The Annals of Statistics
1999, Vol. 27, No. 5, 1600–1615

OPTIMAL CONVERGENCE RATES FOR GOOD'S NONPARAMETRIC MAXIMUM LIKELIHOOD DENSITY ESTIMATOR

BY P. P. B. EGGERMONT AND V. N. LARICCIA

University of Delaware

We study maximum penalized likelihood density estimation using the first roughness penalty functional of Good. We prove a simple pointwise comparison result with a kernel estimator based on the two-sided exponential kernel. This leads to L^1 convergence results similar to those for kernel estimators. We also prove Hellinger distance bounds for the roughness penalized estimator.

1. Introduction. We are interested in nonparametric density estimation from independent identically distributed observations. Two standard examples are "plain" density estimation and the semiparametric deconvolution problem. The standard estimation procedures for problems like these are based on either kernel density estimators or maximum penalized likelihood estimators (MPLEs). For "plain" density estimation the success of kernel estimators is just about uncontested, but for the deconvolution problem mple's seem much more appropriate. Unfortunately, since in the deconvolution problem explicit expressions for the MPLEs are lacking, it appears close to impossible to prove reasonable convergence rates under reasonable conditions for these estimators. Even for "plain" density estimation MPLEs have resisted all attempts—until now. However, the actual purpose of this paper is to show that there may be more to MPLEs than meets the eye and in the process create a renewed interest in maximum penalized likelihood estimation for indirect estimation problems.

Let $X_1, X_2, \ldots, X_n$ be univariate independent identically distributed (iid) random variables drawn from an unknown distribution F_o with density f_o. We wish to estimate f_o without assuming a parametric model, by maximum penalized likelihood estimation; that is, we estimate f_o by the solution of

$$\begin{aligned} &\text{minimize} \quad -\frac{1}{n}\sum_{i=1}^{n} \log f(X_i) + \int_{\mathbb{R}} f(x)\,dx + h^2 R(f) \\ &\text{subject to} \quad f \text{ is a pdf,} \end{aligned} \tag{1.1}$$

Received January 1998; revised July 1999.
AMS 1991 *subject classification*. 62G07.
Key words and phrases. Nonparametric density estimation, maximum likelihood, roughness penalization, Hellinger distance.

where $R(f)$ is the roughness penalization functional. The one that interests us here is the first roughness penalty functional of Good (1971) given by

$$R(f) = \int_{\mathbb{R}} \frac{|f'(x)|^2}{f(x)}\, dx, \tag{1.2}$$

but other choices are possible. We mention those of Silverman (1982) and Cox and O'Sullivan (1990) involving derivatives of the log-density. The drawbacks of maximum penalized likelihood estimators are apparent. The computation of these estimators is nontrivial (although at present this is less of a concern than it was 20 years ago), and proving (L^1) convergence rates under reasonable conditions was unexpectedly hard: either the rates or the conditions are nowhere near optimal. See Klonias (1982, 1984), Silverman (1982) and Cox and O'Sullivan (1990). So, when Wahba (1981), Rudemo (1982), Hall (1983) and Stone (1984) proposed cross validation for the selection of the window parameter, the flood gates were opened and the kernel estimators of Akaike (1954), Rosenblatt (1956) and Parzen (1962) became the sole object of study. The success of the kernel estimators is easy to explain; for known smoothing parameter h they are easy to compute, and the large sample asymptotic behavior (L^1 error) is straightforward to determine, the kernel estimator being a nice linear transformation of the empirical distribution function F_n. The kernel estimators may be written as

$$f^{nh}(x) = A_h * dF_n(x) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} A_h(x - X_i), \qquad x \in \mathbb{R}, \tag{1.3}$$

where $A_h(x) = h^{-1}A(h^{-1}x)$ for a pdf A. However, despite the fact that maximum penalized likelihood estimators are such a pain, there are areas where they are much more appropriate than kernel estimators, for example, for the nonparametric deconvolution problem as demonstrated in Eggermont and LaRiccia (1997). In this paper we discuss maximum penalized likelihood density estimation using the Good (1971) penalization. Although it would seem that the progress made here is mostly technical in nature, we claim that this was precisely what stood in the way of the possible success of maximum penalized density estimators even for the (generalized) deconvolution problem. Thus we give a full analysis of the MPLEs using the penalization of Good (1971), but in the course of doing so we stumbled upon an unexpected result.

In this paper we show that the maximum penalized likelihood estimator with Good's roughness penalization literally compares quite nicely with the kernel estimator with two-sided exponential kernel, which implies that it has the same L^1 consistency behavior under the same minimal conditions on smoothness and tail behavior (integrable second derivative, and existence of a moment of order greater than 1). We use the old analysis of Klonias (1982), but improve on it at a crucial juncture and override it all with the pointwise comparison with kernel estimators. At the end of the paper we show some simulation results comparing the mple under discussion with the kernel

density estimator using the Epanechnikov kernel. By and large, the estimators are indistinguishable.

We finish the introduction with the initial step that seems to lead naturally to everything that follows. In problem (1.1) instead of minimizing over all pdfs, we minimize over all nonnegative (continuous) functions. So we consider

$$\begin{aligned} &\text{minimize} \quad -\frac{1}{n}\sum_{i=1}^{n}\log f(X_i) + \int_{\mathbb{R}} f(x)\,dx + h^2 R(f) \\ &\text{subject to} \quad f \geq 0. \end{aligned} \tag{1.4}$$

The solution is denoted by f_{nh}. It appears that something is lost by ignoring the pdf constraint. In fact, a simple scaling argument shows that

$$\int_{\mathbb{R}} f_{nh}(y)\,dy = 1 - h^2 \left\| \left\{ (f_{nh})^{1/2} \right\}' \right\|_2^2, \tag{1.5}$$

so that f_{nh} is always a sub pdf. [This is a trick of Silverman (1982): in (1.4) minimize over $f = tf_{nh}$, with scalar $t > 0$. The minimum is attained at $t = 1$, and setting the derivative with respect to t equal to 0 gives (1.5).] In the course of the paper it will become clear that ignoring the pdf constraint is of minimal importance and that it is actually quite natural and advantageous to do so.

