Probab. Th. Rel. Fields 90, 149–173 (1991)

Probability Theory and Related Fields

Lower bounds for bandwidth selection in density estimation*

Peter Hall ** **and J.S. Marron**
Department of Statistics, Brown University and University of North Carolina, Chapel Hill, NC 27599–3260, USA

Received June 7, 1989; in revised form March 7, 1991

Summary. This paper establishes asymptotic lower bounds which specify, in a variety of contexts, how well (in terms of relative rate of convergence) one may select the bandwidth of a kernel density estimator. These results provide important new insights concerning how the bandwidth selection problem should be considered. In particular it is shown that if the error criterion is Integrated Squared Error (ISE) then, even under very strong assumptions on the underlying density, relative error of the selected bandwidth cannot be reduced below order $n^{-1/10}$ (as the sample size grows). This very large error indicates that any technique which aims specifically to minimize ISE will be subject to serious practical difficulties arising from sampling fluctuations. Cross-validation exhibits this very slow convergence rate, and does suffer from unacceptably large sampling variation. On the other hand, if the error criterion is Mean Integrated Squared Error (MISE) then relative error of bandwidth selection can be reduced to order $n^{-1/2}$, when enough smoothness is assumed. Therefore bandwidth selection techniques which aim to minimize MISE can be much more stable, and less sensitive to small sampling fluctuations, than those which try to minimize ISE. We feel this indicates that performance in minimizing MISE, rather than ISE, should become the benchmark for measuring performance of bandwidth selection methods.

1. Introduction

Nonparametric density estimation provides a very useful tool for investigating the distribution structure of unknown populations. See Silverman (1986) for a large collection of interesting real data examples, where this method provides inference essentially unavailable from other approaches.

A useful mathematical formulation of the density estimation problem is to think of estimating a probability density f using a random sample, $X_1, \ldots, X_n$,

* Research partially supported by National Science Foundation Grants DMS-8701201 and DMS-8902973

** Research of the first author was done while on leave from the Australian National University

from f. Given a bandwidth h, and a kernel function K, the kernel density estimator is

$$\hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i) ,$$

where $K_h(\cdot) = K(\cdot/h)/h$. Kernel estimators have been at the center of density estimation because their simplicity and easily understood intuition makes them very attractive to those whose main concern is seeing distributional structure in data sets.

The bandwidth h of the kernel estimator controls the smoothness of the resulting curve estimate, with the result that bandwidth choice is crucial to performance of the estimator. See for example Devroye and Györfi (1984) or Silverman (1986). The most effective method for bandwidth selection currently in use is subjective choice by the data analyst. While much insight has been gained by this approach, it is clear that a more objective method would greatly enhance the usefulness of density estimation, so considerable research effort has been invested in finding data based methods of choosing the bandwidth h, see the survey by Marron (1988).

The usual approach to formalizing this problem is to first decide on a method of assessing the performance of $\hat{f}_h$. Widely considered means of doing this include the Integrated Squared Error (ISE)

$$\Delta(h,f) = \int (\hat{f}_h - f^2) ,$$

and the Mean Integrated Squared Error (MISE)

$$M(h,f) = E\Delta(h,f) .$$

(See Devroye and Györfi (1984) for another viewpoint.) The minimizers of these criteria, denoted by $\hat{h}_f$ and h_f respectively, are both intuitively reasonable choices of "optimal bandwidth". Which of these has been called optimal has in the past been determined by the researcher, and there appears to have been little concern (with some exceptions as noted below) over differences between these. The main thrust of this paper is development of asymptotic arguments that we feel demonstrate clearly that:

(1) there is an important difference between these two methods of assessing accuracy, and
(2) h_f and $M(h,f)$ should play the more prominent role in future developments, from both analytical and simulation viewpoints.

The basis for this last conclusion is that, under the common assumption that f has two or more derivatives, the goal of $\hat{h}_f$ cannot be attained without serious sampling fluctuations which can cause grave practical problems. In fact, there do not exist bandwidth selection methods which come closer to $\hat{h}_f$ than $n^{-1/10}$ in terms of relative rate of convergence, no matter how many more than two derivatives are assumed. Analogous results hold for attempts to estimate Δ. However, when f has only slightly more than four derivatives, there are bandwidth selection methods which come within $n^{-1/2}$ of h_f. This relatively low error makes such bandwidth selection methods less sensitive to sampling fluctuations than those which aim to

estimate $\hat{h}_f$. Essentially, the sample variability of $\hat{h}_f$ makes this quantity extraordinarily difficult to estimate, even under very strong smoothness assumptions, while the stability of h_f makes it a reasonable goal.

Many of the famous early theoretical papers in density estimation, such as Rosenblatt (1956), Watson and Leadbetter (1963) and Rosenblatt (1972), all chose $M(f, h)$ as their criterion, because it is far simpler to work with analytically. Three early simulation studies, Anderson (1969), Wegman (1972) and Fryer (1977), expressed a desire to use $M(f, h)$ as a criterion but pointed to numerical difficulties and instead approximated $M(f, h)$ by an average (over the simulations) of a quantity closely related to $\Delta(f, h)$. The first researcher to express concern over the difference between M and Δ was Steele (1978). But Steele's arguments were based on a pathological type of estimator, and in fact it is usually the case that *pointwise in h* (or in deterministic sequences of h's), $M(f, h)$ and $\Delta(h, f)$ are asymptotically equivalent in the sense that their ratio converges to 1 as $n \to \infty$. See Hall (1982), Marron (1986); Marron and Härdle (1986).

However when considering data-based bandwidth selection, the above results need cautious interpretation because then the bandwidth is really considered to be a random variable $\hat{h}$. The first indication that random bandwidths make a substantial difference was provided by Hall and Marron (1987a), who showed that the relative difference between h_f and $\hat{h}_f$ is of the extremely large order of magnitude $n^{-1/10}$ as $n \to \infty$.

Earlier workers paid very little attention to this issue. It is interesting to note that in two fundamental papers on modern data-based bandwidth selection, Rudemo (1982) and Bowman (1984) (who independently proposed the influential idea of Least Squares cross-validation), different viewpoints are taken. Rudemo drew his motivation from thinking in terms of estimating $M(f, h)$, while Bowman arrived at essentially the same bandwidth selector by striving to approximate $\Delta(f, h)$. The ensuing development has been divided as well. More complete literature on this available in an earlier version of this paper available from the authors. See Härdle et al. (1988) and Marron (1988, 1989) for further discussion concerning which of h_f and $\hat{h}_f$ should be called the "optimal" bandwidth.

As noted above, the present paper addresses the issue of h_f versus $\hat{h}_f$ by exploring the best possible rates of convergence of any automatically selected bandwidth to these targets. For $\hat{h}_f$ this rate is a very slow $n^{-1/10}$, while for h_f it is the far faster $n^{-1/2}$. We argue from these results that h_f, not $\hat{h}_f$, should be the goal in bandwidth selection problems. The reason is that relative errors of the order $n^{-1/10}$ in estimation of $\hat{h}_f$ indicate extreme susceptibility to sampling fluctuations—note for example the notorious difficulties which have been experienced with cross-validation (Scott and Terrell 1987). On the other hand, the availability of $n^{-1/2}$ convergence (under strong enough assumptions), and much better sample stability, make h_f much more appealing.

We acknowledge that there are some strong philosophical arguments against this viewpoint. For example, Mammen (1988) has pointed out that h_f and $M(f, h)$ have no interpretation in a classical decision-theoretic sense. Note that for bandwidth selection, the usual notion of "loss" would be $\Delta(f, \hat{h})$. It might appear at first glance that the corresponding "risk" would be $M(f, \hat{h})$, but more careful thought shows that it is in fact $E\Delta(f, \hat{h})$, which is different because of the randomness of $\hat{h}$.

The reason that there is room for notions besides the usual loss and risk is due to the fact that there is a richer structure in the present problem than in the classical

decision-theoretic setup. For this reason we must be careful not to dogmatically apply the rules of classical decision theory, and instead consider possible new viewpoints, such as h_f and $M(f, h)$, in their own right.

There is an entirely different angle from which the above theoretical results also provide a great deal of insight. That viewpoint concerns how one should assess the performance of bandwidth selectors in simulation studies. When one considers the problem in this setting, a strong argument can be made for $\hat{h}_f$ and $\Delta(f, h)$ along the lines of "I want to estimate f well for my set of data, not in an average sense over all possible sets of data (as done by consideration of h_f and $M(f, h)$)". While this argument is sensible, our results show that it is not realistic, because $\hat{h}_f$ has too much variability to be accurately estimated. Furthermore, the high level of noise that is now seen to be inherent to $\hat{h}_f$ and $\Delta(f, h)$ render them less effective than h_f and $M(f, h)$, because the noise tends to obscure important differences between various bandwidth selectors. If the Monte Carlo variability of a bandwidth selection method is assessed by confidence intervals, they will be far wider for $\hat{h}_f$ or Δ, because of the excess noise). See Marron (1989) for an actual demonstration of this phenomenon. In our opinion, the greater resolution possible from the use of h_f and $M(f, h)$ outweighs the intuitive disadvantages mentioned above.

To study how close any bandwidth selector $\hat{h}$ (thought of now as being any measurable function of the data, so as to take into account any future methods, as well as those known now) may be to either notion of optimum, it is necessary to consider more than one underlying density. A convenient means of doing this is through a minimax structure. A point of similarity of the results in Sect. 2 is that, for each n, only two alternative densities need be considered. Donoho and Liu (1987) have termed such settings "problems whose difficulty is determined by the hardest one dimensional subproblem."

From this viewpoint, Theorem 2.2 is a major improvement over the results of Hall and Marron (1987b). The improvement is not just technical in character, but is very important from a statistical point of view. This is because the previous result essentially assumed that f "has no more than two derivatives," and left open the problem of what happens when f is smoother. Theorem 2.2 shows conclusively that smoothness is simply an artifact of the method of proof used there, and in fact the $n^{-1/10}$ bound holds even in the presence of parametric knowledge about the underlying density f.

