OPTIMAL PLUG-IN ESTIMATORS FOR
NON-PARAMETRIC FUNCTIONAL ESTIMATION

by

L. Goldstein and K. Messer

Technical Report No. 277
August 1990

Prepared Under the Auspices
of
National Science Foundation Grant
DMS 87-08083
Ingram Olkin, Project Director

Department of Statistics
Stanford University
Stanford, California
OPTIMAL PLUG-IN ESTIMATORS FOR
NON-PARAMETRIC FUNCTIONAL ESTIMATION

by

L. Goldstein and K. Messer
University of Southern California and California State University, Hayward

Technical Report No. 277

Prepared Under the Auspices
of
National Science Foundation Grant
DMS 87-08083
Ingram Olkin, Project Director

Department of Statistics
Stanford University
Stanford, California
Optimal plug-in estimators for non-parametric functional estimation

L. Goldstein*
Department of Mathematics
University of Southern California

K. Messer†
Department of Mathematics
California State University, Fullerton

August 3, 1990

*Research partially supported by NSF Grant DMS 90-05833.
†Research partially supported by NSF Grant DMS 87-08083.
Abstract

A simple functional calculus can be used to compute optimal rates of convergence for a broad class of linear and non-linear functionals, in the sense of Stone (1980). We produce a straightforward “plug-in” kernel estimator which achieves the optimal rate of convergence over bounded subsets of a Sobolev space. For many functionals the rate calculation depends on a Fréchet-like derivative of the functional, which may be obtained using elementary calculus. We give examples including $L^q$ norms of regression or density functions and their derivatives, the expected integrated squared bias, and the negentropy of a density.

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


Key words and phrases: Nonparametric regression, functionals, optimal rates, plug in estimators.
1 Introduction

In a regression or density estimation setting one is often interested in the value of a functional $\Lambda$ 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 $f(x) = E[Y \mid X = x]$; 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. The $X$’s are assumed to 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

$$\Lambda(f) = f^{(r)}(x_0)$$  \hspace{1cm} (1.1)

for some $r$. Other examples include the mean squared error of an estimate of $f^{(r)}(x)$,

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

integrated mean squared error

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

or the $L^2$ norm of the unknown function $f$ or the $r$th derivative of $f$

$$\Lambda(f) = \left\{ \int (f^{(r)}(x))^2 \right\}^{\frac{1}{2}},$$  \hspace{1cm} (1.2)

these last being of interest for the choice of smoothing parameter. The $L^2$ norm is of additional importance in the density estimation setting, as for example, it appears in both 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 pp. 362,363). Of interest as well in the density estimation setting is the Fisher information of a location parameter $\theta$ in the location family $f_\theta(x) = f(x - \theta)$,

$$\Lambda(f) = \int (f')^2 / f,$$
and the negentropy of $f$

$$\Lambda(f) = -\int f \log f.$$ 

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:

1. What is the theoretical upper bound on the rate of convergence of an estimate $\hat{\Lambda}$ of $\Lambda(f)$?

2. Is there a convenient class of estimators, in particular plug-in estimators, that achieve this upper bound?

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

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

where $K_b(x, t)$ is a certain boundary corrected kernel which satisfies a scaling property in $b$. This kernel is studied in Messer and Goldstein (1989). A formula for $K_b(x, t)$ is given in the Appendix below. The bandwidth $b$ is the same in the numerator and denominator of $\hat{f}$.

We take our definition of optimality from Stone (1980), and consider optimality over a class of functions we denote $\mathcal{W}_p$, roughly the class of functions with $p$ continuous derivatives all of which are bounded by a constant. (We dispense with fractional derivatives, as this sort of refinement is not central to our main argument.) We obtain optimality results over subsets $\mathcal{W}$ of $\mathcal{W}_p$ which may satisfy extra conditions. The most common practical examples are functions which satisfy boundary conditions, which are bounded away from 0, or which integrate to 1.

As a simple example, 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 which is uniform over functions \( f \in \mathcal{W}_p \) of the squared error of an estimator of (1.1) is \( n^{-\frac{2r}{p+1}} \). On the other hand, for a "smooth" functional such as \( \int f(x)^2 \, dx \), it has long been known that the parametric rate of \( n^{-\frac{1}{2}} \) is achievable. For the functional given by (1.2), the same rate \( n^{-1/2} \) is achievable under certain circumstances. The purpose of this paper is to provide an easy method by which functionals may be divided into one of these two groups (a slow "pointwise" group, and a fast "smooth" group), to provide a convenient "plug-in" kernel estimator which will achieve the optimal rate, and to determine the rate of the optimal bandwidth for that estimator.

Our method uses a straightforward Taylor series expansion of the functional \( \Lambda \) 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 Fernholtz (1983). The derivative we use may often be computed using elementary calculus; we give examples below. The remainder term in the expansion is shown to be of smaller order in probability than the leading linear term. Investigation of the linear term then yields the optimal rate for estimating \( \Lambda \) over the class of functions \( \mathcal{W}_p \).

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 number \( r \) which appears in (1.1) then determines the optimal rate \( n^{-\frac{2r}{p+1}} \) achievable over \( \mathcal{W}_p \). The rate is achieved by our "plug-in" estimator so long as the appropriate bandwidth is chosen. We call \( r \) the index of the functional; in this setting it may depend on the boundary behavior of \( f \). These results are given in Theorem 4.1.

Another class of functionals which we call smooth has derivative which is given by integrating against a well-behaved weight function. Smooth functionals will be seen to be estimable at rate \( n^{-1/2} \) over \( \mathcal{W}_p \), using our "plug in" kernel estimator with an undersmoothed bandwidth.