Following the suggestion of de Montricher, Tapia and Thompson (1975), we use the transformation $f \mapsto u^2$ which leads to the problem (with h replaced by $h/2$)

$$\begin{aligned} &\text{minimize} \quad L_h(u, F_n) \overset{\text{def}}{=} -2\int_{\mathbb{R}} \{\log u(x)\}\,dF_n(x) + \|u\|_2^2 + h^2\|u'\|_2^2 \\ &\text{subject to} \quad u \geq 0. \end{aligned} \tag{1.6}$$

Here $\|\cdot\|_p$ denotes the L^p norm on the line (we need $p = 1$ and 2). Indeed, de Montricher, Tapia and Thompson (1975) show that problems (1.4) and (1.6) are equivalent (actually, they show this with the pdf constraint, but the same argument goes through without it). Both problems have unique, positive solutions, which are related in the obvious way. Finally, it should come as no surprise that the large sample asymptotic problem

$$\begin{aligned} &\text{minimize} \quad L_h(u, F_o) \\ &\text{subject to} \quad u \geq 0 \end{aligned} \tag{1.7}$$

plays an important role as well. We denote the solution of (1.6) by u_{nh} and the solution of (1.7) by u^h. The transformation $f \mapsto u^2$ makes it quite natural to study

$$\|u_{nh} - (f_o)^{1/2}\|_2^2 = \|(f_{nh})^{1/2} - (f_o)^{1/2}\|_2^2 = \mathrm{H}(f_{nh}, f_o), \tag{1.8}$$

which is the Hellinger distance between $f_{nh} = (u_{nh})^2$ and f_o. Indeed, we shall derive upper bounds for (1.8) involving $\mathrm{H}(T_h * dF_n, T_h * dF_o)$ for suitable kernels T_h.

Good (1971), Good and Gaskins (1971), Thompson and Tapia (1990) also discuss a second roughness penalization functional $R(f) = R(u^2) = \alpha\|u'\|_2^2 + \beta\|u''\|_2^2$ for which the associated problems (1.4) and (1.6) are not equivalent in general, but for the special choice

$$h^2R(u^2) = R_h(u^2) \stackrel{\text{def}}{=} 2h^2\|u'\|_2^2 + h^4\|u''\|_2^2 \tag{1.9}$$

they are, and this may be treated similarly to the first roughness penalization, for example, the estimator is given implicitly by $u = S_h * S_h * (dF_n/u)$; compare (4.2) below.

2. The main result. We make the standard assumptions about the density f_o in the context of kernel density estimation, that is, that f_o'' is integrable, and that f_o has a moment of order > 1, but need to make a small concession to the penalization functional involved, that is, $R(f_o) < \infty$. So we assume

$$\|f_o''\|_1 < \infty, \qquad \left\|\left\{(f_o)^{1/2}\right\}'\right\|_2 < \infty, \tag{2.1}$$

$$\mathbb{E}[|X|^m] < \infty \quad \text{for some } m > \kappa > 1. \tag{2.2}$$

These conditions are sufficient for the optimal asymptotic rate of convergence for the L^1 error $\|f_{nh} - f_o\|_1$. For better bounds on the Hellinger distance we need to assume finite moments of higher order, and to get good rates essentially exponential decay is required, that is,

$$\mathbb{E}[e^{r|X|}] < \infty \quad \text{for some } r > 0, \tag{2.3}$$

as well as the stronger smoothness condition

$$\left\|\left\{(f_o)^{1/2}\right\}''\right\|_2 < \infty. \tag{2.4}$$

Indeed, (2.4) implies the second inequality of (2.1), by the standard interpolation inequalities [see Adams (1975)] and it implies the first inequality of (2.1) by writing

$$\{f_o\}'' = \left\{\left((f_o)^{1/2}\right)^2\right\}'' = 2(f_o)^{1/2}\left\{(f_o)^{1/2}\right\}'' + 2\left|\left\{(f_o)^{1/2}\right\}'\right|^2,$$

and using Cauchy–Schwarz.

The bounds arrived at for the errors in various norms arise from various bounds for kernel density estimators with one-sided and two-sided exponential kernels,

$$S_h(x) = (2h)^{-1}\exp(-h^{-1}|x|), \qquad -\infty < x < \infty, \tag{2.5}$$

$$T_h(x) = h^{-1}\exp(-h^{-1}x), \qquad x > 0, \tag{2.6}$$

and $T_h(x) = 0$ for $x < 0$.

The main theorem is as follows.

THEOREM 2.7. *Under the assumptions* (2.1), (2.2),

$$\|f_{nh} - f_o\|_1 =_{\text{as}} \mathcal{O}(n^{-2/5}),$$

provided $h = h_n \asymp n^{-1/5}$.

There is no hint yet as to why Theorem (2.7) should be true. The one result that makes it possible is the direct comparison with kernel density estimators:

THEOREM 2.8. *For all* $h > 0$ *and all* $x \in \mathbb{R}$,

$$\tfrac{1}{2}S_{h/2} * dF_n(x) \le f_{nh}(x) \le S_{h/\sqrt{2}} * dF_n(x)$$

and

$$\|f_{nh} - S_{h/\sqrt{2}} * dF_n\|_1 = h^2 \left\| \left\{ (f_{nh})^{1/2} \right\}' \right\|_2^2.$$

Note that the upper bound on f_{nh} is very good, since f_{nh} (an estimator for a pdf) is bounded above by a pdf. This almost immediately implies the bound on the L^1 error given. The next step in getting good bounds on the L^1 error $\|f_{nh} - f_o\|_1$ is bounding $\|\{(f_{nh})^{1/2}\}'\|_2^2$, for appropriate h. Then the same rates of convergence as for kernel density estimation would follow.