Theorem 2.3 provides a lower bound of $n^{-1/2}$ to the relative error in estimating h_f. It is based on a two point discrimination problem, which is essentially parametric in nature. Theorem 3.2 shows that in a nonparametric setting, when the underlying density is not sufficiently smooth, the bound of $n^{-1/2}$ can be sharpened. In Sect. 4, it is shown that the combined bounds on the rate of convergence of Sects. 2 and 3 are sharp, by exhibiting bandwidth selectors which attain the same rates of convergence, for the various amounts of smoothness assumed on the underlying density. The arguments which give the latter results involve an improvement of a result in Hall and Marron (1987c), given in Theorem 4.1.

A very different application of larger alternative classes will be given in Sect. 3.3, where this technique will be applied to gain some improvements of the results in Sect. 2.

All proofs are postponed to Sect. 5.

Observe that for our main results we assume the kernel K is a probability density, which entails that for estimation of f the rate of convergence can not exceed $n^{-4/5}$, which is achieved with only two bounded derivatives. When more

smoothness is assumed (as done in this paper) faster rates are possible, but these require the use of higher order kernels as defined in Remark 2.2.4. However we feel it is most relevant to study nonnegative kernels (even when their asymptotic rate of convergence is slightly less than optimal) because these are what are used in practice. The reason for this is that higher order kernels lose the most attractive feature of the kernel density estimator: its beautifully simple and compelling intuitive content. Even beginners immediately understand what the nonnegative kernel is doing to the data, but this becomes far harder when the kernel takes on negative values. We feel this consideration outweighs the rather minimal (at least for reasonable sample sizes) gains that we have observed for higher order kernels. Extensions of our results to higher order kernels are briefly discussed in Remarks 2.2.4, 2.3.4, 3.2.5 and 4.5, although there also more smoothness is typically assumed than the kernel can use.

2. Bounds involving two alternatives

2.1. Introduction and summary

To obtain the lower bounds in the current section, it is enough to consider (for each n) only two alternative densities. A means of constructing these (in a way which yields useful lower bounds) is to start with a fixed density $f_0(x)$, and a function $\alpha(x)$, and consider the alternative density

$$f_1(x) \equiv \{1 + n^{-1/2}\alpha(x)\} f_0(x) .$$

The fact that f_0 and f_1 are distant only $n^{-1/2}$ apart means that our bounds will apply even in a parametric setting, not solely to nonparametric classes of densities. It also means that most of the usual norms will be of order $n^{-1/2}$ when α is a reasonable function. This will emerge particularly clearly in Sect. 2.4, where the cases of shifts and rescalings of a fixed f will be dicussed.

To ensure that f_1 is a proper density (for n large enough), assume

(2.1.1) $\int \alpha f_0 = 0$ and $f_1 \geqq 0$.

Also assume

(2.1.2) f_0 and $|\alpha| f_0$ are bounded,

(2.1.3) f_0 and αf_0 have five bounded derivatives,

(2.1.4) $0 < \sigma^2 \equiv \int \alpha^2 f_0 < \infty$.

Convenient technical assumptions concerning the estimator are:

(2.1.5) K is nonnegative and symmetric, with $\int K = 1$,

(2.1.6) K is compactly supported with a Hölder-continuous second derivative.

Assumption (2.1.5) is important to the effective behavior of the kernel estimator. It implies that K is a "second order kernel". Versions of our results for higher order kernels will be presented in Remark 2.2.4. Assumption (2.1.6) is made more for convenience. It is straightforward to weaken this assumption through the use of

various truncation arguments, but this is not done explicitly because the increased complexity of proof would detract from the main points.

Further useful notation is,

$$p \equiv 1 - \Phi(\sigma/2)\,,$$

where Φ denotes the standard normal cumulative distribution function.

Sections 2.2 and 2.3 will provide lower bounds to convergence rates of general estimators of $\hat{h}_f$ and h_f, respectively. Section 2.4 will illustrate the main features of these results by considering density estimation in parametric problems where either scale or location is unknown. In particular, the fact that the bounds obtained in Sects. 2.2 and 2.3 apply even in the presence of parametric knowledge is underscored.

2.2. *Bounds in the case of* ISE

In addition to the technical assumptions made in Sect. 2.1, also assume that the alternative densities, f_0 and f_1, are distinct in the sense that

$$\int \{(\mathrm{d}/\mathrm{d}x)^2 \alpha(x) f_0(x)\} f_0(x)\,\mathrm{d}x \neq 0\,, \tag{2.2.1}$$

where here and below $(\mathrm{d}/\mathrm{d}x)^2$ denotes two applications of the usual derivative operator, applied to the product function $\alpha \cdot f$ in this case. The implications of this condition will be made clear in Sect. 2.4. The following theorem shows that it is impossible to find a data-based bandwidth which is closer to $\hat{h}_f$, the minimizer of the Integrated Squared Error $A(h,f)$, than $n^{-1/10}$ in a relative error sense.

Theorem 2.2. *Under the Assumptions* (2.1.1)–(2.1.6) *and* (2.2.1), *for* $\hat{h}$ *any measurable function of the data,*

$$\lim_{\varepsilon \to 0} \liminf_{n \to \infty} \max_{f \in \{f_0, f_1\}} P_f(|\hat{h} - \hat{h}_f|/\hat{h}_f > \varepsilon n^{-1/10}) \geqq p\,, \tag{2.2.2}$$

$$\lim_{\varepsilon \to 0} \liminf_{n \to \infty} \max_{f \in \{f_0, f_1\}} P_f\{|A(\hat{h},f) - A(\hat{h}_f,f)|/A(\hat{h}_f,f) > \varepsilon n^{-1/5}\} \geqq p\,. \tag{2.2.3}$$

The proof of Theorem 2.2 will be given in Sect. 5.1.

Remark 2.2.1. If $\hat{h}$ is taken to be the bandwidth chosen by cross-validation then the convergence rates in (2.2.2) and (2.2.3) are achievable; see Hall and Marron (1987a). Therefore the convergence rates described by Theorem 2.2 are best possible. Theorem 2.2 is a substantial strengthening of Theorems 2.1 and 4.1 of Hall and Marron (1987b). Although the bound is the same, the class of alternatives is much smaller and simpler here, with resulting benefits as discussed in Sect. 1.

Remark 2.2.2. The probability p may be increased to 1 if more than just the two alternatives f_0 and f_1 are considered. A method of doing this will be described in Sect. 3.3.

Remark 2.2.3. If there were really only two densities f_0 and f_1 under consideration, then "$\geqq p$" would become "$= p$" if one took $\hat{h}$ to be the "likelihood ratio bandwidth", which chooses between $\hat{h}_{f_0}$ and $\hat{h}_{f_1}$ depending on whether the likelihood ratio is bigger or smaller than one. The proof of the theorem is based on the

fact that no discrimination rule can distinguish between f_0 and f_1 more effectively than the likelihood ratio rule.

Remark 2.2.4. If the kernel function K is allowed to take on negative values, then the rate of convergence of $\hat{f}_h$ to f may be improved (see, for example, Sect. 3.6 of Silverman 1986). In particular the kernel function K is said to be of order r when

$$\int x^j K(x)\,dx = \begin{cases} 1 & \text{if } j = 0 \\ 0 & \text{if } 1 \leqq j \leqq r-1 \\ \kappa \neq 0 & \text{if } j = r \end{cases}$$

Assumption (2.1.5) ensures that K is of order 2. The advantage of K being of order r is that, when f is assumed to have r continuous derivatives and $h \sim n^{-1/(2r+1)}$, both $\Delta(h,f)$ and $M(h,f)$ are of size $n^{-2r/(2r+1)}$. Theorem 2.2 continues to hold under this type of assumption, with the rates of $n^{-1/10}$ and $n^{-1/5}$ replaced by $n^{-1/2(2r+1)}$ and $n^{-1/(2r+1)}$, respectively. The differential operator $(d/dx)^2$ in condition (2.2.1) should be replaced by $(d/dx)^r$.

2.3. *Bounds in the case of* MISE

In this case, the Assumption (2.2.1) concerning the difference between the alternative densities, f_0 and f_1, should be replaced by

$$\int \{(d/dx)^4 \alpha(x) f_0(x)\} f_0(x)\,dx \neq 0\,. \tag{2.3.1}$$

See Sect. 2.4 for an investigation of the implications of this condition. Our next result shows that it is impossible to use a data-based bandwidth which is closer to h_f, the minimizer of the Mean Integrated Squared Error $M(h,f)$, than $n^{-1/2}$ in a relative error sense.

Theorem 2.3. *Under the Assumptions* (2.1.1)–(2.1.6) *and* (2.3.1), *for* $\hat{h}$ *any measurable function of the data,*

$$\lim_{\varepsilon \to 0} \liminf_{n \to \infty} \max_{f \in \{f_0, f_1\}} P_f(|\hat{h} - h_f|/h_f > \varepsilon n^{-1/2}) \geqq p\,, \tag{2.3.2}$$

$$\lim_{\varepsilon \to 0} \liminf_{n \to \infty} \max_{f \in \{f_0, f_1\}} P_f\{|M(\hat{h},f) - M(h_f,f)|/M(h_f,f) > \varepsilon n^{-1}\} \geqq p\,. \tag{2.3.3}$$

The proof of Theorem 2.3 will be given in Sect. 5.2.

Remark 2.3.1. Section 4 will discuss bandwidth selectors $\hat{h}$ which achieve the convergence rates described in Theorem 2.3. In fact, the convergence rates are achievable uniformly over classes of densities having 4.25 derivatives. In this sense the convergence rates described by Theorem 2.3 are the best possible.

Remark 2.3.2. As in Sect. 2.2, the probability p may be increased to 1 if more than just the two alternatives f_0 and f_1 are considered. A method of doing this will be given in Sect. 3.3.

Remark 2.3.3. Also as in Sect. 2.2, the likelihood ratio bandwidth (adapted for M instead of Δ) gives equality in (2.3.2) and (2.3.3).

Remark 2.3.4. There is a version of Theorem 2.3, with exactly the same rates $n^{-1/2}$ and n^{-1} in (2.3.2) and (2.3.3) respectively, for higher order kernels. The only change required is suitable modification of (2.3.1).

