THE DOUBLE KERNEL METHOD IN DENSITY ESTIMATION

Luc Devroye
School of Computer Science
McGill University
Montreal, Canada H3A 2K6
luc@cs.mcgill.ca

Abstract. Let f_{nh} be the Parzen-Rosenblatt kernel estimate of a density f on the real line, based upon a sample of n i.i.d. random variables drawn from f, and with smoothing factor h. Let g_{nh} be another kernel estimate based upon the same data, but with a different kernel. We choose the smoothing factor H so as to minimize $\int |f_{nh} - g_{nh}|$, and study the properties of f_{nH} and g_{nH}. It is shown that the estimates are consistent for all densities provided that the characteristic functions of the two kernels do not coincide in an open neighborhood of the origin. Also, for some pairs of kernels, and all densities in the saturation class of the first kernel, we show that

$$\limsup_{n\to\infty} \frac{\mathbf{E}\left\{\int |f_{nH} - f|\right\}}{\mathbf{E}\left\{\inf_h \int |f_{nh} - f|\right\}} \le C,$$

where C is a constant depending upon the pair of kernels only. This constant can be arbitrarily close to one.

Keywords and phrases. Density estimation. Asymptotic optimality. Nonparametric estimation. Strong convergence. Kernel estimate. Automatic choice of the smoothing factor.

1991 Mathematics Subject Classifications: 62G05, 62H99, 62G20.

Author's address: School of Computer Science, McGill University, 3480 University Street, Montreal, Canada H3A 2K6. The author's research was sponsored by NSERC Grant A3456 and FCAR Grant 90-ER-0291. FAX number: 1–514–398-3883.

TABLE OF CONTENTS

Section 1: INTRODUCTION

We consider the standard problem of estimating a density f on R^1 from an i.i.d. sample $X_1, ..., X_n$ drawn from f. The density estimates considered in this note are the well-known kernel estimates

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

where $h > 0$ is a smoothing factor, K is an absolutely integrable function called the kernel, $\int K = 1$, and $K_h(x) = (1/h)K(x/h)$ (Parzen, 1962; Rosenblatt, 1956). We are particularly interested in smoothing factors that are functions of the data (which are denoted by H to reflect that they are random variables). Most proposals for functions H found in the literature minimize some criterion; for example, many attempt to keep $\int (f_{nH} - f)^2$ as small as possible. Extending Stone (1984), we say that f_n is asymptotically optimal for f if H is such that

$$\frac{\mathrm{E}\left\{\int |f_{nH} - f|\right\}}{\inf_h \mathrm{E}\left\{|f_{nh} - f|\right\}} \to 1$$

as $n \to \infty$. Stone defined this notion with L_2 errors instead of L_1 errors, without the expected values, and with almost sure convergence to one for the ratio. In Theorem S1, we will show that for our way of choosing H, $\mathrm{E} \inf_h$ and $\inf_h \mathrm{E}$ can be used interchangeably, and that $\int |f_{nH} - f| / \mathrm{E} \int |f_{nH} - f| \to 1$ almost surely, so that all definitions are equivalent for our method.

It should be noted that f_n may be asymptotically optimal for some f and K and not for other choices. A perhaps too trivial example is that in which K is the uniform density on $[-1, 1]$, and $H = cn^{-1/5}$ where c is known to be optimal for the normal $(0,1)$ density and the given K. Obviously, such a choice leads in general to asymptotic optimality for the normal $(0,1)$ density, and to suboptimality in nearly all other cases. It is clearly of interest to the practitioner to make the class of densities on which asymptotic optimality is obtained as large as possible. Remarkably, in the L_2 work of Stone (1984), asymptotic optimality was established for all bounded densities and all bounded kernels of compact support if H is chosen by the L_2 cross-validation method of Rudemo (1982) and Bowman (1984). This H is of little use in L_1, and the method is dangerous: for some unbounded densities, we have $\liminf \mathbf{E}\left\{\int |f_{nH} - f|\right\} \geq 1$ (Devroye, 1988). Since data-based techniques for choosing H are supposed to be automated and inserted into software packages, it is important that the method be consistent.

It is perhaps useful to reflect on the possible strategies. Hall and Wand (1987) have proposed a plug-in adaptive method, in which unknown quantities in the theoretical formula for the asymptotically optimal h are estimated from the data using other nonparametric estimates, and then plugged back in the formula to obtain H. Similar strategies have worked in the past for L_2 (see, e.g., Woodroofe (1970), Nadaraya (1974) and Bretagnolle and Huber (1979)). The advantages of this approach are obvious: the designer clearly understands what is going on, and the problem is conceptually cut in clearly identifiable subproblems. On the other hand, how does one choose the smoothing parameters needed for the secondary nonparametric estimates? And, assuming that the conditions for the theoretical formula for h are not fulfilled, isn't it possible to obtain inconsistent density estimates? To avoid the latter drawback, it is imperative to go back to first principles. Cross-validated maximum likelihood products have been studied by many: Duin (1976) and Habbema, Hermans and Vandenbroek (1974) proposed the method, and Chow, Geman and Wu (1983), Schuster and Gregory (1981), Hall (1982), Devroye and Györfi (1985), Marron (1985) and Broniatowski, Deheuvels and Devroye (1988) studied the consistency and rate of convergence. Unfortunately, the maximum likelihood methods for choosing h pertain to the Kullback-Leibler distances between densities, and bear little relation to the L_1 criterion under investigation here. The L_2 cross-validation method proposed by Rudemo (1982) and Bowman (1984) has no straightforward extension to L_1. Its properties in L_2 are now well understood, see, e.g., Hall (1983,1985), Stone (1984), Burman (1985), Scott and Terrell (1987) and Hall and Marron (1987). This seems to leave us empty-handed were it not for the versatility of the kernel estimate itself. Indeed, the method we are about to propose does not easily generalize beyond the class of kernel estimates.

The estimator proposed below has two advantages:

A. It is consistent for all f, i.e., $\mathbf{E}\left\{\int |f_{nH} - f|\right\} \to 0$ for all f.

B. For a large family of nice densities, we have C-optimality, i.e., there exists a constant C such that for all f in the class,

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

The constant C can be as close to one as desired. We define our H simply as the h that minimizes $\int |f_{nh} - g_{nh}|$, where g_{nh} is the kernel estimate based upon the same data, but with kernel L instead of kernel K. The key idea is that most kernels considered in practice have built-in limitations, including the class of all kernels with compact support. For any such kernel K, it is fairly easy to construct another kernel L whose bias is asymptotically superior in the sense that

$$\lim_{h \downarrow 0} \frac{\int |f * L_h - f|}{\int |f * K_h - f|} = 0,$$

where $*$ is the convolution operator. The class of densities f for which this happens coincides roughly speaking with the class of densities for which $\int |f * K_h - f|$ tends to zero at the best possible rate (or: saturation rate) for the given K. These classes are rich, but they won't satisfy everyone. What the improved kernel can do for us is simple: it is very likely that g_{nh}, the kernel estimate with K replaced by L, is much closer to f than f_{nh}, and thus that $\int |f_{nh} - g_{nh}|$ is of the order of magnitude of $\int |f_{nh} - f|$.

We won't worry here about the numerical details. First of all, if K and L are polynomial and of compact support (as they often are), then the integral to be minimized can be rewritten conveniently as a finite sum with $O(n)$ terms, by considering that each kernel estimate is piecewise polynomial with $O(n)$ pieces at most. The minimization with respect to h is a bit harder to do. Observing that the function to be optimized is uniformly continuous on any interval $(a, b) \subseteq [0, \infty)$, we see that the minimum exists and is a random variable. For more general non-polynomial K and L, under some smoothness conditions, we still have

$$\left| \int |f_{nh} - f| - \int |f_{nh'} - f| \right| \le c \left| \frac{h - h'}{h} \right|$$

for some $c > 0$. For h, h' close enough, this can be made smaller than $1/n$, which is much smaller than the smallest possible L_1 error (which is $1/\sqrt{528n}$ by Devroye, 1988). Thus, the minimization can be carried out over a grid of points, and in any case, it is possible to define a random variable H with the property that $\int |f_{nH} - g_{nH}| \sim \inf_h \int |f_{nh} - g_{nh}|$.

There is another interesting by-product of this method, namely that we end up with two kernel estimates f_{nH} and g_{nH}, where for the class of densities under consideration, g_{nH} is probably better than f_{nH}. Interestingly, f_{nH} is asymptotically optimal for K but g_{nH} is usually not asymptotically optimal for L. One way of looking at our method is as a technique for creating a better estimate (g_{nh}) without imposing additional smoothness conditions on the densities. Another by-product of the method is that $\int |f_{nH} - g_{nH}|$ is a rough estimate of the actual error $\int |f_{nH} - f|$. Unfortunately, if one decides to use g_{nH} instead of f_{nH}, then $\int |f_{nH} - g_{nH}|$ provides little information about the actual error obtained with g_{nH}.

Not all pairs (K, L) are useful. The most important property needed to be fulfilled is that the characteristic functions of K and L do not coincide in an open neighborhood of the origin. This often forces K and L to be kernels of a different order. In addition, we will see that the constant C can be chosen equal to $(1 + u)/(1 - u)$, where $u = 4\sqrt{\int L^2 / \int K^2}$.

The length of the paper is partially explained by the fact that we wanted to state as many properties of the estimate as possible in a density-free manner. This also renders the results more useful for future work on the same topic. Among the density-free results, we cite

- The consistency (Theorem C1).
- The complete convergence (Lemma O2).

- The strong relative stability of the estimate (Lemma O5).
- Bounds on the error that are uniform over all h (Theorem C2 and Lemma O1).
- Necessary conditions of convergence (Lemmas C1 and C2).
- A universal lower bound for the expected error (Lemma O5).
- Universal lower bounds for the variation in the error (Lemmas O7, O8).

While it is good to know that the estimate always converges and is C-optimal for virtually all densities in the saturation class of a kernel K, it is informative to find out what we have not been able to achieve. First of all, the kernels K considered here for C-optimality are class s kernels, i.e., all their moments up to but not including the s-th moment vanish. This implies, as we will see, that the expected error can go to zero no faster than a constant times $n^{-s/(2s+1)}$ for any density. In this respect, we are severely limited, since it is well-known that for very smooth densities kernels can be found that yield error rates that are $O(1/\sqrt{n})$ or come close to it (see, e.g., Watson and Leadbetter (1963) for an equivalent statement in the L_2 setting). We can exhibit explicit constants C and D such that for n large enough,

$$\mathrm{E}\left\{\int |f_{nH} - f|\right\} \le C \inf_h \mathrm{E}\left\{\int |f_{nh} - f|\right\} + D\sqrt{\frac{\log n}{n}}.$$

This inequality implies that C-optimality can only be hoped for when the best possible error rate for the present K and f is at least be $\sqrt{\log n/n}$. Unfortunately, this would exclude such interesting densities as the normal density, for which we can get $O(\log^{1/4} n/\sqrt{n})$ (Devroye, 1988). In particular, it seems that for analytic densities in general, the techniques presented here need some strengthening.

But perhaps the biggest untackled question is what happens to the expected error for densities f that are not in the saturation class of K; these are usually densities that are not smooth enough or not small-tailed enough to attain the rate $n^{-s/(2s+1)}$. Despite this, it may still be possible to apply the present minimization technique to obtain good asymptotic performance for most of them. All that is needed is to verify the fact that for the pair (K, L), $\int |f * L_h - f| = o(\int |f * K_h - f|)$. On the other hand, it is also possible that a general result such as the one obtained by Stone (1984) for L_2 errors does not exist in the L_1 setting.

Section 2: CONSISTENCY

The purpose. The purpose of this section is to prove the following

Theorem C1. *Let K and L be absolutely integrable kernels such that their (generalized) characteristic functions do not coincide on any open neighborhood of the origin, and let f be an arbitrary density. Then $\mathrm{E}\int |f_{nH} - f|$ and $\mathrm{E}\int |g_{nH} - f|$ tend to zero as $n \to \infty$.*

One of the difficulties with this sort of Theorem is that it needs to be shown for all densities f, even those f for which the procedure for selecting H is not specifically designed. Furthermore, the

L_1 errors are not easily decomposed into bias and variance terms, since H depends upon the data, so that conditional on H, the summands in the definition of the kernel estimate are not independent. We will provide a mechanism for decoupling H and the data. In the final analysis, the proof of Theorem C1 rests on an exponential inequality of Devroye (1988) and some other properties of the random function $\int |f_{nh} - f|$ (considered as a function of h). The proof will be cut into many lemmas, some of which will be useful outside this paper and in other sections.