THEOREM 2.9. *Under the assumptions* (2.1), (2.2), *for* $h \asymp n^{-1/5}$,

$$\left\| \left\{ (f_{nh})^{1/2} \right\}' \right\|_2 =_{\text{as}} \mathcal{O}(1).$$

In proving Theorem 2.9 it is quite natural to prove bounds on the Hellinger distance.

THEOREM 2.10. (a) *Under assumption* (2.1), *if* f_o *has a finite moment of order* $m > \kappa > 1$ *then for deterministic* h *with* $h \asymp n^{-1/5}$,

$$\mathrm{H}(f_{nh}, f_o) =_{\text{as}} \mathcal{O}(n^{-2/5}).$$

(b) *Under assumption* (2.1), *if* f_o *has a finite moment of order* $m > \kappa > 2$ *then for deterministic* h *with* $h \asymp n^{-\kappa/(5\kappa+4)}$, *for all* $s > 1$,

$$\mathrm{H}(f_{nh}, f_o) =_{\text{as}} \mathcal{O}\left(n^{-4\kappa/(5\kappa+1)} (\log n)^s\right).$$

(c) *If*, *moreover*, *assumption* (2.3) *holds*, *then for all* $s > 2$,

$$\mathrm{H}(f_{nh}, f_o) =_{\text{as}} \mathcal{O}\left(n^{-4/5} (\log n)^s\right).$$

Finally we state the universal consistency of f_{nh} under minimal conditions.

THEOREM 2.11. *For every density* f_o, *with* $(f_o)^{1/2}$ *integrable*, *if* $h \to 0$, $nh \to \infty$,

$$\|f_{nh} - f_o\|_1 \to 0.$$

3. The main approach. To prove the results in Section 2 it may be possible to take shortcuts here and there, but as argued in the Introduction, the interest is in a full analysis of the maximum penalized likelihood estimator.

The main tool in the analysis is provided by the Euler equations for the maximum penalized likelihood estimation problem (1.6) and by their implicit solution.

THEOREM 3.1 [de Montricher, Tapia and Thompson (1975)]. *Let* Φ *be a distribution function, and let* v *be the minimizer of* $L_h(u, \Phi)$ *over all* $u \in L^2(\mathbb{R})$, $u \geq 0$. *Then* v *solves the boundary value problem*

$$-h^2 v'' + v = \frac{d\Phi}{v}, \qquad x \in \mathbb{R}, \tag{3.2}$$
$$v(x) \to 0, \qquad x \to \pm\infty.$$

Moreover, v *then satisfies*

$$v = S_h * \frac{d\Phi}{v} = \int_{\mathbb{R}} S_h(x - y)\frac{d\Phi(y)}{v(y)}. \tag{3.3}$$

Decomposing as usual the error $u_{nh} - (f_o)^{1/2}$ into an asymptotic variance and asymptotic bias term,

$$\|u_{nh} - (f_o)^{1/2}\|_2 \leq \|u_{nh} - u^h\|_2 + \|u^h - (f_o)^{1/2}\|_2, \tag{3.4}$$

the above theorem allows us to determine bounds for each.

THEOREM 3.5 (The asymptotic bias). *Let* $w = (f_o)^{1/2}$ *and* $u = u^h$. (a) *If* w' *is square integrable, then*

$$\|(u^h)'\|_2 \leq \|w'\|_2.$$

(b) *If* w'' *is square integrable, then*

$$h^2\|u' - w'\|_2^2 + \|u - w\|_2^2 \leq h^4\|w''\|_2^2.$$

THEOREM 3.6 (The asymptotic variance). *There exists a constant* c *such that for all* h, *and with* $\lambda = h/2$,

$$\|u_{nh} - u^h\|_2^2 + h^2\left\|\left(u_{nh} - u^h\right)'\right\|_2^2 \leq cH(T_\lambda * dF_n, T_\lambda * dF_o).$$

The well-known bound

$$\mathrm{H}(T_\lambda * dF_n, T_\lambda * dF_o) \leq \|T_\lambda * dF_n - T_\lambda * dF_o\|_1, \tag{3.7}$$

then yields the same asymptotic rates of convergence for the Good estimator as for kernel density estimation.

In the following sections the claims made in Sections 2 and 3 are substantiated.

4. Comparison with kernel estimators. Here we prove the comparison Theorem 2.8. We consider problem (1.6), but the argument applies to arbitrary distribution functions. From Theorem 3.1 it follows that $u = u_{nh}$

solves

$$-h^2u'' + u = \frac{dF_n}{u}, \qquad x \in \mathbb{R}, \tag{4.1}$$
$$u(x) \to 0, \qquad x \to \pm\infty$$

and

$$u = S_h * \frac{dF_n}{u} = \frac{1}{n}\sum_{i=1}^{n} \frac{S_h(x - X_i)}{u(X_i)}, \qquad x \in \mathbb{R}. \tag{4.2}$$

