A UNIVERSALLY ACCEPTABLE SMOOTHING FACTOR FOR KERNEL DENSITY ESTIMATES

Luc Devroye
School of Computer Science
McGill University
Montreal, Canada H3A 2A7

and

Gábor Lugosi
Department of Mathematics and Computer Science
Technical University of Budapest
Stoczek u. 2-4
Budapest, H-1521 Hungary

ABSTRACT. We define a minimum distance estimate of the smoothing factor for kernel density estimates, based upon a methodology first developed by Yatracos (1985). It is shown that if f_{nh} denotes the kernel density estimate on $\mathbb{R}^d$ for an i.i.d. sample of size n drawn from an unknown density f, where h is the smoothing factor, and if f_n is the kernel estimate with the same kernel and with the proposed new data-based smoothing factor, then, under a regularity condition on the kernel K,

$$\sup_f \limsup_{n\to\infty} \frac{\mathbf{E}\int |f_n - f|\,dx}{\inf_{h>0} \mathbf{E}\int |f_{nh} - f|\,dx} \le 3\ .$$

This is the first published smoothing factor that can be proven to have this property.

KEYWORDS AND PHRASES. Density estimation, kernel estimate, convergence, smoothing factor, minimum distance estimate, asymptotic optimality.

1991 MATHEMATICS SUBJECT CLASSIFICATIONS: Primary 62G05.

RUNNING HEAD: UNIVERSAL SMOOTHING FACTOR

The first author's work was supported by NSERC Grant A3456 and by FCAR Grant 90-ER-0291. The second author's work was supported by OTKA Grant F 014174.

1. Introduction.

We are given an i.i.d. sample $X_1, \ldots, X_n$ drawn from an unknown density f on $\mathbb{R}^d$. We consider the Akaike-Parzen-Rosenblatt density estimate

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

where $K \geq 0$ is a fixed density (the kernel), $K_h(x) = (1/h^d)K(x/h)$, and $h > 0$ is the smoothing factor (Akaike, 1954; Parzen, 1962; Rosenblatt, 1956). Much ink has been spilled regarding the choice of h as a function of the data (for surveys, see Devroye and Györfi (1985), Marron (1988, 1989a), Park and Turlach (1992), Turlach (1993), Cao, Cuevas and González-Manteiga (1994), or Berlinet and Devroye (1994)). Despite the flurry of activity, one has not been able to date to exhibit a single data-dependent smoothing factor H (in which the dependence upon $X_1, \ldots, X_n$ is dropped) for which for a finite constant γ,

$$\sup_f \limsup_{n \to \infty} \frac{\mathbf{E} \int |f_{nH} - f|}{\inf_h \mathbf{E} \int |f_{nh} - f|} \leq \gamma \ .$$

This problem has been mentioned and discussed in the introduction of Devroye (1987) and in the more recent papers of Berlinet and Devroye (1994), and Devroye (1989, 1994). Particular choices of smoothing factors may do if the supremum above is restricted to a subclass of well-behaved densities. In those cases, one often has $\gamma = 1$. For example, if f is restricted to a class of univariate densities in which only a translation and scale parameter is unknown, using $h = a_n \hat{\sigma}$ for a function a_n (depending upon the family), where $\hat{\sigma}$ is a data-based estimate of the scale factor, will do (see Deheuvels (1977a, 1977b) or Deheuvels and Hominal, 1980). The smoothing factor h can also be based upon a plug-in of estimates of unknown functionals into a given formula. This method has the given property if the supremum is taken over classes of univariate densities restricted by smoothness and small tails (Hall and Wand, 1988). The double kernel estimate (Devroye, 1989) satisfies the property mentioned above when the supremum is restricted as in the work of Hall and Wand. Except for trivially restricted classes of densities, none of the L_2-cross-validated estimates in the literature (see Rudemo (1982), Bowman (1984) or Stone (1984) for the early papers on this) possesses the property mentioned above.

In this paper, we present the first smoothing factor that is known to be universally expedient in the sense defined above. The estimate may not be best possible by other criteria, and we are sure improvements will follow soon. Nevertheless, we believe that the mere existence of such a smoothing factor is worth reporting.

2. Relationship with minimax theory.

The performance of an estimate depends upon f. It is quite a task to compare estimates with one another, because of this dependence. To aid in this task, one could consider the minimax error

$$M_n(\mathcal{F}) \stackrel{\text{def}}{=} \inf_{\widehat{f}} \sup_{f \in \mathcal{F}} \mathbf{E} \int |\widehat{f} - f| \ ,$$

where $\widehat{f}$ is an estimate and $\mathcal{F}$ is a given class of densities. The error $M_n(\mathcal{F})$ is the error any estimate has to make on at least one density in $\mathcal{F}$. Unfortunately, if $\mathcal{F}$ is too large, $M_n(\mathcal{F})$ does not tend to zero with n. Examples are given in the minimax chapters of Devroye and Györfi (1985) or Devroye (1987). They include the class of all densities on $[0,1]$ bounded by 2, or all the unimodal densities with infinitely many absolutely continuous derivatives, or the class of all monotone densities on $[0,\infty)$ bounded by 1. The same is true for all convex-shaped densities on $[0,1]$, or the class of all Lipschitz densities on the real line with given Lipschitz constant. The class of all densities that are normal scale mixtures is also rich in the sense that $\liminf_{n\to\infty} M_n(\mathcal{F}) > 0$. For smaller classes $\mathcal{F}$, $M_n(\mathcal{F})$ tends to zero and one may meaningfully look for estimates f_n for which

$$\sup_{f \in \mathcal{F}} \mathbf{E} \int |f_n - f| \approx M_n(\mathcal{F})$$

where $\approx$ means either "$\sim$" or "$= O(.)$". For key minimax bounds, see e.g. Bretagnolle and Huber (1979), Ibragimov and Khasminskii (1982), Birgé (1985, 1986, 1987a, 1987b, 1989), and Assouad (1983). A survey and additional results can be found in Devroye (1987) and Hall (1989). Donoho, Kerkyacharian and Picard (1995) review more recent work.

