# Geometrizing Rates of Convergence I

*David L. Donoho*
*Richard C. Liu*

Department of Statistics
University of California, Berkeley

## ABSTRACT

We introduce a new bound on the rate of convergence for estimating functionals of a density. The rate bound is always at least as strong as the bounds obtained by applying existing techniques due to Farrell, Stone, and Hasminskii, respectively. In addition, the bound is conceptually easy to understand and explain, and treats the nonparametric, semiparametric, and parametric cases in a unified way.

Let $T$ be the functional of interest and let $\mathbf{F}$ be the class of regular (e.g. smooth) densities under consideration. Let $H(F,G)$ denote Hellinger distance, and let $b(\varepsilon)$ be the *modulus of continuity* of the functional $T$ at $F_0$ with respect to Hellinger distance:

$$b(\varepsilon) = \sup\{|T(F) - T(F_0)|: \quad H(F,F_0) \le \varepsilon, \quad F,F_0 \in \mathbf{F}\}. \tag{1}$$

In this paper, we show that the rate of convergence of an estimate $T_n$ to $T(F)$ does not exceed $b(n^{-1/2})$, in a local minimax sense. Thus, if $b(\varepsilon) = \varepsilon^r$, the rate of convergence of $T_n$ to $T(F)$ cannot exceed $b(n^{-1/2}) = n^{-r/2}$ uniformly over a Hellinger neighborhood of $F_0$ in $\mathbf{F}$.

We compute the modulus for 5 interesting cases: estimating functionals such as the location of the mode, the density at a point, and $\int f^2$. In these cases, the optimization problem (1) has concrete explicit solutions, and so the bound $b(n^{-1/2})$ is easy to apply.

The $b(n^{-1/2})$ bound also applies in classical settings; in regular parametric cases it gives Pitman's form of the Cramèr-Rao inequality; in semiparametric cases it generalizes Stein's formula for semiparametric information.

**American Mathematical Society 1980 subject classifications:** Primary 62G20, secondary 62G05, 62F35.

**Key Words and Phrases:** Functionals of a density, Modulus of continuity, Rates of convergence, Nonparametric estimation, Decreasing densities, Estimation of mode, Fisher information, Hellinger distance, Least Informative families.

**Acknowledgements:** The authors would like to thank Lucien LeCam, Lucien Birgé, Alex Samarov, Rudy Beran, and Charles Stone for many helpful discussions. A conversation with S. Marron spurred the writing of this paper. This work was supported by NSF grant DMS-84-51753

# 1. Introduction

*Contents*

### 1.1. Background

Let $T$ be a functional of the unknown probability distribution $F$, and suppose we have a sample $X_1,.....X_n$ i.i.d $F$. If $T$ is continuous in one of the "weak" topologies (e.g. Prohorov, Kolmogorov) then $T(F)$ can be estimated by $T(F_n)$ where $F_n$ is the empirical distribution $F_n(x) = \frac{1}{n}\sum I_{\{X_i \le x\}}$. For example, if $T(F) = \int \Psi dF$, where $\Psi$ is continuous and bounded, then $T(F_n)$ converges to $T(F)$. As an added bonus, in this example convergence occurs at a root-$n$ rate: $n^{1/2}(T(F_n) - T(F)) = O_p(1)$.

On the other hand, many functionals of interest are not continuous in such "weak" topologies. The typical cases are those which are defined in terms of the density $f$ rather than the distribution $F$; as examples, one could cite $T(F) = f(0)$ (the density of $F$ at 0), and $T(F) = \int f^2$. For such functionals, one has no "natural" estimate. Instead, one (for example) computes a kernel density estimate $\hat{f}_n$ and plugs it into an appropriate formula to get an estimate $T_n$ of $T(F)$. In general, such an estimate will converge, as $n \to \infty$, to $T(F)$; however, the rate may be arbitrarily slow (Farrell 1967, Devroye and

Györfi 1985).

This sort of "slow convergence" result does not deter practical use of such methods, which may be justified theoretically by the fact that under supplementary *regularity conditions* on the unknown density, an appropriately tuned estimator will attain a well-defined rate of convergence. Thus, for example, if $F$ has a density $f$ with two continuous, bounded derivatives, an appropriately tuned kernel estimate of $T(F) = f(0)$ can attain a "2/5" rate of convergence: $n^{2/5}(T_n - T(F)) = O_p(1)$.

There are many different estimators of nonparametric functionals, and they attain a variety of rates of convergence — depending on the functional being estimated, on the method of estimation and on the assumed regularity of $F$. If $F$ is very smooth, in estimating $f(0)$, one can attain the rate $n^{-1/3}$ by histogram estimators, $n^{-2/5}$ by kernel density estimators with positive kernel, and rates close to $n^{-1/2}$ by kernel density estimators with special kernels. For other functionals one attains other rates: for estimating the mode one attains $n^{-1/5}$ by a kernel procedure (Venter,1967), but by special techniques under special regularity one attains even faster rates - close to $n^{-1/2}$ (Eddy, 1980). For integral functionals such as the $L_2$ -norm $\int f^2$, one sees results, for example of Ahmad (1976), citing rates of $n^{-1/2}$ for kernel-based estimates.

The wide variety of cited rates and regularity conditions leads to the question whether a certain rate of convergence is "optimal" for the given functional and regularity assumption. Several authors have determined optimal rates of convergence for functionals of a density; notably, Farrell (1972), Wahba (1975), Meyer (1977b), Hasminskii (1979), Stone (1980), Hall and Welsh (1984), Hall and Marron (1987), Ritov (1986); see also the review by Kiefer (1982). The idea is to show, under certain regularity conditions on the unknown density $f$, that a certain rate of convergence cannot be exceeded uniformly for all $f$ satisfying the condition. This places a bound on the rate of uniform convergence, which can then be shown to be attainable by exhibiting an estimator which attains this bound.

There are, in our view two main approaches for bounding rates of convergence for functionals of a density: the "testing method" and the "parametric method", respectively.

Stone (1980) makes quite clear the basic idea behind the testing method. One constructs a sequence $\{F_{1,n}\}$ of distributions in F that approach a fixed $F_0 \in F$. The sequence must be chosen so that the hypotheses $H_1 : F_{1,n}$ and $H_0 : F_0$ are hard to tell apart based on the best test using $n$-observations $X_1,.....X_n$. Then $T_n - T(F)$ cannot usually be smaller than $\delta_n = |T(F_{1,n}) - T(F_0)|/2$, uniformly in F; otherwise the test that decides in favor of $H_i$ when $|T_n - T(F_{i,n})| < \delta_n$ would reliably test $H_1$ against $H_0$.

Stone appears to be the first to make a clear exposition of the testing method; in retrospect, though, the work of Farrell (1972) may be seen as an instance of the testing method (relying on a different inequality at a key point, however; a close reading of Meyer (1977a) should convince the reader that Farrell's method is based on testing).