We need to repeat another observation of de Montricher, Tapia and Thompson (1975), easily enough verified after the fact, that is, that $|(S_h)'(x)| \le h^{-1}S_h(x)$ for all x, so that

$$|u'(x)| \le h^{-1}u(x), \qquad x \in \mathbb{R}. \tag{4.3}$$

The comparison with kernel estimators is now at the surface. Observe that for all u,

$$(u^2)'' = 2uu'' + 2(u')^2,$$

so that the differential equation (4.1) becomes

$$-\tfrac{1}{2}h^2(u^2)'' + u^2 = dF_n - h^2(u')^2. \tag{4.4}$$

It follows from Theorem 3.1 that $u = u_{nh}$ satisfies

$$u^2 = S_{h/\sqrt{2}} * \{dF_n - h^2(u')^2\}, \tag{4.5}$$

whence

$$u^2 \le S_{h/\sqrt{2}} * dF_n. \tag{4.6}$$

The rather surprising consequence is that $f_{nh} = u^2 = (u_{nh})^2$ satisfies

$$\int_{\mathbb{R}} |f_{nh} - S_{h/\sqrt{2}} * dF_n| = \int_{\mathbb{R}} S_{h/\sqrt{2}} * dF_n - f_{nh}$$
$$= 1 - \|f_{nh}\|_1 = h^2\left\|\{(f_{nh})^{1/2}\}'\right\|_2^2,$$

the last inequality by (1.5). This proves the most important part of Theorem 2.8. For the remaining lower bound we observe from (4.4) that

$$-\tfrac{1}{4}h^2(u^2)'' + u^2 = \tfrac{1}{2}dF_n + \tfrac{1}{2}(1 - V)u^2,$$

where $V = h^2(u'/u)^2$. Then, with $S_{h/2}$ being Green's function this time, we get that

$$u^2 = \tfrac{1}{2}S_{h/2} * \{dF_n + (1 - V)u^2\}. \tag{4.7}$$

In view of (4.3) we have that $V \le 1$, so that

$$u^2 \ge \tfrac{1}{2}S_{h/2} * dF_n. \tag{4.8}$$

This is the lower bound of Theorem 2.8.

5. Proof of the asymptotic bias theorem. By Theorem 3.1, solving the large sample asymptotic problem (1.8) is equivalent to solving the boundary value problem

$$\begin{aligned} -h^2u'' + u &= \frac{f_o}{u}, \qquad x \in \mathbb{R}, \\ u(x) &\to 0, \qquad x \to \pm\infty, \end{aligned} \tag{5.1}$$

Let $w = (f_o)^{1/2}$. Note that $(f_o/u) - w = (w - u)(w/u)$. With assumption (2.3) one obtains the differential equation for $u - w$,

$$-h^2(u - w)'' + (u - w)\left(1 + \frac{w}{u}\right) = h^2w''. \tag{5.2}$$

Now multiply by $u - w$, and observe that integration by parts gives

$$\int_{\mathbb{R}}(u - w)''(u - w) = -\int_{\mathbb{R}}|(u - w)'|^2$$

and conclude that

$$h^2\|u' - w'\|_2^2 + \int_{\mathbb{R}}(u - w)^2\left(1 + \frac{w}{u}\right) = h^2\int_{\mathbb{R}}w''(u - w). \tag{5.3}$$

Now, the right-hand side is bounded by $h^2\|w''\|_2\|u - w\|_2$. Upon ignoring the term involving w/u on the left of (5.3), we get the nice inequality

$$h^2\|u' - w'\|_2^2 + \|u - w\|_2^2 \le h^2\|w''\|_2\|u - w\|_2. \tag{5.4}$$

Now ignoring the first term on the left yields

$$\|u - w\|_2 \le h^2\|w''\|_2. \tag{5.5}$$

This also shows that the right-hand side of (5.4) is dominated by $h^4\|w''\|_2^2$, and proves Theorem 3.5(b). To prove (a), we note that $L_h(u^h, F_o) \le L_h(w, F_o)$. This may be rewritten as

$$D\left((u^h)^2, f_o\right) + h^2\|(u^h)'\|_2^2 \le h^2\|w'\|_2^2,$$

where $D(\varphi, \psi) = \int_{\mathbb{R}}\varphi\log(\varphi/\psi) + \psi - \varphi$ is the Kullback–Leibler divergence. Since $D(\varphi, \psi) \ge 0$, the conclusion $\|(u^h)'\|_2 \le \|w'\|_2$ follows. □

6. Proof of the variance theorem. For the proof of Theorem 3.6 we try to repeat the material from Section 5. We start with the boundary value problems (5.1) and (4.1). Subtraction yields the differential equation

$$-h^2\left(u_{nh} - u^h\right)'' + \left(u_{nh} - u^h\right) = \frac{dF_n}{u_{nh}} - \frac{dF_o}{u^h}, \qquad x \in \mathbb{R}. \tag{6.1}$$

Upon multiplication by $u_{nh} - u^h$ and integration over $\mathbb{R}$, with integration by parts on the first term, we get

$$h^2\left\|\left(u_{nh} - u^h\right)'\right\|_2^2 + \|u_{nh} - u^h\|_2^2 = \int_{\mathbb{R}}dm_{n,h}, \tag{6.2}$$

with

$$dm_{n,h} = \left\{ \frac{dF_n}{u_{nh}} - \frac{dF_o}{u^h} \right\} (u_{nh} - u^h).$$

