

# On nonparametric function estimation with infinite-order flat-top kernels

Dimitris N. Politis  
Department of Mathematics  
University of California at San Diego  
La Jolla, CA 92093-0112, USA  
politis@euclid.ucsd.edu

## **Abstract**

The problem of nonparametric estimation of a smooth, real-valued function of a vector argument. is addressed. In particular, we focus on a family of infinite-order smoothing kernels that is characterized by the flatness near the origin of the Fourier transform of each member of the family; hence, the term ‘flat-top’ kernels. Smoothing with the proposed infinite-order flat-top kernels has optimal Mean Squared Error properties. We review some recent advances, as well as give two new results on density estimation in two cases of interest: (i) case of a smooth density over a finite domain, and (ii) case of infinite domain with some discontinuities.

# 1 Introduction: a general family of flat-top kernels of infinite order.

Let  $f : R^d \rightarrow R$  be an unknown function to be estimated from data. In the typical nonparametric set-up, nothing is assumed about  $f$  except that it possesses a certain degree of smoothness. Usually, a preliminary estimator of  $f$  can be easily calculated that, however, lacks the required smoothness; e.g., in the case where  $f$  is a probability density. Often, the preliminary estimator is even inconsistent; e.g., in the case where  $f$  is a spectral density and the preliminary estimator is the periodogram. Rosenblatt (1991) discusses these two cases in an integrated framework.

In order to obtain an estimator (denoted by  $\hat{f}$ ) with good properties, for example, large-sample consistency and smoothness, one can smooth the preliminary estimator by convolving it with a function  $\Lambda : R^d \rightarrow R$  called the ‘kernel’, and satisfying  $\int \Lambda(x)dx = 1$ ; unless otherwise noted, integrals will be over the whole of  $R^d$ . It is convenient to also define the Fourier transform of the kernel as  $\lambda(s) = \int \Lambda(x)e^{i(s \cdot x)}dx$ , where  $s = (s_1, \dots, s_d)$ ,  $x = (x_1, \dots, x_d) \in R^d$ ,  $(s \cdot x) = \sum_k s_k x_k$  is the inner product between  $s$  and  $x$ .

Typically, as the sample size increases, the kernel  $\Lambda(\cdot)$  becomes more and more concentrated near the origin. To achieve this behavior, we let  $\Lambda(\cdot)$  and  $\lambda(\cdot)$  depend on a real-valued, positive ‘bandwidth’ parameter  $h$ , that is, we assume that  $\Lambda(x) = h^{-d}\Omega(x/h)$ , and  $\lambda(s) = \omega(hs)$ , where  $\Omega(\cdot)$  and  $\omega(\cdot)$  are some fixed (not depending on  $h$ ) bounded functions, satisfying  $\omega(s) = \int \Omega(x)e^{i(s \cdot x)}dx$ ; the bandwidth  $h$  will be assumed to be a decreasing function of the sample size.

If  $\Omega$  has finite moments up to  $q$ th order, and moments of order up to  $q - 1$  equal to zero, then  $q$  is called the ‘order’ of the kernel  $\Omega$ . If the unknown function  $f$  has  $r$  bounded continuous derivatives, it typically follows that

$$Bias(\hat{f}(x)) = E\hat{f}(x) - f(x) = c_{f,\Omega}(x)h^k + o(h^k), \quad (1)$$

where  $k = \min(q, r)$ , and  $c_f(x)$  is a bounded function depending on  $\Omega$ , on  $f$ , and on  $f$ ’s derivatives. Note that existence and boundedness of derivatives up to order  $r$  includes existence and boundedness of mixed derivatives of total order  $r$ ; cf. Rosenblatt (1991, p. 8).

This idea of choosing a kernel of order  $q$  in order to get the  $Bias(\hat{f}(x))$  to be  $O(h^k)$  dates back to Parzen (1962) and Bartlett (1963); see also Cacoullos (1966) for the multivariate case. Some more recent references on ‘higher-order’ kernels include the following: Devroye (1987), Gasser, Müller, and Mammitzsch (1985), Granovsky and Müller (1991), Jones (1995), Jones and Foster (1993), Marron (1994), Marron and Wand (1992), Müller (1988), Nadaraya (1989), Silverman(1986), and Scott (1992).

Note that the asymptotic order of the bias is limited by the order of the kernel if the true density is very smooth, i.e., if  $r$  is large. To avoid this limitation, one can define a ‘superkernel’ as a kernel whose order can be any positive integer; Devroye (1992) contains a detailed analysis of superkernels in the case of (univariate) probability density estimation. Thus, if  $f$  has  $r$  bounded continuous derivatives, a superkernel will result in an estimator with bias of order  $O(h^r)$ , no matter how large  $r$  may be; so, we might say that a superkernel is a kernel with ‘infinite order’.