Hasminskii's bound on rates of convergence uses ideas from parametric theory. In brief, if F is the class of all densities satisfying the given regularity condition, and $F_0$ is a particular distribution of interest, one constructs a sequence of parameter families $\{F_{\theta,n}\} \subset F$: the families all have $F_{0,n} = F_0$, i.e. they "pass through" the point $F_0$. The idea is that the parameter $\theta$ should be hard to estimate, for large $n$. One does this by showing that $F_{\theta,n}$ is a locally asymptotically normal family, and applying an asymptotic minimax theorem in the spirit of Hajek-LeCam. One then establishes a link between estimating $\theta$ and estimating $T$, i.e. an equation of the form $T(F_{\theta,n}) - T(F_0) = c\ n^{-\gamma}\ \theta$; using this link one shows that because $\theta$ is hard to estimate, $T$ cannot be estimated at a rate exceeding $n^{-\gamma}$.

The technique of Farrell (1967) is also based on parametric theory. Starting from the same setup as the Hasminskii method, the parameter $\theta$ is shown to be difficult to estimate via the Cramér-Rao inequality rather than a local asymptotic minimax bound. This promising idea appears to have been little used, however. The note of Boyd and Steele (1978) and the unpublished manuscript of Farrell (1980) are the only other applications of this technique we have found.

(We should also mention that "testing" and "parametric" approaches to bounding rates of convergence have been used in other problems - e.g. in estimating functionals of the spectral density of a

stationary time series (Samarov, 1976). Also, in the problem of estimating an entire density (and not just a single functional) new ideas arise; we mention, in alphabetical order, key papers of Birgè (1983), Bretagnolle and C. Huber (1979), Centsov (1962), Hasminskii (1978), LeCam (1975) and Stone (1983). However, in the present work we focus on bounds for functionals $T(F)$ of a density, rather than on functionals of spectral densities or on the entire density).

Both the "testing" and "parametric" methods are powerful, and have produced very interesting and illuminating results. On the other hand, the methods are not really easy to use or explain to others. They depend, for one thing, on proper choice of testing alternatives $\{F_{1,n}\}$ or of parameter families $\{F_{\theta,n}\}$ to get interesting results. Often, one employs a perturbation argument. For example, in the parametric approach, one invents a "perturbing" function $g$ and applies a recipe such as

$$F_{\theta,n}(t) = F_0(t) + \theta\, c_n\, g((t - t_0)/s_n)$$

where $c_n$ and $s_n$ are "normalizing factors". The factors can be adjusted subject to the constraints of keeping $F_{\theta,n} \in \mathbf{F}$, of keeping $\theta$ hard to estimate, and of preserving the link between estimation of $\theta$ and $T$. Within these constraints, one adjusts the factors to give the strongest possible conclusions on estimation of $T$. Because of the technique required, one gets the impression that rates of convergence are isolated, magical quantities, whose value is only deduced after the dust settles in the aftermath of virtuoso calculation.

### 1.2. The New Approach

This paper presents an alternative to the Farrell/Stone/Hasminskii (F/S/H) procedures for bounding rates of convergence. This alternative is conceptually simple, can be explained to graduate students in midcareer, and gives, according to section 6 below, bounds at least as strong as F/S/H. In contrast to F/S/H, which require that one invent a sequence of testing alternatives or parameter families, and by a process of trial and error find normalizations leading to the most pessimistic conclusions, the procedure

we discuss here involves solving a clearly-stated optimization problem. There may be guesswork involved in solving the problem, but trial and error is not central to the approach at a conceptual level, as it is in the Farrell/Stone/Hasminskii methods.

The procedure we propose is to evaluate the *modulus* of the functional of interest in the *Hellinger* metric. This modulus is, of course, the generalized rate of change of $T$ over Hellinger $\varepsilon$-neighborhoods of a distribution $F_0$ in $\mathbf{F}$:

$$b(\varepsilon) = \sup \{|T(F) - T(F_0)| : H(F,F_0) \le \varepsilon, F, F_0 \in \mathbf{F}\}$$

where the Hellinger distance $H(F,G)$ is defined (assuming $f$ and $g$ are densities of $F$ and $G$ respectively w.r.t. some dominating measure $\mu$)

$$H(F,G) = \sqrt{\int (f^{1/2} - g^{1/2})^2 d\mu}$$

With this modulus in hand, the lower bound is easy to get: one just plugs $n^{-1/2}$ into the modulus, getting $b(n^{-1/2})$.

For example, as we see in section 3 below, if $\mathbf{F}$ is the class of unimodal densities and $T(F) = \text{mode}(F)$, then at an appropriate $F_0$, $b(\varepsilon) = A\ \varepsilon^{2/5} + o(\varepsilon^{2/5})$. Thus $b(n^{-1/2}) = n^{-1/5}$ is a bound on the rate of uniform convergence of estimators of the mode.

The sense in which $b(n^{-1/2})$ provides a bound is the traditional local minimax one. That is, for any estimate $T_n$ of $T$, the difference $T_n - T$ cannot be much smaller than $b(n^{-1/2})$ uniformly in a small neighborhood of $F_0$ in $\mathbf{F}$. For example, we conclude in section 2 below that

$$\liminf_{n \to \infty} \inf_{T_n} \sup_{F \in N_\delta(F_0)} E_F \left| \frac{T_n - T(F)}{b(n^{-1/2})} \right|^2 \ge C > 0 \qquad (1.1)$$

where $N_\delta(F_0)$ denotes a Hellinger $\delta$-neighborhood of $F_0$ in $\mathbf{F}$. In words, (1.1) says that "for any estimator $T_n$, there is an $F$ near $F_0$ at which $|T_n - T(F)|$ has a significant chance of being as big as $b(n^{-1/2})$". This lower bound holds quite generally and requires, for example, no linearity assumptions

on the functional $T$.

Of course, a bound of this kind would be quite useless if one could not compute $b(\varepsilon)$. In sections 3 - 5 below we compute the modulus for 5 different functionals of interest, getting quite explicit and concrete answers. The Hilbert space structure of the Hellinger distance makes this possible. In general, we find moduli of the form $b(\varepsilon) = \varepsilon^r$, $0 < r < 1$, and so the expression $b(n^{-1/2})$ translates into $O(n^{-r/2})$.

We are generally able to explicitly compute the $F_\varepsilon$ solving $|T(F_\varepsilon) - T(F_0)| = b(\varepsilon)$. This $F_\varepsilon$ is a kind of least-favorable distribution; it changes $T$ the most without going more than $\varepsilon$ away from $F$ in Hellinger distance. Inspection of $F_\varepsilon$ often shows the "reason" that certain rates occur. For example, in evaluating the modulus of $T(F) = f(\frac{1}{2})$ for a decreasing density, the rate $b(\varepsilon) = \varepsilon^{2/3}$ is seen to come from the relation between the height of a right triangle figure and the Hellinger norm. In evaluating the modulus of $T(F)$ = the mode of $F$, the rate $b(\varepsilon) = \varepsilon^{2/5}$ derives from the relation between the width of a parabolic arc and the Hellinger norm of the figure it inscribes. Thus rates of convergence tie in with simple intuitive facts of analysis.

