The Annals of Statistics
2000, Vol. 28, No. 3, 922–947

MAXIMUM LIKELIHOOD ESTIMATION OF SMOOTH MONOTONE AND UNIMODAL DENSITIES

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

University of Delaware

We study the nonparametric estimation of univariate monotone and unimodal densities using the maximum smoothed likelihood approach. The monotone estimator is the derivative of the least concave majorant of the distribution corresponding to a kernel estimator. We prove that the mapping on distributions Φ with density φ,

$$\varphi \longmapsto \text{the derivative of the least concave majorant of } \Phi,$$

is a contraction in all L^P norms ($1 \le p \le \infty$), and some other "distances" such as the Hellinger and Kullback–Leibler distances. The contractivity implies error bounds for monotone density estimation. Almost the same error bounds hold for unimodal estimation.

1. Introduction. We investigate the nonparametric estimation of monotone and unimodal densities from the maximum smoothed likelihood point of view. The estimation of monotone densities has a long history, dating back to Grenander (1956), and has many applications; see Barlow, Bartholomew, Bremner and Brunk (1972). It is one of only a few instances of density estimation problems where nonparametric maximum likelihood estimation works without smoothing or roughness penalization. On the down side, for smooth densities these estimators do not achieve the usual L^1 convergence rates. In this paper we set out to repair this deficiency by using a smoothed version of the maximum likelihood procedure, which gives the usual convergence rates of kernel density estimators for both nonsmooth and smooth densities. The natural extension to unimodal density estimation is considered also.

Let $X_1, X_2, \ldots, X_n$ be nonnegative iid random variables with common probability density function (pdf) f_o, assumed to be monotone on $(0, \infty)$. Let F_n denote the empirical distribution function of the X_i. The maximum likelihood estimator f^n of f_o, is the (unique) solution to

$$\begin{aligned} &\text{minimize} && -\int_0^\infty \log f(x)\, dF_n(x), \\ &\text{subject to} && f \text{ is a monotone pdf on } (0, \infty). \end{aligned} \tag{1.1}$$

Received October 1998; revised March 2000.
AMS 1991 *subject classification*. 62G07.
Key words and phrases. Maximum likelihood estimation, monotone and unimodal densities, least concave majorants, contractions, L^1 error bounds.

The pdf constraint in (1.1) is somewhat of a pain, but an old trick of Silverman (1982) comes to the rescue: problem (1.1) has the same solution as

$$\begin{aligned}(1.2)\qquad &\text{minimize} \quad -\int_0^\infty \log f(x)\, dF_n(x) + \int_0^\infty f(x)dx,\\ &\text{subject to} \quad f \in L^1(0,\infty),\ f \geq 0, f \text{ monotone on } (0,\infty).\end{aligned}$$

The estimator f^n is characterized in terms of its distribution function, denoted by F^n; F^n is the least concave majorant of the empirical distribution F_n. This is due to Grenander (1956), whence f^n is usually designated as the Grenander estimator. It is well known that f^n is a step function, with jumps at (some of) the order statistics. This makes sense for an arbitrary monotone density, but less so for a *continuous* monotone density. Birgé (1989) shows that if f_o has compact support $[0, B]$ and $(f_o)'$ is continuous, then

$$(1.3)\qquad \mathbb{E}[\|f^n - f_o\|_1] \leq c_n \lambda(f_o) n^{-1/3},$$

where $c_n \longrightarrow 3$, for $n \to \infty$, and

$$(1.4)\qquad \lambda(f) = \int_0^\infty |\tfrac{1}{2} f(x) f'(x)|^{1/3} dx.$$

See also Birgé (1987a, b). Prior to this, Groeneboom (1985) had indicated that if in addition $(f_o)' < 0$ everywhere and $(f_o)''$ is bounded, then

$$(1.5)\qquad n^{1/6}\big\{n^{1/3}\|f^n - f_o\|_1 - c\lambda(f_o)\big\} \longrightarrow_d Y \sim N(0,\sigma^2),$$

for known c ($c \approx 0.82$) and σ ($\sigma^2 \approx 0.17$). For the actual proof see Groeneboom, Hooghiemstra and Lopuhaä (1999). So in this case, the convergence rate is $n^{-1/3}$, and the extra smoothness implies asymptotic normality, but not better rates.

The remedy proposed here is to start with any "good" estimator for smooth densities, and use its distribution function in (1.2), instead of F_n. For definiteness, say we take a boundary kernel estimator $\mathscr{A}_h\, dF_n$ depending on a smoothing paramater h,

$$(1.6)\qquad \mathscr{A}_h\, dF_n(x) = \int_0^\infty a_h(x,y)\, dF_n(y), \qquad x \geq 0,$$

with $a_h(x, y)$ nonnegative; see, for example, Devroye (1987) or Jones (1993). The choice

$$(1.7)\qquad a_h(x,y) = h^{-1}\big\{A\big(h^{-1}(x-y)\big) + A\big(h^{-1}(x+y)\big)\big\}, \qquad x, y > 0,$$

with A a nonnegative kernel on the line (symmetric, finite moments of all orders) will work, though not optimally; see Hall and Wehrly (1991), Müller (1993) and references therein. Then

$$(1.8)\qquad \|\mathscr{A}_h\, dF_n - f_o\|_1 =_{\text{as}} \mathscr{O}(n^{-2/5}), \qquad n \to \infty,$$

for $h \asymp n^{-1/5}$, provided f_o satisfies the usual nonparametric assumptions $[(f_o)'' \in L^1(0,\infty),$ and $\mathbb{E}[X_1^\kappa] < \infty$ for some $\kappa > 1]$; see, for example, Devroye (1991), Devroye, Györfi and Lugosi (1996). Here $=_{\text{as}}$ denotes almost sure equality, and likewise for $\le_{\text{as}}$, $\ge_{\text{as}}$. Moreover, one's favorite procedure for selecting the smoothing parameter will work here as well.

We assume that $\mathscr{A}_h\, dF_n$ is continuous and nonnegative. The maximum *smoothed* likelihood estimator of f_o is now given as the solution to

$$\begin{aligned} &\text{minimize} \quad -\int_0^\infty \mathscr{A}_h\, dF_n(x)\ \log f(x) d(x) + \int_0^\infty f(x)\, dx, \\ &\text{subject to} \quad f \in L^1(0,\infty),\ f \ge 0,\ f \text{ monotone on } (0,\infty). \end{aligned} \tag{1.9}$$

The solution of (1.9) is denoted by f^{nh}. Note that without the monotonicity constraint the solution of (1.9) would be $f = \mathscr{A}_h\, dF_n$; see Eggermont and LaRiccia (1995). This interpretation of the kernel estimator is at least part of the motivation for the study of (1.9). Analogous to the nonsmoothed case (1.2), it is well known that the solution of (1.9) is such that its corresponding distribution is the least concave majorant of the distribution corresponding to $\mathscr{A}_h\, dF_n$; see, for example, Mammen (1991), Bickel and Fan (1996).

We now come to the crux of the matter. First some notation. For Φ a distribution function that is, Φ nonnegative, increasing and bounded on $(0,\infty)$, we let

$$\text{LCM}(\Phi) \equiv \text{the least concave majorant of } \Phi, \tag{1.10}$$

and if Φ has density φ then

$$\text{lcm}(\varphi) = \begin{cases} \text{the left continuous function which is} \\ \text{equal to the derivative of LCM}(\Phi) \text{ a.e.} \end{cases} \tag{1.11}$$

The operation LCM has some remarkable contractivity properties, the first one of which hinted at by Marshall (1970), that is, for distributions Φ, Ψ,

$$\|\text{LCM}(\Phi) - \text{LCM}(\Psi)\|_\infty \le \|\Phi - \Psi\|_\infty. \tag{1.12}$$

Here $\|\cdot\|_\infty$ denotes the sup norm. However, it holds for many other distances, the most useful of which is the total variation norm, which is just the L^1 norm on the corresponding densities (if they exist). If the distributions Φ and Ψ have densities φ and ψ (with respect to Lebesgue measure), then

$$\|\text{lcm}(\varphi) - \text{lcm}(\psi)\|_1 \le \|\varphi - \psi\|_1. \tag{1.13}$$

An obvious and important implication is that

$$\|f^{nh} - f_o\|_1 \le \|\mathscr{A}_h\, dF_n - f_o\|_1, \tag{1.14}$$

so that the monotone estimator is at least as accurate as the kernel estimator. In particular, if (1.8) holds then

$$\|f^{nh} - f_o\|_1 =_{\text{as}} \mathscr{O}(n^{-2/5}), \qquad n \to \infty. \tag{1.15}$$

It is interesting in its own right that the analogue of (1.13) holds for all L^p norms ($1 \le p \le \infty$), as well as the Hellinger, Kullback–Leibler, and Pearson's φ^2 distances. The proofs of all these are based on the pool-adjacent-violators-algorithm for simple densities (step functions), followed by the usual limiting argument. We note that Fougères (1997) observes these properties for monotone rearrangements, referring to Lieb and Loss (1996). It should be noted that Brunk (1965) proves a general result which contains the contractivity in the L^2 norm and Kullback–Leibler distance as special cases. For a precise description, see Section 3.

The results concerning monotone density estimation extend to unimodal density estimation, *almost*. Let the univariate density f_o be unimodal with mode m_o, that is, f_o is increasing on $(-\infty, m_o)$ and decreasing on (m_o, ∞). Here we use increasing to mean nondecreasing, and decreasing to mean nonincreasing. Thus, the mode of a unimodal density need not be unique, but the set of all modes of a unimodal density is a closed interval. For later reference, the distribution corresponding to f_o is denoted by F_o. Now let $X_1, X_2, \dots, X_n$ be an iid random sample, with common density f_o, and let

$$A_h * dF_n(x) = \int_{\mathbb{R}} h^{-1} A\big(h^{-1}(x-y)\big)\, dF_n(y), \qquad x \in \mathbb{R}, \tag{1.16}$$

be a kernel density estimator of f_o, where A is a symmetric, continuous log-concave pdf, and h is the smoothing parameter. The log-concavity requirement is sensible since then $A_h * dF_o$ is unimodal whenever f_o is unimodal, by the celebrated result of Ibragimov (1956). Some examples of continuous log-concave kernels are the Epanechnikov kernel; see, for example, Devroye and Györfi (1985), and the Gaussian and the two-sided exponential densities.