In some settings, there are many functionals whose character will vary depending on the class of functions \( \mathcal{W} \subset \mathcal{W}_p \) under consideration. For example, we consider subsets \( \mathcal{W} \) of \( \mathcal{W}_p \) which satisfy certain boundary restrictions. For \( f \) in \( \mathcal{W} \), the linear term \( T_f \) in the Taylor expansion may be smooth, whereas for \( f \not\in \mathcal{W} \), \( T_f \) may have an atomic component. This sort of behavior usually arises as a result of an integration by parts. Hence a better uniform rate of
convergence may obtain over $\mathcal{W}$ than over the full space $\mathcal{W}_p$. If $f$ is known to lie in $\mathcal{W}$, one may take advantage of the extra information, and find an estimator which achieves this better rate. In many cases the plug in estimator with a different choice of bandwidth will do the job. This is surprising, because the estimate $\hat{f}$ of $f$ will not usually lie in the restricted set $\mathcal{W}$. For functionals which are smooth over $\mathcal{W}$ but not necessarily over $\mathcal{W}_p$, the plug in estimator with an undersmoothed bandwidth will almost always be optimal (under very mild conditions on $\mathcal{W}$). Results for smooth functionals are presented in Theorem 4.2.

Similarly, one may consider functionals which are atomic of different index over $\mathcal{W}$ than $\mathcal{W}_p$. For many such functionals, the plug in estimator is still optimal. For some naturally occurring kinds of degeneracy, however, the plug in estimator cannot achieve the optimal rate. For example, $\Lambda(f) = \int f'^2$ is of the former kind, and the plug-in estimator is optimal; $\Lambda(f) = \int f'g$ for a fixed function $g$ is of the latter kind, and the plug-in estimator is not optimal. We provide conditions on the functional $\Lambda$ and the function class $\mathcal{W}$ which characterize when a plug-in estimator is and is not optimal, for this kind of functional. When a plug in estimate is not optimal, we discuss an appropriate modification. A discussion of this kind of behavior is provided in example 7.

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 latter case undersmoothing is advantageous. Consider $\Lambda(f) = \int_0^1 (f')^2$. We have the expansion

$$\Lambda(f + h) = \Lambda(f) + T_f(h) + O(\|h\|_{2,1}^2),$$

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

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

Hence $T_f$ will determine the rate of convergence of $\Lambda(\hat{f})$ to $\Lambda(f)$ if the remainder term is negligible. This rate will depend on whether $T_f$ is smooth or atomic. Investigating $T_f$, we have

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

$$= -\int_0^1 h f'' + f'(0)h(0) - f'(1)h(1).$$

6
Heuristically, estimating the functional should be at least as hard as estimating \( f''(0) \). That is, \( T_f \) is atomic: its worst behavior for \( f \in \mathcal{W}_p \) produces a point evaluation component of index 0, and hence the best rate obtainable over \( \mathcal{W}_p \) is the slow pointwise rate \( n^{-p/(2p+1)} \). Theorem 4.1 shows that the plug in estimator \( \Lambda(\hat{f}) \) achieves this rate over \( \mathcal{W}_p \), and that this rate is optimal.

On the other hand, suppose we know that \( f \in \mathcal{W} \) where

\[
\mathcal{W} = \{ f \in \mathcal{W}_p : f(0) = f(1) = 0 \}.
\]

In this case we have the representation,

\[
T_f(h) = -\int_0^1 h f'' \quad \text{for } f \in \mathcal{W},
\]

and so \( \Lambda \) is smooth over \( \mathcal{W} \). Theorem 4.2 then says the plug in estimator \( \Lambda(\hat{f}) \) with an "undersmoothed" bandwidth achieves the optimal rate \( n^{-1/2} \) over \( \mathcal{W} \).

In the smooth case, we may almost always undersmooth and achieve the rate of convergence \( n^{-1/2} \). To see why, we consider the plug in estimator in more detail. Suppose for convenience that the data are equally spaced, and we use a simple translation kernel:

\[
\hat{f}(x) = \frac{1}{n} \sum Y_i \frac{1}{b} K\left( \frac{x - \frac{i}{n}}{b} \right).
\]

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

\[
T_f(\hat{f} - f) = \int \left\{ \frac{1}{n} \sum f(\frac{i}{n}) \frac{1}{b} K\left( \frac{x - \frac{i}{n}}{b} \right) - f(x) \right\} w_f(x) \, dx
\]

\[
+ \frac{1}{n} \sum \epsilon_i \int \frac{1}{b} K\left( \frac{x - \frac{i}{n}}{b} \right) w_f(x) \, dx.
\]

The first sum is not stochastic, and by standard properties of the kernel is \( O(b^p) \) uniformly in \( x \), under certain conditions. The second term, in which we have interchanged summation and integration, is a sum of independent mean zero terms whose variance may be bounded independently of \( b \), as \( b \) may be absorbed by a change of variable in the integral. Hence the second
term is $O_p(n^{-1/2})$, and we need only insure that $b^p \sim n^{-1/2}$ in order to achieve the desired rate.

We consider that we have "undersmoothed" by comparison with the usual bandwidth which is optimal for estimating $f$. To see this, consider the expression for $\hat{f}(x) - f(x)$ given by dropping the integrals and $w_f$ from the expression above. Then the first term is $O(b^p)$ as before, but the second term is now $O(1/\sqrt{n\delta})$. The optimal bandwidth becomes $b \sim n^{-1/(2p+1)}$, which gives the optimal rate $n^{-p/(2p+1)}$. By comparison, in the smooth case we choose $b$ of order $n^{-1/2p}$. A smooth functional does some of the smoothing itself; we are free to choose a smaller bandwidth and decrease the bias of the estimate, as the variance of the first order term is independent of $b$. Later we shall see that the second order term of the expansion will provide a lower bound for $b$.