This may be rewritten as

$$dm_{n,h} = -\frac{(u_{nh} - u^h)^2}{u_{nh} u^h}\, dF_n + \frac{u_{nh} - u^h}{u^h} (dF_n - dF_o),$$

and thus we get the inequality

$$dm_{n,h} \le \frac{u_{nh} - u^h}{u^h} (dF_n - dF_o). \tag{6.3}$$

Alternatively, by interchanging the roles of F_o and F_n, we get

$$dm_{n,h} = -\frac{(u_{uh} - u^h)^2}{u_{nh} u^h} dF_o + \frac{u_{nh} - u^h}{u_{nh}} (dF_n - dF_o)$$

and so

$$dm_{n,h} \le \frac{u_{nh} - u^h}{u_{nh}} (dF_n - dF_o). \tag{6.4}$$

Now these two inequalities may be combined into one inequality as follows. Split the integration range in (6.2) into $\{u_{nh} \le u^h\}$ and $\{u_{nh} > u^h\}$. On $\{u_{nh} \le u^h\}$ use (6.3) and on $\{u_{nh} > u^h\}$ use (6.4). Then

$$\int_{\mathbb{R}} dm_{n,h} \le \int_{\mathbb{R}} \vartheta (dF_n - dF_o), \tag{6.5}$$

with

$$\vartheta = \frac{u_{nh} - u^h}{u_{nh} \vee u^h}. \tag{6.6}$$

Here $a \vee b$ is the maximum of a and b. Klonias (1982) uses (6.3) over the whole integration range, but this leads to trouble later on.

The next step in analyzing (6.2)–(6.6) is a well-known integration by parts trick [or a reproducing kernel Hilbert space trick in the setting of Klonias (1982)]. Let $T_\lambda(x)$ be the one-sided exponential distribution (2.6). The integration by parts trick is as follows: for smooth functions φ and distributions Ψ which decay fast enough at $\pm\infty$,

$$\int_{\mathbb{R}} \varphi(x)\, d\Psi(x) = \int_{\mathbb{R}} (\lambda \varphi'(x) + \varphi(x))(T_\lambda * d\Psi(x))\, dx. \tag{6.7}$$

One would expect to use $\lambda = h$, but it turns out that $\lambda = h/2$ is what is needed. We just keep writing λ though. Applying (6.7) to the integral at hand

gives (the functions in question decay fast enough)

$$\int_{\mathbb{R}} \vartheta\,(dF_n - dF_o) = \int_{\mathbb{R}} (\lambda\vartheta' + \vartheta)(T_\lambda * (dF_n - dF_o)). \tag{6.8}$$

Now we have that

$$\left(u_{nh} \vee u^h\right)\vartheta' = \left(u_{nh} - u^h\right)' - \frac{\left(u_{nh} \vee u^h\right)'}{u_{nh} \vee u^h}\left(u_{nh} - u^h\right) \tag{6.9}$$

and, except possibly on sets of measure 0 where $u_{nh}(x) = u^h(x)$ [open intervals on which $u_{nh}(x) = u^h(x)$ cause no problem],

$$\frac{\left(u_{nh} \vee u^h\right)'}{u_{nh} \vee u^h} = \begin{cases} \dfrac{(u_{nh})'}{u_{nh}}, & \text{if } u_{nh} \geq u^h, \\ \dfrac{(u^h)'}{u^h}, & \text{if } u^h \geq u_{nh}. \end{cases}$$

Thus (4.3) and its analogue for u^h imply

$$\left|\frac{\left(u_{nh} \vee u^h\right)'}{u_{nh} \vee u^h}\right| \leq h^{-1}. \tag{6.10}$$

Consequently, the right-hand side of (6.8) is bounded by

$$\begin{aligned} &\int_{\mathbb{R}} \left\{\lambda\left|\left(u_{nh} - u^h\right)'\right| + (1 + \lambda h^{-1})\left|u_{nh} - u^h\right|\right\} \frac{|T_\lambda * (dF_n - dF_o)|}{u_{nh} \vee u^h} \\ &\qquad \leq \left\{\lambda\left\|\left(u_{nh} - u^h\right)'\right\|_2 + (1 + \lambda h^{-1})\left\|u_{nh} - u^h\right\|_2\right\}\mathfrak{F}, \end{aligned} \tag{6.11}$$

where

$$\mathfrak{F}^2 = \int_{\mathbb{R}} \frac{|T_\lambda * (dF_n - dF_o)|^2}{\left(u_{nh} \vee u^h\right)^2}. \tag{6.12}$$

At this point we need the lower bound from Theorem (2.8), as well as its analogue for u^h. Since $S_\lambda(x) \geq \frac{1}{2}T_\lambda(x)$, with $\lambda = h/2$ and Theorem (2.8) we get that $(u_{nh})^2 \geq \frac{1}{2}S_\lambda * dF_n \geq \frac{1}{4}T_\lambda * dF_n$, and likewise for u^h. Thus

$$\frac{(T_\lambda * dF_n) \vee (T_\lambda * dF_o)}{\left(u_{nh} \vee u^h\right)^2} \leq 4. \tag{6.13}$$

Then it is easy to see that

$$\mathfrak{F}^2 \leq 4\int_{\mathbb{R}} \frac{|T_\lambda * (dF_n - dF_o)|^2}{(T_\lambda * dF_n) \vee (T_\lambda * dF_o)} \leq 16H_{n\lambda},$$

in which $H_{n\lambda}$ is a Hellinger distance, defined by

$$H_{n\lambda} = \left\|(T_\lambda * dF_n)^{1/2} - (T_\lambda * dF_o)^{1/2}\right\|_2^2.$$

Now going back to (6.11) with this bound for $\mathfrak{F}$, and then back to (6.2) gives the inequality for $e_{nh} = u_{nh} - u^h$,

$$h^2\|e'_{nh}\|_2^2 + \|e_{nh}\|_2^2 \le 4\{\tfrac{1}{2}h\|e'_{nh}\|_2 + \tfrac{3}{2}\|e_{nh}\|_2\}(H_{n\lambda})^{1/2}.$$