The unimodal estimator f^{nh} is now defined as the solution to the maximum smoothed likelihood problem,

$$\begin{aligned} &\text{minimize} \quad -\int_{\mathbb{R}} A_h * dF_n(x) \log f(x) dx + \int_{\mathbb{R}} f(x)\, dx, \\ &\text{subject to} \quad f \in L^1(\mathbb{R}),\ f \ge 0,\ f \text{ unimodal}. \end{aligned} \tag{1.17}$$

Actually, the solution need not be unique, though in practice it usually is. The theoretical nonuniqueness is best illustrated by the case $n = 2$ with A having compact support, such as the Epanechnikov kernel. If h is small enough then (1.17) has two solutions, one with mode at X_1, the other with mode at X_2. The practical significance is that (1.17) may have many local minima. It also suggests that these local minimizers have their modes at (some of) the local maxima of $A_h * dF_n$, which is indeed the case.

Concerning error bounds, we show that for any of the solutions f^{nh} of (1.17), with mode m_{nh},

$$\|f^{nh} - f_o\|_1 \le \|A_h * dF_n - f_o\|_1 + c_{nh}|m_{nh} - m_o| \|A_h * (dF_n - dF_o)\|_\infty, \tag{1.18}$$

with $c_{nh} \longrightarrow_{\text{as}} \sqrt{32}$. Under the usual nonparametric assumptions and a very mild sharpening of the unimodality assumption this implies (roughly) that for $h \asymp n^{-\beta}$ $(0 < \beta < 1)$,

$$\|f^{nh} - f_o\|_1 \le_{\text{as}} \big(1 + o(1)\big) \|A_h * dF_n - f_o\|_1, \qquad n \to \infty. \tag{1.19}$$

It is clear from (1.18) that for estimating a unimodal density the mode must be estimated, but that it need not be very accurate. The optimal estimation of the mode is entirely different from estimating the density. Estimating the mode of f_o by the mode of $A_h * dF_n$ goes back all the way to Parzen (1962), who used kernel estimators, and Chernoff (1964), who used clustering ideas. Eddy (1980), improving on Chernoff (1964), shows asymptotic normality of this estimator of the mode when the third derivative of f_o is absolutely continuous, and $f_o''(m_o) < 0$ (so the mode is unique). The optimal smoothing parameter for mode estimation is $h \asymp n^{-1/7}$, as opposed to $h \asymp n^{-1/5}$ for density estimation; see Eddy (1980) and Grund and Hall (1995). Finally, it is not clear to the authors whether the bound (1.18) in terms of the sup norm of $A_h * (dF_n - dF_o)$ is sharp. As pointed out by a referee, it would seem that with a bit of work the sup norm over the line can be replaced by the sup norm over a small interval around m_o, but we shall not pursue this.

Finally, a comment on the choice of L^1 error to ascertain the effectiveness of the estimators. For the monotone density estimation problem this choice seems to be traditional, and to a lesser extent this applies to unimodal density estimation. For general density estimation, Devroye and Györfi (1985) make an eloquent case for the advantages of the use of the L^1 error, not the least one of which is its invariance under monotone transformations. The disadvantage is that what happens in the tails is largely ignored. If tail estimation were more important, then perhaps the Hellinger, Kullback–Leibler or Pearson's φ^2 distances, defined, respectively, by

$$H(\varphi, \psi) = \|\sqrt{\varphi} - \sqrt{\psi}\|_2^2, \tag{1.20}$$

$$D(\varphi, \psi) = \int_{\mathbb{R}} \left(\varphi(x) \log\left\{ \frac{\varphi(x)}{\psi(x)} \right\} - \psi(x) + \varphi(x) \right) dx, \tag{1.21}$$

$$P(\varphi, \psi) = \int_{\mathbb{R}} \frac{|\varphi(x) - \psi(x)|^2}{\psi(x)} dx \tag{1.22}$$

could be used. They, too, are invariant under monotone transformations, but are not easy to work with. Note that Kullback–Leibler distance could be $+\infty$. See Devroye and Györfi (1985). We should mention the general L^p norms $(p \geq 1)$,

$$\|\varphi - \psi\|_p^p = \int_{\mathbb{R}} |\varphi(x) - \psi(x)|^p \, dx, \tag{1.23}$$

with the supremum norm being the limiting case $p \to \infty$.

In Section 3 we formulate the assumptions and results alluded to above. The contractivity property (1.13) is stated in full generality, and proved in Section 4 and 5 via a detailed analysis of the pool-adjacent-violators algorithm and a limiting argument. In Section 6 we consider unimodal estimation, and prove (1.18). Some ridiculously slow convergence rates for the estimator of the mode are shown under minimal conditions, which are good enough to show essentially (1.19). In the next section we discuss some alternative estimators for unimodal densities and report on some simulations.

2. Comparisons with other unimodal estimators. In this section we discuss some alternative estimators of unimodal densities. Some experimental comparisons are presented at the end of this section, but the conclusions are briefly alluded to throughout. The first two concern alternative methods for the estimation of the mode.

(a) One can consider estimating the mode of f_o by the mode m of the kernel density estimator, and then estimate the unimodal density by solving (1.17) with the mode fixed at m. This amounts to two Grenander estimators, one to the left of m and one to the right. We denote the solution by $f^{nh}(\cdot, m)$. It appears that the resulting density estimator satisfies the same asymptotic and just about the same small sample behavior as the smoothed maximum likelihood estimator.

(b) One can also consider various minimum distance estimators of the mode. One choice is to use the solution to

$$\text{minimize } \|f^{nh}(\cdot\,; m) - A_h * dF_n\|_1 \quad \text{subject to } m \in \mathbb{R}. \tag{2.1}$$

The solution, denoted by ψ^{nh}, exists since the local minima occur again at some of the modes of $A_h * dF_n$, and (1.18) and (1.19) hold for ψ^{nh} also, by the same arguments under the same assumptions as for the smoothed maximum likelihood estimator. The accuracy of this estimator, too, is just about the same as for the maximum smoothed likelihood estimator.

(b) The approach of Fougères (1997) suggests replacing $f^{nh}(\cdot\,; m)$ by what may be called the unimodal rearrangement of $A_h * dF_n$, that is, the decreasing rearrangement of $A_h * dF_n(x)$, $x \geq m$, on (m, ∞) and the increasing rearrangement of $A_h * F_n(x)$, $x < m$, on $(-\infty, m)$. Here for nonnegative $\varphi \in L^1(m, \infty)$, the decreasing rearrangement of φ is defined to be the nonincreasing function φ^* on (m, ∞) for which the sets

$$\{x > m\colon \varphi(x) > \alpha\} \quad \text{and} \quad \{x > m\colon \varphi^*(x) > \alpha\}$$

have the same Lebesgue measure for all $\alpha > 0$; see Hardy, Littlewood, and Pólya (1951). The increasing rearrangement of an $L^1(-\infty, m)$ function on $(-\infty, m)$ is defined similarly. The resulting unimodal estimator is denoted as $\varphi^{nh}(\cdot\,; m)$. It seems to be both a strength and a weakness of this estimator that it is extremely insensitive to outliers: in its simplest form, when A is symmetric and has compact support, for $n = 1$ and $X_1 \gg m$, the unimodal rearrangement with mode m of $A_h(x - X_1)$, $x \in \mathbb{R}$, is given by

$$\varphi^{1,h}(x; m) = \begin{cases} A_h\big(\tfrac{1}{2}(x - m)\big), & x > m, \\ 0, & x < m, \end{cases}$$

and the outlier has disappeared without trace. As before, two natural choices for the mode are the mode of $A_h * dF_n$ and the solution of

$$\text{minimize } \|\varphi^{nh}(\cdot\,; m) - A_h * dF_n\|_1 \quad \text{subject to } m \in \mathbb{R}. \tag{2.2}$$

Simulations suggest that both of these estimators are quite similar to each other, but behave quite differently from the smoothed maximum likelihood estimator.

(d) Birgé (1997) considers any estimator $f^{nh}(\cdot\,;\mu)$ (as per the above) with distribution $F^{nh}(\cdot\,;\mu)$ for $\mu \neq X_i,\ i = 1, \dots, n$, which satisfies

$$\|F^{nh}(\cdot\,;\mu) - F_n\|_\infty \leq \min_m \|F^{nh}(\cdot\,;m) - F_n\|_\infty + \eta, \tag{2.3}$$

for $\eta = o(n^{-1/2})$, say. For this estimator Birgé (1997) shows bounds similar to (1.3). This method is not competitive for smooth unimodal densities.

(e) Bickel and Fan (1996) propose solving the maximum likelihood-like problem

$$\begin{aligned} &\text{minimize} \quad -\frac{1}{n-1}\sum_{\substack{i=1\\ i\neq j}}^{n} \log f(X_i) + \int_{\mathbb{R}} f(x)\,dx,\\ &\text{subject to} \quad f \in L^1(\mathbb{R}),\ f \geq 0,\ f \text{ unimodal, with mode at } X_j, \end{aligned} \tag{2.4}$$

where the minimization is also over $j = 1, 2, \dots, n$. They prove the pointwise consistency of the estimator, except at the mode m_o, analogous to the monotone case; see Wegman (1970) and Woodroofe and Sun (1993). However, the L^1 convergence rate is still $\mathscr{O}(n^{-1/3})$, and, indeed, the method is not competitive for smooth densities.

(f) Bickel and Fan (1996) also consider "grouping" the observations followed by maximum likelihood estimation. This is quite close to considering (1.17) with the kernel estimator $A_h * dF_n$ (for small h) approximated by a step function on a fine partition of $\mathbb{R}$. The resulting estimator f^n is $\mathscr{O}(n^{-1/3})$ accurate, so is not competitive for smooth unimodal densities. What is *very* competitive is its smoothed version $A_h * f^n$. This is analogous to results for isotone regression; see Mammen (1991).

On to the simulation experiments comparing the unimodal density estimators. Regarding computations, the authors find it expedient to replace the kernel estimator $A_h * dF_n$ by a step function on a fine grid, as follows. Let $\delta > 0$, and define the intervals

$$\omega_j = [\,j\delta, (j+1)\delta), \qquad j \in \mathbb{Z}. \tag{2.5}$$

Now approximate $A_h * dF_n$ by $B_h\,dF_n$, defined as

$$B_h\,dF_n(x) = \frac{1}{\delta}\int_{\omega_j} A_h * dF_n(y)\,dy, \qquad x \in \omega_j,\ j \in \mathbb{Z}. \tag{2.6}$$

For $h \ll \delta$ there is a quite a difference between $A_h * dF_n$ and its approximation, but for $h \gg \delta$ the difference is negligible. Thus, δ should be chosen to be smaller than "good" values of h. In practice, only a finite number of ω_j need consideration, depending possibly on the sample $X_1, X_2, \dots, X_n$. The computational advantage of the above approximation is then obvious: $B_h\,dF_n$ is a finite-dimensional object; it is always a pdf, and the complexity of its (straightforward) computation does not depend on the specific value of $h > 0$. If $h = 0$, then $B_h\,dF_n$ may be interpreted as a histogram estimator on the partition ω_j, $j \in \mathbb{Z}$. Finally, for unimodal estimation we may restrict the choice of modes to the points $j\delta$, and then, as is readily seen, $f^{nh}(\cdot\,; j\delta)$ and $\varphi^{nh}(\cdot\,; j\delta)$ are step functions on the partition (2.5) as well. The computations involve the determination of the mode according to the various criteria, which we implement by inspection of the local modes of $B_h dF_n$. For the estimators based on $f^{nh}(\cdot\,; m)$ this is proved later on; see Theorem (6.5). For the unimodal rearrangements we just assume it.