The condition on K and L implies that $\int |K - L| > 0$. It is possible to have consistency even if the characteristic functions of K and L coincide on some open neighborhood of the origin, but such consistency would not be universal; it would apply only to densities whose characteristic function vanishes off a compact set. The details for such cases can be deduced from the proof. We have also unveiled where it is possible to go wrong: it suffices to have the said coincidence of the two characteristic functions, while f has a characteristic function with an infinite tail. In those cases, the H may actually end up tending to a positive constant as $n \to \infty$. Unfortunately, as is well known, for such densities, it is impossible to have consistency unless $H \to 0$. From this, we retain that the behavior of the characteristic function of $K - L$ near the origin is somehow a measure of the discriminatory power of the method. Usually, we take a standard nonnegative kernel for K whose characteristic function varies as $1 - at^2$ near $t = 0$, whereas for L we can take a kernel whose characteristic function is flatter near the origin, behaving possibly as $1 - bt^4$ or even identically 1 on an open neighborhood of the origin.

THE DECOUPLING DEVICE. We have seen that $X_1, \ldots, X_n$ are i.i.d. random variables with density f, and that $H = H(n)$ is a sequence of random variables where $H(n)$ is measurable with respect to the σ-algebra generated by $X_1, \ldots, X_n$, i.e., it is a function of $X_1, \ldots, X_n$. Consider now independent identically distributed copies of the two sequences, denoted by $\widehat{X}_1, \ldots, \widehat{X}_n, \ldots$ and $\hat{H} = \hat{H}(n)$ respectively. Density estimates based upon the former data are denoted by f_{nh},g_{nh},f_{nH} and g_{nH} typically, while for the latter data, we will write $\hat{f}_{nh}$ and so forth.

In our decoupling, we will show that $\int |f_{nH} - f|$ and $\int |\hat{f}_{nH} - f|$ are close in a very strong sense. Note that the second error is that committed if H is used in a density estimate constructed with a new data set. The independence thus introduced will make the ensuing analysis more manageable. To keep the notation simple, we will write $\mathbf{E}_n$ for the conditional expectation given $X_1, \ldots, X_n$, and $\hat{\mathbf{E}}_n$ for the conditional expectation given $\hat{X}_1, \ldots, \hat{X}_n$. With this notation, note that $\mathbf{E}_n \int |\hat{f}_{nH} - f|$ is distributed as $\hat{\mathbf{E}}_n \int |f_{n\hat{H}} - f|$, and thus that $\mathbf{E} \int |\hat{f}_{nH} - f| = \mathbf{E} \int |f_{n\hat{H}} - f|$.

Uniform convergence with respect to h

The first auxiliary result is so crucial that we are permitting ourselves to elevate it to a Theorem:

THEOREM C2. *Let M be an arbitrary absolutely integrable function, and define $m_{nh} = \frac{1}{n} \sum_{i=1}^{n} M_h(x - X_i)$ where $X_1, \ldots, X_n$ are i.i.d. random variables with an arbitrary density f, and $h > 0$ is a real number. Then*

$$\sup_h \left| \int |m_{nh}| - \mathbf{E} \int |m_{nh}| \right| \to 0$$

almost surely as $n \to \infty$. In fact, for every $\epsilon > 0$, there exists a constant $\gamma > 0$ possibly depending upon f, M and ϵ, such that for all n large enough,

$$\mathrm{P}\left\{\sup_h \left| \int |m_{nh}| - \mathrm{E}\int |m_{nh}| \right| > \epsilon \right\} \le e^{-\gamma n}.$$

To see how Theorem C2 exactly provides us with the required decoupling between H and the data, consider the following

COROLLARIES OF THEOREM C2. *Let f_{nh},g_{nh} be kernel estimates with kernels K and L respectively, and define $J_{nh} = \int |f_{nh} - g_{nh}|$ and $\hat{J}_{nh} = \int |\hat{f}_{nh} - \hat{g}_{nh}|$. Then*

A. $\sup_{h>0} |J_{nh} - \mathrm{E}J_{nh}| \to 0$ *almost surely as $n \to \infty$.*

B. *For any random variable H (possibly not independent of the data), $J_{nH} - \mathrm{E}_n\hat{J}_{nH} \to 0$ almost surely as $n \to \infty$.*

C. *For any random variable H (possibly not independent of the data), $J_{nH} \to 0$ in probability implies $\mathrm{E}J_{nH} \to 0$, $\mathrm{E}_n\hat{J}_{nH} \to 0$ in probability, $\mathrm{E}\hat{J}_{nH} \to 0$, and $\mathrm{E}J_{n\hat{H}} \to 0$.*

PROOF. Note first that $|m_{nh} - u_{nh}| \le \int |M - M'|$ when u_{nh} is the kernel estimate with kernel M'. The fact that the bound does not depend upon h and that M' is arbitrary means that we need only show the Theorem for all M that are continuous and of compact support (since the latter collection is dense in the space of L_1 functions).

The first auxiliary result is the following inequality, valid for all fixed h,n,M and f:

$$\mathrm{P}\left\{\left| \int |m_{nh}| - \mathrm{E}\int |m_{nh}| \right| > \epsilon \right\} \le 2e^{-\frac{n\epsilon^2}{32\int^2 |M|}}$$

(Devroye, 1988). It is this inequality that will be extended to an interval of h's using a rather standard grid technique. Set $\Delta(h) = |\int |m_{nh}| - \mathrm{E}\int |m_{nh}||$, and $\Delta(a,b) = \sup_{h,h' \in [a,b]} |\Delta(h) - \Delta(h')|$. Then the following inclusion of events is valid:

$$\left[\sup_{a \le h \le b} \Delta(h) > \epsilon\right] \subseteq \cup_{i=0}^{k} [\Delta(ac^i) > \epsilon/2] \ \cup_{i=1}^{k} [\Delta(ac^i - 1, ac^i) > \epsilon/2]$$

where k is an integer so large that $ac^k \ge b$, and $c > 1$ is such that

$$\sup_{1 \le h \le c} \int |M_1 - M_h| \le \epsilon/4.$$

Such a c can indeed be found, since for all absolutely integrable M, $\lim_{h\to 1}\int |M_1 - M_h| = 0$ (see, e.g., Devroye, 1987, pp. 38-39). The second union in the inclusion inequality is a union of empty events since

$$\Delta(ac^i-1, ac^i) \le \sup_{ac^i-1\le h,h'<ac^i}\left|\int |m_{nh}| - \int |m_{nh'}|\right| + \sup_{ac^i-1\le h,h'<ac^i}\left|\mathbf{E}\int |m_{nh}| - \mathbf{E}\int |m_{nh'}|\right|$$

$$\le 2\sup_{ac^i-1\le h,h'<ac^i}\int |M_h - M_h'| = 2\sup_{1\le h\le c}\int |M_1 - M_h| < \epsilon/2.$$

Thus,

$$\mathbf{P}\left\{\sup_{a\le h\le b}\Delta(h) > \epsilon\right\} \le \sum_{i=0}^{k}\mathbf{P}\left\{\Delta(ac^i) > \epsilon/2\right\} \le 2(k+1)e^{-\frac{n\epsilon^2}{128\int^2 |M|}}.$$

This tends to 0 when k increases at a polynomial rate in n. To see how large k just is, note that $ac^k \ge b$, so that $k \ge \log(b/a)/\log(c)$. Since c is a constant depending upon ϵ and M only, it suffices to have limits a and b that are such that $b/a \le d^n$ for some constant d. We will only need the present inequality with $b/a = O(n)$, so that k can be taken smaller than $d + \log(n)$ for some constant d. Recapping, we need only establish that for some sequences $a = a(n)$ and $b = b(n)$ with $b/a = O(n)$ that

$$\lim_{n\to\infty}\left(\mathbf{P}\left\{\sup_{h<a}\Delta(h) > \epsilon\right\} + \mathbf{P}\left\{\sup_{h>b}\Delta(h) > \epsilon\right\}\right) = 0.$$

Assume that M vanishes off $[-s, s]$. Take $a = s\delta/2n$ where $\delta > 0$ is a constant to be picked further on. Note that $1 \ge \int |m_{nh}|/\int |M| \ge \frac{n-N}{n}$ where N is the number of X_i's for which $[X_i - 2a, X_i + 2a]$ has at least one X_j with $j \ne i$. The inequality is uniform over $h \le a$. We have

$$\mathbf{P}\left\{\sup_{h<a}\Delta(h) > \epsilon\right\} \le \mathbf{P}\left\{\sup_{h<a}\left|\int |m_{nh}| - \int |M|\right| > \epsilon/2\right\} + \mathbf{P}\left\{\sup_{h<a}\left|\mathbf{E}\int |m_{nh}| - \int |M|\right| > \epsilon/2\right\}$$

$$\le \mathbf{P}\left\{\frac{N}{n} > \frac{\epsilon}{2\int |M|}\right\} + \mathbf{P}\left\{\frac{\mathbf{E}N}{n} > \frac{\epsilon}{2\int |M|}\right\}.$$

By Markov's inequality, this can be made smaller than a given small constant ϵ' if

$$\frac{\mathbf{E}N}{n} \le \min(1, \epsilon')\frac{\epsilon}{2\int |M|}.$$

But $\mathbf{E}N/n \le \int f(x)\min\left(1, n\int_{x-s\delta/n}^{x+s\delta/n} f(y)dy\right)dx$ and the right-hand-side tends to $\int f(x)\min(1, 2s\delta f(x))dx$ by the Lebesgue dominated convergence theorem and the Lebesgue density theorem (see, e.g., Wheeden and Zygmund, 1977). It can be made as small as desired by our choice of δ. We conclude that for any $\epsilon, \epsilon' > 0$, we can find $\delta > 0$ small enough such that

$$\limsup_{n\to\infty}\mathbf{P}\left\{\sup_{h<a}\Delta(h) > \epsilon\right\} \le \epsilon'.$$

We should note here that if Markov's inequality is replaced by an exponential bounding method, then it should be obvious that for δ small enough the probability in question can be bounded by e^{-dn} for some constant $d > 0$.

We finally proceed to show that

$$\mathbf{P}\left\{\sup_{h>b}\Delta(h) > \epsilon\right\}$$

can be made arbitrarily small by choosing a large enough constant b. This would conclude the proof of the theorem since $b/a = O(n)$ as required. We have

$$\int |m_{nh}| \geq \frac{n-N}{n}\left(\int_{|x|\leq Th} |M_h(x)|dx - \int_{|x|\leq Th} \sup_{|y|\leq T} |M_h(x) - M_h(x-y)|dx\right) - \frac{N}{n}\int |M|,$$

where T is a large constant, and N is the number of X_i's with $|X_i| > T$. Let ω be the modulus of continuity of M defined by $\omega(u) = \sup_x \sup_{|y|\leq u} |M(x) - M(x+y)|$. By our assumptions on M, $\omega(u) \to 0$ as $u \downarrow 0$. Then

$$\int_{|x|\leq Th} \sup_{|y|\leq T} |M_h(x) - M_h(x-y)|dx \leq \int_{|x|\leq Th} \frac{1}{h}\omega(T/h)dx \leq 2T\omega(T/h) \leq 2T\omega(T/b),$$

when $h \geq b$. Furthermore, by Hoeffding's inequality (Hoeffding, 1963),

$$\mathrm{P}\left\{\frac{N}{n} > 2\int_{|y|>T} f\right\} \leq e^{-2n\int_{|y|>T}^2 f}$$

so that, combining all this,

$$\mathrm{P}\left\{\sup_{h\geq b} |\int |m_{nh}| - \int |M|| > 2T\omega(T/b) + 4\int |M| \int_{y|\geq T} f\right\} \leq e^{-2n\int_{|y|>T}^2 f}.$$

From this, we have since $\int |m_{nh}| \leq \int |M|$,

$$\sup_{h\geq b}\left|\mathrm{E}\int |m_{nh}| - \int |M|\right| \leq 2T\omega(T/b) + 4\int |M| \int_{|y|\geq T} f + \int |M| e^{-2n\int_{|y|>T}^2 f}.$$

For fixed $\epsilon > 0$, we choose T and b so large that the terms on the right hand side are $< \epsilon/6$, $< \epsilon/6$ and $o(1)$ respectively. Then

$$\begin{aligned}\mathrm{P}\left\{\sup_{h\geq b} \Delta(h) > \epsilon\right\} &\leq \mathrm{P}\left\{\sup_{h\geq b}\left|\int |m_{nh}| - \int |M|\right| > \epsilon/2\right\} + \mathrm{P}\left\{\sup_{h\geq b}\left|\mathrm{E}\int |m_{nh}| - \int |M|\right| > \epsilon/2\right\} \\ &\leq e^{-2n\int_{|y|>T}^2 f}\end{aligned}$$

for all n large enough. This concludes the proof of Theorem C2. □

2 Behavior of the minimizing integral

Although this seems rather obvious, we will nevertheless state and prove the following property:

