# AN ASYMPTOTICALLY EFFICIENT SOLUTION TO THE BANDWIDTH PROBLEM OF KERNEL DENSITY ESTIMATION

James Stephen Marron
University of North Carolina at Chapel Hill

## Abstract

A data-driven method of choosing the bandwidth, h, of a kernel density estimator is proposed. It is seen that this means of selecting h is asymptotically equivalent to taking the h that minimizes a certain weighted version of the mean integrated square error. Thus, for a given kernel function, the bandwidth can be chosen optimally without making precise smoothness assumptions on the underlying density. The proposed technique is a modification of cross-validation.

AMS 1980 Subject Classification: Primary 62G05, Secondary 62G20

Keywords: Nonparametric density estimation, kernel estimator, bandwidth, smoothing parameter, cross-validation.

## 1. Introduction

Consider the problem of estimating a univariate probability density function, $f$, using a sample $X_1,\ldots, X_n$ from $f$. Let $\hat{f} = \hat{f}(x, X_1,\ldots, X_n)$ denote an estimator. A common error norm is Mean Integrated Square Error, which is defined as follows. Let $w(x)$ be some nonnegative "weight function." Define

$$\text{(1.1)} \qquad \text{MISE} = E\int [\hat{f}(x)-f(x)]^2 w(x)\, dx.$$

An estimator which has been studied extensively (see, for example, the survey by Wertz (1978)) is the kernel estimator which is defined as follows. Given a "kernel function," $K$ (with $\int K(x)\, dx = 1$), and a "bandwidth," $h>0$, let

$$\text{(1.2)} \qquad \hat{f}(x,h) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) .$$

The "bandwidth problem" consists of specifying $h=h(n)$ in some asymptotically (as $n\to\infty$) optimal fashion. Under very precise assumptions on the amount of smoothness of $f$, there are many results where $h(n)$ is given deterministically to asymptotically minimize MISE or some other error norm. See, for example, Rosenblatt (1956), Parzen (1962), or Watson and Leadbetter (1963). Unfortunately, this type of result is virtually useless to the practitioner because the optimal $h(n)$ is a function of the (unknown) smoothness of $f$. This may be seen especially clearly from the results of Stone (1980) who deals with a continuum of smoothness classes. Thus there has been a considerable search for techniques which use the data to specify $h$.

A popular technique of this type is the "cross-validated" or "pseudo-maximum-likelihood" method introduced by Habbema, Hermans, and van den Broek (1974). This is defined as follows. For $j=1,\ldots, n$ form the "leave one out" kernel estimator,

$$\text{(1.3)} \qquad \hat{f}_j(x,h) = \frac{1}{(n-1)h} \sum_{\substack{i=1 \\ i\neq j}}^{n} K\left(\frac{x-X_i}{h}\right).$$

Then take $\hat{h}_1$ to maximize the "estimated likelihood,"

$$\hat{L}_1(h) = \prod_{j=1}^{n} \hat{f}_j(X_j,h) \ .$$

A recent paper by Chow, Geman and Wu (1983) contains some interesting heuristics and a consistency theorem for the estimator $\hat{f}(x,\hat{h}_1)$. Despite these encouraging results, this estimator can be very poorly behaved. In section 2, examples and heuristics are given to motivate the modification of cross-validation that is proposed in this paper (see (2.12)).

In section 4 an optimality theorem is stated which shows that, under very mild conditions on f, the modified choice of h is asymptotically equivalent to choosing h to minimize the MISE (for a particular choice of the weight function w(x)). Thus this estimator achieves the "optimal rate of convergence" of Stone (1980) without making the usual use of the precise amount of smoothness assumed on f. In fact, not only is the exponent of algebraic convergence optimized (as in Stone (1980)), but also the constant coefficient is the best possible for kernel estimation with the given kernel.

In section 3, a useful means of measuring the "smoothness" of a density function is introduced. This is seen to be very natural for the density estimation problem because it measures precisely the quantity which has traditionally been given the name "smoothness." This measure of smoothness provides a powerful analytical tool because the more conventional Lipschitz conditions on derivatives or Taylor expansion conditions, yield, in general, only an upper bound on MISE, while the results of this paper require both upper and lower bounds on MISE.

Section 5 contains some remarks. The remaining sections contain the proof of the optimality theorem.

