{log(n)} = \frac{n^{1+\delta}C}{log(n)}$. *By L'Hôpital's rule and the fact that* $-1 < \delta < 0$ *we have* $\frac{nh_n}{log(n)} \to \infty$ *as* $n \to \infty$.

*5. If* $h_n = Cn^{\delta}$ *then* $nh_n^3 = n^{1+3\delta}C^3 \to \infty$ *if* $-1/3 < \delta$. *As in remark 3, by L'Hôpital's rule and the fact that* $-1/3 < \delta < 0$ *we have* $\frac{nh_n^3}{log(n)} \to \infty$ *as* $n \to \infty$.

Remarks 10.1 and 10.2 together with Theorem 13 provide the following corollary.

**Corollary 1.** *Assume a)* $|K(u)| \leq C$ *for all* $u \in \mathbb{R}$*; b)* $\int |K(u)|du < \infty$*; c)* $|u||K(u)| \to 0$ *as* $|u| \to \infty$*; d) for any* $u, u' \in \mathbb{R}$ *with* $u \neq u'$ *we have* $|K(u') - K(u)| < C|u' - u|$ *for some arbitrary constant* $C < \infty$ *(Lipschitz condition of order 1 in* $\mathbb{R}$*); e)* $\int |u||K(u)|du < \infty$*; f)* $h_n \to 0$ *and* $\frac{nh_n^3}{log(n)} \to \infty$ *as* $n \to \infty$*. Then, if* $f$ *is continuous and if it satisfies a Lipschitz condition of order 1 in* $G$*, a compact subset of* $\mathbb{R}$*,*

$$\frac{1}{h_n} \sup_{x \in G} |\hat{f}(x) - f(x)| = O_p(1).$$

# 2 Univariate density estimation over $[0, \infty)$

In practice, it is common to encounter non-negative random variables where the support of the density $f$ is the half-line $[0, \infty)$. In this case, the Rosenblatt kernel density estimator $\hat{f}(x)$ generally has bad properties. To see the source of the problem, suppose $f \in \Sigma(\beta, C)$, $K$ is even and of order $\lfloor \beta \rfloor$ with compact support $[-1, 1]$. Then, for $x \in [h_n, \infty)$ the bias of $\hat{f}(x)$ is given by

$$\begin{aligned} B(\hat{f}(x)) &= \int_{-1}^{x/h_n} K(u)(f(x - h_n u) - f(x))du - f(x) \int_{x/h_n}^{\infty} K(u)du \\ &= \int_{-1}^{1} K(u)(f(x - h_n u) - f(x))du - f(x) \times 0 \leq h_n^{\beta} C. \end{aligned}$$

But if $x \in [0, h_n)$ (or $x \in [0, n^{-1/(2\beta+1)})$) and for $\lambda \in (0, 1)$,

$$\begin{aligned} B(\hat{f}(x)) &= -f(x) \int_{x/h_n}^{1} K(u)du - h_n f^{(1)}(x) \int_{-1}^{x/h_n} uK(u)du + \cdots \\ &\quad + \frac{(-h_n)^{\lfloor \beta \rfloor}}{\lfloor \beta \rfloor !} \int_{-1}^{x/h_n} f^{(\lfloor \beta \rfloor)}(x - \lambda h_n u) u^{\lfloor \beta \rfloor} K(u)du \neq O(h_n^{\beta}) \\ V(\hat{f}(x)) &\leq \frac{C}{nh_n} \int_{-1}^{x/h_n} K^2(u)du. \end{aligned}$$

Unless $f(x) = 0$, $\hat{f}(x)$ is inconsistent. In the typical case where $\beta = 2$, even if $f(x) = 0$, the rate of decay of the bias in the vicinity of the boundary is $h_n$, rather than $h_n^2$, away from the boundary. Hence, the rate of decay of the mean squared error and the optimal bandwidth differ in the vicinity of the boundary. We will now define a new kernel density estimator that does not suffer from this boundary problem.

## 2.1 Hestenes extension and a new estimator for $f$

Let $w_1, ..., w_{s+1}$ be pairwise different positive numbers for $s = 0, 1, \cdots$. Of special interest are the decreasing sequence $w_i = 1/i$, $i = 1, ..., s+1$ (used by Hestenes (1941)) and the increasing sequence $w_i = i$. Let the numbers $k_1, ..., k_{s+1}$ be defined from the following system

$$\sum_{i=1}^{s+1}(-w_i)^j k_i = 1, \ j = 0, ..., s. \tag{8}$$

Since this system has the Van-der-Monde determinant

$$\begin{vmatrix} 1 & 1 & ... & 1 \\ -w_1 & -w_2 & ... & -w_{s+1} \\ ... & ... & ... & \\ (-w_1)^s & (-w_2)^s & ... & (-w_{s+1})^s \end{vmatrix} \neq 0,$$

$k_1, ..., k_{s+1}$ are uniquely defined. Let $s \in \mathbb{N}$ and $\Omega \subseteq \mathbb{R}$. The class of functions $f : \Omega \to \mathbb{R}$ which are $s$-times differentiable with $\left|f^{(s)}(x)\right| \leq C$ for some $0 < C < \infty$ and continuous is denoted by $\mathcal{C}_b^s(\Omega)$. If $f \in \mathcal{C}_b^s$ on its domain $D_f = [0, \infty)$, its Hestenes extension to $(-\infty, 0)$ is given by

$$\phi_s(x) = \sum_{j=1}^{s+1} k_j f(-w_j x), \ x < 0. \tag{9}$$

Note that if $f$ is a density function, $\phi_s$ is not a density, but a linear combination of densities $w_j f(-w_j x)$ with coefficients $k_j/w_j$. Assuming that $f$ has $s$ right-hand derivatives $f(0+), ..., f^{(s)}(0+)$ at zero ($s = 0$ means continuity), we see that the following sewing conditions at zero are satisfied due to (8):

$$\phi_s^{(m)}(0-) = \sum_{j=1}^{s+1}(-w_j)^m k_j f^{(m)}(0+) = f^{(m)}(0+), \ m = 0, 1, \cdots, s.$$

Now, define $g_s$ on $\mathbb{R}$ by

$$g_s(x) = \begin{cases} f(x), & x \geq 0 \\ \phi_s(x), & x < 0 \end{cases}, \tag{10}$$

with $g_s$ being $s$ times differentiable.

**Definition 7.** *Let $f \in \mathcal{C}_b^s$, $K$ be an even kernel and $\{X_i\}_{i=1}^n$ be an IID sequence of random variables with density $f$. We define the estimator $\hat{f}_s(x)$ of $f(x)$, for $x \geq 0$, by*

$$\hat{f}_s(x) = \frac{1}{nh_n}\sum_{i=1}^{n}\left[K\left(\frac{x - X_i}{h_n}\right) + \sum_{j=1}^{s+1}\frac{k_j}{w_j}K\left(\frac{x + X_i/w_j}{h_n}\right)\right]. \tag{11}$$

**Remark 11.** *1. When $s = 0$ in equation (11), $\hat{f}_0(x)$ is the "reflection estimator" from Schuster (1985), i.e.,*

$$\hat{f}_0(x) = \frac{1}{nh_n} \sum_{i=1}^{n} \left[ K\left( \frac{x - X_i}{h_n} \right) + K\left( \frac{x + X_i}{h_n} \right) \right].$$

*2. We note that Schuster's estimator does not depend on $s$, the index on $\mathcal{C}_b^s$. Thus, knowledge that $s > 0$ is not used in constructing his estimator, whereas it is central in the definition of $\hat{f}_s(x)$.*

**Theorem 14.** *Suppose $f \in \mathcal{C}_b^s$ on $D_f = [0, \infty)$ and $K$ is a kernel with $\int K(u)du = 1$. Then, the bias of $\hat{f}_s(x)$ has the representation*

$$E(\hat{f}_s(x)) - f(x) = \int_{\mathbb{R}} K(t) \left[ g_s(x - h_n t) - g_s(x) \right] dt, \ x \in D_f. \tag{12}$$

*Proof.* By the IID assumption

$$\begin{aligned} E\left(\hat{f}_s(x)\right) &= \frac{1}{h_n} E\left[ K\left( \frac{x - X_1}{h_n} \right) + \sum_{j=1}^{s+1} \frac{k_j}{w_j} K\left( \frac{x + X_1/w_j}{h_n} \right) \right] \\ &= \frac{1}{h_n} \left[ \int_0^\infty K\left( \frac{x - t}{h_n} \right) f(t)dt + \sum_{j=1}^{s+1} \frac{k_j}{w_j} \int_0^\infty K\left( \frac{x + t/w_j}{h_n} \right) f(t)dt \right]. \end{aligned} \tag{13}$$

In the first integral let $u = \frac{x-t}{h_n}$, in the others $u = \frac{x+t/w_j}{h_n}$. Then

$$\begin{aligned} E\left(\hat{f}_s(x)\right) &= \frac{1}{h_n} \left[ -h_n \int_{x/h_n}^{-\infty} K(u) f(x - h_n u)du + h_n \sum_{j=1}^{s+1} k_j \int_{x/h_n}^{\infty} K(u) f\left(-w_j(x - h_n u)\right) du \right] \\ &= \left[ \int_{-\infty}^{x/h_n} K(u) f(x - h_n u)du + \int_{x/h_n}^{\infty} K(u) \sum_{j=1}^{s+1} k_j f\left(-w_j(x - h_n u)\right) du \right]. \end{aligned}$$

In the first integral we have $x - h_n u > 0$ and $f(x - h_n u) = g_s(x - h_n u)$; in the second one $x - h_n u < 0$, so $\sum_{j=1}^{s+1} k_j f\left(-w_j(x - h_n u)\right) = g_s(x - h_n u)$. Hence,

$$E\left(\hat{f}_s(x)\right) = \int_{\mathbb{R}} K(u) g_s(x - h_n u)du. \tag{14}$$

Since $\int_{\mathbb{R}} K(t)dt = 1$, this implies (12). □

**Remark 12.** *The integral representation for bias obtained in Theorem 14 depends on the extension $g_s$, not the density $f$. Consequently, existing results for smooth functions (not densities) on $\mathbb{R}$ allow us to obtain bias estimates.*

**Theorem 15.** *Assume that* $\int K(t)dt = 1$, $\int_{\mathbb{R}} K(t)t^j dt = 0$, *for* $j = 1, ..., s-1$, $\int_{\mathbb{R}} |K(t)t^s|\, dt < \infty$ *and* $\left|f^{(s)}(x)\right| < C$ *for all* $x \geq 0$, *then*

$$E\left(\hat{f}_s(x)\right) - f(x) = O(h^s) \text{ for all } x \in D_f. \tag{15}$$

*Proof.* Since $\int K(t)dt = 1$ we have from Theorem 14

$$E(\hat{f}_s(x)) - f(x) = \int K(t)(g_s(x - h_n t) - g_s(x))dt = \int K(t)\left(g_s^{(1)}(x)(-h_n t) + \frac{1}{2!}g_s^{(2)}(x)(-h_n t)^2 + \cdots + \frac{1}{s!}g_s^{(s)}(x - th\tau)(-h_n t)^s\right)dt$$

for some $\tau \in (0, 1)$. If $\int_{\mathbb{R}} K(t)t^j dt = 0$, for $j = 1, ..., s-1$ then

$$|E(\hat{f}_s(x)) - f(x)| \leq \frac{h_n^s}{s!}\int |t|^s |K(t)||g^{(s)}(x - th\tau)|dt \leq Ch_n^s,$$

where the last inequality follows from the assumptions that $\int_{\mathbb{R}} |K(t)t^s|\, dt < \infty$, $\left|f^{(s)}(x)\right| < C$ for all $x \geq 0$ and the structure of $g^{(s)}$. □

**Remark 13.** *1. In the density estimation literature, it is usually assumed that* $f \in \mathcal{C}_b^2$. *In this case, under the conditions in Theorem 15 we have* $E\hat{f}_2(x) - f(x) = \frac{h_n^2}{2}f^{(2)}(x)\int_{\mathbb{R}} t^2 K(t)dt + o(h_n^2)$ *for all* $x \in D_f$. *This expression, similar to what is obtained for the classical Rosenblatt-Parzen estimator when* $D_f = \mathbb{R}$, *contrasts with what is obtained for Schuster's reflection estimator. In particular, for boundary points, viz.,* $x = ch$ *for* $0 \leq c < \infty$, *we have (see Marron and Ruppert (1994))*

$$\begin{aligned} E\hat{f}_0(x) - f(x) &= \int K(t)(g_0(x - h_n t) - g_0(x))dt = \int_{-\infty}^{x/h_n} K(t)f(x - h_n t)dt \\ &+ \int_{x/h_n}^{\infty} K(t)f(-(x - h_n t))dt - \int K(t)g_0(x)dt \\ &= \frac{h_n^2}{2}f^{(2)}(x)\int_{\mathbb{R}} t^2 K(t)dt - 2hf^{(1)}(x)\left(c\mu_{0,-c} + \mu_{1,-c}\right) \\ &+ 2h^2 f^{(2)}(x)(c^2\mu_{0,-c} + c\mu_{1,-c}) + o(h_n^2) \end{aligned}$$

*where* $\mu_{\ell,-c} = \int_{-\infty}^{-c} u^\ell K(u)du$ *for* $\ell = 0, 1, 2$. *The additional bias terms result from the fact that Schuster's estimator does not use the additional smoothness (beyond continuity) of* $f$ *in its construction.*

2. *Theorems 14 and 15, require the existence of densities and their derivatives up to order $s$ at $x = 0$. As such, in the case of densities that either diverge to infinity, or have derivatives that diverge to infinity near the boundary, the bias order we have derived do not hold. In particular, when $f(x) \to \infty$ as $x \to 0$ (a pole at $x = 0$) the bias diverges to infinity at $x = 0$. If $f(0)$ is finite, but $f^{(1)}(x) \to \infty$ as $x \to 0$, the bias decays at a speed that is slower than $h^s$.*

The following theorem provides the order of the variance for $\hat{f}_s(x)$. It is necessary to consider the boundary $x = 0$ and interior points ($x > 0$) separately. Let $k_0 = w_0 = -1$. The quantities

$$\Gamma = \int_{\mathbb{R}} K^2(t)dt \text{ and } \Gamma_l = \int_0^\infty \left[\sum_{j=0}^{s+1} \frac{k_j}{w_j} K\left(\frac{t}{w_j}\right)\right]^2 dt,$$

will appear in variance expressions inside the domain and at the left boundary, respectively.