As mentioned above, the expansion we consider for functionals of regression or density functions is similar to those of Von-Mises (1947) and Fernholz (1983) for distribution functions. Hasminskij and Ibragimov (1978) and Pfanzagl (1982) have also studied estimating functionals of regression or density functions. 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; the class of functions he considers is much larger than that considered here and optimality of rates is not discussed. Under certain "renormalization" conditions, Donoho and Low (1990) obtain general optimality results for linear functionals on classes of functions themselves specified by conditions on functionals.

The remainder of the paper is organized as follows: Section 2 presents the notation and hypotheses used throughout the paper. In Section 3 we define the class of functionals which 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 give examples. Proofs are in Section 6, and a formula for the kernel and its properties is in the Appendix.
2 Notation and hypotheses

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 $1 \leq q \leq \infty$ let

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

(2.4)

where $\lambda$ is a vector of measures $(\lambda_1, \ldots, \lambda_s)$. When $\lambda$ is the vector of Lebesgue measures, we shall drop the dependence on $\lambda$ and write $\|f\|_{(q, s)}$. Here $f^{(j)}$ denotes the $j^{th}$ derivative of $f$.

Let $p \geq 2$ be a nonnegative integer. The kernel $K_b(x, t)$ of order $p$ we use below is given by (A.23) of the Appendix. $K_b^{(i,j)}(x, t)$ will denote the $i, j^{th}$ mixed partial derivative in $x$ and $t$. For a function $h$,

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

(2.5)

The regression and density functions considered below will be assumed to lie in a subset $\mathcal{W}$ of the set

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

(2.6)

for $M$ a constant fixed throughout.

The functionals $\Lambda$ we consider are real valued functions defined on $\mathcal{W}_p$, that is

$$\Lambda : \mathcal{W}_p \to \mathbb{R}.$$

All estimators of $\Lambda(f)$ considered below are of the form $\hat{\Lambda}_n = \Lambda(\hat{f}_n)$, where $\hat{f}_n$ is the following kernel estimate of $f$ using kernel $K_b(x, t)$ from (A.23), and $b = b_n$:

1. In the regression case,

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

(2.7)

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

\[ \hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x, x_i), \]

(2.8)

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

We construct estimators that are optimal over \( \mathcal{W} \) in the sense of Stone (1980). Let \( \Lambda \) be a functional on \( \mathcal{W}_p \) 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 \( f \in \mathcal{W} \). A positive number \( \gamma \) is called an upper bound to the rate of convergence over \( \mathcal{W} \) if for every sequence \( \{ \hat{\Lambda}_n \} \) of estimators

\[ \liminf_{n} \sup_{f \in \mathcal{W}} P(|\hat{\Lambda}_n - \Lambda(f)| > cn^{-\gamma}) > 0 \]

(2.9)

for all \( c > 0 \) and

\[ \lim_{c \to 0} \liminf_{n} \sup_{f \in \mathcal{W}} P(|\hat{\Lambda}_n - \Lambda(f)| > cn^{-\gamma}) = 1. \]

(2.10)

Also \( \gamma \) is called an achievable rate of convergence over \( \mathcal{W} \) if there is a sequence \( \{ \hat{\Lambda}_n \} \) of estimators such that

\[ \lim_{c \to 0} \limsup_{n} \sup_{f \in \mathcal{W}} P(|\hat{\Lambda}_n - \Lambda(f)| > cn^{-\gamma}) = 0; \]

(2.11)

The number \( \gamma \) is called the optimal rate of convergence over \( \mathcal{W} \) if it is both an upper bound to the rate of convergence and an achievable rate of convergence.

We now state the assumptions of our models.

**Assumption 2.1** In the density estimation case, we observe \( X_1, X_2, \ldots, X_n \) independent with density \( f \in \mathcal{W} \). In the atomic case, Theorem 4.1, we assume further that \( \inf_{x \in [0,1]} f(x) > 0 \).

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

**Assumption 2.2** 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 \(\mathcal{W}\).

2. The distribution of \(X\) is absolutely continuous with respect to Lebesgue measure on \([0, 1]\), with density \(p \in C^p[0, 1]\), and there exist constants \(\beta \) and \(\eta\) 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 \(X\), \(\text{Var}[Y|X = x] = \sigma^2(x)\) satisfies \(0 < \kappa < \sigma(x) < \zeta < \infty\).

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 \(\delta\) depending only on \(M\) of equation (2.6) (and hence, in particular, which may be chosen independently of \(f \in \mathcal{W}\)) such that

\[|t| \leq C|q| \text{ for all } q \text{ with } |q| \leq \delta.\]

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

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

In what follows \(p\) will denote the degree of smoothness of \(f\) and the order of the kernel \(K_b(x, t)\), \(m\) the order of the highest derivative of \(f\) that appears in the remainder term in an expansion of \(\Lambda\), \(r\) the "index" of \(\Lambda\), and \(s\) a dummy variable only. \(C\) will denote a positive constant which is not necessarily the same at each occurrence. Integrals without explicit limits will be taken over the unit interval. We often will write \(b\) for \(b_n\).

3 Classification and Differentiability of Functionals.

In this section we give the formal definitions of atomic and smooth, first for bounded linear functionals \(T\) on \(\mathcal{W}_p\), and then for the class of differentiable
functionals $\Lambda$ which we study. Section 3.1 presents a Reisz-type representation for bounded linear functionals $T$ and defines atomic and smooth. 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 check differentiability and to classify functionals as atomic or smooth.

3.1 Classification

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

\[ |||T|||_p = \sup_{||h||_{(\infty,p)} \leq 1} |Th|. \]

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

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

\[ \Lambda(f + h) = \Lambda(f) + T_f(h) + O(||h||^2_{(2,m,\lambda)}), \quad (3.12) \]

where $T_f$, the derivative of $\Lambda$ at $f$, is a bounded linear functional on $C^p[0,1]$, $m$ is an integer $0 \leq m < p$, and $\lambda$ is a vector of finite measures that may depend on $\Lambda$ 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

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

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

We now define the notions of atomic, index and smooth for a bounded linear functional $T$.