THEOREM C3. *Let H minimize $\int |f_{nh} - g_{nh}|$. For all f and all absolutely integrable K and L, $\int |f_{nH} - g_{nH}|$ tends to 0 almost surely and in the mean. Also, $\mathrm{E}_n \int |\hat{f}_{nH} - \hat{g}_{nH}| \to 0$ almost surely, and $\mathrm{E}\int |f_{n\hat{H}} - g_{n\hat{H}}| \to 0$.*

PROOF. Let the sequence $h^* = h^*(n)$ be such that $\mathrm{E}\int |f_{nh^*} - g_{nh^*}| \sim \inf_h \mathrm{E}\int |f_{nh} - g_{nh}|$. We know that whenever $h \to 0$ and $nh \to \infty$, it follows that $\int |f_{nh} - f| \to 0$ almost surely and in the mean (see, e.g., Devroye, 1983). In particular, $\inf_h \int |f_{nh} - f| \to 0$ almost surely, and $\mathrm{E}\int |f_{nh^*} - g_{nh^*}| \to 0$ as $n \to \infty$. Assume that n is so large that $\mathrm{E}\int |f_{nh^*} - g_{nh^*}| < \epsilon/2$. Then, since $\int |f_{nH} - g_{nH}| \le \int |f_{nh^*} - g_{nh^*}|$ by definition, we see that for such n,

$$\begin{aligned} \mathrm{P}\left\{\int |f_{nH} - g_{nH}| > \epsilon\right\} &\le \mathrm{P}\left\{\int |f_{nh^*} - g_{nh^*}| - \mathrm{E}\int |f_{nh^*} - g_{nh^*}| > \epsilon/2\right\} \\ &\le 2e^{-\frac{n\epsilon^2}{128\int^2 |K-L|}}. \end{aligned}$$

We can now apply the corollaries of Theorem C2, to conclude that $\mathrm{E}_n \int |\hat{f}_{nH} - \hat{g}_{nH}| \to 0$ almost surely, and $\mathrm{E}\int |f_{n\hat{H}} - g_{n\hat{H}}| \to 0$. $\square$

The decoupling necessary for the proof of Theorem C1 is now complete.

Necessary conditions of convergence.

Proof of Theorem C1. From Theorem C3, we see that $\mathrm{E}\int |f_{n\hat{H}} - g_{n\hat{H}}| \to 0$. From Lemmas C1 and C2 we retain that $\hat{H} \to 0$ and $n\hat{H} \to \infty$ in probability as $n \to \infty$. Since $\hat{H}$ and H are identically distributed, the same statement is true for H. By Theorem 6.1 of Devroye and Györfi (1985) or Theorem 3.3 of Devroye (1987), this implies that $\int |f_{nH} - f| \to 0$ in probability for all f and all absolutely integrable K, and similarly for $\int |g_{nH} - f|$. This in turn implies convergence in the mean of both quantities. $\square$

Section 3: C-OPTIMALITY

The main result. This is the main body of the paper, even though it is concerned only with specific subclasses of densities. The kernel estimates considered here are class s-estimates (s is an even positive integer), i.e., estimates based upon class s-kernels, which are kernels K having the following properties:

A. $\int (1 + |x|^s)|K(x)|dx < \infty$.

B. $\int K = 1$, $\int x^i K(x)dx = 0$ for $0 < i < s$, and $\int x^s K(x)dx \ne 0$.

C. K is symmetric.

D. $\int K^2 < \infty$.

Note that nonnegative kernels are at best class 2 kernels. It is worth recalling that for every density f, no matter how h is picked as a function of n,

$$\liminf_{n\to\infty} n^{\frac{s}{2s+1}} \mathrm{E}\int |f_n - f| \ge c$$

where $c > 0$ is a constant depending upon K only (for $s = 2$ and $K \geq 0$, see Devroye and Penrod (1984), and for general s, see Devroye, 1988). This lower bound is not achievable for many densities. The rate $n^{-s/(2s+1)}$ can however be attained for densities with $s-1$ absolutely continuous derivatives (i.e., f, $f^{(1)}$, $\ldots$, $f^{(s-1)}$ all exist and are absolutely continuous), satisfying the tail condition $\int \sqrt{f} < \infty$ (see, e.g., Rosenblatt (1979), Abou-Jaoude (1977), or Devroye and Györfi (1985)). The class of such densities will be called $\mathcal{F}_s$ (or $\mathcal{F}$ when no confusion is possible).

To handle the tails of f satisfactorily, it is necessary to introduce a minor tail condition, slightly stronger than $\int \sqrt{f} < \infty$: we let $\mathcal{W}$ be the class of all f for which $\int |x|^{1+\epsilon} f(x)dx < \infty$ for some $\epsilon > 0$, and let $\mathcal{V}$ be the class of all f for which $\int \sqrt{u_f(x)}dx < \infty$, where $u_f(x) \stackrel{\text{def}}{=} \sup_{|y|\leq 1} f(x+y)$. Devroye and Györfi (1985) noted that $\int \sqrt{f} < \infty$ is virtually equivalent to $\int |x|f(x)dx < \infty$ although there are exceptions both ways. Thus, $\mathcal{F}_s \cap \mathcal{W}$ is not much smaller than $\mathcal{F}_s$. The same is true for $\mathcal{V}$, since $\int \sqrt{f} < \infty$ implies $\int \sqrt{u_f} < \infty$ for most smooth densities (e.g. it is always implied when f is monotone in the tails).

Next, we will impose a weak smoothness condition on an absolutely integrable function M: M is said to be smooth if there exists a constant C such that

$$\sup_{1 \leq h \leq c} \int |M - M_h| \leq C(c-1)$$

for all $c > 1$. Smoothness of a kernel implies that small changes in h induce proportionally small changes on f_{nh} with regard to the L_1 distance. It seems vital to control these changes for any method that is based upon the minimization of a criterion involving h. Consider now the problem of picking the smoothness constant C. For example, if $M \geq 0$ is unimodal at the origin and $\int M = 1$, then we can always take $C = 2$ (see Devroye and Györfi (1985, p. 187)). However, this is not interesting for us, since we need to have smoothness for the difference function $K - L$, which takes on negative values. When M is absolutely continuous, we can take $C = \int |x||M'(x)|dx + \int |M|$. This can be seen as follows, if $h > 1$:

$$\begin{aligned}
\int_0^\infty |M(x) - M_h(x)|dx &= \int_0^\infty \left| \int_x^\infty M' - \int_x^\infty (M_h)' \right| dx \\
&= \int_0^\infty \left| \int_x^\infty M' - \int_x /h^\infty \frac{1}{h} M' \right| dx \\
&\leq \int_0^\infty \left| \int_{x/h}^x \frac{1}{h} M' \right| dx + \int_0^\infty \left| \int_x^\infty (1 - 1/h) M' \right| dx \\
&= \int_0^\infty |M'(z)| \frac{1}{h} \int_z^{zh} dx dz + (1 - 1/h) \int_0^\infty |M| \\
&= \frac{h-1}{h} \int_0^\infty z|M'(z)|dz + \frac{h-1}{h} \int_0^\infty |M|.
\end{aligned}$$

The claim is now obtained by considering $\int_{-\infty}^0$ as well.

Finally, a kernel K is said to be regular if it is bounded, and if there exists a symmetric unimodal integrable nonnegative function M such that $|K| \leq M$.

Our main result can now be announced as follows:

THEOREM O1. *Let K be a smooth regular class s kernel. Assume that $f \in \mathcal{F}_s \cap \mathcal{W}$, and that L is chosen such that to pick L such that*

A. $\int (1+|x|^s)|L(x)|dx < \infty$.

B. $\int L = 1$, $\int x^i L(x)dx = 0$ *for* $0 < i \leq s$.

C. L is symmetric, smooth, and regular.

D. The generalized characteristic functions of K and L do not coincide on any open neighborhood of the origin. (This implies that $\int |K - L| > 0$.)

E. $\int L^2 < \int K^2/16$. *Then, the kernel estimate with smoothing factor H minimizing $\int |f_{nh} - g_{nh}|$, is C-optimal where $C = (1+u)/(1-u)$ and $u = 4\sqrt{\int L^2 / \int K^2}$. When $f \in \mathcal{F}_s \cap \mathcal{V} \cap \mathcal{W}^c$, the same is true, provided that, additionally, L has compact support.*

A pair (K, L) is chosen as a function of s and the constant C only. Rescaling L changes $\int L^2$, and can thus be used to push the value of C as close to one as desired.

EXAMPLE 1. Let us illustrate the choice of L on a simple but important example with $s = 2$ and nonnegative kernels K. There is plenty of evidence in favor of choosing Bartlett's kernel $K(x) = (3/4)(1-x^2)_+$ in those cases, see, e.g., Bartlett (1963), Epanechnikov (1969) and Devroye and Penrod (1984). This 2-kernel is smooth, regular, absolutely integrable and of compact support. It is easy to find 4-kernels with the same properties. A little algebra shows that we can take, for example, the continuous kernel $L(x) = (75/16)(1-x^2)_+ - (105/32)(1-x^4)_+$. Interestingly, this coincides with an optimal 4-kernel described, e.g., in Gasser, Müller and Mammitzsch (1985, Table 1).

EXAMPLE 2. It is interesting to note that the functional form of L can be fixed for all s and K once and for all. It suffices for example to consider bounded smooth symmetric kernels L whose characteristic function is zero in an open neighborhood of the origin satisfying the moment condition A of Theorem O1 for all s. This can be done by defining the characteristic function of L as the convolution of the uniform function on $[-1, 1]$ with a symmetric bounded infinitely many times continuously differentiable function with support on $[-1/2, 1/2]$.

EXAMPLE 3. There is a systematic way of creating higher order kernels. Stuetzle and Mittal (1979) pointed out that $2K - K * K$ is a class $2s$ kernel whenever K is a class s kernel. This can be iterated at will. There are other tricks. For example, if M is a class r kernel, and K is a class s kernel, then $K + M - K * M$ is a class $s + r$ kernel. We can also use higher convolutions for creating better kernels. One can verify that $3K - 3K * K + K * K * K$ is a class $3s$ kernel whenever K is a class s kernel. Good sources of possible definitions of families of kernels that are optimal in certain senses are Müller (1984), Gasser, Müller and Mammitzsch (1985), Su-Wong, Prasad and Singh (1982) and Singh (1979).

A better estimate. We have mentioned that g_{nh} is probably preferable over f_{nh}, Even though it is not asymptotically optimal in any sense for its kernel L, it has a smaller error than that of f_{nh}, basically because the class s kernel used in f_{nh} limits its performance. It is interesting to observe that for any absolutely integrable compact support kernel K, we must have $\int x^s K(x) \neq 0$ for some finite s (see, e.g., Theorem 22 of Hardy and Rogosinski (1962)). Thus, for all such kernels we have the saturation phenomenon, and a judiciously picked L for g_{nh} is potentially better.

THEOREM O2. *Let L and K be as in Theorem O1, and assume that all the conditions of Theorem O1 are satisfied. Then*

$$\limsup_{n\to\infty} \frac{\mathbf{E}\int |g_{nH} - f|}{\mathbf{E}\int |f_{nH} - f|} \leq c \stackrel{\text{def}}{=} 4\sqrt{\int L^2 / \int K^2},$$

and

$$\limsup_{n\to\infty} \frac{\mathbf{E}\int |g_{nH} - f|}{\inf_h \mathbf{E}\int |f_{nh} - f|} \leq \frac{c(1+c)}{1-c}.$$

Theorem O2 basically states that since c can be chosen arbitrarily small, the asymptotic performance of g_{nH} can be made any desired fraction of $\inf_h \mathbf{E}\int |f_{nh} - f|$. The argument that g_{nH} itself is not asymptotically optimal for L can be countered with the observation that no smoothing parameters have to be chosen for g_{nH} either. Unless, of course, one considers the spread of L (measured by $\int L^2$) as a hidden smoothing factor of sorts.

Complete convergence of the L1 error. The first result is related to Theorem C2, but differs in that it is more specific in its error estimates.

LEMMA O1. *Let f be an arbitrary density and let K and L be smooth absolutely integrable kernels with $\int |K - L| > 0$. Then, for arbitrary fixed $\epsilon, u > 0$,*

$$\mathbf{P}\left\{ \sup_{\epsilon/n \leq h \leq 1/\epsilon} \left| \int |f_{nh} - g_{nh}| - \mathbf{E}\int |f_{nh} - g_{nh}| \right| \geq \frac{\sqrt{128}\int |K-L| \sqrt{(1+u)\log(n)}}{\sqrt{n}} \right\}$$

$$\leq (1+o(1)) \frac{C\sqrt{n\log(n)}}{\sqrt{2}\int |K-L| \sqrt{1+u}} n^{-(1+u)},$$

where C is the smoothness constant for $K - L$ (see definition of smoothness above).