In the experiments we compare the various estimators to each other. The estimators are based on $f^{nh}(\cdot\,; m)$ and on $\varphi^{nh}(\cdot\,; m)$, which we denote by MLE and UR (unimodal rearrangement). The mode m may be chosen by maximum likelihood, minimum L^1 distance, or as the mode of $B_h\,dF_n$. We emphasize that all estimators are based on $B_h\,dF_n$ rather than $A_h * dF_n$. The specific methods considered are:

(i) MLE-MLE: f^{nh}, the solution of (1.11).
(ii) MODE-MLE: $f^{nh}(\cdot, \mu_{nh})$, with μ_{nh} the mode of $B_h\,dF_n$.
(iii) DIST-MLE: $f^{nh}(\cdot, \lambda_{nh})$, the solution of the minimum L^1 distance problem (2.1).
(iv) $A_h *$MLE, the smoothed version of the maximum likelihood estimator using grouped data, analogous to the regression case of Mammen (1991); see Section 2.6 above.
(v) MODE-UR: $\varphi^{nh}(\cdot, \mu_{nh})$, the unimodal rearrangement of $B_h\,dF_n$ with mode μ_{nh} as the mode of $B_h dF_n$.
(vi) DIST-UR: $\varphi^{nh}(\cdot\,; \eta_{nh})$, the solution of the minimum L^1 distance problem (2.2).

In the experiments, the simulated data correspond to various unimodal densities, restricted to the interval [0, 10]. Here $\phi(\cdot, \mu, \sigma)$ denotes the Gaussian density with mean μ and standard deviation σ, and $\psi(x; \alpha)$ is the density corresponding to the Weibull distribution $\Psi(x; \alpha) = 1-\exp(-x_+^\alpha)$, and $\theta(x; \alpha, \beta) = x_+^{\alpha-1}(1-x)_+^{\beta-1}/B(\alpha, \beta)$ the standard Beta density. The specific densities under consideration are:

(i) A normal density $\phi(\cdot\,; 5, 1)$.
(ii) The Weibull density $f_o(x) = \sigma^{-1}\psi(x - \gamma : \alpha)$; with $\alpha = 1.1$, $\gamma = 1$ and $\sigma = 0.5$.
(iii) The uniform density on $[3, 8]$.
(iv) The mixture of two Beta densities $\frac{9}{10}\theta_\sigma(x-\gamma; \alpha, \beta) + \frac{1}{10}\theta_s(x-c; a, b)$, with $\alpha = 2.1$, $\beta = 2.45$, $\gamma = 1$ and $\sigma = 1.5$, and $a = 1.0$, $b = 1.5$, $c = 1$ and $s = 8$. Here, $\theta_\sigma(x; \cdots) = \sigma^{-1}\theta(\sigma^{-1}x; \cdots)$.
(v) The mixture of two normals $\frac{4}{5}\phi(x; 5, .1) + \frac{1}{5}\phi(x; 5, 1.8)$.

TABLE 1
Estimated minimum L^1 errors, for various unimodal estimators of various densities, for sample size 100, *based on* 1000 *replications*[1]

	normal	Weibull	uniform	Beta mix	normal mix
opt. $A_h * dF_n$	0.126	0.250	0.216	0.200	0.319
corr.MLE-MLE	0.126	0.232	0.207	0.165	0.211
corr.DIST-UR	0.126	0.220	0.215	0.188	0.268
MLE-MLE	0.126	0.214	0.181	0.162	0.200
DIST-MLE	0.126	0.214	0.164	0.162	0.200
MODE-MLE	0.126	0.210	0.160	0.162	0.200
DIST-UR	0.122	0.175	0.215	0.186	0.267
MODE-UR	0.121	0.178	0.215	0.186	0.267
A_h*MLE	0.108	0.212	0.207	0.160	0.192

[1]The minimality refers to the minimum over h. The interval [0,10] was divided into 400 subintervals. The entry "corr.MLE-MLE" refers to the L^1 error of the MLE-MLE estimator for the value of h optimal for $B_h\, dF_n$. Likewise for corr.DIST-UR.

In Table 1 we report on the estimated quantities $\mathbb{E}[\min_h \|f^{nh}(\cdot\,; m) - f_o\|_1]$ for the various choices of the mode m and likewise for φ^{nh}. The sample standard deviations of the L^1 loss were similar for each density and were all in the 0.4–0.7 range.

The following conclusions may be drawn. For smooth densities with light tails, the unimodal rearrangement method à la Fougères (1997) works best, although we are at a loss to explain the large improvement for the Weibull distribution. For smooth densities with heavy tails, simulated by the mixture of Beta densities and the mixture of normals, the smoothed maximum likelihood method (the A_h*MLE method) works the best, with the maximum smoothed likelihood method a close second. The wonderful performance of the A_h*MLE method for the normal density is noteworthy, as is its less than wonderful performance on the uniform density. Perhaps it should be recalled [see Groeneboom(1985)] that when the Grenander (monotone) estimator is used to estimate a uniform density, then $\sqrt{n}\|f_n - f_o\|_1$ converges in distribution. The same is true for the unimodal (unsmoothed) maximum likelihood estimator, *but only for a deterministic choice of the mode*. For the uniform density on $[3, 8]$, with mode fixed at $m = 3$ the estimated expected L^1 error of $f^{nh}(\cdot\,, m)$ is 0.117, as opposed to 0.210 for $f^{nh}(\cdot\,; m_{nh})$, that is, when the mode is chosen by maximum likelihood.

A final word about the selection of the smoothing parameter. In Table 1 we also report on the errors of the MLE-MLE and DIST-UR methods corresponding to the optimal $B_h\, dF_n$, that is, for the value of h that minimizes the error $\|B_h\, dF_n - f_o\|_1$ in each replication. See the headings "corr.MLE-MLE" and "corr.DIST-UR" in Table 1. This shows that the smoothing parameter should be different when the unimodal estimator is used. In fact, the optimal h was always substantially smaller for the various unimodal estimators. However, it is not clear to the authors how this "optimal" h may be estimated in a rational way.

3. Assumptions and theorems.

Monotone densities. We begin by stating the contractivity of the mapping $\varphi \longmapsto \mathrm{lcm}(\varphi)$ in full generality. Let $\mathbb{R}_+ = [0, \infty)$. We consider "distances" of the form

$$\mathbb{J}(\varphi, \psi) = \int_0^\infty J\big(\varphi(x), \psi(x)\big)\,dx, \tag{3.1}$$

where $J\colon \mathbb{R}_+ \times \mathbb{R}_+ \longrightarrow \mathbb{R}_+ \cup \{+\infty\}$ satisfies

$$J(x, y) \text{ is continuous on } \mathbb{R}_+ \times (0, \infty), \text{ and} \tag{3.2}$$

$$\frac{J(b, p) - J(a, p)}{b - a} \le \frac{J(c, q) - J(b, q)}{c - b} \quad \begin{array}{l}\text{for all } 0 \le a < b < c,\\ \text{and all } p \ge q > 0.\end{array} \tag{3.3}$$

Upon taking $p = q$, condition (3.3) says that $J(x, y)$ is convex in x for every $y > 0$. When $J(x, y)$ is twice continuously differentiable, the condition is equivalent to

$$\frac{\partial^2 J}{\partial x^2}(x, y) \ge 0, \quad \frac{\partial^2 J}{\partial x\,\partial y}(x, y) \le 0 \qquad \text{for all } x, y > 0. \tag{3.4}$$

We also require condition (3.3) for the reverse function $(x, y) \longmapsto J(y, x)$ (but with different boundaries),

$$\frac{J(p, b) - J(p, a)}{b - a} \le \frac{J(q, c) - J(q, b)}{c - b} \quad \begin{array}{l}\text{for all } p \ge q \ge 0,\\ \text{and all } 0 < a < b < c,\end{array} \tag{3.5}$$

as well as

$$J(x, y) \text{ is convex in } x, y \text{ jointly.} \tag{3.6}$$

It is not clear to the authors whether all these conditions are independent, but we shall not pursue the point. Abusing notation somewhat, we say that $\mathbb{J}$ satisfies (3.2) if $\mathbb{J}$ is given by (3.1) and J satisfies (3.2), etc.

The standard examples of functions J satisfying (3.2) through (3.6) are constructed as follows. Let f be a nonnegative, convex function on $\mathbb{R}$, with $f(0) = 0$, and define J by

$$J(x, y) = f(x - y), \qquad x \ge 0,\ y \ge 0. \tag{3.7}$$

The choice $f(x) = |x|^p$ $(p \ge 1)$ covers the L^p distances (1.23). Alternatively, for f a nonnegative, convex function on $\mathbb{R}_+$, with $f(1) = 0$, define J by

$$J(x, y) = \begin{cases} yf(x/y), & x \ge 0,\ y > 0,\\ x \liminf\limits_{t\to\infty} t^{-1} f(t), & x > 0,\ y = 0,\\ 0, & x = 0,\ y = 0. \end{cases} \tag{3.8}$$

Thus, $J(x,0)=\infty$ for all $x>0$ or $J(x,0)<\infty$ for all x. Note that

$$x \liminf_{t\to\infty} t^{-1} f(t) = \liminf_{y\to 0+} y\, f(x/y), \qquad x>0.$$

The choices $f(x)=(\sqrt{x}-1)^2$, $x\log x+1-x$, and $(x-1)^2$ cover the Hellinger, Kullback–Leibler and Pearson φ^2 distances; see (1.20)–(1.22).

A general class of functions J satisfying (3.2) and (3.3), but not necessarily (3.5) and (3.6), are functions of the form

$$J(x,y)=f(x)-f(y)-f'(y)(x-y) \tag{3.9}$$

for (differentiable) convex functions f; for example, if $f(x)=-\sqrt{x}$, $x\ge 0$, then (3.5) and (3.6) fail, since $J(x,y)$ is not convex in y.