2. Modification of cross-validation.

To see how $\hat{f}(x,\hat{h}_1)$ can be poorly behaved, consider the following example. Suppose the density f has cumulative distribution function F so that for some $\varepsilon>0$,

$$F(x) = e^{-1/x} \quad \text{for } x\in(0,\varepsilon) .$$

Such an F could easily be constructed to be infinitely differentiable. Let $X_{(1)}$ and $X_{(2)}$ denote the first two order statistics of $X_1,\ldots, X_n$. It follows from example 1.7.5 and Theorem 2.3.2 of Leadbetter, Lindgren and Rootzén (1983) that,

$$\lim_{\delta\to 0} \lim_{n} P[X_{(2)}-X_{(1)} > \frac{\delta}{(\log n)^2}] = 1 .$$

But for compactly supported K (such as, for example, the "optimal kernels" of Epanechnikov (1969) or Sacks and Ylvisaker (1981)), $\hat{L}_1(h)=0$ unless $h\geq c(X_{(2)}-X_{(1)})$ for some constant c. Thus, the cross-validated $\hat{h}_1$ must converge to 0 slower than any algebraic rate.

By the familiar variance and bias$^2$ decomposition (see (3.5) or Rosenblatt (1971)) the mean square error may be written:

$$E[\hat{f}(x,h)-f(x)]^2 = O(\frac{1}{nh}) + O(h^{2s}) ,$$

where s represents the amount of smoothness that is assumed on f. Hence, it is apparent that the estimator $\hat{f}(x,\hat{h}_1)$ can behave very poorly in the mean square sense.

Analogous, though not so dramatic, examples can be constructed by taking, for k large,

$$F(x) = x^k \quad \text{for } x \in (0,\varepsilon) ,$$

or by taking K no longer compactly supported, but with suitably "light tails."

These examples indicate that, even when f is very smooth and compactly supported, ordinary cross-validated estimators can be drastically affected by data points where f is close to 0.

A reasonable way to eliminate the above difficulty is the following. Find an interval [a,b] on which f is known to be bounded above 0. The assumption of the existence of such an interval seems easy for the practitioner to accept. Next redefine the estimated likelihood

$$\hat{L}_2(h) = \prod_{j=1}^{n} \hat{f}_j(X_j,h)^{1_{[a,b]}(X_j)} ,$$

and take $\hat{h}_2$ to maximize $\hat{L}_2(h)$. Note that cross-validation is performed only over those observations which lie in [a,b].

The estimator $\hat{f}(x,\hat{h}_2)$ has been studied by Hall (1982), although he seems to have arrived at it by considerations different from the above. The notation used here (different from that of Hall) is due to Peter Bloomfield and will facilitate the rest of this discussion. Hall's results show that, while the above pathologies cause no problems, this version of cross-validation still behaves suboptimally with respect to the rate of convergence of mean square error. It is interesting to note that the dominant term in his expansions depends only on the behavior of f at the endpoints of [a,b].

David Ruppert has suggested the following heuristic explanation of this endpoint effect. Note that if $f'(a)<0$, there will be more $X_j$'s "just to the left" of a than "just to the right." Hence if h is taken to be relatively large, more probability mass (of the density $\hat{f}(x,h)$) will be moved into the interval [a,b] which will thus increase $\hat{L}_2(h)$. Hence there will be a tendency for cross-validation to "oversmooth" (i.e., take h too large). On the other hand, if $f'(a)>0$, then, by the same argument, cross-validation will tend to "undersmooth" in order

to keep as much probability mass inside [a,b] as possible. When this effect is taken into account at both endpoints simultaneously, it is not surprising that Hall reports oversmoothing when $f'(b)-f'(a)>0$ and undersmoothing when $f'(b)-f'(a)<0$.

With this insight, Ruppert has proposed eliminating this effect in the following way. First for $j=1,\ldots, n$ define

$$\hat{p}_j = \int_a^b \hat{f}_j(x,h)\, dx \ . \tag{2.1}$$

Next redefine the estimated likelihood

$$\hat{L}_3(h) = \prod_{j=1}^{n} \left(\frac{\hat{f}_j(X_j,h)}{\hat{p}_j}\right)^{1_{[a,b]}(X_j)}$$

and take $\hat{h}_3$ to maximize $\hat{L}_3(h)$.