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

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

12
where \( \mu_s \) has a discrete component \( \delta_s \), then we say that \( T \) is atomic. We define \( r \), the index of \( T \) to be the largest such \( s \) that can appear in the above representation.

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

**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 \( \Lambda \) has smooth derivative at \( f \) we say \( \Lambda \) is smooth at \( f \), and that \( \Lambda \) is smooth on \( \mathcal{W} \) if \( T_f \) is smooth for all \( f \in \mathcal{W} \).

Note that smooth is not the same as index 0.

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

**Definition 3.3** Given \( \mathcal{W} \), define \( \mathcal{W}^* \), the space of local variations for \( \mathcal{W} \) by

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

We state the following degeneracy conditions which will be used to rule out exceptional cases. Basically, the first says that degeneracy exists if the rate determining measure is annihilated on the set of local variations. The second will be used to exclude "thin" classes of functions by insuring that \( \mathcal{W} \) is a rich enough set so that \( \mathcal{W}^* \) contains small perturbations ("bump functions"). We shall prove that when \( \Lambda \) is not degenerate, a plug-in estimator is optimal. Any non-constant functional on \( \mathcal{W} = \mathcal{W}_p \) is automatically not degenerate.

**Definition 3.4** Let \( \Lambda \) be a functional of index \( r \) on \( \mathcal{W} \). We say \( (\Lambda, \mathcal{W}) \) is strongly degenerate if for all \( f \in \mathcal{W} \) such that \( T_f \) is of index \( r \)

\[
\int_0^1 h^{(r)}d\delta_{r,f} = 0 \text{ for all } h \in \mathcal{W}^*.
\]
We say \((\Lambda, \mathcal{W})\) is not degenerate if \(\Lambda\) is of index \(r\), and for all \(x_0 \in [0, 1]\), there is a function \(\psi\) of compact support, with \(\psi^{(r)}(0) \neq 0\) such that
\[
a^{-r} \psi(a(x - x_0))1(x \in [0, 1]) \in \mathcal{W}^*
\] (3.13)
for all a sufficiently large.

It is clear that \((\Lambda, \mathcal{W})\) is degenerate whenever it is strongly degenerate. The degeneracy condition (3.13) rules out finite dimensional classes of functions. For example take \(\mathcal{W}\) to be all polynomials of degree less than or equal to \(k\). Then any atomic functional \(\Lambda\) is degenerate, but not necessarily strongly degenerate, on \(\mathcal{W}\): for example take \(\Lambda(f) = f^{(r)}(0)\) with \(r \leq k + 1\). That this pair \((\Lambda, \mathcal{W})\) 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 of \(n^{-1/2}\).

For a discussion of strong degeneracy, see example 7.

**Definition 3.5** Let \(\Lambda\) be a smooth functional on \(\mathcal{W}\). We say \((\Lambda, \mathcal{W})\) is degenerate if for all \(f \in \mathcal{W}\)
\[
\int_0^1 h \omega_f = 0 \quad \text{for all } h \in \mathcal{W}^*.
\]

For example, \(\Lambda(f) = \int_A f\) is degenerate on \(\mathcal{W} = \{f : f 1_A = 0\}\).

### 3.2 Differentiability of functionals

Definition 3.6 below gives our notion of differentiability. The Propositions which follow may be used to verify that a given functional is sufficiently differentiable.

**Definition 3.6** Let \(0 \leq m < p\).

1. We say \(\Lambda\) is differentiable of order \(m\) on \(\mathcal{W} \subseteq \mathcal{W}_p\) if for every \(f \in \mathcal{W}\), \(\Lambda\) has an expansion of the form (3.12) and
\[
\sup_{f \in \mathcal{W}} \|\|T_f\|\|_p < \infty.
\]
2. We say $\Lambda$ is smooth of order $m$ on $\mathcal{W} \subseteq \mathcal{W}_p$ if $\Lambda$ is differentiable of order $m$ on $\mathcal{W}$ and for all $f \in \mathcal{W}$, $\Lambda$ has smooth derivative $T_f$ with

$w = w_f$ and

$$\sup_{f \in \mathcal{W}} \|w_f\|_\infty < \infty.$$  \hfill (3.14)

In many examples, $\Lambda$ is of the special form

$$\Lambda(f) = \int_0^1 g \left( u, f^{(0)}(u), f^{(1)}(u), \ldots, f^{(m)}(u) \right) \, du \quad \hfill (3.15)$$

for some $g : \mathbb{R}^{m+2} \to \mathbb{R}$.

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

**Proposition 3.7** Let $\Lambda$ be given by (3.15) for $0 \leq m < p$. Let $W$ be an open set in $\mathbb{R}^{m+2}$ such that

$$\left\{ (u, f^{(0)}(u), \ldots, f^{(m)}(u)) : u \in [0,1], f \in \mathcal{W} \right\} \subseteq W.$$

Suppose $g$ has two continuous derivatives on the convex closure of $W$. Then $\Lambda$ is differentiable of order $m$ on $\mathcal{W}$.

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

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

**Proposition 3.9** Let $\Lambda$ be differentiable of order $m$ and [of index $r$/smooth] on $\mathcal{W}$ and suppose that $g$ is twice continuously differentiable on $I$, a bounded, closed interval containing $\{ \Lambda(f) : f \in \mathcal{W} \}$ with $g'(x) \neq 0$ on $I$. Then $\Gamma(f) = g(\Lambda(f))$ is differentiable of order $m$ on $\mathcal{W}$, and [of index $r$/smooth].