Of course, even though a bound is computable, it would be of little use if it gave results less powerful than those supplied by the Farrell procedure and its competitors. In section 6 we show that, under mild conditions on the modulus, the strongest conclusion attainable from the Farrell, Stone, or Hasminskii procedures is at best $O(b(n^{-1/2}))$. In a sense, the user who pushes one of these procedures to give the sharpest conclusions is simply trying to compute $b(n^{-1/2})$, to within constant factors, without knowing that he is trying to do so. Our geometric approach seems clearer conceptually.

An interesting feature of our approach is the use of the modulus of continuity of the functional $T$. In several branches of applied mathematics the modulus plays a key role: in numerical analysis it governs the difficulty of numerical integration; in complexity theory, it governs the expense of approximate algorithms; in inverse problems, it governs the difficulty of recovering a functional from inexact remote sensing data. Here we have shown that the modulus controls the rate at which a functional can be estimated in a statistical problem.

A second key feature of our approach is the use of Hellinger distance. Of course, the work of LeCam (1970,1973), and Beran (1977,1978) has exhibited clearly the crucial role of Hellinger distance to understand efficient estimation in parametric settings. The work of Levit (1974,1975) and the little textbook of Pitman (1979) contribute as well to our feeling that Hellinger distance is the right way to understand efficiency. This paper shows that Hellinger distance helps understand nonparametric efficiency as well. Our debt to LeCam, to Beran, and to Pitman, but especially LeCam, will be clear.

### 1.3. $b(n^{-1/2})$ and the information inequality

We have talked so far about rates only, and not constants. If we consider the application of $b(n^{-1/2})$ to classical problems, however, it appears that $b(n^{-1/2})$ is correct not just as to rate, but actually as to constants – provided we divide by 2 (i.e. use $b(n^{-1/2})/2$). To begin, suppose we are interested in estimating the parameter $\theta \in \mathbf{R}$ of the parametric family $\mathbf{F} = \{F_\theta\}$. Let $b(\varepsilon)$ be the Hellinger modulus of the functional $T$ defined for $F_\theta \in \mathbf{F}$ by $T(F_\theta) = \theta$. If the family $\{F_\theta\}$ is quadratic mean differentiable in LeCam's sense, we can interpret a result of Beran (1977) as saying that

$$\lim_{\varepsilon \to 0} \frac{b(\varepsilon)}{2\varepsilon} = (I_{Fisher})^{-1/2} \tag{1.2}$$

where $I_{Fisher}$ is the usual Fisher information for the parameter $\theta$ at $\theta = 0$. In short the Fisher information is "encoded" in the modulus $b(\varepsilon)$: it determines the slope of $b(\varepsilon)$ at $\varepsilon = 0$.

Levit (1974,1975) has defined a notion of information for nonparametric estimation of regular functionals. For such functionals, the modulus $b(\varepsilon)$ "encodes" the information as well. For example, consider the linear functional $T(F) = \int \Psi \, dF$ where $\Psi$ is a bounded continuous function; to simplify matters suppose $\int \Psi \, dF = 0$. Levit's definition of nonparametric information about $T$ gives in this case

$$I_{Levit} = \frac{1}{\int \Psi^2 \, dF} .$$

On the other hand $T$ is Frechet differentiable in Hellinger norm; letting $\mathbf{F} = \{\textit{all distributions}\}$ and computing the modulus, one gets

$$\lim_{\varepsilon \to 0} \frac{b(\varepsilon)}{2\varepsilon} = (I_{Levit})^{-1/2} . \tag{1.3}$$

Again the Levit information is encoded in the modulus $b(\varepsilon)$.

In a sense, $b(\varepsilon)$ generalizes traditional notions of information. To see this, we need notions from asymptotic decision theory. Let $l(t)$ denote a loss function: a monotone increasing function of $|t|$ starting from $l(0) = 0$. Now local asymptotic minimax (L.A.M.) theory justifies the claims of $I_{Fisher}$ and $I_{Levit}$ to measure "information", by showing they place a bound on the difficulty of estimation - a so-called L.A.M. lower bound. Let $\{T_n\}$ be a sequence of estimators; this bound says

$$\liminf_{n \to \infty} \sup_{F \in N_\varepsilon(F_0)} E_F \, l\left(\frac{T_n - T(F)}{(n\, I)^{-1/2}}\right) \geq E \, l(Z) \tag{1.4}$$

where $Z$ is a standard Gaussian random variable, and $I$ is either $I_{Fisher}$ or $I_{Levit}$, depending on the functional $T$ considered (parametric or regular nonparametric). (For more information, see Levit (1974,1975), Millar (1981)). We emphasize that this bound applies for all estimate $\{T_n\}$. Equality can hold in this expression if $\{T_n\}$ is efficient (i.e. locally asymptotically minimax), although perhaps only in a limiting sense, as $\varepsilon \to 0$.

Now using (1.2) and (1.3), we can write (1.4) using $b(n^{-1/2})$ rather than $(n\, I)^{-1/2}$:

$$\liminf_{n \to \infty} \sup_{F \in N_\varepsilon(F_0)} E_F \, l\left(\frac{T_n - T(F)}{b(n^{-1/2})/2}\right) \geq E \, l(Z) . \tag{1.5}$$

In the present paper we show that an inequality like this holds for nearly arbitrary functionals, and for all sorts of nonregular rates. Our Corollary 2.3 below implies (under mild condition on $b(\varepsilon)$) the general inequality

$$\lim_{\varepsilon \to 0} \liminf_{n \to \infty} \sup_{F \in N_\varepsilon(F_0)} E_F \, l\left(\frac{T_n - T(F)}{b(n^{-1/2})/2}\right) > 0 . \tag{1.6}$$

This relation covers nonregular functionals such as the density at a point, the mode, and integral functionals such as $\int f^2$. It covers nonregular rates, as $b(n^{-1/2})$ need not be $O(n^{-1/2})$. In density estimation we will see rates $n^{-1/3}$, $n^{-1/4}$, and $n^{-1/5}$ popping out of $b(n^{-1/2})$.

We can summarize the last few paragraphs as follows. In the classical setting of parameter estimation, the "amount of information about $\theta$ in $n$ observations" is $n\, I_{Fisher}$. In the setting of

nonparametric estimation of a regular functional $T$ the information amount is $n\ I_{Levit}$. In each case the claim to represent information comes from the fact that the corresponding intrinsic difficulty of estimation is $(n\ I)^{-1/2}$, in the sense that the L.A.M. "information inequality" (1.4) holds. Now (1.2), (1.3), (1.5) and (1.6) show that $b(n^{-1/2})/2$ agrees with $(n\ I)^{-1/2}$ in regular cases while providing the same sort of "information inequality" in much more general settings. Because of the level of generality, unfortunately, we are unable to give pretty formulas for the constants on the right-hand side of the inequality (1.6); we are also unable to say that the bound may generally be attained (but see section 7). Nevertheless (1.6), viewed as a nonparametric L.A.M. lower bound, represents the most general answer we know to the question "How much information about the functional $T$ is contained in $n$ observations ($n$ large)?"

### 1.4. On "Best Constants", I

We have proposed a general technique of determining rates of convergence which seems also to give correct constants in the classical cases where those constants are known. Is it possible that the method gives the best constants in other cases?

