The Annals of Statistics
1987, Vol. 15, No. 1, 163–181

ON THE AMOUNT OF NOISE INHERENT IN BANDWIDTH SELECTION FOR A KERNEL DENSITY ESTIMATOR

BY PETER HALL[1] AND J. S. MARRON[2]

University of North Carolina, Chapel Hill

In the setting of kernel density estimation, data-driven bandwidth, i.e., smoothing parameter, selectors are considered. It is seen that there is a well-defined, and surprisingly restrictive, bound on the rate of convergence of *any* automatic bandwidth selection method to the optimum. The method of least squares cross-validation achieves this bound.

1. Introduction. A widely studied method of using a sample $X_1, \dots, X_n$ to estimate their common density function, f, is the kernel estimator,

$$\hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right),$$

where K is called the kernel function and h is called the bandwidth or smoothing parameter. The choice of h is crucial to the performance of this estimator. Too small an h gives a curve that is too noisy in that it is quite dependent on the particular realization of the data at hand, showing features that are not shared by the density f. Too large an h creates a bias that can eliminate, by oversmoothing, some interesting features of f.

A considerable amount has been written about "optimal" selection of the bandwidth (see, e.g., Fryer (1977) and Wegman (1972)). In particular there are well-known asymptotic formulas for the minimizer of mean integrated square error (see Parzen (1962) and Rosenblatt (1971)). Unfortunately, these depend intimately on the unknown density, f, while any practical method of choosing a bandwidth must depend only on the sample. In this paper it is demonstrated that there are well-defined, and surprisingly restrictive, limits to the accuracy of any possible data-driven bandwidth selector.

Among the most promising automatic methods of choosing a bandwidth, are those based on cross-validation. Several recent papers (see, e.g., Hall (1983), Stone (1984, 1985) Burman (1985) and Marron (1985)) have shown that if $\hat{h}$ is some suitably chosen cross-validated bandwidth, then under surprisingly mild conditions, $\hat{h}$ is "asymptotically optimal," in the sense that

$$\Delta(\hat{h}, f)/\Delta(\hat{h}_f, f) \to 1 \tag{1.1}$$

Received May 1985; revised March 1986.

[1] On leave from Australian National University. Work supported by AFOSR grant No. F-49620-82-C-0009.

[2] Work supported by NFS grant DMS-8400602.

AMS 1980 *subject classifications.* Primary 62G05; secondary 62E20, 62H99.

Key words and phrases. Bandwidth, cross-validation, data-driven estimate, noise, smoothing parameter selection, window width.

or

$$\hat{h}/\hat{h}_f \to 1 \tag{1.2}$$

in some mode of convergence, where $\hat{h}_f$ is the minimizer of integrated square error,

$$\Delta(h, f) = \int[\hat{f}_h(x) - f(x)]^2\, dx.$$

Among the most promising of this type of result is that of Hall (1983) and Stone (1984), where $\hat{h}$ is chosen by least squares cross-validation. One means of defining this is to let $\hat{h}_c$ be the minimizer of

$$CV(h) = \int\hat{f}_h(x)^2\, dx - n^{-1}\sum_{i=1}^{n}\hat{f}_{h,i}(X_i),$$

where $\hat{f}_{h,i}$ denotes the kernel density estimator with the ith observation deleted from the sample. This bandwidth was proposed and motivated by Rudemo (1982) and Bowman (1984).

While asymptotic optimality is very encouraging, there remains the important question of exactly what this implies for the set of data at hand. In particular it could be that impossibly large samples are necessary before the asymptotics effectively describe what is happening. One means of addressing this issue is to consider the rates of convergence in (1.1) and (1.2). This has been done in Hall and Marron (1985) and in the related regression setting by Rice (1984). A remarkable and rather depressing feature of these results is that the convergence is very slow. In particular, under very common assumptions (such as K a smooth symmetric probability density and f twice differentiable),

$$\frac{\hat{h}_c - \hat{h}_f}{\hat{h}_f} = O_p(n^{-1/10}), \tag{1.3}$$

$$\frac{\Delta(\hat{h}_c, f) - \Delta(\hat{h}_f, f)}{\Delta(\hat{h}_f, f)} = O_p(n^{-1/5}). \tag{1.4}$$

Hence, very large samples indeed are required before one may be sure that $\hat{h}$ is giving reasonable performance.

The rates given in (1.3) and (1.4) would be very discouraging except for the fact that

$$\frac{h_f - \hat{h}_f}{\hat{h}_f} = O_p(n^{-1/10}), \tag{1.5}$$

$$\frac{\Delta(h_f, f) - \Delta(\hat{h}_f, f)}{\Delta(\hat{h}_f, f)} = O_p(n^{-1/5}), \tag{1.6}$$

where h_f is the minimizer of the mean integrated square error,

$$M(h, f) = E_f[\Delta(h, f)].$$

In other words, the level of noise involved in selecting the bandwidth by cross-validation is of the same order as the difference between the two reasonable choices of "optimal bandwidth," $\hat{h}_f$ and h_f. In this paper, we take $\hat{h}_f$ as the optimum, because this is the bandwidth that makes $\hat{f}$ as close to f as possible for the particular set of data at hand, as opposed to close for the average over all possible data sets.

The fact that the rates in (1.5) and (1.6) are the same as those in (1.3) and (1.4) leads one to suspect that these rates are the best possible. Section 2 contains a theorem that demonstrates that this is indeed the case.

Another way of thinking about this is the following. The excruciatingly slow rates in (1.3) and (1.4) motivate the question: Is it possible that we may find a bandwidth selector that will be substantially closer to the optimum than cross-validation? The theorem of Section 2 shows that this is not possible, so there is no point searching for one.

In Section 3 there are some remarks and extensions of the theorem of Section 2. Section 4 gives additional insight into the theorem of Section 2 by putting it in a bigger framework. The proofs are in the remaining sections.

2. Main theorem. To get a result that includes not only bandwidth selectors that have been proposed, but all possible selectors, we need to show that, for any measurable function of the data, $\hat{h}(X_1, \dots, X_n)$, the rates in (1.3) and (1.4) are the best possible. For such a broad class of possible selectors, a mechanism for ruling out trivialities is required, e.g., observe that $\hat{h}_f$ itself is a function of the data. Hence, a minimax approach, somewhat similar to that first used in density estimation by Farrell (1972), will be used here.

The idea is to insist that $\hat{h}$ perform well uniformly over a collection of underlying densities, $\mathscr{F}$. Given $B > 0$, let $\mathscr{F}$ be the class of all densities f, which have two derivatives with

$$|f''(x)| \le B,$$

for all real x.

In addition suppose that K is a compactly supported probability density with a Hölder continuous second derivative. These assumptions are far from the weakest possible, and are made with a view toward keeping the proofs from being unacceptably long. The interested reader will find it easy to dispense with many of these.

The rates (1.3) and (1.4) are the best possible in the sense that:

THEOREM 2.1. *Under the above assumptions, for $\hat{h}$ any measurable function of $X_1, \dots, X_n$,*

$$\lim_{\varepsilon \to 0} \liminf_{n \to \infty} \sup_{f \in \mathscr{F}} P_f\left[\left|\frac{\hat{h} - \hat{h}_f}{\hat{h}_f}\right| > \varepsilon n^{-1/10}\right] = 1,$$

$$\lim_{\varepsilon \to 0} \liminf_{n \to \infty} \sup_{f \in \mathscr{F}} P_f\left[\left|\frac{\Delta(\hat{h}, f) - \Delta(\hat{h}_f, f)}{\Delta(\hat{h}_f, f)}\right| > \varepsilon n^{-1/5}\right] = 1.$$

3. Remarks.

REMARK 3.1. There are some important differences between our results (and their proofs) and traditional work on "optimal rates of convergence" for nonparametric density estimators (see Farrell (1972), Wahba (1975), Meyer (1977a, b), Bretagnolle and Huber (1979), Stone (1980), Ibragimov and Has'minskii (1981), Sacks and Ylvisaker (1981), Sacks and Strawderman (1982) and Stone (1982)). The classical argument involves showing that there is a bound on the rate of convergence of any estimator to the density f. This rate is achieved by several estimators, including kernel methods. In this paper we confine attention not only to kernel estimators, but to kernel estimators using a specific fixed kernel K. The only variable is the bandwidth for that particular estimator. We are, in effect, switching attention from the problem of "best estimates" of a density to that of "best estimates" of the bandwidth $\hat{h}_f$. Finally, keep in mind that $\hat{h}_f$ is a random variable, so the notion of "estimating" it is different than for estimating f.