We are now ready for the statement of the contractivity theorem.

THEOREM 3.1 (Contractivity Theorem) (a) *Suppose that* $\mathbb{J}$ *satisfies* (3.2) *and* (3.3). *Then for all nonnegative* $\varphi \in L^1(0,\infty)$ *and for all decreasing* $\psi \in L^1(0,\infty)$,

$$\mathbb{J}\big(\mathrm{lcm}(\varphi),\psi\big) \le \mathbb{J}(\varphi,\psi).$$

(b) *If* $\mathbb{J}$ *satifies* (3.2) *through* (3.6) *then for all nonnegative* $\varphi,\psi \in L^1(0,\infty)$,

$$\mathbb{J}\big(\mathrm{lcm}(\varphi),\mathrm{lcm}(\psi)\big) \le \mathbb{J}(\varphi,\psi).$$

We note that Brunk (1965) proves part (a) of the above theorem for functionals J of the form (3.9).

The contractivity theorem implies error bounds on monotone density estimators. For reasons indicated in the introduction, we concentrate on L^1 error. Note that the solution of (1.9) is given by $f^{nh}=\mathrm{lcm}(\mathscr{A}_h\,dF_n)$.

COROLLARY 3.1. *Let* f_o *be monotone on* $(0,\infty)$. *If* $\mathscr{A}_h\,dF_n \ge 0$ *a.e., then*

$$\|f^{nh}-f_o\|_1 \le \|\mathscr{A}_h\,dF_n - f_o\|_1.$$

Convergence rates thus follow from the known rates for boundary kernels, under the usual assumptions on f_o; see, for example, Devroye (1987) or Jones (1993).

Unimodal densities. The basic assumptions are

(3.10) f_o is a continuous unimodal density, and

(3.11) the kernel A is symmetric, continuous and log-concave.

The purpose of (3.11) is to guarantee that $A_h * f_o$ is continuous and unimodal, and that $A_h * dF_n$ is continuous. At times we shall explicitly repeat this assumption. To obtain convergence rates, the usual nonparametric assumptions are required [see, e.g., Devroye and Györfi (1985)],

$$\|(f_o)''\|_1 < \infty, \tag{3.12}$$

$$\mathbb{E}[|X_1|^\kappa] < \infty \quad \text{for some } \kappa > 1. \tag{3.13}$$

Also, the density f_o is required to drop off from its set of modes in a certain manner. For any unimodal density f let

$$M(f) \stackrel{\text{def}}{=} \{m : m \text{ is a mode of } f\} \tag{3.14}$$

be the interval of modes of f. We assume that

$$\begin{aligned}&\text{there exists an open neighborhood } \Omega \text{ of } M(f_o),\\ &\text{such that } f_o''(x) \le 0 \text{ for all } x \in \Omega\end{aligned} \tag{3.15}$$

and

$$\liminf_{|x-m|\to 0} \big(f_o(m) - f_o(x)\big) \exp\left(\frac{\varepsilon}{|x-m|}\right) > 0 \quad \text{for all } \varepsilon > 0, \tag{3.16}$$

where m is the point in $M(f_o)$ closest to x.

Some simple sufficient conditions for (3.15) and (3.16) to hold are

$$f_o'' \text{ is continuous on } \Omega \text{ and } f_o''(m_o) < 0 \tag{3.17}$$

or

$$f_o''(x) = -c|x - m_o|^p \quad \text{for all } x \in \Omega, \tag{3.18}$$

for some $c > 0$, $p > 0$ (and then the mode is unique).

THEOREM 3.2. *Let $h \asymp n^{-\beta}$ for some $0 < \beta < 1$, and let f^{nh} be the solution of* (1.17). *Under assumptions* (3.10) *and* (3.11), *for all $m_{nh} \in M(f^{nh})$ and $m_o \in M(f_o)$,*

$$\|f^{nh} - f_o\|_1 \le \|A_h * dF_n - f_o\|_1 + c_{nh}|m_{nh} - m_o|\, \|A_h * (dF_n - dF_o)\|_\infty,$$

with $c_{nh} \to_{\text{as}} \sqrt{32}$.

THEOREM 3.3. *Let $h \asymp n^{-\beta}$, with $0 < \beta < 1$. Under assumptions* (3.10), (3.11), (3.13), (3.15) *and* (3.16), *for all $m_{nh} \in M(f^{nh})$,*

$$|m_{nh} - m| =_{\text{as}} o\big((\log n)^{-1}\big), \qquad n \to \infty,$$

where m is the point in $M(f_o)$ closest to m_{nh}.

Now (3.12) implies the a.s. bound of Silverman (1978), for $h \asymp n^{-\beta}$ $(0 < \beta < 1)$,

$$\|A_h * (dF_n - dF_o)\|_\infty =_{\text{as}} \mathscr{O}\big((nh)^{-1/2} \log(1/h)\big),$$

and then Theorem 3.2 implies the corollary [see Devroye (1991)].

COROLLARY 3.2. *Under assumptions* (3.10), (3.11), (3.12), (3.13), (3.15) *and* (3.16), *for $h \asymp n^{-1/5}$,*

$$\|f^{nh} - f_o\|_1 \le_{\text{as}} \|A_h * dF_n - f_o\|_1 + o(n^{-2/5}) =_{\text{as}} \mathscr{O}(n^{-2/5}).$$

The corollary says that asymptotically at least the unimodal estimator is not worse than the kernel estimator.

4. The contractivity for simple functions. Here we set out to prove the contractivity theorem. We begin by proving it first for simple step functions, defined here as nonnegative, integrable step functions on $(0, \infty)$. Precisely, consider adjacent intervals $A_1, A_2, \dots, A_n$ of the form

$$A_j = (x_{j-1}, x_j], \qquad j = 1, 2, \dots, n, \tag{4.1}$$

where $0 = x_0 < x_1 < \cdots < x_n < \infty$. A function is a *simple* function if it is nonnegative and can be written as

$$\varphi(x) = \begin{cases} \varphi_j, & x \in A_j, \quad j = 1, 2, \dots, n, \\ 0, & x > x_n, \end{cases} \tag{4.2}$$

or, equivalently,

$$\varphi(x) = \sum_{j \in I} \varphi_j \, \mathbb{1}(x \in A_j), \qquad x \in (0, \infty), \tag{4.3}$$

where $I = \{1, 2, \dots, n\}$, and in which $\varphi_j \geq 0$ for all j. The corresponding distributions are called simple distributions. We do not insist on the simple functions being probability density functions, nor on the φ_j being distinct.

The proofs to follow are based on a judicious use of the pool-adjacent-violators-algorithm (p-a-v-a); see Barlow, Bartholomew, Bremner and Brunk (1972). We recall some of its properties. Let the density φ be given by (4.3), on adjacent intervals $A_1, A_2, \dots, A_n$ as in (4.1). The "basic step" of the algorithm consists of locating a violation of the assumption that φ is decreasing, and fixing it. That is, find a pair of adjacent intervals A_j, A_{j+1} on which $\varphi_j < \varphi_{j+1}$, and replace φ by φ^{new},

$$\varphi^{\text{new}}(x) = \begin{cases} \varphi(x), & x \notin A_j \cup A_{j+1}, \\ \dfrac{|A_j|\varphi_j + |A_{j+1}|\varphi_{j+1}}{|A_j| + |A_{j+1}|}, & x \in A_j \cup A_{j+1}. \end{cases} \tag{4.4}$$

The basic step is concluded by pooling the intervals A_j and A_{j+1} into one interval.

It is helpful to introduce notation to indicate where the "new" function came from, that is,

$$\varphi^{\text{new}} = \operatorname{lcm}(\varphi, A_j \cup A_{j+1}), \tag{4.5}$$

and the corresponding distribution Φ^{new} by $\operatorname{LCM}(\Phi, A_j \cup A_{j+1})$. Likewise, after a number of steps of p-a-v-a we denote the computed φ as $\operatorname{lcm}(\varphi, B)$, where B is the union of all the 'violating' intervals A_j, A_{j+1}. In general, if Φ is a distribution function with density φ,

$$\operatorname{lcm}(\varphi, B) = \begin{cases} \text{the left continuous function which} \\ \text{is a.e. equal to the derivative of the} \\ \text{least majorant of } \Phi \text{ which is concave} \\ \text{on each interval contained in } B. \end{cases} \tag{4.6}$$

The p-a-v-a works because

$$\operatorname{lcm}\bigl(\operatorname{lcm}(\varphi, A_j \cup A_{j+1})\bigr) = \operatorname{lcm}(\varphi), \tag{4.7}$$

so after repeated applications of the basic step until there are no more violations of the monotonicity requirement, then $\mathrm{lcm}(\varphi)$ has been computed. Finally, the p-a-v-a terminates after a finite number of steps, since, due to the "pooling," the number of intervals (the A_j) is decreased by one after each step.

We are now ready to prove Theorem 3.1 for simple functions.

LEMMA 4.1. *Theorem* 3.1(a) *holds for all simple functions* φ, ψ, *with* ψ *decreasing*.

PROOF. Suppose φ and ψ are supported on $(0, T)$, for some $T > 0$, and write φ as

$$\varphi(x) = \sum_{j \in I} \varphi_j \mathbb{1}(x \in A_j), \qquad x \in (0, \infty),$$

for adjacent intervals $A_1, A_2, \ldots, A_n$, with $\bigcup_j A_j = (0, T)$. We may suppose that $\mathbb{J}(\varphi, \psi) < \infty$, since otherwise there is nothing to prove. Then also $J\big(\varphi(x), \psi(x)\big) < \infty$ for all $x > 0$.