Since for all a, b we have $a + 3b \le \sqrt{10(a^2 + b^2)}$, it follows that

$$h^2\|e'_{nh}\|_2^2 + \|e_{nh}\|_2^2 \le 40H_{n\lambda}.$$

This proves Theorem 3.6. □

7. Almost sure bounds for the Hellinger distance. In this section we prove the statements of Theorem 2.10 regarding the Hellinger distance. We begin with $\mathrm{H}(A_h * dF_n, A_h * dF_o)$ with $A_h(x) = h^{-1}A(h^{-1}x)$ for arbitrary pdf A with a finite exponential moment,

$$\int_{\mathbb{R}} A(x)e^{r|x|}\,dx < \infty \quad \text{for some } r > 0. \tag{7.1}$$

We cannot really handle $\mathrm{H}(A_h * dF_n, A_h * dF_o)$, other than by using the inequality [see Devroye and Györfi (1985)],

$$\mathrm{H}(A_h * dF_n, A_h * dF_o) \le D(A_h * dF_n, A_h * dF_o) \tag{7.2}$$

and using the following result.

THEOREM 7.3 [Eggermont and LaRiccia (1999)]. *Assume that $h \asymp n^{-\beta}$ for some $0 < \beta < 1$.* (a) *If f_o has a finite moment of order $> \kappa > 2$ then for all $s > 1$,*

$$D(A_h * dF_n, A_h * dF_o) =_{\mathrm{as}} \mathscr{O}\big((nh)^{-\kappa/(\kappa+1)}(\log n)^s\big).$$

(b) *If f_o has a finite exponential moment, then for every $s > 1$,*

$$D(A_h * dF_n, A_h * dF_o) =_{\mathrm{as}} \mathscr{O}\big((nh)^{-1}\log(nh)^{-1}(\log n)^s\big).$$

PROOF OF THEOREM 2.10. (b) With $w = (f_o)^{1/2}$ the triangle inequality gives

$$\mathrm{H}(f_{nh}, f_o) \le 2\|u_{nh} - u^h\|_2^2 + 2\|u^h - w\|_2^2.$$

Then by Theorems 3.5 and 3.6 for appropriate constants c,

$$\mathrm{H}(f_{nh}, f_o) \le c\mathrm{H}\big(T_{h/2} * dF_n, T_{h/2} * dF_o\big) + ch^4\|w''\|_2^2, \tag{7.4}$$

and it follows that

$$\mathrm{H}(f_{nh}, f_o) \le cD\big(T_{h/2} * dF_n, T_{h/2} * dF_o\big) + ch^4\|w''\|_2^2. \tag{7.5}$$

Now from Theorem 7.3(a), if f_o has a finite moment of order $> \kappa > 2$ then for all $s > 1$,

$$\mathrm{H}(f_{nh}, f_o) =_{\mathrm{as}} \mathscr{O}\big((nh)^{-\kappa/(\kappa+1)}(\log n)^s + h^4\big),$$

and the asymptotic choice $h \asymp n^{-\kappa/(5\kappa+4)}$ gives the requisite bound.

Part (c) is proved similarly, from (7.5) and Theorem 7.3(b). Part (a) is proved starting from (7.4) by

$$\mathrm{H}(f_{nh}, f_o) \le c\|T_{h/2} * dF_n - T_{h/2} * dF_o\|_1 + ch^4\|w''\|_2^2,$$

and using the well-known bound for the L^1 error; see Theorem (8.1) below, due to Devroye (1991). □

8. Wrapping it all up: the proof of the main theorem. To prove the main theorem we use the following result, a direct consequence of Devroye (1991), section 4.3. See also section 9.1 in Devroye, Györfi and Lugosi (1996). The very last statement of the theorem is of course well known.

THEOREM 8.1 [Devroye (1991)]. *Let A be a pdf, with finite moments of all orders. For h deterministically varying with n,*

$$\|A_h * (dF_n - dF_o)\|_1 - \mathbb{E}\big[\|A_h * (dF_n - dF_o)\|_1\big] =_{\text{as}} \mathscr{O}\big(n^{-1/2}(\log n)^{1/2}\big).$$

Moreover, $\mathbb{E}[\|A_h * (dF_n - dF_o)\|_1] = \mathscr{O}((nh)^{-1/2})$, provided f_o has a moment of order > 1.

PROOF OF THEOREM 2.10. Theorem 3.6 and inequality (3.7) imply that

$$h^2\|(u_{nh})' - (u^h)'\|_2^2 \le c\|T_{h/2} * dF_n - T_{h/2} * dF_o\|_1,$$

so that with Theorem 8.1,

$$h^2\|(u_{nh})' - (u^h)'\|_2^2 =_{\text{as}} \mathscr{O}\big((nh)^{-1/2} + n^{-1/2}(\log n)^{1/2}\big).$$

It follows that for deterministic $h \asymp n^{-1/5}$,

$$\|(u_{nh})' - (u^h)'\|_2 =_{\text{as}} \mathscr{O}(n^{-1/20}),$$

whence

$$\|(u_{nh})'\|_2 =_{\text{as}} \|(u^h)'\|_2 + o(1) \le_{\text{as}} \|w'\|_2 + o(1), \tag{8.2}$$

where we used Theorem 3.5(a). □

PROOF OF MAIN THEOREM 2.7. From the triangle inequality,

$$\begin{aligned}\|f_{nh} - f_o\|_1 \le{}& \|f_{nh} - S_{h/\sqrt{2}} * dF_n\|_1 \\ &+ \|S_{h/\sqrt{2}} * (dF_n - dF_o)\|_1 + \|f_o - S_{h/\sqrt{2}} * f_o\|_1.\end{aligned} \tag{8.3}$$