For this we have no answer. We have focussed in this paper on easy results of a geometric character. Guiding our thinking all along have been general results such as those of Donoho (1985), which suggest that topology and geometry determine attainable rates for functionals, and of course the basic work of LeCam (1975) and Birgé (1983) showing that geometry determines optimal rates for estimating an entire density. It seems to us that a general formula for precise constants is a long way off. In this regard an enigmatic comment which Lucien Birgé made to us is apropros. Asked why rates of convergence was such a technical subject, Birgé scoffed. Refering to global density estimation rather than functional estimation, he said (in free translation) "one can easily see what the rate of convergence ought to be - the geometry gives you that quite easily. It is only when you want the precise values of the constants that difficult technique is called for."

However, in view of section 1.3, it seems that $b(n^{-1/2})/2$ is as good a starting place for obtaining precise constants as any. In section 9.4 below we are able to show that, for root-mean-square error, something close to $b(n^{-1/2})/2$ holds as a lower bound in great generality. Indeed, in section 9.4 below we get the lower bound

$$\sup_{N_B(F_0)} \sqrt{E_F(T_n - T(F_0))^2} \geq b(n^{-1/2})/4$$

valid for a wide range of exponents $q$ in moduli $b(\varepsilon) = A\varepsilon^q$, and for $n > n_0(q, b)$. Perhaps the right hand side can be improved to $b(n^{-1/2})/2$ using "difficult technique".

## 2. Preliminaries

Let $T$ be a functional of interest; for example $T(F) = f(0)$ (density at zero) or $T(F) = \int f^2$. Typically it makes no sense to discuss estimation of $T$ without a priori assumption about the unknown distribution $F$ (regularity conditions). Let $\mathbf{F}$ denote a class of distributions satisfying a given regularity condition (e.g. $\mathbf{F} = \{F : \int (f')^2 \leq 16\}$).

We define the modulus of $T$ at $F_0 \in \mathbf{F}$ by

$$b_T(\varepsilon) = \sup\{|T(F) - T(F_0)| : \quad H(F, F_0) \leq \varepsilon; \quad F, F_0 \in \mathbf{F}\}$$

This is the ordinary modulus of continuity with respect to the Hellinger distance. Typically, $b(\varepsilon) = \varepsilon^p$ where $1 \geq p > 0$. When $T$ is Frechet differentiable in Hellinger norm, one has $b(\varepsilon) = \varepsilon$; so the closer $p$ is to 1, the "smoother" $T$ is (at $F_0$).

We will need the following terminology:

- $b(\varepsilon)$ is *regular* if $\dfrac{b(c\,\varepsilon)}{b(\varepsilon)} = O(1)$ as $\varepsilon \to 0$, for each $c > 0$.
- $b(\varepsilon)$ is *regularly increasing* if, for each $c_0 > 1$, there is a $c_1 > 1$ with

$$c_0 < \liminf_{\varepsilon \to 0} \frac{b(c_1\,\varepsilon)}{b(\varepsilon)}. \tag{2.1}$$

- $b(\varepsilon)$ is Hölderian if for some $p > 0$

$$b(\varepsilon) = c\,\varepsilon^p + o(\varepsilon^p).$$

As indicated above, the Hölderian case is the usual one; a Hölderian modulus is both regular and regularly increasing.

**Theorem 2.1.** *For any estimates* $\{T_n\}$, and for each $\delta > 0$,

$$\liminf_{n \to \infty} \inf_{T_n} \sup_{F \in N_\delta(F_0)} P_F\{|T_n - T(F)| > b(c\,n^{-1/2})/2\} > \frac{1}{2}\,e^{-c^2/2} \quad \textit{for each } c > 0. \tag{2.2}$$

We will see that this theorem bounds the rate of convergence of $T_n$ to $T$. The theorem depends crucially on a lemma due to LeCam.

**Lemma 2.2. (LeCam)** *Suppose that $H(F, F_0) \le a/\sqrt{n}$. Then the minimum sum of type I and type II errors of any test between $F$ and $F_0$ based on $n$ observations is not better than $(1 - a^2/2n)^n \approx e^{-a^2/2}$.*

In short, distributions within about $O(1/\sqrt{n})$ of a fixed distribution $F_0$ in Hellinger distance are difficult to distinguish from $F_0$ by means of hypothesis tests. The expression $b(n^{-1/2})$ therefore represents the largest difference $T(F) - T(F_0)$ attainable when $F$ and $F_0$ are not distinguishable by the best test.

The formal proof of theorem 2.1 is rather short, and we give it here. Fix $\varepsilon \in (0,2)$; assume (only to simplify notation) that there is a distribution $F_\varepsilon$ with $H(F_0, F_\varepsilon) = \varepsilon$ and $T(F_\varepsilon) - T(F_0) = b(\varepsilon)$. Let $X_1, \ldots, X_n$ i.i.d. $F$, $F$ unknown, and let the estimate $T_n$ of $T(F)$ be given. Consider the statistic $S_n$ for testing the hypothesis $H_0 : F = F_0$ against $H_1 : F = F_\varepsilon$, defined by

$$S_n = T_n - T(F_0) ,$$

and the test that rejects $H_0$ if $S_n > b(\varepsilon)/2$.

From LeCam's lemma we know that for this test

$$\textit{Type I error} + \textit{Type II error} \ge (1 - \varepsilon^2/2)^n$$

so that

$$\max(\textit{Type I}, \textit{Type II}) \ge (1 - \varepsilon^2/2)^n/2 .$$

By definition

$$\textit{Type I} = P_{F_0}\{T_n - T(F_0) > b(\varepsilon)/2\}$$

and

$$\textit{Type II} = P_{F_1}\{T_n - T(F_1) \le -b(\varepsilon)/2\}$$

so that

$$\max_{F \in (F_0, F_1)} P_F\{|T_n - T(F)| \ge b(\varepsilon)/2\} \ge (1 - \varepsilon^2/2)^n/2 .$$

This bound is valid for any statistic $T_n$ and for any $\varepsilon \in (0,2)$.

Now given $c > 0$, for $n$ large enough, $c/\sqrt{n} < 2$, and so the bound applies with $\varepsilon = c/\sqrt{n}$. Also, for fixed $\delta > 0$, eventually $F_{c/\sqrt{n}} \in N_\delta(F_0)$. It follows that for all large enough $n$,

$$\inf_{T_n} \sup_{N_\delta(F_0)} P_F\{|T_n - T(F)| > b(n^{-1/2})/2\}$$

$$\geq \inf_{T_n} \max_{\{F_0, F_{c/\sqrt{n}}\}} P_F\{|T_n - T(F)| > b(c\ n^{-1/2})/2\}$$

$$\geq (1 - c^2/2n)^n/2 .$$

The result (2.1) now follows from $(1 - c^2/2n)^n \to e^{-c^2/2}$.