This estimator will now be investigated using heuristics developed by Chow, Geman and Wu (1983). First it will be convenient to define

$$p = \int_a^b f(x)\, dx \ , \tag{2.2}$$

$$\hat{p} = \int_a^b \hat{f}(x,h)\, dx \ .$$

For these heuristics assume K is nonnegative and $f(x)\log f(x)$ is integrable. By a Law of Large Numbers,

$$\begin{aligned} \frac{1}{n}\log\hat{L}_3(h) &\approx \frac{1}{n}\sum_{j=1}^{n} 1_{[a,b]}(X_j)[\log\hat{f}(X_j,h)-\log\hat{p}] \\ &\approx \int_a^b f(x)\log\hat{f}(x,h)\, dx - p\log\hat{p} \ . \end{aligned} \tag{2.3}$$

But now by Jensen's Inequality,

$$\int_a^b \frac{f(x)}{p}\log\left(\frac{\hat{pf}(x,h)}{\hat{p}f(x)}\right) dx \le \log\left(\int_a^b \frac{\hat{f}(x,h)}{\hat{p}}\, dx\right) = 0 \ , \tag{2.4}$$

with equality if and only if,

$$\frac{\hat{f}(x,h)}{\hat{p}} = \frac{f(x)}{p} , \quad \text{a.e. on } [a,b] .$$

Hence

$$(2.5) \qquad \int_a^b f(x)\log\hat{f}(x,h)dx - p\log\hat{p} \le \int_a^b f(x)\log f(x)dx - p\log p.$$

Thus, $\hat{L}_3$ is essentially using the conditional Kullback-Leibler information (the left hand side of (2.4)) as a measure of how well $\hat{f}(x,h)$ approximates $f(x)$. But this measure has the disturbing property that it fails to distinguish between $\hat{f}$ and $f$ when they are unequal but proportional to each other.

Peter Bloomfield has suggested overcoming this difficulty by sharpening the inequality (2.5) using the following device. Note that for $x,y>0$,

$$(2.6) \qquad y\log(x/y) \le x - y ,$$

with equality only when $x = y$. Hence

$$p\log\hat{p} - p\log p \le \hat{p} - p .$$

It now follows from (2.5) that

$$(2.7) \qquad \int_a^b f(x)\log\hat{f}(x,h)dx - \hat{p} \le \int_a^b f(x)\log f(x)dx - p ,$$

with equality if and only if $f(x) = \hat{f}(x,h)$ for almost all $x \in [a,b]$. Now reversing the heuristic argument (2.3) it is apparent that the estimated likelihood should be redefined as

$$\hat{L}_4(h) = \prod_{j=1}^{n} [\hat{f}_j(X_j,h)e^{-\hat{p}_j/p}]^{1_{[a,b]}(X_j)}$$

and $\hat{h}_4$ taken to maximize $\hat{L}_4(h)$.

Peter Bloomfield has pointed out that $\hat{L}_4(h)$ may be somewhat simplified, from the computational viewpoint, in the following way. Define

$$(2.8) \qquad A = \{j=1,\ldots, n: \ X_j \in [a,b]\} ,$$

and let #(A) denote the cardinality of A. Now note that

$$\prod_{j=1}^{n} \exp(-1_{[a,b]}(X_j)\hat{p}_j/p) \approx \exp(-\#(A)\hat{p}/p)$$

$$= \exp(-\frac{\#(A)}{p} \int_a^b \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x-X_i}{h})\,dx)$$

$$= \prod_{i=1}^{n} \exp(-\frac{\#(A)}{np} \int_a^b \frac{1}{h} K(\frac{x-X_i}{h})\,dx)$$

$$\approx \prod_{i=1}^{n} e^{-\rho(X_j)} ,$$

where

$$(2.9) \qquad \rho(x) = \int_a^b \frac{1}{h} K(\frac{y-x}{h})\,dy .$$

Thus redefine the estimated likelihood

$$\hat{L}_5(h) = \prod_{j=1}^{n} \hat{f}_j(X_j,h)^{1_{[a,b]}(X_j)} e^{-\rho(X_j)} .$$

Note this also avoids difficulties about the fact that p in $\hat{L}_4(h)$ is unknown.

