NEW MULTIVARIATE PRODUCT DENSITY ESTIMATORS

Luc Devroye
School of Computer Science
McGill University
Montreal, Canada H3A 2K6

and

Adam Krzyżak
Department of Computer Science
Concordia University
Montreal, Canada H3G 1M8

Abstract. Let X be an $\mathbb{R}^d$-valued random variable with unknown density f. Let $X_1, \ldots, X_n$ be i.i.d. random variables drawn from f. The objective is to estimate $f(x)$, where $x = (x_1, \ldots, x_d)$. We study the pointwise convergence of two new density estimates, the Hilbert product kernel estimate

$$\frac{d!}{n} \sum_{i=1}^{n} \prod_{j=1}^{d} \frac{1}{2 \log n |x_j - X_{ij}|} ,$$

where $X_i = (X_{i1}, \ldots, X_{id})$, and the Hilbert k-nearest neighbor estimate

$$\frac{k\,(d-1)!}{2^d n \log^{d-1}(n/(k(d-1)!) \prod_{j=1}^{d} |x_j - X_{(k)j}|} ,$$

where $X_{(k)} = (X_{(k)1}, \ldots, X_{(k)d})$, and $X_{(k)}$ is the k-th nearest neighbor of x when points are ordered by increasing values of the product $\prod_{j=1}^{d} |x_j - X_{(k)j}|$, and $k = o(\log n)$, $k \to \infty$. The auxiliary results needed permit us to formulate universal consistency results (pointwise and in L_1) for product kernel estimates with different window widths for each coordinate, and for rectangular partitioning and tree estimates. In particular, we show that locally adapted smoothing factors for product kernel estimates may make the kernel estimate inconsistent even under standard conditions on the bandwidths.

Keywords and phrases. Density estimation, kernel estimate, convergence, bandwidth selection, nearest neighbor estimate, Saks rarity theorem, Jessen-Marcinkiewicz-Zygmund theorem, nonparametric estimation.

1991 Mathematics Subject Classifications: Primary 62G05.

The first author's work was supported by NSERC Grant A3456 and by FCAR Grant 90-ER-0291. The second author was supported by a grant from the Humboldt Foundation and by NSERC grant OGP0000270.

1. Introduction.

The objective of this note is to study two new multivariate density estimates that avoid the messy problem of smoothing factor selection (in one case, at least), are invariant to affine transformations of the coordinates, and provide easy means to jointly estimate all $2^d - 1$ marginal densities of a density f on $\mathbb{R}^d$. As a by-product, we will be able to study the universal consistency of product kernel estimates and of rectangular partition estimates.

Let $X_1, \ldots, X_n$ be independent observations of an $\mathbb{R}^d$-valued random vector X with unknown density f. The classical kernel estimate of f is

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) \ ,$$

where $h > 0$ is a smoothing factor depending upon n, K is an absolutely integrable function (the kernel), and $K_h(x) = (1/h^d)K(x/h)$ (Akaike, 1953; Parzen, 1962; Rosenblatt, 1956). Observe that for the kernel $K(u) = 1/\|u\|^d$, the smoothing factor h is canceled and we obtain

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\|x - X_i\|^d} \ .$$

One may wonder what happens in this situation, now that the smoothing factor is absent. This problem is dealt with by Devroye and Krzyżak (1998), who showed the following.

THEOREM 1. *The Hilbert estimate*

$$f_n(x) = \frac{1}{V_d n \log n} \sum_{i=1}^{n} \frac{1}{\|x - X_i\|^d}$$

where V_d is the volume of the unit ball in $\mathbb{R}^d$, is weakly consistent at almost all x, that is,

$$f_n(x) \to f(x)$$

in probability at almost all x.

We use the name Hilbert estimate because of the related Hilbert integral with a similar kernel. The Hilbert estimate is not invariant under affine transformations of the coordinates. That is, in transparent notation, where $f_n(x; Y)$ denotes the density estimate at $x \in \mathbb{R}^d$ given data $Y \in (\mathbb{R}^d)^n$,

$$\frac{f_n((x_1 - b_1)/a_1, \ldots, (x_d - b_d)/a_d; Y)}{\prod_{i=1}^{d} a_i} \neq f_n(x_1, \ldots, x_n; b + aY) \ ,$$

where $b + aY$ denotes the sample of size n in which each i-th coordinate of each observation is transformed by $b_i + a_i x$. In kernel density estimation, the invariance may be obtained if one uses product kernels. However, this causes additional problems as each component kernel requires its own smoothing factor. If we apply that principle to the Hilbert estimate, we obtain the product Hilbert estimate

$$f_n(x) = \frac{d!}{n} \sum_{j=1}^{n} \prod_{i=1}^{d} \left(\frac{1}{2 \log n |x_i - X_{ji}|} \right) \ .$$

With this estimate, which once again has no smoothing factor, the invariance mentioned above follows readily. However, the weak pointwise consistency does not hold in $\mathbb{R}^d$ for $d > 1$ for all densities. Counterexamples will be provided below. In fact, for the product kernel estimate

$$f_n(x) = \frac{1}{n} \sum_{j=1}^{n} \prod_{i=1}^{d} \frac{1}{h_i} K\left(\frac{x_i - X_{ji}}{h_i}\right)$$

where $h_1, \ldots, h_d$ are positive smoothing parameters, it is known that if $h_1 = \cdots = h_d \stackrel{\text{def}}{=} h$, $h \to 0$ and $nh^d \to \infty$, then $f_n \to f$ in probability at almost all x when K is a bounded compact support density (Devroye and Györfi, 1985), and $\int |f_n - f| \to 0$ almost surely for any K with $\int K = 1$, $\int |K| < \infty$. However, it is less known that if we allow the individual smoothing factors to tend to zero at different rates and are allowed to depend upon x, this universality is lost! The natural conditions on h_i would seem to be

$$\lim_{n\to\infty} \max_i h_i = 0$$

and

$$\lim_{n\to\infty} n \prod_{i=1}^{d} h_i = \infty .$$

Within these conditions, and with K the uniform density on $[-1, 1]$, there exist densities f for which $f_n \not\to f$ almost everywhere in probability. It turns out that a sufficient condition on f for almost everywhere pointwise convergence is $\int f \log^{d-1}(f + 1) < \infty$ (we write $f \in L(1 + \log^+ L)^{d-1}$). The proof and the counterexamples follow in section 3. The paper then explores sufficient conditions for various types of convergence of the estimates, and discusses in this context the Jessen-Marcienkiewicz-Zygmund condition. Consistency is proved for the multivariate Hilbert product estimate, the product nearest neighbor estimate (which hasa nearest neighbor ranking that is invariant under scaling of the axes, a very desirable feature in high-dimensional data collection!), the ordinary product kernel estimate and finally, tree-based and rectangular partitioning estimates.

