The Annals of Statistics
1997, Vol. 25, No. 6, 2547–2554

ON NONPARAMETRIC CONFIDENCE INTERVALS

BY MARK G. LOW

University of Pennsylvania

An inequality is given for the expected length of a confidence interval given that a particular distribution generated the data and assuming that the confidence interval has a given coverage probability over a family of distributions. As a corollary, attempts to adapt to the regularity of the true density within derivative smoothness classes cannot improve the rate of convergence of the length of the confidence interval over minimax fixed-length intervals and still maintain uniform coverage probability. However, adaptive confidence intervals can attain improved rates of convergence in some other classes of densities, such as those satisfying a shape restriction.

1. Introduction. One of the basic problems in nonparametric function estimation is the construction of confidence intervals and bands for an unknown function based on noisy data. For example, in density estimation problems a confidence interval for the value of the density at a particular point can be constructed based on $X_1, X_2, \ldots, X_n$, i.i.d. observations each with density f. Typically some regularity of the function f is assumed, usually specifying that the unknown function has a given number of derivatives, say k. Two basic approaches to the construction of a confidence interval for f at a particular point x_0 have been given in the literature.

One approach is to specify a particular bound on the kth derivative. Suppose that the data are real valued. Then, for some M and some a,

$$(1)\quad f \in \mathscr{F}(a, k, M) = \left\{ f\colon f \geq 0,\ \int f = 1,\ f(x_0) \leq a,\ \| f^{(k)}(x) \|_\infty \leq M \right\}.$$

We can then seek fixed-length confidence intervals with a given minimum level of coverage over $\mathscr{F}(a, k, M)$ that minimize the length of the interval. This point of view has been studied in depth. In a white-noise setting, Donoho (1994) has developed a precise optimality theory. In the density estimation setting, this theory leads to linear estimators $\hat{f}_n$ and constants $D(\alpha)$ such that, for all $f \in \mathscr{F}(a, k, M)$,

$$(2)\quad \begin{aligned} P_f\big(\hat{f}_n - D(\alpha) a^{k/(2k+1)} M^{1/(2k+1)} n^{-k/(2k+1)} \\ \leq f(x_0) \leq \hat{f}_n + D(\alpha) a^{k/(2k+1)} M^{1/(2k+1)} n^{-k/(2k+1)} \big) \geq 1 - \alpha. \end{aligned}$$

Note that if $a > (M/k!)^{1/(k+1)}$ and $f(x_0) = a$, then a simple variational argument shows that $f \notin \mathscr{F}(a, k, M)$, so we shall assume that $a \leq (M/k!)^{1/(k+1)}$.

Received September 1995; revised March 1997.
AMS 1991 *subject classification*. Primary 62G07.
Key words and phrases. Confidence intervals, density estimation.

These fixed-length intervals are close to optimal among all fixed-length intervals. More precisely Donoho (1994) has shown that if C_n is any confidence interval for $f(x_0)$ based on the observations $X_1, X_2, \ldots, X_n$, with coverage probability over the class of densities $\mathscr{F}(a, k, M)$ of at least $1 - \alpha$, then there is a constant $\tilde{D}(\alpha)$ such that the length of the confidence intervals is at least $\tilde{D}(\alpha)a^{k/(2k+1)}M^{1/(2k+1)}n^{-k/(2k+1)}$. One of the consequences of the bounds given in Theorem 1 of this paper is that these intervals are in fact (rate) optimal even among random-length intervals. More precisely suppose that C_n is a confidence interval for $f(x_0)$ based on the observations $X_1, X_2, \ldots, X_n$, with coverage probability

$$\inf_{\mathscr{F}(a,k,M)} P_f(f(x_0) \in C_n) \geq 1 - \alpha. \tag{3}$$

Let μ be Lebesgue measure. Then it will follow from Theorem 1 of Section 2 that there is a constant $D_1(\alpha)$ such that

$$\sup_{\mathscr{F}(a,k,M)} E_f \mu(C_n) \geq D_1(\alpha)a^{k/(2k+1)}M^{1/(2k+1)}n^{-k/(2k+1)}. \tag{4}$$

Hence, optimal random-length confidence intervals have the same minimax rate of convergence as optimal fixed-length confidence intervals. Of course, two complaints can be made about the fixed-length confidence interval given in (2). If $\|f^{(k)}(x)\|_\infty \ll M$, then the confidence interval is much longer than if we had chosen a much smaller value of M. On the other hand, if $\|f^{(k)}(x)\|_\infty \gg M$, then it is quite likely that the confidence interval defined by (2) will have poor coverage probability. The same remarks can also be made regarding the value of $f(x_0)$.