One last refinement will now be made. Many authors, starting with Parzen (1962) and Watson and Leadbetter (1964), have noticed that the asymptotic properties of K can be greatly improved by allowing K(x) to be negative for some x. The results of this paper apply to either this type of kernel or the non-negative kernels which guarantee that $\hat{f}$ is "range-preserving." However the proofs in this paper involve taking logarithms, so it is necessary to do some truncation. Define, for $x \in \mathbf{R}$,

$$(2.10) \qquad \hat{f}^{+}(x,h) = \max(\hat{f}(x,h),0) ,$$

and for j=1,..., n,

$$(2.11) \qquad \hat{f}_j^{+}(x,h) = \max(\hat{f}_j(x,h),0) .$$

Now redefine the estimated likelihood

$$(2.12)\qquad \hat{L}(h) = \prod_{j=1}^{n} \hat{f}_j^{+}(X_j,h)\, 1_{[a,b]}(X_j)\, e^{-\rho(X_j)}$$

and take $\hat{h}$ to maximize $\hat{L}(h)$. It will be seen in section 4 that the estimator $\hat{f}(x,\hat{h})$ has excellent asymptotic properties.

An interesting side effect of the above truncation is the following. If for some h there is an $X_j \in [a,b]$ for which $\hat{f}_j(X_j,h)<0$, then $\hat{L}(h) = 0$. Hence, such an h can not be chosen to be $\hat{h}$. Thus, since

$$\hat{f}(X_j,h) = \frac{n-1}{n}\hat{f}_j(X_j,h) + \frac{1}{nh}K(0)\ ,$$

if $K(0)\geq 0$, then for $j\in A$, $\hat{f}(X_j,\hat{h})>0$. Hence, the estimator $\hat{f}(x,\hat{h})$ has the property that it is "range-preserving" (i.e.: $>0$) at each data point in [a,b]. Of course, the experimenter who requires that f be "range preserving" outside the interval [a,b] can guarantee this by taking K nonnegative.

## 3. A measure of smoothness.

It is assumed that the kernel, K(x), is a measurable function for which

$$(3.1)\qquad \int K(x)\,dx = 1\ ,$$

(3.2) K is of bounded variation,

(3.3) K is uniformly continuous and letting $\eta(x)$ denote the square root of the modulus of continuity,

$$\int_0^1 [-\log x]^{1/2}\, d\eta(x) < \infty\ ,$$

(3.4) K is bounded and supported inside [-1,1],

Assumption (3.1) guarantees $\int \hat{f}(x,h)\,dx = 1$. Assumptions (3.2)-(3.4) allow the application of the powerful theorem A of Silverman (1978). It should be noted that a sufficient condition for (3.3) is that K satisfy any Generalized Lipschitz Condition. Assumption (3.4) is equivalent to assuming K is bounded with compact support, because h may be rescaled. The compact support requirement is really only needed in the proof of Lemma 5, otherwise some integrability conditions

would suffice. The reader who is familiar with kernel estimation may be surprised at the lack of the vanishing moment assumption introduced by Parzen (1962). The theorem of this paper says $\hat{f}(x,\hat{h})$ gives the best possible MISE for the chosen K. Of course the experimenter will want to choose a K for which the best MISE is reasonably good, but that is irrelevant to the theorem of this paper.

Now recall the definition of MISE from (1.1). In the case of kernel estimation, MISE is usually decomposed into variance and bias$^2$ terms as follows (assuming f is uniformly continuous)

$$\text{var } \hat{f}(y,h) = \frac{1}{n^2h^2}\sum_{i=1}^{n}\left(EK\left(\frac{y-X_i}{h}\right)^2 - \left[EK\left(\frac{y-X_i}{h}\right)\right]^2\right)$$

$$= \frac{1}{nh^2}\int K\left(\frac{y-x}{h}\right)^2 f(x)\,dx - \frac{1}{nh^2}\left[\int K\left(\frac{y-x}{h}\right) f(x)\,dx\right]^2$$

$$= \frac{1}{nh}\int K(u)^2 f(y-hu)\,du - \frac{1}{n}\left[\int K(u) f(y-hu)\,du\right]^2$$