PROOF. From the proof of Theorem C2, we recall the following inequality:

$$\mathrm{P}\left\{\sup_{\epsilon/n\le h\le 1/\epsilon}\left|\int|f_{nh}-g_{nh}|-\mathrm{E}\int|f_{nh}-g_{nh}|\right|\ge t\right\}\le 2\left(\frac{\log(n/\epsilon^2)}{\log(c)}+2\right)e^{-\frac{nt^2}{128\int^2|K-L|}},$$

valid for all $t>0$. Here c depends upon t and $K-L$ in the following manner: it is so small that

$$\sup_{1\le h\le c}\int|(K-L)-(K_h-L_h)|\le t/4.$$

Now, upon replacing t by $\sqrt{128}\int|K-L|\sqrt{(1+u)\log(n)/n}$, we obtain as upper bound

$$2\left(\frac{\log(n/\epsilon^2)}{\log(c)}+2\right)n^{-(1+u)}.$$

We are left with the sole problem of choosing c. From our assumption on $K-L$, we see that it suffices to take $c=1+t/(4C)$. Using the fact that $\log(c)\ge 2t/(8C+t))$, the upper bound becomes

$$2\left(\frac{(8C+t)\log(n/\epsilon^2)}{2t}+2\right)n^{-(1+u)}\sim\frac{C\sqrt{n\log(n)}}{\sqrt{2}\int|K-L|\sqrt{1+u}}n^{-(1+u)}$$

when u and ϵ are held fixed and $n\to\infty$. □

LEMMA O2. *Let f be an arbitrary density. Let K and L be smooth absolutely integrable kernels with $\int|K-L|>0$, and let H minimize $\int|f_{nh}-g_{nh}|$. Assume also that the generalized characteristic functions of K and L do not coincide on any open neighborhood of the origin. Then, for arbitrary fixed $\epsilon>0$,*

$$\mathrm{P}\{H\notin[1/(n\epsilon),\epsilon]\}<1/n^2$$

for all n large enough. Furthermore, $H\to 0$ and $nH\to\infty$ completely. Finally, $\int|f_{nH}-f|\to 0$ completely.

PROOF. We will inherit the notation of Lemma O1. Define $t=\sqrt{128}\int|K-L|\sqrt{3\log(n)}/\sqrt{n}$. Then, by Lemma O1,

$$\mathrm{P}\left\{\sup_{1/(n\epsilon)\le h\le\epsilon}\left|\int|f_{nh}-g_{nh}|-\mathrm{E}\int|f_{nh}-g_{nh}|\right|\ge t\right\}\le(1+o(1))\frac{C\sqrt{\log(n)}}{\sqrt{6}\int|K-L|}n^{-5/2}.$$

This will be combined with the fact that

$$\liminf_{n\to\infty}\inf_{h\notin[1/(n\epsilon),\epsilon]}\mathrm{E}\left\{\int|f_{nh}-g_{nh}|\right\}>0$$

(Lemmas C1 and C2), and with the observation that for fixed $u>0$,

$$\mathrm{P}\left\{\sup_h\left|\int|f_{nh}-g_{nh}|-\mathrm{E}\int|f_{nh}-g_{nh}|\right|>u\right\}\le e^{-\gamma n}$$

where $\gamma = \gamma(u) > 0$ (Theorem C2). Let A be the set $[1/(n\epsilon), \epsilon]$. For any $\delta > 0$, we have the following inclusion of events:

$$[H \notin A] \subseteq \left[\sup_{h\in A} |\int |f_{nh} - g_{nh}| - \mathrm{E}\int |f_{nh} - g_{nh}|| \geq t\right] \cup \left[\inf_{h} \mathrm{E}\int |f_{nh} - g_{nh}| + t \geq \delta\right]$$
$$\cup \left[\inf_{h\notin A} \mathrm{E}\int |f_{nh} - g_{nh}| \leq 2\delta\right] \cup \left[\inf_{h\notin A} \int |f_{nh} - g_{nh}| - \mathrm{E}\int |f_{nh} - g_{nh}| \leq -\delta\right].$$

Since $t \to 0$, the second event on the right-hand-side is vacuous for large enough n. Also, for δ small enough and n large enough, the third event is vacuous as we have pointed out above. Hence, for such δ and such large n,

$$\mathrm{P}\{H \notin A\} \leq (1 + o(1))\frac{C\sqrt{\log(n)}}{\sqrt{6}\int |K - L|} n^{-5/2} + e^{-\gamma(\delta)n} < 1/n^2$$

for n large enough. The last statement of Lemma O2 follows from Theorem 6.1 of Devroye and Györfi (1985). □

Relative stability of the estimate.

Lemma O3. *Let f be an arbitrary density. Let K and L be smooth absolutely integrable kernels with $\int |K - L| > 0$, and let H minimize $\int |f_{nh} - g_{nh}|$. Assume also that the generalized characteristic functions of K and L do not coincide on any open neighborhood of the origin.*

$$\mathrm{P}\left\{\left|\int |f_{nH} - g_{nH}| - \mathrm{E}_n\left\{\int |\hat{f}_{nH} - \hat{g}_{nH}|\right\}\right| \geq \frac{\sqrt{128}\int |K - L|\sqrt{3\log(n)}}{\sqrt{n}}\right\} \leq \frac{2}{n^2}$$

for all n large enough. Also,

$$\begin{aligned}&\left|\mathrm{E}\left\{\int |f_{nH} - g_{nH}|\right\} - \mathrm{E}\left\{\int |f_{n\hat{H}} - g_{n\hat{H}}|\right\}\right| \\ &\leq \mathrm{E}\left\{\left|\int |f_{nH} - g_{nH}| - \mathrm{E}_n\int |\hat{f}_{nH} - \hat{g}_{nH}|\right|\right\} \\ &\leq \frac{\sqrt{129}\int |K - L|\sqrt{3\log(n)}}{\sqrt{n}}\end{aligned}$$

for all n large enough.

Proof. Define $t = \frac{\sqrt{128}\int |K-L|\sqrt{3\log(n)}}{\sqrt{n}}$. Let A be as in the proof of Lemma O2 for arbitrary $\epsilon > 0$. Define the random variable H_A as the projection to A of H. We have

$$\begin{aligned}&\mathrm{P}\left\{\left|\int |f_{nH} - g_{nH}| - \mathrm{E}_n\int |\hat{f}_{nH} - \hat{g}_{nH}|\right| \geq t\right\} \\ &\leq \mathrm{P}\left\{\left|\int |f_{nH_A} - g_{nH_A}| - \mathrm{E}_n\int |\hat{f}_{nH_A} - \hat{g}_{nH_A}|\right| \geq t\right\} + \mathrm{P}\{H \notin A\} \\ &< \frac{1}{n^2} + \mathrm{P}\{H \notin A\} < \frac{2}{n^2}\end{aligned}$$

for all n large enough, where we used estimates from Lemma O2. Also,

$$
\begin{aligned}
&\mathbf{E}\left\{\left|\int |f_{nH} - g_{nH}| - \mathbf{E}_n \int |\hat{f}_{nH} - \hat{g}_{nH}|\right|\right\} \\
&\leq t + \int |K-L| \mathbf{P}\left\{\left|\int |f_{nH} - g_{nH}| - \mathbf{E}_n \int |\hat{f}_{nH} - \hat{g}_{nH}|\right| \geq t\right\} \\
&\leq \frac{\sqrt{129}\int |K-L|\sqrt{3\log(n)}}{\sqrt{n}}
\end{aligned}
$$

for all n large enough. □

LEMMA O4. *Let f be an arbitrary density. Let K and L be smooth absolutely integrable kernels with $\int |K-L| > 0$, and let H minimize $\int |f_{nh} - g_{nh}|$. Assume also that the generalized characteristic functions of K and L do not coincide on any open neighborhood of the origin.*

$$
\mathbf{P}\left\{\left|\int |f_{nH} - f| - \mathbf{E}_n \int |\hat{f}_{nH} - f|\right| \geq \frac{\sqrt{128}\int |K|\sqrt{3\log(n)}}{\sqrt{n}}\right\} \leq \frac{2}{n^2}
$$

for all n large enough. Also,

$$
\begin{aligned}
&\left|\mathbf{E}\left\{\int |f_{nH} - f|\right\} - \mathbf{E}\left\{\int |f_{n\hat{H}} - f|\right\}\right| \leq \mathbf{E}\left\{\left|\int |f_{nH} - f| - \mathbf{E}_n \int |\hat{f}_{nH} - f|\right|\right\} \\
&\leq \frac{\sqrt{129}\int |K|\sqrt{3\log(n)}}{\sqrt{n}}
\end{aligned}
$$

for all n large enough.

PROOF. We mimick the proof of Lemma O1 first, replacing g_{nh} throughout by f, and $K-L$ by K. From this, we conclude that for arbitrary fixed $\epsilon, u > 0$,

$$
\begin{aligned}
&\mathbf{P}\left\{\sup_{\epsilon/n \leq h \leq 1/\epsilon}\left|\int |f_{nh} - f| - \mathbf{E}\int |f_{nh} - f|\right| \geq \frac{\sqrt{128}\int |K|\sqrt{(1+u)\log(n)}}{\sqrt{n}}\right\} \\
&\leq (1+o(1))\frac{C\sqrt{n\log(n)}}{\sqrt{2}\int |K|\sqrt{1+u}} n^{-(1+u)},
\end{aligned}
$$

where C is the smoothness constant for K (see definition of smoothness just before Lemma O1). Then turn to the proof of Lemma O3, replacing $K-L$ in the definition of t by K. Furthermore, replace all the references to g_{nH} and g_{nH_A} by f, and note that $\int |f_{nh} - f| \leq 1 + \int |K|$. This concludes the proof of Lemma O4. □

LEMMA O5. *Let f_{nh} be a kernel estimate with class s kernel K. Then there exists a constant $c = c(K) > 0$ such that for any f and for any sequence $h = h(n)$,*

$$
\liminf_{n\to\infty} n^{\frac{s}{2s+1}} \mathbf{E}\int |f_{nh} - f| \geq c > 0,
$$

The same bound is valid if f_{nh} is replaced by f_{nH}, where $H = H(n)$ is any sequence of positive random variables independent of the data sequence.

If H is obtained by minimizing $\int |f_{nh} - g_{nh}|$ and K and L are smooth absolutely integrable kernels with $\int |K - L| > 0$, such that the generalized characteristic functions of K and L do not coincide on any open neighborhood of the origin, then the asymptotic bound is also valid. In that case, we also have

$$\frac{\int |f_{nH} - f|}{\mathrm{E} \int |f_{nH} - f|} \to 1$$

almost surely as $n \to \infty$ for all densities f.

PROOF. The asymptotic bound for deterministic $h(n)$ is obtained in Devroye (1988). It is clear that for any sequence of random variables $\{H = H(n)\}$, that

$$\liminf_{n\to\infty} \frac{\mathrm{E} \int |f_{n\hat{H}} - f|}{\mathrm{E} \int |f_{nh^*} - f|} \geq 1,$$

where $h^* = h^*(n)$ is such that $\mathrm{E} \int |f_{nh^*} - f| \sim \inf_h \mathrm{E} \int |f_{nh} - f|$. Since the asymptotic lower bound is valid for f_{nh^*}, it must be valid for $f_{n\hat{H}}$.

Let H now be found by minimization as indicated in the statement of Lemma O5. Then, as we have seen in Lemma O4,

$$\mathrm{E} \int |f_{nH} - f| - \mathrm{E} \int |f_{n\hat{H}} - f| = O\left(\sqrt{\frac{\log n}{n}}\right).$$

Thus, the asymptotic bound also applies to $\mathrm{E} \int |f_{nH} - f|$. The last statement of the Lemma is obtained from the probability bound of Lemma O4, the asymptotic lower bound of Lemma O5, and the Borel-Cantelli lemma (the sequence $2/n^2$ is summable in n). $\square$

It is perhaps worthwhile to pause here to see what Lemma O5 implies for us. First of all, f_{nH} is strongly relatively stable, as shown in the last statement of the Lemma. Thus, the random variable $\int |f_{nH} - f|$ is very close to its mean. This is true for all densities f. For general theorems on the relative stability of f_{nH}, with arbitrary f, K and H, see, e.g., Devroye (1988). What this means for us is that $\int |f_{nH} - g_{nH}|$, a known quantity, is probably close to its mean, which, as we shall see below, is not too far away from $\mathrm{E} \int |f_{nH} - f|$. By relative stability again, the last quantity is about equal to $\int |f_{nH} - f|$. In other words, we have another useful by-product of the minimization, i.e., a rough estimate of the actual L_1 error $\int |f_{nH} - f|$.