However, it might be more appropriate to say that a kernel has ‘infinite order’ if it results in an estimator with bias of order  $O(h^r)$  no matter how large  $r$  may be *regardless of whether the kernel has finite moments*. It seems that the finite-moment assumption for  $\Omega$  is just a technical one, and that existence of the Lebesgue integrals used to calculate the moments is *not* necessarily required in order that a kernel has favorable bias performance; rather, it seems that if the integrals defining the moments of  $\Omega$  have a Cauchy principal value of zero then the favorable bias performance follows, and this is in turn ensured by setting  $\omega$  to be constant over an open neighborhood of the origin.

A preliminary report on a specific type of such infinite order kernel in the univariate case (that corresponds to an  $\omega$  of ‘trapezoidal’ shape) was given in Politis and Romano (1993); consequently, in Politis and Romano (1996, 1999) a general family of multivariate flat-top kernels of infinite order was proposed, and the favorable bias (and Mean Squared Error) properties of the resulting estimators were shown in the cases of probability and spectral density estimation.

Presently, we will propose a slightly bigger, more general class of multivariate flat-top kernels of infinite order with similar optimality properties—as will be shown in section 2. Finally, in section 3 we will address the interesting case where the unknown function  $f$  possesses the required smoothness only over a subset of the domain; in particular, we will investigate to what extent the performance of  $\hat{f}(x)$  is affected by a discontinuity of  $f$  (or its derivatives) at points away from  $x$ .

The general family of multivariate flat-top kernels of infinite order can be defined as follows.

**Definition 1.** *Let  $C$  be a compact, convex subset of  $R^d$  that contains an open neighborhood of the origin; in other words, there is an  $\epsilon > 0$  such that  $\epsilon D \subset C \subset \epsilon^{-1} D$ , where  $D$  is the Euclidean unit ball in  $R^d$ .*

*The kernel  $\Omega_C$  is said to be a member of the general family of multivariate flat-top kernels of infinite order if*

$$\Omega_C(x) = (2\pi)^{-d} \int \omega_C(s) e^{-i(s \cdot x)} ds,$$

*where the Fourier transform  $\omega_C(s)$  satisfies the following properties:*

(i)  $\omega_C(s) = 1$  for all  $s \in C$ ;

(ii)  $\int |\omega_C(s)|^2 ds < \infty$ ; and

(iii)  $\omega_C(s) = \omega_C(-s)$ , for any  $s \in R^d$ .

Property (i) guarantees the favorable bias properties and the ‘infinite’ order, while property (ii) ensures a finite variance of the resulting estimator  $\hat{f}$ ; finally, property (iii) guarantees that  $\Omega_C$  is real-valued.

In practically working with such a flat-top kernel, one must choose  $C$ . A typical choice for  $C$  is the unit ball in  $l_p$ , with some choice of  $p$  satisfying  $1 \leq p \leq \infty$ ; see Politis and Romano (1999). In addition, it is natural to impose the condition that  $\omega_C$  be a continuous function with the property  $|\omega_C(s)| \leq 1$ , for any  $s \in R^d$ . Nevertheless, only properties (i), (ii), (iii) are required for our results.

## 2 Multivariate density estimation: a review

Suppose  $X_1, \dots, X_N$  are independent, identically distributed (i.i.d.) random vectors taking values in  $R^d$ , and possessing a probability density function  $f$ ; the assumption of independence is not crucial here. The arguments apply equally well if the observations are stationary and weakly dependent, where weak dependence can be quantified through the use of mixing coefficients—see, for example, Györfi *et al.* (1989).

The objective is to estimate  $f(x)$  for some  $x \in R^d$ , assuming  $f$  possesses a certain degree of smoothness. In particular, it will be assumed that the characteristic function  $\phi(s) = \int e^{i(s \cdot x)} f(x) dx$  tends to zero sufficiently fast as  $\|s\|_p \rightarrow \infty$ ; here  $s = (s_1, \dots, s_d)$ ,  $x = (x_1, \dots, x_d) \in R^d$ ,  $(s \cdot x) = \sum_k s_k x_k$  is the inner product between  $s$  and  $x$ , and  $\|\cdot\|_p$  is the  $l_p$  norm, i.e.,  $\|s\|_p = (\sum_k |s_k|^p)^{1/p}$ , if  $1 \leq p < \infty$ , and  $\|s\|_\infty = \max_k |s_k|$ .

