Upper Bounds for the I-MSE and max-MSE of Kernel Density Estimators

Nils Lid Hjort[1] and Nikolai G. Ushakov[2,*]

December 1999

[1]Department of Mathematics, University of Oslo, Norway
[2]Russian Academy of Sciences, Chernogolovka, Russia

Abstract. The performance of kernel density estimators is usually studied via Taylor expansions and asymptotic approximation arguments, in which the bandwidth parameter tends to zero with increasing sample size. In contrast, this paper focusses directly on the finite-sample situation. Informative upper bounds are derived both for the integrated and the maximal mean squared error function. Results are reached for the traditional case, where the kernel is a probability density function, under various sets of assumptions on the underlying density to be estimated. Results are also derived for the important non-conventional case of the sinc kernel, which is not integrable and also takes negative values. We pin-point ways in which the sinc-based estimator performs better than the conventional kernel estimators. When proving our results we rely on methods related to characteristic and empirical characteristic functions.

Key words: *characteristic functions, density estimation, finite-sample performance, max-MSE, sinc kernel, upper bounds*

1. Introduction

In this article we derive some rigorous upper bounds for the estimation error of kernel density estimators for finite values of the sample size n, in terms of choices of the kernel function K and the bandwidth $h = h_n$. These bounds are by construction non-asymptotic, and are useful when one needs to secure a certain precision of an estimate for a given (finite) value of n, for broad classes of densities. We study both smooth cases (where the density to be estimated is one or more times differentiable) and non-smooth cases (the underlying density function is not supposed to be differentiable or even continuous). The machinery of characteristic and empirical characteristic functions is used, and relevant general results are established in Section 2.

In Section 3 conventional kernel estimators will be considered, i.e. estimators whose kernels are probability density functions. These estimators always produce estimates which are densities. We term a kernel density estimator non-conventional if its kernel function is not a probability density, i.e. it may take negative values

* Partially supported by RFBR Grants 97-01-00273, 98-01-00621 and 98-01-00926, and by INTAS-RFBR Grant IR-97-0537

or/and does not integrate to one (or even is not integrable). Such non-conventional estimators are studied in Section 4, with particular attention to the sinc kernel; see also Glad, Hjort and Ushakov (1999a). Such estimators, based on higher order kernels, superkernels or the sinc kernel, often provide better estimation precision, but have an essential disadvantage: they produce estimates which are not probability density functions, i.e. may take negative values or/and do not integrate to one. However, this defect can be corrected afterwards without loss of their performance properties (see Glad, Hjort and Ushakov, 1999b).

A discussion of our results, with a view towards their use in density estimation problems, is given in the final Section 5. Topics there include new strategies for bandwidth selection.

2. Auxiliary results, via characteristic functions

In this paper we use the characteristic function approach to studying performance of density estimators, rather than the traditional Taylor expansions and asymptotic approximations. Therefore we first express some basic concepts of kernel density estimators in terms of characteristic functions.

Let $X_1, \dots, X_n$ be independent and identically distributed random variables with absolutely continuous distribution function $F(x)$, density function $p(x)$, and characteristic function $f(t)$. The kernel density estimator associated with the sample $X_1, \dots, X_n$ is defined as

$$p_n(x) = p_{n,h}(x) = \frac{1}{n}\sum_{j=1}^{n} K_h(x - X_j), \tag{2.1}$$

where $K(x)$ is the kernel function with scaled version $K_h(x) = h^{-1}K(h^{-1}x)$ and $h = h_n$ is a positive number (depending on n) called the bandwidth or the smoothing parameter. We do not necessarily demand that K is integrable (sometimes the best estimators correspond to nonintegrable kernels). However, we suppose that K is square integrable, and in addition that it is integrable in the sense of the Cauchy principal value with v.p. $\int_{-\infty}^{\infty} K(x)\,\mathrm{d}x = 1$, in which

$$\text{v.p.}\int_{-\infty}^{\infty} = \lim_{T\to\infty}\lim_{\epsilon\to 0}\Big[\int_{-T}^{-\epsilon} + \int_{\epsilon}^{T}\Big].$$

Under these assumptions the Fourier transform of K can be defined as

$$\varphi(t) = \text{v.p.}\int_{-\infty}^{\infty} e^{itx}K(x)\,\mathrm{d}x$$

(see Chapter 4 of Titchmarsh, 1937). In the following we will omit integration limits when the integral is to be taken over the full real line.

Let $\hat{p}_n$ be an estimator (not necessarily a kernel estimator) of p associated with the sample $X_1, \ldots, X_n$. The bias, the mean squared error (MSE) and the mean integrated squared error (MISE) of $\hat{p}_n$ are defined, respectively, as

$$B_n(\hat{p}_n(x)) = \mathrm{E}\hat{p}_n(x) - p(x),$$

$$\mathrm{MSE}(\hat{p}_n(x)) = \mathrm{E}\{\hat{p}_n(x) - p(x)\}^2,$$

and

$$\mathrm{MISE}(\hat{p}_n) = \int \mathrm{MSE}(\hat{p}_n(x))\,\mathrm{d}x = \mathrm{E}\int \{\hat{p}_n(x) - p(x)\}^2\,\mathrm{d}x. \tag{2.2}$$

In case of the kernel estimator p_n, defined by (2.1), the bias may be expressed via the convolution as

$$B_n(p_n(x)) = (K_h \star p)(x) - p(x) = \int K_h(x-y)p(y)\,\mathrm{d}y - p(x).$$

Since convolution is a kind of smoothing, the bias of the kernel estimator is the difference between a smoothed density and the density itself. The mean squared error admits a well-known decomposition into variance and squared bias, with consequent MISE representation

$$\mathrm{MISE}(\hat{p}_n) = \int B_n^2(\hat{p}_n(x))\,\mathrm{d}x + \int \mathrm{Var}(\hat{p}_n(x))\,\mathrm{d}x.$$

Note that together with MSE and MISE other measures of deviation may be used. Among them, the mean absolute error $\mathrm{E}|\hat{p}_n(x) - p(x)|$ and its integral are especially important (see Devroye and Györfi, 1985). In the present article we restrict attention to MSE and MISE, however.

For a real valued function g we will use the following notation, provided the integrals exist:

$$\mu_k(g) = \int |x|^k g(x)\,\mathrm{d}x \quad \text{and} \quad R(g) = \int g^2(x)\,\mathrm{d}x.$$

If the kernel K is a probability density function, and the density to be estimated is twice differentiable and with square integrable second order derivative, then it is well known that the best order of estimation accuracy in terms of MISE is $O(n^{-4/5})$; see also Section 5.1. However, if we permit the kernel not to be a density, then the order can be improved. For example, if p is the normal density and K is the sinc kernel, i.e. $K(x) = \sin x/(\pi x)$, then

$$\min_{h>0} \mathrm{MISE}(p_n) = O\Big(\frac{\sqrt{\log n}}{n}\Big) \quad \text{as } n \to \infty\,;$$

see Section 4 below and Glad, Hjort and Ushakov (1999a).

We now express basic characteristics of density estimators in terms of Fourier transforms and establish some auxiliary results.