REMARK 3.2. Perhaps it is worth noting that even in the absence of the results of Hall and Marron (1985) (and some related results of Rice (1984) in the regression setting), the theorems of this paper would still be of statistical significance. This is because, even if one did not have a bandwidth procedure that could achieve the rates given in (1.3) and (1.4), it is still important to know that the remarkably restrictive bound on these rates, presented in this paper, is present.

REMARK 3.3. It should be pointed out that the only error criteria considered in this paper are of L^2 type. Some rather compelling reasons for considering the L^1 norm to be more natural are given in Devroye and Györfi (1984). Unfortunately there is a trade-off to be made in choice of norms, because the L^1 norm seems to be far less tractable analytically than the L^2 norm. Indeed as far as we know, there is not even a known practical method of selecting the bandwidth to give L^1 analogs of (1.3) and (1.4), so deeper properties such as those of Hall and Marron (1985) and the present paper seem, at least for the moment, too much to ask for. We use L^2 norms here because, in our opinion, it is more important to understand the deeper aspects of bandwidth selection than it is to worry about mathematical details such as precisely how one measures error.

REMARK 3.4. Several extensions of Theorem 2.1 are completely straightforward. In Hall and Marron (1985), it is seen that the rates of convergence in (1.3) change if one assumes that f has more derivatives and K has a corresponding number of vanishing moments (see Parzen (1962) or Rosenblatt (1971) to see how this affects the rate of convergence of h_f to 0). The rates also change if one assumes that the data are random vectors instead of random variables. In both of these settings, Theorem 2.1 is essentially the same, except the rates change such that the rates in the analogs of (1.3) and (1.4) are again the best possible.

REMARK 3.5. If one considers results of the type (1.1) and (1.2) to be "first order optimality," then a reasonable notion of "second order optimality" is optimization of the rate of convergence in results like (1.3) and (1.4). Observe that Theorem 2.1 says that the least squares cross-validated bandwidth, $\hat{h}_c$, is second order optimal.

4. A bigger view. One of the less attractive features about Theorem 2.1 is that the suprema are over the very large class $\mathscr{F}$. In fact, a much smaller class, $\mathscr{F}_n$, which may depend on n, is all that is necessary.

These classes consist of several small perturbations of some fixed density. Start with, for convenience, a compactly supported density f_0, which has four bounded derivatives and satisfies $f_0(x) \equiv c^{(0)} > 0$ for $x \in [0,1]$. Define

$$c^{(1)} \equiv \sup_{x;\ j \le 4} |f_0^{(j)}(x)|/2.$$

Let ψ be any function on $[0, \frac{1}{2}]$, which has four derivatives and satisfies

$$|\psi(\tfrac{1}{4})| > 0, \qquad \sup_{0 \le x \le 1/2} |\psi^{(j)}(x)| < c^{(0)}/2,$$

$$\psi^{(j)}(0) = \psi^{(j)}(\tfrac{1}{2}) = 0,$$

for $0 \le j \le 4$. Set $\psi(x) = -\psi(1-x)$ for $x \in [\frac{1}{2}, 1]$, and extend ψ from $[0,1]$ to $(-\infty, \infty)$ by periodicity. Let m equal the integer part of $n^{1/5}$ and define

$$\gamma(x) = \gamma(x, n) = m^{-2}\psi(mx).$$

For $v = 0, \dots, m-1$, let $\gamma_v(x) = \gamma(x)$ on $C_v \equiv [vm^{-1}, (v+1)m^{-1}]$, and $\gamma_v(x) = 0$ off C_v.

The elements of $\mathscr{F}_n$ are indexed by sequences of length m, all of whose elements are zeros and ones, $\{\tau_v\colon 0 \le v \le m-1\}$, and are given by

$$f(x) = f(\tau_0, \dots, \tau_{m-1})(x) = f_0(x)[1 + \Sigma_v \tau_v \gamma_v(x)],$$

for $x \in (-\infty, \infty)$. These are all probability densities with support equal to the support of f_0, and satisfy

$$\sup_{x;\ j \le 2} |f^{(j)}(x)| \le c^{(1)}.$$

In particular, the second derivatives of the densities in $\mathscr{F}_n$ are all uniformly bounded. The kernel K of Section 2 has been designed for just this kind of density.

Note that Theorem 2.1 is an immediate consequence of:

THEOREM 4.1. *Under the above assumptions, for $\hat{h}$ any measurable function,*

$$\lim_{\varepsilon \to 0} \liminf_{n \to \infty} \sup_{f \in \mathscr{F}_n} P_f\left[\left|\frac{\hat{h} - \hat{h}_f}{\hat{h}_f}\right| > \varepsilon n^{-1/10}\right] = 1, \tag{4.1}$$

$$\lim_{\varepsilon \to 0} \liminf_{n \to \infty} \sup_{f \in \mathscr{F}_n} P_f\left[\left|\frac{\Delta(\hat{h}, f) - \Delta(\hat{h}_f, f)}{\Delta(\hat{h}_f, f)}\right| > \varepsilon n^{-1/5}\right] = 1. \tag{4.2}$$

Another reason to consider $\mathscr{F}_n$, instead of $\mathscr{F}$, is that this easily allows a companion result about the fact that the rates of convergence in (1.3) and (1.4) can be achieved.

THEOREM 4.2. *Under the above assumptions, there is a bandwidth selection method, $\hat{h}_c$, so that,*

$$\lim_{\lambda \to \infty} \limsup_{n \to \infty} \sup_{f \in \mathscr{F}_n} P_f \left[\left| \frac{\hat{h}_c - \hat{h}_f}{\hat{h}_f} \right| > \lambda n^{-1/10} \right] = 0, \tag{4.3}$$

$$\lim_{\lambda \to \infty} \limsup_{n \to \infty} \sup_{f \in \mathscr{F}_n} P_f \left[\left| \frac{\Delta(\hat{h}_c, f) - \Delta(\hat{h}_f, f)}{\Delta(\hat{h}_f, f)} \right| > \lambda n^{-1/5} \right] = 0. \tag{4.4}$$

Theorem 4.2 is formulated in terms of $\mathscr{F}_n$, instead of $\mathscr{F}$, because the very large size of $\mathscr{F}$ allows some paradoxical (and not statistically relevant) behavior. For example, it is easy to construct a sequence $\{f_n\}$ in $\mathscr{F}$ so that $h_{f_n} \to \infty$, by considering rescalings of a fixed density f.

Theorem 4.2 is an extension, to the case of uniformity over $\mathscr{F}_n$, of Theorems 2.1 and 2.2 of Hall and Marron (1985). The bandwidth $\hat{h}_c$ may be taken to be the one chosen by least squares cross-validation, as indicated in that paper. The technical details of this extension for (4.3) are summarized in Lemmas 6.2 and 6.3. Formula (4.4) is a consequence of (4.3) by computations of the type appearing at the end of Section 5.

REMARK 4.1. Ferguson (1985) has asked whether the minimax character of Theorem 4.1 can be replaced by some notion of "averaging." More specifically, can the trivialities discussed in Section 2 be ruled out by a mechanism that looks at a "typical" case instead of at the worst possible case? An inspection of the proofs reveals that this is indeed the case. In particular the suprema over $\mathscr{F}_n$ in Theorem 4.1 can be replaced by simple averages over $\mathscr{F}_n$. This same idea could have been used in many of the more traditional works on optimal rates of convergence given in Remark 3.1.