Estimates that achieve the above minimax optimality may not necessarily adapt well to all densities in the class. For adaptation, we may follow ideas along the lines of those shown for example by Low (1992), and look at ratios like

$$\sup_{\mathcal{F} \in \mathcal{G}} \limsup_{n\to\infty} \frac{\sup_{f\in\mathcal{F}} \mathbf{E} \int |f_n - f|}{\inf_{\widehat{f}} \sup_{f\in\mathcal{F}} \mathbf{E} \int |\widehat{f} - f|}$$

where $\mathcal{G}$ is a class of classes. For example, if $\mathcal{F}$ is the class of Lipschitz densities with constant C and support on $[0,1]$, $\mathcal{G}$ may the class of all $\mathcal{F}$'s with all possible values of C. If the ratio above remains bounded, f_n truly "adapts" to all classes $\mathcal{F}$ in $\mathcal{G}$. Even though every $\mathcal{F}$ must be "small", (by our remark above), if we nest them, the union of $\mathcal{F}$'s may be sufficiently large, but we won't really get a lot of information for the behavior with respect to all densities.

In this paper, we really want to focus on density estimation without restrictions on the densities. As we will see, this can be done if we concentrate on a particular class of estimates, such as the kernel estimates (which are known to be optimal in several senses, and are optimal for the minimax error over several smoothness classes). The only problem with the kernel estimate is the selection of the bandwidth h. If we let $\mathcal{F}$ denote the class of all densities on $\mathbb{R}^d$, and let f_{nH} denote a kernel estimate with data-based bandwidth H, we look at

$$\sup_{f \in \mathcal{F}} \limsup_{n \to \infty} \frac{\mathbf{E} \int |f_{nH} - f|}{\inf_h \mathbf{E} \int |f_{nh} - f|} \ .$$

The boundedness of this supremum shows that the bandwidth selector works well for all f, without exception. As pointed out by a referee, there is a much stronger uniform criterion one might want to keep bounded,

$$\limsup_{n \to \infty} \sup_{f \in \mathcal{F}} \frac{\mathbf{E} \int |f_n - f|}{\inf_h \mathbf{E} \int |f_{nh} - f|} \ .$$

This criterion is very strong as it measures how well the bandwidth selector adapts to every density in f uniformly over all f. For the estimate given in this paper, this limit supremum is ∞ because several parameters of the bandwidth selector including a given interval range $[a_n, b_n]$ (defined below) are picked as a function of n and not as a function of the data, as they should. Additional adaptation (and further analysis) is necessary. Thus, this paper does not answer the question of the boundedness of the last limit supremum. However, if $\mathcal{F}$ is a given small subclass, the limit supremum is known to stay bounded. All that is needed are precise uniform performance bounds on the L_1 error. Under sufficient smoothness and tail assumptions on f, the behavior of the L_1 error is well-known. Bandwidth selectors for which the last limit supremum remains bounded for such classes may be based on plug-in (Hall and Wand, 1988) or on the double kernel method (Devroye, 1989). For the early work along these lines, see the adaptive bandwidth selector of Bretagnolle and Huber (1977).

3. The estimate.

In our estimate, we split the data set into a test set of size $m \ll n$ and a remainder. Then we use Yatracos' minimum distance projection of the empirical measure based upon these m points to the class of densities defined by the kernel estimates based on the remaining $n-m$ points to find an optimal h (Yatracos, 1985).

Let $m < n$ be a positive integer, let K be a nonnegative kernel with $\int K(x)dx = 1$, and let $\mathcal{F}_n$ be the class of densities

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

with $h \in [a_n, b_n]$, where the nonnegative numbers a_n, b_n will be specified later such that the optimal smoothing factor eventually falls in $[a_n, b_n]$ for all densities. Next we cover the class $\mathcal{F}_n$ by finitely many densities as follows: let $\delta_n > 0$ be a parameter to be specified later, let $h_1 = a_n$, and $h_i = h_{i-1}(1+\delta_n)$ for all $i = 2, \ldots, N$, where N is the largest integer with $a_n(1+\delta_n)^{N-1} \le b_n$. The finite class of densities $\{f_{n-m,h_\mathtt{i}} : i = 1, \ldots, N\}$ is denoted by $\mathcal{G}_n$.