We state one last result to be applied to example 5 involving the asymptotic variance of $U$-statistics; there, the case $d = 3$ will apply.

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

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

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

15
**Proof:** The proof is a consequence of the relations

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

and

\[ \Lambda(f + h) - \Lambda(f) - T_f(h) = \int \cdots \int_A \psi(x) \sum_{i=2}^{d} \prod_{|J|=i, J \in J} h(x_J) f(x_k) \, dx \]

\[ \leq C \left[ \int_0^1 |h(x)| \, dx \right]^2 \leq C \|h\|_1^2 \leq C \|h\|_2, \]

for \( \|h\|_\infty \leq 1 \).

4 **Statement of Results**

For atomic functionals we have the following result.

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

For smooth functionals:

**Theorem 4.2** Let \( \Lambda \) be a smooth functional of order \( m \) on \( W \subseteq W_p \), and \( p \geq \max\{2, 2m+1\} \). Let \( c \) be a positive constant, and let \( b_n \in [cn^{-\frac{1}{m+2}}, cn^{-\frac{1}{p+1}}] \). If \( (\Lambda, W) \) is not degenerate, then \( \Lambda(f_n) \) achieves the optimal rate \( n^{-1/2} \) on \( W \).

5 **Examples.**

Here we present some commonly occurring examples of differentiable functionals. In the first six examples, the natural subclasses of functions \( W \) we consider satisfy the conditions of Theorem 4.1. Hence for these examples plug-in estimators are always optimal, even over restricted classes of functions in the atomic case. The last example is for the sake of illustration. It is
a case for which the plug-in estimator is not optimal over a certain restricted function class. The functional is strongly degenerate over the class. In the example we show a modification to the estimator may be constructed so as to achieve the optimal rate.

1. We recover the results about optimal pointwise estimation of regression and density functions in Stone (1980). By considering the point evaluation functional

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

as a differentiable functional of index $r$, we see that $\Lambda$ is estimable at the optimal rate $n^{-(p-r)/(2p+1)}$.

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

Differentiating the function

$$\gamma(t) = \int_0^1 [(f + th)^{(m)}]^q$$

with respect to $t$, we see that the derivative $T_f$ of $\Lambda$ at $f$ is given by

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

If $m = 0$ this demonstrates that $\Lambda$ is smooth. However, $\Lambda$ may be smooth even when $m \geq 1$ for certain $W$. For example, take $q = 2$, and integrate by parts.

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

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

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

$$f^{(j)}(0) = f^{(j)}(1) = 0, \quad m \leq j \leq 2m - 1,$$ 

17
then all atomic terms in $T_f$ drop out in an integration by parts. In fact,

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

hence $T_f$ is smooth. In this case the functional is estimable at rate $n^{-1/2}$ over $\mathcal{W}$, and the rate is achieved by the undersmoothed plug-in estimator. Generally, if $f$ lies in the class of functions $\mathcal{W}^m$ 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 $\Lambda$ 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.

The functional of this example is the subject of Hall and Marron (1987) and Bickel and Ritov (1988). All these authors deal with the case of a density function on the entire real line. These authors have investigated 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. 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 of $n^{-1/2}$.

3. Consider the negentropy functional $\Lambda(f) = -\int_0^1 f \log f$ on $\mathcal{W} = \{f \in \mathcal{W}_p : \inf_{x \in [0,1]} f(x) > \delta\}$ for some $\delta > 0$. This functional is smooth of order 0 on $\mathcal{W}$ by Proposition 3.8, for $p \geq 1$. Hence it is estimable on $\mathcal{W}$ at the optimal rate of $n^{-1/2}$, by an undersmoothed plug-in estimator.

4. Abramson and Goldstein (1991) study the equidistribution functional $\Delta$ of two densities $f$ and $g$,

$$\Delta[f, g] = 2 \int_0^1 \frac{f(u)g(u)}{f(u) + g(u)} du.$$

Aside from being invariant under smooth invertible transformations of the data, the functional enjoys the following property: $\Delta = 1$ if and only if $f = g$; if $f \neq g$ then $\Delta < 1$. In the one sample problem to test whether or not observed data is generated from the known density $g$, it is therefore of interest to estimate $\Delta$. When $\inf_{x} g(x) > 0$, Proposition 3.8 shows that $\Delta$ is smooth of order 0 on $\mathcal{W}_p$; hence $n^{-1/2}$ is optimal.
5. Consider the $U$-statistic
\[ U = \binom{n}{2}^{-1} \sum_{i<j} \psi(X_i, X_j), \]
where $X_1, \ldots, X_n$ are i.i.d. with density $f$ on $[0,1]$ and $\psi$ is a bounded measurable function; take $EU = 0$ for convenience. Then
\[ \Lambda(f) = \int_0^1 [\int_0^1 \psi(y, x)f(y)dy]^2f(x)dx \]
is the asymptotic variance of $U$ (Lehmann, 1975 page 367). That $\Lambda(f)$ is smooth of order 0 on $W_p$ follows from Propostion 3.10 for $d = 3$, and is therefore estimable at the optimal rate $n^{-1/2}$ by Theorem 4.2.