5. Proof of Theorem 4.1. Section 6 contains the statement of a number of lemmas, which will be used in this section. Their proofs are given in the Appendix. The symbols $C, C_1, C_2, \dots$ denote generic positive constants. The complement of an event $\mathscr{E}$ will be denoted $\tilde{\mathscr{E}}$. Superscript notation in $\Delta^{(j)}$, $M^{(j)}$, etc., denotes differentiation with respect to h. The classification argument used by Stone (1982) and Marron (1983) is an important element of this proof.

First observe that it is enough to consider only $\hat{h}$ that take on values that coincide with the $\hat{h}_f$ (where $f \in \mathscr{F}_n$). To see this, given a data-driven bandwidth $\hat{h}$, define $\hat{f}$ to be any element of $\mathscr{F}_n$ such that

$$|\hat{h}_{\hat{f}} - \hat{h}| = \inf_{f \in \mathscr{F}_n} |\hat{h}_f - \hat{h}|.$$

Then $\hat{h}_{\hat{f}}$ is also a data-driven bandwidth, i.e., it is a function of n and $X_1, \dots, X_n$ alone and does not employ any knowledge about the unknown density. For each $f_1 \in \mathscr{F}_n$,

$$|\hat{h}_{\hat{f}} - \hat{h}_{f_1}| \le |\hat{h}_{\hat{f}} - \hat{h}| + |\hat{h} - \hat{h}_{f_1}| \le 2|\hat{h} - \hat{h}_{f_1}|.$$

Therefore, the first conclusion of Theorem 4.1, (4.1), follows from

$$\lim_{\varepsilon \to 0} \liminf_{n \to \infty} \sup_{f \in \mathscr{F}_n} P_f\left[|\hat{h}_{\hat{f}} - \hat{h}_f| > \varepsilon n^{-3/10}\right] = 1, \tag{5.1}$$

(6.1) of Lemma 6.1, and Lemma 6.2.

To prove (5.1), first write

$$\begin{aligned} \Delta(h, \hat{f}\,) &= \Delta(h, f\,) - 2\int [\hat{f}_h(x) - f(x)][\hat{f}(x) - f(x)]\,dx \\ &\quad + \int [\hat{f}(x) - f(x)]^2\,dx. \end{aligned} \tag{5.2}$$

Using the formula

$$(\partial/\partial h)\hat{f}_h(x) = h^{-1}[\hat{g}_h(x) - \hat{f}_h(x)],$$

where

$$\hat{g}_h(x) \equiv (nh)^{-1} \sum_{i=1}^{n} L\left(\frac{x - X_i}{h}\right),$$

$$L \equiv -zK'(z),$$

differentiation of (5.2) with respect to h and evaluation at $h = \hat{h}_{\hat{f}}$ give

$$0 = \Delta^{(1)}(\hat{h}_{\hat{f}}, f\,) + \hat{h}_{\hat{f}}^{-1} 2\xi(\hat{h}_{\hat{f}}, f\,),$$

where

$$\xi(h, f\,) \equiv \int [\hat{f}_h(x) - \hat{g}_h(x)][\hat{f}(x) - f(x)]\,dx.$$

Expand $\Delta^{(1)}(\hat{h}_{\hat{f}}, f\,)$ in a Taylor series about $\hat{h}_f$ (the minimizer of $\Delta^{(1)}(h, f\,)$),

$$0 = \Delta^{(1)}(\hat{h}_{\hat{f}}, f\,) = (\hat{h}_{\hat{f}} - \hat{h}_f)\Delta^{(2)}(h^*, f\,),$$

where h^* lies between $\hat{h}_{\hat{f}}$ and $\hat{h}_{f_1}$. The key to the proof of (5.1) is the following combination of the last two expressions:

$$\hat{h}_{\hat{f}} - \hat{h}_f = \frac{-2\xi(\hat{h}_{\hat{f}}, f\,)}{\hat{h}_{\hat{f}} \Delta^{(2)}(h^*, f\,)}. \tag{5.3}$$

To use this, first choose $0 < a_1 < b_1 < \infty$ so that $2a_1 \le n^{1/5} h_f \le b_1/2$ for all n and all $f \in \mathscr{F}_n$, by (6.1) of Lemma 6.1. Define $\hat{h}'_{\hat{f}} = \hat{h}_{\hat{f}}$ if $a_1 n^{-1/5} < \hat{h}_{\hat{f}} < b_1 n^{-1/5}$, and $\hat{h}'_{\hat{f}} = (a_1 + b_1)n^{-1/5}/2$ otherwise. The fact that the numerator of

(5.3) is pivotal there is demonstrated by:

LEMMA 5.1. *Given $\eta_1 > 0$, we may choose $\eta_2 > 0$ and a sequence $f_1 = f_{1n} \in \mathscr{F}_n$ such that, for all large n,*

$$P_{f_1}\left[|\xi(\hat{h}'_{\hat{f}}, f_1)| > \eta_2 n^{-9/10}\right] > 1 - \eta_1.$$

The proof of Lemma 5.1 is in Section 7.

We now work with the "worst-case" density $f_1 = f_{1n}$ for some fixed $\eta_1, \eta_2 > 0$. For this choice of f, define $\hat{h}^\dagger = h^*$ (h^* as in (5.3)) if $|\hat{h}_{\hat{f}} - \hat{h}_{f_1}| \leq n^{-1/4}$, and $\hat{h}^\dagger = h_{f_1}$, otherwise. The fact that the denominator of (5.3) is no problem follows from:

LEMMA 5.2. *There is an $\eta_3 > 0$, so that*

$$P_{f_1}\left[0 < \Delta^{(2)}(\hat{h}^\dagger, f_1) < \eta_3 n^{-2/5}\right] \to 1.$$

The proof of Lemma 5.2 is in Section 8.

To finish the proof of (5.1), let $\eta_4 \equiv 2(b_1\eta_3)^{-1}\eta_2$. Define the events

$$\begin{aligned}
\mathscr{E}_1 &= \left\{|\hat{h}_{\hat{f}} - \hat{h}_{f_1}| \leq n^{-1/4}\right\},\\
\mathscr{E}_2 &= \left\{|\Delta^{(2)}(\hat{h}^\dagger, f_1)| \leq \eta_3 n^{-2/5}\right\},\\
\mathscr{E}_3 &= \left\{\hat{h}_{\hat{f}} \in n^{-1/5}(a_1, b_1)\right\}.
\end{aligned}$$

By (5.3),

$$\begin{aligned}
&P_{f_1}\left[|\hat{h}_{\hat{f}} - \hat{h}_{f_1}| > \eta_4 n^{-3/10};\ \mathscr{E}_1\right]\\
&\quad\geq P_{f_1}\left[\left|\frac{-2\xi(\hat{h}_{\hat{f}}, f_1)}{\hat{h}_{\hat{f}}}\right| > \eta_3\eta_4 n^{-7/10};\ \mathscr{E}_1\right] - P_{f_1}\left[\tilde{\mathscr{E}}_2\right]\\
&\quad\geq P_{f_1}\left[|2\xi(\hat{h}_{\hat{f}}, f_1)| > b_1\eta_3\eta_4 n^{-9/10};\ \mathscr{E}_1\right] - P_{f_1}\left[\tilde{\mathscr{E}}_2\right] - P_{f_1}\left[\tilde{\mathscr{E}}_3\right]\\
&\quad\geq P_{f_1}\left[|\xi(\hat{h}'_{\hat{f}}, f_1)| > \eta_2 n^{-9/10};\ \mathscr{E}_1\right] - P_{f_1}\left[\tilde{\mathscr{E}}_2\right] - P_{f_1}\left[\tilde{\mathscr{E}}_3\right]\\
&\quad\geq P_{f_1}[\mathscr{E}_1] - P_{f_1}\left[\tilde{\mathscr{E}}_2\right] - P_{f_1}\left[\tilde{\mathscr{E}}_3\right] - \eta_1,
\end{aligned} \tag{5.4}$$