2. Sums of products of inverse uniforms.

Let $Z = \prod_{i=1}^{d} U_i$ be the product of d i.i.d. uniform $[0, 1]$ random variables. Its density is given by

$$f(z) = \frac{\log^{d-1}(1/z)}{(d-1)!} \;, 0 < z \le 1 .$$

Let $Z_1, \ldots, Z_n$ be i.i.d. and distributed as Z. Then we have:

Lemma S1.

$$\frac{1}{n \log^d n} \sum_{i=1}^{n} \frac{1}{Z_i} \to \frac{1}{d!}$$

in probability.

PROOF. The result follows from Theorem 2 of Rogozin, 1971 and Theorem 8.8.1 of Bingham, Goldie and Teugels, 1987 about stability of sums of i.i.d. random variables $S_n = \sum_{i=1}^n Y_i$: if $Y = Y_1$ has distribution function F, then $S_n/a_n \to 1$ in probability if $\int_0^x y dF(y) \sim l(x)$, where $l(x)$ is a slowly varying function and a_n is chosen such that $l(a_n)/a_n \sim 1/n$. Take $Y = 1/Z$ with density $f_Y(y) = y^{-2}(\log y)^{d-1}/(d-1)!, y > 1$. We have $\int_0^x y dF(y) = \log^d x/d!$ and so we can take $l(x) = \log^d x/d!$ and $a_n = n(\log^d n)/d!$. Indeed,

$$\frac{l(a_n)}{a_n} = \frac{(\log^d(n(\log^d n)/d!))/d!}{(n \log^d)/d!} \sim \frac{1}{n} \ . \square$$

3. The Saks rarity theorem and its implications.

To understand the reasons for defining conditions on f, it helps to understand why we have to do so. The reasons go back to the theory of differentiation (de Guzman, 1981, is a good reference). Consider a function f on $\mathbb{R}^d$, together with a collection $\mathcal{B}$ of bounded measurable sets with the property that for every $x \in \mathbb{R}^d$, there exists a sequence $B_1, B_2, \ldots$ from $\mathcal{B}$ with diameters (written diam(.)) decreasing to zero and such that $x \in \cap_i B_i$. Such a collection is called a basis. The collection of sets in $\mathcal{B}$ containing x is denoted by $\mathcal{B}(x)$. We define upper and lower derivatives of f by

$$D_+(f, x) = \limsup_{B \in \mathcal{B}(x):\mathrm{diam}(B)\downarrow 0} \frac{\int_B f(x)\,dx}{\int_B dx}$$

and

$$D_-(f, x) = \liminf_{B \in \mathcal{B}(x):\mathrm{diam}(B)\downarrow 0} \frac{\int_B f(x)\,dx}{\int_B dx}$$

respectively (de Guzman, 1981, p. 105). we say that $\mathcal{B}$ differentiates f if $D_+(f, x) = D_-(f, x)$ almost everywhere (x). For example, it is known that if $\mathcal{B}$ is the collection of all balls or all hypercubes, then $\mathcal{B}$ differentiates all integrable functions f. This is a form of the celebrated Lebesgue density theorem, and is at the basis of the pointwise convergence properties of kernel estimates and indeed most density estimates. Let $\mathcal{B}_2$ denote the interval basis, that is, the collection of all products of d finite intervals containing at least two points. This is the collection of all bounded rectangles of positive measure aligned with the axes. We will require the following result:

LEMMA A1. (THE JESSEN-MARCINKIEWICZ-ZYGMUND THEOREM, 1935). *$\mathcal{B}_2$ differentiates $L(1+\log^+ L)^{d-1}(\mathbb{R}^d)$, that is, all functions f on $\mathbb{R}^d$ for which*

$$\int |f| \log^{d-1}(1 + |f|) < \infty \ .$$

While this includes most densities f, there are indeed exceptions. A good account of these is chapter 7 of de Guzman (1981, p. 167). First of all, according to the Saks rarity theorem (Saks, 1935), there exists a nonnegative function f on $\mathbb{R}^2$ such that $D_+(f, x) = \infty$ almost everywhere (x) (with respect to $\mathcal{B}_2$). Later, Marstrand (1977) found an f with this property that works for all orientations of the axes (each orientation of the axes has a different collection $\mathcal{B}_2$). El Helou (1978) found an f with the latter property and in addition

$$\int |f| \log^a(1 + |f|) < \infty$$

for all $a \in (0, 1)$. Thus, with respect to $\mathcal{B}_2$, the logarithmic condition on f is nearly necessary.

Consider now a standard kernel estimate on $\mathbb{R}^2$ with product kernel $K(x) = \frac{1}{4} I_{|x_1| \le 1} I_{|x_2| \le 1}$. If we have different smoothing factors for each coordinate, the form is

$$f_n(x) = \frac{1}{nh_1h_2} \sum_{j=1}^{n} I_{|x_1 - X_{j1}| \le h_1} I_{|x_2 - X_{j2}| \le h_2}$$

where $X_j = (X_{j1}, X_{j2})$ and $x = (x_1, x_2)$. It is easy to see that

$$\mathbf{E} f_n(x) = \frac{1}{h_1 h_2} \int_{|x_1 - y_1| \le h_1; |x_2 - y_2| \le h_2} f(y_1, y_2)\, dy_1\, dy_2 \ .$$

But by the result mentioned above, there exists a density $f \in \cap_{0<a<1} L(1 + \log^+ L)^a$ such that at almost all x, there exist $h_1 = h_1(n, x) \downarrow 0$, $h_2 = h_2(n, x) \downarrow 0$ as $n \to \infty$, such that $\mathbf{E} f_n(x) \to \infty$. Note that the results do not imply this when h_1 and h_2 are not allowed to depend upon x. As the variance of f_n is $O(\mathbf{E} f_n(x)/(nh_1h_2))$, it is easy to see that if $nh_1h_2 \to \infty$, then $f_n(x)/\mathbf{E} f_n(x) \to 1$ in probability at almost all x, and thus, $f_n(x) \to \infty$ in probability at almost all x. Therefore, if one adapts the smoothing factors to x, it is not true any longer that kernel estimates are pointwise consistent for all densities!

We turn finally to the basis $\mathcal{B}_3$ of all rectangles in $\mathbb{R}^2$ (rotated with respect to all possible orientations). Here the situation is extremely volatile (de Guzman, 1981, p. 224), as $\mathcal{B}_3$ does not even differentiate the characteristic functions of bounded measurable sets. The counterexamples on which differentiability fails include densities $f = cI_B$ where B is a bounded measurable set of area $1/c$ that is a specially selected subset of the Nikodym set N on $[0, 1]^2$ (N has measure one, but for each $x \in N$, there exists a line $l(x)$ through x such that $l(x) \cap N = \{x\}$). Take such an f. The implication is that there exist inconsistent kernel estimates of the rotation kind: for almost every x, there exist smoothing factors $h_1(n, x) \to 0$ and $h_2(n, x) \to 0$ (with $nh_1(n, x)h_2(n, x) \to \infty$, the standard conditions on smoothing factors) and orthonormal rotation matrices $A(n, x)$ such that the kernel estimate

$$f_n(x) = \frac{1}{nh_1h_2} \sum_{j=1}^{n} K(A(n, x)(x - X_j)) \to \infty$$

at almost all x, where as before $K(x) = \frac{1}{4} I_{|x_1| \le 1} I_{|x_2| \le 1}$. Thus, adaptation to both x and allowing adaptive rotations makes kernel estimates potentially inconsistent even on bounded densities with compact support.