$$= \frac{1}{nh} f(y)\int K(u)^2\,du + o\left(\frac{1}{nh}\right) ,$$

as $nh\to\infty$. Similarly,

$$E\,\hat{f}(y,h) = \int K(u) f(y-hu)\,du .$$

Thus

$$E\left[\hat{f}(y,h)-f(y)\right]^2 = \frac{1}{nh}f(y)\int K(u)^2\,du + o\left(\frac{1}{nh}\right) + \left[\int K(u) f(y-hu)\,du - f(y)\right]^2 , \tag{3.5}$$

from which it follows that

$$\begin{aligned} \text{MISE} = {} & \frac{1}{nh}\left(\int f(y)w(y)\,dy\right)\left(\int K(u)^2\,du\right) + o\left(\frac{1}{nh}\right) \\ & + \int\left[\int K(u) f(y-hu)\,du - f(y)\right]^2 w(y)\,dy . \end{aligned} \tag{3.6}$$

For the bias$^2$ term in this expression, adopt the notation

(3.7) $$s_f(h) = \int [\int K(u)f(y-hu)du - f(y)]^2 w(y)dy \ .$$

To get some feel for how the tail behavior of $s_f(h)$ measures the smoothness of f, note that if f satisfies a Lipschitz condition of order $\gamma$ and $w(x)$ is integrable, then

(3.8) $$s_f(h) = O(h^{2\gamma}) \ .$$

Similarly, if, in addition, it is assumed that f has a bounded kth derivative and at least the first k-1 moments of K vanish, then by Taylors Theorem,

$$s_f(h) = O(h^{2k}) \ .$$

The reason $s_f(h)$ is a very natural measure of the quantity which is usually termed "smoothness" is that it is precisely the bias$^2$ term of the MISE, and hence correctly deals with difficulties about more smoothness at one point than at another. As pointed out in section 1, $s_f(h)$ also provides a useful tool in that it easily allows both upper and lower bounds to be placed on the MISE.

## 4. Optimality Theorem.

It will be assumed that the underlying density, f, satisfies

(4.1) f is bounded,

(4.2) f is bounded above 0 on [a,b],

(4.3) there are constants $M,\gamma>0$ so that for all x,y

$$|f(x)-f(y)| \leq M|x-y|^{\gamma}.$$

Assumption (4.1) is a consequence of (4.3), but it is listed separately here for convenience in the upcoming proofs. The need for assumption (4.2) was discussed in section 2. By (3.8), a consequence of (4.3) is

(4.4) $$\limsup_{h\to 0} h^{-2\gamma} s_f(h) < \infty \ .$$

The optimality theorem of this paper can now be stated.

Theorem:

Assume (3.1)-(3.4) and (4.1)-(4.3). Also assume

(4.5) $\quad w(x) = 1_{[a,b]}(x)/f(x)$.

Then choosing h to maximize $\hat{L}(h)$ is asymptotically (as $n\to\infty$) the same as picking h to minimize

$$\text{MISE} + o_p(\text{MISE}) \ .$$

5. Remarks.

Remark 5.1. Stone (1982) poses (see Question 3) the problem of finding a regression estimator which achieves his "optimal rate of convergence" without making use the precise amount of smoothness of the unknown regression function. The theorem of this paper solves the analogous problem in the setting of density estimation. In fact the results of this paper do much more than this because Stone was concerned with only the exponent of algebraic convergence, while here the constant multiplier is also optimized.

Remark 5.2. The fact that optimality is achieved only for a particular weight function should not be too disappointing. The one used here is quite natural because MISE is proportional to the expected relative square error:

$$E[(\frac{\hat{f}(X)-f(X)}{f(X)})^2 | X\in[a,b]] \ .$$

It is seen in Marron (1982) that this error norm is precisely the one required for the application of density estimation to the classification problem. It may be seen without too much effort that the indicator function in (2.12) may be replaced by any bounded, measurable nonnegative function q(x), which is supported inside [a,b], and the theorem will still be true with

$$w(x) = q(x)/f(x) \ .$$

Finally, note that by the expansion (3.6), it is apparent that w(x) affects only the constant multiplier in the rate of convergence. Hence, for any bounded, measurable nonnegative weight function w(x), supported inside [a,b], the analog of Question 3 of Stone (1982) (see Remark 5.1) is still answered.