6. Consider the asymptotic integrated mean squared error $E\int (\hat{f} - f)^2$, and in particular the bias term
\[ \frac{1}{n\hat{\rho}} \sum K_b^{(m,0)}(x, X_i)f(X_i) - f^{(m)}(x). \]
For convenience, we shall take $\rho(x) \equiv 1$, and omit $\hat{\rho}$ from the above; the general case is similar but algebraically more complicated. Under these conditions, taking the expectation of the above gives $\int f_b^{(m)} - f^{(m)}$, where $f_b^{(m)}$ is as in (2.5). Squaring and integrating (usually against the density $\rho$) gives the expected bias functional
\[ \Lambda(f) = \int_0^1 (f_b^{(m)}(x) - f^{(m)}(x))^2 dx. \]
A variational argument shows $\Lambda$ to be differentiable of order $m$ on $W_p$ for $p > m$, with derivative
\[ T_f(h) = 2 \int_0^1 (f_b^{(m)}(x) - f^{(m)}(x))(h_b^{(m)}(x) - h^{(m)}(x)) dx. \]

Integrating by parts as in the previous example, we see that $T_f$ is of index $m - 1$ over $W_p$. 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 $W_p$, for $p \geq 2m - (m - 1) = m + 1$. This rate is achieved by the plug in estimator
\[ \int_0^1 (\hat{f}_b^{(m)}(x) - \hat{f}^{(m)}(x))^2 dx. \]
As in example 2, if we restrict attention to the proper subset \( \mathcal{W} \) of \( \mathcal{W}_p \) of functions which satisfy the appropriate boundary conditions, then the optimal rate of \( n^{-1/2} \) is achieved by the undersmoothed plug-in estimator by Theorem 4.2. Finally, Theorem 4.1 shows that for appropriate boundary restrictions, the plug-in estimator achieves an optimal intermediate rate.

We remark that the bandwidth used in the estimate \( \tilde{f}^{(m)} \) is not necessarily the same as that used in the functional \( \Lambda \). If the bandwidth in \( \Lambda \) is chosen to estimate \( f^{(m)} \) optimally, it will in general be of the same order in \( n \), but larger than, the optimal bandwidth for the plug in estimator \( \tilde{f}^{(m)} \). The latter will be the optimal bandwidth for estimating \( f^{(m-1)} \), as can be seen by the usual expansion of \( \Lambda \). Hence in order to estimate the integrated squared bias for a given bandwidth \( b \), one undersmooths relative to \( b \) and plugs the resulting estimate into the functional.

Finally, note that if \( \rho(x) \) is unknown the integrals appearing in \( \Lambda \) may be approximated by \( 1/n \) times the sum over the observations \( X_i \) without perturbing the order of convergence of the estimate.

7. This example is chosen to illustrate a case for which the plug-in estimator is not optimal over a certain restricted function class. We show how to construct an appropriate modified estimator. Let

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

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

\[
T_f(h) = \int_0^1 h'g = -\int_0^1 hg' + (h(1) - h(0)).
\]  

Hence \( \Lambda \) is differentiable of order 0 and is atomic of index 0 over \( \mathcal{W}_p \). The optimal rate over \( \mathcal{W}_p \) is the pointwise rate \( n^{-p/(2p+1)} \), and is achieved by the plug in estimator \( \Lambda(\tilde{f}) \), as given by Theorem 4.1.

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

\[
\mathcal{W} = \{ f \in \mathcal{W}_p : f(0) = f(1) \}.
\]

Since the derivative of \( \Lambda \) does not depend on \( f \), \( \Lambda \) is still of index \( r \) over \( \mathcal{W} \). (This is unlike all our previous examples, where the derivative depends on
But notice the functional is strongly degenerate on \( \mathcal{W} \): the atomic part of (5.16) is identically zero for \( h \in \mathcal{W}^* \). Hence the hypothesis of Theorem 4.1 is not satisfied. Indeed, we can achieve the smooth rate of \( n^{-1/2} \) over \( \mathcal{W} \) by considering 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 \( n^{-p/(2p+1)} \). Essentially this is because the estimate \( \hat{f} \) lies in \( \mathcal{W}_p \), not \( \mathcal{W} \), and for this example, in contrast to the previous examples, the difference is crucial. For \( f \in \mathcal{W} \), the higher terms of the derivative \( T_{r}(h) \) do not vanish for all \( h \in \mathcal{W}_p \) but only for \( h \in \mathcal{W}^* \).

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

### 6 Proofs

We give the proofs in this section for the regression setting. The proofs in the density estimation case are similar.

**Proof of theorem 4.1.**

**Upper bound.**

Since \( \Lambda \) is of index \( r \) we may choose \( f_0 \in \mathcal{W} \) such that \( T_{f_0} \) has a representation for \( h \in C^p[0,1] \)

\[
T_{f_0}(h) = \sum_{j=0}^{r} \int_0^1 h^{(j)}(d\mu_j),
\]

where without loss of generality \( \mu_0(\{x_0\}) = \eta > 0 \), for \( x_0 \in [0,1] \). We argue as in Stone (1980). Since \( \Lambda \) is not degenerate on \( \mathcal{W} \), we may choose \( \psi \) of compact support so that it satisfies the rescaling condition (3.13), taking \( \psi^{(r)}(0) > 0 \) without loss of generality.

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

\[
g_n(x) = \delta N^p n^{-\tau} \psi(N^{-1}n^{\tau}(x - x_0))1(x \in [0,1]).
\]
Let $f_n = f_0 + g_n$. Condition (3.13) gives that $f_n \in \mathcal{W}$ for $n$ sufficiently large.

Equations (2.2) and (2.3) in Stone (1980) follow as in Stone (1980). Put prior probabilities of $1/2$ each on $f_0$ and $f_n$; as in [11], any method of deciding between $f_0$ and $f_n$ based on $(X_1, Y_1), \ldots, (X_n, Y_n)$ must have overall error probability at least $\epsilon$ for some $\epsilon > 0$. When we apply this result to the classifier

$$f_n = \begin{cases} f_0 & \text{if } \hat{\Lambda}_n \leq \frac{\Lambda(f_0) + \Lambda(f_n)}{2} \\ f_n & \text{if } \hat{\Lambda}_n > \frac{\Lambda(f_0) + \Lambda(f_n)}{2} \end{cases}$$

it follows as in Stone (1980) that

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

Using the differentiability of $\Lambda$ we now compute that

$$\frac{\Lambda(f_n) - \Lambda(f_0)}{2} = \frac{1}{2} T(g_n) + O(||g_n||_{L(\mu, \lambda)}).$$