In the remainder of the paper, a rectangle is a set from $\mathcal{B}_2$, and not from $\mathcal{B}_3$.

4. Weak pointwise consistency of the multivariate Hilbert product kernel estimate.

In this section, we prove the main consistency theorem:

THEOREM 2. *Assume that f is a density with $\int f \log^{d-1}(1+f) < \infty$, and for which $\int g \log^{s-1}(1+g) < \infty$ for all its marginal densities g, where s denotes the dimension of the domain of g. Then the multivariate Hilbert product kernel estimate is weakly pointwise consistent almost everywhere: at almost all x,*

$$f_n(x) \to f(x)$$

in probability.

PROOF. Let $\mathcal{M}$ denote the space of all $d \times d$ diagonal matrices with diagonal elements 1 or -1. Clearly, $|\mathcal{M}| = 2^d$. Let $x = (x_1, \ldots, x_d)$ and $y = (y_1, \ldots, y_d)$ be vectors from $\mathbb{R}^d$. For fixed $x \in \mathbb{R}^d$, define the flipped density

$$f_x^*(y) = \begin{cases} \sum_{M \in \mathcal{M}} f(x + M(y-x)) & \text{if } y \geq x \\ 0 & \text{otherwise,} \end{cases}$$

where $y \geq x$ means that $y_i \geq x_i$ for all i. Observe in particular that $f_x^*(x) = 2^d f(x)$ and that f^* is a bona-fide density with support on the positive quadrant with origin x. By the Jessen-Marcinkiewicz-Zygmund theorem (Lemma A1), we have, if $\mathcal{B}_2$ denotes the interval basis on $\mathbb{R}^d$, for almost all x,

$$\lim_{B \in \mathcal{B}(x) : \mathrm{diam}(B) \downarrow 0} \frac{\int_B f(y)\, dy}{\int_B dy} = f(x) \ .$$

This implies that for almost all x,

$$\lim_{B = [x,y], y > x : \mathrm{diam}(B) \downarrow 0} \frac{\int_B f_x^*(z)\, dz}{\int_B dz} = f_x^*(x) \ .$$

Here $[x, y]$ denotes the rectangle $\prod_{i=1}^d [x_i, y_i]$. To see this, note the following: let B be of the form $[x, y]$. Then as $\mathrm{diam}(B) \downarrow 0$ (while varying y, not x),

$$\int_B f_x = \sum_{M \in \mathcal{M}} \int_{MB} f = \int_{\cup_{M \in \mathcal{M}} MB} f \sim f(x) \int_{\cup_{M \in \mathcal{M}} MB} dz = f(x) 2^d \int_B dz = f_x^*(x) \int_B dz \ .$$

This little excursion allows us to study the behavior of f_x^* instead of f. Interestingly, as $\prod_{i=1}^d |x_i - X_{ji}| = \prod_{i=1}^d |M(x_i - X_{ji})|$ for all matrices M, we see that the same flipping applied to our estimator f_n leaves f_n unaltered. Thus, to show that $f_n \to f$ at almost all x is equivalent to showing that $f_n(x) \to f_x^*(x)/2^d$ at almost all x.

The remainder of the proof requires the introduction of the marginal densities f_S and $f_{x,S}^*$, where $S \subseteq \{1, 2, \ldots, d\}$: thus, f_S is the marginal density of f with respect to all components whose index is in S, and similarly, $f_{x,S}^*$ is the marginal density of f_x^* with respect to all components whose index is in S. We call x a JMZ point (after Jessen, Marcinkiewicz and Zygmund) if for all $S \neq \emptyset$,

$$\lim_{B \in \mathcal{B}(x_S) : \mathrm{diam}(B) \downarrow 0} \frac{\int_B f_S(y)\, dy}{\int_B dy} = f_S(x_S) \ .$$

Here x_S is the $|S|$-dimensional vector composed of components of x whose index is in S, and $\mathcal{B}$ is the collection of rectangles in this $|S|$-dimensional space. By Lemma A1, almost all points are JMZ points. Fix such an x for the remainder of the proof. For $\epsilon > 0$, we can thus find $\delta > 0$ such that simultaneously for all $S \neq \emptyset$, for all $y_S \leq x_S + (\delta, \delta, \ldots, \delta)_S$,

$$\left| \frac{\int_{[x_S, y_S]} f_{x,S}^*(z)\, dz}{\int_{[x_S, y_S]} dz} - f_{x,S}^*(x_S) \right| < \epsilon f_{x,S}^*(x_S) \ .$$

Also, recall that $f_{x,S}^*(x_S) = 2^{|S|} f_S(x_S)$.

LEMMA B1. *Let $x = (x_1, \ldots, x_d)$ be a JMZ point of f with $f(x) > 0$. Let ϵ and δ be as above. Let $X = (X_1, \ldots, X_d)$ be a random vector with density f^*. Let $S \subseteq \{1, \ldots, d\}$ be a nonempty set of indices. Let A be the event that $x_i < X_i < x_i + \delta$ for all $i \in S$. Then, conditional on A, there exist i.i.d. uniform $[0,1]$ random variables U_i such that*

$$\frac{1-\epsilon}{(1+\epsilon)\prod_{i\in S} U_i} \preceq \prod_{i\in S} \frac{\delta}{X_i - x_i} \preceq \frac{1+\epsilon}{(1-\epsilon)\prod_{i\in S} U_i} ,$$

where $\preceq$ denotes stochastic domination.

PROOF. We assume without loss of generality that $S = \{1, \ldots, d\}$ and that $x = 0$. We write f^* instead of f^*_x and f^*_S instead of $f^*_{x,S}$. Conditional on A, X has density f^*/p, where $p = \int_B f^*$ and $B = [0, \delta]^d$. Note that

$$p \in [1-\epsilon, 1+\epsilon] f^*_S(0) \delta^d .$$

Furthermore, if F is the multivariate distribution function for the conditional X, then as x is a JMZ point,

$$\frac{(1-\epsilon) f^*_S(0)}{p} \le \frac{F(X_1, \ldots, X_d)}{\prod_{i=1}^d X_i} \le \frac{(1+\epsilon) f^*_S(0)}{p} .$$

and therefore, as it is well-known that $F(X_1, X_2, \ldots, X_d) \stackrel{\mathcal{L}}{=} \prod_{i=1}^d U_i$, which can be seen by applying the probability integral transform to each conditional distribution function in the conditional decomposition of F, we see that

$$\frac{1-\epsilon}{(1+\epsilon)\delta^d} \preceq \prod_{i=1}^d \frac{U_i}{X_i} \preceq \frac{1+\epsilon}{(1-\epsilon)\delta^d} . \square$$