**Theorem 16.** *Suppose the conditions in Theorem 15 hold,* $\sup_{x\geq 0} f(x) < \infty$, $\sup_x |K(x)| < \infty$ *and* $\int_{\mathbb{R}} |K(t)|\, dt < \infty$. *Then, I) for fixed* $x > 0$

$$V\left(\hat{f}_s(x)\right) = \frac{1}{nh}\left\{f(x)\Gamma + o(1)\right\}, \tag{16}$$

*II) at the left boundary*

$$V\left(\hat{f}_s(0)\right) = \frac{1}{nh}\left\{f(0)\Gamma_l + o(1)\right\}. \tag{17}$$

*Proof.* By the IID assumption, the variance of $\hat{f}_s(x)$ is given by $V(\hat{f}_s(x)) = \frac{1}{n}\left(E(u_1^2) - E(u_1)^2\right)$ where

$$u_1 = I_{\{X_1\geq 0\}} \frac{1}{h_n}\left[K\left(\frac{x - X_1}{h_n}\right) + \sum_{j=1}^{s+1} \frac{k_j}{w_j} K\left(\frac{x + X_1/w_j}{h_n}\right)\right].$$

Since $E(u_1) = E(\hat{f}_s(x)) \to f(x)$, we have that $\frac{1}{n}E(u_1)^2 = O(n^{-1})$. Now,

$$\begin{aligned}
\frac{1}{n}E(u_1^2) = &\frac{1}{nh_n}\left(\frac{1}{h_n}E\left(I_{\{X_1\geq 0\}}K^2\left(\frac{x - X_1}{h_n}\right)\right)\right. \\
&+ \frac{2}{h_n}\sum_{j=1}^{s+1} E\left(I_{\{X_1\geq 0\}}\frac{k_j}{w_j}K\left(\frac{x - X_1}{h_n}\right)K\left(\frac{x + X_1/w_j}{h_n}\right)\right) \\
&\left.+ \frac{1}{h_n}E\left(\sum_{j=1}^{s+1}\frac{k_j}{w_j}I_{\{X_1\geq 0\}}K\left(\frac{x + X_1/w_j}{h_n}\right)\right)^2\right) \\
= &\frac{1}{nh_n}\left(I_{1n} + I_{2n} + I_{3n}\right)
\end{aligned}$$

By a change of variable in integration

$$I_{1n} = \int_{-\infty}^{x/h_n} K^2(u) f(x - h_n u) du \to \begin{cases} f(0) \int_{-\infty}^{0} K^2(u) du, & \text{if } x = 0 \\ f(x) \int_{-\infty}^{\infty} K^2(u) du & \text{if } x > 0 \end{cases}$$

as $n \to \infty$.

$$\begin{aligned} I_{3n} &= \sum_{j=1}^{s+1} \sum_{l=1}^{s+1} \frac{k_j}{w_j} \frac{k_l}{w_l} \frac{1}{h_n} E\left( I_{\{X_1 \geq 0\}} K\left( \frac{x + X_1/w_j}{h_n} \right) K\left( \frac{x + X_1/w_l}{h_n} \right) \right) \\ &= \sum_{j=1}^{s+1} \sum_{l=1}^{s+1} \frac{k_j}{w_j} \frac{k_l}{w_l} I_{jln}. \end{aligned}$$

Now, for $j \neq l$ and $x > 0$, a change of variable $u = \frac{x + X_1/w_j}{h_n}$ we have

$$I_{jln} = w_j \int_{x/h_n}^{\infty} K(u) K\left( u \frac{w_j}{w_l} + \frac{x(1 - w_j/w_l)}{h_n} \right) f((h_n u - x) w_j) du.$$

Since $|K(u)| = o(1)$ as $u \to \infty$ and $h_n \to 0$, we have that $I_{jln} \to 0$ as $n \to \infty$. If $x = 0$, then

$$I_{jln} = w_j \int_{0}^{\infty} K(u) K\left( u \frac{w_j}{w_l} \right) f(h_n u w_j) du \to f(0) w_j \int_{0}^{\infty} K(u) K\left( u \frac{w_j}{w_l} \right) du.$$

For $j = l$ and $x > 0$

$$I_{jjn} = w_j \int_{x/h_n}^{\infty} K^2(u) f((h_n u - x) w_j) du \to 0$$

and for $x = 0$, $I_{jjn} \to w_j f(0) \int_0^{\infty} K^2(u) du$. In all,

$$I_{3n} = \begin{cases} o(1) & \text{if } x > 0 \\ \sum_{j=1}^{s+1} \frac{k_j^2}{w_j} f(0) \int_0^{\infty} K^2(u) du + f(0) \sum_{j=1}^{s+1} \sum_{l=1, l \neq j}^{s+1} \frac{k_l}{w_l} k_j \int_0^{\infty} K(u) K\left( u \frac{w_j}{w_l} \right) du & \text{if } x = 0 \end{cases}$$

Similar arguments give,

$$I_{2n} = \begin{cases} o(1) & \text{if } x > 0 \\ 2 f(0) \sum_{j=1}^{s+1} k_j \int_0^{\infty} K(u) K(-u w_j) du & \text{if } x = 0 \end{cases}.$$

Combining the orders and expressions for $I_{1n}$, $I_{2n}$, $I_{3n}$ we have, for $x > 0$

$$V(\hat{f}_s(x)) = \frac{f(x)}{n h_n} \int_{\mathbb{R}} K^2(u) du + o((n h_n)^{-1})$$

and for $x = 0$

$$V(\hat{f}_s(0)) = \frac{f(0)}{nh_n}\left(\int_{-\infty}^{0} K^2(u)du + \sum_{j=1}^{s+1}\frac{k_j^2}{w_j}\int_0^\infty K^2(u)du \right.$$
$$+\sum_{j=1}^{s+1}\sum_{l=1,l\neq j}^{s+1}\frac{k_l}{w_l}k_j\int_0^\infty K(u)K\left(u\frac{w_j}{w_l}\right)du$$
$$\left.+2\sum_{j=1}^{s+1}k_j\int_0^\infty K(u)K(-uw_j)du\right) + o((nh_n)^{-1}).$$

Now, since

$$\frac{k_0}{w_0}K\left(\frac{t}{w_0}\right) = K(-t), \qquad (k_0 = w_0 = -1),$$

and $w_j > 0$ for $j = 1, \ldots, s+1$ and

$$\Gamma_l \;=\; \int_0^\infty \left[\sum_{j=0}^{s+1}\frac{k_j}{w_j}K\left(\frac{t}{w_j}\right)\right]^2 dt.$$

we have

$$\Gamma_l = \int_0^\infty \left[K(-t) + \sum_{j=1}^{s+1}\frac{k_j}{w_j}K\left(\frac{t}{w_j}\right)\right]^2 dt = \sum_{j=0}^{s+1}\sum_{l=0}^{s+1}\frac{k_jk_l}{w_jw_l}\int_0^\infty K\left(\frac{t}{w_j}\right)K\left(\frac{t}{w_l}\right)\,dt.$$

Split the double sum into four groups.

**1. The $j = l = 0$ diagonal.** With the $j = 0$ term equal to $K(-t)$, substituting $v = -t$,

$$\int_0^\infty K(-t)^2\,dt \;\overset{v=-t}{=}\; \int_{-\infty}^{0} K^2(v)\,dv,$$

which is the first term $\displaystyle\int_{-\infty}^{0} K^2(u)\,du$.

**2. The $j = l \geq 1$ diagonals.** Substituting $u = t/w_j$ (so $dt = w_j\,du$),

$$\frac{k_j^2}{w_j^2}\int_0^\infty K^2\left(\frac{t}{w_j}\right)\,dt = \frac{k_j^2}{w_j}\int_0^\infty K^2(u)\,du,$$

and summing over $j = 1, \ldots, s+1$ yields the second term $\displaystyle\sum_{j=1}^{s+1}\frac{k_j^2}{w_j}\int_0^\infty K^2(u)\,du$.

**3. Off-diagonal pairs with both indices $\geq 1$.** For $j \neq l$, substituting $u = t/w_l$,

$$\frac{k_j k_l}{w_j w_l} \int_0^\infty K\left(\frac{t}{w_j}\right) K\left(\frac{t}{w_l}\right) dt = \frac{k_j k_l}{w_j} \int_0^\infty K(u)\, K\left(u\frac{w_l}{w_j}\right) du.$$

Relabelling the dummy pair $j \leftrightarrow l$ turns this into $\dfrac{k_l}{w_l}\, k_j \displaystyle\int_0^\infty K(u)\, K\left(u\frac{w_j}{w_l}\right) du$, i.e. exactly the third term.

**4. Cross pairs between $j = 0$ and $l \geq 1$.** These occur twice, as $(0, l)$ and $(l, 0)$, giving the factor 2. Using the $j = 0$ term $K(-t)$ and substituting $u = t/w_l$,

$$2\sum_{l=1}^{s+1} \frac{k_l}{w_l} \int_0^\infty K(-t)\, K\left(\frac{t}{w_l}\right) dt = 2\sum_{l=1}^{s+1} k_l \int_0^\infty K(u)\, K(-u\, w_l)\, du,$$

which is the last term.

□

**Remarks.** 1. It is a direct consequence of Theorem 23 that, if $\int |K(u)|^{2+\delta} du < C$ for some $\delta > 0$, by Liapounov's Central Limit Theorem we have for $x > 0$

$$\sqrt{nh}\left(\hat{f}_s(x) - E\left(\hat{f}_s(x)\right)\right) \xrightarrow{d} N\left(0, f(x)\Gamma\right), \tag{18}$$

and for $x = 0$,

$$\sqrt{nh}\left(\hat{f}_s(0) - E\left(\hat{f}_s(0)\right)\right) \xrightarrow{d} N\left(0, f(0)\Gamma_l\right). \tag{19}$$

2. For $f \in \mathcal{C}_b^2$, equation (19) together with the expression for $E(\hat{f}_2(x))$ gives

$$\sqrt{nh}\left(\hat{f}_2(0) - \left(f(0) + \frac{h^2}{2} f^{(2)}(0+) \int_{\mathbb{R}} t^2 K(t) dt + o(h^2)\right)\right) \xrightarrow{d} N\left(0, f(0)\Gamma_l\right),$$

which can be used to construct confidence intervals and conduct hypothesis testing.

# 3 Distribution function estimation over $\mathbb{R}$

It is natural to define an estimator for the distribution function at $F(x)$ at $x$ by integrating $\hat{f}$. Hence, we have

**Definition 8.** *Let $\hat{f}(x)$ be the kernel density estimator. The kernel distribution function estimator is given by*

$$\hat{F}(x) = \int_{-\infty}^{x} \hat{f}(z)dz. \tag{20}$$

**Theorem 17.** *Assume that $\{X_i\}_{i=1}^n$ is an IID sequence and that $K$ is a density function. Then,*

$$E(\hat{F}(x)) - F(x) = \int (F(x + h_n\gamma) - F(x))K(\gamma)d\gamma \tag{21}$$

*and*

$$V(\hat{F}(x)) = \frac{1}{n}\left(2\int F(x + h_n\gamma)(1 - G(\gamma))K(\gamma)d\gamma - \left(\int F(x + h_n\gamma)K(\gamma)d\gamma\right)^2\right) \tag{22}$$

*where $G(\gamma) = \int_{-\infty}^{\gamma} K(z)dz$.*

*Proof.* By definition $\hat{F}(x) = \frac{1}{nh_n}\sum_{i=1}^n \int_{-\infty}^{x} K\left(\frac{X_i - y}{h_n}\right) dy$. By a change of variable in integration we write $\hat{F}(x) = \frac{1}{nh_n}\sum_{i=1}^n h_n \int_{\frac{X_i - x}{h_n}}^{\infty} K(z)dz$. Consequently, by the assumption that the sequence $\{X_i\}_{i=1}^n$ is IID we have

$$\begin{aligned}
E(\hat{F}(x)) &= \frac{1}{n}\sum_{i=1}^{n} \int \int_{\frac{y-x}{h_n}}^{\infty} K(z)dz f(y)dy = \int \int_{\frac{y-x}{h_n}}^{\infty} K(z)dz f(y)dy \\
&= 1 - \int \int_{-\infty}^{\frac{y-x}{h_n}} K(z)dz f(y)dy \text{ given that } \int K(z)dz = 1 \\
&= 1 - h_n \int \int_{-\infty}^{\gamma} K(z)dz f(x + h_n\gamma)d\gamma \text{ by letting } \gamma = \tfrac{y-x}{h_n} \\
&= 1 - \int G(\gamma)dF(x + h_n\gamma)
\end{aligned}$$

Integrating by parts, we have $\int G(\gamma)dF(x+h_n\gamma) = \lim_{\gamma\to\infty} G(\gamma)F(x+h_n\gamma) - \lim_{\gamma\to-\infty} G(\gamma)F(x+h_n\gamma) - \int F(x+h_n\gamma)dG(\gamma) = 1 - \int F(x+h_n\gamma)K(\gamma)d\gamma$. Thus, $E(\hat{F}(x)) = \int F(x+h_n\gamma)K(\gamma)d\gamma$ and given that $\int K(\gamma)d\gamma = 1$

$$E(\hat{F}(x)) - F(x) = \int (F(x + h_n\gamma) - F(x))K(\gamma)d\gamma. \tag{23}$$

Now,

$$
\begin{aligned}
V(\hat{F}(x)) &= E\left(\left(\left(1-\frac{1}{n}\sum_{i=1}^{n} G\left(\frac{X_i-x}{h_n}\right)\right)-\left(1-E\left(G\left(\frac{X_i-x}{h_n}\right)\right)\right)\right)^2\right) \\
&= E\left(\frac{1}{n^2}\left(\sum_{i=1}^{n}\left(G\left(\frac{X_i-x}{h_n}\right)-E\left(G\left(\frac{X_i-x}{h_n}\right)\right)\right)\right)^2\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E\left(\left(G\left(\frac{X_i-x}{h_n}\right)-E\left(G\left(\frac{X_i-x}{h_n}\right)\right)\right)\right. \\
&\times \left.\left(G\left(\frac{X_j-x}{h_n}\right)-E\left(G\left(\frac{X_j-x}{h_n}\right)\right)\right)\right) \\
&= \frac{1}{n}V\left(G\left(\frac{X_i-x}{h_n}\right)\right) = \frac{1}{n}\left(E\left(G^2\left(\frac{X_i-x}{h_n}\right)\right)-\left(E\left(G\left(\frac{X_i-x}{h_n}\right)\right)\right)^2\right)
\end{aligned}
$$

where the penultimate equality follows from the assumption that $\{X_i\}_{i=1}^n$ is IID. Using integration by parts and a change of variable as in obtaining the expression for $E(\hat{F}(x))$ we have

$$E\left(G^2\left(\frac{X_i-x}{h_n}\right)\right) = 1 - 2\int F(x+h_n\gamma)K(\gamma)G(\gamma)d\gamma.$$

Hence, given that $E\left(G\left(\frac{X_i-x}{h_n}\right)\right) = 1-\int F(x+h_n\gamma)K(\gamma)d\gamma$ we have

$$V(\hat{F}(x)) = \frac{1}{n}\left(2\int F(x+h_n\gamma)(1-G(\gamma))K(\gamma)d\gamma - \left(\int F(x+h_n\gamma)K(\gamma)d\gamma\right)^2\right). \quad (24)$$

□

**Remarks.** 1. Note that $E(\hat{F}(x)) - F(x)$ is expressed with the same algebraic structure as the right-hand side of $E(\hat{f}(x)) - f(x) = \int (f(x+h_n\gamma) - f(x))K(\gamma)d\gamma$ with $F$ replacing $f$. In the case of $\hat{f}$ we used Theorem 7, which required $\int |f(\gamma)|d\gamma < \infty$, to show that $E(\hat{f}(x)) - f(x) \to 0$ as $n \to \infty$. In the case of $\hat{F}$, it is not generally true that $\int F(\gamma)d\gamma < \infty$, thus we cannot make use of Theorem 7. Additional assumptions are needed to study the behavior of the bias and the variance of $\hat{F}$.