Remark 5.3. At first glance one might be disturbed by the fact that the MISE that is minimized here is limited to the interval [a,b]. In somewhat similar settings, in the case of estimating a regression function, Gasser and Müller (1979) and Rice and Rosenblatt (1983) have observed that such a MISE is strongly affected by the behavior of the unknown function at the endpoints and hence the bandwidth which minimizes MISE can provide relatively poor estimates in the interior of [a,b]. However, with very little effort, one may see that such an "endpoint effect" does not occur in the present setting. This is because the density f extends (and is smooth) outside the interval [a,b] and observations outside [a,b] are employed in the estimator of this paper. Hence, the MISE of this paper provides a very reasonable error criterion.

## 6. Proof of Theorem.

This proof uses techniques developed in Hall (1982) and Chow, Geman and Wu (1983). For sequences $\{a_n\}$ and $\{b_n\}$ it will be convenient to let the phrase "h=h(n) is between $a_n$ and $b_n$" mean:

$$\lim_{n\to\infty} \frac{h}{a_n} = \infty \quad \text{and} \quad \lim_{n\to\infty} \frac{h}{b_n} = 0 \ .$$

It will also be useful to define, for j=1,..., n,

$$(6.1) \qquad \Delta_j = \frac{\hat{f}_j(X_j,h) - f(X_j)}{f(X_j)} \ , \quad \Delta_j^+ = \frac{\hat{f}_j^+(X_j,h) - f(X_j)}{f(X_j)} \ .$$

By Theorem A of Silverman (1978), for h between logn/n and 1,

$$(6.2)\qquad \sup_x |\hat{f}^+(x,h) - f(x)| \le \sup_x |\hat{f}(x,h) - f(x)| \to 0 ,$$

in probability. But by (1.2), (1.3) and (3.4),

$$(6.3)\qquad \hat{f}_j(x,h) - \hat{f}(x,h) = \frac{1}{n(n-1)h} \sum_{i\neq j} K(\frac{x-X_i}{h}) - \frac{1}{nh} K(\frac{x-X_j}{h})$$

$$= O(\frac{1}{nh}) , \quad \text{uniformly in } x \text{ and } j.$$

Hence, by (4.2), using the notation (2.8), for h between logn/n and 1,

$$\sup_{j\in A} |\Delta_j^+| \le \sup_{j\in A} |\Delta_j| \to 0 ,$$

in probability. Now, for n=1,2,..., define the event

$$U_n = \{\Delta_j^+ = \Delta_j \quad \text{for each } j\in A\} .$$

It follows from the above that for h between logn/n and 1,

$$(6.4)\qquad \lim_{n\to\infty} P[U_n] = 1.$$

It will be convenient to define

$$(6.5)\qquad L = \prod_{j=1}^{n} (f(X_j)/e)^{1_{[a,b]}(X_j)}$$

Note that maximizing $\hat{L}(h)$ is the same as maximizing $n^{-1}\log(\hat{L}(h)/L)$. From (2.12) and the above it follows that, for h between logn/n and 1, on the event $U_n$,

$$\frac{1}{n}\log(\frac{\hat{L}(h)}{L}) = \frac{1}{n} \sum_{j=1}^{n} [1_{[a,b]}(X_j)\log(1+\Delta_j) - \rho(X_j) + 1_{[a,b]}(X_j)]$$

$$(6.6)\qquad = \frac{1}{n} \sum_{j=1}^{n} [1_{[a,b]}(X_j)(\Delta_j - \frac{1}{2}\Delta_j^2 + o_p(\Delta_j^2)) - \rho(X_j) + 1_{[a,b]}(X_j)]$$

$$= [\frac{1}{n} \sum_{j=1}^{n} 1_{[a,b]}(X_j)\Delta_j - \frac{1}{n} \sum_{j=1}^{n} \rho(X_j) + \frac{\#(A)}{n}] - \frac{1}{2n} \sum_{j\in A} \Delta_j^2 + o_p(\frac{1}{n} \sum_{j\in A} \Delta_j^2) .$$

The claim of the theorem for the case of h between logn/n and 1 is now a consequence of (6.4) and the following lemmas.