Let us see how theorem 2.1 bounds the rate of uniform convergence. Let $l(t)$ be a loss function: nonnegative ($l(t) \geq 0$), symmetric ($l(-t) = l(t)$), and increasing ($l(t) \leq l(t+h)$ for $t, h > 0$). For example, put $l(t) = t^2$. The *sequence* $\{\delta_n\}$ *bounds the rate of convergence if*

**Definition**

$$\liminf_{n\to\infty} \inf_{T_n} \sup_{F \in N_\varepsilon(F_0)} E_F\, l\left[\frac{T_n - T(F)}{\delta_n}\right] > 0 \tag{2.3}$$

*for every nonzero loss function.*

Because every nonzero loss function is bounded from below by a multiple of some indicator function $I_{\{|t| > a\}}$, (2.3) holds with $\delta_n = b(n^{-1/2})$ if

$$\liminf_{n\to\infty} \inf_{T_n} \sup_{F \in N_\varepsilon(F_0)} P_F\{|T_n - T(F)| > a\ b(n^{-1/2})\} > 0. \tag{2.4}$$

for each $a > 0$. Now if $b(\varepsilon)$ is regularly increasing, then by (2.1) we can find $c_1 > 0$ with $b(c_1/\sqrt{n})/2 > a\ b(n^{-1/2})$ for all large enough $n$; then (2.2) will imply (2.4). We have proved

Corollary 2.3. *Let $b(\varepsilon)$ be regularly increasing. Then $b(n^{-1/2})$ bounds the rate of convergence.*

This notion of rate bound has been used by Samarov (1976,1977) and Hasminskii (1979). Stone (1980) uses a more demanding notion of rate bound, which requires in addition to (2.2), something like

$$\lim_{\delta \to 0} \lim_{n \to \infty} \inf_{T_n} \sup_{F \in N_\varepsilon(F_0)} P_F \{ |T_n - T(F)| > \delta \, b(n^{-1/2}) \} = 1 . \tag{2.5}$$

For linear functionals at least this result can be proved following the sketch given in Stone (1980), by using Fano's lemma. We omit this exercise. Farrell (1972) uses an apparently different notion of rate bound which, however, is equivalent to (2.3).

We now turn to the evaluation of $b(\varepsilon)$ in some interesting cases. *Note to the reader:* Below we generally prove only the simplest and least technical results in the running text; most proofs are given in section 10.

### 3. Estimating the mode

Here let $\mathbf{F} = \mathbf{U} \equiv \{$*All distributions with unimodal densities*$\}$. Let $T(F)$ = (rightmost) mode of $F$. Let $F_0$ be a distribution with unique mode $m$, and with a density having two continuous derivatives at $m$, with $f_0''(m) < 0$. — so that $f_0$ is as in figure 1.

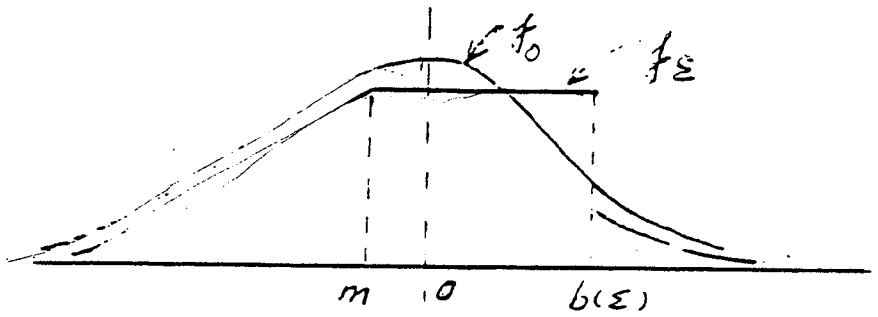


Figure 1

**Theorem 3.1. (Extremal Function)** *The distribution $F_\varepsilon \in \mathbf{U}$ maximizing $T(F_\varepsilon) - T(F_0)$ subject to $H(F_0, F_\varepsilon) \le \varepsilon$ has the density*

$$f_\varepsilon = \begin{cases} c_1^2 f(x) & x \le m \ (\le 0) \\ c_1^2 f(m) & m \le x \le b(\varepsilon) \\ c_1^2 f(x) & x > b(\varepsilon) \end{cases}$$

*where $c_1^2$ satisfies*

$$c_1^2(F(m) + [1 - F(b(\varepsilon))] + (b(\varepsilon) - m) f(m)) = 1 .$$

In short, the extremal density has a very flat top, extending the modal interval for as long as possible to the right, and thereby pulling the mode to the right.

It is not hard using this extremal function to derive an expression for the modulus:

**Theorem 3.2. (Modulus)** *Under the assumption of this section*

$$b_T(\varepsilon) = A_1 \varepsilon^{2/3} + o(\varepsilon^{2/3})$$

*where $A_1 = A_1(f_0(m), f_0'(m))$.*

The "2/5" result is due to the contributions of regions $A$ and $B$ to the Hellinger distance - the contributions from $(-\infty, m)$ and $(m + b(\varepsilon), \infty)$ being relatively unimportant. Roughly speaking, $A + B$ make up the discrepancy between a parabola and a constant. If the constant is chosen correctly, on a short interval of length $b(\varepsilon)$, the root-mean square difference between these will be only $O(b(\varepsilon)^{5/2})$. Thus pulling the mode to the right by $b(\varepsilon)$ need only cost $O(b(\varepsilon)^{5/2})$ in Hellinger distance.

Plugging $n^{-1/2}$ into this expression one obtains this

Corollary *For no estimator $T_n$ can the rate of convergence exceed $n^{-1/5}$ uniformly throughout any neighborhood of* U.

Hasminskii (1979) originally obtained the rate bound $n^{-1/5}$ via his "parametric" argument. Results of Venter (1967) show that $n^{-1/5}$ is attainable by a nearest-neighbor estimator. Our approach, by exhibiting the least favorable distribution $F_\varepsilon$, shows the kind of "stretch" of the mode that causes the rate to occur.

## 4. Estimation of a density at a point

In this section, we study density estimation under various "smoothness assumptions". This is historically one of the most studied problems in nonparametric estimation. See also Farrell (1972) and Stone (1980).

### 4.1. Decreasing density

For our first example, let **F** be

$$\mathbf{D} \equiv \{\textit{All distributions with support } [0,\infty) \textit{ and decreasing densities on } [0,\infty)\}.$$

A typical member of **D** is the exponential distribution with density $e^{-x}$, $x \geq 0$. Let $T(F) = f(\frac{1}{2})$, the density at $x = 1/2$. Let $f_0$ have a positive density at 1/2 with a continuous derivative at 1/2 that is strictly negative.

**Theorem 4.1. (Extremal Function)** *The distribution $F_\varepsilon \in \mathbf{D}$ maximizing $T(F_0) - T(F_\varepsilon)$ subject to $H(F_\varepsilon, F_0) \leq \varepsilon$ has the density*

$$f_\varepsilon(x) = \begin{cases} c^2 f_0(x) & x \notin [1/2, 1/2 + w) \\ c^2 f_0(\frac{1}{2} + w) & x \in [1/2, 1/2 + w) \end{cases}$$

*where $c^2$, $w$ satisfy*

to guarantee $F(x)$ is c.d.f.

$$c^2(F_0(\tfrac{1}{2}) + \{1 - F_0(\tfrac{1}{2} + w)\} + w\, f_0(\tfrac{1}{2} + w)) = 1$$

$$c^2 f_0(\tfrac{1}{2} + w) = f_0(\tfrac{1}{2}) - b(\varepsilon)$$

As shown by figure 2, the extremal function makes a downward jump, $f_0(\frac{1}{2}) - b(\varepsilon)$ at 1/2; the flat spot is necessary to preserve monotonicity of $f_\varepsilon$.