We compute $\mathrm{lcm}(\varphi)$ using the p-a-v-a. Thus we find two adjacent intervals A_j and A_{j+1} such that $\varphi_j < \varphi_{j+1}$ and compute φ^{new} by (4.4). We show that

$$\mathbb{J}(\varphi^{\mathrm{new}}, \psi) \le \mathbb{J}(\varphi, \psi). \tag{4.8}$$

It suffices to consider the contributions of $A_j \cup A_{j+1}$ only, that is, to show that

$$\int_{A_j \cup A_{j+1}} J(\varphi^{\mathrm{new}}(x), \psi(x))\, dx \le \int_{A_j} J(\varphi_j, \psi(x))\, dx + \int_{A_{j+1}} J(\varphi_{j+1}, \psi(x))\, dx,$$

which may be rewritten as

$$\begin{aligned} &\int_{A_j} \big\{J\big(\varphi^{\mathrm{new}}(x), \psi(x)\big) - J\big(\varphi_j, \psi(x)\big)\big\}\, dx \\ &\qquad \le \int_{A_{j+1}} \big\{J\big(\varphi_{j+1}, \psi(x)\big) - J\big(\varphi^{\mathrm{new}}(x), \psi(x)\big)\big\}\, dx. \end{aligned} \tag{4.9}$$

It is annoying that the two integrals may be over intervals of different lengths. Using a simple change of variables to transform the intervals of integration A_j and A_{j+1} into, respectively, $[-1, 0]$ and $[0,1]$, gives

$$\begin{aligned} &|A_j| \int_{-1}^{0} \big\{J\big(\varphi^{\mathrm{new}}, \psi_1(x)\big) - J\big(\varphi_j, \psi_1(x)\big)\big\}\, dx \\ &\qquad \le |A_{j+1}| \int_0^1 \big\{J\big(\varphi_{j+1}, \psi_2(x)\big) - J\big(\varphi^{\mathrm{new}}, \psi_2(x)\big)\big\}\, dx, \end{aligned} \tag{4.10}$$

where $\psi_1(x)$ is the appropriate scaled version of $\psi(x)$ on A_j and $\psi_2(x)$ the scaled version of $\psi(x)$ on A_{j+1}, and where we dropped the argument of φ^{new}, since it is constant on $A_j \cup A_{j+1}$. Now note that

$$\varphi^{\mathrm{new}} - \varphi_j = \frac{|A_{j+1}|\big(\varphi_{j+1} - \varphi_j\big)}{|A_j| + |A_{j+1}|}, \qquad \varphi_{j+1} - \varphi^{\mathrm{new}} = \frac{|A_j|\big(\varphi_{j+1} - \varphi_j\big)}{|A_j| + |A_{j+1}|},$$

and $\varphi_{j+1} - \varphi_j > 0$, so that (4.10) is equivalent to

$$\begin{aligned} &\int_{-1}^{0} \frac{J(\varphi^{\text{new}}, \psi_1(x)) - J(\varphi_j, \psi_1(x))}{\varphi^{\text{new}} - \varphi_j}\, dx \\ &\quad \le \int_0^1 \frac{J(\varphi_{j+1}, \psi_2(x)) - J(\varphi^{\text{new}}, \psi_2(x))}{\varphi_{j+1} - \varphi^{\text{new}}}\, dx. \end{aligned} \tag{4.11}$$

Thus, we are done if (4.11) holds. Since ψ is decreasing, then

$$\psi_1(-x) \ge \psi_1(0) \ge \psi_2(0) \ge \psi_2(t) \quad \text{for all } x,\ t \in [0, 1]$$

and $\varphi_j \le \varphi^{\text{new}} \le \varphi_{j+1}$, condition (3.3) implies that (4.11) holds, and so does (4.8).

The proof is concluded by running the p-a-v-a to completion. By (4.8), each step decreases $\mathbb{J}(\varphi, \psi)$, and after finitely many steps we are done. □

We now turn our attention to proving Theorem 3.1(b), but still for simple functions φ, ψ. Of course, Lemma 4.1 covers the case where ψ is decreasing. It also covers the reverse case; that is,

(4.12) if φ is decreasing and J satisfies (3.2) and (3.5), then $\mathbb{J}(\varphi, \operatorname{lcm}(\psi)) \le \mathbb{J}(\varphi, \psi)$.

The other extreme case, where ψ is increasing, is useful in establishing the general result.

LEMMA 4.2. *Suppose* $\mathbb{J}$ *satisfies* (3.2), (3.3), (3.5) *and* (3.6). *Let* φ *and* ψ *be simple functions on the same partition*, *that is*, *there exist adjacent intervals* $A_1, A_2, \dots, A_n$, *such that* $\bigcup_j A_j = (0, T)$ *for some* $T > 0$ *and*

$$\varphi(x) = \sum_{j \in I} \varphi_j \mathbb{1}(x \in A_j), \quad \psi(x) = \sum_{j \in I} \psi_j \mathbb{1}(x \in A_j), \qquad x \in (0, \infty).$$

Let $0 \le P < Q \le T$, *and assume that* ψ *is increasing on* (P, Q). *Then*

$$\mathbb{J}\big(\operatorname{lcm}(\varphi, (P, Q)), \operatorname{lcm}(\psi, (P, Q))\big) \le \mathbb{J}(\varphi, \psi).$$

PROOF. We may assume that $\mathbb{J}(\varphi, \psi) < \infty$, so that $J(\varphi(x), \psi(x)) < \infty$ for all x. We may also assume that $(P, Q) = (0, T)$.

We apply the p-a-v-a to φ and ψ simultaneously. That is, we find intervals A_j, A_{j+1} such that $\varphi_j < \varphi_{j+1}$, and compute $\varphi^{\text{new}} = \operatorname{lcm}(\varphi, A_j \cup A_{j+1})$. Since ψ is increasing, then $\psi_j \le \psi_{j+1}$, and we do the "same" step on ψ, so we compute $\psi^{\text{new}} = \operatorname{lcm}(\psi, A_j \cup A_{j+1})$. Note that after pooling the intervals A_j and A_{j+1}, both φ and ψ are simple functions on the new set of adjacent intervals.

Now we show that

$$\mathbb{J}\big(\varphi^{\text{new}}, \psi^{\text{new}}\big) \le \mathbb{J}(\varphi, \psi). \tag{4.13}$$

Again, only the contributions of the intervals A_j and A_{j+1} need to be considered, and all functions in question are constant on these intervals. Thus it suffices to show that

$$\big(|A_j| + |A_{j+1}|\big) J(\varphi^{\text{new}}, \psi^{\text{new}}) \le |A_j| J(\varphi_j, \psi_j) + |A_{j+1}| J(\varphi_{j+1}, \psi_{j+1}),$$

and this holds by the convexity of $J(x, y)$ in x, y jointly.

This describes one step of p-a-v-a applied to φ, and simultaneously on ψ. After finitely many steps we have thus computed $\mathrm{lcm}(\varphi)$. Denote the corresponding ψ by ψ^{last}. By induction, using (4.13), it follows that

$$\mathbb{J}\big(\mathrm{lcm}(\varphi), \psi^{\mathrm{last}}\big) \le \mathbb{J}(\varphi, \psi).$$

Now, if $\psi^{\mathrm{last}} = \mathrm{lcm}(\psi)$ we are done. If $\psi^{\mathrm{last}} \ne \mathrm{lcm}(\psi)$, then we still have $\mathrm{lcm}(\psi) = \mathrm{lcm}(\psi^{\mathrm{last}})$, and since $\mathrm{lcm}(\varphi)$ is obviously decreasing, (4.12) implies

$$\mathbb{J}\big(\mathrm{lcm}(\varphi), \mathrm{lcm}(\psi)\big) \le \mathbb{J}\big(\mathrm{lcm}(\varphi), \psi^{\mathrm{last}}\big) \le \mathbb{J}(\varphi, \psi).$$

This concludes the proof. □

THEOREM 4.1. *Theorem* 3.1(b) *holds for all simple functions* φ, ψ.

PROOF. Assume that $\mathbb{J}(\varphi, \psi) < \infty$. Suppose that ψ is a step function on the adjacent intervals A_j, $j = 1, 2, \dots, n$. We apply a basic step of p-a-v-a to ψ and find adjacent intervals A_j and A_{j+1} such that $\psi_j < \psi_{j+1}$. Thus ψ is increasing on $A_j \cup A_{j+1}$. Since φ is a simple function, φ and ψ are simple functions on the adjacent intervals B_i, $i = 1, 2, \dots, m$, with

$$\bigcup_{1\le i\le m} B_i = A_j \cup A_{j+1}.$$

So Lemma 4.2 applies, and gives

$$\mathbb{J}\big(\mathrm{lcm}(\varphi, A_j \cup A_{j+1}), \mathrm{lcm}(\psi, A_j \cup A_{j+1})\big) \le \mathbb{J}(\varphi, \psi). \tag{4.14}$$

After finitely many steps of p-a-v-a applied to ψ, the algorithm terminates, and we have computed $\mathrm{lcm}(\psi)$ as well as $\mathrm{lcm}(\varphi, B)$, where B is the union of all the violating subintervals A_j, A_{j+1}. By induction, using (4.14), we thus obtain that

$$\mathbb{J}\big(\mathrm{lcm}(\varphi, B), \mathrm{lcm}(\psi)\big) \le \mathbb{J}(\varphi, \psi).$$

Since $\mathrm{lcm}(\psi)$ is obviously decreasing, Lemma 4.1 applies, so that

$$\mathbb{J}\big(\mathrm{lcm}(\varphi), \mathrm{lcm}(\psi)\big) \le \mathbb{J}\big(\mathrm{lcm}(\varphi, B), \mathrm{lcm}(\psi)\big) \le \mathbb{J}(\varphi, \psi). \qquad \square$$

5. The general contractivity result. The next step is to extend Lemmas (4.8) and (4.16) to arbitrary densities. Going from simple distributions to arbitrary ones is of course effected by a limiting process. The L^1 case is a simple consequence of the following lemma and is instrumental in proving the general case.

LEMMA 5.1. *Let* Φ_o *be a distribution on* $(0, \infty)$ *with density* φ_o, *and let* $\{\Phi_n\}_n$ *be a sequence of simple distributions with densities* φ_n, *with* $\|\varphi_n - \varphi_o\|_1 \longrightarrow 0$. *Then*

$$\|\mathrm{lcm}(\varphi_n) - \mathrm{lcm}(\varphi_o)\|_1 \longrightarrow 0.$$

PROOF. Let $\Psi_n = \mathrm{LCM}(\Phi_n)$, and $\psi_n = \mathrm{lcm}(\varphi_n)$. Now Lemma 4.1 gives

$$||\psi_n - \psi_m||_1 \le ||\varphi_n - \varphi_m||_1,$$

so that $\{\psi_n\}_n$ is a Cauchy sequence. Thus, there exists a $\psi_o \in L^1(0,\infty)$ with $||\psi_n - \psi_o||_1 \longrightarrow 0$, and all we need to show is that $\psi_o = \mathrm{lcm}(\varphi_o)$, or that $\Psi_o = \mathrm{LCM}(\Phi_o)$, where Ψ_o is the distribution corresponding to ψ_o.

Since

$$||\Psi_n - \Psi_o||_\infty \le ||\psi_n - \psi_o||_1 \longrightarrow 0, \tag{5.1}$$

then Ψ_o is the pointwise limit of a sequence of concave distributions, and it follows that Ψ_o is concave. Also, since $\Psi_n \ge \Phi_n$ for all n, then $\Psi_o \ge \Phi_o$. Thus, ψ_o is a concave majorant of Φ_o.