We now return to the proof of our theorem, where ϵ and δ remain as defined earlier. We enlarge the data by considering an infinite i.i.d. sequence $Y_1, Y_2, \ldots$, all distributed as X. Let $B(Y_j) \subseteq \{1, \ldots, d\}$ be the collection of indices i in Y_{ji} with $Y_{ji} \in [x_i, x_i + \delta]$, where $x = (x_1, \ldots, x_d)$ is a JMZ point. Let $S \ne \emptyset$, and let T be the collection of the first n indices j with $B(Y_j) \equiv S$. We have

$$\sum_{j\in T} \frac{1}{\prod_{i=1}^d (Y_{ji} - x_i)} \le \frac{1}{\delta^d} \sum_{j\in T} \frac{\delta^{|S|}}{\prod_{i\in S}(Y_{ji} - x_i)} \preceq \frac{1}{\delta^d} \sum_{j\in T} \frac{1+\epsilon}{(1-\epsilon)\prod_{i\in S} U_i}$$

by Lemma B1, where $U_1, \ldots, U_d$ are as in Lemma B1. By Lemma S1, the right-hand-side is in probability asymptotic to

$$\frac{(1+\epsilon) n \log^{|S|} n}{(1-\epsilon)|S|!\delta^d} .$$

Thus, the contribution of all Y_j, $j \in T$, when $|S| < d$, is asymptotically negligible. Assume first $f^*_x > 0$. Then, taking $|S| = d$, we can no longer afford to artificially increase the data size as we did above. Thus, let T now be those $j \le n$ for which $|B(Y_j)| = d$. Note that T is binomial (n, p), where $p = \mathbf{P}\{X - x \in [0, \delta]^d\}$. By the independence of T and the U_i's in Lemma B1, it is easy to see that

$$\sum_{j\in T} \frac{1}{\prod_{i=1}^d (Y_{ji} - x_i)} \preceq \frac{1}{\delta^d} \sum_{j\in T} \frac{1+\epsilon}{(1-\epsilon)\prod_{i=1}^d U_i}$$

and the right-hand-side is in probability asymptotic to

$$\begin{aligned} \frac{(1+\epsilon)(np)\log^d(np)}{\delta^d(1-\epsilon)d!} &\le \frac{(1+\epsilon)^2 n f^*_x(x)\delta^d \log^d n}{\delta^d(1-\epsilon)d!} \\ &\le \frac{(1+\epsilon)^2 n 2^d f(x) \log^d n}{(1-\epsilon)d!} . \end{aligned}$$

A similar weak lower bound is obtained, and by letting $\epsilon \downarrow 0$, we obtain the sought result. The contribution of those terms in the density estimate with $|S| = d$ is $o(n \log^d n)$ when $f(x) = 0$ (as seen from the last chain of inequalities as well), just as with all S of size less than d. Thus, the proof of theorem 2 is complete. □

5. Lack of strong convergence.

For all f, it is true that at almost all x with $f(x) > 0$, the Hilbert product kernel estimates cannot possibly converge to f in a strong sense. Rather than to prove the full-blown universal theorem, we restrict ourselves to the uniform density on the real line and recall the following result from Devroye and Krzyżak (1998), which is applicable as for $d = 1$, the Hilbert product kernel estimate coincides with the standard Hilbert kernel estimate.

THEOREM 3. *Let f be the uniform density on $[0,1]$. Then, for any $x \in [0,1]$, $\mathbf{P}\{f_n(x) \geq \log\log n \text{ i.o.}\} = 1$, so that there is no strong convergence at any point in the support.*

The poor rate of convergence of the estimate is best seen by considering points outside the support of f. If x is at least distance s away from the support of f, then $f_n(x) \geq c/(s^d(\log n)^d)$ for some constant c only depending upon k and d.

6. A product nearest neighbor estimate.

The k-nearest neighbor density estimate of Fix and Hodges (1951) and Loftsgaarden and Quesenberry (1965) is

$$g_{k,n}(x) \stackrel{\text{def}}{=} \frac{k}{nV_d \|x - X_{(k,x)}\|^d}$$

where $X_{(k,x)}$ is the k-th nearest neighbor of x among $X_1, \ldots, X_n$. Its properties are well-understood (Moore and Yackel, 1977; Devroye and Wagner, 1977; Mack, 1980; Bhattacharya and Mack, 1987; Mack and Rosenblatt, 1979). For example, at almost all x, we have $g_{k,n}(x) \to f(x)$ as $n \to \infty$ if $k = o(n)$ and $k \to \infty$. The k-nearest neighbor density is not scale-invariant because the relative order of the distances $\|x - X_j\|, 1 \leq j \leq n$ changes when the coordinates are linearly scaled. To remedy this, we introduce the product k-nearest neighbor density estimate

$$g_{k,n}(x) = \frac{k\,(d-1)!}{n \log^{d-1}(n/k(d-1)!) 2^d \prod_{i=1}^d |x_i - X_{(k)i}|}$$

where $x = (x_1, \ldots, x_d)$, and $X_{(1)}, \ldots, X_{(n)}$ is a permutation of $X_1, \ldots, X_d$ according to increasing values of $\prod_{i=1}^d |x_i - X_{(j)i}|, 1 \leq j \leq n$. This permutation is invariant under linear transformations of the coordinate axes (but not rotations). Interestingly, this product estimate has not been considered before except in the trivial case $d = 1$, where we obtain the standard univariate k-th nearest neighbor estimate. As the choice of scale is a perpetual cause of concern in estimation, the product k-th nearest neighbor estimate should be particularly useful. We will prove its weak consistency:

THEOREM 4. *If $k \to \infty$ such that $k \to \infty, k/\log n \to 0$, and if $\int f \log^{d-1}(f+1) < \infty$, and if $\int g \log^{d-1}(g+1) < \infty$ for all lower-dimensional marginals of f, then at almost all x (that is, at all* JMZ *points for f and all lower-dimensional marginals f), $g_{k,n}(x) \to f(x)$ in probability.*

PROOF OF THEOREM 4. We only sketch a rough outline. Mimicking the proof of Theorem 2, we note first that we may wish to consider the flipped density at x, which is $2^d f(y)$, with $y \geq x$, all coordinates of y are at least equal to those for x. We will consider a small square of size ϵ in each coordinate with bottom lower vertex at x. We will show that the k-th nearest neighbor of x is with probability tending to one in this square. Indeed, if $Y + (Y_1, \ldots, Y_d)$ has density f, and x is a JMZ point, then $\prod_{i=1}^d |Y_i - x_i|$ is approximately distributed as $\prod_{i=1}^d U_i / f^*(x)$, where the U_i's are i.i.d. uniform $[0,1]$ random variables, and $f^*(x) = 2^d f(x)$. As $\prod_{i=1}^d U_i$ has density $\log^{d-1}(1/z)/(d-1)!$, $z > 0$, and distribution function $\sim z \log^{d-1}(1/z)/(d-1)!$ as $z \to 0$, we see that the order statistics of a sample of size n drawn from $\prod_{i=1}^d |Y_i - x_i|$ are approximately distributed as $F^{\text{inv}}(1/n)/f^*(x)$, $F^{\text{inv}}(2/n)/f^*(x)$, and so forth, where F^{inv} is the inverse distribution function of $\prod_{i=1}^d U_i$. A good approximation is $F^{\text{inv}}(u) \sim u(d-1)!/(\log^{d-1}(1/(u(d-1)!)))$. Thus, the k-th nearest neighbor has a value $\prod_{i=1}^d |Y_i - x_i|$ concentrated in probability about