Let $\hat{f}_n$ denote the Fourier transform of an estimator $\hat{p}_n$. Making use of the inversion formula for densities and the Parseval–Plancherel identity we easily obtain the following formulae:

$$B_n(\hat{p}_n(x)) = \frac{1}{2\pi}\int e^{-itx}\{\mathrm{E}\hat{f}_n(t) - f(t)\}\,\mathrm{d}t, \tag{2.3}$$

$$\begin{aligned}\mathrm{MSE}(\hat{p}_n(x)) &= \mathrm{E}\Big\{\frac{1}{2\pi}\int e^{-itx}[\hat{f}_n(t) - f(t)]\,\mathrm{d}t\Big\}^2 \\ &= \frac{1}{(2\pi)^2}\int\int e^{-i(u+v)x}\mathrm{E}\{(\hat{f}_n(u) - f(u))(\hat{f}_n(v) - f(v))\}\,\mathrm{d}u\,\mathrm{d}v,\end{aligned} \tag{2.4}$$

and

$$\mathrm{MISE}(\hat{p}_n) = \frac{1}{2\pi}\int \mathrm{E}|\hat{f}_n(t) - f(t)|^2\,\mathrm{d}t. \tag{2.5}$$

In the remainder of this section, we will consider only kernel estimators and suppose that the kernel K is a probability density function, i.e. it is nonnegative and integrates to one. Study the empirical characteristic function associated with sample $X_1, \ldots, X_n$,

$$f_n(t) = \frac{1}{n}\sum_{j=1}^{n} e^{itX_j}.$$

The characteristic function of the estimator $p_{n,h}(x)$ is $f_n(t)\varphi(ht)$, where $\varphi(t) = \int e^{itu}K(u)\,\mathrm{d}u$ is the characteristic function of the kernel. And the kernel estimator (2.1) can be expressed in terms of f_n as

$$p_{n,h}(x) = \frac{1}{2\pi}\int e^{-itx} f_n(t)\varphi(ht)\,\mathrm{d}t.$$

Now, taking into account that

$$\mathrm{E}f_n(u)f_n(v) = (1 - 1/n)f(u)f(v) + (1/n)f(u+v)$$

and

$$\mathrm{E}|f_n(t)|^2 = (1 - 1/n)|f(t)|^2 + 1/n,$$

we can write (2.3)–(2.5) in the form

$$B_n(p_n(x)) = \frac{1}{2\pi}\int e^{-itx} f(t)\{\varphi(ht) - 1\}\,\mathrm{d}t, \tag{2.6}$$

$$\begin{aligned}\mathrm{MSE}(p_n(x)) =& \frac{1}{(2\pi)^2}\int\int e^{-i(u+v)x}\Big[\frac{1}{n}\varphi(hu)\varphi(hv)f(u+v) \\ &+ \Big\{\Big(1 - \frac{1}{n}\Big)\varphi(hu)\varphi(hv) - 2\varphi(hu) + 1\Big\}f(u)f(v)\Big]\,\mathrm{d}u\,\mathrm{d}v\end{aligned} \tag{2.7}$$

and

$$\mathrm{MISE}(p_n) = \frac{1}{2\pi}\Big[\int |f(t)|^2|1-\varphi(ht)|^2\,\mathrm{d}t + \frac{1}{n}\int |\varphi(ht)|^2\{1-|f(t)|^2\}\,\mathrm{d}t\Big]. \quad (2.8)$$

From (2.6) we immediately obtain

$$|B_n(p_n(x))| \le \frac{1}{2\pi}\int |f(t)|\,|1-\varphi(ht)|\,\mathrm{d}t. \quad (2.9)$$

Lemma 1. *For each x,*

$$\mathrm{MSE}(p_n(x)) \le \Big\{\frac{1}{2\pi}\int |f(t)|\,|1-\varphi(ht)|\,\mathrm{d}t\Big\}^2 + \frac{a(x)}{\pi n h}\int |\varphi(t)|\,\mathrm{d}t, \quad (2.10)$$

where $a(x) = (K_h \star p)(x)$. If p is bounded by a, then

$$\sup_x \mathrm{MSE}(p_n(x)) \le \Big\{\frac{1}{2\pi}\int |f(t)|\,|1-\varphi(ht)|\,\mathrm{d}t\Big\}^2 + \frac{a}{\pi n h}\int |\varphi(t)|\,\mathrm{d}t.$$

Proof. It suffices to prove the first statement, since $a(x) = \int p(x-y)K_h(y)\,\mathrm{d}y$ $\le a$ for all x. Making use of relation (2.7), we obtain

$$\begin{aligned}\mathrm{MSE}(p_n(x)) =& \Big[\frac{1}{2\pi}\int e^{-itx}f(t)\{1-\varphi(ht)\}\,\mathrm{d}t\Big]^2\\ &+\frac{1}{n}\frac{1}{(2\pi)^2}\int\int e^{-i(u+v)x}\varphi(hu)\varphi(hv)f(u+v)\,\mathrm{d}u\,\mathrm{d}v\\ &-\frac{1}{n}\frac{1}{(2\pi)^2}\int\int e^{-i(u+v)x}\varphi(hu)\varphi(hv)f(u)f(v)\,\mathrm{d}u\,\mathrm{d}v.\end{aligned}$$

The first term on the right hand side is dominated by the first term of the right hand side of (2.10). Let us then estimate the absolute value of the second (denoted by T_2) and third (denoted by T_3) terms. We have

$$T_2 = \frac{1}{n}\frac{1}{2\pi}\int \varphi(hu)\Big\{\frac{1}{2\pi}\int e^{-i(u+v)x}\varphi(hv)f(u+v)\,\mathrm{d}v\Big\}\,\mathrm{d}u.$$

The term in brackets, being transformed to the form

$$\frac{1}{2\pi}\int e^{-itx}\varphi(h(t-u))f(t)\,\mathrm{d}t,$$

is equal to $\int p(x-y)K_h(y)e^{-iuy}\,\mathrm{d}y$ (since $\varphi(h(t-u))f(t)$ is the Fourier transform of the convolution of functions $p(x)$ and $K_h(x)e^{-iux}$), and clearly

$$\Big|\int p(x-y)K_h(y)e^{-iuy}\,\mathrm{d}y\Big| \le a(x).$$

Hence

$$|T_2| \le \frac{a(x)}{n} \frac{1}{2\pi} \int |\varphi(ht)|\,\mathrm{d}t.$$

Furthermore,

$$\begin{aligned}
|T_3| =& \frac{1}{n}\Big|\frac{1}{2\pi}\int e^{-iux}\varphi(hu)f(u)\,\mathrm{d}u\,\frac{1}{2\pi}\int e^{-ivx}f(v)\varphi(hv)\,\mathrm{d}v\Big| \\
&\le \frac{1}{n}(K_h \star p)(x)\frac{1}{2\pi}\int |f(v)|\,|\varphi(hv)|\,\mathrm{d}v \le \frac{a(x)}{n\,2\pi}\int |\varphi(hv)|\,\mathrm{d}v.
\end{aligned}$$

Thus we finally obtain (2.10). ∎

Lemma 2.

$$\mathrm{MISE}(p_{n,h}) \le \frac{1}{2\pi}\Big\{\int |f(t)|^2\,|1-\varphi(ht)|^2\,\mathrm{d}t + \frac{1}{nh}\int |\varphi(t)|^2\,\mathrm{d}t\Big\}.$$

This lemma immediately follows from relation (2.8).

We conclude this section with some inequalities for characteristic functions and which will be used below.