2.4. Example. Scale and location changes

Additional insight into the structure of the minimax bounds of Theorems 2.2 and 2.3 can be gained by consideration of some specific choices of the alternative densities f_0 and f_1. Particularly interesting features are emphasized if α is chosen to make f_1 approximately a scale or location change of f_0. Of course in such a context, one should never consider estimating a density with a kernel estimator, but it is never the less worth studying because of the interesting implications for the bandwidth selection problem. In the scale-change case, $f_1(x)$ may be represented as

$$(2.4.1) \qquad (1+n^{-1/2})f_0\{(1+n^{-1/2})x\}$$

$$= f_0(x) + n^{-1/2}\{f_0(x) + xf_0'(x)\} + O(n^{-1}) .$$

Thus define $\alpha(x) = 1 + x\{f_0'(x)/f_0(x)\}$. Straightforward calculations show that for any f_0 satisfying (2.1.1)–(2.1.4), conditions (2.2.1) and (2.3.1) hold for this f_1, and so this "scale alternative" may be used in Theorems 2.2 and 2.3. In the terminology of Donoho and Liu, this is a "hard direction" in which to estimate either $\hat{h}_f$ or h_f.

In this context, Theorems 2.2 and 2.3 are perhaps most vividly illustrated by considering the problem of estimating a normal $N(\mu, \sigma^2)$ density using a non-parametric density estimator, as follows. Suppose μ is known, but σ^2 is unknown. Estimate σ^2 using the sample variance $\hat{\sigma}^2$, and take $\tilde{f}$ to be the $N(\mu, \hat{\sigma}^2)$ density. Note that $n^{-1/2}$ is the order of magnitude of the distance between $\hat{\sigma}^2$ and σ^2, so we are essentially in the context of the previous paragraph. Take $\hat{h}_{\tilde{f}}$ (the bandwidth which minimizes $\Delta(h, \tilde{f})$) as our estimate of $\hat{h}_f$ (the bandwidth which minimizes $\Delta(h, f)$). Likewise, take $h_{\tilde{f}}$ as an estimate of h_f. Then $(\hat{h}_{\tilde{f}} - \hat{h}_f)/\hat{h}_f$ is of precise order $n^{-1/10}$, as indicated by Theorem 2.2, and $(h_{\tilde{f}} - h_f)/h_f$ is of precise order $n^{-1/2}$, as indicated by Theorem 2.3. This simple example brings home strikingly the fact that, *even in the presence of parametric knowledge* about f, we cannot hope to estimate $\hat{h}_f$ with a relative error of less than $n^{-1/10}$. The goal of estimating h_f is clearly very different, because in the presence of such parametric knowledge we can achieve the usual parametric rate of $n^{-1/2}$.

However, the situation changes markedly if the unknown parameter is one of location rather than scale. In the location-change case, $f_1(x)$ may be represented as

$$f_0(x + n^{-1/2}) = f_0(x) + n^{-1/2}f_0'(x) + O(n^{-1}) .$$

Hence, define $\alpha(x) = f_0'(x)/f_0(x)$. Again it is simple to check that, for any f_0 allowed by (2.1.1)–(2.1.4), the conditions (2.2.1) and (2.3.1) are not satisfied by this choice of α. Indeed, not only are these assumptions not valid, but the conclusions of Theorems 2.2 and 2.3 fail. i.e. this subproblem is not as hard as that for the scale change, in the terminology of Donoho and Liu (1987).

Again, these features are perhaps best brought out by considering the problem of estimating a normal $N(\mu, \sigma^2)$ density. On this occasion, suppose μ is unknown and σ^2 is known. Estimate μ using the sample mean $\hat{\mu}$, and take $\tilde{f}$ to be the $N(\hat{\mu}, \sigma^2)$ density. Then $n^{-1/2}$ is the order of magnitude of the distance between $\hat{\mu}$ and μ, and

so we are in the context of the previous paragraph. Let $\hat{h}_{\tilde{f}}, h_{\tilde{f}}$ be our parametric estimates of $\hat{h}_f$, h_f respectively. Then $h_{\tilde{f}} = h_f$, so that our estimate of h_f is error-free. However, it may be shown that $(\hat{h}_{\tilde{f}} - \hat{h}_f)/\hat{h}_f$ is of precise order $n^{-3/5}$, which is considerably better than the error of the order $n^{-1/10}$ encountered in the scale-change problem, and even better than the error $n^{-1/2}$ which might have been expected, but not quite error-free. It turns out that a relative error of $n^{-3/5}$ is intrinsic to bandwidth selection for the ISE problem in this setting. In an earlier version of this paper, available from the authors, this fact is formulated as a theorem along the lines of Theorem 2.2.

3. Bounds involving multiple alternatives

3.1. Introduction and summary

There are two points at which deeper insight can be gained by modifying the above two-alternative minimax results to make use of multiple alternatives. The first point is in establishing a better bound on how well a bandwidth selector $\hat{h}$ may approximate h_f, the minimizer of the Mean Integrated Squared Error $M(h,f)$, in situations where the underlying density is not too smooth. The second point is in strengthening Theorems 2.2 and 2.3 by replacing p by 1 on the right hand sides of (2.2.2), (2.2.3), (2.3.2), and (2.3.3).

When the underlying density is not too smooth, the lower bounds of Theorem 2.3 may be sharpened. In such cases, the rates of convergence depend on the amount of smoothness assumed about the underlying density. To quantify this in a form convenient for minimax lower bound results, consider smoothness classes indexed by a parameter $v \geqq 0$. In particular, given $B > 0$ let l be the largest integer strictly less than $2 + v$, and define $G_v(B)$ to be the set of all probability densities which vanish outside $(-B, B)$, have l derivatives, and satisfy

$$\sup_{x,y} |f^{(l)}(x) - f^{(l)}(y)|/|x - y|^{2+v-l} \leqq B. \tag{3.1.1}$$

A minimax lower bound for the relative rate of convergence of $\hat{h}$ to h_f, in terms of the smoothness index v, will be stated in Sect. 3.2. The issue of increasing p to 1 will be treated in Sect. 3.3.

3.2. Bounds in the case of MISE

The minimax lower bound of Theorem 2.3 may be sharpened, when the underlying density is not too smooth, to:

Theorem 3.2. *Under the Assumptions* (2.1.5) *and* (2.1.6), *for* $v, B \geqq 0$ *and* $\hat{h}$ *any measurable function of the data,*

$$\lim_{\varepsilon \to 0} \liminf_{n \to \infty} \sup_{f \in G_v(B)} P_f(|\hat{h} - h_f|/h_f > \varepsilon n^{-\rho}) = 1\,, \tag{3.2.1}$$

$$\lim_{\varepsilon \to 0} \liminf_{n \to \infty} \sup_{f \in G_v(B)} P_f\{|M(\hat{h},f) - M(h_f,f)|/M(h_f,f) > \varepsilon n^{-2\rho}\} = 1\,, \tag{3.2.2}$$

where

$$\rho = \max\{1/10, 4\nu/(4\nu + 9)\} . \tag{3.2.3}$$

The proof of Theorem 3.2 will be given in Sect. 5.2.

Remark 3.2.1. It is important to note that Theorems 2.3 and 3.2 each provide useful information for different values of ν, with $\nu = 2.25$ being the boundary point. In particular, for $\nu > 2.25$ we have $\rho > 1/2$, and then the lower bound of Theorem 2.3 is more informative. On the other hand, for $\nu < 2.25$ we have $\rho < 1/2$, so the present bound is more useful. Thus the overall lower bound is $n^{-\rho'}$, where

$$\rho' = \min(\rho, 1/2) = \min[\max\{1/10, 4\nu/(4\nu + 9)\}, 1/2] .$$

Remark 3.2.2. Observe that $\rho = 1/10$ for $0 \leqq \nu \leqq 0.25$, and $\rho = 4\nu/(4\nu + 9)$ for $0.25 \leqq \nu \leqq 2.25$. The relative error convergence rate of $n^{-\rho}$ for $\hat{h}$ is best possible in both these cases, in the sense that there exist bandwidth selectors which achieve the bound of Theorem 3.2. The convergence rate of $n^{-1/2}$ given by Theorem 2.3 is also optimal, when $\nu \geqq 2.25$. If $0 \leqq \nu \leqq 0.25$, the optimal rate of $n^{-1/10}$ is achieved by cross-validation, see Hall and Marron (1987a). For $\nu \geqq 0.25$ the optimal rate is achieved by a plug-in estimator, as will be shown in Sect. 4.

Remark 3.2.3. The optimal convergence rate for the closely related problem of estimating $\int (f'')^2$, is $n^{-\rho''}$ where

$$\rho'' = \min\{4\nu/(4\nu + 9), 1/2\} ,$$

see Bickel and Ritov (1988) (the relationship between these two problems will be discussed in some detail in Sect. 4). Since $\rho' = \rho''$ for $\nu \geqq 0.25$, but $\rho' > \rho''$ for $0 \leqq \nu < 0.25$, it follows that the problem of estimating $\int (f'')^2$ is of equivalent difficulty to that of estimating h_f when $\nu \geqq 0.25$, but is harder otherwise.

Remark 3.2.4. The class of densities $G_\nu(B)$ is actually far bigger than is required to obtain the bound stated in the theorem. In particular, in the proof a much smaller class (finite for each n) of alternatives is constructed, and this is all that is necessary. The more general result is not stated here because it involves the introduction of considerably more notation, which has a tendency to obscure the main point of this section.

Remark 3.2.5. If the kernel K is of order r, then $G_\nu(B)$ should denote a class of densities with $r + \nu$ "derivatives", instead of $2 + \nu$ as above. Then the only change to Theorem 3.2 is that ρ should be changed to

$$\rho = \max[1/\{2(2r + 1)\}, 4\nu/(4\nu + 4r + 1)] .$$

If this ρ is used in the formula $\rho' = \min(\rho, 1/2)$, then optimal convergence rate (both bound and achievable, is still $n^{-\rho'}$); compare Remark 3.2.2.