Let Θ be any concave majorant of Φ_o. Then $\Theta \ge \Psi_n - \varepsilon_n$, where $\varepsilon_n = ||\Psi_n - \Psi_o||_\infty$. This implies that $\Theta + \varepsilon_n$ is a majorant of Ψ_n, and it is concave. So, it dominates the least concave majorant, $\Theta + \varepsilon_n \ge \mathrm{LCM}(\Phi_n)$, and

$$\Theta + \varepsilon_n \ge \Psi_n \ge \Psi_o - \varepsilon_n,$$

with ε_n as before. This implies that $\Theta \ge \Psi_o - 2\varepsilon_n$. Since $\varepsilon_n \to 0$ [see (5.1)], then $\Theta \ge \Psi_o$, and so $\Psi_o = \mathrm{LCM}(\Phi_o)$. $\square$

The L^1 case of Theorem 3.1(b) follows immediately from Lemma 5.1. The proof for general distances is somewhat more complicated.

PROOF OF THEOREM 3.1(b). Assume that $\mathbb{J}(\varphi,\psi) < \infty$. The proof is an exercise in measure theory; see Wheeden and Zygmund (1977). That is, we approximate φ_o and ψ_o by simple functions φ_n and ψ_n, and take limits in the inequality

$$\mathbb{J}\big(\mathrm{lcm}(\varphi_n), \mathrm{lcm}(\psi_n)\big) \le \mathbb{J}(\varphi_n, \psi_n).$$

Let $n \in \mathbb{N}$. Since φ_o, ψ_o and $J_o \equiv J(\varphi_o(\cdot), \psi_o(\cdot))$ are nonnegative elements of $L^1(0,\infty)$, there exists a $T = T(n) > 0$ such that

$$\int_T^\infty \big(\varphi_o(x) + \psi_o(x) + J_o(x)\big)\,dx < \frac{1}{n}.$$

By Lusin's theorem, there exists a closed set $F \subset (0,T)$, with relative complement $G = \{x \in (0,T)\colon x \notin F\}$, such that φ_o, ψ_o and J_o are uniformly continuous relative to F, and

$$|G| < \frac{1}{n}, \qquad \int_G \big(\varphi_o(x) + \psi_o(x) + J_o(x)\big)\,dx < \frac{1}{n}.$$

The uniform continuity relative to F means that

$$\forall\, \varepsilon > 0\ \exists\, \delta > 0\ \forall\, x, y \in F\colon |x-y| < \delta \Longrightarrow |\varphi_o(x) - \varphi_o(y)| < \varepsilon.$$

Now φ_o may be approximated by step functions in the following manner. For $k \in \mathbb{N}$, let

$$A_{ik} = \big(i/k, (i+1)/k\big], \qquad i = 0, 1, 2, \ldots,$$

and let $I(k) = \{i: |A_{ik} \cap F| > 0\}$. Then for arbitrary $\theta_{ik} \in A_{ik} \cap F$,

$$\lim_{k\to\infty} \sum_{i\in I(k)} \int_{A_{ik}\cap F} |\varphi_o(\theta_{ik}) - \varphi_o(x)|\, dx = 0.$$

Also, since φ_o is bounded on F, being a continuous function on a compact set,

$$\limsup_{k\to\infty} \sum_{i\in I(k)} \bigl|\{x \in A_{ik}: x \notin F\}\bigr|\, \varphi_o(\theta_{ik}) \le \frac{c}{n},$$

where $c = \max\{\varphi_o(x): x \in F\}$. Now choose $k = k(n)$ so large that

$$\sum_{i\in I(k)} \int_{A_{ik}\cap F} |\varphi_o(\theta_{ik}) - \varphi_o(x)|\, dx < \frac{1}{n},$$

$$\sum_{i\in I(k)} |\{x \in A_{ik}: x \notin F\}|\, \varphi_o(\theta_{ik}) < \frac{2c}{n}.$$

Arranging things so that the same inequalities hold for ψ_o and J_o, we define

$$\begin{aligned} \varphi_n(x) &= \sum_{i\in I(k)} \varphi_o(\theta_{ik})\, \mathbb{1}(x \in A_{ik}), \\ \psi_n(x) &= \sum_{i\in I(k)} \psi_o(\theta_{ik})\, \mathbb{1}(x \in A_{ik}), \\ J_n(x) &= J\bigl(\varphi_n(x), \psi_n(x)\bigr), \qquad x > 0. \end{aligned} \tag{5.2}$$

This completes the construction of the approximating sequences.

We now take limits. First

$$\begin{aligned} \|\varphi_n - \varphi_o\|_1 \le{}& \int_T^\infty \varphi_o(x)\, dx + \int_G \varphi_o(x)\, dx \\ &+ \int_F |\varphi_k(x) - \varphi_o(x)|\, dx + \sum_{i\in I(k)} \bigl|\{x \in A_{ik}: x \notin F\}\bigr|\, \varphi_o(\theta_{ik}) \end{aligned}$$

and thus

$$\|\varphi_n - \varphi_o\|_1 < \frac{1}{n} + \frac{1}{n} + \frac{1}{n} + \frac{2c}{n},$$

so $\varphi_n \longrightarrow \varphi_o$ in L^1. The same arguments give $\psi_n \longrightarrow \psi_o$ and $J_n \longrightarrow J_o$ in L^1. By Lemma (5.1), then also $\mathrm{lcm}(\varphi_n) \longrightarrow \mathrm{lcm}(\varphi_o)$ in L^1, and so, for a subsequence, $\mathrm{lcm}(\varphi_n) \longrightarrow \mathrm{lcm}(\varphi_o)$ almost everywhere. Likewise, along a subsequence of the subsequence,

$$\mathrm{lcm}(\psi_n) \longrightarrow \mathrm{lcm}(\psi_o), \quad J_n \longrightarrow J_o \qquad \text{almost everywhere.}$$

By Fatou's lemma along the appropriate subsequence,

$$\mathbb{J}\bigl(\mathrm{lcm}(\varphi_o), \mathrm{lcm}(\psi_o)\bigr) \le \liminf_{n\to\infty} \mathbb{J}\bigl(\mathrm{lcm}(\varphi_n), \mathrm{lcm}(\psi_n)\bigr).$$

Finally, since $J_n \longrightarrow J_o$ in L^1, then $\lim_{n\to\infty} \mathbb{J}(\varphi_n, \psi_n) = \mathbb{J}(\varphi_o, \psi_o)$. The theorem follows. □

The proof of Theorem 3.1(a) follows likewise. We finish this section with useful and obvious results.

LEMMA 5.2. *Let Φ be a distribution on $(0,\infty)$ with continuous density φ. Then the following three statements hold*:

(a) $\mathrm{lcm}(\varphi)$ *is continuous*.
(b) $[\mathrm{lcm}(\varphi)](0) \geq \varphi(0)$.
(c) *For all* $x > t > 0$: $[\mathrm{lcm}(\varphi; (t,\infty))](x) \geq [\mathrm{lcm}(\varphi)](x)$.

6. Estimating smooth unimodal densities. In this section we apply the above results on monotone density estimation to the estimation of unimodal densities

We first show the existence of solutions to (1.9). In the process some nice qualitative observations regarding the estimator of the mode are made. It is useful to consider the estimation problem with an a priori fixed mode m, that is,

$$\text{minimize} \quad L_{nh}(f) \text{ over } f \in \mathscr{U}(m), \tag{6.1}$$

where

$$\mathscr{U}(m) = \{f \in L^1(\mathbb{R}): f \geq 0,\ f \text{ is unimodal, with mode at } m\}. \tag{6.2}$$

It is obvious that (6.1) has a unique solution, which we denote by $f^{nh}(\cdot; m)$. Thus, the existence of solutions to (1.9) reduces to the existence of solutions of

$$\begin{aligned} &\text{minimize} \quad L^*_{nh}(m) \stackrel{\text{def}}{=} \min\{L_{nh}(f): f \in \mathscr{U}(m)\},\\ &\text{subject to} \quad -\infty < m < \infty. \end{aligned} \tag{6.3}$$

Any solution of (6.3) will be denoted by m_{nh}. Since (6.3) is a *parametric* problem, matters should simplify considerably, but apparently they do not.

The existence of solutions to (6.3) is established as follows. We begin with an observation regarding solutions of (6.1). We recall that A is log-concave and continuous, so that $A_h * dF_n$ is continuous.

LEMMA 6.1. *Let $A_h * dF_n$ be continuous, and let $f(\cdot; m)$ be the solution of* (6.1). *Then* (*at least*) *one of the following three statements holds*:

(a) $f(\cdot; m)$ *is continuous, and* $f(m; m) = A_h * dF_n(m)$.
(b) $\exists\, \delta > 0$: $f(\cdot; m)$ *is constant on* $(m-\delta, m)$ *and* $f(m-0; m) \geq f(m+0; m)$.
(c) $\exists\, \delta > 0$: $f(\cdot; m)$ *is constant on* $(m, m+\delta)$ *and* $f(m+0; m) \geq f(m-0; m)$.

PROOF. By Lemma 5.2 we have that $f(\cdot; m)$ is continuous everywhere except possibly at $x = m$, and that $f(m \pm 0; m) \geq A_h * dF_n(m)$. If $f(m+0; m) = f(m-0; m) = A_h * dF_n(m)$ then case (a) holds. If it does not then $f(m+0;m) > A_h * dF_n(m)$ or $f(m-0;m) > A_h * dF_n(m)$. So suppose $f(m+0;m) > A_h * dF_n(m)$. By the continuity of $f(\cdot;m)$ on (m,∞), and the continuity of $A_h * dF_n$ there exists a $\delta > 0$ such that

$$f(x;m) > A_h * dF_n(x) \quad \text{for all } x \in (m, m+\delta).$$

This can only happen if $f(\cdot;m)$ is constant on $(m, m+\delta)$. If now $f(m-0;m) = A_h * dF_n(m)$, then case (c) holds. If on the other hand $f(m-0;m) > A_h * dF_n(m)$, then by the same reasoning as above $f(\cdot;m)$ is constant on some interval $(m-\varepsilon, m)$. Then at least one of the cases (b) and (c) holds. The case $f(m-0;m) > A_h * dF_n(m)$ goes the same way. $\square$

The lemma immediately establishes the following useful (theoretically and practically) result regarding the maximum smoothed likelihood estimator of the mode.

