The Annals of Statistics
1992, Vol. 20, No. 3, 1306–1328

OPTIMAL PLUG-IN ESTIMATORS FOR NONPARAMETRIC FUNCTIONAL ESTIMATION

BY LARRY GOLDSTEIN[1] AND KAREN MESSER[2]

University of Southern California and
California State University, Fullerton

Consider the problem of estimating the value of a functional $\Lambda(f)$ for f an unknown density or regression function. The straightforward plug-in estimator $\Lambda(\hat{f})$ with $\hat{f}$ a particular estimate of f achieves the optimal rate of convergence in the sense of Stone over bounded subsets of a Sobolev space for a broad class of linear and nonlinear functionals. For many functionals the rate calculation depends on a Fréchet-like derivative of the functional, which may be obtained using elementary calculus. For some classes of functionals, $\hat{f}$ is undersmoothed relative to what would be used to estimate f optimally. Examples for which a plug-in estimator is optimal include L^q norms of regression or density functions and their derivatives and the expected integrated squared bias.

When interested in computing estimates over classes of functions which satisfy certain restrictions, such as strict positivity or boundary conditions, the plug-in estimator may or may not be optimal, depending on the functional and the function class. The functional calculus establishes conditions under which the plug-in estimator remains optimal, and sometimes suggests an appropriate modification when it does not.

1. Introduction. In a regression or density estimation setting, one is often interested in the value of a functional Λ of an unknown function f. In the regression context, one observes n independent and identically distributed copies of the pair of random variables (X, Y) with $Y = f(x) + \varepsilon$, where ε is a mean zero error term; in the density estimation framework, one observes $X_1, \ldots, X_n$, independent and identically distributed random variables with density f. In both instances f is assumed to satisfy some smoothness restrictions, but is otherwise unknown. We assume the X's lie in $[0, 1]$. From the observations one wishes to form an estimate $\hat{\Lambda}(f)$ of the value $\Lambda(f)$.

For example, one may wish to estimate the value of f or a derivative of f at a point x_0; in this case

$$\Lambda(f) = f^{(r)}(x_0) \tag{1.1}$$

for some r. Other examples include the mean-squared error of an estimate

Received September 1990; revised December 1991.
[1]Research partially supported by NSF Grant DMS-90-05833.
[2]Research partially supported by NSF Grant DMS-87-08083.
AMS 1980 *subject classifications*. Primary 62G05; secondary 62G20.
Key words and phrases. Nonparametric regression, functionals, optimal rates, plug-in estimators.

$\hat{f}^{(r)}(x)$ of $f^{(r)}(x)$:

$$\Lambda(f) = E\left[\left(\hat{f}^{(r)}(x) - f^{(r)}(x)\right)^2\right],$$

integrated mean-squared error:

$$\Lambda(f) = E\left[\int\left(\hat{f}^{(r)} - f^{(r)}\right)^2\right]$$

or the L^2 norm of the unknown function f or of the rth derivative of f:

$$\Lambda(f) = \left\{\int\left(f^{(r)}\right)^2\right\}^{1/2}. \tag{1.2}$$

The last two functionals are of interest for the choice of smoothing parameter, and the L^2 norm is of additional importance in the density estimation setting; see Example 2. For nonparametric tests of independence and equidistribution, Abramson and Goldstein (1991) study the equidistribution functional of a pair of densities f and g:

$$\Delta[f, g] = 2\int \frac{fg}{f+g}.$$

This paper addresses two questions: What is the theoretical upper bound on the rate of convergence of an estimate Λ to $\Lambda(f)$; and, Is there a convenient class of estimators that achieve this upper bound?

The estimator $\hat{\Lambda}$ we consider is the plug-in estimator $\hat{\Lambda}(f) = \Lambda(\hat{f})$, where $\hat{f}$ is a specific kernel estimator of f. In the regression setting we use a Nadaraya–Watson-type construction:

$$\hat{f}(x) = \frac{\sum Y_i K_b(x, X_i)}{\sum K_b(x, X_i)}, \tag{1.3}$$

where $K_b(x, t)$, given in the Appendix, is a certain boundary-corrected kernel which satisfies a scaling property in b. The kernel is studied in Messer and Goldstein (1992).

We take our definition of optimality from Stone (1980), and consider optimality over a class of functions $\mathscr{W} \subseteq \mathscr{W}_p$, where $\mathscr{W}_p$ is roughly the class of functions with p continuous derivatives all of which are bounded by a constant. We obtain optimality results over subsets $\mathscr{W}$ of $\mathscr{W}_p$ which may satisfy extra conditions, for example, classes of functions which satisfy boundary conditions, which are bounded away from 0, or which integrate to 1.

Consider the contrast between the functionals (1.1) and (1.2). For the point evaluation functional (1.1), Stone (1980) has established the well-known optimal rate of convergence over Sobolev classes of functions: Roughly, the best rate of convergence in probability of an estimator of $f^{(r)}(x_0)$ which is uniform over functions $f \in \mathscr{W}_p$ is $n^{-(p-r)/(2p+1)}$. On the other hand, for a "smooth" functional such as $\int f^2$, it is well known that the parametric rate of $n^{-1/2}$ is achievable. For the functional (1.2), the same rate $n^{-1/2}$ is achievable under certain circumstances.

This paper provides an easy method to divide functionals into two groups (a slow "pointwise" group, and a fast "smooth" group), presents a convenient plug-in kernel estimator which will achieve the optimal rate and determines the order of the optimal bandwidth for that estimator. Our method uses a straightforward Taylor series expansion of the functional Λ about f over an appropriate Sobolev space, using a Fréchet-like derivative which we denote T_f. This expansion is similar to the von Mises (1947) type expansions for functionals of a distribution function which are studied in Fernholz (1983) and Pfanzagl (1985).

For a class of functionals which we call *atomic*, the linear term has a point evaluation component, of which (1.1) is the canonical example. Determining the *index* of the functional [the number r for functional (1.1)] then determines the optimal rate $n^{-(p-r)/(2p+1)}$ achievable over $\mathscr{W}$. In some settings the index may depend on the boundary behavior of f. These results are given in Theorem 4.1.

For another class of functionals which we call *smooth*, the linear term is given by integrating against a well-behaved weight function. Smooth functionals will be seen to be estimable at rate $n^{-1/2}$ over $\mathscr{W}$, using our plug-in kernel estimator with an undersmoothed bandwidth.

In many cases the plug-in estimator will achieve the optimal rate over $\mathscr{W}$ for an appropriate choice of bandwidth. This may be unexpected since the estimate $\hat{f}$ of f will usually not lie in the restricted set $\mathscr{W}$.

For some naturally occurring kinds of degeneracy, however, the plug-in estimator cannot achieve the optimal rate. This is true for the functional $\Lambda(f) = \int f'g$ for certain g and $\mathscr{W}$, as discussed in Example 7. We find that a plug-in estimator is optimal as long as certain degenerate cases are excluded. The conditions in Definitions 3.3, 3.4 and 3.5 guarantee that $\mathscr{W}$ is a rich enough set so that the derivative of Λ does not vanish on $\mathscr{W}^*$, the set of local variations for $\mathscr{W}$. Such a condition is necessary for the existence of an upper bound of the correct order. In Example 6, where a plug-in estimate is not optimal, we discuss a method of finding a modified estimator that achieves the optimal rate.

We now give a simple example of an atomic and a smooth functional in the regression setting and a heuristic discussion of why in the smooth case undersmoothing is advantageous. For an atomic functional consider $\Lambda(f) = \int_0^1 (f')^2$ over $\mathscr{W} = \mathscr{W}_p$. We have the expansion

$$\Lambda(\hat{f}) - \Lambda(f) = T_f(\hat{f} - f) + O\left(\|\hat{f} - f\|_{(2,1)}^2\right),$$

where the linear functional $T_f(h) = 2\int_0^1 h'f'$ is the (Fréchet) derivative of Λ evaluated at f, and $\|\cdot\|_{(2,1)}$ is an appropriate Sobolev norm.