Lemma 3. *Let F be a distribution function with characteristic function f. If the first order absolute moment $\beta_1 = \int |x|\,\mathrm{d}F(x)$ is finite, then*

$$|1 - f(t)| \le \beta_1 |t| \quad \textit{for all real } t.$$

If F has null expectation and finite variance σ^2, then

$$|1 - f(t)| \le \tfrac{1}{2}\sigma^2 t^2 \quad \textit{for all real } t.$$

Proof. Observe that for any positive integer n and any $x > 0$,

$$\Big|e^{ix} - 1 - \frac{ix}{1!} - \ldots - \frac{(ix)^{n-1}}{(n-1)!}\Big| \le \frac{x^n}{n!} \tag{2.11}$$

(see for example Feller, 1971, Chapter 15). The first inequality of the lemma follows quickly via

$$|1 - f(t)| \le \int |1 - e^{itx}|\,\mathrm{d}F(x) \le \int |tx|\,\mathrm{d}F(x) = \beta_1 |t|.$$

To prove the second inequality, we obtain, again making use of (2.11),

$$\begin{aligned}
|1 - f(t)| =& \Big|\int \{e^{itx} - 1\}\,\mathrm{d}F(x)\Big| = \Big|\int (e^{itx} - 1 - itx)\,\mathrm{d}F(x)\Big| \\
&\le \int |e^{itx} - 1 - itx|\,\mathrm{d}F(x) \le \int \tfrac{1}{2}t^2x^2\,\mathrm{d}F(x) = \tfrac{1}{2}\sigma^2 t^2.
\end{aligned}$$

Along the same lines one may prove for example that $|f(t) - (1 - \frac{1}{2}\sigma^2 t^2)| \leq \frac{1}{6}|t|^3 \int |x|^3 \, \mathrm{d}F(x)$. ■

Let g be a real-valued function defined on an interval $[a, b]$ of the real line. The total variation of g on $[a, b]$ is defined as

$$\mathbf{V}_a^b(g) = \sup \sum_{i=1}^{n} |g(x_i) - g(x_{i-1})|$$

where the supremum is taken over all n and all collections $x_0, \ldots, x_n$ such that $a = x_0 < \cdots < x_n = b$. The total variation on the whole real line is defined as

$$\mathbf{V}_{-\infty}^{\infty}(g) = \lim_{x \to \infty} \mathbf{V}_{-x}^{x}(g).$$

In the case $\mathbf{V}_{-\infty}^{\infty}(g)$ we omit limits and write $\mathbf{V}(g)$. A function g is said to be a function of bounded total variation if $\mathbf{V}(g) < \infty$ (or $\mathbf{V}_a^b(g) < \infty$ if it is considered on an interval $[a, b]$). Note that if g has an integrable derivative, then $\mathbf{V}_a^b(g) = \int_a^b |g'| \, \mathrm{d}x$.

Lemma 4. *Let p be a probability density and f the corresponding characteristic function. If p is $m-1$ times differentiable, and $p^{(m-1)}$ is a function of bounded variation, then*

$$|f(t)| \leq \mathbf{V}(p^{(m-1)})/|t|^m$$

for all real t (by definition, $p^{(0)} = p$).

A proof of this lemma is contained in Ushakov and Ushakov (1999).

3. Density estimators with conventional kernels

First we study the 'smooth' case, i.e. when the density to be estimated is one or several times differentiable.

Theorem 1. *Let p be twice differentiable, with p'' a function of bounded variation, $\mathbf{V}(p'') = V_2 < \infty$. If the kernel K has null expectation, and $h_n = h_0 n^{-1/5}$ (h_0 being some constant), then*

$$\mathrm{MISE}(p_n) \leq \left\{ \frac{3\mu_2^2(K) V_2^{5/3}}{10\pi} h_0^4 + \frac{R(K)}{h_0} \right\} n^{-4/5}. \tag{3.1}$$

Proof. Due to Lemma 4, we have

$$|f(t)| \leq \begin{cases} 1 & \text{for } |t| \leq V_2^{1/3}, \\ V_2/|t|^3 & \text{for } |t| > V_2^{1/3}, \end{cases} \tag{3.2}$$

and, due to Lemma 3, $|1-\varphi(h_nt)| \le \frac{1}{2}\mu_2(K)h_n^2t^2$ for all t. Hence

$$\int |f(t)|^2|1-\varphi(h_nt)|^2\,\mathrm{d}t \le \tfrac{1}{2}\mu_2^2(K)h_n^4\int_0^{V_2^{1/3}} t^4\,\mathrm{d}t + \tfrac{1}{2}\mu_2^2(K)h_n^4V_2^2\int_{V_2^{1/3}}^{\infty}\frac{\mathrm{d}t}{t^2}$$
$$= (3/5)\mu_2^2(K)V_2^{5/3}h_0^4\,n^{-4/5}. \tag{3.3}$$

Further, using the Parseval–Plancherel identity, we get

$$\frac{1}{nh_n}\frac{1}{2\pi}\int|\varphi(t)|^2\,\mathrm{d}t = \frac{1}{h_0}n^{-4/5}\int K^2(x)\,\mathrm{d}x = \frac{R(K)}{h_0}n^{-4/5}. \tag{3.4}$$

From (3.3), (3.4) and Lemma 2, we obtain (3.1). ■

Corollary. *Let the conditions of Theorem 1 be satisfied. Then for each n,*

$$\min_{h>0}\mathrm{MISE}(p_n) \le \Big(\frac{3\cdot 5^4}{2^9\pi}\Big)^{1/5}\{\mu_2^2(K)R^4(K)\}^{1/5}V_2^{1/3}n^{-4/5},$$

with minimum of the upper bound attained for

$$h_n = \Big\{\frac{5\pi}{6}\frac{R(K)}{\mu_2^2(K)}\Big\}^{1/5}V_2^{-1/3}n^{-1/5}.$$

If p is only one time differentiable or/and the expectation of K does not equal zero, then results are weaker.

Theorem 2. *Let p be differentiable with p' a function of bounded variation, $\mathbf{V}(p') = V_1 < \infty$. If $h_n = h_0n^{-1/3}$, then*

$$\mathrm{MISE}(p_n) \le \Big\{\frac{4}{3\pi}\mu_1^2(K)V_1^{3/2}h_0^2 + \frac{R(K)}{h_0}\Big\}n^{-2/3}. \tag{3.5}$$

Proof. Due to Lemmas 3 and 4,

$$|f(t)| \le \begin{cases} 1 & \text{for } |t| \le V_1^{1/2}, \\ V_1/|t|^2 & \text{for } |t| > V_1^{1/2}, \end{cases}$$

and $|1-\varphi(h_nt)| \le \mu_1(K)h_n|t|$ for all t. Hence (see the proof of Theorem 1),

$$\int |f(t)|^2|1-\varphi(h_nt)|^2\,\mathrm{d}t \le (8/3)\mu_1^2(K)V_1^{3/2}h_0^2n^{-2/3}. \tag{3.6}$$

And, as in the proof of Theorem 1,

$$\frac{1}{2\pi}\frac{1}{nh_n}\int|\varphi(t)|^2\,\mathrm{d}t = \frac{R(K)}{h_0}n^{-2/3}. \tag{3.7}$$

From (3.6), (3.7) and Lemma 2, we obtain (3.5). ■

Corollary. *Let the conditions of Theorem 2 be satisfied. Then for each n,*

$$\min_{h>0} \mathrm{MISE}(p_n) \le (9/\pi)^{1/3} \mu_1^{2/3}(K) \sqrt{V_1} R(K)^{2/3}\, n^{-2/3}.$$