Such thoughts have led to another point of view. Confidence intervals should adjust so that their expected length depends on the magnitude of $f^{(k)}(x)$ in the neighborhood of the point x_0 and on the value of $f(x_0)$. Of course, the length of such an interval must depend on the data; in particular, it cannot be of fixed length. Such a point of view is not in conflict with the bound given in (4) because this bound measures only the maximum expected length over $\mathscr{F}(a, k, M)$.

Random-length confidence intervals are usually based on fairly complicated resampling schemes. See, for example, Hall (1992) and Härdle and Marron (1991). It has been noted in these papers that the resulting confidence intervals often have poor coverage probability. This leads to the question of whether such schemes are ever likely to improve substantially on fixed-length procedures.

We shall give bounds on the size of confidence intervals that show that without assuming some extra regularity for the kth derivative attempts to adjust to the value of the kth derivative are doomed to failure. More specifically suppose that C_n is a confidence interval for $f(x_0)$ with coverage probability of at least $1 - \alpha$ over $\mathscr{F}(a, k, M)$. Then for any $\varepsilon > 0$ the expected length of this confidence interval must satisfy (3) for *every* $f \in \mathscr{F}(a, k, M - \varepsilon)$. Hence, adaptation is severely limited in this problem. The construction of these bounds is based on a simple inequality which is given in a general

setting in Section 2. In Section 3 we apply this bound to the density estimation problem.

2. Lower bounds. In this section a general lower bound based on the L-distance between probability measures is given for the length of confidence intervals. This bound can then be applied easily to density estimation, nonparametric regression and white noise models. We only give the abstract result in this section. The connection to a density estimation example is made in Section 3.

Suppose that a random variable X is generated by a probability measure P_f, where $f \in \mathscr{F}$, a convex set of parameter points. Let $C(X)$ be a confidence interval for a linear functional Tf, with coverage probability of at least $1 - \alpha$. Assuming that X has distribution P_f write $P_f A$ for the probability of a set A, and $E_f S(X)$ for the expectation of the random variable $S(X)$. Also write $\|Q\|$ for the total variation of a signed measure Q. That is, $\|Q\| = \sup_\nu |\int \nu \, dQ|$, where the supremum is taken over all measurable ν with $|\nu| \le 1$. For any pair of parameter points $f_0 \in \mathscr{F}$ and $f_1 \in \mathscr{F}$ let f_λ be the member of the affine family joining them given by $f_\lambda = f_0 + \lambda(f_1 - f_0)$. Since $\mathscr{F}$ is convex $f_\lambda \in \mathscr{F}$ whenever $0 \le \lambda \le 1$. Lower bounds for the expected length of a confidence interval for a linear functional Tf can be described easily in terms of a modulus of continuity ω, defined by

$$\omega(f_0, \varepsilon, \mathscr{F}) = \sup\{|Tf_1 - Tf_0| : \|P_{f_1} - P_{f_0}\| \le \varepsilon, \, f_1 \in \mathscr{F}\}. \tag{5}$$

THEOREM 1. *Suppose that $C(X)$ is a confidence set for the linear functional Tf, with coverage probability of at least $1 - \alpha$ over the convex parameter space $\mathscr{F}$. If P_{f_0} is the actual distribution generating X, it follows that, for any $\varepsilon > 0$,*

$$E_{f_0} \mu(C(X)) \ge \left(1 - \alpha - \frac{\varepsilon}{4}\right) \omega(f_0, \varepsilon, \mathscr{F}). \tag{6}$$