**Theorem 18.** *Assume that a) $\{X_i\}_{i=1}^n$ is an IID sequence; b) $K$ is a density function with $\int \gamma K(\gamma)d\gamma = 0$ and $\int \gamma^2 K(\gamma)d\gamma < \infty$; c) $F(x)$ has bounded second derivative on $\mathbb{R}$ and $F^{(1)}(x)$ is continuous on $\mathbb{R}$; d) $0 < h_n \to 0$ as $n \to \infty$. Then, $E(\hat{F}(x)) = F(x) + O(h_n^2)$ and $V(\hat{F}(x)) = \frac{F(x)(1-F(x))}{n} - \frac{h_n}{n}f(x)\alpha + o\left(\frac{h_n}{n}\right)$ where $\alpha = 2\int \gamma K(\gamma)G(\gamma)d\gamma$.*

*Proof.* By c), for any $x \in \mathbb{R}$ and some $0 \leq \lambda \leq 1$ we have $F(x+h_n\gamma) = F(x) + F^{(1)}(x)h_n\gamma + \frac{1}{2}F^{(2)}(x+\lambda h_n\gamma)h_n^2\gamma^2$. Then, from Theorem 17

$$\int F(x+h_n\gamma)K(\gamma)d\gamma = F(x) + \frac{h_n^2}{2}\int F^{(2)}(x+\lambda h_n\gamma)\gamma^2 K(\gamma)d\gamma = F(x) + O(h_n^2)$$

where the last integral exists given that $F^{(2)}$ is bounded and $\int \gamma^2 K(\gamma)d\gamma < \infty$. Similarly,

$$\begin{aligned}
\int 2F(x+h_n\gamma)K(\gamma)(1-G(\gamma))d\gamma &= F(x)\int 2K(\gamma)(1-G(\gamma))d\gamma \\
&\quad + h_n F^{(1)}(x)\int 2\gamma K(\gamma)(1-G(\gamma))d\gamma \\
&\quad + h_n^2 \int F^{(2)}(x+\lambda h_n\gamma)(1-G(\gamma))\gamma^2 K(\gamma)d\gamma \\
&= 2F(x) - F(x)\,2\int G(\gamma)K(\gamma)d\gamma \\
&\quad + 2h_n F^{(1)}(x)\int \gamma K(\gamma)d\gamma - h_n F^{(1)}(x)2\int \gamma K(\gamma)G(\gamma)d\gamma \\
&\quad + h_n^2 \int F^{(2)}(x+\lambda h_n\gamma)\gamma^2 K(\gamma)d\gamma \\
&\quad - h_n^2 \int F^{(2)}(x+\lambda h_n\gamma)\gamma^2 K(\gamma)G(\gamma)d\gamma.
\end{aligned}$$

Note that all integrals on the right hand side of the last equality exist since: integrating by parts we have that $2\int G(\gamma)K(\gamma)d\gamma = 1$, by b) $\int \gamma K(\gamma)d\gamma = 0$, $\left|\int \gamma K(\gamma)G(\gamma)d\gamma\right| \leq \int |\gamma|K(\gamma)d\gamma < \infty$ by b) and the fact that $G(\gamma) \leq 1$, $\left|\int F^{(2)}(x+\lambda h_n\gamma)\gamma^2 K(\gamma)G(\gamma)d\gamma\right| \leq C\int \gamma^2 K(\gamma)d\gamma < \infty$ by b), c) and $G(\gamma) \leq 1$. Hence, noting that $F^{(1)}(x) = f(x)$ and letting $\alpha = 2\int \gamma K(\gamma)G(\gamma)d\gamma$ we have

$$\int 2F(x+h_n\gamma)K(\gamma)(1-G(\gamma))d\gamma = F(x) - h_n f(x)\alpha + O(h_n^2). \tag{25}$$

Consequently, $E(\hat{F}(x)) - F(x) = O(h_n^2)$ and

$$\begin{aligned}
V(\hat{F}(x)) &= \frac{1}{n}\left(F(x) - h_n f(x)\alpha + O(h_n^2) - (F(x) + O(h_n^2))^2\right) \\
&= \frac{F(x)(1-F(x))}{n} - \frac{h_n}{n}f(x)\alpha + \frac{h_n^2}{n}O(1) + \frac{h_n^4}{n}O(1) \\
&= \frac{F(x)(1-F(x))}{n} - \frac{h_n}{n}f(x)\alpha + o\left(\frac{h_n}{n}\right) \text{ since } h_n \to 0.
\end{aligned}$$

□

**Remarks.** 1. Under the conditions of Theorem 18 $\hat{F}(x)$ converges in quadratic mean to $F(x)$ and consequently $\hat{F}(x) \xrightarrow{p} F(x)$ for all $x$.

2. Different from the case of density estimation, there is no requirement that $nh_n \to \infty$ in the case of distribution estimation.

3. Given that $\{X_i\}_{i=1}^n$ is IID, the empirical distribution function $F_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_i \leq x)$ is such that $E(F_n(x)) = F(x)$, $V(F_n(x)) = \frac{1}{n}F(x)(1-F(x))$ and using Lévy's central limit theorem $\sqrt{n}(F_n(x) - F(x)) \xrightarrow{d} N(0, F(x)(1-F(x))$. Hence, the use of $\hat{F}(x)$ as an estimator for $F(x)$ introduces a bias term compared to $F_n(x)$. However, since $V(F_n(x)) - V(\hat{F}(x)) = \frac{h_n}{n}f(x)\alpha - o\left(\frac{h_n}{n}\right)$, it is possible to obtain a reduction of the variance. Consequently, $E((\hat{F}(x) - F(x))^2)$ could be smaller than $E((F_n(x) - F(x))^2)$. An advantage of $\hat{F}(x)$ over $F_n(x)$ is its smoothness.

4. $E((\hat{F}(x) - F(x))^2) \leq \frac{F(x)(1-F(x))}{n} - \frac{h_n}{n}f(x)\alpha + o\left(\frac{h_n}{n}\right) + h_n^4 C$. Choosing $h_n$ to minimize the bound on the righthand side of the inequality we obtain $h_n^* = n^{-1/3}\left(\frac{f(x)\alpha}{4C}\right)^{1/3}$. This rate of decay for $h_n$ is different from that obtained for the estimation of $f$ when $\beta \geq 2$ in Theorem 12.

**Theorem 19.** *Assume that a) $\{X_i\}_{i=1}^n$ is an IID sequence; b) $K$ is a density function with $\int \gamma K(\gamma)d\gamma = 0$ and $\int \gamma^2 K(\gamma)d\gamma < \infty$; c) $F(x)$ has bounded second derivative on $\mathbb{R}$ and $F^{(1)}(x)$ is continuous on $\mathbb{R}$; d) $0 < h_n \to 0$ as $n \to \infty$. Then,*

$$\sqrt{n}(\hat{F}(x) - E(\hat{F}(x))) \xrightarrow{d} N(0, F(x)(1-F(x))). \tag{26}$$

*Proof.* $-(\hat{F}(x) - E(\hat{F}(x))) = \frac{1}{n}\sum_{i=1}^n \left(G\left(\frac{X_i - x}{h_n}\right) - E\left(G\left(\frac{X_i - x}{h_n}\right)\right)\right)$. Let $Z_{ni} = \frac{1}{n}G(\frac{X_i - x}{h_n})$, $E(Z_{ni}) = \mu_n$ and $s_n^2 = \sum_{i=1}^n E(Z_{ni} - \mu_n)^2$. Note that

$$s_n^2 = \sum_{i=1}^n \frac{1}{n^2}E\left(G\left(\frac{X_i - x}{h_n}\right) - E\left(G\left(\frac{X_i - x}{h_n}\right)\right)\right)^2 = \frac{1}{n}V\left(G\left(\frac{X_i - x}{h_n}\right)\right) = V\left(\hat{F}(x)\right).$$

Hence, $\frac{\hat{F}(x) - E(\hat{F}(x))}{\sqrt{V(\hat{F}(x))}} = \sum_{i=1}^n \left(\frac{Z_{ni} - \mu_n}{s_n}\right) = \sum_{i=1}^n X_{ni}$ with $E(X_{ni}) = 0$, $E(X_{ni}^2) = \frac{1}{s_n^2}E(Z_{ni} - \mu_n)^2$ and $\sum_{i=1}^n E(X_{in}^2) = 1$. Consequently, by Liapounov's CLT (see Davidson (1994), p. 372)

$$\frac{\hat{F}(x) - E(\hat{F}(x))}{\sqrt{V(\hat{F}(x))}} \xrightarrow{d} N(0,1)$$

provided $lim_{n\to\infty}\sum_{i=1}^{n} E|X_{ni}|^{2+\delta} = 0$ for some $\delta > 0$. We now verify that this limit is in fact zero.

$$\begin{aligned}\sum_{i=1}^{n} E|X_{ni}|^{2+\delta} &= \sum_{i=1}^{n} E\left(\left|\frac{Z_{ni}-\mu_n}{s_n}\right|^{2+\delta}\right) = \sum_{i=1}^{n} V(\hat{F}(x))^{-1-\frac{\delta}{2}} E\left(|Z_{ni}-\mu_n|^{2+\delta}\right) \\ &= V(\hat{F}(x))^{-(1+\frac{\delta}{2})} nE\left(|Z_{ni}-\mu_n|^{2+\delta}\right)\end{aligned}$$

By the $c_r$ inequality and the fact that by Theorem 18 $\mu_n = O(n^{-1})$, we have $E\left(|Z_{ni}-\mu_n|^{2+\delta}\right) \leq 2^{1+\delta}E\left(|Z_{ni}|^{2+\delta}\right)$. Therefore,

$$\begin{aligned}\sum_{i=1}^{n} E|X_{ni}|^{2+\delta} &\leq V(\hat{F}(x))^{-(1+\frac{\delta}{2})} n2^{1+\delta} E\left(|Z_{nt}|^{2+\delta}\right) \\ &= V(\hat{F}(x))^{-(1+\frac{\delta}{2})} 2^{1+\delta} n^{-1-\delta} E\left(\left|G\left(\frac{X_t-x}{h_n}\right)\right|^{2+\delta}\right) \\ &\leq V(\hat{F}(x))^{-(1+\frac{\delta}{2})} n^{-1-\delta} \text{ since } |G(x)| \leq 1 \text{ and } \int f(\gamma)d\gamma = 1 \\ &= (nV(\hat{F}(x)))^{-1-\frac{\delta}{2}} n^{-\delta/2}\end{aligned}$$

By Theorem 18 as $n \to \infty$, $nV(\hat{F}) \to F(x)(1-F(x))$ and since $n^{-\delta/2} \to 0$ for all $\delta > 0$ we have $\sum_{i=1}^{n} E|X_{ni}|^{2+\delta} = o(1)$. Since $V(\hat{F}(x)) - \frac{F(x)(1-F(x))}{n} = -\frac{h_n}{n}f(x)\alpha + o\left(\frac{h_n}{n}\right) = o(1)$ we have

$$\frac{\hat{F}(x) - E(\hat{F}(x))}{\sqrt{\frac{F(x)(1-F(x))}{n}}} \xrightarrow{d} N(0,1). \tag{27}$$

□

**Remark.** $\sqrt{n}(\hat{F}(x) - F(x)) = \sqrt{n}(\hat{F}(x) - E(\hat{F}(x))) + \sqrt{n}(E(\hat{F}(x)) - F(x)) = \sqrt{n}(\hat{F}(x) - E(\hat{F}(x))) + n^{1/2}h_n^2 O(1)$, where the last equality follows from Theorem 18. Hence, if $n^{1/2}h_n^2 \to 0$ we have $\sqrt{n}(\hat{F}(x) - F(x)) \xrightarrow{d} N(0, F(x)(1-F(x)))$. If $h_n = h_n^*$ on remark 4 following Theorem 18 then $n^{1/2}(h_n^*)^2 = n^{-1/6}C \to 0$ as $n \to \infty$ and $\sqrt{n}(\hat{F}(x) - F(x)) \xrightarrow{d} N(0, F(x)(1-F(x)))$. Hence, under suitable choice for the order of the bandwidth, $F_n$ and $\hat{F}$ have the same asymptotic distribution.

# 4 Estimation of discontinuous distributions and their jumps

We will now consider the estimation of distribution functions that are continuous, but not absolutely continuous, or even discontinuous at $x$. We first study the case where $F$ is

continuous but does not have a density at $x$. In what follows, since $F$ is a distribution function, we take it to be right-continuous, $L_1$ denotes the space of functions on $\mathbb{R}$ with a finite norm $\|f\|_{L_1} = \int_{\mathbb{R}} |f(t)|dt$, $C_b$ is the space of uniformly bounded continuous functions on $\mathbb{R}$ with norm $\|f\|_C = \sup_{t\in\mathbb{R}} |f(t)|$.

## 4.1 Estimation at points of continuity of $F$

For $x$ a point of continuity of $F$, we start by defining a general class of estimators for $F(x)$. Given a function $U$ and a random sample $\{X_i\}_{i=1}^n$ from $F$ we consider estimators defined as

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} U\left(\frac{X_i - x}{h}\right) \tag{28}$$

where $h := h_n > 0$ is a bandwidth, or smoothing parameter, satisfying $h \to 0$ as $n \to \infty$. In addition, we place the following restriction on $U$.

**Assumption 1.** *$U \in C$, $\lim_{x\to-\infty} U(x) = 1$ and $\lim_{x\to\infty} U(x) = 0$.*

If $K : \mathbb{R} \to \mathbb{R}$ is integrable with $\int_{-\infty}^{\infty} K(t)dt = 1$ and we set $U(x) = \int_x^{\infty} K(t)dt$, then it follows that $U$ satisfies Assumption 1. In addition, if $K$ is an even function, then $\hat{F}(x) = \frac{1}{n}\sum_{i=1}^n L\left(\frac{x-X_i}{h}\right)$ where $L(x) = \int_{-\infty}^x K(t)dt$.

Letting

$$\phi_U(N) = \max\left\{\sup_{z<-N} |U(z) - 1|, \sup_{z>N} |U(z)|\right\}$$

for $N > 0$, we note that Assumption 1 implies $\phi_U(N) \to 0$ as $N \to \infty$. Also, letting $\omega(x, \epsilon) \equiv \sup_{0<h\leq\epsilon} (F(x + h) - F(x - h))$, we note that if $x$ is a continuity point of $F$, then $\omega(x, \epsilon) \to 0$ as $\epsilon \to 0$. We now state the following lemma.

**Lemma 3.** *If $U$ satisfies Assumption 1 and $\delta \in (0, 1)$, then for all $h > 0$*

$$\left|\int_{\mathbb{R}} U\left(\frac{y-x}{h}\right) dF(y) - F(x)\right| \leq \phi_U(h^{\delta-1}) + \omega(x, h^\delta)(1 + \|U\|_C).$$

*Proof.* Since $U$ is bounded and continuous and $F$ is of bounded variation, $\int_{\mathbb{R}} U\left(\frac{y-x}{h}\right) dF(y)$ exists (Natanson, 1955, p. 238). For $\epsilon > 0$ we have

$$\begin{aligned}\left|\int_{\mathbb{R}} U\left(\frac{y-x}{h}\right) dF(y) - \int_{-\infty}^{x} dF(y)\right| \leq& \left|\int_{y<x-\epsilon} \left[U\left(\frac{y-x}{h}\right) - 1\right] dF(y)\right| \\ &+ \left|\int_{|x-y|\leq\epsilon} U\left(\frac{y-x}{h}\right) dF(y)\right| \\ &+ \left|\int_{y>x+\epsilon} U\left(\frac{y-x}{h}\right) dF(y)\right| + \int_{x-\epsilon}^{x} dF(y).\end{aligned}$$

Then, we obtain the following bounds

$$\left|\int_{y<x-\epsilon}\left[U\left(\frac{y-x}{h}\right)-1\right]dF(y)\right| \leq \sup_{z<-\epsilon/h}|U(z)-1|\int_{y<x-\epsilon}dF(y) \leq \phi_U(\epsilon/h)F(x)$$

$$\left|\int_{|x-y|\leq\epsilon}U\left(\frac{y-x}{h}\right)dF(y)\right| \leq \|U\|_C\left[F(x+\epsilon)-F(x-\epsilon)\right] \leq \|U\|_C\,\omega(x,\epsilon),$$

$$\left|\int_{y>x+\epsilon}U\left(\frac{y-x}{h}\right)dF(y)\right| \leq \sup_{z>\epsilon/h}|U(z)|\int_{y>x+\epsilon}dF(y) \leq \phi_U(\epsilon/h)(1-F(x)),$$

and $\int_{x-\epsilon}^{x}dF(y)\leq\omega(x,\epsilon)$. Collecting the above inequalities we have

$$\left|\int_{\mathbb{R}}U\left(\frac{y-x}{h}\right)dF(y)-F(x)\right| \leq \phi_U(\epsilon/h)+(1+\|U\|_C)\,\omega(x,\epsilon).$$

To finish the proof, we put $\epsilon=h^\delta$.

□

Lemma 3 is the basis for the following theorem.

**Theorem 20.** *Suppose $U$ satisfies Assumption 1 and let $x$ be a continuity point of $F$. Then,*
*a)* $\left|E(\hat{F}(x))-F(x)\right| \leq \phi_U(h^{\delta-1})+\omega(x,h^\delta)(1+\|U\|_C)$,
*b) If* $F(x)\in(0,1)$ *then* $nV(\hat{F}(x))=F(x)(1-F(x))+o(1)$.

*Proof.* a) follows directly from Lemma 3 since

$$\left|E\left(\hat{F}(x)\right)-F(x)\right| = \left|\int_{\mathbb{R}}U\left(\frac{t-x}{h}\right)dF(t)-F(x)\right|.$$

b) We have $V(\hat{F}(x))=\frac{1}{n}\left\{E\left(U^2\left(\frac{X-x}{h}\right)\right)-\left[E\left(U\left(\frac{X-x}{h}\right)\right)\right]^2\right\}$. From part a) we know that $E\left(U\left(\frac{X-x}{h}\right)\right)=F(x)+o(1)$ as $h\to0$. $U^2$ satisfies Assumption 1, so by Lemma 3

$$\left|E\left(U^2\left(\frac{X-x}{h}\right)\right)-F(x)\right| \leq \phi_{U^2}(h^{\delta-1})+\omega(x,h^\delta)(1+\left\|U^2\right\|_C)\to0,\ h\to0.$$

If $F(x)\in(0,1)$ this proves the statement. □

The following corollary is a direct consequence of Theorem 20.