The derivative T_f will determine the rate of convergence of $\Lambda(\hat{f})$ to $\Lambda(f)$ if the remainder term is negligible. Investigating T_f, we have

$$\frac{1}{2}T_f(h) = \int_0^1 h'f' = -\int_0^1 hf'' + f'(0)h(0) - f'(1)h(1). \tag{1.4}$$

Then $\Lambda(\hat{f}) - \Lambda(f)$ has a term $f'(0)(\hat{f}(0) - f(0))$, taking $h = \hat{f} - f$ in the above. Hence if $f'(0) \neq 0$ estimating the functional should be at least as hard as estimating $f(0)$. That is, T_f is atomic: Its worst behavior for $f \in \mathscr{W}_p$ produces a point evaluation component of "index 0." Hence the best rate over $\mathscr{W}_p$ is the usual pointwise rate $n^{-p/(2p+1)}$ with the bandwidth $b \sim n^{-1/(2p+1)}$. Theorem 4.1 shows that the plug-in estimator $\Lambda(\hat{f})$ achieves this rate over $\mathscr{W}_p$, and that this rate is optimal.

For a smooth functional consider $\Lambda(f) = \int_0^1 f^2$. We have an expansion as before where now $(1/2)T_f(h) = \int_0^1 hf$, and so Λ is smooth over $\mathscr{W}_p$. Theorem 4.2 then says the plug-in estimator $\Lambda(\hat{f})$ with an "undersmoothed" bandwidth achieves the optimal rate $n^{-1/2}$ over $\mathscr{W}_p$.

In the smooth case we may almost always undersmooth and achieve the rate of convergence $n^{-1/2}$. To see why, suppose for convenience that the data are equally spaced, and we use a simple translation kernel:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} Y_i \frac{1}{b} K\left(\frac{x - i/n}{b}\right).$$

Let w_f be the weight function of T_f, that is, $T_f(h) = \int h w_f$. In the example $\Lambda(f) = \int f^2$ above, $w_f = 2f$. Taking $h = \hat{f} - f$,

$$T_f(\hat{f} - f) = \int \left\{ \frac{1}{n}\sum f\left(\frac{i}{n}\right)\frac{1}{b}K\left(\frac{x - i/n}{b}\right) - f(x)\right\} w_f(x)\,dx$$
$$+ \frac{1}{n}\sum \varepsilon_i \int \frac{1}{b} K\left(\frac{x - i/n}{b}\right) w_f(x)\,dx.$$

The first sum is not stochastic, and by the properties of the kernel is $O(b^p)$. The second term, in which we have interchanged summation and integration, is an average of independent mean zero terms; the variance of each term may be bounded independently of b as b may be absorbed in the integral by a change of variable. Hence the second term is $O_p(n^{-1/2})$, and we need only ensure that $b^p \sim n^{-1/2}$ in order to achieve the desired rate. To make this argument rigorous requires the bias to be $O(b^p)$ uniformly in x. Hence we use the boundary-corrected kernel $K_b(x, t)$ of the Appendix.

A functional may be atomic or smooth over $\mathscr{W}$, depending on $\mathscr{W}$. Consider the atomic example $\Lambda(f) = \int_0^1 (f')^2$ given above, but now take

$$\mathscr{W} = \{f \in \mathscr{W}_p : f'(0) = f'(1) = 0\}.$$

In this case, using (1.4), we have the representation

$$\frac{1}{2}T_f(h) = -\int_0^1 hf'' \quad \text{for } f \in \mathscr{W},$$

and so Λ is smooth over $\mathscr{W}$. This suggests that if f is known to lie in a restricted class $\mathscr{W}$, the extra information may be used to advantage.

Levit (1979) considers a type of plug-in estimator for the estimation of functionals that can be written in the special form given in (3.16). He is able to show the plug-in type estimator is efficient in this case.

Has'minskii and Ibragimov (1979) have also studied estimating Fréchet differentiable functionals of a density function f, in general function spaces and with general loss functions. Their functionals do not involve higher derivatives of f, and in our terminology are smooth. They begin with the plug-in estimator $\Lambda(\hat{f})$, with $\hat{f}$ an arbitrary estimator which attains the optimal point-wise rate. As they do not undersmooth, they find their plug-in estimator to be "as a rule, very bad."

Pfanzagl (1985) considers differentiable statistical functionals of a distribution function in a general framework. The focus of his efforts is on obtaining asymptotic bounds for the performance of statistical procedures; his emphasis is not on the existence of such procedures. He briefly considers modified plug-in estimators in Section 10.7, following Has'minskii and Ibragimov (1979). Undersmoothing is not considered.

Hall and Marron (1987) and Bickel and Ritov (1988) consider estimation of the functional (1.2); overlap with our approach is discussed in Example 2. Donoho (1988) gives a method for computing one-sided confidence bounds for some functionals of densities; optimality of rates is not discussed. Under certain "renormalization" conditions, Donoho and Low (1992) obtain general optimality results for linear functionals on classes of functions themselves specified by conditions on functionals. Donoho and Liu (1991) consider optimal rates of convergence for linear functionals in general and include three nonlinear examples as well; as in Donoho (1988), the class of underlying functions considered there is broader than here.

The remainder of the paper is organized as follows: Section 2 presents the notation and model assumptions used throughout the paper. In Section 3 we specify conditions on the functionals we study, classify them as atomic or smooth and provide simple propositions useful in verifying whether a functional satisfies certain technical differentiability conditions. In Section 4 we state our results on optimal rates of convergence, and in Section 5 we give examples. Proofs are presented in Section 6, and a formula for the kernel and its properties are given in the Appendix.

2. Notation and model assumptions. For s a nonnegative integer, let $C^s[0,1]$ denote the set of continuous functions on $[0,1]$ with s or more continuous derivatives. For $f \in C^s[0,1]$ and $1 \le q < \infty$, let

$$\|f\|_{(q,s,\lambda)} = \sum_{j=0}^{s} \left\{ \int_0^1 \left(f^{(j)} \right)^q d\lambda_j \right\}^{1/q}, \tag{2.5}$$

where λ is a vector of measures $(\lambda_1, \ldots, \lambda_s)$. When λ is the vector of Lebesgue measures, we shall drop the dependence on λ and write $\|f\|_{(q,s)}$; $\|f\|_{(\infty,s)}$ is defined similarly.

Let $p \ge 2$ be a nonnegative integer. The kernel $K_b(x,t)$ of order p we use is given by (A.24) of the Appendix. $K_b^{(i,j)}(x,t)$ will denote the i,jth mixed partial

derivative in x and t. For a function h,

$$h_b(x) = \int_0^1 K_b(x,t)h(t)\,dt. \tag{2.6}$$

The regression and density functions considered will be assumed to lie in a subset $\mathscr{W}$ of a set $\mathscr{W}_p$. For regression functions

$$\mathscr{W}_p = \{f \in C^p[0,1]: \|f\|_{(\infty,p)} < M\} \tag{2.7}$$

for M a constant fixed throughout, and in an abuse of notation clear in context, for density functions

$$\mathscr{W}_p = \left\{ f \in C^p[0,1]:\ f \geq 0,\ \int_0^1 f = 1,\ \|f\|_{(\infty,p)} < M \right\}.$$

The functionals we consider are (perhaps extended) real-valued functions on $C^p[0,1]$ and are well behaved on $\mathscr{W}$.

All estimators of $\Lambda(f)$ considered are of the form $\hat{\Lambda}_n = \Lambda(\hat{f}_n)$, where $\hat{f}_n$ is the following kernel estimate of f with $b = b_n$:

1. In the regression case

$$\hat{f}_n(x) = \frac{1}{n\hat{\rho}_n(x)} \sum_{i=1}^n Y_i K_b(x, X_i), \tag{2.8}$$

when observing $(X_1, Y_1), \ldots, (X_n, Y_n)$ independent and identically distributed random variables, where X has density ρ, and $f(x) = E[Y|X = x]$. The quantity $\hat{\rho}_n$ above is the density estimate given in case 2 using the same bandwidth b.

2. In the density estimation case

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^n K_b(x, X_i), \tag{2.9}$$

when observing $X_1, \ldots, X_n$ independent and identically distributed random variables with density f.

We construct estimators that are optimal over $\mathscr{W}$ in the sense of Stone (1980). Let Λ be a functional on $\mathscr{W}_p$ and let $\{\hat{\Lambda}_n\}$ denote a sequence of estimators of $\Lambda(f)$ such that $\hat{\Lambda}_n$ is based on a sample of size n from an unknown distribution that depends on $f \in \mathscr{W}$. A positive number γ is called an *upper bound* to the rate of convergence over $\mathscr{W}$ if for every sequence $\{\hat{\Lambda}_n\}$ of estimators

$$\liminf_n \sup_{f \in \mathscr{W}} P\left(\left|\hat{\Lambda}_n - \Lambda(f)\right| > cn^{-\gamma}\right) > 0 \quad \text{for all } c > 0 \tag{2.10}$$

and

$$\lim_{c \to 0} \liminf_n \sup_{f \in \mathscr{W}} P\left(\left|\hat{\Lambda}_n - \Lambda(f)\right| > cn^{-\gamma}\right) = 1. \tag{2.11}$$

The number γ is called an *achievable* rate of convergence over $\mathscr{W}$ if there is a sequence $\{\hat{\Lambda}_n\}$ of estimators such that

$$\lim_{c\to\infty} \limsup_n \sup_{f\in\mathscr{W}} P\left(\left|\hat{\Lambda}_n - \Lambda(f)\right| > cn^{-\gamma}\right) = 0. \tag{2.12}$$

The number γ is called the *optimal* rate of convergence over $\mathscr{W}$ if it is both an upper bound to the rate of convergence and an achievable rate of convergence.