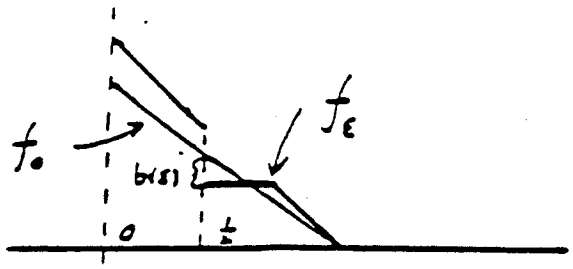


Figure 2

Theorem 4.2. (Modulus) *Under the assumptions of this subsection,*

$$b_T(\varepsilon) = A_2\varepsilon^{2/3} + o(\varepsilon^{2/3})$$

*where* $A_2 = A_2(f(\frac{1}{2})f'(\frac{1}{2}))$; *For example, if* $f_0(x) = \sqrt{2} - x$ *for* $0 \le x \le \sqrt{2}$ *then* $A_2 = 2.2199$.

The "2/3" results here derives from the contribution of the "triangle" $A$ to the distance $H(F_0, F_\varepsilon)$. - the contribution from $x \in [1/2, 1/2 + w)$ being relatively unimportant. Roughly speaking, this triangle has sides $b(\varepsilon)$, $b(\varepsilon)/f_0(1/2)$, and it contributes $O(b(\varepsilon)^{3/2})$ to $H(F_0, F_\varepsilon)$. Thus, perturbing by the downward jump changes $T(F)$ by an amount $b(\varepsilon)$; at the (small) cost of only $= b(\varepsilon)^{3/2}$ in Hellinger distance.

Corollary *The minimax rate for estimating a decreasing density at a point is no better than* $n^{-1/3}$.

This result is due originally to Kiefer (1982), who used a testing argument (he calls this a Bayes argument). That this rate is attainable follows from results of Prakasa Rao (1969).

### 4.2. Decreasing density with smoothness constraint

The extremal function of Theorem 4.1 may be considered unrealistic because it makes a jump at 1/2, and one might ordinarily expect $f$ to have a bounded derivative. Consider then $D_c = \{f \in D, f$ *abs. cont.*, $f' \ge -c$ *a.e.*$\}$ The extremal function for the new family will have to satisfy $0 \ge f_\varepsilon' \ge -c$.

**Theorem 4.3. (Extremal Function)** *The distribution $F_\varepsilon$ attaining $b(\varepsilon)$ has the density*

$$f_\varepsilon = \begin{cases} c_1^2 f_0(x) & x \in [w_1, w_2) \\ f_0(\frac{1}{2}) - b(\varepsilon) & 1/2 \le x \le w_2 \\ (f_0(\frac{1}{2}) - b(\varepsilon)) + c(x - \frac{1}{2}) & w_1 \le x \le 1/2 \end{cases}$$

*(see figure 3) where*

$$c_1^2(1 + F_0(w_1) - F_0(w_2) + (w_2 - \tfrac{1}{2})(f_0(\tfrac{1}{2}) - b(\varepsilon)) + \{(f_0(\tfrac{1}{2}) - b(\varepsilon)) + c_1^2 f_0(w_1)\}(\tfrac{1}{2} - w_1)/2 = 1$$

$$c_1^2 f_0(w_2) = f_0(\tfrac{1}{2}) - b(\varepsilon)$$

$$c_1^2 f_0(w_1) = (f_0(\tfrac{1}{2}) - b(\varepsilon)) + c(w_1 - \tfrac{1}{2})$$

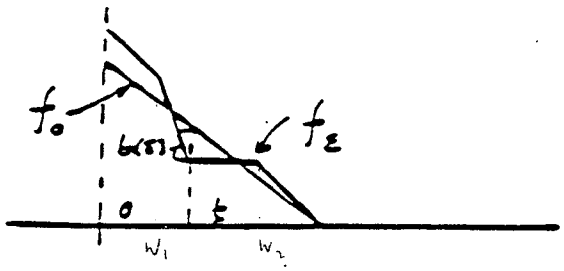


Figure 3

So instead of a jump at 1/2, the extremal function has a steep line segment just to the left of 1/2. While this segment may make the result more plausible or pleasing, however, it does not change the rates involved — only the constants.

**Theorem 4.4. (Modulus)**

$$b_T(\varepsilon) = A_3 \varepsilon^{2/3} + o(\varepsilon^{2/3})$$

*where* $A_3 = A_3(c, f_0, f_0')$ *and* $A_3 < A_2$.

Thus bounding the derivative explicitly will not improve the rates, but instead only constants, in the modulus. Kiefer (1982) contains discussion related to this effect.

### 4.3. Density under integral smoothness

Qualitative conditions on the density — such as monotonicity — are unrealistic in many situations. If so, one can use smoothness constraints involving norms on derivatives of the density to form the class F.

Let $I(F)$ denote the Fisher information of $F$;

$$I(F) = 4\int((f^{1/2})')^2$$

and let $\mathrm{I}_c$ be the set of distributions with $I(F) \le c$. Let $f_0 \in \mathrm{I}_c$ have $f_0(0) > 0$, and let $T(F) = f(0)$. Here we, unfortunately, have to settle for a result that is approximate, not exact.

**Theorem 4.5. (Extremal Function)** *The extremal density $f_\varepsilon$ satisfies*

$$\sqrt{f_\varepsilon}(x) = (1 - \varepsilon^2/2)\sqrt{f_0}(x) + \sqrt{\varepsilon^2 + \varepsilon^4/4}\,h(x) + o(\varepsilon^2)$$

*where* $h(x) = c_0 e^{-c_0^2|x|}$.

In short, the worst perturbation to $f_0$ is of the form $h$, in the sense of changing $f(0)$ a lot without changing much in Hellinger distance (figure 4).

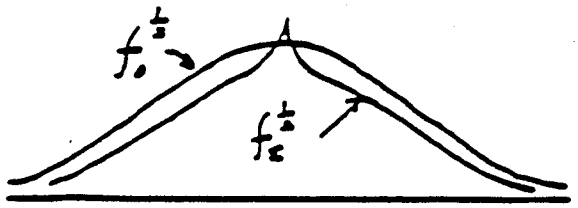


Figure 4

**Theorem 4.6. (Modulus)**

$$b_T(\varepsilon) = \sqrt{2 f_0(0)}(c - I(F_0))^{1/4}\sqrt{\varepsilon} + O(\varepsilon).$$

**Corollary** *The minimax rate for* $\mathbf{I}_e$ *is not better than* $n^{-1/4}$.

This class does not seem to have been studied before; the bound is new.