The Green's function property of S_h implies that $f_o - S_{h/\sqrt{2}} * f_o = -\frac{1}{2}h^2 S_{h/\sqrt{2}} * (f_o)''$, and hence

$$\|f_o - S_{h/\sqrt{2}} * f_o\|_1 \le \tfrac{1}{2}h^2\|f_o''\|_1. \tag{8.3}$$

Together with Theorems 2.8 and 2.9 we then get, for an appropriate constant c,

$$\|f_{nh} - f_o\|_1 \le_{\text{as}} ch^2\|w'\|_2^2 + \|S_{h/\sqrt{2}} * (dF_n - dF_o)\|_1 + ch^2\|(f_o)''\|_1. \tag{8.4}$$

Now applying Theorem 8.1 gives

$$\|f_{nh} - f_o\|_1 =_{\text{as}} \mathcal{O}\big(h^2 + (nh)^{-1/2} + n^{-1/2}(\log n)^{1/2} + h^2\big),$$

and with $h \asymp n^{-1/5}$ the main theorem follows. □

9. Universal consistency.

PROOF OF THEOREM 2.11. We recall that u_{nh} solves (1.7), and u^h is the solution to (1.8). In view of Theorem 3.6 and the bound (3.7), Theorem 8.1 implies that the variance part of the error tends to 0, that is,

$$\|u_{nh} - u^h\|_2 \to 0 \quad \text{provided } nh \to \infty.$$

It thus suffices to consider the bias term $\|u^h - (f_o)^{1/2}\|_2$. The way we go about this is by smoothing f_o and applying the results for smooth f_o.

Let $\lambda > 0$ to be chosen later, and set $f_\lambda = (S_\lambda * S_\lambda * f_o)^{1/2}$. Then $f_\lambda \in L^2(\mathbb{R})$, and

$$(f_\lambda)'' = \frac{(S_\lambda * S_\lambda * f_o)''}{2f_\lambda} - \frac{|(S_\lambda * S_\lambda * f_o)'|^2}{4(S_\lambda * S_\lambda * f_o)^{3/2}}.$$

Now the argument leading to (4.3) shows

$$\begin{aligned}
|(S_\lambda * S_\lambda * f_o)'(x)| &\le \lambda^{-1} S_\lambda * S_\lambda * f_o(x),\\
|(S_\lambda * S_\lambda * f_o)''(x)| &\le \lambda^{-2} S_\lambda * S_\lambda * f_o(x),
\end{aligned}$$

and thus

$$|(f_\lambda)''(x)| \le \tfrac{3}{4}\lambda^{-2} f_\lambda(x).$$

It follows that

$$\|(f_\lambda)''\|_2 \le \tfrac{3}{4}\lambda^{-2}, \tag{9.1}$$

and thus $f_\lambda \in W^{2,2}(\mathbb{R})$. Now let $\vartheta_{h,\lambda}$ be the minimizer of $L_h(\vartheta, S_\lambda * S_\lambda * F_o)$, that is, $\vartheta_{h,\lambda}$ is the solution to the large sample asymptotic problem (1.8) with the true density f_o replaced by $S_\lambda * S_\lambda * f_o$. Now it is rather surprising that the Asymptotic Variance Theorem 3.6 applies to yield

$$\begin{aligned}
&\|u^h - \vartheta_{h,\lambda}\|_2^2 + h^2\|\big(u^h - \vartheta_{h,\lambda}\big)'\|_2^2\\
&\qquad \le cH\big(T_{h/2} * f_o, T_{h/2} * S_\lambda * S_\lambda * f_o\big) \le c\|f_o - S_\lambda * S_\lambda * f_o\|_1
\end{aligned}$$

and this last expression tends to 0 as $\lambda \to 0$. Thus

$$\|u^h - \vartheta_{h,\lambda}\|_2 \to 0 \quad \text{for } \lambda \to 0, \text{ uniformly in } h > 0. \tag{9.2}$$

Finally, we apply the Asymptotic Bias Theorem 3.5 and obtain

$$\|\vartheta_{h,\lambda} - f_\lambda\|_2 \le h^2\big\|\{(f_\lambda)''\big\|_2 \le h^2\lambda^{-2},$$

the last inequality by (9.1). Thus, setting $\lambda = h^{1/2}$ does the trick:

$$\|u^h - (f_o)^{1/2}\|_2 \le \|u^h - \vartheta_{h,\lambda}\|_2 + \|\vartheta_{h,\lambda} - (S_\lambda * S_\lambda * f_o)^{1/2}\|_2 + \|(S_\lambda * S_\lambda * f_o)^{1/2} - (f_o)^{1/2}\|_2,$$

where the last term may be bounded from above by $\|S_\lambda * S_\lambda * f_o - f_o\|_1$, and for $h \to 0$ and $\lambda = h^{1/2}$, each term tends to 0. □

10. Some simulations. In this section we offer some simulation experiments comparing kernel estimators with the Good estimator. Actually, since the Good estimator f_{nh} is a sub pdf, in the simulations to follow we replace it by

$$\varphi^{nh} = f_{nh}/p, \tag{10.1}$$

with $p = \int_{\mathbb{R}} f_{nh}(x)\,dx$. It is well known that this improves the error, that is,

$$\|\varphi^{nh} - f_o\|_1 \le \|f_{nh} - f_o\|_1. \tag{10.2}$$

We consider the kernel estimators with the Epanechnikov kernel and the two-sided exponential kernel. The criterion by which the estimators are judged is the smallness of the L^1 error, using the optimal smoothing parameter (with respect to L^1 error).