the last line following from Lemma 5.1. Lemma 5.2 implies that $P_{f_1}[\tilde{\mathscr{E}}_2] \to 0$, and $P_{f_1}[\tilde{\mathscr{E}}_3] \to 0$ follows from (6.1) of Lemma 6.1 together with Lemma 6.2. If n is so large that $\eta_4 n^{-3/10} < n^{-1/4}$, then by (5.4),

$$\begin{aligned}
&P_{f_1}\left[|\hat{h}_{\hat{f}} - \hat{h}_{f_1}| > \eta_4 n^{-3/10}\right]\\
&\quad= P_{f_1}\left[|\hat{h}_{\hat{f}} - \hat{h}_{f_1}| > \eta_4 n^{-3/10};\ \mathscr{E}_1\right] + P_{f_1}\left[\tilde{\mathscr{E}}_1\right]\\
&\quad\geq \left\{P_{f_1}[\mathscr{E}_1] - P_{f_1}\left[\tilde{\mathscr{E}}_2\right] - P_{f_1}\left[\tilde{\mathscr{E}}_3\right] - \eta_1\right\} + P_{f_1}\left[\tilde{\mathscr{E}}_1\right]\\
&\quad= 1 - P_{f_1}\left[\tilde{\mathscr{E}}_2\right] - P_{f_1}\left[\tilde{\mathscr{E}}_3\right] - \eta_1 \to 1 - \eta_1
\end{aligned}$$

as $n \to \infty$. This proves (5.1) and hence (4.3). To establish (4.2), note that

$$\Delta(\hat{h}, f) - \Delta(\hat{h}_f, f) = \tfrac{1}{2}(\hat{h} - \hat{h}_f)\Delta^{(2)}(h^*, f),$$

where h^* is between $\hat{h}$ and $\hat{h}_f$. Now apply (6.3) of Lemma 6.1, Lemma 6.5, and the preceding methods.

6. Statement of additional lemmas. The proofs of the following lemmas are in the Appendix:

LEMMA 6.1. *For some* $n_0 > 0$,

$$0 < \inf_{n \geq n_0,\, f \in \mathscr{F}_n} n^{1/5} h_f \leq \sup_{n \geq n_0,\, f \in \mathscr{F}_n} n^{1/5} h_f < \infty; \tag{6.1}$$

for any $\varepsilon > 0$ *there exists* $\eta = \eta(\varepsilon)$ *such that*

$$\inf_{|h - h_f| > \varepsilon n^{-1/5}} M(h_f, f) \geq (1 + \eta) M(h_f, f) \tag{6.2}$$

for all $f \in \mathscr{F}_n$ *and all large* n; *for some* $n_0 > 0$,

$$0 < \inf_{n \geq n_0,\, f \in \mathscr{F}_n} n^{2/5} M^{(2)}(h_f, f) \leq \sup_{n \geq n_0,\, f \in \mathscr{F}_n} n^{2/5} M^{(2)}(h_f, f) < \infty; \tag{6.3}$$

for any $\varepsilon > 0$ *there exists* $\eta = \eta(\varepsilon)$ *such that*

$$\inf_{|h - h_f| \leq \varepsilon n^{-1/5}} |M^{(2)}(h, f) - M^{(2)}(h_f, f)| \leq \eta(\varepsilon) n^{-2/5} \tag{6.4}$$

for all $f \in \mathscr{F}_n$ *and all large* n, *and* $\eta(\varepsilon) \to 0$ *as* $\varepsilon \to 0$.

LEMMA 6.2.

$$\lim_{\lambda \to \infty} \limsup_{n \to \infty} \sup_{f \in \mathscr{F}_n} P_f\left[|\hat{h}_f - h_f| > \lambda n^{-3/10}\right] = 0.$$

LEMMA 6.3.

$$\lim_{\lambda \to \infty} \limsup_{n \to \infty} \sup_{f \in \mathscr{F}_n} P_f\left[|\hat{h}_c - h_f| > \lambda n^{-3/10}\right] = 0.$$

LEMMA 6.4. *For each* $0 < a < b < \infty$ *and each* $\varepsilon > 0$,

$$\sup_f P_f\left\{ \sup_{v;\, a \leq t \leq b} \left| \int_{C_v} [\hat{f}_h(x) - f(x)] \gamma(x)\, dx \right| > n^{-1+\varepsilon} \right\} \to 0,$$

$$\sup_f P_f\left\{ \sup_{v;\, a \leq t \leq b} \left| \int_{C_v} [\hat{g}_h(x) - g(x)] \gamma(x)\, dx \right| > n^{-1+\varepsilon} \right\} \to 0.$$

LEMMA 6.5.

$$\sup_{f \in \mathscr{F}_n} P_f\left[n^{2/5} |\Delta^{(2)}(h_f, f) - M^{(2)}(h_f, f)| > \varepsilon\right] \to 0.$$

7. Proof of Lemma 5.1. The proof is via a sequence of three lemmas. Let P_0 be the probability measure defined by

$$P_0[\mathscr{E}] \equiv 2^{-m} \sum_{f \in \mathscr{F}_n} P_f[\mathscr{E}],$$

and let E_0 denote expectation with respect to P_0. Under P_0, f should be regarded as a random variable. There are precisely 2^m elements in $\mathscr{F}_n$. Writing

$$f = (1 + \Sigma_v \tau_v \gamma_v) f_0, \qquad \hat{f} = (1 + \Sigma_v \hat{\tau}_v \gamma_v) f_0,$$

for sequences $\{\tau_v\}$ and $\{\hat{\tau}_v\}$ of 0's and 1's, we see that

$$\xi\big(\hat{h}_f', f\big) = -c_0 S,$$

where c_0 is the constant value taken by f_0 on $\bigcup_v C_v$,

$$S \equiv \Sigma_v(\tau_v - \hat{\tau}_v)\hat{w}_v$$

and

$$\hat{w}_v \equiv \int_{C_v} \Big[\hat{f}_{\hat{h}_f'}(x) - \hat{g}_{\hat{h}_f'}(x)\Big]\gamma(x)\,dx.$$

(The function γ was defined in Section 4.) Notice that S depends on f only through the indicators τ_v; this observation is crucial to our argument.

Let $\mathscr{X}$ denote the sample $X_1, \dots, X_n$. Under the probability measure P_0, and conditional on $\mathscr{X}$, the τ_v's are independent Bernoulli random variables with

$$(7.1) \quad \hat{q}_v \equiv P_0[\tau_v = 1|\mathscr{X}] = \{\Pi^{(v)}[1 + \gamma(X_1)]\}/\{1 - \Pi^{(v)}[1 + \gamma(X_1)]\},$$

where $\Pi^{(v)}$ denotes the product over indices i with $X_i \in C_v$. Thus,

$$\hat{\mu} \equiv E_0[S|\mathscr{X}] = \Sigma_v(\hat{q}_v - \hat{\tau}_v)\hat{w}_v,$$

$$\hat{\sigma}^2 \equiv \mathrm{var}_0[S|\mathscr{X}] = \Sigma_v \hat{q}_v(1 - \hat{q}_v)\hat{w}_v^2,$$

$$\hat{\beta} \equiv \Sigma_v E_0\Big[|(\tau_v - \hat{q}_v)\hat{w}_v|^3 |\mathscr{X}\Big] \le \Sigma_v |\hat{w}_v|^3.$$

The next two lemmas describe asymptotic properties of $\hat{\sigma}^2$ and $\hat{\beta}$.

LEMMA 7.1. *There exist fixed constants* $0 < d_1 < d_2 < \infty$ *such that*

$$P_0\big[d_1 n^{-9/5} < \hat{\sigma}^2 < d_2 n^{-9/5}\big] \to 1.$$

LEMMA 7.2. *For each* $\varepsilon > 0$,

$$P_0\big[\hat{\beta} > n^{\varepsilon - 14/5}\big] \to 0.$$

The proofs of these lemmas, and the next lemma, are postponed to the end of this section. Let Φ be the standard normal distribution function.