PROOF. Let I be the indicator function defined by

$$I(a \in B) = \begin{cases} 1, & \text{if } a \in B, \\ 0, & \text{if } a \notin B. \end{cases}$$

Now, for a given ε, let $\delta > 0$ be arbitrary. Then, by the definition of $\omega(f_0, \varepsilon, \mathscr{F})$, there is an $f_1 \in \mathscr{F}$ such that

$$|Tf_1 - Tf_0| \ge \omega(f_0, \varepsilon, \mathscr{F}) - \delta \tag{7}$$

and

$$\|P_{f_1} - P_{f_0}\| \le \varepsilon. \tag{8}$$

Since $C(X)$ has probability of coverage of at least $1 - \alpha$ it follows that, for $0 \le \lambda \le 1$,

$$P_{f_\lambda} I(Tf_\lambda \in C(X)) \ge 1 - \alpha, \tag{9}$$

and it follows from (8) and (9) that

$$\|P_{f_\lambda} - P_{f_0}\| \le \lambda\varepsilon; \tag{10}$$

hence

$$P_{f_0} I(Tf_\lambda \in C(X)) \ge 1 - \alpha - \frac{\lambda\varepsilon}{2}. \tag{11}$$

Now

$$E_{f_0}\,\mu(C(X)) \ge E_{f_0}\int_0^1 (Tf_1 - Tf_0)\, I(Tf_\lambda \in C(X))\, d\lambda \tag{12}$$

$$= (Tf_1 - Tf_0)\int_0^1 P_{f_0} I(T(f_\lambda) \in C(X))\, d\lambda \tag{13}$$

$$\ge (Tf_1 - Tf_0)\int_0^1 \left(1 - \alpha - \frac{\lambda\varepsilon}{2}\right) d\lambda \tag{14}$$

$$= (\omega(f_0, \varepsilon, \mathscr{F}) - \delta)\left(1 - \alpha - \frac{\varepsilon}{4}\right). \tag{15}$$

Hence (6) holds, since $\delta > 0$ is arbitrary. □

3. Bounds for a density estimation example. The bounds given in Section 2 can be easily applied in a variety of function estimation problems such as nonparametric regression and density estimation. In particular, the density estimation problem discussed in the Introduction is treated easily. Let $X_1, X_2, \ldots, X_n$ be i.i.d. f, where $f \in \mathscr{F}(a, k, M)$. Without loss of generality focus attention on finding a confidence interval for $f(0)$. The bound given in Section 2 is based on the L_1-distance between probability measures. In product situations the L_1-distance is often difficult to calculate, and it is often more convenient to bound the L_1-distance by the Hellinger distance. Such bounds are by now standard and are given, for example, by Le Cam (1986). For two measures P and Q write $H(P, Q)$ for the Hellinger distance between P and Q, where

$$H^2(P, Q) = \int \left(\sqrt{dP} - \sqrt{dQ}\right)^2. \tag{16}$$

Then results in Chapter 4 of Le Cam (1986) immediately show that the L_1-distance between the product measures P^n and Q^n can be bounded by

$$L_1(P^n, Q^n) \le 2\left(2 - 2\left(1 - \frac{H^2(P, Q)}{2}\right)^n\right)^{1/2}. \tag{17}$$

This bound can be combined with Theorem 1 to yield lower bounds for the size of confidence intervals in density estimation problems. For example, suppose that C_n is a variable-length confidence interval for $f(0)$ based on i.i.d. data $X_1, X_2, \ldots, X_n$, each with density f. If C_n has coverage probability of at least $1 - \alpha$ over $\mathscr{F}(a, k, M)$,

$$\inf_{\mathscr{F}(a, k, M)} P_f(f(0) \in C_n) \ge 1 - \alpha. \tag{18}$$

As mentioned in the Introduction, Theorem 1 can be used to show that there is a constant $D_1(\alpha)$ such that

$$\sup_{\mathscr{F}(a,k,M)} E_f \mu(C_n) \geq D_1(\alpha) a^{k/(2k+1)} M^{1/(2k+1)} n^{-k/(2k+1)}. \tag{19}$$

In fact Theorem 1 gives bounds for the expected length of a confidence interval for *each* density f and not just for the maximum length over $\mathscr{F}(a, k, M)$.