Let μ_m be the empirical measure defined by the rest of the data points: $X_{n-m+1}, \ldots, X_n$, i.e., for any Borel set $A \subseteq \mathbb{R}^d$,

$$\mu_m(A) = \frac{1}{m} \sum_{i=n-m+1}^{n} I_A(X_i),$$

where I_A denotes the indicator function of A. As is well-known, the L_1 distance is equivalent to the twice the total variation distance. If we are to use the empirical measure, we would thus be tempted to select h so as to minimize the total variation

$$T \overset{\text{def}}{=} \sup_A \left| \int_A f_{n-m,h} - \mu_m(A) \right| .$$

As μ_m is an atomic measure, $T \equiv 2$ for all h. Following a clever idea of Yatracos (1985), we take the supremum instead over a specially picked rich class of subsets $\mathcal{A}$, defined as the family of sets

$$\{x : f_{n-m,h_\mathtt{i}}(x) > f_{n-m,h_\mathtt{j}}(x)\}, \quad i,j \le N.$$

The estimate f_n is defined to be that $f_{n-m,h_\mathtt{i}} \in \mathcal{G}_n$ for which

$$\sup_{A \in \mathcal{A}} \left| \int_A f_{n-m,h_\mathtt{i}} - \mu_m(A) \right|$$

is minimal. If the minimum is not unique, we choose among the minimizing densities according to a prespecified rule, e.g., we choose the one with smallest index. Note that in any case, our estimate optimizes over a given finite set, and is thus defined with computional efficiency in mind.

Choice of the parameters. It helps at this stage to pin down choices for a_n, b_n and δ_n. We stress that our choices are surely not the only ones—they make the remainder a bit more readable. The choice is determined by the choice of the kernel K. We assume the following: a kernel is said to be *elegant* if it is nonnegative, if it is Lipschitz of constant C (i.e., $|K(x) - K(y)| \le C\|x - y\|$ for all x, y), and if $K = 0$ outside $[-1, 1]^d$. Then define $a_n = e^{-n}$, $b_n = e^n$, $\delta_n = c/\sqrt{n}$ for a fixed constant c. This class contains the standard Deheuvels kernel that is optimal in $\mathbb{R}^d$ and is of the form $C'(1 - \|x\|^d)_+$. For more general classes, we will show how to take the parameters in remarks below. □

A computational remark. In most univariate cases, the sets A above are finite unions of intervals. The number of such intervals can be rigorously controlled if the kernel is polynomial on a compact set (such as with the celebrated Epanechnikov-Bartlett kernel $3/4(1 - x^2)_+$). The computations are much more involved for $d > 1$, unless K is the indicator function of a unit square. The class $\mathcal{A}$ has N^2 members. A quick calculation shows that

$$N - 1 \le \frac{\log(b_n/a_n)}{\log(1 + \delta_n)} \le \frac{n(2 + \delta_n)}{\delta_n} = n + 2n^{3/2}/c \; .$$

A lower bound on the number of integrals over sets A (if we were to naively minimize) would be of the order of n^3. However, clever shortcuts are possible. □

The set $\mathcal{A}$. The set $\mathcal{A}$ cannot be replaced by the set of all rectangles of $\mathbb{R}^d$. This class is simply not rich enough, and Lemma 2 below would not be valid. □

4. The main result.

THEOREM. *Let K be an elegant kernel. Let a_n, b_n be such that $na_n \to 0$ and $b_n \to \infty$. Assume that $\delta_n = c/\sqrt{n}$ for some constant c and that $\log(b_n/a_n) \le c'n^a$ for some finite $c', a > 0$. If*

$$\frac{m}{n} \to 0 \quad \text{and} \quad \frac{m}{n^{4/5}\log n} \to \infty \quad \text{as } n \to \infty,$$

then the estimate f_n defined above satisfies

$$\sup_f \limsup_{n\to\infty} \frac{\mathbf{E}\int |f_n - f|}{\inf_h \mathbf{E}\int |f_{n,h} - f|} \le 3.$$

This result is valid for any multivariate density. It may be possible to improve the constant in the bound. With a bit of work, one may also be able to replace f_n in the result by f_{nH}, where H is the smoothing factor used in f_n. The difference here is that f_{nH} uses all n data points, while f_n is the kernel estimate based on H and $X_1, \ldots, X_{n-m}$.

One may argue that the selected smoothing factor is not scale-invariant. This is easily taken care of by letting M_n denote the median of the $\binom{n-m}{2}$ distances $\|X_i - X_j\|$, $1 \le i, j \le n - m$, and setting $a_n = M_n e^{-n}$ and $b_n = M_n e^n$. As M_n is almost surely bounded away from 0 and infinity, one can verify that the Theorem holds for this choice of interval. By not letting M_n depend upon $X_{n-m+1}, \ldots, X_n$, we are not jeopardizing Lemmas 1 through 3 below. The remaining changes in the proof are minor and are left to the reader.

For convenience we assumed that the kernel K is nonnegative. It is well known, however, that some kernels taking negative values provide smaller L_1 errors for smooth densities. The above theorem is easily extended to such kernels at the expense of further restrictions on the growth of m, depending on the order of the kernel. The key properties required are those analogous to Lemmas 1 and 5. Lemma 4 applies to bounded compact support negative kernels as well. With these results in hand (see, e.g., Devroye, 1988, for a negative kernel version of Lemma 5), the proof may easily be modified to generalize the theorem for such kernels. We do not pursue this question further, as this will add unnecessary noise to the development of the ideas.

Finally, there is quite a bit of freedom in the choice of all the parameters. For example, δ_n does not have to tend to zero at the rate $1/\sqrt{n}$. Various constants also need to be picked. For example, suitable data-based choices for a_n and b_n are necessary. It is clear that further analysis is needed to pin the parameters down. We are afraid however that this can only be done by putting conditions on the densities, and this is what we tried to avoid in the first place.

We prove the Theorem by building on a series of Lemmas.