$$\frac{(d-1)!k/n}{f^*(x) \log^{d-1}(n/(k(d-1)!))} \,. \tag{D}$$

The concentration follows from $k \to \infty$. We only need to show that the probability that the k-th nearest neighbor is in the square tends to one. To this end, we show that the nearest neighbor among all points that have m coordinates outside the square and $d - m$ coordinates in the square is at distance (always measured by $\prod_{i=1}^d |Y_i - x_i|$) asymptotically much larger than (D). Without loss of generality, fix the first $m \geq 1$ ($m < d$) coordinates. If x is also a JMZ point for the marginal density for the last $d - m$ coordinates, then given that the first m coordinates are outside the square and the remaining ones inside (and assuming that ϵ is small enough), the nearest neighbor distance is asymptotically of the order of $1/(n \log^{d-m-1} n)$ and thus it is improbable that the k-th nearest neighbor point can have any coordinate outside the small square. This concludes the sketch of the proof. $\square$

7. Product kernel estimates.

In this section, we consider product kernel estimates defined by

$$f_n(x) = \frac{1}{n} \sum_{j=1}^n \prod_{i=1}^d \frac{1}{h_{n,i}} K_i \left(\frac{x_i - X_{j,i}}{h_{n,i}} \right)$$

where $x = (x_1, \ldots, x_d)$, and $K_1, \ldots, K_d$ are univariate kernels with $\int K_i = 1$. The smoothing factors $h_{n,i}$ are for now deterministic sequences. Product kernel estimates are of interest because they offer scale invariance if the $h_{n,i}$'s are proportional to scale (e.g., make $h_{n,i}$ proportional to a weighted sum of the pairwise distances $|X_{j,i} - X_{m,i}|, 1 \leq j, m \leq n$). This may introduce big differences between the $h_{n,i}$'s. In fact, such differences may be desirable in situations like this one: let $d = 2$ and let f be the product of two univariate densities, a smooth one and a jagged one. For each density, one may want to pick different smoothing factors, and even different orders for the kernel. In those cases, $h_{n,1}$ and $h_{n,2}$ may tend to zero at different rates. As each density is locally a product density (as it resembles a uniform density), it is really important to consider product kernels with d individually picked smoothing parameters.

It is interesting that the literature offers little help with respect to the universal consistency properties of these estimates with respect to pointwise or L_1 convergence. The "natural" conditions on the $h_{n,i}$'s would appear to be

$$\lim_{n \to \infty} \max_i h_{n,i} = 0 \tag{1}$$

and

$$\lim_{n\to\infty} n \prod_{i=1}^{d} h_{n,i} = \infty \ . \tag{2}$$

In this section, we prove two basic consistency results, which we have not been able to find in the vast literature.

THEOREM 5. *Let f_n be the product kernel estimate, and let each component kernel K_i be absolutely integrable. Then under the natural conditions (1) and (2),*

$$\lim_{n\to\infty} \mathbf{E} \int |f_n - f| = 0$$

for all f.

THEOREM 6. *Let f_n be the product kernel estimate, and let all kernels K_i be bounded, of compact support, and Riemann approximable, that is, in the L_∞ sense, each K_i is in the closure of the space of functions that are finite weighted sums of indicators of finite intervals. Then, under the natural conditions (1) and (2),*

$$\lim_{n\to\infty} \mathbf{E}|f_n(x) - f(x)| = 0$$

at almost all x provided that $\int f \log^{d-1}(1+f) < \infty$.

NOTE. The conditions on the kernels are satisfied by kernels K_i that are continuous, bounded and of compact support. With a bit of effort, we can extend Theorem 6 to include kernels K_i with $K_i(x) = O(1/|x|)$ as $|x| \to \infty$

Perhaps the easiest proof of Theorem 5, and the most transparent one, uses the embedding device from Devroye (1985) (where the embedding is used to handle the L_1 consistency of variable kernel estimates), which may be summarized in the following Lemma:

LEMMA B2. *Let f and g be two densities on $\mathbb{R}^d$, and let $f_n = f_n(x; x_1, \ldots, x_n)$ be a density estimate based on an i.i.d. samples of size n drawn from f. Assume that*

$$\sup_{j, x_1, \ldots, x_n, x_j'} \int |f_n(x; x_1, \ldots, x_{j-1}, x_j, x_{j+1}, \ldots, x_n) - f_n(x; x_1, \ldots, x_{j-1}, x_j', x_{j+1}, \ldots, x_n)| \le \frac{C}{n}$$

for some constant C. Then, if g_n is the density estimate based upon an i.i.d. sample of size n drawn from g, we have

$$\mathbf{E} \int |f_n - f| \le (C+1) \mathbf{E} \int |g_n - g| + \int |f - g| \ .$$

PROOF. Consider a uniform Poisson point process on $\mathbb{R}^d \times [0,\infty)^2$. Let (U,V,T) be a typical point in this process. Keep only those points with $V < f(U)$ or $V < g(U)$. For $T < t$, there is almost surely a finite number of such points (with a Poisson $(2t)$ distribution), so that we may order the points according to increasing values of T, obtaining $(U_1, V_1, T_1), \ldots$. Let $X_1, X_2, \ldots, X_n$ be the first n values of U_j for which $V_j < f(U_j)$. Let $Y_1, Y_2, \ldots, Y_n$ be the first n values of U_j for which $V_j < g(U_j)$. Now, throw away the Poisson point process, which was only needed to couple the two samples. Interestingly, $X_1, \ldots, X_n$ is i.i.d. and drawn from f and $Y_1, \ldots, Y_n$ is i.i.d. and drawn from g. Also, for every i,

$$
\begin{aligned}
\mathbf{P}\{V_j < \min(f(U_j), g(U_j))\} &= \mathbf{E}\left\{\frac{\min(f(U_j), g(U_j))}{\max(f(U_j), g(U_j))}\right\} \\
&= \int \frac{\max(f(x), g(x))}{\int \max(f,g)} \frac{\min(f(x), g(x))}{\max(f(x), g(x))}\, dx \\
&= \int \frac{\int \min(f,g)}{\int \max(f,g)} \\
&\stackrel{\text{def}}{=} p\ .
\end{aligned}
$$

Each one of the triples (U_j, V_j, T_j) with $V_j < \min(f(U_j), g(U_j))$ generates a common point in both samples, and thus, the number of common points is stochastically greater than a binomial (n,p) random variable. The number of points in one sample not seen in the other sample is not more than a binomial $(n, 1-p)$ random variable, and its expected value does not exceed $n(1-p) = n\int|f-g|/(1+(1/2)\int|f-g|) \le n\int|f-g|$. We may assume that f_n and g_n are based on the (coupled) X_j and Y_j samples respectively. Then

$$
\mathbf{E}\int|f_n - f| \le \mathbf{E}\int|f_n - g_n| + \mathbf{E}\int|g_n - g| + \mathbf{E}\int|g-f| \le (C+1)\int|f-g| + \mathbf{E}\int|g_n - g|
$$

as $\int|f_n - g_n| \le \frac{C(n-N)}{n}$ by applying the triangle inequality $n-N$ times. □