**Corollary 2.** *If $U$ satisfies Assumption 1 and $F\in C$, then*

$$\sup_{x\in\mathbb{R}}\omega(x,\epsilon)\to0,\ \ as\ \epsilon\to0, \tag{29}$$

*and* $\sup_{x\in\mathbb{R}}\left|E(\hat{F}(x))-F(x)\right| \leq \phi_U(h^{\delta-1})+\sup_{x\in\mathbb{R}}\omega(x,h^\delta)(1+\|U\|_C)\to0,\ h\to0.$

*Proof.* Since $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$, the variation of $F$ on $(-\infty, -N]$ and $[N, \infty)$ will be small for large $N > 0$. On $[-N, N]$, $F$ is uniformly continuous and its variation on segments $[x - \epsilon, x + \epsilon]$ will be small for small $\epsilon$ and all $x \in [-N, N]$. Thus (29) is true and the statement follows from Theorem 20 a). □

The following theorem gives the asymptotic distribution of $\hat{F}(x)$ under suitable normalization and centering.

**Theorem 21.** *Suppose $U$ satisfies Assumption 1 and $x$ is a point of continuity of $F$ such that $F(x) \in (0, 1)$. For $\delta \in (0, 1)$, if $h \to 0$, $\sqrt{n}\,\phi_U(h^{\delta-1}) \to 0$ and $\sqrt{n}\,\omega(x, h^\delta) \to 0$ as $n \to \infty$, we have*

$$\sqrt{n}(\hat{F}(x) - F(x)) \xrightarrow{d} N(0, (1 - F(x))F(x)). \tag{30}$$

*Proof.* Let $Z_{in} = \frac{1}{n} U\left(\frac{X_i - x}{h}\right)$, $\mu_n = E(Z_{in})$ and $s_n^2 = \sum_{i=1}^n E(Z_{in} - \mu_n)^2$. Then, $s_n^2 = V(\hat{F}(x))$ and

$$\frac{\hat{F}(x) - E(\hat{F}(x))}{\sqrt{V(\hat{F}(x))}} = \sum_{i=1}^n \frac{Z_{in} - \mu_n}{s_n} = \sum_{i=1}^n X_{in}$$

where $X_{in} = \frac{Z_{in} - \mu_n}{s_n}$, $E(X_{in}) = 0$, $V(X_{in}) = \frac{1}{s_n^2} E(Z_{in} - \mu_n)^2$ and $\sum_{i=1}^n V(X_{in}) = 1$. By Liapounov's Central Limit Theorem $\sum_{i=1}^n X_{in} \xrightarrow{d} \mathcal{Z} \sim N(0, 1)$ provided $\lim_{n\to\infty} \sum_{i=1}^n E\left(|X_{in}|^{2+\theta}\right) = 0$ for some $\theta > 0$ (see Davidson, 1994, p. 369-372). It follows from the fact that $\{X_i\}_{i=1,2,\cdots}$ is assumed to be independent and identically distributed that

$$\sum_{i=1}^n E\left(|X_{in}|^{2+\theta}\right) = nV(\hat{F}(x))^{-1-\theta/2} E(|Z_{in} - \mu_n|^{2+\theta})$$

and by the $c_r$ inequality $E(|Z_{in} - \mu_n|^{2+\theta}) \le 2^{1+\theta}(E(|Z_{in}|^{2+\theta}) + |\mu_n|^{2+\theta})$. From Theorem 20, if $h \to 0$ we have that $n|\mu_n|^{2+\theta} = \frac{1}{n^{1+\theta}} |F(x) + o(1)|^{2+\theta}$. Consequently,

$$\begin{aligned}
\sum_{i=1}^n E\left(|X_{in}|^{2+\theta}\right) &\le nV(\hat{F}(x))^{-1-\theta/2} 2^{1+\theta} \left( \frac{1}{n^{2+\theta}} E\left( \left| U\left(\frac{X_1 - x}{h}\right)\right|^{2+\theta}\right) + |\mu_n|^{2+\theta} \right) \\
&= (nV(\hat{F}(x)))^{-1-\theta/2} 2^{1+\theta} n^{-\theta/2} \left( E\left( \left| U\left(\frac{X_1 - x}{h}\right)\right|^{2+\theta}\right) + |F(x) + o(1)|^{2+\theta} \right) \\
&\to 0 \text{ as } n \to \infty
\end{aligned}$$

since $nV(\hat{F}(x)) \to F(x)(1 - F(x))$ and $E\left(\left|U\left(\frac{X_1 - x}{h}\right)\right|^{2+\theta}\right) = O(1)$ by the fact that $U^{2+\theta}$ satisfies assumption 1. Now, noting that $\sqrt{n}(\hat{F}(x) - F(x)) = \sqrt{n}(\hat{F}(x) - E(\hat{F}(x))) +$

$\sqrt{n}(E(\hat{F}(x)) - F(x))$ we have that $\sqrt{n}(\hat{F}(x) - F(x)) \xrightarrow{d} N(0, (1 - F(x))F(x))$ provided that $\sqrt{n}(E(\hat{F}(x)) - F(x)) \to 0$. But from Theorem 20,

$$|\sqrt{n}(E(\hat{F}(x)) - F(x))| \leq \sqrt{n}\phi_U(h^{\delta-1}) + \sqrt{n}\,\omega(x, h^\delta)(1 + ||U||_C).$$

Since, $\sqrt{n}\,\phi_U(h^{\delta-1}) \to 0$ and $\sqrt{n}\,\omega(x, h^\delta) \to 0$ by assumption, the proof is complete. □

**Remark.** The requirements that $\sqrt{n}\,\phi_U(h^{\delta-1}) \to 0$ and $\sqrt{n}\,\omega(x, h^\delta) \to 0$ can always be satisfied by an appropriate choice of the sequence $\{h_n\}$. If $U(x) = \int_x^\infty K(t)dt$ and $K$ has compact support, then $\sqrt{n}\,\phi_U(h^{\delta-1}) \equiv 0$ for any $\delta \in (0, 1)$ and all $0 < h < h(\delta)$. If $K$ has compact support, the choice of the bandwidth is dictated solely by the condition that $\sqrt{n}\,\omega(x, h^\delta) \to 0$. If, in addition to continuity at $x$, $F$ satisfies a local Lipschitz condition of order $\alpha > 0$, we have that for some positive constant $M_x$ and $u \in (0, h^\delta]$

$$F(x + u) - F(x - u) < 2M_x u^\alpha \leq 2M_x h^{\alpha\delta}$$

and $\sqrt{n}\,\omega(x, h^\delta) \leq 2M_x\sqrt{n}\,h^{\alpha\delta} \to 0$ if $nh^{2\alpha\delta} = o(1)$. If, as usual, $h \propto n^{-\gamma}$ for $\gamma > 0$ then $nh^{2\alpha\delta} = o(1)$ if $\gamma > 1/(2\alpha\delta)$.

Now, we turn to the estimation of the probability $F(x, y) := F(y) - F(x)$ for $x < y$. From Theorem 20 we can derive the rate of convergence of $\hat{F}(y) - \hat{F}(x)$ to $F(x, y)$. However, the variance of $\hat{F}(y) - \hat{F}(x)$ cannot be easily obtained from Theorem 20. To obtain it we make the following assumption.

**Assumption 2.** *$G = G([a, b])$ is a bounded continuous function of intervals $[a, b] \subset \mathbb{R}$ or, more precisely, of an ordered pair $(a, b)$, where $a < b$ with*

$$\lim_{b\to-\infty} G([a, b]) = 0, \quad \lim_{a\to-\infty,\ b\to\infty} G([a, b]) = 1 \text{ and } \lim_{a\to\infty} G([a, b]) = 0.$$

Define $\phi_G(N) = \max\left\{\sup_{b<-N} |G([a, b])|, \sup_{a>N} |G([a, b])|, \sup_{a<-N,\ b>N} |G([a, b]) - 1|\right\}$ and note that Assumption 2 implies that $\phi_G(N) \to 0$ as $N \to \infty$. If a kernel $K$ satisfies $\int_{\mathbb{R}} K(t)dt = 1$, then $G([a, b]) = \int_a^b K(t)dt$ satisfies Assumption 2. Besides, if $U_1$, $U_2$ satisfy Assumption 1, then $G([a, b]) = U_1(a) - U_2(b)$, $-\infty < a < b < \infty$, satisfies Assumption 2. We estimate $F(x, y)$ by

$$\hat{F}(x, y) = \frac{1}{n}\sum_{i=1}^n G\left(\left[\frac{X_i - y}{h}, \frac{X_i - x}{h}\right]\right).$$

**Lemma 4.** *If $G$ satisfies Assumption 2 and $\delta \in (0, 1)$, then for all $h > 0$*

$$\left|\int_{\mathbb{R}} G\left(\left[\frac{z - y}{h}, \frac{z - x}{h}\right]\right) dF(z) - F(x, y)\right| \leq \phi_G(h^{\delta-1}) + (1 + \|G\|_C)\left[\omega(x, h^\delta) + \omega(y, h^\delta)\right].$$

*Proof.* We put $\left[\frac{z-y}{h}, \frac{z-x}{h}\right] \equiv [a,b]$ and write

$$\begin{aligned}\int_{\mathbb{R}} G([a,b])\,dF(z) - \int_x^y dF(z) = &\int_{z<x-\epsilon} G([a,b])\,dF(z) + \int_{x+\epsilon}^{y-\epsilon} [G([a,b])-1]\,dF(z)\\ &+ \int_{z>y+\epsilon} G([a,b])\,dF(z) + \int_{|z-x|\le\epsilon} G([a,b])\,dF(z)\\ &+ \int_{|z-y|\le\epsilon} G([a,b])\,dF(z) + \int_x^{x+\epsilon} dF(z) + \int_{y-\epsilon}^{y} dF(z).\end{aligned}$$

In the first integral $\frac{z-y}{h} < \frac{z-x}{h} < -\frac{\epsilon}{h}$, so

$$\left|\int_{z<x-\epsilon} G([a,b])\,dF(z)\right| \le \sup_{b<-\epsilon/h} G([a,b])\,F(x) \le \phi_G(\epsilon/h)F(x).$$

In the second integral $\frac{z-y}{h} < -\frac{\epsilon}{h}$, $\frac{z-x}{h} > \frac{\epsilon}{h}$ and

$$\left|\int_{x+\epsilon}^{y-\epsilon} [G([a,b])-1]\,dF(z)\right| \le \sup_{a<-\epsilon/h,\ b>\epsilon/h} |G([a,b])-1|(F(y)-F(x)) \le \phi_G(\epsilon/h)(F(y)-F(x)).$$

In the third integral $\frac{z-x}{h} > \frac{z-y}{h} > \frac{\epsilon}{h}$ and

$$\left|\int_{z>y+\epsilon} G([a,b])\,dF(z)\right| \le \sup_{a>\epsilon/h} G([a,b])\,(1-F(y)) \le \phi_G(\epsilon/h)(1-F(y)).$$

The sum of the remaining four integrals is obviously bounded by $(1+\|G\|_C)\,[\omega(x,\epsilon)+\omega(y,\epsilon)]$. Thus, to finish the proof it remains to add the bounds and put $\epsilon = h^\delta$. □

**Theorem 22.** *Suppose $G$ satisfies Assumption 2, $x < y$ are continuity points of $F$ and let $\delta \in (0,1)$. Then*

*a)* $\left|E\hat{F}(x,y) - F(x,y)\right| \le \phi_G(h^{\delta-1}) + (1+\|G\|_C)\left[\omega(x,h^\delta)+\omega(y,h^\delta)\right]$ *for $h > 0$.*

*b) If $F(x,y) \in (0,1)$ then $V(\hat{F}(x,y)) = \frac{1}{n}\left[F(x,y)(1-F(x,y)) + o(1)\right]$.*

*Proof.* Part a) immediately follows from Lemma 4 because

$$E\left(\hat{F}(x,y)\right) = \int_{\mathbb{R}} G\left(\left[\frac{z-y}{h}, \frac{z-x}{h}\right]\right) dF(z).$$

Part b) follows from

$$V(\hat{F}(x,y)) = \frac{1}{n}\left\{EG^2\left(\left[\frac{X-y}{h}, \frac{X-x}{h}\right]\right) - \left[EG\left(\left[\frac{X-y}{h}, \frac{X-x}{h}\right]\right)\right]^2\right\}$$

where $G^2$ satisfies Assumption 2 and, by Lemma 4,

$$\left|EG^2\left(\left[\frac{X-y}{h}, \frac{X-x}{h}\right]\right) - F(x,y)\right| \le \phi_{G^2}(h^{\delta-1}) + (1+\left\|G^2\right\|_C)\left[\omega(x,h^\delta)+\omega(y,h^\delta)\right] \to 0,$$

as $h \to 0$. □

# 5 Estimation of jumps of a distribution function

Let the jump of $F$ at point $x$ be given by $p_x := F(x) - F(x-0)$, where $F(x-0) \equiv \lim_{\epsilon \downarrow 0} F(x-\epsilon)$. We estimate it by

$$\hat{p}_x = \frac{1}{n} \sum_{i=1}^{n} W\left(\frac{X_i - x}{h}\right), \tag{31}$$

where the function $W$ satisfies the following assumption.

**Assumption 3.** *The function $W$ is bounded and continuous, satisfies $W(0) = 1$ and $\lim_{|x| \to \infty} W(x) = 0$.*

We let $\omega_W(\epsilon) = \sup_{|x| \leq \epsilon} |W(x) - 1|$, $\phi_W(N) = \sup_{|x| \geq N} |W(x)|$, $\delta_F(\epsilon) = \int_{|z-x|<\epsilon} dF(z) - p_x \geq 0$ and note that Assumption 3 implies $\omega_W(\epsilon) \to 0$ as $\epsilon \to 0$ and $\phi_W(N) \to 0$ as $N \to \infty$. Since $\bigcap_{h>0} (x-h, x+h) = \{x\}$, by continuity of probability $\lim_{\epsilon \to 0} \delta_F(\epsilon) = 0$.

Denote by $\tilde{f}$ the exponential Fourier transform of $f$. Thus, for a distribution function $F$ we have $\tilde{F}(t) = \int_{\mathbb{R}} e^{itz} dF(z)$ and for the kernel $K$ we have $\tilde{K}(t) = \int_{\mathbb{R}} e^{itz} K(z) dz$. If $K$ is a kernel, its exponential Fourier transform $\tilde{K}$ satisfies Assumption 3.