5. Auxiliary results.

We begin with a property of elegant kernels.

LEMMA 1. *If K is an elegant kernel, then for any $f_{n-m,h} \in \mathcal{F}_n$, there is h_i, $i \in \{1,\ldots,N\}$ such that*

$$\int |f_{n-m,h} - f_{n-m,h_{\mathtt{i}}}| \le C'\delta_n \ ,$$

where $C' = C2^d\sqrt{d} + d$.

PROOF. Clearly, for $u > v > 0$, $u/v = b > 1$,

$$\begin{aligned} |f_{n-m,u} - f_{n-m,v}| &\le \int |K_u - K_v| = \int |K - K_{v/u}| = \int \left|K(y) - b^d K(by)\right| dy \\ &\le \int |K(y) - K(by)|\, dy + \int \left|K(by) - b^d K(by)\right| dy \\ &\le C2^d\sqrt{d}|1-b| + d|1-b|/b \\ &\le C'|1-b| \ . \end{aligned}$$

Let h_i be the smallest of the h_j's at least equal to h. Then by the above,

$$\int |f_{n-m,h} - f_{n-m,h_{\mathtt{i}}}| \le C' \left|\frac{h_i - h}{h}\right| \le C'\delta_n \ . \square$$

We have the following crucial property of the estimate:

Lemma 2. *For each density f and elegant kernel K, if $\delta_n = c/\sqrt{n}$, $C' = C2^d\sqrt{d} + d$, and $f_n \in \mathcal{G}_n$,*

$$\int |f_n - f| \le 3 \inf_{h\in[a_n,b_n]} \int |f_{n-m,h} - f| + \frac{5C'c}{\sqrt{n}} + 4 \sup_{A\in\mathcal{A}} \left| \mu_m(A) - \int_A f \right| .$$

Proof. $\mathcal{G}_n$ is an infinite class of densities (kernel estimates) covered by a finite number (N) of L_1 balls of radius $C'\delta_n$ centered at the densities in $\mathcal{F}_n$ (Lemma 1). A lemma of Yatracos (1985) (see Lemma 6.1 in Devroye, 1987) states that for any pair f_{nh} and $f_{nh^{\odot}}$ in $\mathcal{G}_n$, we must have

$$\int |f_{nh} - f_{nh^{\odot}}| \le 4C'\delta_n + 2 \sup_{A\in\mathcal{A}} \left| \int_A f_{nh} - \int_A f_{nh^{\odot}} \right| .$$

Let $\bar{f}$ be a member of $\mathcal{F}_n$ such that

$$\int |f - \bar{f}| \le \int |f - g| \qquad \text{for all } g \in \mathcal{F}_n.$$

Then

$$\begin{aligned}
\int |f_n - f| &\le \int \left|f - \bar{f}\right| + \int \left|f_n - \bar{f}\right| \\
&\le \int \left|f - \bar{f}\right| + 4C'\delta_n + 2\sup_{A\in\mathcal{A}} \left| \int_A f_n - \int_A \bar{f} \right| \\
&\le \int \left|f - \bar{f}\right| + 4C'\delta_n + 2\sup_{A\in\mathcal{A}} \left| \int_A f_n - \mu_m(A) \right| + 2\sup_{A\in\mathcal{A}} \left| \mu_m(A) - \int_A \bar{f} \right| \\
&\le \int \left|f - \bar{f}\right| + 4C'\delta_n + 2\sup_{A\in\mathcal{A}} \left| \int_A \widehat{f} - \mu_m(A) \right| + 2\sup_{A\in\mathcal{A}} \left| \mu_m(A) - \int_A \bar{f} \right|
\end{aligned}$$

(by the definition of f_n, where $\widehat{f} \in \mathcal{G}_n$ satisfies
$2\sup_{A\in\mathcal{A}} \left| \int_A \widehat{f} - \int_A \bar{f} \right| \le C'\delta_n$; Lemma 1 guarantees that
such an $\widehat{f}$ exists, since by Scheffé's theorem
$2\sup_{A\in\mathcal{A}} \left| \int_A \widehat{f} - \int_A \bar{f} \right| \le \int |\widehat{f} - \bar{f}|$)

$$\le \int \left|f - \bar{f}\right| + 4C'\delta_n + 2\sup_{A\in\mathcal{A}} \left| \int_A \widehat{f} - \int_A \bar{f} \right| + 4\sup_{A\in\mathcal{A}} \left| \mu_m(A) - \int_A \bar{f} \right|$$

(by the triangle inequality)

$$\le \int \left|f - \bar{f}\right| + \frac{5C'c}{\sqrt{n}} + 4\sup_{A\in\mathcal{A}} \left| \mu_m(A) - \int_A f \right| + 4\sup_{A\in\mathcal{A}} \left| \int_A \bar{f} - \int_A f \right|$$

(by the above property of $\widehat{f}$ and by the triangle inequality)

$$\le 3 \inf_{h\in[a_n,b_n]} \int |f - f_{n-m,h}| + \frac{5C'c}{\sqrt{n}} + 4\sup_{A\in\mathcal{A}} \left| \mu_m(A) - \int_A f \right|$$

(by Scheffé's theorem),

proving Lemma 2. □

Lemma 3. *If $\delta_n = c/\sqrt{n}$ and $\log(b_n/a_n) \le c'n^a$ for positive constants c, c', a, then, if K is an elegant kernel,*

$$\mathbf{E}\int |f_n - f| \le 3\mathbf{E}\left\{\inf_{h\in[a_n,b_n]}\int |f_{n-m,h} - f|\right\} + \frac{5C'c}{\sqrt{n}} + \frac{8\sqrt{\log c'' + (2a+1)\log n} + 1}{\sqrt{2m}},$$

where $c'' = (1 + c'/2 + c'/c)^2$, and $C' = C2^d\sqrt{d} + d$.