PROOF OF THEOREM 5. We first verify that Lemma B2 applies to the kernel estimate with kernel K. Indeed, if $y = (y_1, \ldots, y_d)$ and $x_j = (x_{j1}, \ldots, x_{jd})$, then

$$
\begin{aligned}
&\int |f_n(y; x_1, \ldots, x_{j-1}, x_j, x_{j+1}, \ldots, x_n) - f_n(y; x_1, \ldots, x_{j-1}, x'_j, x_{j+1}, \ldots, x_n)| \\
&\le \frac{1}{n}\int\left|\prod_{i=1}^d h_{n,i}^{-1} K_i\left(y_i - x_{ji}\right)\right| + \frac{1}{n}\int\left|\prod_{i=1}^d h_{n,i}^{-1} K_i\left(y_i - x'_{ji}\right)\right| \\
&= \frac{2}{n}\prod_{i=1}^d \int |K_i|
\end{aligned}
$$

so that Lemma B2 applies with $C = 2\prod_{i=1}^d \int|K_i|$. If all kernels K_i are nonnegative, then simply, $C=2$. By Lemma 2 then, it suffices to prove theorem 5 for all continuous densities g of compact support, as those densities are dense in the L_1 space of all densities. Indeed, pick g continuous and of compact support such that $\int|f-g| < \epsilon$. Let f_n and g_n denote the product kernel estimate, but based on samples drawn from densities f and g respectively. Applying Lemma B2, we have

$$
\mathbf{E}\int|f_n - f| \le (C+1)\int|f-g| + \mathbf{E}\int|g_n - g|\ .
$$

As ϵ is arbitrary, it suffices therefore to prove Theorem 5 for all such g.

Since each K_i is measurable and absolutely integrable, we may approximate it in the L_1 sense by a sum of indicator functions of intervals. Thus, for $\epsilon > 0$, we find a finite number of intervals $[a_{ij}, b_{ij}]$

and coefficients k_{ij} such that

$$\int |K_i - L_i| < \epsilon \ ,$$

where

$$L_i \stackrel{\text{def}}{=} \sum_j k_{ij} I_{[a_{ij}, b_{ij}]} < \epsilon \ .$$

The L_i can even be picked such that $\int L_i = 1$. Note also that $\int |L_i| < \int |K_i| + \epsilon$, and that $\int |K_i| \geq 1$ (as $\int K_i = 1$). Define the constant

$$D = \prod_{i=1}^{d} \left(\int |K_i| + \epsilon \right) \ .$$

But if f_n and g_n are two product kernel estimates with the same data but different product kernels K_i, L_i, $1 \leq i \leq d$, then

$$\begin{aligned}
\int |f_n - g_n| &\leq \int \left| \prod_{i=1}^{d} K_i - \prod_{i=1}^{d} L_i \right| \\
&\leq \int \left| \prod_{i=1}^{d} K_i - L_1 \prod_{i=2}^{d} L_i \right| + \int \left| L_1 \prod_{i=2}^{d} K_i - L_1 L_2 \prod_{i=3}^{d} L_i \right| + \cdots + \int \left| K_d \prod_{i=1}^{d-1} L_i - \prod_{i=1}^{d} L_i \right| \\
&\leq \int |K_1 - L_1| \prod_{i=2}^{d} \int |K_i| + \int |K_2 - L_2| \int |L_1| \prod_{i=3}^{d} \int |K_i| + \cdots + \int |K_d - L_d| \int |K_d| \prod_{i=1}^{d-1} \int |L_i| \\
&\leq \epsilon D \ .
\end{aligned}$$

Again, by the arbitrary nature of ϵ, it suffices to consider kernels that are products of finite sums of weighted indicator functions of intervals. Let N be a bound on the number of indicators for any of the component kernels. It is easy to see then, by forming the Cartesian grid of these intervals, that such an estimate is equivalent to a kernel estimate with kernel

$$L(x) = \sum_{j=1}^{N^d} \alpha_j I_{[a_j, b_j]}(x) \ ,$$

where $[a_j, b_j]$ is the shorthand notation for a rectangle of $\mathbb{R}^d$ with vertices a_j and b_j, and α_j is a real number. Also, $\int L = 1$ and $\int |L| \leq D$. Let p_j denote the volume of $[a_j, b_j]$. Introduce the notation $M_j = I_{[a_j, b_j]}/p_j$, and note that M_j is a bona fide kernel with integral one. Then we have, letting f_n denote the kernel estimate with kernel L, and f a continuous density of compact support, as $\sum_j \alpha_j p_j = 1$,

$$\int |f_n - f| \leq \sum_{j=1}^{N^d} |\alpha_j| p_j \int |f_{nj} - f| \ ,$$

where f_{nj} is the kernel estimate with kernel M_j. The upper bound tends to zero in the mean if each individual term tends to zero. Thus, it suffices to prove the Theorem for kernels that are indicator functions of rectangles.

But by Theorem 6, we have $\mathbf{E}|f_{nj} - f| \to 0$ for all bounded f at almost all x. But then

$$\mathbf{E} \int |f_{nj} - f| = 2\mathbf{E} \int (f - f_{nj})_+ = 2 \int \mathbf{E}(f - f_{nj})_+ \to 0$$

by dominated convergence. This concludes the proof of Theorem 5. □

Proof of Theorem 6. Let x be a JMZ point for f. Let the kernel $K = cI_R$ be proportional to an indicator a bounded rectangle $R = [a, b]$ not necessarily containing the origin, where $c = 1/\lambda(R)$ and

$a = (a_1, \ldots, a_d)$ and $b = (b_1, \ldots, b_d)$ denote the vertices of R. By varying each coordinate in turn, we see that

$$K = c \sum_{i=1}^{d} s_i I_{[0,z_i]}$$

where z_i are points of $\mathbb{R}^d$ and $s_i = \pm 1$. Denote the volume of $[0, z_i]$ by p_i. Let M_n denote the $d \times d$ diagonal matrix with elements $h_{n,1}, \ldots, h_{n,d}$ on the diagonal. By Lemma A1 and condition (1),

$$\frac{\int_{x+M_n[0,z_i]} f}{\lambda(x + M_n[0, z_i])} \to f(x)$$

and thus

$$\frac{c \int_{x+M_n R} f}{\lambda(x + M_n R)} \to c f(x) \sum_{i=1}^{d} \frac{s_i \lambda(x + M_n[0, z_i])}{\lambda(x + M_n R)} = f(x) \ ,$$

because $\int K = 1$. We conclude that $\mathbf{E} f_n(x) = \int_{x+M_n R} f / \lambda(x + M_n R) \to f(x)$, and that