Theorems 1 and 2 give bounds for the integral deviation of the mean squared error of a kernel estimator from zero. Now we obtain bounds for the sup deviation, in terms of

$$A(K) = \frac{1}{2\pi} \int |\varphi(t)|\, \mathrm{d}t.$$

Theorem 3. *Let p be three times differentiable with p''' a function of bounded variation, $\mathbf{V}(p''') = V_3 < \infty$, and let p be bounded by a. If $h_n = h_0 n^{-1/5}$, then*

$$\sup_x \mathrm{MSE}(p_n(x)) \le \Big\{ \frac{4}{9\pi^2} \mu_2^2(K) V_3^{3/2} h_0^4 + \frac{2aA(K)}{h_0} \Big\}\, n^{-4/5}.$$

Proof. Due to Lemma 4,

$$|f(t)| \le \begin{cases} 1 & \text{for } |t| \le V_3^{1/4}, \\ V_3/|t|^4 & \text{for } |t| > V_3^{1/4}, \end{cases}$$

and, due to Lemma 3, $|1 - \varphi(h_n t)| \le \frac{1}{2}\mu_2(K) h_n^2 t^2$ for all t. Hence

$$\begin{aligned} \int |f(t)|\, |1 - \varphi(h_n t)|\, \mathrm{d}t &\le \mu_2(K) h_n^2 \Big(\int_0^{V_3^{1/4}} t^2\, \mathrm{d}t + V_3 \int_{V_3^{1/4}}^{\infty} \frac{\mathrm{d}t}{t^2} \Big) \\ &= (4/3) \mu_2(K) V_3^{3/4} h_0^2 n^{-2/5}. \end{aligned}$$

To get the result it suffices now to apply Lemma 1. ∎

Corollary. *Let the conditions of Theorem 3 be satisfied. Then for each n,*

$$\min_{h>0} \sup_x \mathrm{MSE}(p_n(x)) \le 5(36\pi^2)^{-1/5} \mu_2^{2/5}(K) V_3^{3/10} A^{4/5}(K) a^{4/5}\, n^{-4/5},$$

with minimum of the upper bound being attained for

$$h_n = \Big\{ \frac{9\pi^2}{8} \frac{A(K)}{\mu_2^2(K)} \Big\}^{1/5} a^{1/5} V_3^{-3/10}\, n^{-1/5}.$$

Theorem 4. *Let p be twice differentiable with p'' a function of bounded variation, and let p be bounded by a. If $h_n = h_0 n^{-1/3}$, then*

$$\sup_x \mathrm{MSE}(p_n(x)) \le \Big\{ \frac{9}{4\pi^2} \mu_1^2(K) V_2^{4/3} h_0^2 + \frac{2aA(K)}{h_0} \Big\} n^{-2/3}. \tag{3.8}$$

Proof. Using (3.2) and the second inequality of Lemma 3, we have $|1 - \varphi(h_n t)| \le \mu_1(K) h_n |t|$. This leads to

$$\int |f(t)|\,|1-\varphi(h_n t)|\,\mathrm{d}t \le 2\mu_1(K) h_n \int_0^{V_2^{1/3}} t\,\mathrm{d}t + 2V_2\mu_1(K)h_n \int_{V_2^{1/3}}^{\infty} \frac{\mathrm{d}t}{t^2}$$
$$= 3\mu_1(K) V_2^{2/3} h_n = 3\mu_1(K) V_2^{2/3} h_0 n^{-1/3}.$$

Using this estimate and Lemma 1, we get (3.8). ■

Corollary. *Let the conditions of Theorem 4 be satisfied. Then for each n,*

$$\min_{h>0} \sup_x \mathrm{MSE}(p_n(x)) \le 3\Big(\frac{9}{4\pi^2}\Big)^{1/3} \mu_1^{2/3}(K) V_2^{4/9} B^{2/3}(K) a^{2/3} n^{-2/3}.$$

Next we consider the so-called non-smooth case. This means that the underlying density function is not supposed to be differentiable or even continuous. Some minimum regularity conditions must be introduced, however (otherwise nothing substantial can be derived). Here this minimum condition will be the boundedness of the total variation of the underlying density. Note that this condition is a little less restrictive than those usually assumed when authors work with the non-smooth case (see for example van Eeden, 1985 and van Es, 1997).

Theorem 5. *Let the underlying density p be a function of bounded variation, $V = \mathbf{V}(p) < \infty$. If $h_n = h_0/(\sqrt{n}\log n)$, then*

$$\mathrm{MISE}(p_{n,h}) \le \frac{\log^2 n}{\sqrt{n}} \Big[\frac{4\sqrt{2}}{\pi} \max\{\sqrt{\mu_1(K)}, \mu_1(K)\} \times \max\{V^{3/2}, V^2\}\, \max\{\sqrt{h_0}, h_0\} + \frac{R(K)}{h_0 \log n}\Big] \tag{3.9}$$

for all $n \ge 16$.

Proof. Let us use Lemma 2. For the second term in the square brackets, due to the Parseval–Plancherel identity, we have

$$\frac{1}{nh_n}\int |\varphi(t)|^2\,\mathrm{d}t = \frac{2\pi}{nh_n}\int K^2(x)\,\mathrm{d}x = \frac{2\pi R(K)}{nh_n}. \tag{3.10}$$

Let us estimate the first term. First we establish the following inequality: for any $0 < \alpha < 1$,

$$|1 - \varphi(t)| \le \mu_1(K)^\alpha 2^{1-\alpha} |t|^\alpha \tag{3.11}$$

for all real t. Indeed, due to Lemma 3,

$$|1 - \varphi(t)| \le \mu_1(K)|t|. \tag{3.12}$$

For $|t| \le 2/\mu_1(K)$, the right hand side of (3.11) majorises the right hand side of (3.12), therefore (3.11) holds for these t. If $|t| > 2/\mu_1(K)$, then (3.11) is evident because its right hand side exceeds 2.

Let α be arbitrary inside $(0, \frac{1}{2})$. Making use of (3.11) and Lemma 4, we get

$$\begin{aligned}\int |f(t)|^2|1-\varphi(h_n t)|^2\,\mathrm{d}t &= 2\int_0^V |f(t)|^2|1-\varphi(h_n t)|^2\,\mathrm{d}t \\ &\quad + 2\int_V^\infty |f(t)|^2|1-\varphi(h_n t)|^2\,\mathrm{d}t \\ &\le 2\mu_1^{2\alpha}(K)2^{2(1-\alpha)}h_n^{2\alpha}\Big(\int_0^V t^{2\alpha}\,\mathrm{d}t + V^2\int_V^\infty t^{2\alpha-2}\,\mathrm{d}t\Big) \\ &= \frac{2^{4-2\alpha}}{1-4\alpha^2}\,\mu_1(K)^{2\alpha}V^{2\alpha+1}h_n^{2\alpha}.\end{aligned}$$

From this estimate and (3.10), using Lemma 1, we obtain

$$\begin{aligned}\mathrm{MISE}(p_{n,h}) &\le \frac{2^{3-2\alpha}}{\pi(1-4\alpha^2)}\mu_1^{2\alpha}(K)V^{2\alpha+1}h_n^{2\alpha} + \frac{R(K)}{nh_n} \\ &= \frac{2^{3-2\alpha}}{\pi(1-4\alpha^2)}\mu_1^{2\alpha}(K)V^{2\alpha+1}h_0^{2\alpha}\Big(\frac{1}{\sqrt{n}\log n}\Big)^{2\alpha} + \frac{R(K)}{h_0}\frac{\log n}{\sqrt{n}}\end{aligned} \tag{3.13}$$

for any $\alpha \in (0, \frac{1}{2})$. Put

$$\alpha = \frac{\log n}{2(\log n + 2\log\log n)}.$$

Then $\frac{1}{4} < \alpha < \frac{1}{2}$ (provided that $n \ge e^e$, which translates into $n \ge 16$), and hence

$$2^{3-2\alpha} < 4\sqrt{2}, \quad \mu_1(K) \le \max\{\sqrt{\mu_1(K)}, \mu_1(K)\},$$

$$V^{2\alpha+1} \le \max\{V^{3/2}, V^2\}, \quad h_0^{2\alpha} \le \max\{\sqrt{h_0}, h_0\}.$$