Proof. The class of sets $\mathcal{A}$ has not more than N^2 members, where

$$N - 1 \le \frac{\log(b_n/a_n)}{\log(1+\delta_n)} \le \frac{c'n^a(2+\delta_n)}{2\delta_n} = \frac{c'n^a}{2} + \frac{c'n^{a+1/2}}{c} \ .$$

Thus,

$$N^2 \le (1 + c'/2 + c'/c)^2 n^{2a+1} \stackrel{\text{def}}{=} c''n^{2a+1} \ .$$

Thus, for each $t > 0$,

$$\begin{aligned}\mathbf{P}\left\{\sup_{A\in\mathcal{A}}\left|\mu_m(A) - \int_A f\right| > t\right\} &\le c''n^{2a+1}\sup_{A\in\mathcal{A}}\mathbf{P}\left\{\left|\mu_m(A) - \int_A f\right| > t\right\} \\ &\le 2c''n^{2a+1}e^{-2mt^2}\end{aligned}$$

by Hoeffding's inequality (Hoeffding, 1963). By the inequality $\mathbf{E}Z \le u + \int_u^\infty \mathbf{P}\{Z > t\}dt$ for a nonnegative random variable Z and $u > 0$, standard bounding of the integral, and optimizing for u, we obtain

$$\mathbf{E}\left\{\sup_{A\in\mathcal{A}}\left|\mu_m(A) - \int_A f\right|\right\} \le \frac{2\sqrt{\log c'' + (2a+1)\log n} + 1}{\sqrt{2m}},$$

which in conjunction with Lemma 2 proves Lemma 3. □

We need two fundamental results on kernel estimates.

Lemma 4 (Devroye, 1983). *Assume $K \ge 0$. If $\mathbf{E}\int |f_{nh} - f| \to 0$ for some density f and some sequence h, then $h \to 0$ and $nh^d \to \infty$. Conversely, if $h \to 0$ and $nh^d \to \infty$, then $\mathbf{E}\int |f_{nh} - f| \to 0$ for all densities f.*

Lemma 5 (Devroye and Penrod, 1984). *Assume $K \geq 0$. Then*

$$\inf_f \liminf_{n\to\infty} n^{2/5} \inf_h \mathbf{E} \int |f_{nh} - f| \geq 0.86.$$

We note that Lemma 5 was only proved in the cited paper for $d = 1$. However, if f and f_{nh} are a density and a kernel density estimate on $\mathbb{R}^d$, and if g and g_{nh} denote the marginal densities for f and f_{nh} (with respect to any fixed component), then,

$$\mathbf{E} \int |g_{nh} - g| \leq \mathbf{E} \int |f_{nh} - f|$$

(Devroye and Györfi, 1985). Interestingly, g_{nh} itself is a valid univariate kernel estimate with as kernel the marginal density of the original kernel. Therefore, a universal lower bound for $d = 1$ of the type shown in Lemma 5 then applies equally for all dimensions d.

Lemma 6. *Let X and Y be independent random variables, and let $\mathbf{E}Y = 0$. Then $\mathbf{E}|X + Y| \geq \mathbf{E}|X|$.*

Lemma 7 (Devroye and Györfi, 1985, page 137). *Let $Y_1, \ldots, Y_n$ be i.i.d. zero mean random variables. Then*

$$\mathbf{E}\left\{ \left| \sum_{i=1}^n Y_i \right| \right\} \geq \sqrt{\frac{n}{8}} \mathbf{E}|Y_1| \ .$$

The final missing piece is perhaps the most important one. Define

$$J_{nh} \stackrel{\text{def}}{=} \int |f_{nh} - f| \ .$$

It is absolutely necessary to have good universal bounds on the oscillation of $\inf_{h>0} \mathbf{E}J_{nh}$ as a function of n. The next Lemma shows that this infimum is indeed very stable. For fixed h, it is known that $\mathbf{E}J_{nh}$ is nonincreasing in n (Devroye and Györfi, 1985, page 282). Therefore, $\inf_{h>0} \mathbf{E}J_{nh}$ is nonincreasing in n. Still, as a function of n, the behavior could possibly be erratic; there could for example be sudden big relative drops. In fact, this does not happen for any density! Lemma 8 may seem to be obvious to those with deep intuitions, but it certainly is not a superficial property. It is this Lemma that allows us to get away with the split sample methodology explained earlier, as for $m = o(n)$, $\inf_h \mathbf{E}J_{n-m,h}$ is close to $\inf_h \mathbf{E}J_{n,h}$. We have the following universal inequality, valid for any density f and any dimension d:

LEMMA 8. *Let K be a bounded kernel. If $m > 0$ is a positive integer such that $2m \le n$, then*

$$1 \le \frac{\inf_h \mathbf{E}J_{n-m,h}}{\inf_h \mathbf{E}J_{n,h}} \le 1 + \frac{2m}{n-m} + 8\sqrt{\frac{m}{n}} \ .$$