LEMMA 7.3. *For some fixed* $c > 0$,

$$\liminf_{n \to \infty} \inf_{x > 0} \big\{P_0\big[|S| > n^{-9/10}x\big] - 2[1 - \Phi(cx)]\big\} \ge 0.$$

To finish the proof of Lemma 5.1, choose $x > 0$ so small that

$$2[1 - \Phi(cx)] > 1 - \eta_1/2,$$

and let $\eta_2 = c_0^{-1}x$. Then for large n,

$$1 - \eta_1 \le P_0\left[|c_0 S| > \eta_2 n^{-9/10}\right] = 2^{-m} \sum_{f \in \mathscr{F}_n} P_f\left[|\xi(\hat{h}_f', f)| > \eta_2 n^{-9/10}\right].$$

Therefore, there must exist some $f_1 \in \mathscr{F}_n$ such that

$$1 - \eta_1 \le P_{f_1}\left[|\xi(\hat{h}_f', f)| > \eta_2 n^{-9/10}\right].$$

Proof of Lemma 7.1. Let N_v denote the number of elements of $\mathscr{X}$ within C_v, and notice that the P_f distribution of the sequence $\{N_v\}$ does not depend on f. Observe that for a constant $c > 0$, $E_f(N_v) = E_f(N_1) \sim Cn^{4/5}$. Therefore, for large n,

$$\begin{aligned} P_f\left[N_v > 3cn^{4/5} \text{ for some } v\right] &\le Cn^{1/5} P_f\left[N_1 > 3cn^{4/5}\right] \\ &\le Cn^{1/5} P_f\left[|N_1 - E_f(N_1)| > cn^{4/5}\right] \\ &\le Cn^{1/5}(cn^{4/5})^{-2} E_f\left[|N_1 - E_f(N_1)|^2\right] \\ &= O(n^{-3/5}). \end{aligned}$$

Thus, if $\mathscr{E}$ is the event that no interval C_v contains more than $3cn^{4/5}$ elements of $\mathscr{X}$, then $\inf_{f \in \mathscr{F}_n} P_f[\mathscr{E}] \to 1$.

Let $\Sigma^{(v)}$ denote summation over indices i with $X_i \in C_v$, and observe that

$$\begin{aligned} \Pi^{(v)}[1 + \gamma(X_i)] &= \exp\left\{ \sum_{j=1}^{\infty} (-1)^{j+1} j^{-1} \Sigma^{(v)} [\gamma(X_i)]^j \right\} \\ &= \exp\left(T_1^{(v)} + T_2^{(v)}\right), \end{aligned}$$

where

$$\begin{aligned} T_1^{(v)} &\equiv \Sigma^{(v)} \gamma(X_i), \\ |T_2^{(v)}| &\le \sum_{j=2}^{\infty} j^{-1} \Sigma^{(v)} |\gamma(X_i)|^j \equiv T_3^{(v)}. \end{aligned}$$

Bearing in mind that $\sup|\gamma| \le C_1 n^{-2/5}$, we may easily prove that on the set $\mathscr{E}$ and for all large n, $T_3^{(v)} \le C_2$ uniformly in v.

Thus, for each $z > 0$ there exist numbers $0 < a_2(z) < b_2(z) < \infty$ such that, on the set $\{|T_1^{(v)}| \le z\} \cap \mathscr{E}$,

$$a_2(z) \le \Pi^{(v)}[1 + \gamma(X_1)] \le b_2(z),$$

for all v. Remembering the definition (7.1) of $\hat{q}_v$ in terms of $\Pi^{(v)}[1 + \gamma(X_1)]$, we now deduce the existence of a positive decreasing function $a(z) \le \frac{1}{2}$, such that on $\mathscr{E}$, $|T_1^{(v)}| \le z$ implies $|\hat{q}_v - \frac{1}{2}| \le \frac{1}{2} - a(z)$. Therefore, on $\mathscr{E}$,

$$[a(z)/2] \Sigma_v \hat{w}_v^2 1_{\{|\Sigma^{(v)} \gamma(X_i)| \le z\}} \le \hat{\sigma}_v^2 \le \Sigma_v \hat{w}_v^2/4, \tag{7.2}$$

for all $z > 0$.

Let

$$\hat{w}_v(h, z) \equiv 1_{\{|\Sigma^{(v)}\gamma(X_i)| \le z\}} \int_{C_v} \left[\hat{f}_h(x) - \hat{g}_h(x)\right]\gamma(x)\,dx,$$

$\mu_v(h, z, f) \equiv E_f[\hat{w}_v^2(h, z)]$ and $\mu(h, z, f) \equiv \Sigma_v \mu_v(h, z, f)$. We claim that the function $c(n, h, z, f)$ defined by $\mu(h, z, f) = c(n, h, z, f)n^{-9/5}$, is bounded away from zero and infinity uniformly in $n \ge 1$, $h \in n^{-1/5}(a_1, b_1)$, $z \ge z_0$ and $f \in \mathscr{F}_n$, for some $z_0 > 0$. This is relatively easy to verify if we take $z_0 = \infty$. To see that $z_0 < \infty$ is permissible, notice that

$$\mu_v(h, z, f) \ge \mu_v(h, \infty, f) - \left[P_f\{|\Sigma^{(v)}\gamma(X_i)| > z\}\right]^{1/2}\left[E_f\{\hat{w}_v^4(h, \infty)\}\right]^{1/2};$$

$$E_f\left[\hat{w}_v^4(h, \infty)\right] \le C_1 n^{-4},$$

uniformly in $v = 0, \dots, m-1$, $h \in n^{-1/5}(a_1, b_1)$ and $f \in \mathscr{F}_n$. Also observe that

$$\begin{aligned} P_f\{|\Sigma^{(v)}\gamma(X_i)| > z\} &\le z^{-2}E_f\left[|\Sigma^{(v)}\gamma(X_i)|^2\right] \\ &= z^{-2}E_f\big(N_v E\left[\gamma^2(X_1)|X_1 \in C_v\right] \\ &\qquad -(N_v^2 - N_v)E\left[\gamma(X_1)\gamma(X_2)|X_1, X_2 \in C_v\right]\big) \\ &\le z^{-2}C_2\left[E_f(N_v)\|C_v\|^{-1}\int_{C_v}\gamma^2(x)\,dx\right. \\ &\qquad \left.-E_f(N_v^2)\left(\|C_v\|^{-1}\int_{C_v}\tau_v\gamma^2(x)\,dx\right)^2\right] \\ &\le z^{-2}C_2, \end{aligned}$$

uniformly in v and f, where $\|C_v\|$ denotes the length of C_v. Consequently,

$$\mu_v(h, z, f) \ge \mu_v(h, \infty, f) - (C_1C_2)^{1/2}n^{-2}z^{-1}$$

uniformly in v, h, z and f. Adding this inequality over v, we see that the stated properties of the function c are available for some finite $x_0 > 0$.