Remark 3.2.6. A referee has pointed out an interesting heuristic, to the effect that if f is smooth everywhere, except for a single jump discontinuity, the cross-validated bandwidth will have a faster relative rate of convergence than $n^{-1/10}$. The generalization of this idea to the case of f everywhere smooth, except for finitely many jumps or kinks (where f is continuous, but f' has a jump), has been independently established by van Es (1991). This might be considered surprising, in

view of our results, since it seems that with "less smoothness" one can obtain a faster rate of convergence. The problem lies with the fact that such underlying densities are not really "less smooth", in the sense of being "representative members" of $G_\nu(B)$, for $\nu < 0.25$. In particular when there are only finitely many jumps and kinks, the locations and sizes of these can be estimated very well, so the only unknown part of f is the smooth part, which can be estimated with a faster rate. We feel it is to the credit of cross-validation that it "automatically adapts" to such functions, with no apriori knowledge of the jumps or kinks required.

3.3. Probability one bounds

In Theorems 2.2 and 2.3 the probabilities p may all be sharpened to 1 if a larger class of alternatives is used. A simple way of constructing such a larger class is to consider all convex combinations of the f_0 and f_1 described above. In particular, define

$$C(f_0, f_1) \equiv \{\omega f_0 + (1-\omega) f_1 : \omega \in [0, 1]\} .$$

Then, if the set of alternatives $\{f_0, f_1\}$ is replaced by $C(f_0, f_1)$, the values of p in Theorems 2.2 and 2.3 may all be taken to be 1.

This is intuitively clear, because the minimax bounds calculated in these theorems come from the difficulty in using $X_1, \ldots, X_n$ to choose among the various possible density functions. If the class $\{f_0, f_1\}$ is enlarged by including some convex combinations of f_0 and f_1 (say for ω in some equally spaced grid in $[0, 1]$), then the probability p of misclassifying the underlying density gets larger. The limit of this process is the class $C(f_0, f_1)$, and $p = 1$. We do not include a specific proof of this fact, because the idea is the same as that used to verify (1.2) in Stone (1980).

4. Achievability

The plug-in method is a simple way of demonstrating the existence of bandwidths which achieve fast rates of convergence. This makes use of the fact that

$$h_f \sim h_{f,A} \equiv C_K \{\int (f'')^2\}^{-1/5} n^{-1/5} ,$$

where C_K is a constant depending only on the kernel K. The idea is to plug in an estimate of $\int (f'')^2$. Hall and Marron (1987c) discuss the slightly more general problem of estimating $\theta_m = \int (f^{(m)})^2$. One of the estimators considered there is

$$\hat{\hat{\theta}}_m \equiv (-1)^m n^{-1} (n-1)^{-1} \sum\sum_{i \neq j} K_h^{(2m)}(X_i - X_j) .$$

Bickel and Ritov (1988) have shown that rates of convergence calculated in Hall and Marron (1987c) are not the best possible in all cases, by presenting a more complicated estimator which attains a faster rate of convergence. This motivated us to look more carefully at our results, and we found that in certain instances the upper bound given in Hall and Marron (1987c) can be improved upon. In particular, part (d) of Lemma 3.1 in that paper can be sharpened to:

Theorem 4.1. *Under the assumption that $f \in G_\nu(B)$ for some $B > 0$, and that K is of even order k,*

$$E(\hat{\hat{\theta}}_m) - \theta_m = \begin{cases} O(h^{2(2+\nu-m)}) & \text{if } k \geqq 2(2+\nu-m) \\ C_k h^k + o(h^k) & \text{if } k < 2(2+\nu-m) \end{cases},$$

where if $k < 2(2+\nu-m)$,

$$C_k = (-1)^{k/2} \{ \int (f^{(m+k/2)})^2 \} (\int u^k K)/k! \,.$$

Recall the definition of a kernel of order k from Remark 2.2.4.

Remark 4.1. To appreciate the implications of Theorem 4.1, let us assume that $k > 2(2+\nu-m)$. Then

$$E(\hat{\hat{\theta}}_m) - \theta_m = O(h^{2(2+\nu-m)}) \,,$$

and by Lemma 3.2 of Hall and Marron (1987c),

$$\operatorname{var}(\hat{\hat{\theta}}_m)^2 = D_1 n^{-2} h^{-(4m+1)} + D_2 n^{-1} + o(n^{-2} h^{-(4m+1)} + n^{-1})$$

for positive constants D_1 and $D_2 = \int (f^{(2m)})^2 f - \int (f^{(m)})^2$ (note there is a typographical error where this is stated in Hall and Marron (1987c)). It follows that

$$E(\hat{\hat{\theta}}_m - \theta_m)^2 = D_1 n^{-2} h^{-(4m+1)} + D_2 n^{-1} + O(h^{4(2+\nu-m)}) + o(n^{-2} h^{-(4m+1)} + n^{-1}) \,.$$

Hence we get the following.

Corollary 4.1. *If $\nu > 2(m-1) + (1/4)$, and $h = h(n)$ is chosen so that $h = o(n^{-1/4(2+\nu-m)})$ but $n^{-1/(4m+1)} = o(h)$, then $\hat{\hat{\theta}}_m$ is $n^{1/2}$ consistent for θ_m and*

$$E(\hat{\hat{\theta}}_m - \theta_m)^2 = D_2 n^{-1} + o(n^{-1}) \,.$$

Remark 4.2. If we take $m = 2$ in Remark 4.1 we see that, provided $\nu > 2.25$ (i.e. f has more than 4.25 derivatives), we may construct a $n^{1/2}$ consistent estimator $\hat{\hat{\theta}}_2$ of θ_2. For example if we assume f has 5 derivatives then this may be done with K a sixth order kernel and $h = h(n)$ chosen to be of order between $n^{-1/9}$ and $n^{-1/12}$. Remark 4.1 also shows that if $\nu > 2.25$ then it is possible to construct an estimator $\hat{\hat{\theta}}_3$ of θ_3 having the property $\hat{\hat{\theta}}_3 - \theta_3 = o_p(n^{-1/10})$. We use these versions of $\hat{\hat{\theta}}_2$ and $\hat{\hat{\theta}}_3$ below. Define $k_0 = \int K^2$, $k_2 = \int z^2 K(z)\,dz$, $k_4 = \int z^4 K(z)\,dz$, $A_1 = k_0/(k_2^2 \theta_2^2)$, $A_2 = k_4 \theta_3/(20 k_2 \theta_2)$ and

$$h_f' = n^{-1/5} A_1^{1/5} + n^{-3/5} A_1^{3/5} A_2 \,.$$

Then it may be shown by Taylor expansion of $M(h,f)$ about $M(h_f,f)$ that

$$(h_f' - h_f)/h_f = o(n^{-1/2}) \,.$$

From this it follows that

$$(\hat{h} - h_f)/h_f = O_p(n^{-1/2}) \,. \tag{4.1}$$

where

$$\hat{h} = n^{-1/5}\hat{A}_1^{1/5} + n^{-3/5}\hat{A}_1^{3/5}\hat{A}_2 ,$$

$$\hat{A}_1 = k_0/(k_2^2\hat{\hat{\theta}}_2), \quad \hat{A}_2 = k_4\hat{\hat{\theta}}_3/(20k_2\hat{\hat{\theta}}_2) .$$

The condition $v > 2.25$ is sufficient to grant the existence of $\hat{\hat{\theta}}_2$ and $\hat{\hat{\theta}}_3$ such that (4.1) holds. For a version of the plug-in bandwidth which takes constant coefficients, as well as rates of convergence into account, see Park and Marron (1990).

Remark 4.3. For $0.25 \leqq v \leqq 2.25$, straightforward modification of the plug-in bandwidth selector and the above calculations, shows that

$$(\hat{h} - h_f)/h_f = O_p(n^{-4v/(4v+9)}) .$$

Hence in this case also the bound obtained in Theorem 3.2 is sharp, in the sense that there are bandwidth selectors which achieve the bound on the relative rate of convergence.

Remark 4.4. Note that when $v < 0.25$, the performance of the plug-in bandwidth is actually worse than least squares cross-validation (whose rate of convergence is $n^{-1/10}$). From this perspective, it is a strength of least squares cross-validation that it maintains the rate $n^{-1/10}$ for v all the way down to 0.

Remark 4.5. In the case of an r'th order kernel, (4.1) is the same. The only change is that the assumption $v \geqq 2.25$ is replaced by $v \geqq (4r + 1)/4$.

Remark 4.6. It is straightforward to extend Theorem 4.1 to the case of noncompactly supported f satisfying a tail condition, by a truncation argument. This is not done explicitly here because the statistical insight gained does not seem worth the added length of proof.

Remark 4.7. A referee has pointed out the h_f' representation in Remark 4.2 can be extended to given an explanation of why the rate summarized in Remark 3.2.1 has an abrupt change at $v = 0.25$. In particular, for general v,

$$h_f' = n^{-1/5}C_K\theta_2^{-2/5}(1 + O(n^{-2\beta/5})) ,$$

where $\beta = \min(1, v)$. When $v \leqq 1$, θ_2 can be estimated at the rate $O(n^{-4v/(4v+9)})$. Note that the O term is negligible, i.e. $n^{-2v/5} = o(n^{-4v/(4v+9)})$, exactly when $v > 0.25$. The argument can be easily extended to $v \geqq 1$, using the additional θ_3 term in Remark 4.2.