Therefore from (3.13) we obtain

$$\begin{aligned}\mathrm{MISE}(p_{n,h}) \le &\frac{4\sqrt{2}}{\pi}\max\{\sqrt{\mu_1(K)}, \mu_1(K)\}\ \max\{V^{3/2}, V^2\}\ \max\{\sqrt{h_0}, h_0\} \\ &\times \frac{1}{1-4\alpha^2}\Big(\frac{1}{\sqrt{n}\log n}\Big)^{2\alpha} + \frac{R(K)}{h_0}\frac{\log n}{\sqrt{n}}.\end{aligned} \tag{3.14}$$

Putting now

$$\alpha_0 = \frac{\log n - 2\log\log n}{2(\log n + 2\log\log n)},$$

then $\alpha > \alpha_0$ (if $n \ge e^e$), hence

$$\Big(\frac{1}{\sqrt{n}\log n}\Big)^{2\alpha} < \Big(\frac{1}{\sqrt{n}\log n}\Big)^{2\alpha_0} = \frac{\log n}{\sqrt{n}}. \tag{3.15}$$

It remains to assess the size of $1/(1-4\alpha^2)$. We have

$$\frac{1}{1-4\alpha^2} = \frac{(\log n + 2\log\log n)^2}{(\log n + 2\log\log n)^2 - (\log n)^2} = \frac{(\log n + 2\log\log n)^2}{(2\log n + 2\log\log n)2\log\log n}$$
$$\le \frac{1}{4}\log n + 1 + \frac{\log\log n}{\log n + \log\log n} \le \log n \tag{3.16}$$

if $n \ge e^e$.

From (3.14), (3.15) and (3.16) we finally obtain (3.9). ■

Corollary. *Let p be a unimodal density function, and bounded by a. If $h_n = h_0/(\sqrt{n}\log n)$, then*

$$\mathrm{MISE}(p_{n,h}) \le \frac{\log^2 n}{\sqrt{n}} \cdot \Big[\frac{4\sqrt{2}}{\pi}\max\{\sqrt{\mu_1(K)}, \mu_1(K)\}$$
$$\times \max\{2\sqrt{2}a^{3/2}, a^2\}\ \max\{\sqrt{h_0}, h_0\} + \frac{R(K)}{h_0\log n}\Big].$$

4. The sinc kernel density estimator

The sinc kernel is the function

$$K(x) = \frac{\sin x}{\pi x}$$

with the Fourier transform (defined as the principal value of the corresponding integral)

$$\varphi(t) = \begin{cases} 1 & \text{for } |t| \le 1, \\ 0 & \text{for } |t| > 1. \end{cases}$$

(Sometimes the sinc kernel is defined as $K(x) = \sin(\pi x)/(\pi x)$ with the Fourier transform $\varphi(t) = I\{|t| \le \pi\}$. Both functions $\sin x/(\pi x)$ and $\sin(\pi x)/(\pi x)$ integrate to one in the sense of the principal value, and the difference is only in the scale parameter.)

From now on we focus on the kernel estimator $p_n(x)$ of (2.1) with K being the sinc kernel. It often leads to better performance, and some of its properties are in fact easier to study than for other kernel estimators; see Glad, Hjort and Ushakov (1999a). Its defects – possible negativeness and nonintegrability – can easily be corrected by a certain modification procedure (Glad, Hjort and Ushakov, 1999b). It consists in setting

$$\bar{p}_n(x) = \max\{p_n(x) - \xi, 0\},$$

where the random ξ is chosen so that the integral is 1. After this correction procedure, estimation precision of the estimator is guaranteed to improve.

In terms of the empirical characteristic function $f_n(t)$ the sinc estimator can be expressed as

$$p_n(x) = \frac{1}{2\pi}\int_{-1/h_n}^{1/h_n} e^{-itx} f_n(t)\,\mathrm{d}t. \tag{4.1}$$

Suppose that the characteristic function f of the underlying density p is integrable. First we obtain relations for the sinc estimator, analogous to those of Lemmas 1 and 2 (these cannot be applied directly since now K is not integrable).

Lemma 5. *For the sinc kernel estimator,*

$$\sup_x \mathrm{MSE}(p_n(x)) \le \Big\{\frac{1}{2\pi}\int_{|t|\ge 1/h_n} |f(t)|\,\mathrm{d}t\Big\}^2 + \frac{2}{\pi n h_n}\frac{1}{2\pi}\int |f(t)|\,\mathrm{d}t \tag{4.2}$$

and

$$\mathrm{MISE}(p_n) \le \frac{1}{2\pi}\Big\{\int_{|t|\ge 1/h_n} |f(t)|^2\,\mathrm{d}t + \frac{2}{nh_n}\Big\}. \tag{4.3}$$

Proof. We first prove the first inequality. We have

$$\begin{aligned}
\mathrm{MSE}(p_n(x)) &= \mathrm{E}\Big[\frac{1}{2\pi}\Big\{\int e^{-itx} f(t)\,\mathrm{d}t - \int_{-1/h_n}^{1/h_n} e^{-itx} f_n(t)\,\mathrm{d}t\Big\}\Big]^2 \\
&= \mathrm{E}\Big[\frac{1}{2\pi}\int_{|t|\ge 1/h_n} e^{-itx} f(t)\,\mathrm{d}t + \frac{1}{2\pi}\int_{-1/h_n}^{1/h_n} e^{-itx}\{f(t)-f_n(t)\}\,\mathrm{d}t\Big]^2 \\
&= \Big\{\frac{1}{2\pi}\int_{|t|\ge 1/h_n} e^{-itx} f(t)\,\mathrm{d}t\Big\}^2 + \mathrm{E}\Big[\frac{1}{2\pi}\int_{-1/h_n}^{1/h_n} e^{-itx}\{f(t)-f_n(t)\}\,\mathrm{d}t\Big]^2.
\end{aligned}$$

Let us estimate the second term on the right hand side. Denote it by T_2. Taking into account that

$$\mathrm{E} f_n(u) f_n(v) = (1-1/n) f(u) f(v) + (1/n) f(u+v),$$

we obtain

$$\begin{aligned}
T_2 &= \frac{1}{n}\frac{1}{(2\pi)^2}\int_{-1/h_n}^{1/h_n}\int_{-1/h_n}^{1/h_n} e^{-i(u+v)x}\{f(u+v) - f(u)f(v)\}\,\mathrm{d}u\,\mathrm{d}v \\
&= \frac{1}{2\pi n}\int_{-1/h_n}^{1/h_n}\Big\{\frac{1}{2\pi}\int_{-1/h_n}^{1/h_n} e^{-i(u+v)x} f(u+v)\,\mathrm{d}u\Big\}\,\mathrm{d}v \\
&\qquad\qquad - \frac{1}{n}\Big\{\frac{1}{2\pi}\int_{-1/h_n}^{1/h_n} e^{-itx} f(t)\,\mathrm{d}t\Big\}^2 \\
&= \frac{1}{2\pi n}\int_{-1/h_n}^{1/h_n}\Big\{\frac{1}{2\pi}\int_{-1/h_n+v}^{1/h_n+v} e^{-itx} f(t)\,\mathrm{d}t\Big\}\,\mathrm{d}v - \frac{1}{n}\Big\{\frac{1}{2\pi}\int_{-1/h_n}^{1/h_n} e^{-itx} f(t)\,\mathrm{d}t\Big\}^2.
\end{aligned}$$

Therefore

$$
\begin{aligned}
T_2 &\le \frac{1}{2\pi n}\int_{-1/h_n}^{1/h_n} \mathrm{d}v\,\frac{1}{2\pi}\int |f(t)|\,\mathrm{d}t + \frac{1}{n}\frac{1}{2\pi}\int |f(t)|\,\mathrm{d}t\,\frac{1}{2\pi}\int_{-1/h_n}^{1/h_n} \mathrm{d}s \\
&= \frac{2}{\pi n h_n}\frac{1}{2\pi}\int |f(t)|\,\mathrm{d}t.
\end{aligned}
$$

Thus we obtain (4.2).