Let $\phi(x) = (2\pi)^{-1/2}\exp(-y^2/2)$ denote the standard normal density with mean 0 and variance 1, and set $\phi_\sigma(y) = \sigma^{-1}\phi(\sigma^{-1}y)$. It is also useful to recall the beta density with parameters α and β,

$$\psi_{\alpha,\beta}(x) = B(\alpha,\beta)(x_+)^{\alpha-1}((1-x)_+)^{\beta-1}.$$

The simulations involve the following densities:

nice mix $\quad f_o(y) = \frac{9}{10}\phi_{1/2}(y-5) + \frac{1}{10}\phi_{1/2}(y-7),$

normal $\quad f_o(y) = \phi(y-5),$

uniform $\quad f_o(x) = \frac{1}{5}\mathbb{I}(3 < x < 8),$

beta $\quad f_o(x) = \frac{1}{5}\psi_{\alpha,\beta}\big(\frac{1}{5}(x-3)\big)$ with $\alpha = 1.4,\ \beta = 2.6,$

bad mix $\quad f_o(x) = \frac{1}{5}\phi_{9/5}(x-6) + \frac{4}{5}\phi_{1/10}(x-2),$

bimodal $\quad f_o(x) = \frac{1}{2}\phi_{1/2}(x-3.5) + \frac{1}{2}\phi_{1/2}(x-6.5).$

Now to the actual simulation setup. For each density, random samples of size 100 were generated and the optimal smoothing parameter H determined for each method, that is, $h = H$ was chosen so as to realize $\inf_h \|f_{nh} - f_o\|_1$ for each estimator f_{nh} under consideration. The resulting estimator is denoted by $f^{n,\mathrm{OPT}}$. In all cases this was replicated 500 times, and the (sample) means and standard deviations of the error $\|f^{n,\mathrm{OPT}} - f_o\|_1$ were computed.

The results are tabulated in Table 1. The two-sided exponential kernel does not give good results. It is thus surprising, in view of the comparison theorem (2.8), that the Good estimator is really a remarkably good estimator: on average it gives the same optimal L^1 errors as the Epanechnikov kernels.

TABLE 1
Estimated means and standard deviations of $\|f_{n,OPT} - f_o\|_1$ *for various kernels and the Good estimator applied to various densities, for sample size* 100, *based on* 500 *replications*

	nice mix	**normal**	**uniform**	**beta**	**bad mix**	**bimodal**
good	0.153(0.046)	0.121(0.044)	0.213(0.040)	0.163(0.043)	0.298(0.042)	0.185(0.048)
epan	0.155(0.052)	0.129(0.051)	0.216(0.039)	0.164(0.042)	0.336(0.054)	0.185(0.051)
exp	0.172(0.049)	0.145(0.047)	0.226(0.037)	0.179(0.041)	0.345(0.051)	0.203(0.049)

These observations also apply to larger sample sizes (up to 1000). It remains of course to be seen if the smoothing parameter for the Good estimator can be chosen as effectively as for kernel estimators.

Acknowledgment. The authors thank the referee and Associate Editor for drawing their attention to the Universal Consistency Theorem 2.8 and for pointing out the poor exposition of part of Section 6 in an earlier version of the paper and the omission of the scaling (10.1) and (10.2).

REFERENCES

ADAMS, R. A. (1975). *Sobolev Spaces*. Academic Press, New York.

AKAIKE, H. (1954). An approximation to the density function. *Ann. Inst. Statist. Math.* **6** 127–132.

COX, D. D. and O'SULLIVAN, F. (1990). Asymptotic analysis of penalized likelihood and related estimators. *Ann. Statist.* **18** 1676–1695.

DE MONTRICHER, G. F., TAPIA, R. A. and THOMPSON, J. R. (1975). Nonparametric maximum likelihood estimation of probability densities by penalty function methods. *Ann. Statist.* **3** 1329–1348.

DEVROYE, L. (1991). Exponential inequalities in nonparametric estimation. In *Nonparametric Functional Estimation and Related Topics* (G. Roussas, ed.) 31–44. Kluwer, Dordrecht.

DEVROYE, L. AND L. GYÖRFI (1985). *Density Estimation*: *The* L_1*-View*. Wiley, New York.

DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, New York.

EGGERMONT, P. P. B. and LARICCIA, V. N. (1997). Nonlinearly smoothed EM density estimation with automated smoothing parameter selection for nonparametric deconvolution problems. *J. Amer. Statist. Assoc.* **92** 1451–1458.

EGGERMONT, P. P. B. and LARICCIA, V. N. (1999). Best asymptotic normality of the kernel density entropy estimator for smooth densities. *IEEE Trans. Inform. Theory* **45** 1321–1326.

GOOD, I. J. (1971). A nonparametric roughness penalty for probability densities. *Nature* **229** 29–30.

GOOD, I. J. and GASKINS, R. A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika* **58** 255–277.

HALL, P. (1983). Large sample optimality of least squares cross validation in density estimation. *Ann. Statist.* **11** 1156–1174.

KLONIAS, V. K. (1982). Consistency of two nonparametric maximum penalized likelihood estimators of the probability density function. *Ann. Statist.* **10** 811–824.

KLONIAS, V. K. (1984). On a class of nonparametric density and regression estimators. *Ann. Statist.* **12** 1263–1284.

PARZEN, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33** 1065–1076.

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27** 832–835.

RUDEMO, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* **9** 65–78.

SILVERMAN, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10** 795–810.

STONE, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.* **12** 1285–1297.

THOMPSON, J. R. and TAPIA, R. A. (1990). *Nonparametric Function Estimation, Modeling, and Simulation*. SIAM, Philadelphia.

WAHBA, G. (1981). Data-based optimal smoothing of orthogonal series density estimates. *Ann. Statist.* **9** 146–156.

DEPARTMENT OF MATHEMATICAL SCIENCES
UNIVERSITY OF DELAWARE
NEWARK, DELAWARE 19716
E-MAIL: eggermon@math.udel.edu
lariccia@math.udel.edu