$$\begin{aligned} \mathrm{Var} f_n(x) &= \frac{\mathrm{Var}\{c I_{x+M_n R}(X_1)\}}{n} \\ &\le \frac{\mathbf{E}\{c^2 I_{x+M_n R}(X_1)\}}{n} \\ &= \frac{c^2 \lambda(x + M_n R) \mathbf{E} f_n(x)}{n} \\ &= \frac{f(x) + o(1)}{\lambda(R) n \prod_{i=1}^{d} h_{n,i}} \end{aligned}$$

and therefore,

$$\begin{aligned} \mathbf{E}\left\{|f_n(x) - f(x)|\right\} &\le \mathbf{E}\left\{|f_n(x) - \mathbf{E} f_n(x)|\right\} + |\mathbf{E} f_n(x) - f(x)| \\ &\le \sqrt{\mathrm{Var}\left\{f_n(x) - \mathbf{E} f_n(x)\right\}} + \left| \frac{\int_{x+M_n R} f}{\lambda(M_n R)} - f(x) \right| \\ &= \sqrt{\frac{f(x) + o(1)}{\lambda(R) n \prod_{i=1}^{d} h_{n,i}}} + o(1) \\ &= o(1) \end{aligned}$$

if $f(x) > 0$. When $f(x) = 0$ and x is a JMZ point, then

$$\mathbf{E}\left\{|f_n(x) - f(x)|\right\} = \mathbf{E} f_n(x) = \frac{\int_{x+M_n R} f}{\lambda(x + M_n R)} \to f(x) = 0$$

by Lemma A1 and (2).

We now turn to more general kernels. It will take some work to generalize the above results to a reasonably big class of kernels. It is easy to verify from the last chain of inequalities that under condition (2), for any bounded kernel, $\mathrm{Var}\{f_n(x)\} \to 0$ at points x at which $\mathbf{E} f_n(x) \to f(x)$. Thus, Theorem 6 is valid for all bounded kernels for which (1) implies that $\mathbf{E} f_n(x) \to f(x)$ at almost all x. Define $H = \prod_{i=1}^{d} h_{n,i}$. We find for each $\epsilon > 0$ a finite collection of rectangles R_i and constants α_i such that

$$\sup_x \left| K(x) - \sum_i \alpha_i I_{R_i}(x) \right| < \epsilon I_{[-s,s]^d}(x) \ ,$$

where $s > 0$ is a large positive integer. From this, by integration, we note that

$$\left| 1 - \sum_i \alpha_i \lambda(R_i) \right| = \left| \int K(x) - \sum_i \alpha_i \lambda(R_i) \right| < \epsilon \lambda([-s,s]^d) = \epsilon (2s)^d \ .$$

Then

$$\begin{aligned}
&|\mathbf{E} f_n(x) - f(x)| \\
&\le \frac{\int f(y) \left| K(M_n^{-1}(x-y)) - \sum_i \alpha_i I_{R_i}(M_n^{-1}(x-y)) \right| dy}{H} + \left| \frac{\int f(y) \sum_i \alpha_i I_{R_i}(M_n^{-1}(x-y)) dy}{H} - f(x) \right| \\
&\le \epsilon \frac{\int f(y) I_{[-s,s]^d}(M_n^{-1}(x-y)) dy}{H} + \left| \sum_i \alpha_i \frac{\int (f(y) - f(x)) I_{R_i}(M_n^{-1}(x-y)) dy}{H} \right| + f(x) \left| \sum_i \alpha_i \lambda(R_i) - 1 \right| \\
&= \epsilon (2s)^d (f(x) + o(1)) + o(1) + f(x) \epsilon (2s)^d \ ,
\end{aligned}$$

where in the last step, we used (1) and the first part of the proof for kernels that are indicators of rectangles. Since ϵ was arbitrary, the proof is complete. □

8. Tree-based and rectangular partitioning density estimates.

Let $\mathcal{A}_1, \mathcal{A}_2, \ldots$ be a sequence of partitions of $\mathbb{R}^d$ into rectangles (which do not have to be open or closed). For x, let $A_n(x)$ denote the rectangle in $\mathcal{A}_n$ to which x belongs. Then the partitioning estimate of f is given by

$$f_n(x) = \frac{\# A_n(x)}{n \lambda(A_n(x))} \ ,$$

where $\lambda(.)$ is Lebesgue measure, and $\#(.)$ denotes the number of data points falling in a set. We assume for now that the sequence of partitions is picked before the data are collected. The present estimates contain all standard histogram estimates and indeed most tree-based density estimates. The purpose of this section is to discuss its universal consistency, pointwise and in L_1. For L_1 consistency, there is a rather general theorem by Abou-Jaoude (1976a, 1976c) of which Theorem 7 below is a special case. However, we include it here, as our proof is short and uses new tools. The L_∞ convergence was dealt with by Abou-Jaoude (1976b), but it is irrelevant here. There are just a few general consistency theorems for partitioning estimates, such as estimates that partition the space via order statistics (Hanna and Abou-Jaoude, 1981) or via multivariate rectangular partitions (Gessaman, 1970). The most difficult problem, from a universal convergence point of view, is the pointwise convergence. We give just such a theorem below.

THEOREM 7. *Let f_n be a partitioning density estimate. Assume that* $\mathrm{diam}(A_n(x)) \to 0$ *at almost all x, and that $n\lambda(A_n(x)) \to \infty$ at almost all x. Then*

$$\lim_{n\to\infty} \mathbf{E} \int |f_n - f| = 0 \ .$$

(There is no condition on f.)

THEOREM 8. *Let f_n be a partitioning density estimate. Assume that* $\mathrm{diam}(A_n(x)) \to 0$ *at almost all x, that $n\lambda(A_n(x)) \to \infty$ at almost all x, and that $\int f \log^{d-1}(1+f) < \infty$. Then*

$$\mathbf{E}|f_n(x) - f(x)| \to 0$$

at almost all x.

PROOF. Let x be a JMZ point for f. Then

$$\begin{aligned}
\mathbf{E}\{|f_n(x) - f(x)|\} &\le \mathbf{E}\{|f_n(x) - \mathbf{E}f_n(x)|\} + |\mathbf{E}f_n(x) - f(x)| \\
&\le \sqrt{\mathrm{Var}\{f_n(x) - \mathbf{E}f_n(x)\}} + \left|\frac{\int_{A_n(x)} f}{\lambda(A_n(x))} - f(x)\right| \\
&= \sqrt{\frac{n\int_{A_n(x)} f\left(1 - \int_{A_n(x)} f\right)}{\left(n\int_{A_n(x)} f\right)^2}} + o(1) \\
&\le \frac{1}{\sqrt{n\int_{A_n(x)} f}} + o(1) \\
&= \frac{1}{\sqrt{n(f(x) + o(1))\lambda(A_n(x))}} + o(1) \\
&= o(1)
\end{aligned}$$

if $f(x) > 0$. Thus, by Lemma A1, the theorem is proved at almost all points with $f(x) > 0$. When $f(x) = 0$ and x is a JMZ point, then

$$\mathbf{E}\{|f_n(x) - f(x)|\} = \mathbf{E}f_n(x) = \frac{\int_{A_n(x)} f}{\lambda(A_n(x))} \to f(x) = 0$$