5. Proofs

5.1. *Proof of Theorem 2.2*

We prove only (2.2.2), since the extension to (2.2.3) may be accomplished as in Hall and Marron (1987b, p. 171). Recall that $\hat{h}_f$ is the minimizer of $\Delta(h, f)$ and $\hat{h}$ is a generic data driven bandwidth. Let $\tilde{h} = \hat{h}_{\tilde{f}}$ be an element of $\{\hat{h}_{f_0}, \hat{h}_{f_1}\}$ which minimizes $|\hat{h} - \tilde{h}|$ over those elements. If f is either f_0 or f_1 then

$$|\tilde{h} - \hat{h}_f| \leqq |\tilde{h} - \hat{h}| + |\hat{h} - \hat{h}_f| \leqq 2|\hat{h} - \hat{h}_f| .$$

Therefore result (2.2.2) will follow if we prove that

$$\lim_{\varepsilon \to 0} \liminf_{n \to \infty} \max_{f \in \{f_0, f_1\}} P_f(|\tilde{h} - \hat{h}_f| > \varepsilon n^{-3/10}) \geqq p \,. \tag{5.1.1}$$

Define $L(z) \equiv -zK'(z)$ and

$$\hat{g}_h(x) \equiv (nh)^{-1} \sum_i L\{(x - X_i)/h\} \,.$$

Arguing as in Hall and Marron (1987b, p. 169) we may deduce that

$$\hat{h}_f - \tilde{h} = 2\xi(\tilde{h}, f)/\tilde{h}\Delta^{(2)}(h^*, f) \,, \tag{5.1.2}$$

where $\xi(h, f) \equiv \int (\hat{f}_h - \hat{g}_h)(\tilde{f} - f)$, where $\Delta^{(2)}(h, f)$ denotes the second derivative of $\Delta(h, f)$ with respect to h, and where h^* lies between $\tilde{h}$ and $\hat{h}_f$. It is relatively easy to prove, as in Lemmas 5.2, 6.1 and 6.2 of Hall and Marron (1987b), that

$$\lim_{a \to 0, b \to \infty} \liminf_{n \to \infty} \min_{f \in \{f_0, f_1\}} P_f(an^{-1/5} \leqq \hat{h}_{f_0}, \hat{h}_{f_1} \leqq bn^{-1/5}) = 1 \,,$$

and for all $0 < a < b < \infty$,

$$\lim_{\lambda \to \infty} \limsup_{n \to \infty} \max_{f \in \{f_0, f_1\}} P_f \left\{ \min_{h \in (an^{-1/5}, bn^{-1/5})} |\Delta^{(2)}(h, f)| > \lambda n^{-2/5} \right\} = 0 \,.$$

Therefore result (5.1.1) will follow if we show that

$$\lim_{\varepsilon \to 0} \liminf_{n \to \infty} \max_{f \in \{f_0, f_1\}} P_f \left\{ \min_{h \in (an^{-1/5}, bn^{-1/5})} |\xi(h, f)| > \varepsilon n^{-9/10} \right\} \geqq p \,. \tag{5.1.3}$$

If $\tilde{f} \neq f$ then

$$|\xi(h, f)| = |\int (\hat{f}_h - \hat{g}_h)(\tilde{f} - f) = n^{-1/2} |\int (\hat{f}_h - \hat{g}_h) \alpha f_0| \,.$$

And by the Neyman–Pearson lemma,

$$\begin{aligned} \max_{f \in \{f_0, f_1\}} P_f(\tilde{f} \neq f) &\geqq (1/2)\{P_{f_0}(\tilde{f} = f_1) + P_{f_1}(\tilde{f} = f_0)\} \\ &\geqq (1/2)\{P_{f_0}(\bar{f} = f_1) + P_{f_1}(\bar{f} = f_0)\} \,, \end{aligned}$$

where $\bar{f}$ is the likelihood ratio rule for deciding between f_0 and f_1. Now,

$$\begin{aligned} P_{f_0}(\bar{f} = f_1) &= P_{f_0}\left[\sum_i \log\{1 + n^{-1/2} \alpha(X_i)\} > 0 \right] \\ &= P_{f_0}\left\{ n^{-1/2} \sum_i \alpha(X_i) - \frac{1}{2} n^{-1} \sum_i \alpha(X_i)^2 + o_p(1) > 0 \right\} \\ &\to 1 - \phi(\sigma/2) = p \,, \end{aligned}$$

and similarly $P_{f_1}(\bar{f} = f_0) \to p$. Therefore

$$\liminf_{n \to \infty} \max_{f \in \{f_0, f_1\}} P_f(\tilde{f} \neq f) \geqq p \,, \tag{5.1.4}$$

and so (5.1.3) will follow if we prove that

$$\lim_{\varepsilon\to 0}\liminf_{n\to\infty}\max_{f\in\{f_0,f_1\}}P_f\left\{\min_{h\in(an^{-1/5},bn^{-1/5})}|\int(\hat{f}_h-\hat{g}_h)\alpha f_0|>\varepsilon n^{-2/5}\right\}=1\,. \tag{5.1.5}$$

Put $A\equiv K-L$. Then

$$\begin{aligned} S=S(h)&\equiv\int(\hat{f}_h-\hat{g}_h)\alpha f_0 \\ &=(nh)^{-1}\sum_i\int A\{(x-X_i)/h\}\,\alpha(x)f_0(x)\,\mathrm{d}x \\ &=n^{-1}\sum_i\int A(y)\alpha(X_i+hy)f_0(X_i+hy)\,\mathrm{d}y\,. \end{aligned} \tag{5.1.6}$$

Now,

$$\begin{aligned} E_f(S)&=\int A(y)\,\mathrm{d}y\int\alpha(x+hy)f_0(x+hy)f(x)\,\mathrm{d}x \\ &=h^2\{\int y^2A(y)\,\mathrm{d}y/2\}[\int\{(\mathrm{d}/\mathrm{d}x)^2\alpha(x)f_0(x)\}f_0(x)\,\mathrm{d}x] \\ &\quad+O(h^3+h^2n^{-1/2}) \\ &=h^2c+O(h^3+h^2n^{-1/2})\,, \end{aligned} \tag{5.1.7}$$

say, where $c\neq 0$. (Here we have used (2.2.1) and the fact that $\int y^jA(y)\,\mathrm{d}y=0$ for $j=0,1$.) By Rosenthal's inequality for sums of independent random variables (Burkholder 1973, p. 40),

$$\max_{h\in(an^{-1/5},bn^{-1/5})}\max_{f\in\{f_0,f_1\}}E_f\{|S(h)-ES(h)|^{2r}\}\leqq C(a,b,r)n^{-r}$$

for all $r\geqq 1$. Therefore if H_n is any set of elements of $(an^{-1/5},bn^{-1/5})$ containing no more than n^d elements for any fixed $d>0$, we have for large n,

$$\begin{aligned} &\min_{f\in\{f_0,f_1\}}P_f\left\{\min_{h\in H_n}|S(h)|>(1/3)a^2|c|n^{-2/5}\right\} \\ &\quad\geqq 1-\sum_{h\in H_n}\max_{f\in\{f_0,f_1\}}P_f\{|S(h)-ES(h)|>(1/3)a^2|c|n^{-2/5}\} \\ &\quad\geqq 1-O\{n^d(n^{2/5}n^{-1/2})^{2r}\}\to 1 \end{aligned}$$

as $n\to\infty$, provided we choose $r>5d$. Result (5.1.5) now follows via the continuity argument of Hall and Marron (1987b, p. 175). This completes the proof of Theorem 2.2.

5.2. Proof of Theorem 3.2

The proof of Theorem 3.2 appears before that of Theorem 2.3 because the proof of Theorem 2.3 is greatly reduced in length by using related parts of Theorem 3.2. As for Theorem 2.2, we prove only (3.2.1). The circumstance $0\leqq\nu\leqq 0.25$ follows from $\nu\geqq 0.25$ (because of the nesting of the $G_\nu(B)$ as discussed in Remark 3.2.6) so we assume $\nu\geqq 0.25$. We may further suppose that $\nu\leqq 2.25$, for otherwise $\rho>1/2$ and then Theorem 3.2 follows from Theorem 2.3.

The first step is to construct a class of densities which are "hard to distinguish", yet at the same time are "far apart". Following ideas of Stone (1982) and Bickel and

Ritov (1988), let ψ_0 be a symmetric, six times differentiable function on $(-\infty, \infty)$, vanishing outside $(-1/4, 1/4)$, and having $\sup|\psi_0^{(j)}| \leqq B$ for $0 \leqq j \leqq 6$. Put $m \equiv \eta n^{2/(4\nu+9)}$ where $\eta > 0$, let $\psi = \delta\psi_0$ where $0 < \delta \leqq 1$, let g_0 be a density which is constant at a nonzero value on $(-1/2, 3/2)$ and vanishes outside $(-1, 2)$, let $\tau = (\tau_1, \ldots, \tau_m)$ be a vector, and define

$$\gamma_v \equiv m^{-(2+\nu)}\psi(m(x - v/m)) ,$$

$$f(x) = f(x|\tau) \equiv g_0(x) + \sum_{v=1}^{m} \tau_v \gamma_v(x) , \tag{5.2.1}$$

and

$$\mathscr{F} \equiv \{ f(x|\tau) : \tau \text{ is a sequence of 0's, 1's and } -1\text{'s} \} .$$

Note that for large n, $\mathscr{F}$ is a set of densities vanishing outside $(-1, 2)$ and essentially having uniformly continuous, bounded $(2 + \nu)$'th derivatives. Furthermore, $\mathscr{F} \subset G_\nu(B)$. Note that there are many related constructions possible here. This one has been selected because it seems to require minimal overhead in terms of notation and length of proof.

Let $\tilde{h} = h_{\tilde{f}}$ minimize $|\hat{h} - h_{\tilde{f}}|$ over all $\tilde{f} \in \mathscr{F}$. Then $|\tilde{h} - h_f| \leqq 2|\hat{h} - h_f|$ for all $f \in \mathscr{F}$, and so it suffices to prove that

$$\lim_{\varepsilon \to 0} \lim_{\delta \to 0} \liminf_{n \to \infty} \max_{f \in \mathscr{F}} P_f(|\tilde{h} - h_f| > \varepsilon n^{-(1/5) - 4\nu/(4\nu+9)}) = 1 . \tag{5.2.2}$$

The first step in establishing (5.2.2) is to develop an analogue of (5.1.2).