We define the flat-top kernel smoothed estimator of  $f(x)$ , for some  $x \in R^d$ , by

$$\hat{f}(x) = \frac{1}{N} \sum_{k=1}^N \Lambda_C(x - X_k) = \frac{1}{(2\pi)^d} \int \lambda_C(s) \phi_N(s) e^{-i(s \cdot x)} ds, \quad (2)$$

where  $\lambda_C(s) = \int \Lambda_C(x) e^{i(s \cdot x)} dx$ ,  $\Lambda_C(x) = h^{-d} \Omega_C(x/h)$ , for some chosen bandwidth  $h > 0$ , and  $\Omega_C$  satisfies the properties of Definition 1; also note that  $\phi_N(s)$  is the sample

characteristic function defined by

$$\phi_N(s) = \frac{1}{N} \sum_{k=1}^N e^{i(s \cdot X_k)}.$$

Now it is well known (cf. Rosenblatt (1991, p. 7)) that if  $f$  is continuous at  $x$ , and  $f(x) > 0$ , then

$$\text{Var}(\hat{f}(x)) = \frac{1}{h^d N} f(x) \int \Omega^2(x) dx + O(1/N). \quad (3)$$

Hence, the order of magnitude of the Mean Squared Error (MSE) of  $\hat{f}(x)$  hinges on the order of magnitude of its bias. To quantify the bias (and resulting MSE) of  $\hat{f}(x)$ , we formulate three different conditions based on the rate of decay of  $\phi$  that are in the same spirit as the conditions in Watson and Leadbetter (1963).

*Condition  $C_1$ : For some  $p \in [1, \infty]$ , there is an  $r > 0$ , such that  $\int \|s\|_p^r |\phi(s)| ds < \infty$*

*Condition  $C_2$ : For some  $p \in [1, \infty]$ , there are positive constants  $B$  and  $K$  such that  $|\phi(s)| \leq B e^{-K \|s\|_p}$  for all  $s \in R^d$ .*

*Condition  $C_3$ : For some  $p \in [1, \infty]$ , there is a positive constant  $B$  such that  $|\phi(s)| = 0$ , if  $\|s\|_p \geq B$ .*

Note that if one of Conditions  $C_1$  to  $C_3$  holds for some  $p \in [1, \infty]$ , then, by the equivalence of  $l_p$  norms for  $R^d$ , that same Condition would hold for *any*  $p \in [1, \infty]$ , perhaps with a change in the constants  $B$  and  $K$ .

Conditions  $C_1$  to  $C_3$  can be interpreted as different conditions on the smoothness of the density  $f(x)$  for  $x \in R^d$ ; cf. Katznelson (1968), Butzer and Nessel (1971), Stein and Weiss (1971), and the references therein. Note that they are given in increasing order of strength, i.e., if Condition  $C_2$  holds, then Condition  $C_1$  holds as well, and if Condition  $C_3$  holds, then Conditions  $C_1$  and  $C_2$  hold as well. Also note that if Condition  $C_1$  holds, then  $f$  must necessarily have  $[r]$  bounded, continuous derivatives over  $R^d$ , where  $[\cdot]$  is the positive part; cf. Katznelson (1968, p. 123). Obviously, if Condition  $C_2$  holds, then  $f$  has bounded, continuous derivatives of *any* order over  $R^d$ .

The following theorem quantifies the performance of the proposed family of flat-top estimators. It was first proved in Politis and Romano (1999) in the case where  $C$  is the  $l_p$  unit ball with  $1 \leq p \leq \infty$ ; we restate it below in this more general case.

**Theorem 1** [POLITIS AND ROMANO (1999)] *Assume that  $N \rightarrow \infty$ .*

*(a) Under Condition  $C_1$ , and letting  $h \sim AN^{-1/(2r+d)}$ , for some constant  $A > 0$ , it follows*

that

$$\sup_{x \in \mathbb{R}^d} \text{MSE}(\hat{f}(x)) = O(N^{-2r/(2r+d)}).$$

(b) Under Condition  $C_2$ , and letting  $h \sim A/\log N$ , where  $A$  is a constant such that  $A < 2K$ , it follows that

$$\sup_{x \in \mathbb{R}^d} \text{MSE}(\hat{f}(x)) = O\left(\frac{\log^d N}{N}\right).$$

(c) Under Condition  $C_3$ , and letting  $h$  be some constant small enough such that  $h \leq B^{-1}$ , it follows that

$$\sup_{x \in \mathbb{R}^d} \text{MSE}(\hat{f}(x)) = O(1/N).$$

**Remark 1a.** The special case where  $\omega_C(s) = 0$  for all  $s \notin C$ , has been considered by many authors in the literature, e.g. Parzen (1962), Davis (1977), and Ibragimov and Hasminskii (1982). Nevertheless, the choice  $\omega_C(s) = 0$  for all  $s \notin C$  is *not* recommendable in practice; see Politis and Romano (1999) for more details on such practical concerns, including choosing the bandwidth  $h$  in practice.