Proof of Theorems O1 and O2. Let $h^* = h^*(n)$ be such that

$$\mathbf{E}\left\{\int |f_{nh^*} - g_{nh^*}|\right\} \sim \inf_h \mathbf{E}\left\{\int |f_{nh} - g_{nh}|\right\}.$$

We note the following:

$$\begin{aligned}
\mathbf{E}\left\{\int |f_{nH} - f|\right\} &\le \mathbf{E}\left\{\int |f_{nH} - g_{nH}|\right\} + \mathbf{E}\left\{\int |g_{nH} - f|\right\} \\
&= \mathbf{E}\left\{\inf_h \int |f_{nh} - g_{nh}|\right\} + \mathbf{E}\left\{\int |g_{nH} - f|\right\} \\
&\le (1+o(1))\mathbf{E}\left\{\int |f_{nh^*} - g_{nh^*}|\right\} + \mathbf{E}\left\{\int |g_{n\hat{H}} - f|\right\} + \sqrt{129}\int |K|\sqrt{\frac{3\log(n)}{\sqrt{n}}} \\
&\le (1+o(1))\mathbf{E}\left\{\int |f_{nh^*} - f|\right\} \\
&\quad + (1+o(1))\mathbf{E}\left\{\int |g_{nh^*} - f|\right\} + \mathbf{E}\left\{\int |g_{n\hat{H}} - f|\right\} \\
&\quad + \frac{\sqrt{129}\int |L|\sqrt{3\log(n)}}{\sqrt{n}}
\end{aligned}$$

for all n large enough (Lemma O4, applied to g_{nh}). So far, everything is valid for all densities.

Under the conditions of Theorem O1, with L as suggested in the statement of the Theorem, it is possible to show that (1) through (6) are satisfied with $c_1 = c_2 = 4\sqrt{\int L^2 / \int K^2}$ (Lemmas O6, O11 and O12):

(1)
$$\int |\mathbf{E}g_{nh^*} - f| = o\left(\int |\mathbf{E}f_{nh^*} - f|\right),$$

(2)
$$\mathbf{E}\left\{\int |g_{nh^*} - \mathbf{E}g_{nh^*}|\right\} \le (c_1 + o(1))\mathbf{E}\left\{\int |f_{nh^*} - f|\right\},$$

(3)
$$\mathbf{E}\left\{\int |g_{nh^*} - f|\right\} \le (c_1 + o(1))\mathbf{E}\left\{\int |f_{nh^*} - f|\right\},$$

(4)
$$\mathbf{E}\left\{\int |f * L_{\hat{H}} - f|\right\} = o\left(\mathbf{E}\left\{\int |f * K_{\hat{H}} - f|\right\}\right),$$

(5)
$$\mathbf{E}\left\{\int |g_{n\hat{H}} - f * L_{\hat{H}}|\right\} \le (c_2 + o(1))\mathbf{E}\left\{\int |f_{n\hat{H}} - f|\right\},$$

(6)
$$\le (c_2 + o(1))\mathbf{E}\left\{\int |f_{n\hat{H}} - f|\right\}.$$

We may conclude from (3) and (6) that

$$
\begin{aligned}
\mathrm{E}\left\{\int |f_{nH}-f|\right\} &\le (1+o(1))(1+c_1+o(1))\mathrm{E}\left\{\int |f_{nh^*}-f|\right\} \\
&\quad + (c_2+o(1))\mathrm{E}\left\{\int |f_{n\hat{H}}-f|\right\} + \frac{\sqrt{129}\int |L|\sqrt{3\log(n)}}{\sqrt{n}} \\
&\le (1+c_1+o(1))\mathrm{E}\left\{\int |f_{nh^*}-f|\right\} + (c_2+o(1))\mathrm{E}\left\{\int |f_{nH}-f|\right\} \\
&\quad + (c_2+o(1))\frac{\sqrt{129}\int |K|\sqrt{3\log(n)}}{\sqrt{n}} + \frac{\sqrt{129}\int |L|\sqrt{3\log(n)}}{\sqrt{n}}
\end{aligned}
$$

by Lemma O4. By Lemma O5, we see that the last two terms in the upper bound are asymptotically negligible with respect to the first term. Thus, we can conclude that

$$
\mathrm{E}\left\{\int |f_{nH}-f|\right\} \le \frac{1+c_1+o(1)}{1-c_2+o(1)}\mathrm{E}\left\{\int |f_{nh^*}-f|\right\}.
$$

The right hand side can be made smaller than $1+\epsilon+o(1)$ for any $\epsilon>0$ by the appropriate choice of L, since $c_1=c_2=4\sqrt{\int L^2/\int K^2}$. This concludes the proof of Theorem O1.

Recall the $n^{-s/(2s+1)}$ lower bound for $\mathrm{E}\left\{\int |f_{nH}-f|\right\}$ and $\inf_h \mathrm{E}\left\{\int |f_{nh}-f|\right\}$ (Lemma O5). Theorem O2 follows from this fact, (6), Theorem O1, and the fact that $\mathrm{E}\left\{\int |f_{nH}-f|\right\} - \mathrm{E}\left\{\int |f_{n\hat{H}}-f|\right\}$ is $O(\sqrt{\log n/n})$, and similarly for g_{nh} (Lemma 4). □

Remarks and further work. First we observe that the condition that $f\in\mathcal{F}_s$ is too strong. Theorem O1 holds for a much larger class of densities. It suffices to note that the crucial asymptotic result used in the proof of Lemma O6 remains valid, in the case $s=2$, $K\ge 0$, when f is such that it has a finite functional

$$
D(f) \stackrel{\text{def}}{=} \liminf_{a\downarrow 0} \int |(f*\phi_a)^{(2)}|,
$$

where ϕ is a mollifier, i.e., a kernel with $\int\phi=1$, $\phi\ge 0$, $\phi=0$ outside $[-1,1]$, and ϕ has infinitely many continuous derivatives on the real line. This functional coincides with $\int |f^{(2)}|$ when f and f' are absolutely continuous, and is well-defined (possibly ∞) and independent of the choice of ϕ for all f. For the proofs of this, see, e.g., Devroye (1987), pp. 108-111. To illustrate this, consider the triangular density. It does not have an absolutely continuous derivative, yet $D(f)<\infty$. For smooth regular symmetric nonnegative K with finite second moment, Theorem O1 is valid for all f in $\mathcal{W}$ or $\mathcal{V}$ for which $\int\sqrt{f}<\infty$ and $D(f)<\infty$.

It is possible to get asymptotic optimality for a proper subclass of $\mathcal{F}_s$ by choosing L in such a way that L varies with n by a scale factor only, i.e., L_h is replaced throughout by $L_{a_n h}$ where a_n tends very slowly to ∞ so as not to upset properties (1) and (4). This will allow us to formally take $c_1=c_2=0$, and obtain the asymptotic optimality. The proper subclass of $\mathcal{F}_s$ is determined by the rate of divergence of a_n. This will not be pursued any further here.

In Theorem O1, K is a class s kernel, so that the best possible rate of convergence is $n^{-s/(2s+1)}$ (Lemma O5). If we know that f is very smooth, then this could be an unwelcome restriction. One might wonder if there is nothing that can be said if we employ a class ∞ kernel. We have the following general result, which can be proved along the lines of the proof of Theorem 1, provided that Lemma O6 is replaced by a (trivial) counterpart.

THEOREM O3. *Let K and L be symmetric smooth regular kernels, with $\int L^2 < \int K^2/16$. Assume furthermore that the generalized characteristic functions of K and L do not coincide on any open neighborhood of the origin. (This implies that $\int |K-L| > 0$.) Let $f \in \mathcal{W}$ be such that $\int |f*L_h-f| / \int |f*K_h-f| \to 0$ as $h \downarrow 0$. Then, the kernel estimate in which H is defined by $\int |f_{nH} - g_{nH}| \sim \inf_h \int |f_{nh} - g_{nh}|$, satisfies the following inequality:*

$$\mathrm{E}\left\{\int |f_{nH} - f|\right\} \leq (1+o(1))\frac{1+u}{1-u}\inf_h \mathrm{E}\left\{\int |f_{nh} - f|\right\} + (1+o(1))\frac{u\int |K| + \int |L|}{1-u}\sqrt{\frac{387\log n}{n}},$$

where $u = 4\sqrt{\int L^2 / \int K^2}$. When $f \in \mathcal{V} \cap \mathcal{W}^c$, the same is true, provided that, additionally, L has compact support.

It is easy to see that the H obtained with the pair $(K, L) = (K, 2K - K * K)$ is indistinguishable from the H obtained by the pair $(K, K * K)$. This has an interesting interpretation: indeed, the kernel estimate f_{nh} can formally be considered as $\mu_n * K_h$ where μ_n is the standard empirical measure. With $L = K * K$, the estimate g_{nh} is nothing but $\mu_n * K_h * K_h = f_{nh} * K_h$. Minimizing $\int |f_{nh} - g_{nh}|$ is like asking that the operation $*K_h$ yields a stable point (doesn't change things too much); if h is really good, then $\mu_n * K_h$ should be close to f. But then applying the same operator again should not yield a very different curve, so $\mu_n * K_h * K_h$ should be close to $\mu_n * K_h$. So, what are the properties of the double kernel estimate with the pair $(K, K * K)$?

The final series of lemmas.

LEMMA O6. *Assume that $f \in \mathcal{F}_s$. Let K and L be smooth absolutely integrable kernels whose generalized characteristic functions do not coincide on any open neighborhood of the origin, and let K be a class s kernel. Then facts (1) and (4) are valid provided that L is picked such that L is symmetric, $\int L = 1$, $\int |L| < \infty$, $\int x^i L(x)dx = 0$ for all $0 < i \leq s$, and $\int |x|^s |L(x)|dx < \infty$.*

PROOF. We recall that $h^* \to 0$ as $n \to \infty$ (see the proof of Theorem C3 together with Lemmas C1 and C2). Under the conditions of Theorem O1, we have for $f \in \mathcal{F}_s$, as $h \downarrow 0$,

$$\int |\mathrm{E}f_{nh} - f| = \int |f * K_h - f| \sim h^s \left|\int \frac{x^s}{s!} K(x)dx\right| \int |f^{(s)}|$$

(see, e.g., Devroye and Györfi (1985, p. 209) or Devroye (1987, p. 110)). Also, if $\int |f^{(s)}| < \infty$, and if L is such that $\int L = 1$, $\int |L| < \infty$, $\int x^i L(x)dx = 0$ for all $0 < i \leq s$, and $\int |x|^s |L(x)|dx < \infty$, then, from Devroye (1987, p. 110) we retain that

$$\int |\mathrm{E}g_{nh} - f| = \int |f * L_h - f| = o(h^s).$$

This establishes (1). For the proof of (4), we note that $H \to 0$ in probability (from Theorem C3 and Lemma C1). Thus, if μ is the probability measure for H, and $F(h)$ and $G(h)$ denote the biases $\int |f * K_h - f|$ and $\int |f * L_h - f|$ respectively, then, for $\epsilon > 0$,

$$\begin{aligned}
\frac{\mathrm{E}\left\{\int |f * L_H - f|\right\}}{\mathrm{E}\left\{\int |f * K_H - f|\right\}} &= \frac{\int G(h)\mu(dh)}{\int F(h)\mu(dh)} \\
&\le \frac{\int_0^\epsilon G(h)\mu(dh) + (1 + \int |L|)\mathrm{P}\{H > \epsilon\}}{\int F(h)\mu(dh)} \\
&\le \sup_{h<\epsilon} \frac{G(h)}{F(h)} + \left(1 + \int |L|\right) \frac{\mathrm{P}\{H > \epsilon\}}{\mathrm{P}\{H \in [1/(n\epsilon), \epsilon)\} \inf_{1/(n\epsilon)\le h \le \epsilon} F(h)}.
\end{aligned}$$

The first term in the upper bound tends to zero as $\epsilon \downarrow 0$ (by (1)), while for fixed ϵ, the denominator of the second term is at least equal to a constant times n^{-s} (by Lemma O2 and an estimate for the bias used above). Its denominator is not greater than $n^{-(s+1)}$ (say) for n large enough by a suitable generalization of the bound of Lemma O4 (it suffices to replace the constant 3 in the definition of t there by a larger constant). This proves (4). □

The following lemma provides a uniform lower bound for the expected variation of any kernel estimate with a regular kernel.