Let f, g be densities, and observe that

$$\begin{aligned} M(h, g) &= \int E_g(\hat{f}_h - g)^2 \\ &= \int E_g(\hat{f}_h - f)^2 + 2\int (E_g\hat{f}_h - f)(f - g) + \int (f - g)^2 \\ &= M(h, f) + \int (E_g - E_f)(\hat{f}_h - f)^2 + 2\int (E_g\hat{f}_h - f)(f - g) + \int (f - g)^2 . \end{aligned}$$

Differentiating with respect to h we obtain

$$\begin{aligned} M^{(1)}(h, g) = M^{(1)}(h, f) &+ 2h^{-1}\int (E_f - E_g)(\hat{f}_h - \hat{g}_h)(\hat{f}_h - f) \\ &+ 2h^{-1}\int E_g(\hat{f}_h - \hat{g}_h)(g - f) , \end{aligned}$$

where $M^{(j)}(h, g)$ denotes the jth derivative of $M(h, g)$ with respect to h and where $\hat{g}_h$ is as defined in Sect. 5.1. Therefore with

$$\eta(h, f, g) \equiv \int \{(E_f - E_g)(\hat{f}_h - \hat{g}_h)(\hat{f}_h - f) + E_g(\hat{f}_h - \hat{g}_h)(g - f)\}$$

we have

$$M^{(1)}(h, g) = M^{(1)}(h, f) + 2h^{-1}\eta(h, f, g) .$$

Taking $(h, f, g) = (h_{f_1}, f, f_1)$ for $f_1 \in \mathscr{F}$, we find that

$$\begin{aligned} 0 = M^{(1)}(h_{f_1}, f_1) &= M^{(1)}(h_{f_1}, f) + 2h_{f_1}^{-1}\eta(h_{f_1}, f, f_1) \\ &= M^{(1)}(h_f, f) = M^{(1)}(h_{f_1}, f) + (h_f - h_{f_1})M^{(2)}(h^\dagger, f) , \end{aligned}$$

where $h^\dagger$ lies between h_f and h_{f_1}. In consequence,

$$h_f - h_{f_1} = 2\eta(h_{f_1}, f, f_1)/h_{f_1} M^{(2)}(h^\dagger, f) ,$$

whence

$$h_f - \tilde{h} = 2\eta(\tilde{h}, f, \tilde{f})/\tilde{h} M^{(2)}(\tilde{h}^\dagger, f) , \tag{5.2.3}$$

where $\tilde{h}^\dagger$ lies between $\tilde{h}$ and h_f. This is the desired analogue of (5.1.2).

Arguing as in the proofs of Lemmas 5.2, 6.1 and 6.2 of Hall and Marron (1987b) we may show that given $\xi > 0$ we may choose δ, in the definition of ψ, so small and $n_0 \geqq 1$ so large that if $a > 0$ is the constant such that $h_{g_0} \sim n^{-1/5} a$ then

$$a - \xi < \inf_{n \geqq n_0,\, f \in \mathscr{F}} n^{1/5} h_f \leqq \sup_{n \geqq n_0,\, f \in \mathscr{F}} n^{1/5} h_f < a + \xi ,$$

and for any $0 < b < c < \infty$ and some $\lambda = \lambda(b, c) > 0$,

$$\sup_{h \in (bn^{-1/5}, cn^{-1/5}),\, f \in \mathscr{F}} |M^{(2)}(h, f)| \leqq \lambda n^{-2/5} .$$

In view of these results and (5.2.3) we see that (5.2.2) will follow if we prove that for each sufficiently small $\xi, \eta > 0$ (not depending on δ), and any nonparametric rule $\tilde{f}$ for selecting an element of $\mathscr{F}$,

$$\lim_{\varepsilon \to 0} \liminf_{\delta \to 0} \liminf_{n \to \infty} \max_{f \in \mathscr{F}} P_f \left\{ \min_{h \in H_n} |\eta(h, f, \tilde{f})| > \varepsilon n^{-(4/5) - 4\nu/(4\nu+9)} \right\} = 1 , \tag{5.2.4}$$

where $H_n \equiv ((a - \xi)n^{-1/5}, (a + \xi)n^{-1/5})$.

The next step is to simplify $\eta(h, f, f_1)$. Write

$$f = g_0 + \sum_\nu \tau_\nu \gamma_\nu \quad \text{and} \quad f_1 = g_0 + \sum_\nu \tau_{1\nu} \gamma_\nu ,$$

and let $A = K - L$. Define

$$J_1 \equiv \int \psi(y) \left[\iint A(w) K(x) \psi\{y + hm(w + x)\} \, dw \, dx - \int A(w) \psi(y + hmw) \, dw \right] dy ,$$

$$J_2 \equiv \iiint \psi(y) A(w) K(x) \psi\{y + hm(w + x)\} \, dw \, dx \, dy ,$$

not depending on ν. Then:

Lemma 5.2.1.

$$\eta(h, f, f_1) = m^{-(2\nu+5)} J_1 \sum_\nu (|\tau_\nu| - |\tau_{1\nu}|) - m^{-(2\nu+5)} n^{-1} J_2 \sum_\nu (|\tau_\nu| - \tau_\nu \tau_{1\nu}) .$$

Proof of Lemma 5.2.1. Observe that

$$\begin{aligned} E_g\{\hat{f}_h(\hat{f}_h - \hat{g}_h)\} &= (nh^2)^{-1} E_g[K\{(x - X)/h\} A\{(x - X)/h\}] \\ &\quad + (1 - n^{-1})(E_g \hat{f}_h) E_g(\hat{f}_h - \hat{g}_h) . \end{aligned}$$

Therefore

$$\begin{aligned} \eta(h, f, f_1) &= \int \{(E_f - E_{f_1})(\hat{f}_h - \hat{g}_h)(\hat{f}_h - f) + E_{f_1}(\hat{f}_h - \hat{g}_h)(f_1 - f)\} \\ &= \int [(nh^2)^{-1} (E_f - E_{f_1}) K\{(x - X)/h\} A\{(x - X)/h\} \\ &\quad + (1 - n^{-1})\{(E_f \hat{f}_h) E_f(\hat{f}_h - \hat{g}_h) - (E_{f_1} \hat{f}_h) E_{f_1}(\hat{f}_h - \hat{g}_h)\} \\ &\quad - (E_f - E_{f_1})(\hat{f}_h - \hat{g}_h) f + (f_1 - f) E_{f_1}(\hat{f}_h - \hat{g}_h)] \end{aligned} \tag{5.2.5}$$

$$= \int [(1-n^{-1})\{(E_f - E_{f_1})\hat{f}_h E_f(\hat{f}_h - \hat{g}_h)$$
$$+ (E_f - E_{f_1})(\hat{f}_h - \hat{g}_h) E_{f_1}(\hat{f}_h)\} - (E_f - E_{f_1})(\hat{f}_h - \hat{g}_h) f$$
$$+ (f_1 - f) E_{f_1}(\hat{f}_h - \hat{g}_h)]$$
$$= \sum_v (\tau_v - \tau_{1v}) m^{-(v+3)} I_v ,$$

where

$$(5.2.6) \quad I_v \equiv m^{v+3} \int_{C_v} \gamma_v(y) (\int [(1-n^{-1})\{K((x-y)/h) E_f(\hat{f}_h - \hat{g}_h)(x)$$
$$+ A((x-y)/h) E_{f_1}(\hat{f}_h)(x)\} - h^{-1} A\{(x-y)/h\} f(x)]\,dx$$
$$- E_{f_1}(\hat{f}_h - \hat{g}_h)(y))\,dy$$
$$= \int \psi(y-v) (\int [(1-n^{-1})\{K(x) E_f(\hat{f}_h - \hat{g}_h)(m^{-1}y + hx)$$
$$+ A(x) E_{f_1}(\hat{f}_h)(m^{-1}y + hx)\} - A(x) f(m^{-1}y + hx)]\,dx$$
$$- E_{f_1}(\hat{f}_h - \hat{g}_h)(m^{-1}y))\,dy .$$

If $y \in v + (-1/4, 1/4)$, if K vanishes outside $(-1/4, 1/4)$, and if x is in the support of K, then for n so large that $hm < 1/4$,

$$E_f(\hat{f}_h - \hat{g}_h)(m^{-1}y + hx) = h^{-1} \int A\{(m^{-1}y + hx - w)/h\} f(w)\,dw$$
$$= h^{-1} \sum_u \tau_u \int_{C_u} A\{(m^{-1}y + hx - w)/h\} \gamma_u(w)\,dw$$
$$= h^{-1} \tau_v \int_{C_v} A\{(m^{-1}y + hx - w)/h\} \gamma_v(w)\,dw$$
$$= m^{-(v+2)} \tau_v \int A(w) \psi\{y - v + hm(x-w)\}\,dw .$$

Similarly,

$$E_{f_1}(\hat{f}_h(m^{-1}y + hx)) = m^{-(v+2)} \tau_{1v} \int K(w) \psi\{y - v + hm(x-w)\}\,dw,$$
$$f(m^{-1}y + hx) = m^{-(v+2)} \tau_v \psi(y - v + hmx) ,$$
$$E_{f_1}(\hat{f}_h - \hat{g}_h)(m^{-1}y) = m^{-(v+2)} \tau_{1v} \int A(w) \psi(y - v - hmw)\,dw .$$

Substituting into (5.2.6) we obtain

$$I_v = m^{-(v+2)} \int \psi(y) ([\int (1-n^{-1})\{\tau_v K(x) \int A(w) \psi(y + hm(x-w))\,dw$$
$$+ \tau_{1v} A(x) \int K(w) \psi(y + hm(x-w))\,dw\}$$
$$- \tau_v A(x) \psi(y + hmx)]\,dx - \tau_{1v} \int A(w) \psi(y - hmw)\,dw)\,dy$$
$$= m^{-(v+2)} (\tau_v + \tau_{1v}) J_1 - m^{-(v+2)} n^{-1} \tau_v J_2 ,$$

where J_1 and J_2 are as defined prior to the statement of the lemma. We may now deduce from (5.2.5) that

$$\eta(h,f,f_1) = m^{-(2\nu+5)} J_1 \sum_v (\tau_v^2 - \tau_{1v}^2) - m^{-(2\nu+5)} n^{-1} J_2 \sum_v (\tau_v - \tau_{1v})\tau_v$$

$$= m^{-(2\nu+5)} J_1 \sum_v (|\tau_v| - |\tau_{1v}|) - m^{-(2\nu+5)} n^{-1} \sum_v (|\tau_v| - \tau_v \tau_{1v}) ,$$

completing the proof of the lemma.