**Lemma 5.** *a) Let $\epsilon_1 \in (0,1)$, $\epsilon_2 > \epsilon_1$. For any $\lambda \in (0,1)$*

$$\left| \int_{\mathbb{R}} W\left(\frac{z-x}{\lambda}\right) dF(z) - p_x \right| \leq \omega_W(\lambda^{\epsilon_2 - \epsilon_1}) \left[p_x + \delta_F(\lambda^{1-\epsilon_1})\right] + (1 + \|W\|_C)\, \delta_F(\lambda^{1-\epsilon_1}) + \phi_W(\lambda^{-\epsilon_1}).$$

*b) If $x$ is an isolated point of the support of $F$, that is, $\int_{|z-x|<h} dF(z) = p_x$ for all small $h > 0$, then for all small $\lambda > 0$*

$$\left| \int_{\mathbb{R}} W\left(\frac{z-x}{\lambda}\right) dF(z) - p_x \right| \leq \phi_W(\lambda^{-\epsilon_1}).$$

*In particular, when $W$ has compact support, $\int_{\mathbb{R}} W\left(\frac{z-x}{\lambda}\right) dF(z) = p_x$ for all small $\lambda$.*

*Proof.* a) By additivity of probability for any $h_1 \in (0, h)$

$$p_x \leq \int_{|z-x|<h_1} dF(z) \leq \int_{|z-x|<h} dF(z) = p_x + \delta_F(h). \tag{32}$$

This implies

$$\int_{h_1 \leq |z-x|<h} dF(z) = \int_{|z-x|<h} dF(z) - \int_{|z-x|<h_1} dF(z) \leq \delta_F(h), \tag{33}$$

$$\left|\int_{|z-x|<h_1} dF(z) - p_x\right| \le \delta_F(h). \tag{34}$$

We wish to bound

$$\int_{\mathbb{R}} W\left(\frac{z-x}{\lambda}\right) dF(z) - p_x = \int_{|z-x|<h_1} W\left(\frac{z-x}{\lambda}\right) dF(z) - p_x + \int_{h_1\le|z-x|<h} W\left(\frac{z-x}{\lambda}\right) dF(z) + \int_{|z-x|\ge h} W\left(\frac{z-x}{\lambda}\right) dF(z). \tag{35}$$

Obviously,

$$\left|\int_{|z-x|\ge h} W\left(\frac{z-x}{\lambda}\right) dF(z)\right| \le \phi_W(h/\lambda). \tag{36}$$

By (32) and (34)

$$\begin{aligned}\left|\int_{|z-x|<h_1} W\left(\frac{z-x}{\lambda}\right) dF(z) - p_x\right| &\le \left|\int_{|z-x|<h_1}\left[W\left(\frac{z-x}{\lambda}\right) - W(0)\right] dF(z)\right| \\ &+ \left|W(0)\int_{|z-x|<h_1} dF(z) - p_x\right| \\ &\le \omega_W(h_1/\lambda)\int_{|z-x|<h_1} dF(z) + \left|\int_{|z-x|<h_1} dF(z) - p_x\right| \\ &\le \omega_W(h_1/\lambda)\left[p_x + \delta_F(h)\right] + \delta_F(h). \end{aligned} \tag{37}$$

By (33)

$$\left|\int_{h_1\le|z-x|<h} W\left(\frac{z-x}{\lambda}\right) dF(z)\right| \le \|W\|_C \int_{h_1\le|z-x|<h} dF(z) \le \|W\|_C\, \delta_F(h). \tag{38}$$

Combining (36)-(38) we get

$$\left|\int_{\mathbb{R}} W\left(\frac{z-x}{\lambda}\right) dF(z) - p_x\right| \le \omega_W(h_1/\lambda)\left[p_x + \delta_F(h)\right] + \left(1 + \|W\|_C\right)\delta_F(h) + \phi_W(h/\lambda).$$

It remains to put here $h = \lambda^{1-\epsilon_1}$, $h_1 = h\lambda^{\epsilon_2} = \lambda^{1+\epsilon_2-\epsilon_1} < h$ to finish the proof.
b) Instead of (35) write

$$\int_{\mathbb{R}} W\left(\frac{z-x}{\lambda}\right) dF(z) - p_x = \int_{|z-x|<h} W\left(\frac{z-x}{\lambda}\right) dF(z) - p_x + \int_{|z-x|\ge h} W\left(\frac{z-x}{\lambda}\right) dF(z).$$

For all small $h$ $\int_{|z-x|<h} W\left(\frac{z-x}{\lambda}\right) dF(z) = W(0)p_x = p_x$. By (36) we see that the statement in the enunciation of the lemma is true. □

**Theorem 23.** *Suppose $W$ satisfies Assumption 3.*

*a) Let $\epsilon_1 \in (0,1)$, $\epsilon_2 > \epsilon_1$. Then, for any $h \in (0,1)$,*

$$|E(\hat{p}_x) - p_x| \leq \omega_W(h^{\epsilon_2-\epsilon_1}) \left[p_x + \delta_F(h^{1-\epsilon_1})\right] + (1 + \|W\|_C)\, \delta_F(h^{1-\epsilon_1}) + \phi_W(h^{-\epsilon_1}).$$

*b) If $x$ is an isolated point of the support of $F$, then for all small $h$*

$$|E(\hat{p}_x) - p_x| \leq \phi_W(h^{-\epsilon_1}).$$

*In particular, when $W$ has compact support, $\hat{p}_x$ is unbiased for all small $h$.*

*c) If $p_x \in (0,1)$ then $V(\hat{p}_x) = \frac{1}{n}\left[p_x(1-p_x) + o(1)\right],\ h \to 0.$*

*Proof.* a) and b) follow from Lemma 5. c) obtains from a), given that

$$V(\hat{p}_x) = \frac{1}{n}\left\{EW^2\left(\frac{X-x}{h}\right) - \left[EW\left(\frac{X-x}{h}\right)\right]^2\right\}$$

and the fact that $W^2$ satisfies Assumption 3. □

**Remark.** 1. Suppose that the moments $\int_{\mathbb{R}} t^j K(t)dt$ are zero for $j = 1, ..., m-1$, and the derivative $\tilde{K}^{(m)}$ is bounded. Then, since $\tilde{K}^{(j)}(0) = 0$ for $j = 1, ..., m-1$, by Taylor's Theorem

$$\tilde{K}\left(\frac{z-x}{\lambda}\right) - \tilde{K}(0) \;=\; \tilde{K}^{(m)}(\theta)\left(\frac{z-x}{\lambda}\right)^m \frac{1}{m!}.$$

Using this equation in (37) we get

$$\begin{aligned}\left|\int_{|z-x|<h_1} \tilde{K}\left(\frac{z-x}{\lambda}\right) dF(z) - p_x\right| \;&\leq\; \frac{\sup\left|\tilde{K}^{(m)}\right|}{m!}\left(\frac{h_1}{\lambda}\right)^m \left|\int_{|z-x|<h_1} dF(z)\right| \\ &+\; \left|\int_{|z-x|<h_1} dF(z) - p_x\right| \\ &\leq\; c\lambda^{m(\epsilon_2-\epsilon_1)}\left[p_x + \delta_F(\lambda^{1-\epsilon_1})\right] + \delta_F(\lambda^{1-\epsilon_1})\end{aligned}$$

with the same choice of $h_1, h$ as before. The resulting bound

$$|E(\hat{p}_x) - p_x| \;\leq\; c\lambda^{m(\epsilon_2-\epsilon_1)}\left[p_x + \delta_F(\lambda^{1-\epsilon_1})\right] + \left(1 + \|K\|_{L_1}\right)\delta_F(\lambda^{1-\epsilon_1}) + \sup_{|z|\geq\lambda^{-\epsilon_1}}\left|\tilde{K}(z)\right|$$

shows that imposing higher-order kernel conditions, without restricting the class of distribution functions, has a limited effect on convergence rate, because of the term $\delta_F(\lambda^{1-\epsilon_1})$. The following theorem gives asymptotic normality of $\hat{p}_x$ when suitably centered and normalized.

**Theorem 24.** *Suppose $W$ satisfies Assumption 3 and $F$ has a jump $p_x$ at $x$ where $p_x \in (0,1)$. For $\epsilon_1 \in (0,1)$ and $\epsilon_2 > \epsilon_1$, if $h \to 0$, $\sqrt{n}\,\phi_W(h^{-\epsilon_1}) \to 0$, $\sqrt{n}\,\omega_W(h^{\epsilon_2-\epsilon_1}) \to 0$ and $\sqrt{n}\delta_F(h^{1-\epsilon_1}) \to 0$ as $n \to \infty$, we have*

$$\sqrt{n}(\hat{p}_x - p_x) \xrightarrow{d} N(0, (1-p_x)p_x). \tag{39}$$

*Proof.* In the proof of Theorem 21, let $Z_{in} = \frac{1}{n} W\left(\frac{X_i - x}{h}\right)$ and proceed with the same arguments to conclude that $\sqrt{n}(\hat{p}_x - E(\hat{p}_x)) \xrightarrow{d} N(0, (1-p_x)p_x)$. Use Theorem 23 to obtain

$$|\sqrt{n}(E(\hat{p}_x) - p_x)| \leq \sqrt{n}\left(\omega_W(h^{\epsilon_2-\epsilon_1})\left[p_x + \delta_F(h^{1-\epsilon_1})\right] + (1 + \|W\|_C)\,\delta_F(h^{1-\epsilon_1}) + \phi_W(h^{-\epsilon_1})\right).$$

Hence, given the conditions on the enunciation of the theorem we have $\sqrt{n}(E(\hat{p}_x) - p_x) \to 0$ as $n \to \infty$.

□

**Remark.** The requirements that $\sqrt{n}\,\phi_W(h^{-\epsilon_1}) \to 0$, $\sqrt{n}\,\omega_W(h^{\epsilon_2-\epsilon_1}) \to 0$ as $n \to \infty$ always hold with appropriately chosen $\{h_n\}$. If $W(t) = \int_{\mathbb{R}} e^{itz} K(z)dz$ and $K$ is a kernel such that $W$ has compact support (see, e.g., $W$ and $K$ used in subsection 5.2), then $\sqrt{n}\,\phi_W(h^{-\epsilon_1}) \equiv 0$ for all small $h$. The condition that $\sqrt{n}\delta_F(h^{1-\epsilon_1}) \to 0$ can always be satisfied if $h_n$ tends to 0 sufficiently quickly. However, to obtain a tractable dependence of the bandwidth on $n$, we note that

$$\delta_F(h^{1-\epsilon_1}) = F(x + h^{1-\epsilon_1}) - \left(F(x - h^{1-\epsilon_1}) + p_x\right)$$

and if a Lipschitz condition of order $\alpha$ is imposed on the difference on the righthand side of the equality, we have $\sqrt{n}\delta_F(h^{1-\epsilon_1}) < 2M_x\sqrt{n}h^{(1-\epsilon_1)\alpha})$. As in Remark 1, if $h \propto n^{-\gamma}$, it suffices to choose $\gamma > 1/(2(1-\epsilon_1)\alpha)$ to satisfy the condition.

# 6 Multivariate density estimation over $\mathbb{R}$

Let $X(\omega) : \Omega \to \mathbb{R}^D$ be a random vector defined on $(\Omega, \mathcal{F}, P)$ for $D \in \mathbb{N}$ and consider a random sample of observations $\{X_i\}_{i=1}^n$ on $X$. We assume that $F(x) = \int_{X \leq x} f(u)du$, where $x \in \mathbb{R}^D$ and $X \leq x$ is taken element-wise. As in the univariate case $f$ is called the density function of $X$.

**Definition 9.** *Let $H_n = diag\{h_{d,n}\}_{d=1}^D$ be a nonstochastic matrix with $0 < h_{d,n}$ for all $d$ and $n$, $x \in \mathbb{R}^D$, $K : \mathbb{R}^D \to \mathbb{R}$ be a measurable function. Then, we define*

$$\hat{f}(x) = \frac{1}{ndet(H_n)} \sum_{i=1}^{n} K\left(H_n^{-1}(X_i - x)\right)$$

*where $det(H_n)$ denotes the determinant of the matrix $H_n$ and $H_n^{-1}$ denotes its inverse.*

**Remark.** Since $0 < h_{d,n}$ for all $d$, $det(H_n) > 0$ and $H_n^{-1}$ exists. Under the assumption that $K : \mathbb{R}^D \to \mathbb{R}$ is a measurable function and that $E\left|K\left(H_n^{-1}(X_i - x)\right)\right| < \infty$ and $E\left(K\left(H_n^{-1}(X_i - x)\right)^2\right) < \infty$ it is straightforward to prove a multivariate version of Theorem 6. Hence, we have

$$E(\hat{f}(x)) = \int K(\gamma) f(x + H_n\gamma)d\gamma$$

and

$$V(\hat{f}(x)) = \frac{1}{ndet(H_n)} \int K^2(\gamma) f(x + H_n\gamma)d\gamma - \frac{1}{n}\left(\int K(\gamma) f(x + H_n\gamma)d\gamma\right)^2.$$

The integral sign in both expression are $D$-fold Riemann integrals. We now state an extension of Theorem 7. Throughout let $||x||_1 = \sum_{i=1}^D |x_i|$ be the norm of the vector $x$ and recall that by the Cauchy-Schwarz Inequality $||x||_1 \leq D^{1/2}||x||_E$.

**Theorem 25.** *Assume that a) $\int_{\mathbb{R}^D} |K(\phi)|d\phi < \infty$; b) $||\phi||_1 |K(\phi)| \to 0$ as $||\phi||_1 \to \infty$; c) $0 < h_{d,n} \to 0$ for all $d$ as $n \to \infty$. Let $f(x) : \mathbb{R}^D \to \mathbb{R}$ such that d) $\int_{\mathbb{R}^D} |f(\phi)|d\phi < \infty$. Then, for every continuity point $x$ of $f$,*

$$det(H_n)^{-1} E\left(K\left(H_n^{-1}(X_i - x)\right)\right) \to f(x) \int K(\phi)d\phi \text{ as } n \to \infty. \tag{40}$$

The remark that follows Theorem 7 applies to the multivariate case and therefore if $|K(\phi)| < C$

$$det(H_n)^{-1} E\left(K^r\left(H_n^{-1}(X_i - x)\right)\right) \to f(x) \int K^r(\phi)d\phi$$

as $n \to \infty$ for $r > 1$. Hence, we have the following multivariate version of Theorem 8.

**Theorem 26.** *Let $\{X_i\}_{i=1}^n$ be a sequence of IID random vectors with density $f$. Assume that a) $|K(\gamma)| \leq M$ for all $\gamma \in \mathbb{R}^D$; b) $\int |K(\gamma)|d\gamma < \infty$ and $\int K(\gamma)d\gamma = 1$; c) $||\gamma||_1|K(\gamma)| \to 0$ as $||\gamma||_1 \to \infty$; d) $h_{d,n} > 0$ for all $n$ and $h_{d,n} \to 0$ as $n \to \infty$. Then, $E(\hat{f}(x)) - f(x) \to 0$ as $n \to \infty$ and if $n\,det(H_n) \to \infty$ as $n \to \infty$, $V(\hat{f}(x)) \to 0$ as $n \to \infty$ for every $x$ which is a point of continuity of $f$.*

**Remark 14.** *1. Theorem 26 implies that $\hat{f}(x)$ converges in quadratic mean to $f(x)$ and consequently $\hat{f}(x) \xrightarrow{p} f(x)$.*

*2. If $f$ satisfies a Lipschitz condition of order 1 in $\mathbb{R}^D$, i.e., for all $x, x' \in \mathbb{R}^D$ with $x \neq x'$ we have that for some $C < \infty$, $|f(x) - f(x')| \leq C\|x - x'\|_1$. Then*

$$\begin{aligned}|E(\hat{f}(x)) - f(x)| &\leq \int |f(x + H_n\gamma) - f(x)||K(\gamma)|d\gamma \leq C \int \|H_n\gamma\|_1|K(\gamma)|d\gamma \\ &= C \int \left(\sum_{d=1}^{D} h_{n,d}|\gamma_d|\right)|K(\gamma)|d\gamma \\ &= C \sum_{d=1}^{D} h_{n,d} \int |\gamma_d||K(\gamma)|d\gamma \leq C\ trace(H_n) \text{ if } \int |\gamma_d||K(\gamma)|d\gamma < C.\end{aligned}$$

*The condition that $\int |\gamma_d||K(\gamma)|d\gamma < C$ is verified if $K \geq 0$ and $\int \gamma_d K(\gamma)d\gamma < C$.*

Let $D_{d_1,\cdots,d_m} f(x) = \frac{\partial^m}{\partial x_{d_1} \cdots \partial x_{d_m}} f(x)$ for $d_1, \cdots, d_m = 1, \cdots, D$ whenever the partial derivatives on the righthand side of the equality exist. We now define

**Definition 10.** *Let $T \subseteq \mathbb{R}^D$ and let $\beta$ and $C$ be two positive numbers. The Hölder class $\Sigma(\beta, C)$ on $T$ is defined as the set of $\lfloor\beta\rfloor$ times differentiable functions $f : T \to \mathbb{R}$ whose partial derivatives of order $\lfloor\beta\rfloor$ satisfy*

$$|D_{d_1,\cdots,d_{\lfloor\beta\rfloor}} f(x) - D_{d_1,\cdots,d_{\lfloor\beta\rfloor}} f(x')| \leq C||x - x'||_1^{\beta - \lfloor\beta\rfloor},\ \forall\, x \neq x',\ x, x' \in T.$$

**Theorem 27.** *Assume that $f \in \Sigma(\beta, C)$ for $\beta \geq 2$ and $K$ is a multivariate kernel of order $\lfloor\beta\rfloor$, i.e., $\int \gamma_{d_1} \cdots \gamma_{d_{\lfloor\beta\rfloor}} K(\gamma)d\gamma = 0$ for $d_1, \cdots, d_{\lfloor\beta\rfloor} = 1, \cdots, D$. Then, if $\int |K(\gamma)||\gamma_d|^\beta d\gamma < C$ for $d = 1, \cdots, D$*

$$E(\hat{f}(x)) - f(x) = O\left(\sum_{d=1}^{D} h_{n,d}^\beta\right) = O\left(trace(H_n^\beta)\right). \tag{41}$$