## 5. Integral functionals

Functionals such as $\int f^2$ and even $I(f) = 4\int (f')^2/f$ are important in statistical theory. The possibility of estimating them nonparametrically opens new avenues in development of confidence intervals and estimates of risk - see Donoho (1985)

The following gives a model result, for the functional $T(F) = \int f^2$. Let $\mathbf{F} = \mathbf{L}_3 \equiv \{f : \int f^3 \le C^2\}$.

**Theorem 5.1. (Modulus of $\int f^2$)**

$$b_T(\varepsilon) \le C\varepsilon .$$

In short, the modulus of the functional is *linear* , giving the possibility that the functional may be estimated at the "ordinary" $n^{-1/2}$ rate under enough regularity ($\int f^3 \le$ bound).

A parallel result holds for the functional $T(F) = I(F)$. Let $\mathbf{F} = \mathbf{I}_2(C) = \{f : \int (f^{1/2})'^2 \le C^2\}$.

**Theorem 5.2. (Modulus of $I(F)$)**

$$b_T(\varepsilon) \le C\ \varepsilon .$$

This suggests that the Fisher information might be estimable at a root-$n$ rate if $f^{1/2}$ has two weak derivatives in a quadratic mean.

Theorem 5.1 and 5.2 derive from an abstract result which suggests that $\mathbf{L}_3(C)$ and $\mathbf{I}_2(C)$ are the "natural" regularity classes for these functionals. The key point is that $\int f^2$ and $I(F)$ are both strictly convex functionals of $f^{1/2}$: they both have Gateaux differentials. Thus, for $\int f^2$ the differential is the functional $\lambda_{f_0}(h) = 4\int f_0^{3/2} h$; and for $I(F) = 4\int (f^{1/2})'^2$ the differential is $\lambda_{f_0}(h) = (-8)\int (f_0^{1/2})'' h$. If the differentials are really good approximation to the behavior of the functionals, then we expect that $T(G) - T(F_0) = \lambda_{f_0}(g^{1/2} - f_0^{1/2}) + o(H(G,F_0))$ and so that

$$b_T(\varepsilon) = \sup_{\int h^2 \le \varepsilon^2} \lambda_{f_0}(h)$$

$$= \varepsilon \sup_{\int h^2 \le 1} \lambda_{f_0}(h)$$
$$= \varepsilon \, |\lambda_{f_0}|_*;$$

here $|\lambda_{f_0}|_*$ denotes the norm of the linear functional $\lambda_{f_0}$, i.e. $|\lambda_{f_0}|_* = \sup_{|h| \le 1} \lambda_{f_0}(h)$.

In short, if $\mathbf{F} = \{f : |\lambda_f|_* \le C\}$, then we expect

$$b_T(\varepsilon) \le C\varepsilon$$

for these functionals. Theorem 5.1 is exactly of this form: we recognize $\int f^3$ as $\frac{1}{4}|\lambda_{f_0}|_*$ in this particular case. Similarly, in Theorem 5.2 we recognize $\int (f^{1/2})'^2$ as $\frac{1}{4}|\lambda_{f_0}|_*^2$. We conjecture that $|\lambda_f|_* \le C$ is the minimal regularity condition necessary to have a result of this form.

The abstract result we referred to is:

**Theorem 5.3.** *Let $T(F) = J(f^{1/2})$ where $J$ is a strictly convex functional. Suppose that $J(\alpha r)$ is defined whenever $J(r)$ is, for each $\alpha > 0$, (i.e. suppose dom($J$) is a cone with vertex 0). Suppose also that $\{f^{1/2} : |\lambda_f|_* \le C\}$ is a convex set. Then if $\mathbf{F} = \{F : |\lambda_f|_* \le C\}$, $T$ is Lipschitz over $\mathbf{F}$ in Hellinger norm, with constant $C$*

$$b_T(\varepsilon) \le C \, \varepsilon.$$

For both our applications, the cone condition on dom($J$) is easy to verify. For $\int f^2$ it just means that $\int (\alpha^2 f)^2$ is defined whenever $\int f^2$ is. The convexity of the regularity class $\{f^{1/2} : |\lambda_f|_* \le C\}$ comes from a computation:

$$\textit{for} \quad \int f^2, \quad |\lambda_f|_*^2 = \int f^3; \quad \textit{and}$$

$$\textit{for} \quad \int (f^{1/2})'^2, \quad |\lambda_f|_*^2 = \int (f^{1/2})'^{*2}.$$

In each case $|\lambda_f|_\bullet$ is a convex functional of $f^{1/2}$.

## 6. Comparison with other lower bounds.

In this section we show that $b(n^{-1/2})$ always gives rate bounds as sharp as those provided by the Farrell, Stone, and Hasminskii procedures. Intuitively, the reason is that these other procedures all give bounds $O(\delta_n)$ where $\delta_n = T(F_{1,n}) - T(F_n)$ and $F_{1,n}$ and $F_0$ are chosen subject to particular constraints. Each procedure's constraints imply that $F_{1,n}$ and $F_0$ can only be partially distinguished by a test on $n$ observations. But, due to a dictim of LeCam's (LeCam, 1973), the lack of a perfect test will then imply that $H(F_{1,n}, F_0) \le c/\sqrt{n}$ for an appropriate $c > 0$. Accordingly $\delta_n \le b(c/\sqrt{n})$. For regular $b$, this means $\delta_n = O(b(n^{-1/2}))$. Thus a sequence $\{\delta_n\}$ derived as a bound on the rate of convergence by these other methods will not tend to zero essentially more slowly than $b(n^{-1/2})$.

### 6.1. Good but not perfect tests

Let $F_{1,n}$ be a sequence of distributions converging (in some sense) to a fixed distribution $F_0$. Let $P_{1,n}$ denote the product measure on $\mathbf{R}^n$ with marginal $F_{1,n}$ and $P_{0,n}$ denote the product measure with marginal $F_0$. Below we need the likelihood ratio $L_n = \frac{dP_{1,n}}{dP_{0,n}}$. We also need to quantify the probability of error of the best test between $P_{0,n}$ and $P_{1,n}$. We define

$$\pi(P_{0,n}, P_{1,n}) = \inf_S P_{0,n}(S) + P_{1,n}(S^c). \tag{6.1}$$

$\pi$ is the minimum sum of type I and type II errors of any test between $P_{0,n}$ and $P_{1,n}$. If $S$ achieves the infimum in (6.1) the test which achieves $\pi$ decides $P_{1,n}$ if $S$ is true and $P_{0,n}$ otherwise.