PROOF. As the given ratio in the lemma is at least one, by monotonicity, we need only look at upper bounds for the ratio. Note the following:

$$\begin{aligned}\inf_{h>0} \mathbf{E}J_{n-m,h} &\le \inf_{h>0} \mathbf{E}J_{n,h} \times \sup_{h>0}\left(\frac{\mathbf{E}J_{n-m,h}}{\mathbf{E}J_{n,h}}\right) \\ &= \inf_{h>0} \mathbf{E}J_{n,h} \times \left(1 + \sup_{h>0} \frac{\mathbf{E}J_{n-m,h} - \mathbf{E}J_{nh}}{\mathbf{E}J_{n,h}}\right) \ .\end{aligned}$$

The supremum is rewritten as follows:

$$\sup_{h>0} \frac{\mathbf{E}J_{n-m,h} - \mathbf{E}J_{nh}}{\mathbf{E}J_{n,h}} \le \sup_{h>0} \frac{\mathbf{E}\int |f_{n-m,h} - f_{nh}|\,dx}{\mathbf{E}J_{n,h}} \le 2\sup_{h>0} \frac{\mathbf{E}\int |f_{n-m,h} - f_{nh}|\,dx}{\mathbf{E}\int |f_{nh} - \mathbf{E}f_{nh}|\,dx} \ ,$$

where we used a simple bound from page 23 of Devroye and Györfi (1986) (see also Devroye, 1989). Fix x and h for now. Introduce $Y_i = K_h(x - X_i) - \mathbf{E}K_h(x - X_i)$, and denote the partial sums of Y_i's by $S_j = Y_1 + \cdots + Y_j$. We will need the existence for fixed x and h of the first absolute moment of Y_1. That is satisfied for any bounded kernel. Then observe the following:

$$n|f_{n-m,h} - f_{nh}| = \left|\frac{m}{n-m}(Y_1 + \cdots + Y_{n-m}) - (Y_{n-m+1} + \cdots + Y_n)\right|$$

so that

$$\mathbf{E}\left\{n|f_{n-m,h} - f_{nh}|\right\} \le \frac{m}{n-m}\mathbf{E}|S_{n-m}| + \mathbf{E}|S_m| \ .$$

Also, $n|f_{nh} - \mathbf{E}f_{nh}| = |S_n|$, which implies $\mathbf{E}\left\{n|f_{nh} - \mathbf{E}f_{nh}|\right\} = \mathbf{E}|S_n|$. Still holding x and h fixed, we bound the following ratio:

$$\begin{aligned}\frac{\mathbf{E}|f_{n-m,h} - f_{nh}|}{\mathbf{E}|f_{nh} - \mathbf{E}f_{nh}|} &\le \frac{\frac{m}{n-m}\mathbf{E}|S_{n-m}| + \mathbf{E}|S_m|}{\mathbf{E}|S_n|} \\ &\le \frac{m}{n-m} + \frac{\mathbf{E}|S_m|}{\mathbf{E}|S_n|} \quad \text{(because } \mathbf{E}|S_n| \ge \mathbf{E}|S_{n-m}|\text{)} \\ &\le \frac{m}{n-m} + \frac{\mathbf{E}|S_m|}{\sqrt{\frac{\lfloor n/m \rfloor}{8}}\mathbf{E}|S_m|} \quad \text{(by Lemmas 6 and 7)} \\ &\le \frac{m}{n-m} + 4\sqrt{\frac{m}{n}} \quad \text{(if } 2m \le n\text{)} \ .\end{aligned}$$

This implies that for any fixed h,

$$\mathbf{E}\int |f_{n-m,h} - f_{nh}|\,dx \le \left(\frac{m}{n-m} + 4\sqrt{\frac{m}{n}}\right)\mathbf{E}\int |f_{nh} - \mathbf{E}f_{nh}|\,dx\ .$$

Lemma 8 now follows without work. □

6. Proof of the Theorem.

Write

$$\frac{\mathbf{E}\int |f_n - f|}{\inf_h \mathbf{E}J_{n,h}} = \frac{\mathbf{E}\int |f_n - f|}{\inf_{h\in[a_n,b_n]} \mathbf{E}J_{n-m,h}} \times \frac{\inf_{h\in[a_n,b_n]} \mathbf{E}J_{n-m,h}}{\inf_h \mathbf{E}J_{n-m,h}} \times \frac{\inf_h \mathbf{E}J_{n-m,h}}{\inf_h \mathbf{E}J_{n,h}}$$
$$= I \times II \times III.$$

From Lemma 8 we immediately have that $III \to 1$. From Lemma 4, any smoothing factor h' that minimizes $\mathbf{E}J_{n,h}$ satisfies $h' \to 0$ and $nh'^d \to \infty$, so that $II \to 1$. Use Lemmas 3 and 5, and the condition $m/(n^{4/5}\log n) \to \infty$ to verify that

$$\mathbf{E}\int |f_n - f| \le (3 + o(1)) \inf_{h\in[a_n,b_n]} \mathbf{E}J_{n-m,h}\ ,$$

and to conclude that $\limsup_{n\to\infty} I \le 3$. The theorem is now proved. □

7. Supplements.

The Theorem deals with $f_n = f_{n-m,H}$, not f_{nH}. In other words, the estimate does not use the full sample. In this section, we show that f_{nH} and $f_{n-m,H}$ are asymptotically equivalent for all densities. We will show the following.