Next we prove (4.3). Observe that we may use relation (2.5) with

$$
\hat{f}_n(x) = \begin{cases} f_n(x) & \text{if } |t| \le 1/h_n, \\ 0 & \text{otherwise.} \end{cases}
$$

Therefore

$$
\mathrm{MISE}(p_n) = \frac{1}{2\pi}\Big\{\int_{-1/h_n}^{1/h_n} \mathrm{E}|f_n(t) - f(t)|^2\,\mathrm{d}t + \int_{|t|\ge 1/h_n} |f(t)|^2\,\mathrm{d}t\Big\},
$$

and it suffices to show that

$$
\int_{-1/h_n}^{1/h_n} \mathrm{E}|f_n(t) - f(t)|^2\,\mathrm{d}t \le \frac{2}{nh_n}.
$$

Taking now into account that $\mathrm{E}|f_n(t)|^2 = (1 - 1/n)|f(t)|^2 + 1/n$, we obtain

$$
\int_{-1/h_n}^{1/h_n} \mathrm{E}|f_n(t) - f(t)|^2\,\mathrm{d}t = \frac{1}{n}\int_{-1/h_n}^{1/h_n} (1 - |f(t)^2|)\,\mathrm{d}t \le \frac{1}{n}\int_{-1/h_n}^{1/h_n} \mathrm{d}t = \frac{2}{nh_n}.
$$

This proves the claim. ∎

Now we derive some estimates for MISE and MSE of the sinc estimator in terms of the degree of smoothness of the underlying density. First we consider the non-smooth case, when a density to be estimated is not supposed to be differentiable or even continuous.

Theorem 6. *Let p have bounded variation, $\mathbf{V}(p) = V < \infty$, and let p_n be the sinc estimator. If $h_n = h_0/\sqrt{n}$, then*

$$
\mathrm{MISE}(p_n) \le \frac{1}{\pi\sqrt{n}}\Big(V^2 h_0 + \frac{1}{h_0}\Big).
$$

Proof. Making use of relation (4.3) of Lemma 5 and Lemma 4, we obtain

$$
\begin{aligned}
\mathrm{MISE}(p_n) &\le \frac{1}{2\pi}\Big\{\int_{|t|\ge 1/h_n} |f(t)|^2\,\mathrm{d}t + \frac{2}{nh_n}\Big\} \\
&\le \frac{1}{2\pi}\Big(2V^2\int_{1/h_n}^{\infty}\frac{\mathrm{d}t}{t^2} + \frac{2}{nh_n}\Big) = \frac{1}{\pi\sqrt{n}}\Big(V^2 h_0 + \frac{1}{h_0}\Big),
\end{aligned}
$$

as required. ∎

Corollary 1. *Let the conditions of Theorem 6 be satisfied. Then for each n,*

$$\min_{h>0} \mathrm{MISE}(p_n) \leq \frac{2V}{\pi\sqrt{n}}.$$

Corollary 2. *Let p be a unimodal density function, and let p_n be the sinc estimator. If p is bounded by a, and $h_n = h_0/\sqrt{n}$, then*

$$\mathrm{MISE}(p_n) \leq \frac{1}{\pi\sqrt{n}}\Big(4a^2 h_0 + \frac{1}{h_0}\Big),$$

and

$$\min_{h>0} \mathrm{MISE}(p_n) \leq \frac{4a}{\pi\sqrt{n}}.$$

Now consider the case when the density to be estimated is m times differentiable, $m \geq 1$. It will be shown that in this case the upper bound for MISE of the sinc estimator has order $n^{-2m/(2m+1)}$ that in principal cannot be achieved (for $m > 2$) for kernel estimators with kernels being density functions.

Theorem 7. *Let p be m times differentiable with $p^{(m)}$ a function of bounded variation, $\mathbf{V}(p^{(m)}) = V_m < \infty$. If p_n is the sinc estimator, and $h_n = h_0 n^{-1/(2m+1)}$, then*

$$\mathrm{MISE}(p_n) \leq \frac{1}{2\pi}\Big\{\frac{4(m+1)}{2m+1} V_m^{(2m+1)/(m+1)} h_0^{2m} + \frac{2}{h_0}\Big\} n^{-2m/(2m+1)}. \qquad (4.4)$$

Proof. We have

$$\begin{aligned}\int_{|t|\geq 1/h_n} |f(t)|^2\, \mathrm{d}t &= h_n^{2m} \int_{|t|\geq 1/h_n} \Big(\frac{1}{h_n}\Big)^{2m} |f(t)|^2\, \mathrm{d}t \\ &\leq h_n^{2m} \int_{|t|\geq 1/h_n} |t|^{2m} |f(t)|^2\, \mathrm{d}t \leq h_n^{2m} \int |t|^{2m} |f(t)|^2\, \mathrm{d}t.\end{aligned} \qquad (4.5)$$

Let us estimate the integral on the right hand side, making use of Lemma 4. We have

$$|f(t)| = \begin{cases} 1 & \text{for } |t| \leq V_m^{1/(m+1)}, \\ V_m/|t|^{m+1} & \text{for } |t| > V_m^{1/(m+1)}, \end{cases}$$

therefore

$$\begin{aligned}\int |t|^{2m} |f(t)|^2\, \mathrm{d}t &\leq 2\int_0^{V_m^{1/(m+1)}} t^{2m}\, \mathrm{d}t + 2V_m^2 \int_{V_m^{1/(m+1)}}^{\infty} \frac{\mathrm{d}t}{t^2} \\ &= \frac{4(m+1)}{2m+1} V_m^{(2m+1)/(m+1)}.\end{aligned} \qquad (4.6)$$

Thus, from inequality (4.3) of Lemma 5, and relations (4.5) and (4.6), we obtain (4.4). ∎

Corollary. *Let the conditions of Theorem 7 be satisfied. Then for each n,*

$$\min_{h>0} \mathrm{MISE}(p_n) \le \frac{1}{2\pi}\{4(m+1)\}^{1/(2m+1)}\Big(\frac{2m+1}{m}\Big)^{2m/(2m+1)} V_m^{1/(m+1)} n^{-2m/(2m+1)}.$$

Theorem 8. *Let p be m times differentiable, with $p^{(m)}$ a function of bounded variation, $\mathbf{V}(p^{(m)}) = V_m < \infty$. If p_n is the sinc estimator, and $h_n = h_0 n^{-1/(2m-1)}$, then*

$$\begin{aligned}\sup_x \mathrm{MSE}(p_n(x)) \le& \frac{1}{\pi^2}\Big\{\frac{(m+1)^2}{m^2} V_m^{2m/(m+1)} h_0^{2(m-1)} \\ &+ 2\Big(V_m^{1/(m+1)} + \frac{1}{m} V_m^{m/(m+1)}\Big)\frac{1}{h_0}\Big\} n^{-2(m-1)/(2m-1)}.\end{aligned}$$