*Proof.* Since $f \in \Sigma(\beta, C)$, by Taylor's Theorem, for some $\lambda \in (0,1)$

$$\begin{aligned} f(x + H_n\gamma) - f(x) &= \sum_{d_1=1}^{D} D_{d_1} f(x) h_{n,d_1}\gamma_{d_1} + \frac{1}{2}\sum_{d_1=1}^{D}\sum_{d_2=1}^{D} D_{d_1,d_2} f(x) h_{n,d_1} h_{n,d_2}\gamma_{d_1}\gamma_{d_2} + \cdots \\ &+ \frac{1}{(\lfloor\beta\rfloor - 1)!}\sum_{d_1=1}^{D}\cdots\sum_{d_{\lfloor\beta\rfloor-1}=1}^{D} D_{d_1,\cdots,d_{\lfloor\beta\rfloor-1}} f(x) h_{n,d_1}\cdots h_{n,d_{\lfloor\beta\rfloor-1}}\gamma_{d_1}\cdots\gamma_{d_{\lfloor\beta\rfloor-1}} \\ &+ \frac{1}{\lfloor\beta\rfloor!}\sum_{d_1=1}^{D}\cdots\sum_{d_{\lfloor\beta\rfloor}=1}^{D} D_{d_1,\cdots,d_{\lfloor\beta\rfloor}} f(x + \lambda H_n\gamma) h_{n,d_1}\cdots h_{n,d_{\lfloor\beta\rfloor}}\gamma_{d_1}\cdots\gamma_{d_{\lfloor\beta\rfloor}} \end{aligned}$$

Since $K$ is of order $\lfloor\beta\rfloor$ and $f \in \Sigma(\beta, C)$,

$$\begin{aligned} E(\hat{f}(x)) - f(x) &= \int (f(x + H_n\gamma) - f(x))d\gamma \\ &= \frac{1}{\lfloor\beta\rfloor!}\sum_{d_1=1}^{D}\cdots\sum_{d_{\lfloor\beta\rfloor}=1}^{D}\int D_{d_1,\cdots,d_{\lfloor\beta\rfloor}} f(x + \lambda H_n\gamma) h_{n,d_1}\cdots h_{n,d_{\lfloor\beta\rfloor}}\gamma_{d_1}\cdots\gamma_{d_{\lfloor\beta\rfloor}} K(\gamma)d\gamma \\ &= \frac{1}{\lfloor\beta\rfloor!}\sum_{d_1=1}^{D}\cdots\sum_{d_{\lfloor\beta\rfloor}=1}^{D}\int \left(D_{d_1,\cdots,d_{\lfloor\beta\rfloor}} f(x + \lambda H_n\gamma) - D_{d_1,\cdots,d_{\lfloor\beta\rfloor}} f(x)\right) \\ &\times h_{n,d_1}\cdots h_{n,d_{\lfloor\beta\rfloor}}\gamma_{d_1}\cdots\gamma_{d_{\lfloor\beta\rfloor}} K(\gamma)d\gamma \end{aligned}$$

By the Triangle Inequality, $f \in \Sigma(\beta, C)$ and the fact that $\lambda \in (0,1)$, we have

$$\begin{aligned} |E(\hat{f}(x)) - f(x)| &\le \frac{C}{\lfloor\beta\rfloor!}\sum_{d_1=1}^{D}\cdots\sum_{d_{\lfloor\beta\rfloor}=1}^{D}\int \|\lambda H_n\gamma\|_1^{\beta-\lfloor\beta\rfloor} h_{n,d_1}\cdots h_{n,d_{\lfloor\beta\rfloor}}|\gamma_{d_1}|\cdots|\gamma_{d_{\lfloor\beta\rfloor}}||K(\gamma)|d\gamma \\ &= \frac{C}{\lfloor\beta\rfloor!}\sum_{d_1=1}^{D}\cdots\sum_{d_{\lfloor\beta\rfloor}=1}^{D}\int \sum_{d=1}^{D} h_{n,d}^{\beta-\lfloor\beta\rfloor}|\gamma_d|^{\beta-\lfloor\beta\rfloor} h_{n,d_1}\cdots h_{n,d_{\lfloor\beta\rfloor}}|\gamma_{d_1}|\cdots|\gamma_{d_{\lfloor\beta\rfloor}}||K(\gamma)|d\gamma \end{aligned}$$

where the last inequality follows from the fact that $h_{d,n}, |\gamma_d| > 0$ and $\beta - \lfloor\beta\rfloor \le 1$. The largest power on $h_{n,d}$ occurs when $d_1 = \cdots = d_{\lfloor\beta\rfloor} = d$ and we have

$$|E(\hat{f}(x)) - f(x)| \le \frac{CD}{\lfloor\beta\rfloor!}\sum_{d=1}^{D} h_{n,d}^{\beta}\int |\gamma_d|^{\beta}|K(\gamma)|d\gamma = O\left(\sum_{d=1}^{D} h_{n,d}^{\beta}\right)$$

provided $\int |K(\gamma)||\gamma|^{\beta}d\gamma < C$ and $h_n > 0$. □

The following theorem is an extension of Theorem 10 to the multivariate setting.

**Theorem 28.** *Assume that: a)*$\{X_i\}_{i=1}^n$ *forms an IID sequence of random vectors with density* $f$*; b)* $|K(\gamma)| \leq M$ *for all* $\gamma \in \mathbb{R}^D$*; c)* $\int |K(\gamma)|d\gamma < \infty$*; d)* $||\gamma||_1|K(\gamma)| \to 0$ *as* $||\gamma||_1 \to \infty$*; e)* $\int K(\psi)^{2+\delta}d\psi < \infty$ *for some* $\delta > 0$*; f)* $h_{d,n} > 0$ *for all* $n$*,* $h_{d,n} \to 0$ *and* $ndet(H_n) \to \infty$ *as* $n \to \infty$*. Then,*

$$(ndet(H_n))^{\frac{1}{2}}(\hat{f}(x) - E(\hat{f}(x))) \xrightarrow{d} N\left(0, f(x)\int K^2(\psi)d\psi\right),$$

*where* $x$ *is a point of continuity of* $f$*.*

**Remark 15.** *1. Theorem 28 shows that one of the consequences of estimating multivariate densities is that the rate of convergence in distribution slows (exponentially) as* $D$ *grows. This is known as the "curse" of dimensionality.*

*2. Note that under the conditions of Theorem 27*

$$\begin{aligned}(n\det(H_n))^{\frac{1}{2}}(\hat{f}(x) - f(x)) &= (n\det(H_n))^{\frac{1}{2}}(\hat{f}(x) - E(\hat{f}(x))) \\ &\quad + (n\det(H_n))^{\frac{1}{2}}(E(\hat{f}(x)) - f(x)) \\ &= (n\det(H_n))^{\frac{1}{2}}(\hat{f}(x) - E(\hat{f}(x))) + (n\det(H_n))^{\frac{1}{2}}O\left(\sum_{d=1}^D h_{n,d}^{\beta}\right).\end{aligned}$$

*Hence, if* $(n\det(H_n))^{\frac{1}{2}}\sum_{d=1}^D h_{n,d}^{\beta} \to 0$ *as* $n \to \infty$*,*

$$(ndet(H_n))^{\frac{1}{2}}(\hat{f}(x) - f(x)) \xrightarrow{d} N\left(0, f(x)\int K^2(\psi)d\psi\right).$$

*3. If, in the previous item,* $h_{n,d} = h_n$ *for all* $d$*, then*

$$(n\det(H_n))^{\frac{1}{2}}\sum_{d=1}^D h_{n,d}^{\beta} = D(nh_n^{D+2\beta})^{1/2} = o(1) \text{ if } nh_n^{D+2\beta} = o(1).$$

*If* $h \propto n^{\delta}$*, since* $n\det H_n \to \infty$ *and* $nh_n^{D+2\beta} = o(1)$ *as* $n \to \infty$ *we need*

$$-\frac{1}{D} < \delta < -\frac{1}{2\beta + D}$$

*to conclude that*

$$(nh_n^D)^{\frac{1}{2}}(\hat{f}(x) - f(x)) \xrightarrow{d} N\left(0, f(x)\int K^2(\psi)d\psi\right).$$

# 7 Bandwidth Choice

## 7.1 Minimizing Integrated Squared Error

It should be apparent by now that bandwidth choice is an important part of nonparametric density or distribution estimation. Intuitively, we would like to select a sequence $h_n := h$ so that a suitably defined distance between $\hat{f}(x)$ and $f(x)$ is as small as possible over the set where the random variable $X$ takes values. We are here searching for a global criterion for the choice of $h$.

Distances in function spaces can be defined in various ways, subject to some mathematical technicalities. Here, we search for $h$ that will minimize

$$\int_{\mathbb{R}} (f(x) - \hat{f}(x))^2 dx. \tag{42}$$

In the statistical literature this integral is called the Integrated Square Error ($ISE$) of $\hat{f}$, and we write $ISE(\hat{f})$. Note that $ISE(\hat{f})$ does not depend on where the estimator $\hat{f}$ is being evaluated, but it does depend on the sample $\{X_i\}_{i=1}^n$, hence $ISE(\hat{f})$ is itself a random variable. It is clear that minimization of $ISE(\hat{f})$ with respect to $h$ is equivalent to the minimization of $\int \hat{f}^2(x)dx - 2\int f(x)\hat{f}(x)dx$.

From the definition of $\hat{f}(x)$ and taking $K$ to be an even function,

$$\int_{\mathbb{R}} \hat{f}^2(x)dx = \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n \frac{1}{h_n}(K \star K)\left(\frac{X_i - X_j}{h_n}\right)$$

where $(K \star K)(u) = \int_{\mathbb{R}} K(u-t)K(t)dt$.

The integral $\int f(x)\hat{f}(x)dx$ clearly depends on $f$, which is unknown. Hence, to render the criterion for choosing $h$ operational we need a suitable approximation for $\int_{\mathbb{R}} f(x)\hat{f}(x)dx$. Note that,

$$\begin{aligned}\int_{\mathbb{R}} f(x)\hat{f}(x)dx &= \frac{1}{n}\sum_{i=1}^n \int_{\mathbb{R}} \frac{1}{h_n} K\left(\frac{X_i - x}{h_n}\right) f(x)dx \\ &= \frac{1}{n}\sum_{i=1}^n Z_{ni}\end{aligned}$$

and we define

$$\theta_n := E(Z_{ni}) = \int_{\mathbb{R}}\int_{\mathbb{R}} \frac{1}{h} K\left(\frac{\alpha - x}{h}\right) f(x)f(\alpha)dxd\alpha = E\left(\frac{1}{h}K\left(\frac{\alpha - x}{h}\right)\right),$$

where the expectation is taken with respect to the joint density of two independent and identically distributed random variables. Let $g_n(x,y) = \frac{1}{h}K\left(\frac{x-y}{h}\right)$ and define

$$U_n := \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\frac{1}{h}K\left(\frac{X_i - X_j}{h}\right) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} g_n(X_i, X_j).$$

$U_n$ is a U-statistics of order 2 with $n$-dependent symmetric kernel given by $g_n$. Symmetry follows from the evenness of $K$, and $E(g_n(X_1,X_2)) = \theta_n$. Now, by Hoeffding's decomposition

$$U_n = \theta_n + \underbrace{\frac{2}{n}\sum_{i=1}^{n} g_{1,n}(X_i)}_{L_n} + \underbrace{\frac{1}{n(n-1)}\sum_{i\neq j} g_{2,n}(X_i,X_j)}_{R_n}.$$

where $g_{1,n}(x) = E\, g_n(x,X_2) - \theta_n$ and $g_{2,n}(x,y) = g_n(x,y) - g_{1,n}(x) - g_{1,n}(y) - \theta_n$. Now,

$$E\, g_{1,n}(X_1) = 0, \qquad E\big[g_{2,n}(X_1,X_2) \mid X_1\big] = 0$$

and with $\zeta_{1,n} = \operatorname{Var} g_{1,n}(X_1)$ and $\zeta_{2,n} = \operatorname{Var} g_n(X_1,X_2)$, the variances are:

$$\operatorname{Var}(L_n) = \frac{4\zeta_{1,n}}{n}, \qquad \operatorname{Var}(R_n) = \frac{2\,(\zeta_{2,n} - 2\zeta_{1,n})}{n(n-1)}.$$

Hence,

$$\operatorname{Var}(U_n) = \frac{2}{n(n-1)}\Big[2(n-2)\,\zeta_{1,n} + \zeta_{2,n}\Big].$$

If $X$ has a bounded density $f$ and $K$ is bounded

$E[g_n(X_1,X_2) \mid X_1 = x] = \int K(u)f(x-h_nu)\,du$ is bounded, so $\zeta_{1,n} = O(1)$. $E[g_n^2] = \frac{1}{h}\iint K(u)^2 f(x)f(x-hu)\,du\,dx = O(1/h_n)$, so $\zeta_{2,n} = O(1/h)$. Therefore

$$\operatorname{Var}(U_n) = O\left(\frac{1}{n}\right) + O\left(\frac{1}{n^2 h}\right),$$

and Chebyshev Inequality gives $U_n - \theta_n \xrightarrow{p} 0$ whenever $n^2h \to \infty$.

Hence, we have the following approximation for $\int \hat{f}^2(x)dx - 2\int f(x)\hat{f}(x)$ ,

$$M(h) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{h_n}K\star K\left(\frac{X_j - X_i}{h}\right) - 2\frac{1}{n(n-1)}\frac{1}{h}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}K\left(\frac{X_i - X_j}{h}\right). \quad (43)$$

which can be minimized with respect to $h$. $M(h)$ is called the cross-validation function for $\hat{f}$, and we define

$$h_{cv} := \arg\min_h M(h),$$

which can be obtained by a numerical optimization procedure.
**Example:** Suppose $K(\psi) = \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}\psi^2)$. Then, $K * K(\psi) = \frac{1}{\sqrt{2\pi}\sqrt{2}} exp(-\frac{1}{4}\psi^2)$, and

$$M(\hat{f}) = \frac{1}{h_n n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sqrt{2}} exp\left(-\frac{1}{4}\left(\frac{X_i - X_j}{h_n}\right)^2\right) - 2(n(n-1)h_n)^{-1} \sum_{t=1}^{n} \sum_{i=1, i\neq t}^{n} K\left(\frac{X_t - X_i}{h_n}\right).$$

## 7.2 Cross validation for the Hestenes based estimator

We consider a density $f$ supported on $[0, \infty)$ with $\int_{[0,\infty)} f^2(x)dx < \infty$ and for $x \geq 0$,

$$\hat{f}_s(x) = \frac{1}{nh_n} \sum_{i=1}^{n} \left[ K\left(\frac{x - X_i}{h_n}\right) + \sum_{j=1}^{s+1} \frac{k_j}{w_j} K\left(\frac{x + X_i/w_j}{h_n}\right) \right]. \tag{11}$$

The bandwidth $h_n$ is chosen to minimize an estimate of the integrated squared error over $D_f$,

$$\int_{[0,\infty)} \big(\hat{f}_s(x) - f(x)\big)^2 \, dx = \int_{[0,\infty)} \hat{f}_s^2(x) \, dx - 2 \int_{[0,\infty)} \hat{f}_s(x) f(x) \, dx + \int_{[0,\infty)} f^2(x) \, dx.$$

The last term does not depend on $h_n$, so we need expressions for the first two integrals in terms of the data and the seed kernel $K$. Define $(c_0, a_0) = (1, 1)$ and

$$(c_j, a_j) = \left(\frac{k_j}{w_j}, -\frac{1}{w_j}\right), \quad j = 1, \ldots, s+1.$$

Since $K$ is even, $K\big((x + X_i/w_j)/h_n\big) = K\big((x - a_j X_i)/h_n\big)$, and (11) can be written as

$$\hat{f}_s(x) = \frac{1}{nh_n} \sum_{i=1}^{n} \sum_{j=0}^{s+1} c_j \, K\left(\frac{x - a_j X_i}{h_n}\right), \quad x \geq 0. \tag{44}$$

Thus $\hat{f}_s$ is a weighted kernel sum over the $n(s+2)$ *pseudo-observations* $Y_{ij} = a_j X_i$: the data themselves ($j = 0$) and the reflected points $-X_i/w_j \leq 0$ ($j \geq 1$), with weights $c_j$.