Lemma 1: For h between $n^{-1}$ and 1,

$$\frac{1}{n}\sum_{j=1}^{n} 1_{[a,b]}(X_j)\Delta_j - \frac{1}{n}\sum_{j=1}^{n} \rho(X_j) + \frac{\#(A)}{n} = o_p(\frac{1}{nh} + s_f(h)) .$$

Lemma 2: For h between $n^{-1}$ and 1,

$$\frac{1}{n}\sum_{j\in A} \Delta_j^2 = \frac{b-a}{nh}\int K(u)^2 du + s_f(h) + o_p(\frac{1}{nh} + s_f(h)) .$$

Lemma 3: For $w(x) = 1_{[a,b]}(x)/f(x)$ and h between $n^{-1}$ and 1,

$$\text{MISE} = \frac{b-a}{nh}\int K(u)^2 du + s_f(h) + o(\frac{1}{nh}) .$$

The fact that the theorem is true for all h is now demonstrated by:

Lemma 4: If $\hat{h}=\hat{h}(n)$ is any sequence of maxima of $\hat{L}(h)$, then

$$\lim_{c\to 0} \overline{\lim_{n}} P[\hat{h}<cn^{\frac{-1}{2\gamma+1}}] = 0 .$$

Lemma 5: If $\hat{h}=\hat{h}(n)$ is any sequence of maxima of $\hat{L}(h)$, then $\hat{h} \to 0$ a.s.

The proof of Lemma 1 is in section 7. Lemma 2 is theorem 2 of Marron (1983a) in the special case of d=1 and $w(x) = f(x)^{-2}1_{[a,b]}(x)$. Lemma 3 follows from (3.6), (3.7) and (4.5). Lemmas 4 and 5 are theorems 1 and 2 respectively of Marron (1983b). Lemma 4 makes use of an order statistics result of Cheng (1983).

## 7. Proof of Lemma 1.

This proof uses techniques developed by Hall (1982). It will be convenient to define:

$$\ell(x,y) = \frac{1_{[a,b]}(y)}{(n-1)hf(y)} K(\frac{y-x}{h})$$

$$g_1(x) = E\ell(x,X_j) ,$$

$$g_2(y) = E\ell(X_i,y) \ ,$$

$$G = E\ell(X_i,X_j) \quad , \quad \text{where } i\neq j \ ,$$

Observe that, from (6.1)

$$n^{-1}\sum_{j=1}^{n} 1_{[a,b]}(X_j)\Delta_j = n^{-1}\sum_{j=1}^{n}\sum_{i\neq j} \ell(X_i,X_j) - \#(A)n^{-1}$$

$$= n^{-1}\sum_{j=1}^{n}\sum_{i\neq j}[\ell(X_i,X_j)-g_1(X_i)-g_2(X_j)+G] + (1-n^{-1})\sum_{j=1}^{n}[g_2(X_j)-G] \tag{7.1}$$

$$+ (1-n^{-1})\sum_{i=1}^{n} g_1(X_i) - \#(A)n^{-1} \ .$$

Each term on the right hand side will be approximated in turn.

For the first term, let

$$r(x,y) = \ell(x,y) - g_1(x) - g_2(y) + G \ .$$

Routine computations show that, for $i\neq j$,

$$Er(X_i,X_j)^2 = E[\ell(X_i,X_j)^2-G^2] - E[g_1(X_i)^2-G^2] - E[g_2(X_j)^2-G^2]$$

$$\leq E\ell(X_i,X_j)^2 = \iint [\frac{1_{[a,b]}(y)}{(n-1)hf(y)} K(\frac{y-x}{h})]^2 f(x)f(y)dxdy$$

$$= O(n^{-2}h^{-1}) \ ,$$

and also that

$$E[\sum_{j=1}^{n}\sum_{i\neq j} r(X_i,X_j)]^2 = n(n-1)Er(X_i,X_j)^2 = O(h^{-1}) \ .$$

Hence, by the Markov inequality,

$$n^{-1}\sum_{j=1}^{n}\sum_{i\neq j}[\ell(X_i,X_j)-g_1(X_i)-g_2(X_j)+G] = O_p(n^{-1}h^{-1/2}) \ . \tag{7.2}$$