Let $E_n = \text{support } (g_n)$; note that $E_n \downarrow \{x_0\}$ since $\psi$ has compact support. Consider the final term in the sum (6.17):

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

Let the Jordan decomposition of $\mu_r$ be given by $\mu_r = \mu_r^+ - \mu_r^-$. Since $\mu_r^+$ and $\mu_r^-$ 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

$$\mu_r^-(E_n) \leq \frac{\eta \psi^{(r)}(0)}{2\|\psi^{(r)}\|_{\infty}}.$$

Hence, for $n \geq n_0$

$$\int_0^1 g_n^{(r)} d\mu_r = \delta N^{p-r} n^{-(p-r)r} \left[ \int_0^1 \psi^{(r)}(n^r(x - x_0))d\mu_r^+ - \int_0^1 \psi^{(r)}(n^r(x - x_0))d\mu_r^- \right]$$

$$\geq \delta N^{p-r} n^{-(p-r)r} \left[ \eta \psi^{(r)}(0) - \mu_r^-(E_n)\|\psi^{(r)}\|_{\infty} \right] \geq \frac{1}{2} \delta N^{p-r} n^{-(p-r)r} \eta \psi^{(r)}(0).$$
The remaining terms in the sum (6.17) with $j < r$ and $O(\|g_n\|^2_{2,m,\lambda})$, the remainder term in the expansion of $\Lambda$, are all seen to be $o(n^{-(p-r)r})$. Hence for $n$ sufficiently large

$$\frac{\Lambda(f_n) - \Lambda(f_0)}{2} \leq \frac{\eta \delta N^{p-r} \psi^{(r)}(0)}{8} n^{-(p-r)r}.$$ 

Since $N$ is arbitrary, this demonstrates equation (2.9), and equation (2.10) follows similarly as in Stone (1980).

**Achievability.**

We first state the following lemma.

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

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

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

**Proof:** The proof follows by standard arguments as in Nadaraya (1989, Theorem 1.5, chapter 4, pg. 121).

We turn to the proof of the theorem.

Since $\Lambda$ is differentiable of index $r$ and order $m$, we have the expansion

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

where $\lambda$ is a vector of finite measures by (3.12). From Lemma 6.1 and Markov's inequality, we have that

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

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 non-zero measure:

$$\int_0^1 (\hat{f}^{(j)} - f^{(j)}) \, d\mu_j,$$
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.

We have

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

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

**Proof of Theorem 4.2**

**Upper bounds.**

We argue as in Stone (1980). Since $\Lambda$ is not degenerate, we may choose $f \in W$ and $h \in W^*$ such that $T_f(h) > 0$, without loss of generality. With $N > 0$ and $\delta > 0$ arbitrary let $g_n(x) = \delta Nn^{-1/2}h(x)$ and $f_n = f_0 + g_n; \ f_n \in W$ for $n$ sufficiently large. Equation (2.1) in [11] holds. Proceeding as in the previous theorem and using the differentiability of $\Lambda$ we have that

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

$$= \frac{1}{2} \delta Nn^{-1/2} T_f(h) + O(n^{-1}) \geq \frac{\delta NT_f(h)}{4} n^{-1/2}$$

for $n$ sufficiently large. Since $N > 0$ can be arbitrarily large, equation (2.9) holds.

Equation (2.10) follows as in Stone (1980).

**Achievability**

**Lemma 6.2**

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

and

$$||\hat{\rho} - \rho||_{(\infty, m)} = O_p\left( \frac{1}{\sqrt{nb^{m+1}}} + \frac{1}{\sqrt{nb^{m+1}}} \right).$$
Let $\tilde{f}$ be given by
\[ \tilde{f}(x) = \frac{1}{n \rho(x)} \sum_{i=1}^{n} K_b(x_i, X_i) f(X_i). \] (6.18)

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

Proof: Let $0 \leq j \leq m$.

First,
\[
\left| \frac{1}{n} \sum_{i=1}^{n} K_b^{(j,0)}(x, X_i) - \rho_b^{(j)} \right|
\leq CD_nb^{-(j+1)}
\]
by a change of variable, where $D_n = \sup \{|F_n(t) - F(t)|\}$. Since $D_n = O_p(1/\sqrt{n})$, the term above is $O_p \left( \frac{1}{\sqrt{nb^{j+1}}} \right)$.

Note that the constant in the $O_p$ term does not depend on $x$. Next,
\[
\left| \frac{1}{n} \sum_{i=1}^{n} K_b^{(j,0)}(x, X_i) - \rho^{(j)} \right|
\leq \left| \frac{1}{n} \sum_{i=1}^{n} K_b^{(j,0)}(x, X_i) - \rho^{(j)}_\ast \right| + \left| \rho^{(j)}_\ast - \rho^{(j)} \right|_{\infty},
\]
and invoke equation (A.1). The proof of the first two assertions is now completed by noting that the term $j = m$ is dominant. The proof of the last assertion is entirely similar.

Proof of Theorem 4.2

Consider the expansion
\[ \Lambda(\hat{f}) - \Lambda(f) = T_f(\hat{f} - f) + O(||\hat{f} - f||_{(2,m,\lambda)}^2). \]

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

25
where $\tilde{f}$ is as in (6.18).

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

$$\tilde{f}(x) - f(x) = \frac{1}{n \rho(x)} \sum_{i=1}^{n} \left( K_b(x, X_i) f(X_i) - \int_0^1 K_b(x, t)f(t) \rho(t) \right) + O(b^p). \quad (6.19)$$

Applying $T_T$, we find that $T_T(\tilde{f} - f)$ is the contribution from the last term above, of order $O(b^p) = O(1/\sqrt{n})$, plus a simple average of $n$ mean zero, independent terms. The latter has variance bounded in $n$, and is seen to be $O_p(1/\sqrt{n})$ by (3.14).