Next the size of J_1 and J_2 are described. Put

$$t = (1/4) |\{\int w^2 A(w)\,dw\}\{\int x^2 K(x)\,dx\}| \neq 0 .$$

Lemma 5.2.2. (i) *If* $\nu \geqq 0.25$ *then* $|J_2| = O\{(hm)^2\}$ *as* $n \to \infty$. (iia) *If* $\nu > 0.25$ *and* $h \leqq cn^{-1/5}$, *any fixed* $c > 0$, *then* $|J_1| \sim t(hm)^4$ *as* $n \to \infty$, (iib) *If* $\nu = 0.25$ *and* $a > 0$ *then* ξ *and* η *(the latter in the definition of* m*) may be chosen so small that*

$$\liminf_{n\to\infty} \inf_{h\in H_n} |J_1| > 0 .$$

Proof of Lemma 5.2.2. We shall prove only (iia) and (iib), since the derivation of (i) is simpler. Note that J_1 depends on h and m only through the product, hm. In case (iia) we have $hm \to 0$ (since $hm \leqq$ const $n^{-(1/5)+2/(4\nu+9)}$ and $\nu > 0.25$). In case (iib) the product hm may be made arbitrarily small by choosing η small (note that $m = \eta n^{1/5}$ in this circumstance). Therefore if suffices to prove that with hm replaced by ζ in the formula for J_1, we have $|J_1| \sim t\zeta^4$ as $\zeta \to 0$.

Observe that

$$J_1 = \int \psi(y) \int\int A(w)K(x)\{(1/2)\zeta^2(w+x)^2\psi''(y) + (1/24)\zeta^4(w+x)^4\psi^{(4)}(y)\}\,dw\,dx$$

$$- \int A(w)\{(1/2)\zeta^2 w^2 \psi''(y) + (1/24)\zeta^4 w^4 \psi^{(4)}(y)\}\,dw]\,dy + O(\zeta^5)$$

$$= (1/2)t_2(\int \psi\psi'')\zeta^2 + (1/24)t_4(\int \psi\psi^{(4)})\zeta^4 + O(\zeta^5 + n^{-1}\zeta^2) .$$

where

$$t_2 \equiv \int\int A(w)K(x)(w+x)^2\,dw\,dx - \int A(w)w^2\,dw = 0 ,$$

$$t_4 \equiv \int\int A(w)K(x)(w+x)^4\,dw\,dx - \int A(w)w^4\,dw$$

$$= 6\{\int w^2 A(w)\,dw\}\{\int x^2 K(x)\,dx\} \neq 0 .$$

In consequence,

$$J_1 = (1/24)t_4 \int (\psi'')^2 \zeta^4 + o(\zeta^4) .$$

Since $t = |t_4|/24$ then we have established the desired result, completing the proof of Lemma 5.2.2.

Finally we establish (5.2.4). Since $\tilde{f} \in \mathscr{F}$, $\tilde{f}$ admits a representation of the form (5.2.1),

$$\tilde{f} = g_0 + \sum_{v=1} \tilde{\tau}_v \gamma_v$$

say. Hence by Lemma 5.2.1,

$$\eta(h,f,\tilde{f}) = m^{-(2\nu+5)} J_1 \sum_v (|\tau_v| - |\tilde{\tau}_v|) - m^{-(2\nu+5)} n^{-1} J_2 \sum_v (|\tau_v| - \tau_v \tilde{\tau}_v) .$$

In view of Lemma 5.2, the absolute value of the second term on the right-hand side is dominated by

$$C_1 m^{-(2\nu+5)} n^{-1} (hm)^2 m \leqq C_2 n^{-7/5} m^{-2(\nu+1)}$$

$$\leqq C_3 n^{-(7/5)-4(\nu+1)/(4\nu+9)} = o(n^{-(4/5)-4\nu/(4\nu+9)}) ,$$

where C_1, C_2 and C_3 are constants. Furthermore, $n^{(4/5)+4\nu/(4\nu+9)} m^{-(2\nu+5)} J_1$ is of size

$$n^{(4/5)+4\nu/(4\nu+9)} m^{-(2\nu+1)} h^4 \geqq C m^{-1} .$$

Therefore (5.2.4) will follow if we prove that

$$\lim_{\varepsilon \to 0, \delta \to 0} \liminf_{n \to \infty} \max_{f \in \mathscr{F}} P_f \left\{ \left| \sum_v (|\tau_v| - |\tilde{\tau}_v|) \right| > \varepsilon m \right\} = 1 . \tag{5.2.6}$$

Conditional on the data, let $\tau_0^*, \tau_1^*, \ldots, \tau_m^*$ be a sequence of independent, symmetric variables taking only the values 1 and -1, and put $I = (1 - \tau_0^*)/2$ (so that $I = 0$ or 1, each with probability 1/2). Define

$$f^* = g_0 + \sum_{v=1}^{m} \tau_v^* \gamma_v , \quad f^\dagger = I g_0 + (1 - I) f^* .$$

Both f^* and $f^\dagger$ are random elements of $\mathscr{F}$. Write $E^\dagger$ for expectation with respect to the distribution of the random quantity $f^\dagger$. Consider the problem of discriminating between $f = f^\dagger$ (a random density) and $f = g_0$, in a context where either density may arise with probability 1/2. Note that

$$\left| \sum_v (|\tau_v| - |\tilde{\tau}_v|) \right| = \begin{cases} m - \sum_v |\tilde{\tau}_v| & \text{if } f = f^* \\ \sum_v |\tilde{\tau}_v| & \text{if } f = g_0 . \end{cases}$$

Therefore if m is an odd integer (which we may assume is the case) then the event

$$A(f,\tilde{f}) = \{ |\textstyle\sum (|\tau_v| - |\tilde{\tau}_v|)| > m/2 \}$$

holds for one but not both of $f = f^*$ and $f = g_0$. It may be interpreted as a decision rule which decides in favor of f^* whenever $A(f^*,\tilde{f})$ fails, and in favor of g_0 otherwise. Call this rule R_1, let R_2 be the likelihood ratio rule, and write π_i for the average probability of misclassification under rule R_i. Then $\pi_1 \geqq \pi_2$, and it follows as in Bickel and Ritov (1988) that

$$\lim_{\delta \to 0} \liminf_{n \to \infty} \pi_2 = 1 . \tag{5.2.7}$$

Now,

$$\max_{f\in\mathscr{F}} P_f\left\{\left|\sum_v (|\tau_v| - |\tilde{\tau}_v|)\right| > m/2\right\}$$

$$\geqq E^{\dagger}\left[P_{f'}\left\{\left|\sum_v (|\tau_v| - |\tilde{\tau}_v|)\right| > |m/2 \mid f^{\dagger}\right\}\right] = \pi_2 \geqq \pi_1 .$$

Result (5.2.6) follows from these inequalities and (5.2.7).

5.3. Proof of Theorem 2.3

The context here is related to that in Theorem 3.2, if we think of $m = 1$, $\nu = \infty$, $m^{-\nu} = n^{-1}$, $\psi = \alpha f_0$, and the two densities as being f_0 and $f_1 = f_0 + n^{-1/2}\psi$ instead of g_0 and $g_0 + m^{-(2+\nu)}\gamma_1$. With these changes, the argument is a hybrid of those for Theorems 2.2 and 3.2. In particular, it suffices to prove, instead of (5.1.3) or (5.2.4), that

$$\lim_{\varepsilon\to 0} \liminf_{n\to\infty} \max_{f\in\{f_0, f_1\}} P_f\left\{\min_{h\in(an^{-1/5}, bn^{-1/5})} |\eta(h, f, \tilde{f})| > \varepsilon n^{-13/10}\right\} \geqq p , \tag{5.3.1}$$

where, by an analogue of (5.2.5)

$$|\eta(h, f, \tilde{f})| = \begin{cases} 0 & \text{if } \tilde{f} = f \\ n^{-1/2}|I| & \text{if } \tilde{f} \neq f' \end{cases} , \tag{5.3.2}$$

and where I is given by

$$I = \int \psi(y)\left(\int [(1 - n^{-1})\{K(x)E_{f_0}(\hat{f}_h - \hat{g}_h)(y + hx) + A(x)E_{f_1}(\hat{f}_h)(y + hx)\} - A(x)f_0(y + hx)]\,dx - E_{f_1}(\hat{f}_h - \hat{g}_h)(y)\right) dy .$$

Note that

$$E_{f_0}(\hat{f}_h - \hat{g}_h)(y + hx) = h^{-1}\int A\{(y + hx - w)/h\} f_0(w)\,dw$$
$$= \int A(w) f_0\{y + h(x - w)\}\,dw ,$$
$$E_{f_1}(\hat{f}_h)(y + hx) = \int K(w) f_1\{y + h(x - w)\}\,dw ,$$
$$E_{f_1}(\hat{f}_h - \hat{g}_h)(y) = h^{-1}\int A(w) f_1(y - hw)\,dw .$$

It follows that

$$I = \int \psi(y)[(1 - n^{-1})\iint A(w)K(x)\{f_0(y + h(w + x)) + f_1(y + h(w + x))\}\,dw\,dx - \int A(w)\{f_0(y + hw) + f_1(y + hw)\}\,dw]\,dy$$
$$= 2\{(1/2)t_2(\int \psi f_0'')h^2 + (1/24)t_4(\int \psi f_0^{(4)})h^4 + O(h^5 + n^{-1}h^2)\} ,$$

following the arguments in the proof of Lemma 5.2.2.

Since $t_2 = 0$,

$$I = (1/12)t_4(\int \psi f_0^{(4)})h^4 + o(h^4) .$$

The desired result (5.3.1) follows from this formula, (5.3.2), and the facts $\int \psi f_0^{(4)} = \int \psi^{(4)} f_0$ and

$$\liminf_{n\to\infty} \max_{f\in\{f_0,f_1\}} P_f(\tilde{f} \neq f) \geqq p .$$

This last is established in manner identical to (5.1.4).

5.4. Proof of Theorem 4.1

Write $g = f^{(m)}$. We must show that with

$$J(h) = \int\int K(u)g(z)\{g(z-hu) - g(z)\}\,du\,dz$$

we have $J(h) = C_k h^k + o(h^k) + O(h^{2r})$, where $r = 2 + v - m$.