Take $z = z_0$ in (7.2). In view of the properties of c established in the previous paragraph, Lemma 7.1 will follow from (7.2) if we prove that for each $\varepsilon > 0$, and for $z = z_0$ and $z = \infty$,

$$\sup_{f \in \mathscr{F}_n} P_f\left[\sup_{a_1 \le t \le b_1} |\Sigma_v\left[\hat{w}_v^2(n^{-1/5}t, z) - \mu_v(n^{-1/5}t, z, f)\right]| > \varepsilon n^{-9/5}\right] \to 0. \tag{7.3}$$

Using Hölder continuity of K and L, and the fact that these functions have compact support, we may choose $\lambda > 0$ so large that

$$\begin{aligned} &\Sigma_v\left[|\hat{w}_v^2(n^{-1/5}s, z) - \hat{w}_v^2(n^{-1/5}t, z)|\right. \\ &\qquad \left.+|\mu_v(n^{-1/5}s, z, f) - \mu_v(n^{-1/5}t, z, f)|\right] \\ &\qquad \le Cn^{-2}, \end{aligned} \tag{7.4}$$

uniformly in n, $z = z_0$ and ∞, $f, v, a_1 \le s \le t \le b_1$ with $|s - t| \le n^{-\lambda}$, and samples $X_1, \dots, X_n$. Let $a_1 = t_0 < t_1 < \cdots < t_{\nu-1} \le b_1 < t_\nu$ be a partition of (a_1, b_1) with $t_i - t_{i-1} = n^{-\lambda}$ for each i. In view of (7.4), result (7.3) will follow if we show that for each $\varepsilon > 0$,

$$p_f \equiv \Sigma_i P_f\{|\Sigma_v[\hat{w}_v^2(n^{-1/5}t_i, z) - \mu_v(n^{-1/5}t_i, z, f)]| > \varepsilon n^{-9/5}\}$$

converges to zero uniformly in $f \in \mathscr{F}_n$ for $z = z_0$ and ∞.

Since K and L have compact support, and each $t_i \in (a_i, b_i)$, then for each i we may divide the subscripts v among a fixed finite number k (not depending on i or n) of sets $V_{i1}, \dots, V_{ik}$ such that for each i and j, and for $z = z_0$ and ∞, the variables $\hat{w}_v^2(n^{-1/5}t_i, z)$, $v \in V_{ij}$, are stochastically independent, and for each i, each subscript v is contained in just one set V_{ij}. Consequently, for all integers $l \ge 1$,

$$p_f \le \Sigma_i \Sigma_j P_f\left\{\left|\sum_{v \in V_{ij}} [\hat{w}_v^2(n^{-1/5}t_i, z) - \mu_v(n^{-1/5}t_i, z, f)]\right| > \varepsilon k^{-1} n^{-9/5}\right\}$$

$$\le \Sigma_i \Sigma_j E_f\left\{\left|(\varepsilon k^{-1} n^{-9/5})^{-1} \sum_{v \in V_{ij}} [\hat{w}_v^2(n^{-1/5}t_i, z) - \mu_v(n^{-1/5}t_i, z, f)]\right|^{2l}\right\}.$$

An inequality for moments of sums of independent random variables (see Burkholder (1973), formula (21.4)) now gives

$$p_f \le C_1(l)(\varepsilon^{-1}k)^{2l} n^{18l/5} \Sigma_i \Sigma_j \left\{\left[\sum_{v \in V_{ij}} E_f(Y_{iv}^2)\right]^l + \sum_{v \in V_{ij}} E_f(|Y_{iv}|^{2l})\right\},$$

where $Y_{iv} \equiv \hat{w}_v^2(n^{-1/5}t_i, z) - \mu_v(n^{-1/5}t_i, z, f)$. The same moment inequality gives $E_f(|Y_{iv}|^{2l}) \le C_2 n^{-4l}$ uniformly in i, v and f. Since the number of partition points is of order n^λ,

$$\sup_f p_f \le C_3 n^{18l/5} \Sigma_i[(n^{1/5}n^{-4})^l + n^{1/5}n^{-4l}] = O(n^{\lambda - l/5}) \to 0,$$

provided only that $l > 5\lambda$. This completes the proof of Lemma 7.1. □

PROOF OF LEMMA 7.2. The argument used to prove Lemma 7.1 shows that for some $c_3 > 0$,

$$P_0\{\Sigma_v \hat{w}_v^2 > c_3 n^{-9/5}\} \to 0.$$

Lemma 6.4 gives

$$P_0\left\{\sup_v |\hat{w}_v| > n^{-1+\varepsilon}\right\} \to 0.$$

Lemma 7.2 follows on combining these results. □

PROOF OF LEMMA 7.3. Let $\mathscr{E}$ denote the event that $\hat{\beta}\hat{\sigma}^{-3} \le n^{-1/20}$. According to Lemmas 7.1 and 7.2,

$$P_0(\tilde{\mathscr{E}}) \le P_0[\hat{\sigma}^2 \le d_1 n^{-9/5}] + P_0[\hat{\beta} > n^{-1/20}(d_1 n^{-9/5})^{3/2}] \to 0.$$

On the set $\mathscr{E}$ the Berry–Esseen bound (see Petrov (1978), page 111) gives

$$\sup_{-\infty < x < \infty} |P_0[S \le x|\mathscr{X}] - \Phi((x - \hat{\mu})/\hat{\sigma})| \le A n^{-1/20},$$

where A is an absolute constant. Therefore, on $\mathscr{E}$ and for $x > 0$,

$$\begin{aligned} P_0(|S| > x|\mathscr{X}) &\ge 1 - \Phi((x - \hat{\mu})/\hat{\sigma}) + \Phi((-x - \hat{\mu})/\hat{\sigma}) - 2An^{-1/20} \\ &\ge 2[1 - \Phi(x/\hat{\sigma})] - 2An^{-1/20}. \end{aligned}$$

Taking expectations, and using Lemma 7.1 again, finishes the proof of Lemma 7.3. □

8. Proof of Lemma 5.2. Note that Lemma 5.2 will follow from (6.3) of Lemma 6.1 together with, for any $\varepsilon > 0$,

$$P_{f_1}\left\{n^{2/5}|\Delta^{(2)}(\hat{h}^\dagger, f_1) - M^{(2)}(h_{f_1}, f_1)| > \varepsilon\right\} \to 0.$$

Hence, by Lemma 6.5, it is enough to show that

$$P_{f_1}\left\{n^{2/5}|\Delta^{(2)}(\hat{h}^\dagger, f_1) - \Delta^{(2)}(h_{f_1}, f_1)| > \varepsilon\right\} \to 0. \tag{8.1}$$

To verify (8.1), first recall our definition of $\hat{h}^\dagger$, which gives

$$|\hat{h}^\dagger - h_{f_1}| \le n^{-1/4} + |\hat{h}_{f_1} - h_{f_1}|.$$

Therefore, by Lemma 6.2,

$$P_{f_1}\left\{|\hat{h}^\dagger - h_{f_1}| > 2n^{-1/4}\right\} \to 0. \tag{8.2}$$

It now follows from (6.4) of Lemma 6.1 that

$$P_{f_1}\left\{n^{2/5}|M^{(2)}(\hat{h}^\dagger, f_1) - M^{(2)}(h_{f_1}, f_1)| > \varepsilon\right\} \to 0,$$

and since (by (8.2))

$$P_{f_1}\left\{\hat{h}^\dagger, h_{f_1} \in n^{-1/5}(a_1, b_1)\right\} \to 1,$$

the proof of (8.1), and hence that of Lemma 5.2, will be completed if we show that

$$P_{f_1}\left\{n^{2/5}|D^{(2)}(\hat{h}^\dagger, f_1)| > \varepsilon\right\} \to 0, \tag{8.3}$$

$$P_{f_1}\left\{n^{2/5}|D^{(2)}(h_{f_1}, f_1)| > \varepsilon\right\} \to 0, \tag{8.4}$$

where

$$D(h, f) \equiv \Delta(h, f) - M(h, f). \tag{8.5}$$

Statements (8.3) and (8.4) are immediate consequences of

$$P_{f_1}\left\{\sup_{h \in n^{-1/5}(a_1, b_1)} n^{2/5}|D^{(2)}(h, f_1)| > \varepsilon\right\} \to 0,$$

which may be verified by a partitioning argument that is similar to, but easier than, that used in the verification of (7.3).

APPENDIX

Proofs of Lemmas in Section 6.