**Remark 1b.** A rather surprising observation is that smoothing with flat-top kernels does not seem to be plagued by the ‘curse of dimensionality’ in case the underlying density is ultra-smooth, possessing derivatives of all orders, i.e., under Condition  $C_2$  (or  $C_3$ ). For example, in Theorem 1b under Condition  $C_2$ , the MSE of estimation achieved by flat-top kernel smoothing is of order  $O(\frac{\log^d N}{N})$ , i.e., depending on the dimension  $d$  only through the slowly varying function  $\log^d N$ . A more extreme result obtains under Condition  $C_3$ : Theorem 1c shows that in that case the MSE of estimation becomes exactly  $O(1/N)$  which is identical to the parametric rate of estimation, and does not depend on the dimension  $d$  at all.

**Remark 1c.** It is also noteworthy that, even in the univariate case  $d = 1$ , the MSE of estimation is identical to the  $O(1/N)$  parametric rate of estimation under Condition  $C_3$ , and is very close to  $O(1/N)$  under Condition  $C_2$ . In other words, if a practitioner is to decide between fitting a particular parametric model to the data vs. assuming that the unknown density has derivatives of all orders (i.e., Condition  $C_2$ ) and using our proposed flat-top kernel smoothing, there is no real benefit (in terms of rate of convergence) in favor of the parametric model. As a matter of fact, the smoothness Condition  $C_2$  may be

viewed as defining a huge class of functions that includes all the usual parametric models; the proposed flat-top kernel smoothing can then proceed to estimate the unknown function with accuracy comparable to the accuracy of a parametric estimator.

**Remark 1d.** It is well-known in the literature (see, for example, Müller (1988) or Scott (1992)) that kernel density estimators corresponding to kernels of order bigger than two are not necessarily nonnegative functions; it goes without saying that the same applies for our estimators  $\hat{f}$  that are obtained using kernels of infinite order. Nevertheless, the nonnegativity is not a serious issue as there is a natural fix-up, namely using the modified estimator  $\hat{f}^+(x) = \max(\hat{f}(x), 0)$ ; see also Gajek (1986) and Hall and Murison (1992). Note that the estimator  $\hat{f}^+(x)$  is not only nonnegative, but is more accurate as well, in the sense that  $MSE(\hat{f}^+(x)) \leq MSE(\hat{f}(x))$ , for all  $x$ ; this fact follows from the obvious inequality  $|\hat{f}^+(x) - f(x)| \leq |\hat{f}(x) - f(x)|$ . In addition, if  $f(x) > 0$ , an application of Chebychev's inequality shows that  $Prob\{\hat{f}(x) = \hat{f}^+(x)\} \rightarrow 1$  under the assumptions of our Theorem 1; on the other hand, if  $f(x) = 0$ , then the large-sample distribution of either  $\sqrt{h^d N} \hat{f}^+(x)$ , or  $\sqrt{h^d N} \hat{f}(x)$ , degenerates to a point mass at zero.

**Remark 1e.** By the formal analogy between probability spectral density estimation (see e.g. Rosenblatt (1991)) it should not be surprising that flat-top kernels might be applicable in a context of nonparametric spectral density estimation. In Politis and Romano (1995, 1996), kernels belonging to a subset of the family of flat-top kernels are employed for the purpose of spectral density estimation using data consisting of a realization of a stationary time series or a homogeneous random field. Incidentally, note a typo in the statement of Theorem 2 in Politis and Romano (1996): instead of  $M_i \sim dc_i \log N_i$  it should read  $m_i \sim dc_i \log N_i$ .

### 3 Further issues on density estimation

In this section we will continue the discussion on probability density estimation based on i.i.d. data  $X_1, \dots, X_N$ , and will investigate to what extent the performance of  $\hat{f}(x)$  is affected by a discontinuity of  $f$  (or its derivatives) at points away from  $x$ .

#### 3.1 Case of smooth density over a finite domain

To fix ideas, consider first the univariate case  $d = 1$ ; it is well-known that, if the random variables  $X_1, \dots, X_N$  are bounded, that is, if the density  $f$  has domain the finite interval

$[a, b]$  as opposed to  $R$ , then the characteristic function  $\phi$  will not satisfy the smoothness Conditions  $C_1$ ,  $C_2$ , or  $C_3$ . The situation is exemplified by the smoothest of such densities, namely the uniform density over the interval  $[-\theta, \theta]$  whose characteristic function is given by  $\phi(s) = \frac{\sin \theta s}{\theta s}$ ; cf. Rao (1973, p. 151).

In general, suppose  $\bar{f}$  is a very smooth density function (e.g., satisfying one of the smoothness Conditions  $C_1$ ,  $C_2$ , or  $C_3$ ), and let  $f(x) = c \bar{f}(x) 1_{[-\theta, \theta]}(x)$ , where  $1_{[-\theta, \theta]}(x)$  is the indicator function, and  $c = 1 / \int_{-\theta}^{\theta} \bar{f}(x) dx$ . Then the characteristic function of  $f$  is given by  $\phi(s) = 2\theta c \bar{\phi}(s) * \frac{\sin \theta s}{\theta s}$ , where  $\bar{\phi}$  is the characteristic function of  $\bar{f}$ , and  $*$  denotes convolution. In other words, the term  $\frac{\sin \theta s}{\theta s}$  seems unavoidable, and is due to the truncation of the random variables.

