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 $[a,b]$,*

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

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

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

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 μ, ν be measures on a measurable space $(E, \mathcal{E})$. We write $\mu \ll \nu$ (μ is* absolutely continuous with respect to ν*) 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 $\lambda(U') \leq \lambda(U) < \delta$. $\square$

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 ϵ.

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 ϵ, we conclude $P_X(A) = 0$. $\square$

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

Theorem 2 (Radon–Nikodým)**.** *Let μ and ν be measures on the measurable space $(E, \mathcal{E})$. If ν is σ-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* ν*-almost everywhere unique* $\mathcal{E}$*-measurable* $f \geq 0$

2. $\mu \ll \nu$.

Lemma 2. P_X *is a finite measure and* λ *is* σ*-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*

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

and $\int_{\mathbb{R}} f_X \, d\lambda = 1$. *The function* f_X *is unique up to* λ*-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 λ-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 $g < \infty$ λ-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 λ-null set $N \cup \{|g| = \infty\}$. Since altering the integrand on a λ-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 $f = f'$ λ-a.e. □

Definition 3. *A measurable function $f : \mathbb{R} \to \mathbb{R}$ is said to be locally integrable with respect to Lebesgue measure λ 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 λ-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

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

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 $\lim\limits_{h\to 0}\frac{F(x+h)-F(x-h)}{2h} = f(x)$ 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 λ, 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,*

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

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

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

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.

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

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 $n \to \infty$, $nh_n V(\hat{f}) \to f(x) \int K^2(v)dv$ 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 $n \to \infty$. $\square$

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 β.

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 the kernel function K is of order $m \geq 1$ with $m \in \mathbb{N}$ 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 $0 < |\int u^m K(u)du| < \infty$. 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 $\int u^2 K(u)du < C$.*

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 β 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 δ 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 $E\left((\hat{f}(x) - f(x))^2\right) \leq h_n^4 C_1 + \frac{1}{nh_n}C_2$, 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.*