OUTLINE OF PROOF OF LEMMA 6.1. Write

$$M(h, f) = V(h, f) + B(h, f),$$

where

$$V(h, f) = n^{-1}h^{-1}\int\int K(u)^2 f(x - hu)\,du\,dx$$

$$-n^{-1}\int\left[\int K(u)f(x - hu)\,du\right]^2 dx,$$

$$B(h, f) = \int\left(\int K(u)[f(x - hu) - f(x)]\,du\right)^2 dx.$$

The derivatives $M^{(1)}(h, f)$ and $M^{(2)}(h, f)$ may be studied by differentiating $V(h, f)$, then approximating as in Rosenblatt (1971), and by differentiating $B(h, f)$ and using a Taylor expansion with integral form of the remainder.

PROOFS OF LEMMAS 6.2 AND 6.3. First note that, for D as in (8.5) and δ as in Hall and Marron (1985):

LEMMA A.1. *For each* $0 < a < b < \infty$ *and all positive integers* l,

$$\text{(A.1)}\qquad \sup_{n,\, f\in\mathscr{F}_n,\, a\le t\le b} E_f|n^{7/10}D^{(1)}(n^{-1/5}t, f)|^{2l} \le C_1(a, b, l),$$

$$\text{(A.2)}\qquad \sup_{n,\, f\in\mathscr{F}_n,\, a\le t\le b} E_f|n^{7/10}\delta^{(1)}(n^{-1/5}t, f)|^{2l} \le C_1(a, b, l).$$

Furthermore, there exists $\varepsilon_1 > 0$, *not depending on* a, b *or* l, *such that*

$$\text{(A.3)}\qquad E_f|n^{7/10}[D^{(1)}(n^{-1/5}s, f) - D^{(1)}(n^{-1/5}t, f)|^{2l} \le C_2(a, b, l)|s - t|^{\varepsilon_1 l},$$

$$\text{(A.4)}\qquad E_f|n^{7/10}[\delta^{(1)}(n^{-1/5}s, f) - \delta^{(1)}(n^{-1/5}t, f)|^{2l} \le C_2(a, b, l)|s - t|^{\varepsilon_1 l},$$

for all $f \in \mathscr{F}_n$ *and* $a \le s \le t \le b$.

LEMMA A.2. *For some* $\varepsilon > 0$ *and any* $0 < a < b < \infty$,

$$\text{(A.5)}\qquad \sup_{f\in\mathscr{F}_n} P_f\left\{\sup_{a\le t\le b}\left[|D^{(1)}(n^{-1/5}t, f)| + |\delta^{(1)}(n^{-1/5}t, f)|\right] > n^{-3/5-\varepsilon}\right\} \to 0.$$

Furthermore, for any $\varepsilon_2 > 0$ *and* $\eta_2 > 0$,

$$\text{(A.6)}\qquad \begin{aligned} \sup_{f\in\mathscr{F}_n} P_f\bigg\{ \sup_{|t - n^{1/5}h_f| \le n^{-\varepsilon_2}} & n^{7/10}\Big[|D^{(1)}(n^{-1/5}t, f) - D^{(1)}(h_f, f)| \\ & + |\delta^{(1)}(n^{-1/5}t, f) - \delta^{(1)}(h_f, f)|\Big] > n^{-3/5-\varepsilon}\bigg\} \to 0. \end{aligned}$$

The proofs of Lemmas A.1 and A.2 are omitted because they closely parallel the proofs of Lemmas 3.1 and 3.2 in Hall and Marron (1985).

LEMMA A.3. *For any* $\varepsilon > 0$,

$$\sup_{f \in \mathscr{F}_n} P_f\left[|\hat{h}_f - h_f| > \varepsilon n^{-1/5}\right] \to 0.$$

PROOF. It suffices to show that for any sequence of choices $f_1 = f_{1n} \in \mathscr{F}_n$, and for each $\varepsilon > 0$,

$$P_{f_1}\left[|\hat{h}_{f_1} - h_{f_1}| > \varepsilon n^{-1/5}\right] \to 0. \tag{A.7}$$

We may easily prove that for some $b > 0$, $P_{f_1}[n^{-b} \le \hat{h}_{f_1} \le n^b] \to 1$. Let $H = H_n$ be a set of bandwidths in the range $[n^{-b}, n^b]$, such that $\#(H) \le n^a$ for some $a > 0$. Arguing as in the proofs of Lemmas 2 and 4 of Stone (1984), we may show that for each $\varepsilon > 0$,

$$P_{f_1}\left[\sup_{h \in H} |\Delta(h, f_1) - M(h, f_1)|/M(h, f_1) > \varepsilon\right] \to 0. \tag{A.8}$$

Now use Hölder continuity of K to show that for any (random) bandwidth $\tilde{h}$ with $P_{f_1}[n^{-b} \le \tilde{h} \le n^b] \to 1$,

$$P_{f_1}\left[|\Delta(\tilde{h}, f_1) - M(\tilde{h}, f_1)|/M(\tilde{h}, f_1) > \varepsilon\right] \to 0.$$

Finally invoke (6.2) of Lemma 6.1 to obtain (A.7). □

LEMMA A.4. *For any* $\varepsilon > 0$,

$$\sup_{f \in \mathscr{F}_n} P_f\left[|\hat{h}_c - h_f| > \varepsilon n^{-1/5}\right] \to 0.$$

PROOF. Again, it suffices to prove that for any $\varepsilon > 0$ and sequence $f_1 = f_{1n} \in \mathscr{F}_n$,

$$P_{f_1}\left[|\hat{h}_c - h_{f_1}| > \varepsilon n^{-1/5}\right] \to 0, \tag{A.9}$$

and it is easily shown that for some $b > 0$, $P_f[n^{-b} \le \hat{h}_c \le n^b] \to 1$. Define

$$CV(h, f) \equiv CV(h) + \int f^2 + 2n^{-1}(n-1)^{-1}(n+1) \sum_{i=1}^{n} \left[f(X_i) - E_f f(X_i)\right]$$

and let H be as in the proof of Lemma A.3. Minimising CV is equivalent to minimising $CV(\cdot, f)$, for any f. Using the argument leading to Stone's (1984) Lemmas 2, 3 and 4, we may show that for any $\varepsilon > 0$,

$$P_{f_1}\left[\sup_{h \in H} |CV(h, f_1) - M(h, f_1)|/M(h, f_1) > \varepsilon\right] \to 0.$$

This formula serves as an analogue of (A.8) in the proof of Lemma A.3. The proof of (A.9) may now be completed as was that proof. □

LEMMA A.5. *For some* $\varepsilon > 0$,

$$\sup_{f \in \mathscr{F}_n} P_f\left[|\hat{h}_f - h_f| + |\hat{h}_c - h_f| > n^{-1/5-\varepsilon}\right] \to 0.$$

PROOF. Argue as in Lemma 3.3 of Hall and Marron (1985), but use Lemmas 3.4 and 3.5 to replace the limit theorems $\hat{h}_0/h_0 \to_p 1$ and $\hat{h}_c/h_0 \to_p 1$ (in the notation of that paper), and use our Lemma A.2 in place of Lemma 3.2 there. □

To finish the proof of Lemma 6.2, note that it suffices to show that for any sequences $f_1 = f_{1n} \in \mathscr{F}_n$ and $\lambda_n \to \infty$,

$$P_{f_1}\left[|\hat{h}_{f_1} - h_{f_1}| > \lambda_n n^{-3/10}\right] = 0. \tag{A.10}$$

Observe that

$$\begin{aligned} 0 = \Delta^{(1)}\left(\hat{h}_{f_1}, f_1\right) &= M^{(1)}\left(\hat{h}_{f_1}, f_1\right) + D^{(1)}\left(\hat{h}_{f_1}, f_1\right) \\ &= \left(\hat{h}_{f_1} - h_{f_1}\right)M^{(2)}(h^*, f_1) + D^{(1)}\left(\hat{h}_{f_1}, f_1\right), \end{aligned} \tag{A.11}$$

where h^* lies in between $\hat{h}_{f_1}$ and h_{f_1}. Define $c_1 = c_1(n)$ and $c_2 = c_2(n)$ by $h_{f_1} \equiv c_2 n^{-1/5}$ and $M^{(2)}(h_{f_1}, f_1) \equiv c_2 n^{-2/5}$. Then c_1 and c_2 are bounded away from zero and infinity as $n \to \infty$ (note (6.1) and (6.3) from Lemma 6.1). Given any $\xi > 0$, there exists $\eta(\xi) > 0$ such that $\eta(\xi) \to 0$ as $\xi \to 0$ and for large n,

$$\sup_{|h - h_{f_1}| \le \xi n^{-1/5}} |M^{(2)}(h, f_1) - M^{(2)}(h_{f_1}, f_1)| \le \eta(\xi) n^{-2/5}.$$