The proof of this theorem is analogous to that of Theorem 7, one just needs to use relation (4.2) of Lemma 5 instead of relation (4.3) and take into account that due to Lemma 4,

$$\begin{aligned}A(p) = \frac{1}{2\pi}\int |f(t)|\,\mathrm{d}t \le& \frac{1}{2\pi}\int_{-V_m^{1/(m+1)}}^{V_m^{1/(m+1)}} \mathrm{d}t + \frac{1}{2\pi}\int_{|t|>V_m^{1/(m+1)}} \frac{V_m\,\mathrm{d}t}{|t|^{m+1}} \\ \le& \frac{1}{\pi}\Big\{V_m^{1/(m+1)} + \frac{1}{m} V_m^{m/(m+1)}\Big\}.\end{aligned}$$

Corollary. *Let the conditions of Theorem 8 be satisfied. Then for each n,*

$$\begin{aligned}\min_{h>0}\sup_x \mathrm{MSE}(p_n(x)) \le& \frac{2m-1}{\pi^2}\Big(\frac{m+1}{m}\Big)^{2/(2m-1)} \\ &\times \Big\{\frac{m + V_m^{(m-1)/(m+1)}}{m(m-1)}\Big\}^{\frac{2m-2}{2m-1}} V_m^{2/(m+1)} n^{-2(m-1)/(2m-1)}.\end{aligned}$$

Now we proceed to the 'supersmooth' case which we define in terms of characteristic functions (although this class of distribution can be defined in terms of density functions as well, a description in terms of characteristic functions is simpler, more natural and more convenient for our purposes). A distribution F with characteristic function $f(t)$ is said to be supersmooth if for some $\alpha > 0$ and $\gamma > 0$,

$$B(p;\alpha,\gamma) = \int e^{\gamma|t|^\alpha}|f(t)|\,\mathrm{d}t < \infty.$$

Thus a normal density is supersmooth with $\alpha = 2$ while a Cauchy is supersmooth with $\alpha = 1$, for example.

Theorem 9. *Let the characteristic function f of p have a finite $B(p;\alpha,\gamma)$ value, for some positive α and γ. If p_n is the sinc estimator, and*

$$h_n = \Big\{\frac{1}{\gamma}\log(h_0 n)\Big\}^{-1/\alpha}, \quad \text{with } h_0 \ge \frac{1}{n},$$

then

$$\mathrm{MISE}(p_n) \le \frac{1}{2\pi n}\Big\{\frac{2}{\gamma^{1/\alpha}}(\log h_0 + \log n)^{1/\alpha} + \frac{B(p;\alpha,\gamma)}{h_0}\Big\}. \tag{4.7}$$

Proof. We have

$$\int_{|t|\ge 1/h_n} |f(t)|^2\,\mathrm{d}t \le \int_{|t|\ge 1/h_n} |f(t)|\,\mathrm{d}t = e^{-\gamma/h_n^\alpha}\int_{|t|\ge 1/h_n} e^{\gamma/h_n^\alpha}|f(t)|\,\mathrm{d}t$$
$$\le e^{-\gamma/h_n^\alpha}\int e^{\gamma|t|^\alpha}|f(t)|\,\mathrm{d}t = \frac{B(p;\alpha,\gamma)}{nh_0}.$$

Using this estimate and inequality (4.3) of Lemma 5, we obtain (4.7). ■

Theorem 10. *Let the conditions of Theorem 9 be satisfied, and let again $A(p) = (2\pi)^{-1}\int |f(t)|\,\mathrm{d}t$. Then*

$$\sup_x \mathrm{MSE}(p_n(x)) \le \frac{1}{n}\Big\{\frac{2A(p)}{\pi\gamma^{1/\alpha}}(\log h_0 + \log n)^{1/\alpha} + \frac{B^2(p;\alpha,\gamma)}{4\pi^2 nh_0}\Big\}.$$

The proof of the theorem is similar to that of Theorem 9 (inequality (4.2) is used instead of (4.3)).

Theorems 9 and 10 can be improved for one subclass of supersmooth densities. The result is given by the next theorem and is quite curious. Note that this theorem corresponds to a special case of a result by Ibragimov and Khas'minskii (1982), and we give it here for the completeness.

Theorem 11. *Let the characteristic function f of p satisfy the condition: there exists $\tau > 0$ such that $f(t) = 0$ for $|t| > \tau$. If p_n is the sinc estimator, and $h_n \le 1/\tau$, then*

$$\sup_x \mathrm{MSE}(p_n(x)) \le \frac{2A(p)}{\pi n h_n} \le \frac{2\tau}{\pi^2 n h_n},$$

and

$$\mathrm{MISE}(p_n) \le \frac{1}{\pi n h_n}.$$

In particular, if $h_n = \mathrm{const.} = 1/\tau$, then

$$\sup_x \mathrm{MSE}(p_n(x)) \le \frac{2\tau^2}{\pi^2 n} \quad \text{and} \quad \mathrm{MISE}(p_n) \le \frac{\tau}{\pi n}.$$

A proof of the theorem can be immediately obtained from inequalities (4.2) and (4.3) of Lemma 5: integrals on the right hand sides of these vanish when $h_n \le 1/\tau$.

Theorem 11 implies in particular that if the characteristic function of the underlying distribution vanishes for large values of the argument, and one uses the sinc estimator for approximation, then p_n converges to p as $n \to \infty$ even when h_n does not converge to zero.

5. Discussion and applications

This article has provided upper bounds for both the traditional MISE and also the less worked with max-MSE performance measures of kernel density estimators. A list of such upper bounds has been provided, under various sets of assumptions, for both the traditional kernels as well as for the sinc kernel, which has particularly attractive features. Our finite-sample results have been reached entirely outside the customary framework of asymptotics, Taylor expansions and small bandwidths, through the extensive use of characteristic and empirical characteristic functions. Below we give some concluding remarks, pointing to ways in which the results can be applied in statistics.

5.1. Rule-of-thumb bandwidths for MISE and max-MSE. Consider kernel estimators with a traditional kernel K, a symmetric density. The traditional large-sample approximations lead to an asymptotically optimal bandwidth of size

$$h_n = \{R(K)/\mu_2^2(K)\}^{1/5} R(p'')^{-1/5}\, n^{-1/5},$$

and with consequent minimum approximate MISE of size

$$\min_{h>0} \mathrm{AMISE}(p_n) = (5/4)\{\mu_2^2(K)R^4(K)\}^{1/5} R(p'')^{1/5}\, n^{-4/5},$$

see for example Wand and Jones (1995). When K is standard normal, and p is a normal density with standard deviation σ, this leads to the popular 'normal rule-of-thumb' bandwidth $h_n = 1.0592\,\sigma\, n^{-1/5}$. Note that the structure of these classic results is very similar to that seen in Theorem 1 and its corollary; in particular, the well-known large-sample result about the $n^{-4/5}$ precision rate is here reached entirely without asymptotics machinery or approximations.