Let g be a bounded function with compact support such that $\|g^{(k)}(x)\|_\infty \leq 1$, $g(0) < 0$, $\int g = 0$ and $\int g^2 = \delta$. Then the renormalized function

$$g_{ab}(x) = (a^k b)^{1/(2k+1)} g\left(\left(\frac{b^2}{a}\right)^{1/(2k+1)} x\right) \tag{20}$$

satisfies

$$g_{ab}(0) = (a^k b)^{1/(2k+1)} g(0), \tag{21}$$

$$\int g_{ab}^2 = a\delta \tag{22}$$

and

$$\left\| g_{ab}^{(k)} \right\|_\infty \leq b. \tag{23}$$

Now suppose that f_0 is a density such that $\|f_0^{(k)}(x)\|_\infty < M$ and $f_0(0) > \varepsilon > 0$. Let $b = M - \|f_0^{(k)}(x)\|_\infty$ and $a = f_0(0)$. Define $f_{n,1}$ by

$$f_{n,1}(x) = f_0(x) + n^{k/(2k+1)} g_{ab}(n^{1/(2k+1)} x), \tag{24}$$

and note that, for sufficiently large n, $f_{n,1} \in \mathscr{F}(a, k, M)$. The squared Hellinger distance between f_0 and $f_{n,1}$ satisfies

$$H^2(f_0, f_{n,1}) = \int \left(\sqrt{f_0(x)} - \sqrt{f_{n,1}(x)}\right)^2 dx \tag{25}$$

$$= \int \left(\sqrt{f_0(x)} - \sqrt{f_0(x)\left(1 + \frac{f_{n,1}(x) - f_0(x)}{f_0(x)}\right)}\right)^2 dx \tag{26}$$

$$= \frac{\delta}{4n}(1 + o(1)) \tag{27}$$

uniformly over $f_0 \in \mathscr{F}(a, k, M) \cap \{f\colon f(0) \geq \varepsilon\}$. Hence, by (17),

$$L_1(f_0^n, f_{n,1}^n) \leq 2\left(2 - 2\left(1 - \frac{\delta}{8n}\right)^n\right)^{1/2} (1 + o(1)) \tag{28}$$

$$\leq 2\left(2 - 2\exp - \frac{\delta}{8}\right)^{1/2} (1 + o(1)) \tag{29}$$

uniformly over $f_0 \in \mathscr{F}(a, k, M) \cap \{f\colon f(0) \geq \varepsilon\}$. Now set γ equal to

$$\gamma = 2\left(2 - 2\exp - \frac{\delta}{8}\right)^{1/2} \tag{30}$$

and let $Tf = f(0)$. Note that

$$Tf_{n,1} - Tf_0 = g(0) f_0^{k/(2k+1)}(0)\left(M - \left\| f_0^{(k)}(x)\right\|_\infty\right)^{1/(2k+1)} n^{-k/(2k+1)}, \tag{31}$$

and hence

$$\begin{aligned}&\omega(f_0, \gamma, \mathscr{F}(a, k, M))\\ &\quad \geq g(0) f_0^{k/(2k+1)}(0)\left(M - \left\| f_0^{(k)}(x)\right\|_\infty\right)^{1/(2k+1)} n^{-k/(2k+1)}(1 + o(1)).\end{aligned} \tag{32}$$

It then follows from Theorem 1 that there are constants D and N such that if $n \geq N$, then

$$\begin{aligned}E_{f_0}\mu(C_n(x)) &\geq (1 - \alpha - \gamma) D f_0^{k/(2k+1)}(0)\\ &\quad \times \left(M - \left\| f_0^{(k)}(x)\right\|_\infty\right)^{1/(2k+1)} n^{-k/(2k+1)}.\end{aligned} \tag{33}$$

We can summarize this result in the following theorem.