Nevertheless, it seems intuitive that for the ultra-smooth uniform density over the interval  $[-\theta, \theta]$ , smoothing should give good results; this is indeed true as the following discussion shows. First note that if  $\theta = \pi$ , then  $\frac{\sin \theta s}{\theta s} = 0$  for all  $s \in Z - \{0\}$ ; this observation naturally brings us to Fourier series on the circle defined by ‘wrapping’ the interval  $[-\pi, \pi]$  around on a circle, or—in high dimensions—Fourier series on the  $d$ -dimensional torus.

So, without loss of generality (and possibly having to use a linear/affine transformation in pre-processing the data), assume that  $X_1, \dots, X_N$  are i.i.d. with probability density  $f$  defined on the torus  $T = [-\pi, \pi]^d$ . Now let the characteristic function  $\phi(s) = \int_T e^{i(s \cdot x)} f(x) dx$ , and the sample characteristic function  $\phi_N(s) = \frac{1}{N} \sum_{k=1}^N e^{i(s \cdot X_k)}$ . Recall the Fourier series formula

$$f(x) = (2\pi)^{-d} \sum_{s \in Z^d} e^{-i(s \cdot x)} \phi(s),$$

and define our estimator

$$\hat{f}(x) = (2\pi)^{-d} \sum_{s \in Z^d} \lambda_C(s) e^{-i(s \cdot x)} \phi_N(s),$$

where  $\lambda_C(s)$  was defined in Section 2. As before, we define some smoothness conditions based on the characteristic function  $\phi$ .

*Condition  $K_1$ :* For some  $p \in [1, \infty]$ , there is an  $r > 0$ , such that  $\sum_{s \in Z^d} \|s\|_p^r |\phi(s)| < \infty$

*Condition  $K_2$ :* For some  $p \in [1, \infty]$ , there are positive constants  $B$  and  $K$  such that  $|\phi(s)| \leq B e^{-K \|s\|_p}$  for all  $s \in Z^d$ .

*Condition  $K_3$ :* For some  $p \in [1, \infty]$ , there is a positive constant  $B$  such that  $|\phi(s)| = 0$ , if  $\|s\|_p \geq B$  (with  $s \in Z^d$ ).

The following theorem quantifies the performance of the general family of flat-top estimators; its proof follows closely the proof in Politis and Romano (1999) and is omitted.

**Theorem 2** *Assume that  $N \rightarrow \infty$ .*

(a) *Under Condition  $K_1$ , and letting  $h \sim AN^{-1/(2r+d)}$ , for some constant  $A > 0$ , it follows that*

$$\sup_{x \in T} \text{MSE}(\hat{f}(x)) = O(N^{-2r/(2r+d)}).$$

(b) *Under Condition  $K_2$ , and letting  $h \sim A/\log N$ , where  $A$  is a constant such that  $A < 2K$ , it follows that*

$$\sup_{x \in T} \text{MSE}(\hat{f}(x)) = O\left(\frac{\log^d N}{N}\right).$$

(c) *Under Condition  $K_3$ , and letting  $h$  be some constant small enough such that  $h \leq B^{-1}$ , it follows that*

$$\sup_{x \in T} \text{MSE}(\hat{f}(x)) = O(1/N).$$

**Remark 2.** It is easy to see that the uniform density on  $T$  satisfies condition  $K_3$ , and thus smoothing with a flat-top kernel achieves the parametric  $\sqrt{N}$ -rate in this case (with no dependence on the dimensionality  $d$ ) which is remarkable. Nevertheless, Conditions  $K_1$ ,  $K_2$ ,  $K_3$  are quite stringent as they imply smoothness/differentiability of  $f$  over the whole torus  $T$ ; this is equivalent to assuming that a periodic extension of  $f$  over  $R^d$  is smooth/differentiable over the whole of  $R^d$ .

To fix ideas, we again return to the case  $d = 1$ , and note that Condition  $K_1$  implies that  $f$  has  $[r]$  bounded, continuous derivatives over  $T$ , where  $[\cdot]$  is the positive part; this implies, in particular, that  $f(-\pi) = f(\pi)$ ,  $f'(-\pi) = f'(\pi)$ ,  $f''(-\pi) = f''(\pi)$ , and so forth up to the  $[r]$ -th derivative. If  $f$  is smooth/differentiable over  $(-\pi, \pi)^d$  but not over the whole torus  $T$  then the Fourier series method is not appropriate; rather, a technique of extension of  $f$  over the whole of  $R^d$  might be useful as elaborated upon in the next subsection.