LEMMA O7. *Let f_{nh} be the kernel estimate with regular kernel K. Then, for all f, and for all sequences a_n and b_n with $b_n \to 0$, $na_n \to \infty$, $a_n \le b_n$,*

$$\liminf_{n\to\infty} \inf_{a_n \le h \le b_n} \left\{\frac{4nh}{\int K^2}\right\}^{1/2} \mathrm{E}\left\{\int |f_{nh} - K_h * f|\right\} \ge \int \sqrt{f}.$$

PROOF. Let $h^* = h^*(n)$ be a sequence of positive numbers with $a_n \le h^*(n) \le b_n$ such that

$$\mathrm{E}\left\{\sqrt{nh^*}\int |f_{nh^*} - K_{h^*} * f|\right\} \sim \inf_{a_n \le h \le b_n} \mathrm{E}\left\{\sqrt{nh}\int |f_{nh} - K_h * f|\right\}.$$

Then, since $h^* \to 0$ and $nh^* \to \infty$, we know that

$$\liminf_{n\to\infty} \left\{\frac{4nh^*}{\int K^2}\right\}^{1/2} \mathrm{E}\left\{\int |f_{nh^*} - K_{h^*} * f|\right\} \ge \int \sqrt{f}.$$

(Devroye, 1987, Lemma 5). This proves Lemma O7. □

Lemma O8. *Let f be an arbitrary density. Let K be regular. Let $H = H(n)$ have an arbitrary sequence of distributions. Then, for any sequences $a_n \le b_n$ with $b_n \to 0$ and $na_n \to \infty$, and for any ϵ, we have for n large enough (where the definition of large enough does not depend upon the distribution of the sequence $\{H = H(n)\}$),*

$$\frac{\mathbb{E}\left\{\int |\hat{f}_{nH} - K_H * f|\right\}}{\mathbb{E}\left\{\frac{1}{\sqrt{nH} I_{a_n \le H \le b_n}}\right\}} \ge \begin{cases} \frac{1}{\epsilon} & \text{if } \int \sqrt{f} = \infty; \\ \frac{\sqrt{\int K^2} \int \sqrt{f}}{2-\epsilon} & \text{if } \int \sqrt{f} < \infty. \end{cases}$$

Proof. Let us write $\Delta(n,h)$ for $\mathbb{E}\left\{\int |f_{nh} - K_h * f|\right\}$. Then, since $\hat{f}$ and H are independent,

$$\begin{aligned} \mathbb{E}\left\{\int |\hat{f}_{nH} - K_H * f|\right\} &= \mathbb{E}\Delta(n,H) \ge \mathbb{E}\left\{\Delta(n,H) I_{a_n \le H \le b_n}\right\} \\ &\ge \mathbb{E}\left\{\frac{\sqrt{nH}\Delta(n,H) 1}{\sqrt{nH} I_{a_n \le H \le b_n}}\right\} \\ &\ge \inf_{a_n \le h \le b_n} \sqrt{nh}\, \Delta(n,h) \mathbb{E}\left\{\frac{1}{\sqrt{nH} I_{a_n \le H \le b_n}}\right\}. \end{aligned}$$

Now apply Lemma O7. □

Lemma O9. *Assume that $\int \sqrt{f} < \infty$. Let f_{nh} be the kernel estimate with regular kernel K. Then, for all sequences a_n and b_n with $b_n \to 0$, $na_n \to \infty$, $a_n \le b_n$,*

$$\limsup_{n \to \infty} \sup_{a_n \le h \le b_n} \left\{\frac{nh}{\int K^2}\right\}^{1/2} \mathbb{E}\left\{\int |f_{nh} - K_h * f|\right\} \le \int \sqrt{f},$$

when either $f \in \mathcal{W}$ and K has finite second moment, or when $f \in \mathcal{V}$ and K has compact support.

Proof. Note first that every regular kernel is square integrable. Now apply a standard bound that can be obtained directly via the Cauchy-Schwarz inequality:

$$\mathbb{E}\left\{\int |f_{nh} - f * K_h|\right\} \le \frac{\int \sqrt{f * (K^2)_h}}{\sqrt{nh}}$$

(see, e.g., Devroye, 1987, p. 113). This upper bound is

$$\frac{(1 + o(1))\sqrt{\int K^2} \int \sqrt{f}}{\sqrt{nh}}$$

when $h \to 0$ as $n \to \infty$ when both f and K^2 have finite absolute $1 + \epsilon$ moments for some $\epsilon > 0$ (see, e.g., exercise 7.8 of Devroye (1987)). The latter condition on K is satisfied if K has finite second moment and is regular. For f, the condition is implied when $f \in \mathcal{W}$. The same asymptotic result is valid if K has compact support and is bounded, and $f \in \mathcal{V}$ (Devroye and Györfi 1985, Lemma 5.26).

Let $h^* = h^*(n)$ be a sequence of positive numbers with $a_n \le h^*(n) \le b_n$ such that

$$\mathbf{E}\left\{\sqrt{nh^*}\int |f_{nh^*} - K_{h^*} * f|\right\} \sim \sup_{a_n \le h \le b_n} \mathbf{E}\left\{\sqrt{nh}\int |f_{nh} - K_h * f|\right\}.$$

Then, since $h^* \to 0$ and $nh^* \to \infty$, we know that

$$\limsup_{n\to\infty} \left\{\frac{nh^*}{\int K^2}\right\}^{1/2} \mathbf{E}\left\{\int |f_{nh^*} - K_{h^*} * f|\right\} \le \int \sqrt{f}.$$

This proves Lemma O9. □

Lemma O10. *Let f and K be as in Lemma O9. Then, for any sequences $a_n \le b_n$ with $b_n \to 0$ and $na_n \to \infty$, and for any definition of the random variables $H = H(n)$,*

$$\limsup_{n\to\infty} \frac{\mathbf{E}\left\{\int |\hat{f}_{nH} - K_H * f|\right\}}{\mathbf{E}\left\{\frac{1}{\sqrt{nH}} I_{a_n \le H \le b_n}\right\}} \le \sqrt{\int K^2}\int \sqrt{f} + \limsup_{n\to\infty} \frac{2\sqrt{nb_n}\int |K| \mathbf{P}\{H \notin [a_n, b_n]\}}{\mathbf{P}\{H \in [a_n, b_n]\}}.$$

Proof. Let us write $\Delta(n,h)$ for $\mathbf{E}\left\{\int |f_{nh} - K_h * f|\right\}$. Then, since $\hat{f}_{nh}$ and H are independent,

$$\begin{aligned}
\mathbf{E}\left\{\int |\hat{f}_{nH} - K_H * f|\right\} &= \mathbf{E}\Delta(n,H) \\
&\le \mathbf{E}\left\{\Delta(n,H) I_{a_n \le H \le b_n}\right\} + 2\int |K| \mathbf{P}\{H \notin [a_n,b_n]\} \\
&\le \mathbf{E}\left\{\sqrt{nH}\Delta(n,H)\frac{1}{\sqrt{nH}} I_{a_n \le H \le b_n}\right\} + 2\int |K| \mathbf{P}\{H \notin [a_n,b_n]\} \\
&\le \sup_{a_n \le h \le b_n} \sqrt{nh}\Delta(n,h)\mathbf{E}\left\{\frac{1}{\sqrt{nH}} I_{a_n \le H \le b_n}\right\} + 2\int |K| \mathbf{P}\{H \notin [a_n,b_n]\}.
\end{aligned}$$

Now apply Lemma O9. □

Lemma O11. *Let f_{nh},g_{nh} and H be as in Theorem O1 and assume that $\int \sqrt{f} < \infty$. Assume that K is a smooth regular kernel. Then (2) and (5) hold if we choose a smooth regular L in such a way that the generalized characteristic functions of K and L do not coincide on any open neighborhood of the origin, and that either $f \in \mathcal{W}$ and L has finite second moment, or $f \in \mathcal{V}$ and L has compact support.*

Also, we can take $c_1 = c_2 = 4\sqrt{\int L^2 / \int K^2}$.

Proof. By Lemmas C1 and C2, we have $h^* \to 0$ and $nh^* \to \infty$ when K and L are absolutely integrable kernels whose generalized characteristic functions do not coincide on any open neighborhood of the origin. Thus, from Lemma O7 applied to f_{nh} and K (which requires that K be regular) and Lemma O9 applied to g_{nh} and L (which requires that $\int \sqrt{f} < \infty$, that L be regular, and that either $f \in \mathcal{W}$ and L has finite second moment, or $f \in \mathcal{V}$ and L has compact support),

$$\begin{aligned}
\limsup_{n\to\infty} \sqrt{nh^*}\,\mathbf{E}\left\{\int |g_{nh^*} - \mathbf{E} g_{nh^*}|\right\} &\le \sqrt{\int L^2}\int\sqrt{f} \\
&= \frac{c_1}{2}\frac{1}{2}\sqrt{\int K^2}\int\sqrt{f} \\
&\le \liminf_{n\to\infty} \frac{c_1}{2}\sqrt{nh^*}\,\mathbf{E}\left\{\int |f_{nh^*} - \mathbf{E} f_{nh^*}|\right\} \\
&\le \liminf_{n\to\infty} c_1\sqrt{nh^*}\,\mathbf{E}\left\{\int |f_{nh^*} - f|\right\},
\end{aligned}$$

by the triangle inequality and Jensen's inequality, where $c_1 = 4\sqrt{\int L^2/\int K^2}$. We will see that we can take $c_2 = c_1$. This concludes the proof of (2).

To prove (5), let a_n and b_n be $3/n^2$ and $1-3/n^2$ quantiles of H respectively. We show first that $b_n \to 0$ and $na_n \to \infty$. Take $\epsilon > 0$ arbitrary. Assume for example that for an infinite subsequence, we have $b_n > \epsilon$. Then, on that subsequence,

$$\begin{aligned}
\mathbf{P}\{\epsilon \le H \le b_n\} &= \mathbf{P}\{H \ge b_n\} - \mathbf{P}\{H \ge \epsilon\} \\
&\ge \frac{3}{n^2} - \frac{1}{n^2} \text{ (Lemma O2, all } n \text{ large enough)} \\
&> \frac{1}{n^2},
\end{aligned}$$

which is a contradiction, since $\mathbf{P}\{H \ge \epsilon\} \le 1/n^2$ for n large enough. Hence $b_n < \epsilon$ for all n large enough, and, by symmetry, $na_n > 1/\epsilon$ for all n large enough. Thus, $b_n \to 0$ and $na_n \to \infty$ as required. Note that Lemma O2 required that K and L both be smooth absolutely integrable kernels whose generalized characteristic functions do not coincide on any open neighborhood of the origin.

Since $\int \sqrt{f} < \infty$ and K is regular, we can employ Lemma O8, and since $\int\sqrt{f} < \infty$, L is regular, and either $f \in \mathcal{W}$ and L has finite second moment, or $f \in \mathcal{V}$ and L has compact support, we can use Lemma O10 to conclude that

$$\begin{aligned}
\limsup_{n\to\infty} \frac{\mathbf{E}\left\{\int |\hat{g}_{nH} - L_H * f|\right\}}{\mathbf{E}\left\{\frac{1}{\sqrt{nH}} I_{a_n \le H \le b_n}\right\}} &\le \sqrt{\int L^2}\int\sqrt{f} + \limsup_{n\to\infty} \frac{2\sqrt{nb_n}\int |K| \mathbf{P}\{H \notin [a_n, b_n]\}}{\mathbf{P}\{H \in [a_n, b_n]\}} \\
&\le \sqrt{\int L^2}\int\sqrt{f} + \limsup_{n\to\infty} \frac{2o(\sqrt{n})\int|K|(6/n^2)}{1-6/n^2} \\
&= \frac{c_2}{2}\frac{\sqrt{\int K^2}\int\sqrt{f}}{2} \\
&\le \frac{c_2}{2}\liminf_{n\to\infty} \frac{\mathbf{E}\left\{\int |\hat{f}_{nH} - K_H * f|\right\}}{\mathbf{E}\left\{\frac{1}{\sqrt{nH}} I_{a_n \le H \le b_n}\right\}} \\
&\le c_2 \liminf_{n\to\infty} \frac{\mathbf{E}\left\{\int |\hat{f}_{nH} - f|\right\}}{\mathbf{E}\left\{\frac{1}{\sqrt{nH}} I_{a_n \le H \le b_n}\right\}}
\end{aligned}$$

where we once again used the fact that

$$\mathbb{E}_n\left\{\int |\hat{f}_{nH}-f|\right\} \geq \frac{1}{2}\mathbb{E}_n\left\{\int |\hat{f}_{nH}-\mathbb{E}_n\hat{f}_{nH}|\right\} = \frac{1}{2}\mathbb{E}_n\left\{\int |\hat{f}_{nH}-f*K_H|\right\}.$$

Fact (5) follows from the chain of inequalities derived above and the observation that for sequences of positive numbers u_n, v_n, w_n,

$$\limsup u_n/v_n \leq \frac{\limsup u_n/w_n}{\liminf v_n/w_n}.\square$$

Lemma O12. *In the proof of Theorem O1, (1) and (2) together imply (3), and (4) and (5) together imply (6).*