Roughly then, if γ is an [upper bound to the rate of convergence/achievable rate of convergence], then [for any sequence of estimators $\hat{\Lambda}_n$/there exists a sequence of estimators $\hat{\Lambda}_n$ such that] the true value $\Lambda(f)$ will lie [outside/inside] an interval around $\hat{\Lambda}_n$ with probability tending to 1 when the length of the interval is tending to 0 [faster/slower] than $n^{-\gamma}$.

We now state the assumptions of our models.

In the regression setting we adopt essentially the assumptions of Model 1 of Stone (1980).

ASSUMPTION 2.1. *In the regression setting*:

1. (X_i, Y_i), $i = 1, \ldots, n$, *are independent copies of the pair of real-valued random variables* (X, Y) *with* $f(x) = E[Y|X = x]$, *an unknown member of* $\mathscr{W} \subseteq \mathscr{W}_p$.
2. *The distribution of* X *is absolutely continuous with respect to Lebesgue measure on* $[0, 1]$, *with density* $\rho \in C^p[0, 1]$, *and there exist constants* β *and* η *such that* $0 < \beta < \rho(x) < \eta < \infty$.
3. *The conditional distribution of* Y *given* x *satisfies the assumptions of Model 1 of Stone* (1980). *In particular, the conditional variance of* Y *given* X, $\mathrm{Var}[Y|X = x] = \sigma^2(x)$ *satisfies* $0 < \kappa < \sigma(x) < \zeta < \infty$.

ASSUMPTION 2.2. *In the density estimation case we observe* $X_1, X_2, \ldots, X_n$ *independent with density* $f \in \mathscr{W} \subseteq \mathscr{W}_p$. *In the atomic case, Theorem* 4.1, *we assume further that* $\inf_{x\in[0,1]} f(x) > 0$.

We will use the notation $t = O(q)$ in a stronger sense than is conventional. We will write $t = O(q)$ when there exist constants C and δ depending only on M of (2.7) (and hence, in particular, which may be chosen independently on $f \in \mathscr{W}$) such that

$$|t| \le C|q| \quad \text{for all } q \text{ with } |q| \le \delta.$$

Similarly, for statements in probability about random variables T_n, Q_n whose distributions may depend on f, we write

$$T_n = O_p(Q_n),$$

when

$$\lim_{c\to\infty} \limsup_n \sup_{f\in\mathscr{W}} P(|T_n| > c|Q_n|) = 0.$$

In what follows p will denote the degree of smoothness of f and the order of the kernel $K_b(x, t)$. C will denote a positive constant which is not necessarily the same at each occurrence. We often will write b for b_n.

3. Classification and differentiability of functionals. Section 3.1 presents a Riesz-type representation for bounded linear functionals T and defines *atomic*, *index* and *smooth*, first for bounded linear functionals T on $\mathscr{W}_p$, and then for the class of differentiable functionals Λ which we study.

The functionals Λ are required to admit the "Taylor" expansion (3.13) with first term $T_f(h)$ linear in the perturbation h and remainder depending on only the first m derivatives of h, $0 \le m < p$. Generally the *order* m of the functional keeps track of how many derivatives of h are involved in the remainder.

Essentially, the linear term $T_f(h)$ will classify functionals as smooth if T_f has no point evaluation component and as atomic otherwise. For atomic functionals the *index* r, defined below, is the highest derivative of h for which the derivative of Λ contains a point evaluation.

We state degeneracy conditions which will be used to rule out exceptional cases. Section 3.2 gives our notion of differentiability for functionals (essentially Fréchet differentiability with a slightly different remainder condition), and presents some simple propositions which may be used to verify the differentiability of a given functional.

3.1. *Classification.* For a linear functional T on $C^p[0, 1]$, define

$$\|T\|_p = \sup_{\|h\|_{(\infty, p)} \le 1} |Th|.$$

We say T is bounded on $C^p[0, 1]$ if $\|T\|_p < \infty$.

We consider functionals Λ that have the following expansion for $f \in \mathscr{W}$ and $h \in C^p[0, 1]$ with $\|h\|_{(\infty, m)}$ sufficiently small:

$$\Lambda(f + h) = \Lambda(f) + T_f(h) + O\big(\|h\|^2_{(2, m, \lambda)}\big), \tag{3.13}$$

where T_f, the derivative of Λ at f, is a bounded linear functional on $C^p[0, 1]$, m is an integer $0 \le m < p$ and λ is a vector of finite measures that may depend on Λ but not on f.

By an application of the Hahn–Banach theorem, we extend T_f to the product space $\Pi_{j=1}^{p} C[0, 1]$, and then use the Riesz representation theorem to derive that T has representation(s)

$$T_f(h) = \sum_{j=0}^{p} \int_0^1 h^{(j)}\, d\mu_j \quad \text{for } h \in C^p[0, 1],$$

where the μ_j are finite signed Borel measures on $[0, 1]$ which may depend on f.

We now define the notions of *atomic*, *index* and *smooth*.

DEFINITION 3.1. If for some integer s, $0 \le s \le p$, T has a representation such that for all $h \in C^p[0,1]$,

$$T(h) = \sum_{j=0}^{s} \int_0^1 h^{(j)}\, d\mu_j,$$

where μ_s has a discrete component δ_s, then we say that T is *atomic*. We define r, the *index* of T, to be the largest s that can appear in any such representation.

By extension, if Λ has atomic derivative T_f of index r at f we say Λ *is atomic of index r at f* and write the discrete component of the measure $\mu_r = \mu_{r,f}$ in the above representation as $\delta_{r,f}$. Finally, Λ *is of index r on* $\mathscr{W} \subseteq \mathscr{W}_p$ *if* $\max_{f \in \mathscr{W}} \text{index}(T_f) = r$.

DEFINITION 3.2. If, for all $h \in C^p[0,1]$,

$$T(h) = \int_0^1 hw$$

for some bounded measurable function w, we say that T is *smooth*. By extension, if Λ has smooth derivative at f we say Λ *is smooth at f*, and that Λ *is smooth on* $\mathscr{W}$ if T_f is smooth for all $f \in \mathscr{W}$.

Note that smooth is not the same as index 0.

Next we define $\mathscr{W}^*$, the set of local variations used for taking directional derivatives at any $f \in \mathscr{W}$. The space depends on the class $\mathscr{W}$ in which f is known to lie:

DEFINITION 3.3. Given $\mathscr{W}$, define $\mathscr{W}^*$, the space of local variations for $\mathscr{W}$, by

$$\mathscr{W}^* = \{h \in C^p[0,1]: \forall f \in \mathscr{W},\ f + \varepsilon h \in \mathscr{W} \text{ for all } \varepsilon > 0 \text{ sufficiently small}\}.$$

We state the following degeneracy conditions which will be used to rule out exceptional cases. For a degenerate functional, the upper bound proof breaks down and faster rates may obtain. Basically, the first condition says that degeneracy exists if the rate-determining measure is annihilated on the set of local variations. The second condition will be used to exclude "thin" classes of functions. It ensures that $\mathscr{W}$ is rich enough so that $\mathscr{W}^*$ contains small "bump function" perturbations. We shall prove that when Λ is not degenerate, a plug-in estimator is optimal. Any nonconstant functional on $\mathscr{W} = \mathscr{W}_p$ is automatically not degenerate.

DEFINITION 3.4. Let Λ be an atomic functional of index r on $\mathscr{W}$. We say $(\Lambda, \mathscr{W})$ *is strongly degenerate* if for all $f \in \mathscr{W}$ such that T_f is of index r:

$$\int_0^1 h^{(r)}\, d\delta_{r,f} = 0 \quad \text{for all } h \in \mathscr{W}^*.$$

(Recall that $\delta_{r,f}$ is the discrete component of the measure $\mu_{r,f}$ in the representation of T_f, the derivative of Λ at f, as given in Definition 3.1.) We say $(\Lambda, \mathscr{W})$ is *not degenerate* if for all $x_0 \in [0,1]$ there is a function ψ of compact support with $\psi^{(r)}(0) \neq 0$ such that

$$a^{-p}\psi(a(x - x_0))\mathbf{1}(x \in [0,1]) \in \mathscr{W}^* \tag{3.14}$$

for all a sufficiently large; otherwise we call $(\Lambda, \mathscr{W})$ degenerate.