### 3.2 Case of infinite domain with some discontinuities

We now return to the set-up of Section 2 where  $X_1, \dots, X_N$  are i.i.d. random vectors taking values in  $R^d$  possessing a probability density function  $f$ . The objective is to estimate  $f(x)$  for some  $x$  in the interior of  $I$ , where  $I$  is a compact rectangle in  $R^d$  over

which  $f$  possesses a certain degree of smoothness. Again without loss of generality (and possibly having to use a linear/affine transformation in pre-processing the data), assume that  $I = [-a, a] \times [-a, a] \times \cdots [-a, a]$  for some  $a > 0$ . Outside the rectangle  $I$ ,  $f$  and its derivatives might have discontinuities, and  $f$  might even be zero (bringing us to the set-up of bounded random variables as in the previous subsection). We now define the following condition which is related to our previous Condition  $C_1$ .

*Condition  $C^*[r]$ : For some positive integer  $r$ ,  $f$  has  $r$  bounded, continuous derivatives over the closed region  $I$ .*

Again define the flat-top kernel smoothed estimator of  $f(x)$  by

$$\hat{f}(x) = \frac{1}{N} \sum_{k=1}^N \Lambda_C(x - X_k) \quad (4)$$

where  $\Lambda_C(s)$  is defined in Section 2.

It is intuitive that, if the tails of  $\Omega_C(x)$  were negligible, the influence on  $\hat{f}(x)$  of some  $X_k$  observations that are far away from  $x$  (and may even correspond to a region where  $f$  is not smooth) would be insignificant; this is indeed a true observation, and leads to the following result. To state it, we define the smaller rectangle  $J = [-b, b] \times [-b, b] \times \cdots \times [-b, b]$  where  $0 < b < a$ ;  $b$  should be thought to be close to  $a$  such that the point  $x$  of interest will also belong to  $J$ .

**Theorem 3** *Assume that*

$$\Omega_C(x) = O((1 + \max_i |x_i|)^{-q}), \quad (5)$$

*for some real number  $q > d$ . Let  $p = \text{Prob}\{X_1 \in I\} = \int_I f(x) dx$  be strictly in  $(0, 1)$ . Under Condition  $C^*[r + 1]$ , and letting  $h \sim AN^{-1/(2r^*+d)}$ , for some constant  $A > 0$ , it follows that*

$$\sup_{x \in J} \text{MSE}(\hat{f}(x)) = O(N^{-2r^*/(2r^*+d)})$$

*as  $N \rightarrow \infty$ , where  $r^* = \min(r, q - d)$ .*

**Proof.** The order of  $\text{MSE}(\hat{f}(x))$  again depends on the bias of  $\hat{f}(x)$  since  $\text{Var}(\hat{f}(x))$  is of order  $1/(h^d N)$  as before. To estimate the bias, consider the following argument.

Let  $\bar{f}$  be a probability density that has (at least)  $r + 1$  bounded derivatives over the whole of  $R^d$ , and such that  $\bar{f}(x) = f(x)$  for all  $x \in I$ ; this extension of  $f$  over the whole

of  $R^d$  can be done in many ways—see e.g. Stein (1970). Re-order the  $X_i$ s in such a way that  $X_1, \dots, X_K$  are in  $I$ , whereas  $X_{K+1}, \dots, X_N$  are in  $I^c$ , i.e., the complement of  $I$ .

Construct a new sample  $Y_1, \dots, Y_N$  with the property that  $Y_i = X_i$  for  $i = 1, \dots, K$ , and such that  $Y_{K+1}, \dots, Y_N$  are drawn i.i.d. from density  $\bar{f}(x)1_{I^c}(x)/(1-p)$ , where  $p = \text{Prob}\{X_1 \in I\} = \int_I f(x)dx$ . It is apparent now that the sample  $Y_1, \dots, Y_N$  can be considered as a *bona fide* i.i.d. sample from density  $\bar{f}(x)$  for  $x \in R^d$ .

Now recall that (together with the re-ordering) we have

$$\hat{f}(x) = \frac{1}{N} \sum_{k=1}^N \Lambda_C(x - X_k) = \frac{1}{N} \sum_{k=1}^K \Lambda_C(x - X_k) + \left(\frac{N-K}{N}\right)u_1.$$

Note that, due to assumption (5), and to the fact that an observation outside  $I$  will be at a distance of at least  $a-b$  (in an  $l_\infty$  sense) from the point  $x \in J$ , we have  $u_1 = O\left(\frac{h^{q-d}}{(a-b)^q}\right)$  almost surely.