Proof. We will use the facts that $\int |\mathbb{E} f_{nh^*}-f| \leq \mathbb{E}\left\{\int |f_{nh^*}-f|\right\}$ and that

$$\mathbb{E}\left\{\int |f*K_{\hat{H}}-f|\right\} = \mathbb{E}\left\{\int |\hat{\mathbb{E}}_n f_{n\hat{H}}-f|\right\} \leq \mathbb{E}\left\{\hat{\mathbb{E}}_n\left\{\int |f_{n\hat{H}}-f|\right\}\right\} = \mathbb{E}\left\{\int |f_{n\hat{H}}-f|\right\}.$$

Now, $\mathbb{E}\left\{\int |g_{nh^*}-f|\right\}$ does not exceed the sum of the left-hand-sides of (1) and (2), from which the claim about (3) follows. Similarly, $\mathbb{E}\left\{\int |g_{n\hat{H}}-f|\right\}$ does not exceed the sum of the left-hand-sides of (4) and (5), from which the claim about (5) follows. $\square$

Section 4: PROPERTIES OF THE OPTIMAL SMOOTHING FACTOR

Theorem S1. *Let f be an arbitrary density. Let f_{nh} be a kernel estimate with smooth absolutely integrable class s kernel K. (Note: its characteristic function does not coincide with 1 on any open neighborhood of the origin.) Let $H = H(n)$ be any sequence of random variables for which*

$$\int |f_{nH}-f| \sim \inf_h \int |f_{nh}-f|.$$

Then

$$\lim_{n\to\infty} \frac{\mathbb{E}\left\{\int |f_{nH}-f|\right\}}{\inf_h \mathbb{E}\left\{\int |f_{nh}-f|\right\}} = 1,$$

and

$$\lim_{n\to\infty} \frac{\int |f_{nH}-f|}{\mathbb{E}\left\{\int |f_{nH}-f|\right\}} = 1$$

almost surely.

Theorem S1 reassures us that it is irrelevant whether we study $\inf_h \mathbb{E}\left\{\int |f_{nh}-f|\right\}$ or $\mathbb{E}\left\{\inf_h \int |f_{nh}-f|\right\}$, since both are asymptotically equal for all densities f when K is a class s kernel. The study of the previous section was with respect to the former quantity. We now see that in the definition of C-optimality, it would have been possible to replace the denominator by $\mathbb{E}\left\{\inf_h \int |f_{nh}-f|\right\}$.

Lemma S1. *Let f be an arbitrary density. Let K be an absolutely integrable kernel whose characteristic function does not coincide with 1 on any open neighborhood of the origin. Then, for all $\epsilon > 0$,*

$$\liminf_{n\to\infty} \inf_{h \notin [1/(n\epsilon),\epsilon]} \mathrm{E}\left\{ \int |f_{nh} - f| \right\} > 0.$$

Proof. The first statement is obtained by mimicking the proofs of Lemmas C1 and C2. In Lemma C1, it suffices to replace $f * L_h$ throughout by f (which formally corresponds to taking L with characteristic function identical to one). Hence the need to introduce the condition that K not coincide with 1 on any open neighborhood of the origin. Lemma C2 remains valid with little change, provided that M in that proof is replaced by K. It is necessary there to reverify that

$$\liminf_{n\to\infty} \inf_{0<h\le d/n} \mathrm{E}\left\{ \int |m_{nh} - \mathrm{E}m_{nh}| \right\} > 0,$$

where m_{nh} is in the notation of Lemma C2; it is the f_{nh} of the present Lemma. Still in the notation of Lemma C2, we have

$$\inf_{h\le d/n} \mathrm{E}\left\{ \int |m_{nh} - \mathrm{E}m_{nh}| \right\} \ge \inf_{h\le d/n} \mathrm{E}\left\{ \frac{pN}{n} - \frac{q(n-N)}{n} - \int |\mathrm{E}m_{nh}| - \mathrm{E}\int_B |\mathrm{E}m_{nh}| \right\}$$

where B is the collection of all sets $[X_i - ch, X_i + ch]$ that capture no data point besides X_i. Let us consider the last term. It is surely not greater than $\sup_{h\le d/n}(\int |\mathrm{E}m_{nh} - f| + \mathrm{E}\int_B f)$. This is $o(1) + \sup_{h\le d/n} \mathrm{E}\int_B f$ by the convergence of the bias (Theorem 2.1 of Devroye and Györfi, 1985). The last term is in turn not larger than $\mathrm{E}\left\{\omega(\lambda(B'))\right\}$ where B' is defined as B, but with the interval lengths $2ch$ replaced by the larger values $2cd/n$, λ is Lebesgue measure, and $\omega(u) \stackrel{\text{def}}{=} \sup_{A:\lambda(A)\le u} \int_A f$ (which $\to 0$ as $u \downarrow 0$). We can bound the term by $2cdN/n \le 2cd$. Combined with the lower bounds of Lemma C2, we can conclude that

$$\inf_{h\le d/n} \mathrm{E}\left\{ \int |m_{nh} - \mathrm{E}m_{nh}| \right\} \ge \int f e^{-2cdf} \int |M| - q - \omega(2cd) > 0,$$

for n large enough, c large enough and d small enough. The second part of the proof of Lemma C2 requires no modifying. □

Lemma S2. *Let f be an arbitrary density. Let K be an absolutely integrable kernel. For fixed $u > 0$,*

$$\mathrm{P}\left\{ \sup_h \left| \int |f_{nh} - f| - \mathrm{E}\left\{ \int |f_{nh} - f| \right\} \right| > u \right\} \le e^{-\gamma n}$$

where $\gamma = \gamma(u) > 0$. As a consequence, with $J_{nh} \stackrel{\text{def}}{=} \int |f_{nh} - f|$, we have the following:

A. $\sup_h > 0|J_{nh} - \mathrm{E}J_{nh}| \to 0$ *almost surely as $n \to \infty$.*

B. *For any random variable H (possibly not independent of the data), $J_{nH} - \mathrm{E}_n\hat{J}_{nH} \to 0$ almost surely as $n \to \infty$.*

C. *For any random variable H (possibly not independent of the data), $J_{nH} \to 0$ in probability implies $\mathrm{E}J_{nH} \to 0$, $\mathrm{E}_n\hat{J}_{nH} \to 0$ in probability, $\mathrm{E}\hat{J}_{nH} \to 0$ and $\mathrm{E}J_{n\hat{H}} \to 0$.*

PROOF. We extend the proof of Theorem C2. Note first that $|f_{nh} - u_{nh}| \leq \int |K - K'|$ when u_{nh} is the kernel estimate with kernel K'. The fact that the bound does not depend upon h and that K' is arbitrary means that we need only show the Theorem for all K that are continuous and of compact support (since the latter collection is dense in the space of L_1 functions).

The following inequality is valid for all fixed h,n,K and f:

$$\mathrm{P}\left\{\left|\int |f_{nh} - f| - \mathrm{E}\left\{\int |f_{nh} - f|\right\}\right| > \epsilon\right\} \leq 2e^{-\frac{n\epsilon^2}{32\int^2 |K|}}$$

(Devroye, 1988). We employ the grid technique of Theorem C2 again. Set $\Delta(h) = |\int |f_{nh} - f| - \mathrm{E}\left\{\int |f_{nh} - f|\right\}|$, and $\Delta(a,b) = \sup_{h,h' \in [a,b]} |\Delta(h) - \Delta(h')|$. Then let $c > 1$ be such that

$$\sup_{1 \leq h \leq c} \int |K_1 - K_h| \leq \epsilon/4.$$

Noting that

$$\left|\int |f_{nh} - f| - \int |f_{nh'} - f|\right\} \leq \int |f_{nh} - f_{nh'}| \leq \int |K_h - K'_h|,$$

we see that as in the proof of Theorem C2,

$$\mathrm{P}\left\{\sup_{a \leq h \leq b} \Delta(h) > \epsilon\right\} \leq \sum_{i=0}^{k} \mathrm{P}\left\{\Delta(ac^i) > \epsilon/2\right\} \leq 2\left(1 + \frac{\log(b/a)}{\log c}\right) e^{-\frac{n\epsilon^2}{128\int^2 |K|}}.$$

It suffices to have limits a and b that are such that $b/a = O(n)$, for this upper bound to tend to zero with n at an exponential rate. We need only establish that for some sequences $a = a(n)$ and $b = b(n)$ with $b/a = O(n)$ that

$$\lim_{n\to\infty} \left\{\mathrm{P}\left(\sup_{h<a} \Delta(h) > \epsilon\right\} + \mathrm{P}\left\{\sup_{h>b} \Delta(h) > \epsilon\right\}\right) = 0.$$

Assume that K vanishes off $[-s,s]$. Take $a = s\delta/2n$ where $\delta > 0$ is a constant to be picked further on. Let N be the number of X_i's for which $[X_i - 2a, X_i + 2a]$ has at least one X_j with $j \neq i$, and let A be the union of the sets $[X_i - a, X_i + a]$ for those X_i not counted in N. Note that $1 \geq \int_A |f_{nh}| / \int |K| \geq \frac{n-N}{n}$, uniformly over $h \leq a$. We have

$$\mathrm{P}\left\{\sup_{h<a} \Delta(h) > \epsilon\right\} \leq \mathrm{P}\left\{\sup_{h<a}\left|\int |f_{nh} - f| - (1 + \int |K|)\right\} > \epsilon/2\right\}$$
$$+ \mathrm{P}\left\{\sup_{h<a}\left|\mathrm{E}\left\{\int |f_{nh} - f|\right\} - (1 + \int |K|)\right\} > \epsilon/2\right\}.$$

We claim that

$$\mathrm{P}\left\{\sup_{h<a}\left|\int |f_{nh} - f| - (1 + \int |K|)\right\} \geq \epsilon/2\right\} \leq \mathrm{P}\left\{\sup_{h<a} \int_A |f_{nh}| \leq \int |K| - \epsilon/6\right\}$$

if $\delta < \rho(f,\epsilon)$ for some positive-valued function ρ. Indeed, since $\int |f_{nh} - f| \leq 1 + \int |K|$, it suffices to consider only one kind of signed difference. Take δ so small that uniformly over all sets B with $\lambda(B) < s\delta$,

$\int_B f < \epsilon/12$ where λ is Lebesgue measure (this is always possible). Note in passing that $\lambda(A) \le s\delta$. It suffices to show that $\sup_{B:\lambda(B)<s\delta} \int_B |f_{nh}| \ge \int |K| - \epsilon/6$ implies that $\int |f_{nh} - f| \ge 1 + \int |K| - \epsilon/2$. We have $\int_B |f_{nh} - f| \ge \int |K| - \epsilon/6 - \int_B f$, $\int_{B^c} |f_{nh} - f| \ge \int_{B^c} f - \int_{B^c} |f_{nh}|$, which is at least $1 - \int_B f - \epsilon/6$. Summing this and noting that $\int_B f \le \epsilon/12$ shows that $\int |f_{nh} - f| \ge 1 + \int |K| - \epsilon/2$.

The previous facts can now be combined to conclude that for δ small enough,

$$\mathrm{P}\left\{\sup_{h<a} \Delta(h) > \epsilon\right\} \le \mathrm{P}\left\{\frac{N}{n} > \frac{\epsilon}{6\int |K|}\right\} + \mathrm{P}\left\{\frac{\mathrm{E}N}{n} > \frac{\epsilon}{6\int |K|}\right\}.$$

This is handled precisely as in the proof of Theorem C2. Thus, $\lim_{n\to\infty} \mathrm{P}\left\{\sup_{h<a} \Delta(h) > \epsilon\right\} = 0$ for $a = s\delta/(2n)$ and δ small enough. In fact, the said probability does not exceed e^{-dn} for some constant $d > 0$ depending upon ϵ.