It is interesting to compare the above with what one finds using the upper bounds. For a normal density, $V_2 = \int |p'''|\,\mathrm{d}x$ is found to be the scale factor σ^{-3} times $2(2\pi)^{-1/2}\{1 + 4\exp(-3/2)\} = 1.5100$, and this leads via Theorem 1 to the rule $h_n = 0.8204\,\sigma\, n^{-1/5}$. This has been calculated using upper bound results derived under minimal assumptions, and which hence do not pretend to be very accurate for smooth densities like the normal. It is comforting to see that only a moderate amount is lost in precision, in this very smooth case, since the ratio of the minimised upper bound to the minimised asymptotic MISE is found to be 1.2911.

The max-MSE criterion is a natural venue, seemingly not travelled before. It is difficult to reach applicable results for this criterion based on the traditional approximations. However, Theorem 3 and its corollary provide ways of bounding the max-MSE when there is information on $V_3 = \int |p^{(4)}|\,\mathrm{d}x$. If p is normal, then V_3 can be shown to be the scale factor σ^{-4} times $4\{(3b - b^3)\phi(b) - (3c - c^3)\phi(c)\} = 2.8006$, where $b = (3 - \sqrt{6})^{1/2}$ and $c = (3 + \sqrt{6})^{1/2}$, and ϕ is the standard normal density. The normal rule-of-thumb, when the normal kernel is used, becomes $h_n = 1.1883\,\sigma\,n^{-1/5}$, which again is not far from the traditional rule-of-thumb.

We also point out that some of these results may be sharpened under further assumed constraints on the underlying density. The quite crude bound (3.2) has for example been used for $|f(t)|$, which could be bounded more effectively under such additional restrictions. This is not pursued here, however.

5.2. Cross-validation and normal rule-of-thumb in new light. Results reached in this article, about MISE and upper bounds expressed in terms of characteristic functions, point to new ways in which bandwidths can be selected from data. Expressions (2.8) and its upper bound given in Lemma 2 depend on $q(t) = |f(t)|^2$, but not on other aspects of the underlying density p. A suitable estimate $\hat{q}(t)$ may now be inserted in these expressions, after which one may minimise over the smoothing parameter h. For the normal kernel, this could for example mean minimising

$$Q_n(h) = \frac{1}{2\pi}\int \hat{q}(t)\{1-\exp(-\tfrac{1}{2}h^2t^2)\}^2\,\mathrm{d}t + \frac{1}{2\pi}\frac{1}{n}\int \exp(-h^2t^2)\,\mathrm{d}t$$

over h, after having selected an estimator $\hat{q}(t)$. Interestingly it turns out that this scheme reproduces the well-known 'unbiased cross-validation' rule, see for example Wand and Jones (1995), when one employs the natural nonparametric unbiased estimator

$$\hat{q}(t) = \frac{1}{n(n-1)}\sum_{j\neq k}\exp(it(X_j-X_k)) = \frac{2}{n(n-1)}\sum_{j<k}2\cos(t(X_j-X_k)).$$

Other methods emerge by using alternative estimators for $q(t)$. If one above uses the simplest parametric estimate, namely $\exp(-\hat{\sigma}^2t^2)$ (or a debiased version thereof) corresponding to a normality approximation, then minimising $Q_n(h)$ is a better finite-sample version of the classic rule-of-thumb $1.0592\,\hat{\sigma}\,n^{-1/5}$. Semiparametric versions of this argument would be worth studying; see Hjort (1999) for a similar enterprise.

5.3. Minimax precision control. Another type of application would be following. Assume that an upper bound for $V_2 = \int |p'''|\,\mathrm{d}x$ is established, say $V_2 \le v_2$. This is a statement of the maximum envisaged wigglyness of the density; a small V_2 would mean a density which can be approximated with a quadratic function. The bound v_2 could be set after inspection of data or from prior grounds. – In such a situation the corollary of Theorem 1 can be given to find a sample size n_0 guaranteed to secure a MISE-accuracy below a given threshold, for the big class of all densities with $V_2 \le v_2$.

Variations of this scenario can easily be given, for example selecting a sample size necessary to secure max-MSE below a certain precision threshold for all relevant densities, as constrained with bounds on V_3 and the maximum value a.

5.4. Large-sample superiority of the sinc method. Approximation of densities via the sinc kernel can often be more accurate than with traditional kernels. This has been pointed out early on by Davis (1977), but the method does not seem popular in practice. That it takes negative values and is not (Lebesgue)-integrable is not a real concern, since it can be repaired for these defects in an automatic fashion which also guarantees precision improvement; see Glad, Hjort and Ushakov (1999b). And results from Section 4 above, with further analysis provided in Glad, Hjort and Ushakov (1999a), give clear indications of the strong performance of the sinc method.

Theorem 9 shows that the rate of MISE towards zero is often much better than the $n^{-4/5}$ available with traditional kernels. For any mixture of normals, for example, the rate is $O((\log n)^{1/2}/n)$, while if some Cauchy type tail behaviour is mixed in it becomes $O((\log n)/n)$. The same remarks apply to the max-MSE performance criterion.

That the sinc method also can perform better than the traditional ones for non-smooth cases is made clear by Theorem 6, where a $O(n^{-1/2})$ rate is exhibited for MISE under a minimal assumption on p, compared to the $O(n^{-1/2}\log^2 n)$ rate for the ordinary methods.

Acknowledgments. The second author is grateful for the opportunity to visit the Department of Mathematics at the University of Oslo during the summer of 1998, and the first author equally grateful for being able to visit Механико-математический факультет Московского государственного университета in 1999. These visits have been made possible with partial funding from the Norwegian Research Foundation and the Department of Mathematics at the University of Oslo. The authors also appreciate fruitful discussions with Ingrid Glad.

References

Davis, K.B. (1977). Mean integrated square error properties of density estimates. *Annals of Statistics* **5**, 530–535.

Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L_1 View.* Wiley, New York.

van Eeden, C. (1985). Mean integrated squared error of kernel estimators when the density and its derivatives are not necessarily continuous. *Annals of the Institute of Statistical Mathematics* **37**, Part A, 461–472.

van Es, A.J. (1997). A note on the integrated squared error of a kernel density estimator in non-smooth cases. *Statistics and Probability Letters* **35**, 241–250.

Feller, W. (1971). *An Introduction to Probability Theory and its Applications*, Vol. 2 (2nd ed.). Wiley, New York.

Glad, I.K., Hjort, N.L. and Ushakov, N.G. (1999a). Density estimation using the sinc kernel. *Statistical Research Report*, Department of Mathematics, University of Oslo.

Glad, I.K., Hjort, N.L. and Ushakov, N.G. (1999b). Correction of density estimators which are not densities. *Statistical Research Report*, Department of Mathematics, University of Oslo.

Hjort, N.L. (1999). Towards semiparametric bandwidth selectors for kernel density estimators. *Statistical Research Report*, Department of Mathematics, University of Oslo.

Ibragimov, I.A. and Khas'minskii, R.Z. (1982). Estimation of distribution density belonging to a class of entire functions. *Theory of Probability and their Applications* **27**, No. 3, 551–562.

Titchmarsh, E. (1937). *Introduction to the Theory of Fourier Integrals.* Clarendon Press, Oxford.

Ushakov, N.G. and Ushakov, V.G. (1999). Some inequalities for characteristic functions of densities with bounded variation. *Statistical Research Report*, Department of Mathematics, University of Oslo.

Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing.* Chapman and Hall, London.

Nils Lid Hjort
Department of Mathematics
University of Oslo
P.B. 1053 Blindern
Oslo, Norway

Nikolai G. Ushakov
Institute of Microelectronics Technology
Российская Академия Наук
142 432 Черноголовка
Московская область, Россия

1999
2000
3111
⋮