Finally, observe that

$$\frac{1}{N} \sum_{k=1}^K \Lambda_C(x - X_k) = \frac{1}{N} \sum_{k=1}^K \Lambda_C(x - Y_k) = \frac{1}{N} \sum_{k=1}^N \Lambda_C(x - Y_k) + \left(\frac{N-K}{N}\right)u_2,$$

where again  $u_2 = O\left(\frac{h^{q-d}}{(a-b)^q}\right)$  almost surely.

To summarize:

$$\hat{f}(x) = \frac{1}{N} \sum_{k=1}^N \Lambda_C(x - Y_k) + (u_1 + u_2)\left(\frac{N-K}{N}\right). \quad (6)$$

Taking expectations in equation (6), and recalling that  $E\left(\frac{N-K}{N}\right) = 1-p$ , and

$$E\left(\frac{1}{N} \sum_{k=1}^N \Lambda_C(x - Y_k)\right) = \bar{f}(x) + o(h^r),$$

and that  $\bar{f}(x) = f(x)$  for all  $x \in I$  (and thus for all  $x \in J$  as well), the theorem is proven.  $\square$

**Remark 3a.** Theorem 3 shows that, in the possible presence of discontinuities in  $f$  or its derivatives at points away from the target region  $I$ , it is important to use a flat-top kernel  $\Omega_C(x)$  that is chosen to have small tails. For example, the case where  $\omega_C(s) = 0$  for all  $s \notin C$  that was considered by Parzen (1962), Davis (1977), and Ibragimov and Hasminskii (1982) satisfies equation (5) with  $q = 1$  and is *not* recommendable. By contrast, the simple kernel  $\Lambda_c^{PROD}(x)$  of Politis and Romano (1999) satisfies (5) with  $q = 2$

as long as  $c > 1$ .

**Remark 3b.** It is easy to construct flat-top kernels satisfying equation (5) with  $q > 2$ ; all it takes is to make sure that  $\omega_C(s)$  has a high degree of smoothness (e.g., high number of derivatives) for all  $s$ . To do this, one must pay special attention at the boundary of the region  $C$  since, inside  $C$ ,  $\omega_C(s)$  is infinitely differentiable with all derivatives being zero. We now give an explicit construction of a flat-top kernel satisfying (5) with an arbitrary exponent  $q$  in the case where  $C$  is the  $l_\infty$  unit ball. Let  $\omega_1(s_1)$  be a member of the flat-top family in the case  $d = 1$ , and having the following properties:  $\omega_1(s_1) = \omega_1(-s_1)$  for all  $s_1 \in R$ ;  $\omega_1(s_1) = 1$  for all  $|s_1| \leq 1$ ;  $\omega_1(s_1) = 0$  for all  $|s_1| \geq c$  for some  $c > 1$  ( $c = 2$  is a useful practical choice);  $\omega_1(s_1)$  is monotone decreasing for  $1 < s_1 < c$ ; and  $\omega_1(s_1)$  possesses  $n$  continuous derivatives for all  $s_1 \in R$ . Then, the  $d$ -dimensional flat-top kernel  $\Omega_C(x)$  with Fourier transform equal to  $\omega_C(s) = \prod_{i=1}^d \omega_1(s_i)$  satisfies (5) with exponent  $q > n + 1$ .

As a last remark, note that Condition  $C^*[r + 1]$  in Theorem 3 can be relaxed to  $C^*[r]$  if  $\Omega_C(x)$  is chosen to have  $r$  finite moments which in turn can be guaranteed by having  $q > r + 1$  in equation (5). The following corollary should be compared to Theorem 1a.

**Corollary 1** *Assume Condition  $C^*[r]$ , as well as equation (5) with some  $q > r + 1$  and  $q \geq r + d$ . Let  $p = \text{Prob}\{X_1 \in I\} = \int_I f(x)dx$  be strictly in  $(0, 1)$ . Letting  $h \sim AN^{-1/(2r+d)}$ , for some constant  $A > 0$ , it follows that*

$$\sup_{x \in J} \text{MSE}(\hat{f}(x)) = O(N^{-2r/(2r+d)})$$

as  $N \rightarrow \infty$ .

**Acknowledgement.** This research was partially supported by NSF grant DMS 97-03964. Many thanks are due to Prof. George Kyriazis of the University of Cyprus for many helpful discussions, and to Prof. E. Masry of the University of California at San Diego for pointing out the typo in Politis and Romano (1996).