## The integral of $\hat{f}_s^2$

For $\alpha \in [-\infty, \infty)$ and $\beta \in \mathbb{R}$ define the *incomplete convolution* of $K$ with itself,

$$\bar{K}_2(\alpha, \beta) = \int_{\alpha}^{\infty} K(t) \, K(t + \beta) \, dt. \tag{45}$$

Note that $\bar{K}_2(-\infty, \beta) = \int_{\mathbb{R}} K(t)K(t+\beta) \, dt = (K \star K)(\beta)$, the usual convolution that appears in cross-validation for the Rosenblatt estimator.

**Lemma 6.** *For the estimator in (11),*

$$\int_{[0,\infty)} \hat{f}_s^2(x)\,dx = \frac{1}{n^2 h_n} \sum_{i=1}^{n}\sum_{l=1}^{n}\sum_{j,m=0}^{s+1} c_j c_m\, \bar{K}_2\!\left(-\frac{a_j X_i}{h_n},\, \frac{a_j X_i - a_m X_l}{h_n}\right). \tag{46}$$

*Proof.* Squaring (44),

$$\int_{[0,\infty)} \hat{f}_s^2(x)\,dx = \frac{1}{n^2 h_n^2} \sum_{i=1}^{n}\sum_{l=1}^{n}\sum_{j=0}^{s+1}\sum_{m=0}^{s+1} c_j c_m \int_0^\infty K\left(\frac{x - a_j X_i}{h_n}\right) K\left(\frac{x - a_m X_l}{h_n}\right) dx.$$

For fixed $y, z \in \mathbb{R}$, the change of variables $x = y + h_n t$ gives $(x - z)/h_n = t + (y - z)/h_n$, and $x \geq 0$ if and only if $t \geq -y/h_n$. Hence

$$\begin{aligned}\int_0^\infty K\left(\frac{x-y}{h_n}\right) K\left(\frac{x-z}{h_n}\right) dx &= h_n \int_{-y/h_n}^\infty K(t) K\left(t + \frac{y-z}{h_n}\right) dt \\ &= h_n\, \bar{K}_2\!\left(-\frac{y}{h_n}, \frac{y-z}{h_n}\right) := C_h(y,z).\end{aligned}$$

Taking $y = a_j X_i$ and $z = a_m X_l$ gives (46). □

**Corollary 3** (Gaussian kernel)**.** *Let $K = \phi$, the standard normal density, let $\Phi$ be its distribution function, and let $\phi_{\sqrt{2}}(u) = (4\pi)^{-1/2} e^{-u^2/4}$ be the $N(0,2)$ density. Then*

$$\bar{K}_2(\alpha, \beta) = \phi_{\sqrt{2}}(\beta)\left[1 - \Phi\!\left(\sqrt{2}\,\alpha + \frac{\beta}{\sqrt{2}}\right)\right],$$

*and, writing $Y_{ij} = a_j X_i$,*

$$\int_{[0,\infty)} \hat{f}_s^2(x)\,dx = \frac{1}{n^2 h_n} \sum_{i=1}^{n}\sum_{l=1}^{n}\sum_{j=0}^{s+1}\sum_{m=0}^{s+1} c_j c_m\, \phi_{\sqrt{2}}\!\left(\frac{Y_{ij} - Y_{lm}}{h_n}\right) \Phi\!\left(\frac{Y_{ij} + Y_{lm}}{\sqrt{2}\, h_n}\right). \tag{47}$$

*Proof.* Completing the square, $t^2 + (t+\beta)^2 = 2(t + \beta/2)^2 + \beta^2/2$, so

$$\phi(t)\phi(t+\beta) = \frac{1}{2\pi}\, e^{-\beta^2/4}\, e^{-(t+\beta/2)^2}.$$

With $v = \sqrt{2}\,(t + \beta/2)$,

$$\int_\alpha^\infty e^{-(t+\beta/2)^2}\,dt = \frac{1}{\sqrt{2}} \int_{\sqrt{2}(\alpha+\beta/2)}^\infty e^{-v^2/2}\,dv = \sqrt{\pi}\left[1 - \Phi\big(\sqrt{2}\,\alpha + \beta/\sqrt{2}\big)\right].$$

Multiplying, $\bar{K}_2(\alpha,\beta) = (2\sqrt{\pi})^{-1} e^{-\beta^2/4}\big[1 - \Phi(\sqrt{2}\alpha + \beta/\sqrt{2})\big]$, which is the first claim. For (47), take $\alpha = -y/h_n$ and $\beta = (y-z)/h_n$; then $\sqrt{2}\,\alpha + \beta/\sqrt{2} = -(y+z)/(\sqrt{2}\,h_n)$ and $1 - \Phi(-u) = \Phi(u)$. □

## The integral $\int_{[0,\infty)} \hat{f}_s(x) f(x) dx$

By linearity and (44),

$$\int_{[0,\infty)} \hat{f}_s(x) f(x)\, dx = \frac{1}{n h_n} \sum_{i=1}^{n} \sum_{j=0}^{s+1} c_j \int_0^\infty K\left(\frac{x - a_j X_i}{h_n}\right) f(x)\, dx, \tag{48}$$

which depends on the unknown $f$. Let $\hat{f}_{s,-l}$ denote the estimator (11) computed from the $n-1$ observations $\{X_i\}_{i \neq l}$.

**Lemma 7.** *Let $\hat{f}_s^{(n-1)}$ denote the estimator (11) based on a sample of size $n-1$. Then*

$$E\left[\frac{1}{n}\sum_{l=1}^{n} \hat{f}_{s,-l}(X_l)\right] = E\left[\int_{[0,\infty)} \hat{f}_s^{(n-1)}(x) f(x)\, dx\right],$$

*and the leave-one-out estimator of* (48) *has the explicit form*

$$\begin{aligned}
\frac{1}{n}\sum_{l=1}^{n} \hat{f}_{s,-l}(X_l) &= \frac{1}{n(n-1)h_n} \sum_{l=1}^{n} \sum_{i \neq l} \sum_{j=0}^{s+1} c_j\, K\left(\frac{X_l - a_j X_i}{h_n}\right) \\
&= \frac{1}{n(n-1)h_n} \sum_{l=1}^{n} \sum_{i \neq l} \left[ K\left(\frac{X_l - X_i}{h_n}\right) + \sum_{j=1}^{s+1} \frac{k_j}{w_j} K\left(\frac{X_l + X_i/w_j}{h_n}\right)\right]. \qquad (49)
\end{aligned}$$

*Proof.* The explicit form follows from (44) with the sample $\{X_i\}_{i \neq l}$ and the point $x = X_l \geq 0$. Since $X_l$ is independent of $\{X_i\}_{i \neq l}$, conditioning on $\{X_i\}_{i \neq l}$ gives

$$E\big[\hat{f}_{s,-l}(X_l) \,\big|\, \{X_i\}_{i \neq l}\big] = \int_{\mathbb{R}} \hat{f}_{s,-l}(x) f(x)\, dx = \int_{[0,\infty)} \hat{f}_{s,-l}(x) f(x)\, dx,$$

where the last equality uses $f(x) = 0$ for $x < 0$; this is why $\hat{f}_{s,-l}$ is only ever evaluated on $D_f$. Since $\hat{f}_{s,-l}$ has the same distribution as $\hat{f}_s^{(n-1)}$, taking expectations and averaging over $l$ gives the result. □

## The cross-validation function

Combining (46) and (49), the least-squares cross-validation function for the Hestenes estimator is

$$\begin{aligned}
CV_s(h_n) = \frac{1}{n^2 h_n} \sum_{i=1}^{n} \sum_{l=1}^{n} \sum_{j=0}^{s+1} \sum_{m=0}^{s+1} c_j c_m\, \bar{K}_2\!\left(-\frac{a_j X_i}{h_n},\, \frac{a_j X_i - a_m X_l}{h_n}\right) \\
- \frac{2}{n(n-1)h_n} \sum_{l=1}^{n} \sum_{i \neq l} \sum_{j=0}^{s+1} c_j\, K\left(\frac{X_l - a_j X_i}{h_n}\right),
\end{aligned} \tag{50}$$

and $\hat{h}_{CV} = \arg\min_{h_n>0} CV_s(h_n)$. By lemmas 6 and 7,

$$E\big[CV_s(h_n)\big] = E\int_{[0,\infty)} \hat{f}_s^2(x)\,dx - 2E\int_{[0,\infty)} \hat{f}_s^{(n-1)}(x) f(x)\,dx,$$

which equals the mean integrated squared error over $D_f$ minus $\int_{[0,\infty)} f^2$, up to the replacement of $n$ by $n-1$ in the second term.

**Remark 16.** *1. For $s = 0$ and $w_1 = 1$, $(c_1, a_1) = (1, -1)$ and (53) is the cross-validation function for Schuster's reflection estimator.*

*2. The criterion (53) integrates over $D_f = [0, \infty)$. Using instead the Rosenblatt criterion $\int_{\mathbb{R}} \hat{f}^2 - \frac{2}{n}\sum_l \hat{f}_{-l}(X_l)$ for the uncorrected estimator penalizes the $O(1)$ boundary bias, which pushes the selected bandwidth down.*

*3. For the Gaussian kernel, replace $\bar{K}_2$ in (53) by the closed form of Corollary 3. For a kernel supported on $[-1, 1]$, replace $\bar{K}_2$ by $(K * K)$ for all pairs except those with both pseudo-observations in $(-h_n, h_n)$, and drop all pseudo-observations with $a_j X_i \le -h_n$.*

## 7.3 Cross validation for $\hat{F}(x)$ estimator

The distribution function $F$ satisfies $\int |x|\,dF(x) < \infty$, and $K : \mathbb{R} \to \mathbb{R}$ is integrable with $\int K(t)\,dt = 1$ and $\int |t||K(t)|\,dt < \infty$. With

$$U(x) = \int_x^\infty K(t)\,dt,$$

the estimator in equation (28) is

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} U\left(\frac{X_i - x}{h}\right). \tag{28}$$

We want $h$ to minimize the integrated squared error

$$ISE(\hat{F}) = \int_{\mathbb{R}} \big(\hat{F}(x) - F(x)\big)^2\,dx.$$

### The cross-validation function

Let

$$\hat{F}_{-i}(x) = \frac{1}{n-1}\sum_{j\neq i} U\left(\frac{X_j - x}{h}\right)$$

be the estimator computed without $X_i$, and define

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n} \int_{\mathbb{R}} \Big( I\{X_i \le x\} - \hat{F}_{-i}(x) \Big)^2 \, dx. \tag{51}$$

The bandwidth is chosen as $\hat{h} = \operatorname{argmin}_{h>0} CV(h)$.

**Lemma 8.** *Let $\hat{F}_{n-1}$ denote the estimator (28) based on a sample of size $n-1$. Then*

$$E\big(CV(h)\big) = E \int_{\mathbb{R}} \big(\hat{F}_{n-1}(x) - F(x)\big)^2 \, dx + \int_{\mathbb{R}} F(x)\big(1-F(x)\big)\, dx.$$

*Proof.* Fix $x$ and $i$. Write $I\{X_i \le x\} - \hat{F}_{-i}(x) = \big(I\{X_i \le x\} - F(x)\big) - \big(\hat{F}_{-i}(x) - F(x)\big)$. Since $X_i$ is independent of $\hat{F}_{-i}(x)$ and $E\big(I\{X_i \le x\}\big) = F(x)$, the cross product has expectation zero, so

$$E\Big( I\{X_i \le x\} - \hat{F}_{-i}(x) \Big)^2 = F(x)\big(1-F(x)\big) + E\big(\hat{F}_{-i}(x) - F(x)\big)^2.$$

$\hat{F}_{-i}$ has the distribution of $\hat{F}_{n-1}$. Integrating over $x$ (Tonelli) and averaging over $i$ gives the result. □

The second term does not depend on $h$. Hence $CV(h)$ is, up to a constant, an unbiased estimator of the mean integrated squared error of $\hat{F}_{n-1}$. It plays the role that $M(\hat{f})$ plays for the density.

## A closed form

**Lemma 9.** *Let $H_a(x) = I\{a \le x\}$. For all $a, b, c, d \in \mathbb{R}$,*

$$\int_{\mathbb{R}} \big(H_a(x) - H_b(x)\big)\big(H_c(x) - H_d(x)\big)\, dx = \tfrac{1}{2}\big(|a-d| + |b-c| - |a-c| - |b-d|\big).$$

*Proof.* Since $H_a H_c = H_{a\vee c}$, the integrand is $H_{a\vee c} - H_{a\vee d} - H_{b\vee c} + H_{b\vee d}$, and $\int_{\mathbb{R}} (H_p - H_q)\, dx = q - p$. The integral is therefore $(a\vee d) + (b\vee c) - (a\vee c) - (b\vee d)$. Using $p\vee q = \frac{1}{2}(p+q+|p-q|)$, the linear terms cancel, which gives the claim. □

Note that $U\big((X_j - x)/h\big) = \int I\{X_j - ht \le x\} K(t)\, dt = \int H_{X_j - ht}(x) K(t)\, dt$. Since $\int K = 1$,

$$I\{X_i \le x\} - \hat{F}_{-i}(x) = \frac{1}{n-1}\sum_{j \ne i} \int \big(H_{X_i}(x) - H_{X_j - ht}(x)\big) K(t)\, dt.$$

Square this, integrate in $x$, and use Fubini (justified by $\int |t||K(t)|\,dt < \infty$). Then apply Lemma 9 with $a = c = X_i$, $b = X_j - ht$, $d = X_k - hs$. Define

$$A(u) = \int_{\mathbb{R}} |u+t|K(t)\,dt, \qquad B(u) = \int_{\mathbb{R}}\int_{\mathbb{R}} |u-t+s|K(t)K(s)\,dt\,ds.$$

This gives

$$CV(h) = \frac{h}{n}\sum_{i=1}^{n}\left[\frac{1}{n-1}\sum_{j\neq i} A\left(\frac{X_i - X_j}{h}\right) - \frac{1}{2(n-1)^2}\sum_{j\neq i}\sum_{k\neq i} B\left(\frac{X_j - X_k}{h}\right)\right]. \qquad (52)$$

In the double sum the diagonal $j = k$ is included, and $B(0) = \int\int |t-s|K(t)K(s)\,dt\,ds \neq 0$. Since $K$ is known, $\hat{h}$ can be found by numerical minimization of (52), exactly as for $M(\hat{f})$.

**Remark 17.** *If $K$ is even, then $\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} L\big((x - X_i)/h\big)$ with $L(x) = \int_{-\infty}^{x} K(t)\,dt$, and*

$$A(u) = \int_{\mathbb{R}} |u-t|K(t)\,dt, \qquad B(u) = \int_{\mathbb{R}} |u-v|\,K{*}K(v)\,dv, \quad K{*}K(v) = \int_{\mathbb{R}} K(v-t)K(t)\,dt.$$

**Remark 18.** *Let $B_{jk} = B\big((X_j - X_k)/h\big)$ and $S = \sum_{j=1}^{n}\sum_{k=1}^{n} B_{jk}$. Then $\sum_{j\neq i}\sum_{k\neq i} B_{jk} = S - 2\sum_{k=1}^{n} B_{ik} + B_{ii}$. So (52) needs only $O(n^2)$ kernel evaluations for each $h$.*

## 7.4 Cross validation for the jump estimator

The set of points where $F$ jumps is $A = \{x_k\}_{k\in\mathbb{N}}$, and $p_k := p_{x_k} = F(x_k) - F(x_k - 0)$. We write $\hat{p}_k := \hat{p}_{x_k}$ and want $h$ to minimize

$$J(h) = \sum_k \big(\hat{p}_k - p_k\big)^2 = \sum_k \hat{p}_k^2 - 2\sum_k p_k\hat{p}_k + \sum_k p_k^2.$$

The last term does not depend on $h$ and the first is computable from the data. The cross term $\sum_k p_k\hat{p}_k$ involves the unknown $p_k$ and must be estimated.