THEOREM 6.1. *The minimum of L^*_{nh} occurs at a local mode of $A_h * dF_n$.*

PROOF. Let $m \in \mathbb{R}$, and suppose that $f(\cdot;m)$ is constant on the interval $[m, m+\delta]$ with $\delta > 0$ chosen as large as possible, and $f(m+0;m) \geq f(m-0;m)$. Then $f(\cdot;m) \in \mathscr{U}(\eta)$ for all $\eta \in [m, m+\delta]$. Consequently,

$$L^*_{nh}(m) = L_{nh}(f(\cdot;m)) \geq L^*_{nh}(\eta) \qquad \text{for all } \eta \in [m, m+\delta]. \tag{6.4}$$

Let $\eta_o \in [m, m+\delta]$ be the largest local mode of $A_h * dF_n$ on $[m, m+\delta]$. There are two possibilities: either $A_h * dF_n(\eta_o) = f(m+0;m)$ or $A_h * dF_n(\eta_o) > f(m+0;m)$. In the first case $A_h * dF_n(x) = f(m+0;m)$ for all $x \in [m, m+\delta]$, and m is a local mode of $A_h * dF_n$. In the second case, it is obvious that $f^{nh}(\cdot;\eta_o)$ is different from $f^{nh}(\cdot;m)$, since $f^{nh}(\eta_o;\eta_o) > f^{nh}(m+0;m)$. Then the uniqueness of $f^{nh}(\cdot;\eta_o)$ shows that $L^*_{nh}(m) > L^*_{nh}(\eta_o)$, and in view of (6.4) m is not a local minimum of L^*_{nh}. The same conclusion prevails if $f(\cdot;m)$ is constant on an interval $[m-\delta, m]$, and $f(m-0;m) \geq f(m+0;m)$. Thus, if m is a local minimum of L^*_{nh}, part (a) of Lemma 6.1 must hold, and m is a local mode of $A_h * dF_n$. $\square$

COROLLARY 6.1. *If f^{nh} solves* (1.17), *then $f^{nh}(m_{nh}) = A_h * dF_n(m_{nh})$.*

The consistency of the estimator of the mode is considered next. The moment condition in the lemma can probably be removed, but since it is needed for the bound (1.8), there is not much point to it.

LEMMA 6.2. *Suppose that $A_h * f_o$ is unimodal. Let $h \asymp n^{-\beta}$ for some $0 < \beta < 1$. If $\mathbb{E}[|X|^\kappa] < \infty$ for some $\kappa > 1$ then*

$$\min\{|m_{nh} - m| \colon m \in M(f_o)\} \longrightarrow_{a.s.} 0.$$

PROOF. Let m_o be the point in $M(f_o)$ closest to m_{nh}. Assume that $m_{nh} < m_o$. Now with f^{nh} the solution of (1.9), for all $s > 1$,

$$D(A_h * dF_n, f^{nh}) \leq D(A_h * dF_n, A_h * dF_o) =_{\text{a.s.}} \mathscr{O}((nh)^{-1/2}(\log n)^s),$$

since f_o has a finite moment of order > 1; see Eggermont and LaRiccia (1999). Now the inequality of Kemperman (1967) for all nonnegative L^1 functions φ, ψ,

$$\left\{\int_{\mathbb{R}} \tfrac{2}{3}\varphi + \tfrac{4}{3}\psi\right\}^{-1}\left\{\int_{\mathbb{R}} |\varphi - \psi|\right\}^2 \le D(\varphi, \psi), \tag{6.5}$$

gives for all $s > \frac{1}{2}$,

$$\|f^{nh} - A_h * dF_o\|_1 =_{\mathrm{as}} \mathscr{O}((nh)^{-1/4}(\log n)^s).$$

Since $\|f_o - A_h * dF_n\|_1 \longrightarrow 0$ for $h \to 0$ [see, e.g., Devroye and Györfi (1985)], the triangle inequality then implies $\|f^{nh} - f_o\|_1 \longrightarrow_{\text{a.s.}} 0$, and thus

$$\int_{m_{nh}}^{m_o} |f^{nh}(x) - f_o(x)|\, dx \longrightarrow_{\text{a.s.}} 0. \tag{6.6}$$

Now recall that $m_{nh} < m_o$. On the interval $[m_{nh}, m_o]$ the function f_o is increasing and f^{nh} is decreasing. Thus, the increasing estimator of f^{nh} on this interval is a constant function, and it follows from Corollary 3.1 that

$$\min_{c} \int_{m_{nh}}^{m_o} |f_o(x) - c|\, dx \le \int_{m_{nh}}^{m_o} |f^{nh}(x) - f_o(x)|\, dx \longrightarrow_{\text{a.s.}} 0. \tag{6.7}$$

By the continuity of f_o then $f_o(m_{nh}) - f_o(m_o) \longrightarrow_{\text{a.s.}} 0$, and the lemma follows. □

On to convergence rates. We first derive the bound for the L^1 error.

Proof of Theorem 3.2. Assume that $m_{nh} < m_o$. (The case $m_{nh} > m_o$ goes the same way.) We first split the integral

$$\int_{\mathbb{R}} |f^{nh} - f_o| = \int_{-\infty}^{m_{nh}} \cdots + \int_{m_{nh}}^{m_o} \cdots + \int_{m_o}^{\infty} \cdots. \tag{6.8}$$

For the first integral on the right of (6.8) we have by Corollary 3.1,

$$\int_{-\infty}^{m_{nh}} |f^{nh} - f_o| \le \int_{-\infty}^{m_{nh}} |A_h * dF_n - f_o|.$$

For the last integral in (6.8) the triangle inequality gives

$$\begin{aligned}\int_{m_o}^{\infty} |f^{nh} - f_o| \le \int_{m_o}^{\infty} &|f^{nh}(\cdot\,; m_{nh}) - f^{nh}(\cdot\,; m_o)| \\ &+ \int_{m_o}^{\infty} |f^{nh}(\cdot\,; m_o) - f_o|,\end{aligned} \tag{6.9}$$

and for the last integral of this, again Corollary 3.1 gives

$$\int_{m_o}^{\infty} |f^{nh}(\cdot\,; m_o) - f_o| \le \int_{m_o}^{\infty} |A_h * dF_n - f_o|.$$

The first integral on the right of (6.9) requires some work. First note that

$$f^{nh}(x; m_o) = [\mathrm{lcm}(A_h * dF_n, (m_o, \infty))](x), \qquad x > m_o,$$

and recall Lemma 5.2 to see that not only

$$\int_{m_o}^{\infty} f^{nh}(x; m_o)\, dx = \int_{m_o}^{\infty} A_h * dF_n(x)\, dx, \tag{6.10}$$

but also

$$f^{nh}(x; m_{nh}) = [\mathrm{lcm}(\psi, (m_{nh}, \infty))](x), \qquad x > m_{nh},$$

where

$$\psi(x) = \begin{cases} f^{nh}(x; m_o), & x > m_o, \\ A_h * dF_n(x), & m_{nh} < x < m_o. \end{cases}$$

It then follows from Lemma 5.2 that

$$f^{nh}(x; m_o) \geq f^{nh}(x; m_{nh}) \quad \text{for } x > m_o. \tag{6.11}$$

Thus, for the first integral on the right of (6.9), we obtain

$$\begin{aligned} \int_{m_o}^{\infty} |f^{nh}(x; m_{nh}) - f^{nh}(x; m_o)|\, dx &= \int_{m_o}^{\infty} \{f^{nh}(x; m_o) - f^{nh}(x; m_{nh})\}\, dx \\ &= \int_{m_o}^{\infty} \{A_h * dF_n(x) - f^{nh}(x; m_{nh})\}\, dx \\ &= \int_{m_{nh}}^{m_o} \{f^{nh}(x; m_{nh}) - A_h * dF_n(x)\}\, dx, \end{aligned} \tag{6.12}$$

where the first equality is by (6.11), the second one by (6.10), and the third one again by (6.10) with m_o replaced by m_{nh}.

Finally, for the middle integral on the right of (6.8) the triangle inequality suffices. Putting it all back together gives the inequality

$$\|f^{nh} - f_o\|_1 \leq \|A_h * dF_n - f_o\|_1 + 2\int_{m_{nh}}^{m_o} |f^{nh} - A_h * dF_n|.$$

The last remaining integral is duly bounded in the next rather technical lemma. □

Note that in the above proof no use was made of the fact that f^{nh} is the maximum smoothed likelihood solution, or that m_{nh} is its mode. In the next lemma that changes.

LEMMA 6.3. *Under the same conditions as Lemma* 6.2,

$$\left|\int_{m_{nh}}^{m_o} |f^{nh} - A_h * dF_n|\right| \leq c_{nh}|m_{nh} - m_o|\|A_h * (dF_n - dF_o)\|_\infty,$$

where $c_{nh} \longrightarrow_{a.s.} \sqrt{8}$.

PROOF. We only consider the case $m_{nh} < m_o$. Let

$$\varphi^{nh}(x) = \begin{cases} f^{nh}(x), & x \notin [m_{nh}, m_o], \\ A_h * dF_o(x) + \delta_{nh}, & x \in [m_{nh}, m_o], \end{cases}$$

where $\delta_{nh} = \|A_h * (dF_n - dF_o)\|_\infty$. Below we prove that

$$\varphi^{nh} \quad \text{is unimodal,}$$

so that φ^{nh} is a (presumably unsuccessful) candidate for a solution of (1.17); that is,

$$D(A_h * dF_n, f^{nh}) \leq D(A_h * dF_n, \varphi^{nh}).$$

Both sides of this inequality are integrals over $(-\infty, \infty)$, but the integrands differ only on $[m_{nh}, m_o]$, so that

$$\begin{aligned} &\int_{m_{nh}}^{m_o} \left\{ A_h * dF_n \log \frac{A_h * dF_n}{f^{nh}} + f^{nh} - A_h * dF_n \right\} \\ &\quad \leq \int_{m_{nh}}^{m_o} \left\{ A_h * dF_n \log \frac{A_h * dF_n}{\varphi^{nh}} + \varphi^{nh} - A_h * dF_n \right\}. \end{aligned} \tag{6.13}$$

The integral on the right may be bounded by Pearson's φ^2 distance,

$$\int_{m_{nh}}^{m_o} \frac{|A_h * dF_n - \varphi^{nh}|^2}{\varphi^{nh}},$$

the square root of which may be bounded as

$$\begin{aligned} &\{f_h(\mu^{nh})\}^{-1/2} \left[\left\{ \int_{m_{nh}}^{m_o} |A_h * (dF_n - dF_o)|^2 \right\}^{1/2} + (m_o - m_{nh})^{1/2} \delta_{nh} \right] \\ &\quad \leq 2\{f_h(\mu^{nh})\}^{-1/2} |m_{nh} - m_o|^{1/2} \|A_h * (dF_n - dF_o)\|_\infty, \end{aligned}$$

where $f_h(\mu^{nh}) = \min\{f_h(x) : x \in [m_{nh}, m_o]\}$. The integral on the left of (6.13) may be bounded below by [cf. (6.5)]

$$\left\{ \int_{m_{nh}}^{m_o} \tfrac{2}{3} A_h * dF_n + \tfrac{4}{3} f^{nh} \right\}^{-1} \left\{ \int_{m_{nh}}^{m_o} |f^{nh} - A_h * dF_n| \right\}^2,$$

and the first factor behaves as $\{2f_o(m_o)(m_o - m_{nh})\}^{-1}$ for $nh \to \infty$, $h \to 0$. Here we used that $m_{nh} \longrightarrow_{\text{a.s.}} m_o$; see Lemma 6.2.