THEOREM 2. *Let $\varepsilon > 0$ and suppose that C_n satisfies* (18). *Then $\exists$ N that depends on ε and M, and a constant $D(\gamma) > 0$ independent of n, ε, a and M, such that, for all $n > N$,*

$$\begin{aligned}E_f\mu(C_n(x)) &\geq (1 - \alpha - \gamma) D(\gamma)\left(f^{k/(2k+1)}(0)\right)\\ &\quad \times \left(M - \|f^{(k)}\|_\infty\right)^{1/(2k-1)} n^{-k/(2k+1)}\end{aligned} \tag{34}$$

for all $f \in \mathscr{F}(a, k, M)$ with $f(0) > \varepsilon$.

REMARK. 1. Theory for confidence intervals in nonparametric function estimation problems can be divided into:

(a) fixed-length confidence intervals;
(b) minimax expected length confidence intervals;
(c) pointwise expected length confidence intervals.

It is clear that for a given coverage probability that optimal fixed-length confidence intervals cannot be shorter than the expected length of minimax expected length confidence intervals. Similarly the maximum expected length of minimax expected length confidence intervals cannot be shorter than the expected length of this confidence interval assuming that a particular distribution generated the data.

Donoho (1994) has given a detailed treatment of fixed-length confidence intervals. Theorem 1 of this paper gives lower bounds for the expected length of minimax and pointwise random-length confidence intervals.

REMARK 2. Some of the recent work in adaptive estimation has focused on finding estimators that adapt for the mean squared error over many different function spaces. It is often possible to find estimators that are almost simultaneously minimax over a whole range of parameter spaces. In one sense, this paper shows that although an estimate may be adaptive for squared error loss it may be impossible to make a data dependent claim on how well you have done. In particular, once we admit that f might perhaps have only k derivatives, the expected length of the confidence interval must be of the order $n^{-k/(2k-1)}$ even if f is in fact infinitely differentiable.

REMARK 3. An entirely similar analysis yields corresponding results for both nonparametric regression and white noise models. In fact the analysis is sometimes even easier for those models as the L_1-distance can often be evaluated explicitly.

REMARK 4. Donoho and Liu (1991) and Donoho (1994) reduced the minimax theory for estimating linear functionals Tf over convex parameter spaces $\mathscr{F}$ to finding the one-dimensional subfamily of $\mathscr{F}$ which makes the problem of estimating Tf most difficult. In adaptive estimation several parameter spaces are studied simultaneously. For simplicity consider two convex spaces $\mathscr{F}_1$ and $\mathscr{F}_2$ and let $\mathscr{F}_2 \subset \mathscr{F}_1$.

Suppose that for estimating the linear functional over $\mathscr{F}_1$ that the associated hardest one-dimensional subfamily is $\{g_\theta\colon -1 \le \theta \le 1\}$. Sometimes this hardest one-dimensional subfamily will contain a function, say g_ϕ, which also belongs to $\mathscr{F}_2$. Then applying Theorem 1 with $f_0 = g_\phi$ shows that in this case it is impossible to find adaptive length confidence intervals over $\mathscr{F}_1$ and $\mathscr{F}_2$. In particular, this is the case for estimating a function at a point over Lipschitz classes of different orders.

It is, however, no longer the case for Lipschitz classes of order less than or equal to 1 when these functions are also assumed to be monotone. Under these constraints Hengartner and Stark (1995) have constructed confidence intervals which do adapt to the unknown Lipschitz order.

Acknowledgment. I would like to acknowledge the very detailed and helpful comments given by a referee.

REFERENCES

DONOHO, D. L. (1994). Statistical estimation and optimal recovery. *Ann. Statist.* **22** 238–270.

DONOHO, D. L. and LIU, R. C. (1991). Geometrizing rates of convergence, III. *Ann. Statist.* **19** 668–701.

HALL, P. (1992). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. *Ann. Statist.* **20** 675–694.

HÄRDLE, W. and MARRON, J. S. (1991). Bootstrap simultaneous error bars for nonparametric regression. *Ann. Statist.* **19** 778–798.

HENGARTNER, N. W. and STARK, P. B. (1995). Finite-sample confidence envelopes for shape-restricted densities. *Ann. Statist.* **23** 525–550.
LE CAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York.

DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104
E-MAIL: lowm@compstat.wharton.upenn.edu