In the atomic case, it is clear that $(\Lambda, \mathscr{W})$ is degenerate whenever it is strongly degenerate. Condition (3.14) may be used to rule out finite-dimensional classes of functions. For example, take $\mathscr{W}$ to be all polynomials of degree less than or equal to k. Then any atomic functional Λ is degenerate, but not necessarily strongly degenerate, on $\mathscr{W}$: For example, take $\Lambda(f) = f^{(r)}(0)$ with $r \leq k + 1$. That this pair $(\Lambda, \mathscr{W})$ is degenerate is consistent with the fact that the rate given in Theorem 4.1 is not optimal, and plug-in estimators do not achieve the optimal rate $n^{-1/2}$.

For a discussion of strong degeneracy, see the end of Example 6.

DEFINITION 3.5. Let Λ be a smooth functional on $\mathscr{W}$. We say $(\Lambda, \mathscr{W})$ is *degenerate* if for all $f \in \mathscr{W}$,

$$\int_0^1 h w_f = 0 \quad \text{for all } h \in \mathscr{W}^*.$$

For example, $\Lambda(f) = \int_A f$ is degenerate on $\mathscr{W} = \{f\colon f\mathbf{1}_A = 0\}$.

3.2. *Differentiability of functionals.* Definition 3.6 gives our notion of differentiability, essentially that expansion (3.13) is uniform over $\mathscr{W}$. The propositions which follow may be used to verify the differentiability of a given functional.

DEFINITION 3.6. Let $0 \leq m < p$.

1. We say Λ *is differentiable of order m on* $\mathscr{W} \subseteq \mathscr{W}_p$ if for every $f \in \mathscr{W}$, Λ has an expansion of the form (3.13) and
$$\sup_{f \in \mathscr{W}} \|T_f\|_p < \infty.$$
2. We say Λ *is smooth of order m on* $\mathscr{W} \subseteq \mathscr{W}_p$ if Λ is differentiable of order m on $\mathscr{W}$ and for all $f \in \mathscr{W}$, Λ has smooth derivative T_f with $w = w_f$ and
$$\sup_{f \in \mathscr{W}} \|w_f\|_\infty < \infty. \tag{3.15}$$

In many examples, Λ is of the special form

$$\Lambda(f) = \int_0^1 g(u, f^{(0)}(u), f^{(1)}(u), \ldots, f^{(m)}(u))\, du \tag{3.16}$$

for some $g\colon \mathbf{R}^{m+2} \to \mathbf{R}$.

The following proposition shows that such functionals are differentiable when g is sufficiently smooth.

PROPOSITION 3.7. *Let* Λ *be given by* (3.16) *for* $0 \le m < p$. *Let* W *be a set in* R^{m+2} *such that*

$$\{\langle u, f^{(0)}(u) + h^{(0)}(u), \ldots, f^{(m)}(u) + h^{(m)}(u)\rangle : u \in [0,1],\ f \in \mathscr{W},\ \|h\|_{(\infty,m)} < \varepsilon\} \subseteq W.$$

Suppose for some $\varepsilon > 0$, g *has two continuous derivatives on the convex closure of* W. *Then* Λ *is differentiable of order* m *on* $\mathscr{W}$.

PROPOSITION 3.8. *If* $m = 0$ *in Proposition* 3.7, *then* Λ *is smooth of order* 0 *on* $\mathscr{W}$.

The following proposition shows that differentiability properties are retained under compositions.

PROPOSITION 3.9. *Let* Λ *be differentiable of order* m *and* [*of index* r/*smooth*] *on* $\mathscr{W}$. *Let* $s = \max(m, r)$ *in the atomic case and* $s = m$ *in the smooth case. Suppose that for some* $\varepsilon > 0$, g *is twice continuously differentiable on* I, *a bounded, closed interval containing* $\{\Lambda(f+h)\colon f \in \mathscr{W}, \|h\|_{(\infty,m)} < \varepsilon\}$ *with* $g'(x) \neq 0$ *on* I. *Then* $\Gamma(f) = g(\Lambda(f))$ *is differentiable of order* s *on* $\mathscr{W}$, *and* [*of index* r/*smooth*].

We state one last result to be applied to Example 4 involving the asymptotic variance of U-statistics; there, the case $d = 3$ will apply.

PROPOSITION 3.10. *If* $\psi\colon [0,1]^d \to \mathbf{R}$ *is a bounded measurable function, and* $A \subseteq [0,1]^d$ *is measurable, then the functional*

$$\Lambda(f) = \int \cdots \int_A \psi(x_1, \ldots, x_d) f(x_1) \cdots f(x_d)\, dx_1 \cdots dx_d$$

is smooth of order 0 *for every* $\mathscr{W} \subseteq \mathscr{W}_p$.

PROOF. The proof is a consequence of the relations

$$T_f(h) = \sum_{i=1}^{d} \int \cdots \int_A \psi(\mathbf{x}) h(x_i) \prod_{j \neq i} f(x_j)\, d\mathbf{x}$$

and

$$\begin{aligned} &|\Lambda(f+h) - \Lambda(f) - T_f(h)| \\ &\quad = \left| \int \cdots \int_A \psi(\mathbf{x}) \sum_{i=2}^{d} \prod_{|J|=i,\, j \in J,\, k \notin J} h(x_j) f(x_k)\, d\mathbf{x} \right| \\ &\quad \le C\left[\int_0^1 |h(x)|\, dx\right]^2 \le C\|h\|_1^2 \le C\|h\|_2^2 \end{aligned}$$

for $\|h\|_\infty \le 1$. □

4. Statement of results. For atomic functionals we have the following result.

THEOREM 4.1. *Let Λ be a differentiable functional of order m and index r on $\mathscr{W} \subseteq \mathscr{W}_p$, where $p \geq \max\{m+2, r+1, 2m-r+1\}$. Let $c_2 > c_1 > 0$ be constants and let $b_n \in [c_1 n^{-1/(2p+1)}, c_2 n^{-1/(2p+1)}]$. If $(\Lambda, \mathscr{W})$ is not degenerate, then $\Lambda(\hat{f}_n)$ achieves the optimal rate $n^{-(p-r)/(2p+1)}$ on $\mathscr{W}$.*

For smooth functionals we have the following result.

THEOREM 4.2. *Let Λ be a differentiable smooth functional of order m on $\mathscr{W} \subseteq \mathscr{W}_p$, and $p \geq \max\{2, 2m+1\}$. Let c be a positive constant and let $b_n \in [cn^{-1/(4m+2)}, cn^{-1/2p}]$. If $(\Lambda, \mathscr{W})$ is not degenerate, then $\Lambda(\hat{f}_n)$ achieves the optimal rate $n^{-1/2}$ on $\mathscr{W}$.*

REMARK. One can get a rough idea of the magnitude of the constant term in the $n^{-1/2}$ rate of convergence in the smooth case by considering the case where p is large and ρ is known. In this instance, with $b = o(n^{-1/2p})$ it can be shown that the variance in the linear term dominates and the constant is asymptotic to $\int(f^2 + \sigma^2)w^2\rho^{-1} - (\int fw)^2$.

In the smooth density case, under conditions and using the results of Levit (1979), Goldstein and Khas'minskii (1992) are able to show the plug-in estimator achieves the best constant $\operatorname{Var} w_f(X)$, where X has density f and, moreover, is locally asymptotically minimax.

5. Examples. Here we present five natural applications of our results and one last example to illustrate degeneracy.

EXAMPLE 1 [$\Lambda(f) = f^{(r)}(x_0)$]. We recover results about optimal pointwise estimation of regression and density functions in Stone (1980) by considering the point evaluation functional Λ as a nondegenerate differentiable functional of order 0, say, and index r. Hence, by Theorem 4.1, Λ is estimable at the optimal rate $n^{-(p-r)/(2p+1)}$ over $\mathscr{W}_p$ for $p \geq \max\{2, r+1\}$. Note that Theorem 4.1 yields this same rate for functionals $g(f^{(r)}(x_0))$ such as $(f^{(r)}(x_0))^q$ or $\exp(f^{(r)}(x_0))$ by Proposition 3.9.