Let l denote the largest integer strictly less than $2 + v - m$. By assumption, $g^{(l)}$ exists and is bounded. Assume initially that $g^{(2l)}$ exists and is bounded. From a Taylor expansion with an integral remainder,

$$g(z+\varepsilon) - g(z) = \sum_{i=1}^{2l-1} (\varepsilon^i/i!)g^{(i)}(z) + \{\varepsilon^{2l}/(2l-1)!\}\int_0^1 (1-t)^{2l-1} g^{(2l)}(z+t\varepsilon)\,dt .$$

It follows from this fact, and the identities

$$\int g(z)g^{(2i)}(z+\delta)\,dz = (-1)^i \int g^{(i)}(z)g^{(i)}(z+\delta)\,dz ,$$

$$\int g(z)g^{(2i+1)}(z)\,dz = 0 ,$$

that

$$J(h) = \sum_{i=1}^{l-1} \alpha_i h^{2i} + \beta_l h^{2l} \int_0^1 (1-t)^{2l-1} \int u^{2l}K(u) \int g^{(l)}(z)g^{(l)}(z-hut)\,dz\,du\,dt ,$$

where

$$\alpha_i = (-1)^i \{\int (g^{(i)})^2\}\{\int u^{2i}K(u)\,du\}/(2i)! ,$$

$$\beta_i = (-1)^i/(2i-1)! .$$

This formula only involves derivatives of g up to the l'th, and so must hold under the assumption that $g^{(l)}$ (not $g^{(2l)}$) is bounded. (Approximate g arbitrarily closely by a function with $2l$ derivatives.) Under this assumption,

$$J(h) = \sum_{i=1}^{l} \alpha_i h^{2i} + \beta_l h^{2l} \int_0^1 (1-t)^{2l-1} \int u^{2l}K(u) \times \int g^{(l)}(z)\{g^{(l)}(z-hut) - g^{(l)}(z)\}\,dz\,du\,dt . \tag{5.4.1}$$

Now given $B > 0$, define $H_q(B)$ to be the set of all functions which vanish outside $(-B, B)$ and satisfy (3.1.1) in the special case $l = 0$, $v = q - 2$. A useful lemma at this stage (whose proof is deferred to the end of this section) is

Lemma 5.4. *If $a \in H_q(B)$, for $0 \leqq q \leqq 1$ and for some $B > 0$, then*

$$I(\varepsilon) \equiv \int a(z)\{a(z+\varepsilon) - a(z)\}\,dz = O(|\varepsilon|^{2q})$$

as $\varepsilon \to 0$.

An application of Lemma 5.4, with $q = 2 + \nu - m - l$ and $a = g^{(l)}$, shows that

$$\int g^{(l)}(z)\{g^{(l)}(z+\varepsilon) - g^{(l)}(z)\}\,dz = O(|\varepsilon|^{2q})$$

as $\varepsilon \to 0$. Hence by (5.4.1),

$$J(h) = \sum_{i=1}^{l} \alpha_i h^{2i} + O(h^{2l+2q}) .$$

Since K is a k'th order kernel, it follows that $\alpha_i = 0$ for $2i < k$, and so $J(h) = C_k h^k + O(h^{2r})$, where $r = l + q = 2 + \nu - m$.

Proof of Lemma 5.4.
Without loss of generality, suppose that a vanishes outside (1/3, 2/3) and that $\varepsilon \geqq 0$. Observe that

$$2I(\varepsilon) = \int a(z)\{a(z+\varepsilon) - 2a(z) + a(z-\varepsilon)\}\,dz . \tag{5.4.2}$$

Put $c_i = \varepsilon^{-1} \int_{i\varepsilon}^{(i+1)\varepsilon} a(z)\,dz$. In view of the definition of $G_q(B)$,

$$\sup_{1 \leqq i \leqq \varepsilon^{-1}} |c_{i+1} - c_i| = O(\varepsilon^q) , \tag{5.4.3}$$

$$\sup_{1 \leqq i \leqq \varepsilon^{-1}} \sup_{i\varepsilon \leqq z \leqq (i+1)\varepsilon} |a(z) - c_i| = O(\varepsilon^q) , \tag{5.4.4}$$

$$\sup_{-\infty < z < \infty} |a(z+\varepsilon) - 2a(z) + a(z-\varepsilon)| = O(\varepsilon^q) , \tag{5.4.5}$$

as $\varepsilon \to 0$. From (5.4.4) and (5.4.5) it follows that

$$\int_{i\varepsilon}^{(i+1)\varepsilon} a(z)\{a(z+\varepsilon) - 2a(z) + a(z-\varepsilon)\}\,dz$$

$$= c_i \int_{i\varepsilon}^{(i+1)\varepsilon} \{a(z+\varepsilon) - 2a(z) + a(z-\varepsilon)\}\,dz + O(\varepsilon^{2q+1}) ,$$

uniformly in $1 \leqq i \leqq \varepsilon^{-1}$. The first term on the right side is identically $\varepsilon c_i(c_{i+1} - 2c_i + c_{i-1})$. Adding over $1 \leqq i \leqq \varepsilon^{-1}$, and noting (5.4.2), observe that

$$2I(\varepsilon) = \varepsilon \sum_{1 \leqq i \leqq \varepsilon^{-1}} c_i(c_{i+1} - 2c_i + c_{i-1}) + O(\varepsilon^{2q}) , \tag{5.4.6}$$

where $c_0 \equiv 0$.

Abel's method of summation provides the formula

$$\sum_{i=1}^{N} a_i b_i = \sum_{i=1}^{N-1} A_i(b_i - b_{i+1}) + A_N b_N , \tag{5.4.7}$$

where $A_i = \sum_{j=1}^{i} a_j$. Take $a_i = c_{i+1} - 2c_i + c_{i-1}$ and $b_i = c_i$, and use (5.4.7) to simplify the series on the right hand side of (5.4.6). Note that since a vanishes

outside (1/3, 2/3), it follows that for sufficiently small ε, $c_i = 0$ for $i < 1/(6\varepsilon)$ or $i > 5/(6\varepsilon)$. This entails $A_N = 0$ if $N > 5/(6\varepsilon)$, and $A_i = c_{i+1} - c_i$. Hence

$$2I(\varepsilon) = \varepsilon \sum_{1 \leqq i \leqq \varepsilon^{-1}} (c_{i+1} - c_i)(c_i - c_{i+1}) + O(\varepsilon^{2q}) .$$

It is now a consequence of (5.4.3) that $I(\varepsilon) = O(\varepsilon^{2q})$, as required.

Acknowledgement. We are grateful to Professors P. Bickel and Y. Ritov for providing a copy of their unpublished technical report. The referees made many insightful and helpful comments.

References

Anderson, G.D.: A comparison of methods for estimating a probability density function. Phd Dissertation, University of Washington, 1969

Bickel, P., Ritov, Y.: Estimating integrated squared density derivatives: sharp best order of convergence estimates. Sankhya **50-A**, 381–393 (1988)

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

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

Devroye, L., Györfi, L.: Nonparametric density estimation: the L_1 View. New York; Wiley 1984

Donoho, D., Liu, R.: Geometrizing rates of convergence (unpublished manuscript 1987)

Es, B. van.: Likelihood cross-validation bandwidth selection for nonparametric kernel density estimators. J. Nonparamet. Stat. (in press 1991)

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

Hall, P.: Limit theorems for stochastic measures of the accuracy of density estimators. Stochastic Processes Appl. **13**, 11–25 (1982)

Hall, P., Marron, J.S.: Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. Probab. Th. Rel. Fields **74**, 567–581 (1987a)

Hall, P., Marron, J.S.: On the amount of noise inherent in bandwidth selection for a kernel density estimator. Ann. Stat. **15**, 163–181 (1987b)

Hall, P., Marron, J.S.: Estimation of integrated squared density derivatives. Stat. Probab. Lett. **6**, 109–115 (1987c)

Härdle, W., Hall, P., Marron, J.S.: How far are automatically chosen regression smoothers from their optimum?. J. Am. Stat. Assoc. **83**, 86–95 (1988)

Mammen, E.: A short note on optimal bandwidth selection for kernel estimators. Stat. Probab. Lett. **9**, 23–25 (1988)

Marron, J.S.: Convergence properties of an empirical error criterion for multivariate density estimation. J. Multivariate Anal. **19**, 1–13 (1986)

Marron, J.S.: Automatic smoothing parameter selection: A survey. Emp. Econ. **13**, 187–208 (1988)

Marron, J.S.: Comments on a data-driven bandwidth selector. Comp. Stat. Data Anal. **8**, 155–170 (1989)

Marron, J.S., Härdle, W.: Random approximations to some measures of accuracy in nonparametric curve estimation. J. Multivariate Anal. **20**, 91–113 (1986)

Park, B.U., Marron, J.S.: Comparison of data-driven bandwidth selectors. J. Am. Stat. Assoc. **85**, 66–72 (1990)

Rosenblatt, M.: Remarks on some non-parametric estimates of a density function. Ann. Math. Stat. **27**, 832–837 (1956)

Rosenblatt, M.: Curve estimates. Ann. Math. Stat. **42**, 1815–1842 (1971)

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

Scott, D.W., Terrell, G.R.: Biased and unbiased cross-validation in density estimation. J. Am. Stat. Assoc. **82**, 1131–1146 (1987)

Silverman, B.W.: Density estimation for statistics and data analysis. New York: Chapman and Hall 1986

Steele, J.M.: Invalidity of average squared error criterion in density estimation. Can. J. Stat. **6**, 193–200 (1978)

Stone, C.J.: Optimal convergence rates for nonparametric estimators. Ann. Stat. **8**, 1348–1360 (1980)

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

Watson, G.S., Leadbetter, M.R.: On the estimation of the probability density, I. Ann. Math. Stat. **34**, 480–491 (1963)

Wegman, E.J.: Nonparametric probability density estimation: II. A comparison of density estimation methods. J. Stat. Comput. Simulation **1**, 225–245 (1972)