Since $f_h \longrightarrow f_o$ uniformly, as $nh \to \infty$, $h \to 0$, putting all this together proves the required bound.

We still must prove (6.13). Some of the crucial facts required in its proof are that $f^{nh}(m_{nh}; m_{nh}) = A_h * dF_n(m_{nh})$; see Corollary 6.1, and since $m_{nh} \leq m_o$, also $f^{nh}(m_o; m_{nh}) \leq f^{nh}(m_o; m_o)$; see Lemma 5.2. Let $m_{\infty,h}$ be the mode of $A_h * f_o$, analogous to the interpretation of m_{nh}. We distinguish between the cases $m_{\infty,h} > m_o$ and $m_{\infty,h} \leq m_o$.

Suppose $m_{\infty,h} > m_o$. Then

$$\varphi^{nh}(m_{nh}) = A_h * f_o(m_{nh}) + \delta_{nh} \geq A_h * dF_n(m_{nh}) = f^{nh}(m_{nh}; m_{nh}).$$

Moreover, φ^{nh} is increasing on $[m_{nh}, m_o]$. Thus, φ^{nh} is increasing on $(-\infty, m_o)$. Since it is decreasing on (m_o, ∞), then, as required, φ^{nh} is unimodal with mode at m_o.

Suppose $m_{\infty,h} \le m_o$. Since $A_h * f_o$ is decreasing on (m_o, ∞), we now have that

$$\begin{aligned} f^{nh}(m_o; m_{nh}) \le f^{nh}(m_o; m_o) &\le \max_{x \ge m_o} A_h * dF_n(x) \\ &\le \max_{x \ge m_o} A_h * dF_o(x) + \delta_{nh} \\ &= A_h * dF_o(m_o) + \delta_{nh} = \varphi^{nh}(m_o). \end{aligned}$$

At $x = m_{nh}$ we have

$$\varphi^{nh}(m_{nh}) = A_h * f_o(m_{nh}) + \delta_{nh} \ge A_h * dF_n(m_{nh}) = f^{nh}(m_{nh}; m_{nh}).$$

Now φ^{nh} is increasing on $(-\infty, m_{nh})$, decreasing on (m_o, ∞), and is unimodal on (m_{nh}, m_o). Thus, φ^{nh} is unimodal. $\square$

The remainder of this section consists of suitably bounding $m_{nh} - m_o$.

PROOF OF LEMMA 3.2. Let m_o be the point in $M(f_o)$ closest to m_{nh}, and set

$$\delta_{nh} = (1 + c_{nh}) \mid m_o - m_{nh} \mid \|A_h * (dF_n - dF_o)\|_\infty. \tag{6.14}$$

Then the triangle inequality and Lemma 6.3 imply

$$\int_{m_{nh}}^{m_o} \mid f_o(x) - f^{nh}(x) \mid dx \le \delta_{nh} + \int_{m_{nh}}^{m_o} \mid f_o(x) - A_h * f_o(x) \mid dx.$$

Since $(f_o)'' \in L^1(\mathbb{R})$, we have $(f_o)' \in L^\infty(\mathbb{R})$, and so for the last integral for suitable constants k and γ,

$$(m_o - m_{nh}) \|f_o - A_h * f_o\|_\infty \le k(m_o - m_{nh})h \| (f_o)' \|_\infty \le \gamma(m_o - m_{nh})h$$

[see, e.g., Devroye and Györfi (1985)]. If follows that

$$\int_{m_{nh}}^{m_o} \mid f_o(x) - f^{nh}(x) \mid dx \le \delta_{nh} + \gamma(m_o - m_{nh})h,$$

and then, as in (6.7),

$$\min_c \int_{m_{nh}}^{m_o} \mid f_o(x) - c \mid dx \le \delta_{nh} + \gamma(m_o - m_{nh})h. \tag{6.15}$$

Since $f_o' \le 0$ thenf_o is increasing on $[m_{nh}, m_o]$ and so the optimal c in (6.15) equals $c = f_o(\mu)$, with $\mu = \frac{1}{2}(m_{nh} + m_o)$. Then

$$\int_{m_{nh}}^{m_o} \mid f_o(x) - c \mid dx \ge \int_{\mu}^{m_o} (f_o(x) - f_o(\mu))\, dx.$$

Since f_o is concave on $[\mu, m_o]$, geometric considerations show that the last integral equals at least one-half the area of the rectangle with opposite vertices $(\mu, f_o(\mu))$ and $(m_o, f_o(m_o))$, or

$$\delta_{nh} + \gamma(m_o - m_{nh})h \geq \tfrac{1}{2}(m_o - \mu)(f_o(m_o) - f_o(\mu)).$$

With (6.14), assumption (3.11) implies that for all $\varepsilon > 0$, there exists a constant $c > 0$ such that

$$\exp\Big(-\frac{\varepsilon}{m_o - \mu}\Big) \leq c(\|A_h * (dF_n - dF_o)\|_\infty + \gamma h).$$

In view of the Silverman (1978) bound (3.13) for $h \asymp n^{-\beta}$, this implies asymptotically, for a suitable constant c',

$$m_o - \mu \leq_{\mathrm{as}} c'\varepsilon\,(\log n)^{-1}.$$

Since $\varepsilon > 0$ is arbitrary, the conclusion follows. $\square$

Acknowledgment. The authors thank the referee for insisting on a precise formulation and proof of Theorem 3.1.

REFERENCES

BARLOW, R. E., BARTHOLOMEW, D. J., BREMNER, J. M. and BRUNK, H. D. (1972). *Statistical Inference under Order Restrictions.* Wiley, New York.

BICKEL, P. J. and FAN, I. (1996). Some problems on the estimation of unimodal densities. *Statist. Sinica* **6** 23–45.

BIRGÉ, L. (1987a). Estimating a density under restrictions: nonasymptotic minimax risk. *Ann. Statist.* **15** 995–1012.

BIRGÉ, L. (1987b). On the risk of histograms for estimating decreasing densities. *Ann. Statist.* **15** 1113–1022.

BIRGÉ, L. (1989). The Grenander estimator: a nonasymptotic approach. *Ann. Statist.* **17** 1532–1549.

BIRGÉ, L. (1997). Estimation of unimodal densities without smoothness assumptions. *Ann. Statist.* **25** 970–981.

BRUNK, H. D. (1965). Conditional expectation given a σ lattice and applications. *Ann. Math. Statist.* **36** 1339–1350.

CHERNOFF, H. (1964). Estimation of the mode. *Ann. Inst. Statist. Math.* **16** 31–41.

DEVROYE, L. (1987). *A Course in Density Estimation.* Birkhäuser, Boston.

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 GYÖRFI, L. (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.

EDDY, W. F. (1980). Optimal kernel estimators of the mode. *Ann. Statist.* **8** 870–882.

EGGERMONT, P. P. B. and LARICCIA, V. N. (1995). Maximum smoothed likelihood density estimation. *J. Nonparametr. Statist.* **4** 211–222.

EGGERMONT, P. P. B. and LARICCIA, V. N. (1999). Best asymptotically normal kernel density entropy estimators for smooth densities. *IEEE Trans Inform. Theory* **45** 1321–1326.

FOUGÈRES, A.-L. (1997). Estimation de densités unimodales. *Canad. J. Statist.* **25** 375–387.

GRENANDER, U. (1956). On the theory of mortality measurements II. *Skand. Akt.* **39** 125–153.

GROENEBOOM, P. (1985). Estimating a monotone density. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (L. M. Le Cam and R. A. Olshen, eds.) 539–555. Wadsworth, Belmont, CA.

GROENEBOOM, P., HOOGHIEMSTRA, G. and LOPUHAÄ, H. P. (1999). Asymptotic normality of the L_1 error of the Grenander estimator. *Ann. Statist.* **27** 1316–1347.

GRUND, B. and HALL, P. (1995). On the minimisation of L^p error in mode estimation. *Ann. Statist.* **23** 2264–2284.

HALL, P. and WEHRLY, T. E. (1991). A geometrical method for removing edge effects from kernel-type nonparametric regression estimators. *J. Amer. Statist. Assoc.* **86** 665–672.

HARDY, G. H., LITTLEWOOD, G. E. and PÓLYA, G. (1951). *Inequalities.* Cambridge Univ. Press.

IBRAGIMOV, I. A. (1956). On the composition of unimodal distributions. *Theory Probab. Appl.* **1** 255–260.

JONES, M. C. (1993). Simple boundary corrections for kernel density estimation. *Statist. Comput.* **3** 135–146.

KEMPERMAN, J. H. B. (1967). On the optimum rate of transmitting information. *Lecture Notes in Math.* **23** 126–169. Springer, Berlin.

LIEB, E. H. and LOSS, M. (1996). *Analysis.* Amer. Math. Soc., Providence, RI.

MAMMEN, E. (1991). Estimating a smooth monotone regression function. *Ann. Statist.* **19** 724–740.

MARSHALL, A. W. (1970). Discussion on Barlow and van Zwet's paper. In *Nonparametric Techniques in Statistical Inference* (M. L. Puri, ed.) 175–176. Cambridge Univ. Press.

MÜLLER, H.-G. (1993). On the boundary kernel method for nonparametric curve estimation near endpoints. *Scand. J. Statist.* **20** 313–328.

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

SILVERMAN, B. W. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Ann. Statist.* **6** 177–184.

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

WEGMAN, E. I. (1970). Maximum likelihood estimation of a unimodal density II. *Ann. Math. Statist.* **6** 2169–2174.

WHEEDEN, R. and ZYGMUND, A. (1977). Measure and Integral. North-Holland, Amsterdam.

WOODROOFE, M. and SUN, J. (1993). A penalized maximum likelihood estimate of $f(0+)$ when f is non-increasing *Statist. Sinica* **3** 501–515.

FOOD AND RESOURCE ECONOMICS
UNIVERSITY OF DELAWARE
NEWARK, DELAWARE 19716
E-MAIL: eggermon@udel.edu