EXAMPLE 2 [$\Lambda(f) = \int_0^1 |f^{(m)}|^q$]. The L_q^q norm functional Λ satisfies Proposition 3.7 for $q \geq 2$, and hence is differentiable of order m for any $\mathscr{W} \subseteq \mathscr{W}_p$. This functional is important in the choice of a smoothing parameter for estimation of a density or regression function.

Taking q even for illustration, the derivative T_f of Λ at f is given by

$$T_f(h) = q\int_0^1 h^{(m)}\left(f^{(m)}\right)^{q-1}.$$

If $m = 0$ this demonstrates that Λ is smooth. Hence, for $p \geq 2m+1$, the optimal rate of $n^{-1/2}$ is achieved by the plug-in estimator whenever $(\Lambda, \mathscr{W})$ is

nondegenerate. This is satisfied whenever $\mathscr{W}^* \neq \varnothing$. However, Λ may be smooth even when $m \geq 1$ for certain $\mathscr{W}$. For example, take $q = 2$, and integrate by parts:

$$\frac{1}{2}T_f(h) = \int_0^1 h^{(m)} f^{(m)}$$

$$= h^{(m-1)}(1) f^{(m)}(1) - h^{(m-1)}(0) f^{(m)}(0) - \int_0^1 h^{(m-1)} f^{(m+1)}.$$

We see therefore that T_f is of index $r = m - 1$ on $\mathscr{W}_p$, and that by Theorem 4.1 the optimal rate over $\mathscr{W}_p$ of $n^{-(p-(m-1))/(2p+1)}$ is achieved by the plug-in estimator, for $p \geq m + 2$. However, if $p \geq 2m + 1$ and f is known to lie in the class of functions $\mathscr{W}^{2m}$ which satisfy boundary conditions:

$$\mathscr{W}^{2m} = \left\{ f \in \mathscr{W}_p \colon f^{(j)}(0) = f^{(j)}(1) = 0,\ m \leq j \leq 2m - 1 \right\},$$

then all atomic terms in T_f drop out. In fact,

$$\frac{1}{2}T_f h = (-1)^m \int_0^1 f^{(2m)} h;$$

hence T_f is smooth. In this case the functional Λ is easily seen to be nondegenerate on $\mathscr{W}^{2m}$ and hence is estimable at the optimal rate $n^{-1/2}$ by the undersmoothed plug-in estimator. Generally, if f lies in the class of functions $\mathscr{W}^i$ which satisfy $f^{(j)}(0) = f^{(j)}(1) = 0$ for $m \leq j < i \leq 2m - 1$ but $f^{(i)}$ is nonzero at, say, 0, then Λ is of index $r = 2m - i - 1$ and the intermediate rate $n^{-(p-(2m-1-i))/(2p+1)}$ obtains. This rate is achieved by the plug-in estimator with no undersmoothing, by Theorem 4.1. By Proposition 3.9 similar remarks apply to the estimation of the L^q norm itself on any set that excludes a neighborhood of the zero function.

Related discussions of the functional in this example appear in Has'minskii and Ibragimov (1979), Hall and Marron (1987) and Bickel and Ritov (1988). These last two references deal with the case of a density function on the entire real line and investigate how the rate of convergence depends on the smoothness of the underlying density. They show that rates slower than $n^{-1/2}$ obtain when the density fails to have sufficient derivatives, a case we do not consider. Ritov and Bickel (1990) show in this latter case that without sufficient smoothness the functional is not estimable at rate $n^{-\alpha}$ for any $\alpha > 0$. We show that rates slower than $n^{-1/2}$ obtain on a finite interval if certain boundary conditions are satisfied, a case they do not consider. With sufficient smoothness ($p \geq 2m + 1$) and boundary conditions, all authors obtain the rate $n^{-1/2}$.

With $m = 0$ and $q = 2$ the functional appears in the variance of the Hodges–Lehmann estimator and the Pitman efficiency of the one-sample Wilcoxon test to the one-sample t test [Lehmann (1975)]. It is also investigated in nonparametric sequential ranks by Sen (1981) and Fenstad and Skovlund (1990).

EXAMPLE 3 [$\Lambda(f) = 2\int_0^1 fg/(f+g)$]. When $\inf_x g > 0$ the equidistributional functional Λ is smooth of order 0 on $\mathscr{W}_p$ by Proposition 3.8. Hence Theorem 4.2 shows that the plug-in estimator achieves the optimal rate $n^{-1/2}$.

In Abramson and Goldstein (1991), the equidistribution functional Λ is shown to be invariant under smooth invertible transformations of the data, and to enjoy the following property: $\Lambda = 1$ if and only if $f = g$; if $f \neq g$, then $\Lambda < 1$. In the one-sample problem to test whether or not observed data are generated from the known density g, it is therefore of interest to estimate Λ.

EXAMPLE 4 [$\Lambda(f) = \int_0^1[\int_0^1 \psi(y,x) f(y)\,dy]^2 f(x)\,dx$]. That $\Lambda(f)$ is smooth of order 0 on $\mathscr{W}_p$ follows from Proposition 3.10 for $d = 3$. For $p \geq 1$, Λ is therefore estimable at the optimal rate $n^{-1/2}$ by Theorem 4.2. The functional Λ is the asymptotic variance of the U-statistic

$$U = \binom{n}{2}^{-1} \sum_{i<j} \psi(X_i, X_j),$$

where $X_1, X_2, \ldots, X_n$ are i.i.d. with density f on $[0,1]$, ψ is bounded measurable function and $EU = 0$ for convenience. See Lehmann (1975), page 367.

EXAMPLE 5 [$\Lambda(f) = \int_0^1 (f_{b_0}^{(m)}(x) - f^{(m)}(x))^2\,dx$]. The function f_{b_0} is as in (2.6); take $\mathscr{W} = \mathscr{W}_p$.

This functional arises when considering the conditional integrated mean-squared error $E[\int(\hat{f}^{(m)} - f^{(m)})^2 | \mathbf{X}]$ with $\rho \equiv 1$ for convenience, and in particular the conditional bias term

$$\frac{1}{n}\sum_{i=1}^{n} K_{b_0}^{(m,0)}(x, X_i) f(X_i) - f^{(m)}(x).$$

For large samples the above sum may be approximated by an integral and hence the conditional bias term by $f_{b_0}^{(m)}(x) - f^{(m)}(x)$. Hence the squared bias component of the integrated mean-squared error for estimating $f^{(m)}$ using bandwidth b_0 is approximately $\Lambda(f)$. Therefore this functional is of interest when one wishes to choose a bandwidth b_0 which estimates $f^{(m)}$ well in the integrated mean-squared error sense. An estimator of the optimal bandwidth may be given by minimizing an appropriate function of $\Lambda(\hat{f}, b_0)$ over b_0. Of course, a bandwidth b_0 which estimates $f^{(m)}$ well may not in general be the same as a bandwidth b which estimates $\Lambda(f)$ well.

A variational argument shows Λ to be differentiable of order m on $\mathscr{W}_p$ for $p > m$, with derivative

$$\frac{1}{2}T_f(h) = \int_0^1 \left(h_{b_0}^{(m)}(x) - h^{(m)}(x)\right)\left(f_{b_0}^{(m)}(x) - f^{(m)}(x)\right) dx.$$

For $m = 0$, Fubini's theorem shows Λ to be smooth and therefore estimable at the optimal rate $n^{-1/2}$. For $m \geq 1$, integrating by parts as in Example 2, we

see that T_f is of index $m - 1$ over $\mathscr{W}_p$. Hence we have from Theorem 4.1 that $n^{-(p-m+1)/(2p+1)}$ is the optimal rate of convergence for estimating the asymptotic integrated squared bias over $\mathscr{W}_p$ for $p \geq m + 2$. These rates are achieved by the plug-in estimator $\Lambda(\hat{f})$.

The case with general ρ is tractable but more complicated.

EXAMPLE 6. This example illustrates two cases for which the plug-in estimator is not optimal. In the first we show how to construct a modified estimator.

Let

$$\Lambda(f) = \int_0^1 f'g,$$

where $g \in C^1[0, 1]$ and $g(0) = g(1) = 1$. Then Λ is linear, so for $f \in \mathscr{W}_p$, $h \in C^p[0, 1]$,

$$\begin{aligned} T_f(h) &= \int_0^1 h'g \\ &= -\int_0^1 hg' + (h(1) - h(0)). \end{aligned} \tag{5.17}$$

Hence Λ is differentiable of order 0, say, and is atomic of index 0 over $\mathscr{W}_p$. Since Λ is not degenerate over $\mathscr{W}_p$, the optimal rate is the pointwise rate $n^{-p/(2p+1)}$, and is achieved by the plug-in estimator $\Lambda(\hat{f})$ by Theorem 4.1 for $p \geq 2$.