### Estimating the cross term

For any bounded $g$, $\sum_k p_k\, g(x_k) = E\big(g(X)I\{X \in A\}\big)$. Hence, if $X'$ is independent of the sample and has distribution $F$,

$$\sum_k p_k\hat{p}_k = \frac{1}{n}\sum_{i=1}^{n}\sum_k p_k\, W\left(\frac{X_i - x_k}{h}\right) = \frac{1}{n}\sum_{i=1}^{n} E_{X'}\left(W\left(\frac{X_i - X'}{h}\right) I\{X' \in A\}\right).$$

Leaving one observation out gives the unbiased estimator

$$\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} W\left(\frac{X_i - X_j}{h}\right) I\{X_j \in A\}.$$

Indeed, for $j \neq i$,

$$E\left(W\left(\frac{X_i - X_j}{h}\right) I\{X_j \in A\}\right) = \sum_k p_k \, E\, W\left(\frac{X_i - x_k}{h}\right) = \sum_k p_k \, E(\hat{p}_k).$$

## The cross-validation function

$$CV(h) = \frac{1}{n^2} \sum_k \left(\sum_{i=1}^{n} W\left(\frac{X_i - x_k}{h}\right)\right)^2 - \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} W\left(\frac{X_i - X_j}{h}\right) I\{X_j \in A\}, \tag{53}$$

and $\hat{h} = \operatorname{argmin}_{h>0} CV(h)$.

**Lemma 10.** $E\big(CV(h)\big) = E\big(J(h)\big) - \sum_k p_k^2$.

*Proof.* The first term in (53) equals $\sum_k \hat{p}_k^2$. By the previous subsection, the expectation of the second term is $-2\sum_k p_k E(\hat{p}_k)$. Taking expectations in the expansion of $J(h)$ gives the result. □

So $CV(h)$ is, up to a term that does not depend on $h$, an unbiased estimator of the expected criterion.

## A simpler form

Let $N_k = \#\{j : X_j = x_k\}$ be the number of observations at the jump point $x_k$. Since $I\{X_j \in A\} = \sum_k I\{X_j = x_k\}$ and $\sum_{j \neq i} I\{X_j = x_k\} = N_k - I\{X_i = x_k\}$,

$$\sum_{i=1}^{n} \sum_{j \neq i} W\left(\frac{X_i - X_j}{h}\right) I\{X_j = x_k\} = N_k \sum_{i=1}^{n} W\left(\frac{X_i - x_k}{h}\right) - N_k \, W(0) = N_k\big(n\hat{p}_k - 1\big),$$

using $W(0) = 1$. Hence

$$CV(h) = \sum_k \left[\hat{p}_k^2 - 2\,\frac{N_k}{n} \cdot \frac{n\hat{p}_k - 1}{n-1}\right]. \tag{54}$$

Both factors in the cross term have a direct interpretation:

- $N_k/n$ is the empirical frequency of $x_k$, an unbiased estimator of $p_k$;

- $(n\hat{p}_k - 1)/(n-1)$ is the estimator of $p_k$ computed without one of the observations at $x_k$, which contributes $W(0) = 1$ to $n\hat{p}_k$.

Computing (54) requires $O(n)$ operations per jump point and per $h$.

**Remark 19.** *The criterion requires $A$ to be known. In practice, an observed value that occurs more than once is an atom of $F$, since the continuous part of $F$ produces ties with probability zero. An atom that is observed only once cannot be distinguished from an observation from the continuous part.*

**Remark 20.** *If $A$ is infinite, we need $\sum_k E(\hat{p}_k^2) < \infty$. This holds, for example, if $W$ has compact support and $A$ is locally finite. If $A$ is dense (for example, the rationals) the criterion may diverge.*

## 7.5 Minimizing Mean Integrated Squared Error

One of the remarks made regarding $ISE$ was that it depends on the sample $\{X_i\}_{i=1}^n$. One can imagine a process of repeated sampling that induces a distribution on ISE. Hence, we might consider,

$$\begin{aligned} MISE(\hat{f}) &= E\left(\int (\hat{f}(x) - f(x))^2 dx\right) \\ &= E\left(\int (\hat{f}(x) - E(\hat{f}(x)) + E(\hat{f}(x)) - f(x))^2 dx\right) \\ &= E\left(\int \{(\hat{f}(x) - E(\hat{f}(x)))^2 + (E(\hat{f}(x)) - f(x))^2 \right. \\ &+ \left. 2(\hat{f}(x) - E(\hat{f}(x)))(E(\hat{f}(x)) - f(x))\} dx\right) \\ &= \int \{V(\hat{f}(x)) + E(\hat{f}(x) - f(x))^2\} dx \end{aligned}$$

Note that $V(\hat{f}(x)) = \frac{1}{nh_n} f(x) \int K^2(\psi) d\psi + o((nh_n)^{-1})$ and if $f \in \Sigma(2, C)$ we have that $E(\hat{f}(x)) - f(x)) = \frac{h_n^2}{2} \mu_2 f^{(2)}(x) + o(h_n^2)$ with $\mu_2 = \int x^2 K(x) dx$. Hence, $MISE(\hat{f})$ can be written as

$$MISE(\hat{f}) = (nh_n)^{-1} \int K^2(\psi) d\psi + \frac{h_n^4}{4} \mu_2^2 \int f^{(2)}(x)^2 dx + o((nh_n)^{-1}) + o(h_n^4).$$

Let $\int K^2(\psi) d\psi = \lambda_2$, $\mu_2^2 \int f^{(2)}(x)^2 dx = \lambda_1$ so that

$$MISE(\hat{f}) = (nh_n)^{-1} \lambda_2 + \frac{h_n^4}{4} \lambda_1 + o((nh_n)^{-1}) + o(h_n^4).$$

Let the $h_n$ that minimizes the first two terms of $MISE(\hat{f})$ be denoted by $\hat{h}_n$. It satisfies

$$-n^{-1}(\hat{h}_n)^{-2}\lambda_2 + \hat{h}_n^3\lambda_1 = 0$$

or $\hat{h}_n^5\lambda_1 = \frac{1}{n}\lambda_2$ which implies

$$\hat{h}_n = \left(\frac{\lambda_2}{n\lambda_1}\right)^{\frac{1}{5}} = n^{-\frac{1}{5}}\left(\frac{\int K^2(\psi)d\psi}{(\int \psi^2 K(\psi)d\psi)^2 \int (f^{(2)}(x))^2 dx}\right)^{\frac{1}{5}}$$

**Remark 21.** *1. Although $\lambda_2$ can be obtained from knowledge of $K$, $\lambda_1$ depends on $f^{(2)}(x)$. Hence, in practice $\hat{h}_n$ depends on obtaining an estimate for $\int (f^{(2)}(x))^2 dx$.*

*2. Taking $(\frac{\lambda_2}{\lambda_1})^{\frac{1}{5}} = c$ we see that $\hat{h}_n \propto n^{-\frac{1}{5}}$. Hence, according to the AMISE criterion, there is an optimal rate of convergence of the bandwidth to zero, namely, $n^{-1/5}$. Recall that*

$$(nh_n)^{1/2}(\hat{f}(x) - f(x)) = (nh_n)^{1/2}(\hat{f}(x) - E(\hat{f}(x)) + (nh_n)^{1/2}(E(\hat{f}(x)) - f(x))$$

*and since the first term after the equality converges in distribution to a normal distribution, normality of the left side of the equality results if $(nh_n)^{1/2}(E(\hat{f}(x)) - f(x))$ approaches zero as $n \to \infty$. But*

$$(nh_n)^{1/2}(E(\hat{f}(x)) - f(x)) = n^{1/2}h_n^{5/2}\frac{1}{2}\mu_2 f^{(2)}(x) + n^{1/2}\frac{h_n^{5/2}}{h_n^2}R(n)$$

*where $R(n) = O(h_n^2)$. Hence, for $(nh_n)^{1/2}(E(\hat{f}(x)) - f(x))$ to vanish asymptotically, we need $nh_n^5 \to 0$ as $n \to \infty$ which implies that $h_n \propto n^{\delta}$ where $-1 < \delta < -1/5$. Consequently, asymptotic normality of $(nh_n)^{1/2}(\hat{f}(x) - f(x))$ cannot result when the optimal choice $h_n$ is based on AMISE, oversmoothing is necessary.*

*3. It should be clear that both minimization of the cross validation function and AMISE can be extended to multivariate density estimation with little conceptual effort. For a random vector $X : \Omega \to \mathbb{R}^D$ we have*

$$M(\hat{f}) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{det(H_n)}K * K\left(H_n^{-1}(X_j - X_i)\right)$$
$$- 2\frac{1}{n(n-1)}\frac{1}{det(H_n)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}K\left(H_n^{-1}(X_i - X_j)\right).$$

*Also, in the case where* $f \in \Sigma(2, C)$ *and* $\int \gamma_d K(\gamma) d\gamma = 0$ *and* $\int \gamma_{d_1}\gamma_{d_2} K(\gamma) d\gamma = 0$ *when* $d_1 \neq d_2$ *we can write*

$$\begin{aligned} AMISE(\hat{f}) &= (ndet(H_n))^{-1} \int K^2(\psi) d\psi + \frac{\mu_2^2}{4} \int \left( \sum_{i=1}^{D} D_{ii} f(x) h_{i,n}^2 \right)^2 dx \\ &= (ndet(H_n))^{-1} \int K^2(\psi) d\psi + \frac{\mu_2^2}{4} \sum_{i=1}^{D} \sum_{j=1}^{D} h_{i,n}^2 h_{j,n}^2 \int D_{ii} f(x) D_{jj} f(x) dx. \end{aligned}$$

*Taking derivatives with respect to each of the elements in* $H_n$ *we have that* $\hat{H}_n$ *must satisfy,*

$$\mu_2^2 \left( \hat{h}_{d,n}^3 \int (D_{dd})^2 f(x) dx + \hat{h}_{d,n} \sum_{i=1, i \neq d}^{D} \hat{h}_{i,n}^2 \int D_{dd} f(x) D_{ii} f(x) dx \right) = \frac{\lambda_2}{n \hat{h}_{d,n} det(\hat{H}_n)}$$

*for all* $d$. *When* $H_n = h_n I_D$ *we have*

$$\hat{h}_n^{D+4} = \frac{\lambda_2}{n\mu_2^2 \sum_{i=1}^{D} \sum_{d=1}^{D} \int D_{dd} f(x) D_{ii} f(x) dx} \text{ which implies that } \hat{h}_n = \left( \frac{\lambda_2}{n\mu_2^2 \lambda_1} \right)^{\frac{1}{D+4}}.$$

4. *The simplest procedure to estimate* $\int (f^{(2)}(x))^2 dx$ *is to assume, at this stage, that* $f(x)$ *belongs to a parametric family of densities. For example, if we assume that* $X_i \sim N(\mu, \sigma^2)$, *then for* $r \in \mathbb{N}$

$$\int (f^{(r)}(u))^2 du = \frac{(2r)!}{2^{2r+1} r! \sqrt{\pi} \sigma^{2r+1}}.$$

*For* $r = 2$ *and plugging back into the optimal bandwidth formula gives*

$$h_{AMISE} = \begin{cases} n^{-1/5} 1.06 \sigma & \text{if } K \text{ is a Gaussian kernel} \\ n^{-1/5} (1.06)(2.214) \sigma & \text{if } K \text{ is an Epanechnikov kernel} \\ n^{-1/5} (1.06)(1.740) \sigma & \text{if } K \text{ is an uniform kernel} \end{cases}$$

*where* $\sigma$ *can be estimated by the sample standard deviation. The resulting bandwidths are normally referred to as Silverman's rule-of-thumb bandwidths.*

# Appendix

**Lemma 11.** *Let $C \in \mathbb{R}$, $n \in \mathbb{N}$ and $\{X_1, \cdots, X_n\}$ be a collection of random variables defined in the same probability space. Then,*

$$P\left(\max\{X_1, \cdots, X_n\} > C\right) \leq \sum_{i=1}^{n} P(\{X_i > C\}).$$

*Proof.* First, note that

$$\begin{aligned} P\left(\max\{X_1, \cdots, X_n\} > C\right) &= 1 - P\left(\max\{X_1, \cdots, X_n\} \leq C\right) \\ &= 1 - P\left(\cap_{i=1}^{n}\{X_i \leq C\}\right) \\ &= P\left(\left(\cap_{i=1}^{n}\{X_i \leq C\}\right)^c\right). \end{aligned}$$

Second, by De Morgan's Laws

$$\begin{aligned} P\left(\max\{X_1, \cdots, X_n\} > C\right) &= P\left(\cup_{i=1}^{n}\{X_i \leq C\}^c\right) \\ &\leq \sum_{i=1}^{n} P(\{X_i \leq C\}^c) = \sum_{i=1}^{n} P(\{X_i > C\}). \end{aligned}$$

□

**Lemma 12.** *(Bernstein's Inequality) Let $\{W_i\}_{i=1}^n$ be a sequence of independent random variables satisfying*

1. *$|W_i - E(W_i)| \leq m_i$ for all $i$,*
2. *$V(W_i) = \sigma_i^2 <$ for all $i$*

*Let $M_n := \max_{1 \leq i \leq n}\{m_i\}$, $S_n := \sum_{i=1}^{n} W_i$ and $\sigma_n^2 := \sum_{i=1}^{n} \sigma_i^2$. Then, for all $\Delta > 0$ we have*

$$P\left(|S_n| \geq \Delta\sigma_n\right) < 2\exp\left(-\frac{\Delta^2}{2 + \frac{2}{3}\frac{M_n}{\sigma_n}\Delta}\right)$$

*Proof.* See Bernstein (1924) or Bennett (1962). □

**Lemma 13.** *(Loève's $c_r$ Inequality) For $r > 0$,*

$$E\left(\left|\sum_{i=1}^{n} X_i\right|^r\right) \leq c_r \sum_{i=1}^{n} E(|X_i|^r)$$

*where $c_r = 1$ when $r \leq 1$ and $c_r = n^{r-1}$ when $r > 1$.*

*Proof.* We will prove that

$$\left|\sum_{i=1}^{n} X_i\right|^r \leq c_r \sum_{i=1}^{n} |X_i|^r \tag{55}$$

and then take expectations. By the Triangle Inequality $|\sum_{i=1}^n X_i|^r \leq (\sum_{i=1}^n |X_i|)^r$. Thus, it suffices to consider the case where each $X_i \geq 0$. Let $0 < r \leq 1$ and note that since $\sum_{i=1}^n X_i > 0$ (if $X_i = 0$ for all $i$ then inequality (55) holds with equality) it suffices to show that

$$1 \leq \sum_{i=1}^{n} Z_i^r$$

where $Z_i = \frac{X_i}{\sum_{i=1}^n X_i}$, $\sum_{i=1}^n Z_i = 1$ and $0 \leq Z_i \leq 1$. But since, $Z_i^r \geq Z_i$ and $\sum_{i=1}^n Z_i = 1$, the results follows immediately. For $r > 1$, convexity of $x^r$ immediately gives

$$\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^r \leq \frac{1}{n}\sum_{i=1}^{n} X_i^r \Leftrightarrow \left(\sum_{i=1}^{n} X_i\right)^r \leq n^{r-1}\sum_{i=1}^{n} X_i^r.$$

Taking expectations in both cases gives the desired inequality. □

# References


Bennett, G., 1962. Probability inequalities for the sum of independent random variables. Journal of the American Statistical Association 57, 33–45.

Bernstein, S., 1924. Sur une modification de l'inéqualité de tchebichef. Annals Science Institute Sav. Ukraine Sect. Math. I.

Davidson, J., 1994. Stochastic limit theory. Oxford University Press, New York.

Hestenes, M. R., 1941. Extension of the range of a differentiable function. Duke Mathematical Journal 8, 183–192.

Marron, J. S., Ruppert, D., 1994. Transformations to reduce boundary bias in kernel density estimation. Journal of the Royal Statistical Society - Series B 56, 653–671.

Natanson, I. P., 1955. Theory of functions of a real variable. Vol. 1. Frederick Ungar Publishing Co, New York, NY.

Rosenblatt, M., 1956. Remarks on some nonparametric estimates of a density function. Annals of Mathematical Statistics 27, 832–837.

Schuster, E. F., 1985. Incorporating support constraints into nonparametric estimators of densities. Communications in Statistics - Theory and Methods, 14, 1123–1136.