by Lemma A1 again. □

PROOF OF THEOREM 7. We simply use the result of Theorem 8 and bounded convergence: indeed, if $f_n \to f$ at almost all x, and each f_n is a bona fide (deterministic) density, then $\int |f_n - f| \to 0$. Glick (1974) (see also Devroye and Györfi, 1985) has shown that we may add the phrases "in probability" or "almost surely" on both sides in case f_n is a data-dependent sequence of estimates. Now note that the condition of Lemma B2 is satisfied for any partition estimate in which the partition does not depend upon the data with $C = 2$: indeed, consider two determistic samples differing in only the i-th point, and let the corresponding partitioning estimates be called f_n and g_n. Let $\#A_n(x)$ and $\#\#A_n(x)$ denote the cardinalities of $A_n(x)$ under both samples. Then

$$\begin{aligned}
\int |f_n - g_n| &= \sum_{A\in\mathcal{A}_n} \int_A |f_n - g_n| \\
&= \frac{1}{n}\sum_{A\in\mathcal{A}_n} |\#A - \#\#A| \\
&= \frac{1}{n}|\#A_n(x_i) - \#\#A_n(x_i)| + \frac{1}{n}|\#A_n(x_i') - \#\#A_n(x_i')| \\
&\le \frac{2}{n}\,.
\end{aligned}$$

So, now let f_n and g_n denote the same partitioning estimate, but based on samples drawn from densities f and g respectively, where g is a density that will be picked later. Applying Lemma B2, we have

$$\mathbf{E}\int |f_n - f| \le 3\int |f-g| + \mathbf{E}\int |g_n - g| \ .$$

At this point, we pick $g = \min(f,M)/\int \min(f,M)$ where M is picked so large that that $\int |f-g| \le \epsilon$. By Theorem 8, at almost all x, $\mathbf{E}|g_n(x) - g(x)| \to 0$, as g is bounded. but then $\mathbf{E}\int |g_n - g| = 2\mathbf{E}\int (g-g_n)_+ \to 0$ by bounded convergence. Thus,

$$\mathbf{E}\int |f_n - f| \le 3\epsilon + o(1) \ .$$

As ϵ was arbitrary, the proof is complete. □

9. Acknowledgment

The authors thank an attentive referee for pointing out a shortcut in a proof.

10. References

S. Abou-Jaoude, "Conditions nécessaires et suffisantes de convergence L_1 en probabilité de l'histogramme pour une densité," Annales de l'Institut Henri Poincaré, vol. 12, pp. 213–231, 1976a.

S. Abou-Jaoude, "La convergence L_1 et L_∞ de l'estimateur de la partition aléatoire pour une densité," Annales de l'Institut Henri Poincaré, vol. 12, pp. 299–317, 1976b.

S. Abou-Jaoude, "Sur une condition nécessaire et suffisante de L1-converence presque complète de l'estimateur de la partition fixe pour une densité," Comptes Rendus de l'Académie des Sciences de Paris, vol. 283, pp. 1107–1110, 1976c.

H. Akaike, "An approximation to the density function," Annals of the Institute of Statistical Mathematics, vol. 6, pp. 127–132, 1954.

J. Beirlant and L. Györfi, "On the L_1-Error in Histogram Density Estimation: The Multidimensional Case", Nonparametric Statistics, vol. 9, pp. 197–216, 1999.

N. H. Bingham, C. M. Goldie and J. L. Teugels, Regular Variation, Cambridge University Press, Cambridge, 1987.

M. de Guzman, Real variable Methods in Fourier Analysis, North-Holland, Amsterdam, 1981.

L. Devroye, "The equivalence of weak, strong and complete convergence in L1 for kernel density estimates," Annals of Statistics, vol. 11, pp. 896–904, 1983.

L. Devroye, "A note on the L1 consistency of variable kernel estimates," Annals of Statistics, vol. 13, pp. 1041–1049, 1985.

L. Devroye and L. Györfi, Nonparametric Density Estimation: The L1 View, John Wiley, New York, 1985.

L. Devroye, L. Györfi and A. Krzyżak, "The Hilbert kernel regression estimate," Journal of Multivariate Analysis, vol. 65, pp. 209–227, 1998.

L. Devroye and A. Krzyżak, "On the Hilbert density estimate," Statistics and Probability Letters, vol. 44, pp. 299–308, 1999.

L. P. Devroye and T. J. Wagner, "The strong uniform consistency of nearest neighbor density estimates," Annals of Statistics, vol. 5, pp. 536–540, 1977.

J. El Helou, Recouvrement du tore T^q par des ouverts aléatoires, Technical Report, 1978.

E. Fix and J. L. Hodges, Discriminatory analysis, nonparametric discrimination, consistency properties, Report 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

M. P. Gessaman, "A consistent nonparametric multivariate density estimator based on statistically equivalent blocks," Annals of Mathematical Statistics, vol. 41, pp. 1344–1346, 1970.

N. Glick, "Consistency conditions for probability estimators and integrals of density estimators," Utilitas Mathematica, vol. 6, pp. 61–74, 1974.

B. Hanna and S. Abou-Jaoude, "Sur la vitesse de convergence de l'estimateur de la partition aléatoire d'une densité de probabilité," Publications de l'Institut de Statistique de l'Université de Paris, vol. 26, pp. 51–67, 1981.

B. Jessen, J. Marcinkiewicz and A. Zygmund, "Note on the differentiability of multiple integrals," Fund. Math., vol. 25, pp. 217–234, 1935.

D. O. Loftsgaarden and C. P. Quesenberry, "A nonparametric estimate of a multivariate density function," Annals of Mathematical Statistics, vol. 36, pp. 1049–1051, 1965.

Y. P. Mack, "Asymptotic normality of multivariate k–NN density estimates," Sankhya, vol. A42, pp. 53–63, 1980.

Y. P. Mack and M. Rosenblatt, "Multivariate k-nearest neighbor density estimates," Journal of Multivariate Analysis, vol. 9, pp. 1–15, 1979.

J. M. Marstrand, "A counter-example in the theory of strong differentiation," Bulletin of the London Mathematical Society, vol. 9, pp. 209–211, 1977.

D. S. Moore and J. W. Yackel, "Consistency properties of nearest neighbor density function estimators," Annals of Statistics, vol. 5, pp. 143–154, 1977.

E. Parzen, "On the estimation of a probability density function and the mode," Annals of Mathematical Statistics, vol. 33, pp. 1065–1076, 1962.

B. A. Rogozin, "The distribution of the first ladder moment and height and fluctuation of a random walk," Theory of Probability and Applications, vol. 16, pp. 575–595, 1971.

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

S. Saks, "On the strong derivatives of functions of an interval," Fund. Math., vol. 25, pp. 235–252, 1935.

H. Scheffé, "A useful convergence theorem for probability distributions," Annals of Mathematical Statistics, vol. 18, pp. 434–458, 1947.

G. R. Shorack and J. A. Wellner, Empirical Processes with Applications to Statistics, John Wiley, New York, 1986.

R. L. Wheeden and A. Zygmund, Measure and Integral, Marcel Dekker, New York, 1977.