Now suppose f is known to lie in the restricted class of functions

$$\mathscr{W} = \{f \in \mathscr{W}_p : f(0) = f(1)\}.$$

Since the derivative of Λ does not depend on f, Λ is still of index 0 over $\mathscr{W}$. This is unlike all our previous examples, where the derivative depends on f. This functional is strongly degenerate on $\mathscr{W}$: The atomic part of (5.17) is identically 0 for $h \in \mathscr{W}^*$. Hence the hypothesis of Theorem 4.1 is not satisfied. Indeed, the smooth rate $n^{-1/2}$ is achieved over $\mathscr{W}$ by the estimator

$$\hat{\Lambda} = -\int_0^1 \hat{f}g'.$$

The plug-in estimator $\Lambda(\hat{f})$ by contrast can only achieve the slower pointwise rate $n^{-p/(2p+1)}$. This is because the estimate $\hat{f}$ lies in $\mathscr{W}_p$, not $\mathscr{W}$, and for this example that difference is crucial: For $f \in \mathscr{W}$, the higher terms of the derivative $T_f(h)$ do not vanish for all $h \in \mathscr{W}_p$ but only for $h \in \mathscr{W}^*$.

One can think of the modified estimator $\hat{\Lambda}$ given above as a plug-in estimator of a functional Λ' which is smooth on $\mathscr{W}_p$ and which coincides with the functional Λ on the restricted set $\mathscr{W}$.

Our second example is given by again considering $\Lambda(f) = \int_0^1 (f')^2$ with derivative

$$\frac{1}{2}T_f h = -\int_0^1 hf'' + f'(0)h(0) - f'(1)h(1)$$

as in (1.4). Taking $\mathscr{W} = \{f \in \mathscr{W}_p\colon f(0) = f(1) = 0\}$ forces $\mathscr{W}^* \subset \{h \in C^p[0,1]\colon h(0) = h(1) = 0\}$. Since there exists $f \in \mathscr{W}$ with $f'(0) \neq 0$, Λ is of index 0 on $\mathscr{W}$. The pair $(\Lambda, \mathscr{W})$ is strongly degenerate since the atomic terms of the derivative vanish for $h \in \mathscr{W}^*$, that is,

$$\int_0^1 h\, d\delta_{0,f} = f'(0)h(0) - f'(1)h(1) = 0$$

for all $h \in \mathscr{W}^*$ and $f \in \mathscr{W}$ such that T_f is of index 0. Hence one would not expect the plug-in estimator to achieve the optimal rate.

6. Proofs. We give the proofs for the regression setting. The proofs in the density estimation case are similar, but much simpler in the smooth case.

PROOF OF THEOREM 4.1. *Upper bound.* Since Λ is of index r we may choose $f_0 \in \mathscr{W}$ such that T_{f_0} has a representation for $h \in C^p[0,1]$:

$$T_{f_0}(h) = \sum_{j=0}^{r} \int_0^1 h^{(j)}\, d\mu_j, \tag{6.18}$$

where without loss of generality $\mu_r(\{x_0\}) = \eta > 0$ for $x_0 \in [0,1]$. We argue as in Stone (1980). Since Λ is not degenerate on $\mathscr{W}$, we may choose ψ of compact support so that it satisfies the rescaling condition (3.14), taking $\psi^{(r)} > 0$ without loss of generality.

Let $N > 0$ and $\delta \in (0,1]$ be arbitrary, let $\tau = 1/(2p+1)$, and define g_n by

$$g_n(x) = \delta N^p n^{-\tau p} \psi\big(N^{-1} n^{\tau}(x - x_0)\big)\mathbf{1}(x \in [0,1]).$$

Let $f_n = f_0 + g_n$. Condition (3.14) gives that $f_n \in \mathscr{W}$ for n sufficiently large.

Equations (2.2) and (2.3) in Stone (1980) follow as in Stone (1980). Using the same classification argument as in Stone (1980), it follows that for any sequence of estimators $\{\hat{\Lambda}_n\}$:

$$\liminf_n \sup_{f \in \mathscr{W}_p} P_f\left(\left|\hat{\Lambda}_n - \Lambda(f)\right| \geq \frac{\Lambda(f_n) - \Lambda(f_0)}{2}\right) > 0.$$

Using the differentiability of Λ, we now compute that

$$\frac{\Lambda(f_n) - \Lambda(f_0)}{2} = \frac{1}{2}T_{f_0}(g_n) + O\big(\|g_n\|^2_{(2,m,\lambda)}\big).$$

Let $E_n = \text{support}(g_n)$; note that $E_n \downarrow \{x_0\}$ since ψ has compact support. Consider the final term in the sum (6.18):

$$\int_0^1 g_n^{(r)}\, d\mu_r.$$

Let the Jordan decomposition of μ_r be given by $\mu_r = \mu_r^+ - \mu_r^-$. Since μ_r^+ and μ_r^- are mutually singular $\mu_r^-(\{x_0\}) = 0$. Therefore $\lim_{n\to\infty} \mu_r^-(E_n) = 0$. Take n_0 so that $n \geq n_0$ implies

$$\mu_r^-(E_n) \leq \frac{\eta\psi^{(r)}(0)}{4\|\psi^{(r)}\|_\infty}.$$

Hence, breaking up the integral into two parts corresponding to μ_r^+ and μ_r^-, we derive that for n sufficiently large

$$\int_0^1 g_n^{(r)}\, d\mu_r \geq \frac{1}{4}\delta N^{p-r} n^{-(p-r)\tau}\eta\psi^{(r)}(0).$$

The remaining terms in the sum (6.18), those with $j < r$, and the remainder term in the expansion of Λ, $O(\|g_n\|^2_{(2,m,\lambda)})$, are all $o(n^{-(p-r)\tau})$. Hence for n sufficiently large

$$\frac{\Lambda(f_n) - \Lambda(f_0)}{2} \geq \frac{\eta\delta N^{p-r}\psi^{(r)}(0)}{16} n^{-(p-r)\tau}.$$

Since N is arbitrary, this demonstrates that (2.10) and (2.11) follow as in Stone (1980).

Achievability. We first state the following lemma.

LEMMA 6.1. *Let μ be a finite measure, $0 \leq j < p$. Then*

$$E\left[\int_0^1 \left(\hat{f}^{(j)} - f^{(j)}\right)^2 d\mu\right] = O(b^{2(p-j)} + n^{-1}b^{-(2j+1)}),$$

$$E\left[\int_0^1 \left(\hat{\rho}^{(j)} - \rho^{(j)}\right)^2 d\mu\right] = O(b^{2(p-j)} + n^{-1}b^{-(2j+1)}).$$

PROOF. The proof follows by standard arguments as in Nadaraya (1990), Theorem 1.5, Chapter 4, page 121. □

We now show achievability.

Using the elementary Sobolev inequality $\|f\|_{(\infty,m)} \leq 2\|f\|_{(2,m+1)}$, the conditions on b_n and Lemma 6.1, we have $\|\hat{f} - f\|_{(\infty,m)} = O_p(a_n)$ with $a_n \to 0$. Therefore, since Λ is differentiable of index r and order m, we have the expansion as in (3.13):

$$\Lambda(\hat{f}) - \Lambda(f) = T_f(\hat{f} - f) + O\left(\|\hat{f} - f\|^2_{(2,m,\lambda)}\right)$$

for n sufficiently large.

From Lemma 6.1 and Markov's inequality, we have that

$$\int_0^1 \left(\hat{f}^{(j)} - f^{(j)}\right)^2 d\lambda_j = O_p(b^{2(p-j)} + n^{-1}b^{-(2j+1)}),$$

and then from the bounds on b_n that $\|\hat{f} - f\|^2_{(2,m,\lambda)} = O_p(n^{-2(p-m)/(2p+1)}) = O_p(n^{-(p-r)/(2p+1)})$.