## References

- [1] Bartlett, M.S. (1963), Statistical Estimation of Density Functions, *Sankhya, Ser. A*, 25, 245-54.
- [2] Butzer, P. and Nessel, R. (1971), *Fourier analysis and approximation*, Academic Press, New York.
- [3] Cacoullos, T. (1966), Estimation of a multivariate density, *Annals Inst. Statist. Math.*, vol. 18, pp. 178-189.
- [4] Davis, K.B. (1977), Mean integrated square error properties of density estimates, *Ann. Statist.*, vol. 5, pp. 530-535.
- [5] Devroye, L. (1987), *A course in density estimation*, Birkhäuser, Boston.
- [6] Devroye, L. (1992), A note on the usefulness of superkernels in density estimation, *Ann. Statist.*, vol. 20, no. 4, pp. 2037-2056.
- [7] Gajek, L. (1986), On improving density estimators which are not bona fide functions, *Ann. Statist.*, vol. 14, 1612-1618.
- [8] Gasser, T., Müller, H.G. and Mammitzsch, V. (1985), Kernels for nonparametric curve estimation, *J. Roy. Statist. Soc. B*, vol. 47, pp. 238-252.
- [9] Granovsky, B.L. and Müller, H.G. (1991), Optimal kernel methods: A unifying variational principle, *Internat. Statist. Review*, vol. 59, no. 3, pp. 373-388.
- [10] Györfi, L., Härdle, W., Sarda, P., and Vieu, P. (1989), *Nonparametric Curve Estimation from Time Series*, Lecture Notes in Statistics No.60, Springer-Verlag.
- [11] Hall, P. and Murison, R.D. (1992), Correcting the negativity of high-order kernel density estimators, *J. Multivar. Anal.*, vol. 47, 103-122.
- [12] Ibragimov, I.A. and Hasminskii, R.Z. (1982), Estimation of distribution density belonging to a class of entire functions, *Theor. Probab. Appl.*, vol. 27, pp. 551-562.
- [13] Jones, M.C. (1995), On higher order kernels, *J. Nonparametr. Statist.*, vol. 5 , 215-221.
- [14] Jones, M.C. and Foster, P.J. (1993), Generalized jackknifing and higher order kernels, *J. Nonparametr. Statist.*, vol. 3, 81-94

- [15] Katznelson, Y. (1968), *An Introduction to Harmonic Analysis*, Dover, New York.
- [16] Marron, J.S. (1994), Visual understanding of higher order kernels, *J. Comput. Graphical Statist.*, vol.3, 447-458.
- [17] Marron, J.S. and Wand, M.P. (1992), Exact mean integrated squared error, *Ann. Statist.*, vol. 20, 712-736.
- [18] Müller, H.G. (1988), *Nonparametric regression analysis of longitudinal data*, Springer-Verlag, Berlin.
- [19] Nadaraya, E.A. (1989), *Nonparametric Estimation of Probability Densities and Regression Curves*, Kluwer Academic Publishers, Dordrecht.
- [20] Parzen, E. (1962), On Estimation of a Probability Density Function and its Mode, *Ann. Math. Statist.*, vol. 33, 1065-1076.
- [21] Politis, D.N. and Romano, J.P. (1993), On a Family of Smoothing Kernels of Infinite Order, in *Computing Science and Statistics, Proceedings of the 25th Symposium on the Interface*, San Diego, California, April 14-17, 1993, (M. Tarter and M. Lock, eds.), The Interface Foundation of North America, pp. 141-145.
- [22] Politis, D.N. and Romano, J.P. (1995), Bias-Corrected Nonparametric Spectral Estimation, *J. Time Ser. Anal.*, vol. 16, No. 1, 67-104.
- [23] Politis, D.N. and Romano, J.P. (1996), On flat-top kernel spectral density estimators for homogeneous random fields, *J. Statist. Plan. Infer.*, vol. 51, 41-53.
- [24] Politis, D.N. and Romano, J.P. (1999), Multivariate density estimation with general flat-top kernels of infinite order, *J. Multivar. Anal.*, vol. 68, 1-25.
- [25] Priestley, M.B. (1981), *Spectral Analysis and Time Series*, Academic Press.
- [26] Rao, C. R.(1973), *Linear Statistical Inference and its Applications, 2nd Ed.*, John Wiley, New York.
- [27] Rosenblatt, M. (1991), *Stochastic Curve Estimation*, NSF-CBMS Regional Conference Series vol. 3, Institute of Mathematical Statistics, Hayward.
- [28] Scott, D. W. (1992), *Multivariate density estimation: theory, practice, and visualization*, Wiley, New York.

- [29] Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.
- [30] Stein, E.M. (1970), *Singular integrals and differentiability properties of functions*, Princeton Univ. Press, Princeton, New Jersey.
- [31] Stein, E.M., and Weiss, W. (1971), *Introduction to Fourier analysis on Euclidean spaces*, Princeton Univ. Press, Princeton, New Jersey.
- [32] Wand, M.P. and Jones, M.C. (1993), Comparison of smoothing parameterizations in bivariate kernel density estimation, *J. Amer. Statist. Assoc.*, vol. 88, 520-528.
- [33] Watson, G.S. and Leadbetter, M.R. (1963), On the estimation of the probability density I, *Ann. Math. Statist.*, vol. 33, pp. 480-491.