We finally proceed to show that

$$\mathrm{P}\left\{\sup_{h>b} \Delta(h) > \epsilon\right\}$$

can be made exponentially small in n by choosing a large enough constant b. This would conclude the proof of the Theorem since $b/a = O(n)$ as required. Let ω be the modulus of continuity of K defined by $\omega(u) = \sup_x \sup_{|y|\le u} |K(x) - K(x+y)|$. By our assumptions on K, $\omega(u) \to 0$ as $u \downarrow 0$. Let t and $T > t$ be positive numbers chosen in such a way that $\int_{t\le|x|\le T} |K| \ge \int |K| - \epsilon/8$ and $\sup_z \int_{z-t}^{z+t} |K| \le \epsilon/8$. Also, T should be so large that $\int_{|x|\ge T} f < \epsilon/(12\int |K|)$. This fixes t and T once and for all. Let N be the number of X_i's with $|X_i| \ge T$. We have the following inequality:

$$\int_{th\le|x|\le Th} |f_{nh} - K_h| \le (T-t)\omega(T/b) + \frac{N}{n}\int |K|.$$

This can best be seen by noting that

$$\begin{aligned}
|f_{nh} - K_h| &= \left|\frac{1}{n}\sum_{i=1}^{n}(K_h(x - X_i) - K_h(x))\right\} \\
&\le \frac{1}{n}\sum_{i:|X_i|\le T} \sup_{|y|\le T} |K_h(x-y) - K_h(x)| + \frac{1}{n}\sum_{i:|X_i|>T} |K_h(x - X_i)| \\
&\le \frac{1}{h}\omega(T/h) + \frac{1}{n}\sum_{i:|X_i|>T} |K_h(x - X_i)|.
\end{aligned}$$

Now, integrating over the given interval and noting that $h \ge b$ yields the result. For $h \ge b$,

$$\begin{aligned}
&\int |f_{nh} - f| \\
&\ge \int_{th\le|x|\le Th} |f_{nh}| - \int_{th\le|x|\le Th} f + \int_{|x|\le th} f - \int_{|x|\le th} |f_{nh}| \\
&\ge \int_{th\le|x|\le Th} |K_h| - \int_{th\le|x|\le Th} |f_{nh} - K_h| - \int_{th\le|x|\le Th} f + \int_{|x|\le th} f - \frac{1}{n}\sum_{i=1}^{n}\int_{|x|\le th} |K_h(x - X_i)| \\
&\ge \int |K| - \frac{\epsilon}{8} - (T-t)\omega(T/h) - \frac{\int |K| N}{n} + 1 - 2\int_{th\le|x|} f - \sup_z \int_z -t^z + t|K| \\
&\ge \int |K| - \frac{\epsilon}{8} - (T-t)\omega(T/b) - \frac{\int |K| N}{n} + 1 - 2\int_{tb\le|x|} f - \frac{\epsilon}{8}
\end{aligned}$$

$$\geq \int |K| + 1 - \frac{\epsilon}{3} - \frac{\int |K| N}{n}$$

if b is so large that $(T-t)\omega(T/b) + 2\int_{tb \leq |x|} f \leq \epsilon/12$. Thus,

$$\mathbf{P}\left\{\inf_{h \geq b} \int |f_{nh} - f| \leq 1 + \int |K| - \epsilon/2\right\} \leq \mathbf{P}\left\{\frac{N}{n} \geq \frac{\epsilon}{6\int |K|}\right\}$$

and

$$\inf_{h \geq b} \mathbf{E}\left\{\int |f_{nh} - f|\right\} \geq 1 + \int |K| - \frac{\epsilon}{2}$$

when $\mathbf{E}N/n \leq \epsilon/(6\int |K|)$. Now, $\mathbf{E}N/n = \int_{|x| \geq T} f < \epsilon/(12\int |K|)$ by our choice of T. By Hoeffding's inequality (Hoeffding, 1963),

$$\mathbf{P}\left\{\frac{N}{n} \geq \frac{\epsilon}{6\int |K|}\right\} \leq \mathbf{P}\left\{\frac{N - \mathbf{E}N}{n} \geq \frac{\epsilon}{12\int |K|}\right\} \leq e^{-\frac{2n\epsilon^2}{144\int^2 |K|}}.$$

In conclusion, for our choice of t, T and b,

$$\begin{aligned}
&\mathbf{P}\left\{\sup_{h \geq b} \Delta(h) > \epsilon\right\} \\
&\leq \mathbf{P}\left\{\sup_{h \geq b} \left| \int |f_{nh} - f| - (1 + \int |K|)\right| > \epsilon/2\right\} + \mathbf{P}\left\{\sup_{h \geq b} \left|\mathbf{E}\left\{\int |f_{nh} - f|\right\} - (1 + \int |K|)\right| > \epsilon/2\right\} \\
&\leq e^{-\frac{2n\epsilon^2}{144\int^2 |K|}}.
\end{aligned}$$

This concludes the proof of Lemma S2. □

Proof. Let us first try to prove that for arbitrary fixed $\epsilon > 0$,

$$\mathbf{P}\{H \notin [1/(n\epsilon), \epsilon]\} < 1/n^2$$

for all n large enough, and that $H \to 0$ and $nH \to \infty$ completely. This statement parallels that of Lemma O2. It suffices to replace g_{nh} throughout by f and $K - L$ by K. Also, the constant C now becomes the smoothness constant for K. It is easy to see then that we need only two facts at this stage:

$$\liminf_{n \to \infty} \inf_{h \notin [1/(n\epsilon), \epsilon]} \mathbf{E}\left\{\int |f_{nh} - f|\right\} > 0$$

and for fixed $u > 0$,

$$\mathbf{P}\left\{\sup_h \left|\int |f_{nh} - f| - \mathbf{E}\left\{\int |f_{nh} - f|\right\}\right| > u\right\} \leq e^{-\gamma n}$$

where $\gamma = \gamma(u) > 0$. These were proved in Lemmas S1 and S2.

We note now that the inequalities of Lemma O4 apply without change to the present H. In particular,

$$\mathbf{E}\left\{\int |f_{nH} - f|\right\} - \mathbf{E}\left\{\int |f_{n\hat{H}} - f|\right\} = O\left(\sqrt{\frac{\log n}{n}}\right),$$

where $\hat{H}$ is distributed as H but independent of the data stream. From Lemma O5 we retain that for any f

$$\liminf_{n\to\infty} n^{\frac{s}{2s+1}} \mathbf{E}\left\{\int |f_{n\hat{H}} - f|\right\} \geq c > 0,$$

for some constant $c > 0$. Hence, this bound also applies if $\hat{H}$ is replaced by H. In fact, we have

$$\lim_{n\to\infty} \frac{\mathbf{E}\left\{\int |f_{nH} - f|\right\}}{\mathbf{E}\left\{\int |f_{n\hat{H}} - f|\right\}} = 1$$

for all densities f. Thus,

$$\mathbf{E}\left\{\int |f_{nH} - f|\right\} \sim \mathbf{E}\left\{\int |f_{n\hat{H}} - f|\right\} \leq \inf_h \mathbf{E}\left\{\int |f_{nh} - f|\right\} \leq \mathbf{E}\left\{\int |f_{nH} - f|\right\},$$

which shows the first part of the Theorem. The strong convergence is obtained from the probability bound of Lemma O4 generalized above, the asymptotic lower bound of Lemma O5 (also generalized above), and the Borel-Cantelli lemma (the sequence $2/n^2$ is summable in n). $\square$

Section 5: ACKNOWLEDGMENTS

This research was sponsored by NSERC Grant A3456 and FCAC Grant EQ-1678.

Section 6: REFERENCES

S. Abou-Jaoude, "La convergence L1 et L infini de certains estimateurs d'une densité de probabilité," Thèse de Doctorat d'Etat, Université de Paris VI, France, 1977.

M. S. Bartlett, "Statistical estimation of density functions," *Sankhya Series A*, vol. 25, pp. 245–254, 1963.

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

J. Bretagnolle and C. Huber, "Estimation des densités: risque minimax," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 47, pp. 119–137, 1979.

M. Broniatowski, P. Deheuvels, and L. Devroye, "On the relationship between stability of extreme order statistics and convergence of the maximum likelihood kernel density estimate," *Annals of Statistics*, vol. 0, pp. 0–0, 1988. To appear..

P. Burman, "A data dependent approach to density estimation," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 69, pp. 609–628, 1985.

Y. S. Chow, S. Geman, and L. D. Wu, "Consistent cross-validated density estimation," *Annals of Statistics*, vol. 11, pp. 25–38, 1983.

L. Devroye, "The equivalence of weak, strong and complete convergence in L1 for kernel density estimates," *Annals of Statistics*, vol. 11, pp. 896–904, 1983.

L. Devroye and C. S. Penrod, “Distribution-free lower bounds in density estimation,” *Annals of Statistics*, vol. 12, pp. 1250–1262, 1984.

L. Devroye and L. Györfi, *Nonparametric Density Estimation: The L1 View*, John Wiley, New York, 1985.

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

L. Devroye, “The kernel estimate is relatively stable,” *Probability Theory and Related Fields*, vol. 77, pp. 521–536, 1988.

L. Devroye, “A universal lower bound for the kernel estimate,” *Statistics and Probability Letters*, vol. 0, pp. 0–0, 1988.

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

L. Devroye, *On the non-consistency of the $Lsub2$ cross-validated kernel density estimate*, 1988. Submitted..

R. P. W. Duin, “On the choice of smoothing parameters for Parzen estimators of probability density functions,” *IEEE Transactions on Computers*, vol. C-25, pp. 1175–1179, 1976.

V. A. Epanechnikov, “Nonparametric estimation of a multivariate probability density,” *Theory of Probability and its Applications*, vol. 14, pp. 153–158, 1969.

T. Gasser, H.-G. Müller, and V. Mammitzsch, “Kernels for nonparametric curve estimation,” *Journal of the Royal Statistical Society, Series B*, vol. 47, pp. 238–252, 1985.

J. D. F. Habbema, J. Hermans, and K. Vandenbroek, “A stepwise discriminant analysis program using density estimation,” in: *COMPSTAT 1974*, (edited by G. Bruckmann), pp. 101–110, Physica Verlag, Wien, 1974.

P. Hall, “Cross-validation in density estimation,” *Biometrika*, vol. 69, pp. 383–390, 1982.

P. Hall, “Large-sample optimality of least squares cross-validation in density estimation,” *Annals of Statistics*, vol. 11, pp. 1156–1174, 1983.

P. Hall, “Asymptotic theory of minimum integrated square error for multivariate density estimation,” in: *Multivariate Analysis VI*, (edited by P. R. Krishnaiah), pp. 289–309, North-Holland, Amsterdam, 1985.

P. Hall and M. P. Wand, “Minimizing L1 distance in nonparametric density estimation,” Technical Report, Department of Statistics, Australian National University, 1987.

P. Hall and J. S. Marron, “Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation,” *Probability Theory and Related Fields*, vol. 74, pp. 567–581, 1987.

G. H. Hardy and W. W. Rogosinski, *Fourier Series*, Cambridge University Press, 1962.

W. Hoeffding, “Probability inequalities for sums of bounded random variables,” *Journal of the American Statistical Association*, vol. 58, pp. 13–30, 1963.

J. S. Marron, "An asymptotically efficient solution to the bandwidth problem of kernel density estimation," *Annals of Statistics*, vol. 13, pp. 1011–1023, 1985.

H.-G. Müller, "Smooth optimum kernel estimators of densities, regression curves and modes," *Annals of Statistics*, vol. 12, pp. 766–774, 1984.

E. A. Nadaraya, "On the integral mean square error of some nonparametric estimates for the density function," *Theory of Probability and its Applications*, vol. 19, pp. 133–141, 1974.

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

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

M. Rosenblatt, "Global measures of deviation for kernel and nearest neighbor density estimates," in: *Proceedings of the Heidelberg Workshop*, pp. 181–190, Springer Lecture Notes in Mathematics 757, Springer-Verlag, Berlin, 1979.

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

E. F. Schuster and G. G. Gregory, "On the nonconsistency of maximum likelihood nonparametric density estimators," in: *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, (edited by W. F. Eddy), pp. 295–298, Springer Verlag, New York, N.Y., 1981.

D. W. Scott and G. R. Terrell, "Biased and unbiased cross-validation in density estimation," *Journal of the American Statistical Association*, vol. 82, pp. 1131–1146, 1987.

R. S. Singh, "Mean squared errors of estimates of a density and its derivatives," *Biometrika*, vol. 66, pp. 177–180, 1979.

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

W. Stuetzle and Y. Mittal, "Some comments on the asymptotic behavior of robust smoothers," in: *Proceedings of the Heidelberg Workshop*, (edited by T. Gasser and M. Rosenblatt), pp. 191–195, Springer Lecture Notes in Mathematics 757, Springer-Verlag, Heidelberg, 1979.

H. Y. Su-Wong, B. Prasad, and R. S. Singh, "A comparison between two kernel estimators of a probability density function and its derivatives," *Scandinavian Actuarial Journal*, vol. 0, pp. 216–222, 1982.

S. J. Szarek, "On the best constants in the Khintchine inequality," *Studia Mathematica*, vol. 63, pp. 197–208, 1976.

G. S. Watson and M. R. Leadbetter, "On the estimation of the probability density," *Annals of Mathematical Statistics*, vol. 34, pp. 480–491, 1963.

R. L. Wheeden and A. Zygmund, *Measure and Integral*, Marcel Dekker, New York, 1977.

M. Woodroofe, “On choosing a delta sequence,” *Annals of Mathematical Statistics*, vol. 41, pp. 1665–1671, 1970.