For the second term on the right of (7.1), note that by (3.7) and (4.5),

$$E[g_2(X_j)-(n-1)^{-1}1_{[a,b]}(X_j)]^2$$

$$= \int [\int \frac{1_{[a,b]}(y)}{(n-1)hf(y)} K(\frac{y-x}{h}) f(x)dx - \frac{1_{[a,b]}(y)}{n-1}]^2 f(y)dy$$

$$= (n-1)^{-2}\int [\int K(u)f(y-hu)du-f(y)]^2 1_{[a,b]}(y)f(y)^{-1}dy$$

$$= (n-1)^{-2}s_f(h) \ .$$

Thus, using the fact that $Eg_2(X_j) = G$ twice,

$$E(\sum_{j=1}^{n}[g_2(X_j)-G])^2 = nE[g_2(X_j)-G]^2 = O(n^{-1}s_f(h)) \ .$$

Hence

$$(7.3) \qquad (1-n^{-1})\sum_{j=1}^{n}[g_2(X_j)-G] = O_p(n^{-1/2}s_f(h)^{1/2}) \ .$$

For the third term on the right of (7.1), from (2.9) note that

$$\rho(x) = (n-1)g_1(x) \ .$$

It now follows from (7.1), (7.2) and(7.3) that

$$n^{-1}\sum_{j=1}^{n}1_{[a,b]}(X_j)\Delta_j = O_p(n^{-1}h^{-1/2}) + O_p(n^{-1/2}s_f(h)^{1/2}) + n^{-1}\sum_{j=1}^{n}\rho(X_j) - \#(A)n^{-1}.$$

Lemma 1 is an easy consequence of this and (4.4).

## 8. Acknowledgement.

The author is grateful to David Ruppert, Raymond Carroll and especially to Peter Bloomfield for many interesting and stimulating conversations during the course of the research presented in this paper. The author is also indebted to Charles J. Stone for introducing him to this area of statistics and mentioning the importance of the problem solved in this paper.

9. References.

Cheng, S.H. (1983). On a problem concerning spacings. Center for Stochastic Processes Tech. Rept. #27, Statistics Dept., UNC, Chapel Hill, NC.

Chow, Y.S., Geman, S. and Wu, L.D. (1983). Consistent cross-validated density estimation. Ann. Statist. 11, 25-38.

Epanechnikov, V. (1969). Nonparametric estimates of a multivariate probability density. Theor. Prob. Appl. 14, 153-158.

Gasser, T. and Müller, H.G. (1979). Kernel estimation of regression functions. Smoothing Techniques for Curve Estimation. Lecture Notes in Math. 757, 23-68.

Habbema, J.D.F., Hermans, J. and van den Broek, K. (1974). A stepwise discrimination analysis program using density estimation. Compstat 1974: Proceedings in computational statistics. (G. Bruckman, ed.) 101-110. Vienna: Physica Verlag.

Hall, P. (1982). Cross-validation in density estimation. Biometrika 69, 383-390.

Leadbetter, M.R., Lindgren, G. and Rootzén, H. (1983). Extremes and related properties of random sequences and processes. Springer (New York).

Marron, J.S. (1982). Optimal rates of convergence to Bayes risk in nonparametric discrimination. North Carolina Institute of Statistics, Mimeo Series #1510.

Marron, J.S. (1983a). Convergence properties of an empirical error criterion for multivariate density estimation. North Carolina Institute of Statistics, Mimeo Series #1520.

Marron, J.S. (1983b). Uniform consistency of a cross-validated density estimator. North Carolina Institute of Statistics, Mimeo Series #1519.

Parzen, E. (1962). On the estimation of a probability density and mode. Ann. Math. Statist. 33, 1065-1076.

Rice, J. and Rosenblatt, M. (1983). Smoothing splines: Regression, derivatives and deconvolution. Ann. Statist. 11, 141-156.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. Ann. Math. Statist. 27, 832-837.

Rosenblatt, M. (1971). Curve estimates. Ann. Math. Statist. 42, 1815-1842.

Sacks, J. and Ylvisaker, D. (1981). Asymptotically optimum kernels for density estimation at a point. Ann. Statist. 9, 334-346.

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

Stone, C.J. (1980). Optimal convergence rates for nonparanetric estimators. Ann. Statist. 8, 1348-1360.

Stone, C.J. (1982). Optimal global rates of convergence for nonparametric regression. Ann. Statist. 10, 1040-1053.

Watson, G.S. and Leadbetter, M.R. (1963). On the estimation of a probability density, I. Ann. Math. Statist. 34, 480-491.

Wertz, W. (1978). Statistical density estimation: A survey. Angewandte Statistique und Okonometrie 13, van den Broek and Ruprecht.