It remains to consider $T_T$ applied to

$$\hat{f} - \tilde{f} = \left( \frac{1}{\hat{\rho}(x)} - \frac{1}{\rho(x)} \right) \frac{1}{n} \sum_{i=1}^{n} K_b(x, X_i) f(X_i) + \frac{1}{n \hat{\rho}(x)} \sum_{i=1}^{n} K_b(x, X_i) e_i. \quad (6.20)$$

We treat the first sum in (6.20) first. Let

$$e(x) = \rho(x) \tilde{f}(x) w_f(x),$$

and consider the functional

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

Then $T_T$ applied to the first sum in 6.20 may be written as $(\Gamma(\hat{\rho}) - \Gamma(\rho_b)) + (\Gamma(\hat{\rho}_b) - \Gamma(\rho))$. A Taylor expansion of the function $\Gamma(\rho_b + tq)$ about $t = 0$ evaluated at $t = 1$ for $q = \hat{\rho} - \rho_b$ yields

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

for some $s \in [0,1]$. All on a set of arbitrarily large probability, the first assertion of Lemma 6.2 shows that the denominator in the second term is bounded away from zero for large $n$; the last assertion shows that $||\tilde{f}||_\infty$ and therefore $||e||_\infty$ is bounded. Therefore, the second term above is of the order $||\hat{\rho} - \rho_b||_2^2 = O_p(1/\sqrt{n})$, by lemma 6.1 and the triangle inequality.

The first term is equal to

$$\int (\hat{\rho} - \rho_b) \left( \rho \tilde{f} - (\rho f)_b \right) \frac{w_f}{\rho_b^2} + \int (\hat{\rho} - \rho_b) (\rho f)_b \frac{w_f}{\rho_b^2}. \quad (6.21)$$
Use that \(w_f\) and \(1/\rho_b^2\) are bounded, and apply the Cauchy-Schwartz inequality to the first term. Lemma 6.1 and the triangle inequality bounds \(\|\hat{\rho} - \rho_b\|_2\). The function \(\rho_f^c - (\rho f)_b\) has expectation zero 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 \(\|\rho_f^c - (\rho f)_b\|_2 = O_p(n^{-1/4})\). Taking the product of the two \(L^2\) norm bounds shows that the first term is \(O_p(1/\sqrt{n})\).

The second term in (6.21) is seen to be a sum of \(n\) terms with zero mean and bounded variance, hence of order \(1/\sqrt{n}\).

A similar but simpler argument may be used to handle \(\Gamma^c(\rho) - \Gamma^c(\rho_b)\), using \(b_0^* = O(1/\sqrt{n})\).

Lastly consider \(T_f\) applied to the final term in equation (6.20).

Let \(\Omega_n = \{\omega : \inf_x \hat{\rho}(x) > \beta/2\}\), where \(\beta = \inf_x \rho(x) > 0\) by condition 2.2.2.

By Lemma 6.2, \(P(\Omega_n) \to 1\); hence it suffices to show that

\[
T_f \left\{ \frac{1}{n \hat{\rho}(x)} \sum_{i=1}^n K_b(x, X_i) \epsilon_i \right\} = \frac{1}{n} \sum_{i=1}^n \left[ \int_0^1 \frac{K_b(x, X_i)}{\hat{\rho}(x)} w_f(x) \, dx \right] \epsilon_i = O_p(1/\sqrt{n}).
\]

Conditioning on \(X_1, \ldots, X_n\) shows the terms of the sum to have mean zero; the conditional variance formula and a change of variable show that each has variance bounded in \(n\).

## A Appendix

**Formula for and properties of the kernel.** In this section we present a general formula for the kernel of order \(p \geq 2\), and relevant properties. Further simplification, proofs and discussion may be found in Messer and Goldstein (1989).

### 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

\[
\psi(t) = (e^{itr_0}, \ldots, e^{itr_{p-1}})'.
\]

We denote the components of \( \psi(t) \) by \( \psi_j, \ 0 \leq j \leq p - 1 \).
Let \( \phi(t) = C\psi(t) \), where the \( p \times p \) matrix \( C \) is

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

with \( \Lambda = i\text{diag}(r_0, \ldots, r_{p-1}) \) and \( L \) the Vandermonde matrix

\[
L = [1, \Lambda_1, \ldots, \Lambda_1^{p-1}1].
\]

Here \( 1 = (1, \ldots, 1)' \).
Let \( k(t) \) be the real function

\[
k(t) = \frac{-1}{2p} \sum_{j=0}^{p-1} ir_j \psi_j(|t|). \tag{A.22}
\]

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

\[
K_b(x, t) = b^{-1}k\left(\frac{x-t}{b}\right) + b^{-1} \sum_{j=1}^{p} (-1)^{j+1} \left\{ \phi_{2p-j}(\frac{t}{b})k^{(2p-j)}(\frac{x}{b}) + \phi_{2p-j}(\frac{1-t}{b})k^{(2p-j)}(\frac{1-x}{b}) \right\}. \tag{A.23}
\]

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

\[
2^{\frac{3}{2}}bK_b(2^{\frac{3}{2}}bx, 2^{\frac{3}{2}}bt) = e^{-|x-t|}(\sin(|x-t|) + \cos(x-t)) + \Phi(x+t, x-t) + \Phi((1-x) + (1-t), (1-x) - (1-t)).
\]

**A.2 Properties.**

The following theorem bounds the asymptotic bias of the kernel. 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.23),

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

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

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

$$|k^{(j)}(x;b)| \leq C(j,p)b^{-(j+1)}e^{-\frac{\sin(\frac{x}{b})^p}{x^p}}.$$  

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

References