Proposition. *Let K be an elegant kernel. Let a_n, b_n be such that $na_n \to 0$ and $b_n \to \infty$. Assume that $\delta_n = c/\sqrt{n}$ for some constant c and that $\log(b_n/a_n) \le c'n^a$ for some finite $c', a > 0$. If*

$$\frac{m}{n} \to 0 \quad \textit{and} \quad \frac{m}{n^{4/5}\log n} \to \infty \quad \textit{as } n \to \infty,$$

then

$$\sup_f \limsup_{n\to\infty} \frac{\mathbf{E}\int |f_{nH} - f|}{\inf_h \mathbf{E}\int |f_{n,h} - f|} \le 3.$$

To prove the Proposition, we require one lemma.

Lemma 9. *Let $J_{nh}=\int|f_{nh}-f|$ and $J_{n-m,h}=\int|f_{n-m,h}-f|$. Then, if $h_1,\ldots,h_N$ are N arbitrary smoothing factors,*

$$\mathbf{E}\left\{\max_{1\le i\le N}|J_{nh_i}-\mathbf{E}J_{nh_i}|\right\}\le\sqrt{\frac{2+2\log 2N}{n}}\ .$$

Similarly,

$$\mathbf{E}\left\{\max_{1\le i\le N}|J_{n-m,h_i}-\mathbf{E}J_{n-m,h_i}|\right\}\le\sqrt{\frac{2+2\log 2N}{n-m}}\ .$$

Proof. We begin with the following exponential inequality from Devroye (1991) (which in turn was a rather straightforward application of McDiarmid's inequality, see McDiarmid (1989)):

$$\mathbf{P}\left\{|J_{nh}-\mathbf{E}J_{nh}|>u\right\}\le 2e^{-nu^2/2}\ ,\ u>0\ .$$

From this, we have

$$\mathbf{P}\left\{\max_{1\le i\le N}|J_{nh_i}-\mathbf{E}J_{nh_i}|>u\right\}\le\min\left(1,2Ne^{-nu^2/2}\right)\ ,\ u>0\ .$$

For a positive random variable Z, we have, for $v>0$,

$$\mathbf{E}Z\le\sqrt{\mathbf{E}Z^2}=\sqrt{2\int_0^\infty u\mathbf{P}\{Z>u\}\,du}\le\sqrt{v^2+2\int_v^\infty u\mathbf{P}\{Z>u\}\,du}\ .$$

Replace Z by the maximum we are dealing with, and set $v^2=(2\log 2N)/n$. Deduce that

$$\begin{aligned}\mathbf{E}^2\left\{\max_{1\le i\le N}|J_{nh_i}-\mathbf{E}J_{nh_i}|\right\}&\le\frac{2\log 2N}{n}+2\int_v^\infty 2Nue^{-nu^2/2}\,du\\&=\frac{2\log 2N}{n}+\frac{2}{n}\ .\end{aligned}$$

This concludes the proof of Lemma 9. □

Proof of the Proposition. Let N be the cardinality of $\mathcal{A}$ and let $h_1,\ldots,h_N$ be the smoothing factors from among which H is selected. We use the notation J_{nh} introduced earlier. Note the following:

$$\begin{aligned}J_{nH}&=\sum_{i=1}^N J_{nh_i}I_{H=h_i}\\&=\sum_{i=1}^N\left(J_{nh_i}-\mathbf{E}J_{nh_i}\right)I_{H=h_i}-\sum_{i=1}^N\left(J_{n-m,h_i}-\mathbf{E}J_{n-m,h_i}\right)I_{H=h_i}\end{aligned}$$

$$
\begin{aligned}
&+\sum_{i=1}^{N}\left(\mathbf{E}J_{n,h_{\mathtt{i}}}-\mathbf{E}J_{n-m,h_{\mathtt{i}}}\right)I_{H=h_{\mathtt{i}}}+\sum_{i=1}^{N}J_{n-m,h_{\mathtt{i}}}I_{H=h_{\mathtt{i}}} \\
&\le \max_{i=1}^{N}\left|J_{nh_{\mathtt{i}}}-\mathbf{E}J_{nh_{\mathtt{i}}}\right|+\max_{i=1}^{N}\left|J_{n-m,h_{\mathtt{i}}}-\mathbf{E}J_{n-m,h_{\mathtt{i}}}\right|+J_{n-m,H} \\
&\qquad \text{(by monotonicity of } \mathbf{E}J_{nh} \text{ in } n \text{ for fixed } h\text{)}\ .
\end{aligned}
$$

We take expected values on left and right hand sides and note that by Lemma 9,

$$
\mathbf{E}\left\{J_{nH}-J_{n-m,H}\right\}\le\sqrt{\frac{2+\log 2N}{n-m}}+\sqrt{\frac{2+\log 2N}{n}}=O\left(\sqrt{\frac{\log n}{n}}\right)\ .
$$

Thus, by Lemma 5,

$$
\frac{\mathbf{E}J_{nH}}{\inf_h \mathbf{E}J_{nh}}\le\frac{\mathbf{E}J_{n-m,H}}{\inf_h \mathbf{E}J_{nh}}+O\left(\frac{\sqrt{\log n}}{n^{1/10}}\right)\ .
$$

The Proposition follows from the last inequality and the Theorem. □

8. Acknowledgment.

We thank the referees for their insightful comments and helpful suggestions.

9. References.

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

P. Assouad, "Deux remarques sur l'estimation," *Comptes Rendus de l'Académie des Sciences de Paris*, vol. 296, pp. 1021–1024, 1983.