It remains only to consider the linear term. Choose a representation for T_f as in Definition 3.1. Consider a term with a nonzero measure:

$$\int_0^1 \left(\hat{f}^{(j)} - f^{(j)} \right) d\mu_j, \qquad 0 \le j \le r,$$

and take the Jordan decomposition $\mu_j = \mu_j^+ - \mu_j^-$. Set $|\mu_j| = \mu_j^+ + \mu_j^-$. Without loss of generality we may assume that $|\mu_j|$ is a probability measure on [0, 1]; this will not change the order of the bounds:

$$\left| \int_0^1 \left(\hat{f}^{(j)} - f^{(j)} \right) d\mu_j \right| \le \int_0^1 \left| \hat{f}^{(j)} - f^{(j)} \right| d|\mu_j| \le \left(\int_0^1 \left(\hat{f}^{(j)} - f^{(j)} \right)^2 d|\mu_j| \right)^{1/2}$$

by Jensen's inequality. The above is $O_p(n^{-(p-j)/(2p+1)})$ which for $j \le r$ is $O_p(n^{-(p-r)/(2p+1)})$ as was to be shown. □

PROOF OF THEOREM 4.2. *Upper bound*. We argue as in Stone (1980). Since Λ is not degenerate, we may choose $f \in \mathscr{W}$ and $h \in \mathscr{W}^*$ such that $T_f(h) \neq 0$. Say $T_f(h) > 0$. With $N > 0$ and $\delta > 0$ arbitrary let $g_n(x) = \delta N n^{-1/2} h(x)$ and $f_n = f_0 + g_n$; $f_n \in \mathscr{W}$ for n sufficiently large. Equation (2.1) in Stone (1980) holds. Proceeding as in the previous theorem and using the differentiability of Λ, we have that

$$\begin{aligned} \frac{\Lambda(f_n) - \Lambda(f_0)}{2} &= \frac{1}{2} T_f(g_n) + O\left(\|g_n\|_{(2,m,\lambda)}^2 \right) \\ &= \frac{1}{2} \delta N n^{-1/2} T_f(h) + O(n^{-1}) \ge \frac{\delta N T_f(h)}{4} n^{-1/2} \end{aligned}$$

for n sufficiently large. Since $N > 0$ can be arbitrarily large, (2.10) holds using a classification argument as before.

Equation (2.11) follows as in Stone (1980).

Achievability. We first state the following lemma.

LEMMA 6.2.

$$\|\hat{\rho} - \rho_b\|_{(\infty,m)} = O_p\left(\frac{1}{\sqrt{n}\, b^{m+1}} \right)$$

and

$$\|\hat{\rho} - \rho\|_{(\infty,m)} = O_p\left(b^{p-m} + \frac{1}{\sqrt{n}\, b^{m+1}} \right).$$

Let $\tilde{f}$ be given by

$$\tilde{f}(x) = \frac{1}{n\rho(x)} \sum_{i=1}^{n} K_b(x, X_i) f(X_i). \tag{6.19}$$

Then

$$\| \tilde{f} - f \|_{(\infty, m)} = O_p\left(b^{p-m} + \frac{1}{\sqrt{n}\, b^{m+1}} \right).$$

PROOF. Take $0 \le j \le m$, let F be the distribution function of X, and F_n the empirical distribution function of $X_1, X_2, \ldots, X_n$.

We have

$$\begin{aligned} \left| \hat{\rho}^{(j)}(x) - \rho_b^{(j)}(x) \right| &= \left| \frac{1}{n} \sum_{i=1}^{n} K_b^{(j,0)}(x, X_i) - \rho_b^{(j)}(x) \right| \\ &= \left| \int_0^1 K_b^{(j,0)}(x,t)\, d\big(F_n(t) - F(t)\big) \right| \\ &= \left| \int_0^1 K_b^{(j,1)}(x,t)\big(F_n(t) - F(t)\big)\, dt \right| \\ &\le C D_n b^{-(j+1)} \end{aligned}$$

by absorbing a factor of b by a change of variable. As usual, $D_n = \sup_t |F_n(t) - F(t)|$. Since $D_n = O_p(n^{-1/2})$, the term above is $O_p(1/\sqrt{n}\, b^{j+1})$, where the constant in the O_p term does not depend on x. Since the term $j = m$ is dominant, this proves the first assertion of the lemma.

Next, use that

$$\begin{aligned} \| \hat{\rho}^{(j)} - \rho^{(j)} \|_\infty &= \left\| \frac{1}{n} \sum_{i=1}^{n} K_b^{(j,0)}(\cdot, X_i) - \rho^{(j)} \right\|_\infty \\ &\le \left\| \frac{1}{n} \sum K_b^{(j,0)}(\cdot, X_i) - \rho_b^{(j)} \right\|_\infty + \| \rho_b^{(j)} - \rho^{(j)} \|_\infty \end{aligned}$$

and invoke Theorem A.1. The proof of the second assertion is now completed by noting that the term $j = m$ is dominant. The proof of the last assertion is similar. □

We now show achievability.

Again, using an elementary Sobolev inequality and Lemma 6.1, we have, for n sufficiently large, the expansion

$$\Lambda(\hat{f}) - \Lambda(f) = T_f(\hat{f} - f) + O\big(\| \hat{f} - f \|_{(2, m, \lambda)}\big),$$

where $T_f(h) = \int_0^1 h w_f$ with $\sup_{f \in \mathscr{F}} \| w_f \|_\infty < \infty$.

Lemma 6.1 shows that the remainder term $\| \hat{f} - f \|_{(2, m, \lambda)}^2 = O_p(n^{-1/2})$.

Now

$$T_f(\hat{f} - f) = T_f(\hat{f} - \tilde{f}) + T_f(\tilde{f} - f),$$

where $\tilde{f}$ is as in (6.19).

By the properties of the kernel given in Theorem A.1,

$$\tilde{f}(x) - f(x) = \frac{1}{n\rho(x)} \sum_{i=1}^{n} \left(K_b(x, X_i) f(X_i) - \int_0^1 K_b(x,t) f(t)\rho(t) \right) + O(b^p). \tag{6.20}$$

Applying T_f, we find that $T_f(\tilde{f} - f)$ is the sum of the contribution from the last term above, of order $O(b^p) = O(n^{-1/2})$, plus a simple average of n mean zero, independent terms. In the latter, since a factor of b may be absorbed in

$$\int_0^1 K_b(x, X_i) f(X_i) w_f(x)\, dx,$$

we see that the variance of the summands may be bounded independently of n. Therefore this average is seen to be $O_p(n^{-1/2})$ by (3.15).

It remains to consider T_f applied to

$$\hat{f}(x) - \tilde{f}(x) = \left(\frac{1}{\hat{\rho}(x)} - \frac{1}{\rho(x)} \right) \frac{1}{n} \sum_{i=1}^{n} K_b(x, X_i) f(X_i) + \frac{1}{n\hat{\rho}(x)} \sum_{i=1}^{n} K_b(x, X_i)\varepsilon_i. \tag{6.21}$$

We treat the first sum in (6.21) first. Let

$$e(x) = \frac{1}{n} \sum_{i=1}^{n} K_b(x, X_i) f(X_i) w_f(x)$$

and consider the functional

$$\Gamma(q) = \int_0^1 \frac{e(x)}{q(x)}\, dx.$$

Then T_f applied to the first sum in (6.21) may be written as $(\Gamma(\hat{\rho}) - \Gamma(\rho_b)) + (\Gamma(\rho_b) - \Gamma(\rho))$. Let

$$\Omega_n = \left\{ \omega: \inf_x \hat{\rho}(x) > \frac{\beta}{2}, \|\rho - \rho_b\|_\infty < \frac{\beta}{3}, \|\hat{\rho} - \rho_b\|_\infty < \frac{\beta}{3}, \|\tilde{f} - f\|_\infty < 1 \right\},$$

where $\beta = \inf_x \rho(x) > 0$ as in condition (2.1.2). By Lemma 6.2 and the triangle inequality, $P(\Omega_n) \to 1$. The following assertions hold on Ω_n. A Taylor expansion of the function $\Gamma(\rho_b + tq)$ about $t = 0$ evaluated at $t = 1$ for $q = \hat{\rho} - \rho_b$ yields

$$\Gamma(\hat{\rho}) - \Gamma(\rho_b) = -\int_0^1 \frac{\hat{\rho} - \rho_b}{\rho_b^2} e + \int_0^1 \frac{(\hat{\rho} - \rho_b)^2}{\{\rho_b + s(\hat{\rho} - \rho_b)\}^3} e$$

for some $s \in [0, 1]$. On Ω_n the denominator in the second term is bounded away from 0 and $\|\tilde{f}\|_\infty$ and therefore $\|e\|_\infty$ is bounded. Therefore the second

term above is of the order $\|\hat{\rho} - \rho_b\|_2^2 = O_p(n^{-1/2})$ by Lemma 6.1, Theorem A.1 and the triangle inequality.