(Note (6.4) of Lemma 6.1.) Let a_1, b_1 be fixed positive lower and upper bounds to c_1, respectively, and let a_2 be a fixed positive lower bound to c_2. Choose $\xi \in (0, a_1/2)$ so small that $\eta(\xi) \le a_2/2$. By (A.11),

$$\begin{aligned} |\hat{h}_{f_1} - h_{f_1}| &\le \left(a_2 n^{-2/5}/2\right)^{-1} |D^{(1)}\left(\hat{h}_{,,}, f_1\right)| \\ &\le 2a_2^{-1} n^{2/5} \sup_{a_1/2 \le t \le a_1/2 + b_1} |D^{(1)}\left(n^{-1/5}t, f_1\right)|, \end{aligned}$$

whenever the event $\mathscr{E}_1 \equiv \{|\hat{h}_{f_1} - h_{f_1}| \le \xi n^{-1/5}\}$ holds. Let $\mathscr{E}_2$ be the event $\{\sup_{a \le t \le b} |D^{(1)}(n^{-1/5}t, f_1)| \le n^{-3/5-\varepsilon}\}$, where $a = a_1/2$, $b = a_1/2 + b_1$ and ε is as in (6.4) of Lemma A.2. Whenever $\mathscr{E}_1 \cap \mathscr{E}_2$ holds, so does the event $\mathscr{E}_3 \equiv \{|\hat{h}_{f_1} - h_{f_1}| \le 2a_2^{-1} n^{-1/5-\varepsilon}\}$. Let $\mathscr{E}_4$ be the event that

$$|D^{(1)}\left(\hat{h}_{f_1}, f_1\right) - D^{(1)}\left(h_{f_1}, f_1\right)| > n^{-7/10}.$$

Then

$$\begin{aligned} &P_{f_1}\left[|\hat{h}_{f_1} - h_{f_1}| > \lambda_n n^{-3/10}\right] \\ &\quad \le P_{f_1}\left[\tilde{\mathscr{E}}_1\right] + P_{f_1}\left[\tilde{\mathscr{E}}_2\right] + P_{f_1}\left[\mathscr{E}_3 \cap \mathscr{E}_4\right] \\ &\qquad + P_{f_1}\left[|D^{(1)}\left(h_{f_1}, f_1\right)| > \lambda_n n^{-3/10}\left(2a_2^{-1} n^{2/5}\right)^{-1} - n^{-7/10}\right]. \end{aligned} \tag{A.12}$$

Chebyshev's inequality and (A.1) of Lemma A.1 show that the last-written probability converges to zero as $n \to \infty$. Lemma A.3 gives $P_{f_1}[\mathscr{E}_3 \cap \mathscr{E}_4] \to 0$. Result (A.10) follows from (A.12), which finishes the proof of Lemma 6.2. □

To finish the proof of Lemma 6.3 use essentially the argument employed to prove Lemma 6.2, but replace (A.11) by

$$\begin{aligned} 0 = CV^{(1)}(\hat{h}_c) &= M^{(1)}(\hat{h}_c, h_f) + D^{(1)}(\hat{h}_c, f_1) + \delta^{(1)}(\hat{h}_c, f_1) \\ &= (\hat{h}_c - h_{f_1})M^{(2)}(h^*, f_1) + D^{(1)}(\hat{h}_c, f_1) + \delta^{(1)}(\hat{h}_c, f_1), \end{aligned}$$

where h^* lies in between $\hat{h}_c$ and h_{f_1}. □

Proof of Lemma 6.4. This proof is similar in character to, but much easier than, the proof of (7.3). □

Proof of Lemma 6.5. First use the argument employed to prove (A.1) of Lemma A.1, to show that, for each $0 < a < b < \infty$ and all positive integers l,

$$\sup_{n,\, f \in \mathscr{F}_n,\, a \le t \le b} E_f |n^{1/2} D^{(2)}(n^{-1/5}t, f)|^{2l} \le C(a, b, l). \qquad □$$

REFERENCES

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

Bretagnolle, J. and Huber, C. (1979). Estimation des densities: risque minimax. *Z. Wahrsch. verw. Gebiete* **47** 119–137.

Burkholder, D. L. (1973). Distribution function inequalities for martingales. *Ann. Probab.* **1** 19–42.

Burman, P. (1985). A data dependent approach to density estimation. *Z. Wahrsch. verw. Gebiete* **69** 609–628.

Devroye, L. and Györfi, L. (1984). *Nonparametric Density Estimation: The L_1 View.* Wiley, New York.

Farrell, R. H. (1972). On the best obtainable rates of convergence in estimation of a density function at a point. *Ann. Math. Statist.* **43** 170–180.

Ferguson, T. S. (1985). Personal communication.

Fryer, M. J. (1977). A review of some non-parametric methods of density estimation. *J. Inst. Math. Appl.* **20** 335–354.

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

Hall, P. and Marron, J. S. (1985). Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. To appear in *Probab. Theory Rel. Fields.*

Ibragimov, I. A. and Has'minskii, R. Z. (1981). *Statistical Estimation, Asymptotic Theory.* Springer, New York.

Marron, J. S. (1985). An asymptotically efficient solution to the bandwidth problem of kernel density estimation. *Ann. Statist.* **13** 1011–1023.

Meyer, T. G. (1977a). On fixed or scaled radii confidence sets: the fixed sample size case. *Ann. Statist.* **5** 65–78.

Meyer, T. G. (1977b). Bounds for estimation of density functions and their derivatives. *Ann. Statist.* **5** 136–142.

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

PETROV, V. V. (1978). *Sums of Independent Random Variables*. Springer, New York.

RICE, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12** 1215–1230.

ROSENBLATT, M. (1971). Curve estimates. *Ann. Math. Statist.* **42** 1815–1842.

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

SACKS, J. and STRAWDERMAN, W. (1982). Improvements on linear minimax estimates. In *Statistical Decision Theory and Related Topics III* (S. S. Gupta and J. O. Berger, eds.) **2** 287–304. Academic, New York.

SACKS, J. and YLVISAKER, N. D. (1981). Asymptotically optimal kernels for density estimation at a point. *Ann. Statist.* **9** 334–346.

STONE, C. J. (1980). Optimal convergence rates for nonparametric estimators. *Ann. Statist.* **8** 1348–1360.

STONE, C. J. (1982). Optimal global rates of convergence of nonparametric regression. *Ann. Statist.* **10** 1040–1053.

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

STONE, C. J. (1985). An asymptotically optimal histogram selection rule. In *Proc. of the Berkeley Conference in Honor of Jerzy Neyman and Jack Keifer* (L. M. Le Cam and R. A. Olshen, eds.) **2** 513–520. Wadsworth, Monterey, Calif.

WAHBA, G. (1975). Optimal convergence properties of variable knot, kernel, and orthogonal series methods for density estimation. *Ann. Statist.* **3** 15–29.

WEGMAN, E. J. (1972). Nonparametric probability density estimation: I. A summary of the available methods. *Technometrics* **14** 533–546.

DEPARTMENT OF STATISTICS
FACULTY OF ECONOMICS AND COMMERCE
AUSTRALIAN NATIONAL UNIVERSITY
G.P.O. BOX 4, CANBERRA, A.C.T. 2601
AUSTRALIA

DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA 27514