A. Berlinet and L. Devroye, "A comparison of kernel density estimates," *Publications de l'Institut de Statistique de l'Université de Paris*, vol. 38(3), pp. 3–59, 1994.

L. Birgé, "Non-asymptotic minimax risk for Hellinger balls," *Probability and Mathematical Statistics*, vol. 5, pp. 21–29, 1985.

L. Birgé, "On estimating a density using Hellinger distance and some other strange facts," *Probability Theory and Related Fields*, vol. 71, pp. 271–291, 1986.

L. Birgé, "On the risk of histograms for estimating decreasing densities," *Annals of Statistics*, vol. 15, pp. 1013–1022, 1987a.

L. Birgé, "Estimating a density under order restrictions: nonasymptotic minimax risk," *Annals of Statistics*, vol. 15, pp. 995–1012, 1987b.

L. Birgé, "The Grenander estimator: a nonasymptotic approach," *Annals of Statistics*, vol. 17, pp. 1532–1549, 1989.

A. W. Bowman, "An alternative method of cross-validation for the smoothing of density estimates," *Biometrika*, vol. 71, pp. 353–360, 1984.

R. Cao, A. Cuevas, and W. González-Manteiga, "A comparative study of several smoothing methods in density estimation," *Computational Statistics and Data Analysis*, vol. 17, pp. 153–176, 1994.

P. Deheuvels, "Estimation non paramétrique de la densité par histogrammes generalisés," *Revue de Statistique Appliquée*, vol. 25, pp. 5–42, 1977a.

P. Deheuvels, "Estimation nonparamétrique de la densité par histogrammes generalisés," *Publications de l'Institut de Statistique de l'Université de Paris*, vol. 22, pp. 1–23, 1977b.

P. Deheuvels and P. Hominal, "Estimation automatique de la densité," *Revue de Statistique Appliquée*, vol. 28, pp. 25–55, 1980.

L. Devroye, "The equivalence of weak, strong and complete convergence in L_1 for kernel density estimates," *Annals of Statistics*, vol. 11, pp. 896–904, 1983.

L. Devroye, *A Course in Density Estimation*, Birkhäuser, Boston, 1987.

L. Devroye, "Asymptotic performance bounds for the kernel estimate," *Annals of Statistics*, vol. 16, pp. 1162–1179, 1988.

L. Devroye, "A universal lower bound for the kernel estimate," *Statistics and Probability Letters*, vol. 8, pp. 419–423, 1989.

L. Devroye, "The double kernel method in density estimation," *Annales de l'Institut Henri Poincaré*, vol. 25, pp. 533–580, 1989.

L. Devroye, "Exponential inequalities in nonparametric estimation," in: *Nonparametric Functional Estimation and Related Topics*, ed. G. Roussas, pp. 31–44, NATO ASI Series, Kluwer Academic Publishers, Dordrecht, 1991.

L. Devroye, "A note on the usefulness of superkernels in density estimation," *Annals of Statistics*, vol. 20, pp. 2037–2056, 1992.

L. Devroye, "On good deterministic smoothing sequences for kernel density estimates," *Annals of Statistics*, vol. 22, pp. 886–889, 1994.

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

D. L. Donoho, G. Kerkyacharian, and D. Picard, "Wavelet shrinkage: asymptotia?," *Journal of the Royal Statistical Society, Series B*, vol. 57, pp. 301–369, 1995.

D. L. Donoho and R. C. Liu, "Geometrizing rates of convergence, III," *Annals of Statistics*, vol. 19, pp. 668–701, 1991.

P. Hall, "On convergence rates in nonparametric problems," *International Statistical Review*, vol. 57, pp. 45–58, 1989.

P. Hall and M. P. Wand, "Minimizing L_1 distance in nonparametric density estimation," *Journal of Multivariate Analysis*, vol. 26, pp. 59–88, 1988.

I. A. Ibragimov and R. Z. Khasminskii, "Estimation of distribution density belonging to a class of entire function," *Theory of Probability and its Applications*, vol. 27, pp. 551–562, 1982.

J. S. Marron, "Automatic smoothing parameter selection: a survey," *Empirical Economics*, vol. 13, pp. 187–208, 1988.

J. S. Marron, "Automatic smoothing parameter selection: a survey," in: *Semiparametric and Nonparametric Economics*, ed. A. Ullah, pp. 65–86, Heidelberg, 1989a.

C. McDiarmid, "On the method of bounded differences," in: *Surveys in Combinatorics 1989*, vol. 141, pp. 148–188, London Mathematical Society Lecture Notes Series, Cambridge University Press, Cambridge, 1989.

B.-U. Park and B. A. Turlach, "Practical performance of several data driven bandwidth selectors (with discussion)," *Computational Statistics*, vol. 7, pp. 251–270, 1992.

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

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

M. Rudemo, "Empirical choice of histograms and kernel density estimators," *Scandinavian Journal of Statistics*, vol. 9, pp. 65–78, 1982.

C. J. Stone, "An asymptotically optimal window selection rule for kernel density estimates," *Annals of Statistics*, vol. 12, pp. 1285–1297, 1984.

B. A. Turlach, "Bandwidth selection in kernel density estimation: a review," Technical Report, Université Catholique de Louvain, 1993.

Y. G. Yatracos, “Rates of convergence of minimum distance estimators and Kolmogorov’s entropy,” *Annals of Statistics*, vol. 13, pp. 768–774, 1985.