The first term is equal to

$$\int_0^1 (\hat{\rho} - \rho_b)(\rho\tilde{f} - (\rho f)_b)\frac{w_f}{\rho_b^2} + \int_0^1 (\hat{\rho} - \rho_b)(\rho f)_b \frac{w_f}{\rho_b^2}. \tag{6.22}$$

Consider the first integral. Note that w_f and $1/\rho_b^2$ are bounded on Ω_n and apply the Cauchy–Schwarz inequality to the remaining factors. Lemma 6.1 and the triangle inequality show that $\|\hat{\rho} - \rho_b\|_2^2 = O_p(n^{-1/2})$. The function $\rho\tilde{f} - (\rho f)_b$ has expectation 0 at each $x \in [0, 1]$, and variance of the order $O(1/nb)$ uniformly on $[0, 1]$. We may apply Fubini's theorem, Markov's inequality and the constraints on b to conclude that $\|\rho\tilde{f} - (\rho f)_b\|_2^2 = O_p(1/nb) = O_p(n^{-1/2})$. Taking the product of the two L^2 norm bounds shows that the first integral is $O_p(n^{-1/2})$.

The second integrand in (6.22) is a sum of n terms with zero mean and variance which can be bounded independently of n, and hence is of order $n^{-1/2}$.

A simpler argument may be used to handle $\Gamma(\rho) - \Gamma(\rho_b)$, using $b^p = O(n^{-1/2})$.

Finally, consider T_f applied to the last term in (6.21). It suffices to show that

$$T_f\left\{\frac{\mathbf{1}(\Omega_n)}{n\hat{\rho}(x)}\sum_{i=1}^n K_b(x, X_i)\varepsilon_i\right\} = \frac{1}{n}\sum_{i=1}^n\left[\int_0^1 \mathbf{1}(\Omega_n)\frac{K_b(x, X_i)}{\hat{\rho}(x)}w_f(x)\,dx\right]\varepsilon_i = O_p(n^{-1/2}).$$

Conditioning on $X_1, \ldots, X_n$ shows the terms of the sum to have mean 0; the conditional variance formula and a change of variable now shows that each has variance bounded independently on n. The claim now follows from Chebyshev's inequality. □

APPENDIX

In this section we present a general formula for and relevant properties of the kernel of order $p \geq 2$. Proofs and discussion may be found in Messer and Goldstein (1992).

A.1. The general formula. For $0 \leq j \leq 2p - 1$, let the $2p$th roots of -1 be given by

$$r_j = \exp\left(\frac{i\pi(2j + 1)}{2p}\right)$$

and define the $p \times 1$ column vector

$$\psi(t) = \langle e^{itr_0}, \ldots, e^{itr_{p-1}}\rangle'.$$

We denote the components of $\psi(t)$ by ψ_j, $0 \leq j \leq p - 1$.

Let $\phi(t) = \langle \phi_p(t), \ldots, \phi_{2p-1}(t)\rangle' = C\psi(t)$, where the $p \times p$ matrix C is

$$C = L^{-1}\Lambda^{-p},$$

with $\Lambda = i\,\mathrm{diag}\langle r_0, \ldots, r_{p-1}\rangle$ and L the Vandermonde matrix

$$L = [\mathbf{1}, \Lambda\mathbf{1}, \ldots, \Lambda^{p-1}\mathbf{1}].$$

Here $\mathbf{1} = \langle 1, \ldots, 1\rangle'$.

Let $k(t)$ be the real function

$$k(t) = \frac{-1}{2p} \sum_{j=0}^{p-1}{}^{\cdot} i r_j \psi_j(|t|). \tag{A.23}$$

The kernel $K_b(x,t)$ is given by

$$\begin{aligned} K_b(x,t) = {} & \frac{1}{b}k\left(\frac{x-t}{b}\right) + \frac{1}{b}\sum_{j=1}^{p}(-1)^{j+1}\left\{\phi_{2p-j}\left(\frac{t}{b}\right)k^{(2p-j)}\left(\frac{x}{b}\right)\right. \\ & \left. + \phi_{2p-j}\left(\frac{1-t}{b}\right)k^{(2p-j)}\left(\frac{1-x}{b}\right)\right\}. \end{aligned} \tag{A.24}$$

When $p = 2$ this may be written more simply. Let $\Phi(u,v) = e^{-u}(\cos(u) - \sin(u) + 2\cos(v))$. Then

$$\begin{aligned} & 2^{3/2}bK_b(2^{3/2}bx, 2^{3/2}bt) \\ & \quad = e^{-|x-t|}(\sin(|x-t|) + \cos(x-t)) \\ & \qquad + \Phi(x+t, x-t) + \Phi((1-x)+(1-t), (1-x)-(1-t)). \end{aligned}$$

A.2. Properties. The following theorem establishes a bound in b on the asymptotic bias of the kernel estimator. Notice that the bias bound is independent of x for $x \in [0,1]$. Hence there is no boundary bias, to first order.

THEOREM A.1. *For the kernel $K_b(x,t)$ as given in* (A.24),

$$\left|\int_0^1 K_b^{(j,0)}(x,t)\,f(t)\,dt - f^{(j)}(x)\right| \le C(p,j)b^{p-j}\|f\|_{(\infty,p)}$$

for all $b > 0$, $f \in C^p[0,1]$ and $0 \le j \le p$.

The following proposition allows us to differentiate under the integral and to establish various bounds. Let $k(x;b) = b^{-1}k(xb^{-1})$.

PROPOSITION A.2.

$$|k^{(j)}(x;b)| \le C(j,p)b^{-(j+1)}e^{-\sin(\pi/2p)|x/b|}$$

for all $j \ge 0$. A similar bound holds for $(b^{-1}\phi_l(t/b))^{(j)}$.

REFERENCES

ABRAMSON, I. and GOLDSTEIN, L. (1991). Efficient testing by nonparametric functional estimation. *J. Theoret. Probab.* **4** 137–159.

BICKEL, P. and RITOV, J. (1988). Estimating integrated squared density derivatives: Sharp best order of convergence estimates. *Sankhyā Ser. A* **50** 381–393.

DONOHO, D. L. (1988). One-sided inference about functionals of a density. *Ann. Statist.* **16** 1390–1420.

DONOHO, D. L. and LIU, R. C. (1991). Geometrizing rates of convergence. II. *Ann. Statist.* **19** 633–667.

DONOHO, D. L. and LOW, M. G. (1992). Renormalization exponents and optimal pointwise rates of convergence. *Ann. Statist.* **20** 944–970.

FENSTAD, G. and SKOVLUND, E. (1990). A two-sample sequential rank test by Sen investigated by stochastic simulation. *J. Statist. Comput. Simulation* **36** 129–137.

FERNHOLZ, L. T. (1983). *Von Mises' Calculus for Statistical Functionals*. Springer, New York.

GOLDSTEIN, L. and KHAS'MINSKII, R. Z. (1992). On efficient estimation of smooth functionals. Preprint.

HALL, P. and MARRON, J. (1987). Estimation of integrated squared density derivatives. *Statist. Probab. Lett.* **6** 109–115.

HAS'MINSKII, R. Z. and IBRAGIMOV, I. A. (1979). On the nonparametrical estimation of functionals. In *Proceedings of the Second Prague Symposium on Asymptotic Statistics* (P. Mandl and M. Hušková, eds.) 41–55. North-Holland, Amsterdam.

LEHMANN, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco.

LEVIT, B. (1979). Asymptotically efficient estimation of nonlinear functionals. *Problems Inform. Transmission* **14** 65–72.

MESSER, K. and GOLDSTEIN, L. (1992). A new class of kernels for nonparametric curve estimation. *Ann. Statist.* To appear.

NADARAYA, E. Z. (1990). *Nonparametric Estimation of Probability Densities and Regression Curves*. Kluwer, London.

PFANZAGL, J. (1985). *Contributions to a General Asymptotic Statistical Theory. Lecture Notes in Statist.* **13**. Springer, New York.

RITOV, Y. and BICKEL, P. J. (1990). Achieving information bounds in non and semiparametric models. *Ann. Statist.* **18** 925–938.

SEN, P. K. (1981). *Sequential Nonparametrics: Invariance Principles and Statistical Inference*. Wiley, New York.

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

VON MISES, R. (1947). On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Statist.* **18** 309–348.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF SOUTHERN CALIFORNIA
LOS ANGELES, CALIFORNIA 90089-1113

DEPARTMENT OF MATHEMATICS
CALIFORNIA STATE UNIVERSITY
FULLERTON, CALIFORNIA 92634