For sequences of product measures $(P_{0,n})$ and $(P_{1,n})$, we say that *there is no perfect test between* $(P_{0,n})$ *and* $(P_{1,n})$ *if there is* $\alpha > 0$ *with*

$$\pi(P_{0,n}, P_{1,n}) > \alpha \qquad \textit{for all } n > n_0 \tag{6.2}$$

Note that there is an implied positive constant $\alpha$ here which must bound the sum of errors for every $n > n_0$. We also say that *there is a good test between* $(P_{0,n})$ *and* $(P_{1,n})$ *if there is* $\beta > 0$ *with*

$$1 - \beta > \pi(P_{0,n}, P_{1,n}) \qquad \text{for all } n > n_0 \tag{6.3}$$

Again note that this bounds the sum of errors uniformly in $n$. LeCam (1973) announced the result "..... *if there is a good but not perfect test between .... [$(P_{0,n})$ and $(P_{1,n})$] then [$n^{1/2}H(F_0, F_{1,n})$] stays bounded away from zero and infinity*". The result was not explicitly given in LeCam's paper; we give it here.

**Lemma. (LeCam)** *If there is no perfect test between $(P_{0,n})$ and $(P_{1,n})$ then for $c_0 = c_0(\alpha)$*

$$H(F_0, F_{1,n}) \le c_0/\sqrt{n} \qquad n > n_0. \tag{6.4}$$

*If there is a good test between $(P_{0,n})$ and $(P_{1,n})$ then for $c_1 = c(\beta)$*

$$H(F_0, F_{1,n}) \ge c_1/\sqrt{n} \qquad \text{for all } n > n_0. \tag{6.5}$$

The proof results from inequalities between the Hellinger affinity $\rho = 1 - \frac{1}{2}H$ and $\pi$; it is given in the appendix.

**Theorem 6.1** *If there is no perfect test between $(P_{0,n})$ and $(P_{1,n})$ and if $b$ is regular, then*

$$T(F_{1,n}) - T(F_0) = O(b(n^{-1/2})) \tag{6.6}$$

As indicated above, this will be a basic tool in establishing that the $b(n^{-1/2})$ bound includes other known bounds.

## 6.2. Stone's Procedure

**Theorem 6.2.** *The bounding sequence $\{\delta_n\}$ in Stone's method is of the form $\delta_n = (T(F_{1,n}) - T(F_0))/2$, where $\{F_{1,n}\}$ satisfies the constraint $E_{P_{0,n}} |\log L_n| < c_0$ for some constant $c_0 < \infty$. This constraint implies that there is no perfect test between $(P_{0,n})$ and $(P_{1,n})$. Consequently,*

*the sequence* $\delta_n \le b(c_1/\sqrt{n})/2$ *for some* $c_1 > 0$; *if* $b$ *is regular,* $\delta_n = O(b(n^{-1/2}))$.

The theorem is proved by "walking through" Stone's method, which has these steps:

[ S1 ] Define a family $\{F_{1,n}\}$ converging to $F_0$ in such a way that

$$E_{P_{0,n}} |\log L_n| < c_0 < \infty \tag{6.8}$$

where $L_n$ is the likelihood ratio defined earlier.

[ S2 ] Show that (6.8) implies there is no perfect test between $(P_{0,n})$ and $(P_{1,n})$, for some $\alpha = \alpha(c_0)$.

[ S3 ] Thus the test $\chi_n$ = "accept $F_0$ if $T_n$ is nearer to $T(F_0)$ than to $T(F_{1,n})$; reject in favor of $F_{0,n}$ otherwise" has a sum of type I and type II errors of at least $\alpha$. Define

$$\delta_n = (T(F_{1,n}) - T(F_0))/2 \tag{6.9}$$

and put $S = \{T_n - T(F_0) < \delta_n\}$. Then

$$\begin{aligned} & \textit{Type I + Type II} \\ &= P_{F_0}\{T_n - T(F_0) < \delta_n\} + P_{F_1}\{T_n - T(F_1) > -\delta_n\} \\ &= P_{F_0}(S) + P_{F_1}(S^c) \\ &\ge \pi(P_{0,n}, P_{1,n}) > \alpha. \end{aligned}$$

[ S4 ] Using $\max(a,b) \ge \text{ave}(a,b) = \text{sum}(a,b)/2$, conclude from [S3] that

$$\begin{aligned} & \inf_{T_n} \max_{F \in (F_0, F_{1,n})} P_F\{|T_n - T(F)| > \delta_n\} \\ &\ge \inf_{T_n} \operatorname*{Ave}_{F \in (F_0, F_{1,n})} P_F\{|T_n - T(F)| > \delta_n\} \\ &> \alpha/2 \end{aligned} \tag{6.10}$$

from which it follows that

$$\liminf_{n \to \infty} \inf_{T_n} \sup_{N_\varepsilon(F_0)} P_F\{|T_n - T(F)| > \delta_n\} > \frac{\alpha}{2}. \tag{6.11}$$

At this point, $\delta_n$ has become a candidate for bounding the rate of convergence; it satisfies (2.3) when $l(t) = I\{|t| > 1\}$. To extend (2.3) to all loss functions $l$, it is enough to show that it holds for $l_c(t) = I\{|t| > c\}$, $c > 0$. This requires constructing a family $\{F_{1,n}\}$ with $(T(F_{1,n}) - T(F_0)) \ge 2\, c\, \delta_n$ satisfying (6.8) with some value $c_0$ (If $T$ is linear this is rather easy to do). The crucial observation to make is that any sequence $\{\delta_n\}$ proposed as a rate bound by Stone's method satisfies (6.9) where

$\{F_{1,n}\}$ satisfies (6.8). At this point, we invoke

**Lemma (Stone):** *If* $E_{P_{0,n}} |\log L_n| < c < \infty$

$$\pi(P_{0,n}, P_{1,n}) > \alpha = (4(e^c - 1))^{-1} .$$

This lemma, in conjunction with (6.8), (6.9) and theorem 6.1, implies that $\delta_n = O(b(n^{-1/2}))$.

Stone (1980) actually uses a more demanding definition of rate of convergence than the one we are using. This means that an extra set of conditions on $\{\delta_n\}$ must be checked. However, the sequence $\{\delta_n\}$ being checked always arises from (6.8), (6.9); so it must be $O(b(n^{-1/2}))$. Stone's approach will never announce a rate bound essentially stronger than $b(n^{-1/2})$.

### 6.3. Farrell's Procedure

**Theorem 6.3.** *The bounding sequence* $\{\delta_n\}$ *in Farrell's method is of the form* $\delta_n = (T(F_{1,n}) - T(F_0))/2$ *where* $\{F_{1,n}\}$ *satisfies the constraint* $E_{P_{0,n}} L_n^2 < c_0 < \infty$. *The constraint implies that there is no perfect test between* $(P_{0,n})$ *and* $(P_{1,n})$. *Consequently,* $\delta_n \le b(c_1/\sqrt{n})/2$ *for some* $c_1 > 0$; *if* $b$ *is regular* $\delta_n = O(b(n^{-1/2}))$.

Farrell's method, as originally stated, uses a different notion of rate of convergence that we do. To make discussion simpler, we begin by discussing a "modified - Farrell" procedure, adopted to our notions.

Our modified-Farrell procedure is very similar to Stone's procedure. The key difference between the two comes at the beginning:

[ F1 ] Choose $\{F_{1,n}\} \subset \mathbf{F}$ so that

$$E_{P_{0,n}} L_n^2 < c_0 < \infty \qquad (6.12)$$

[ F2 ] Show that (6.12) implies there is no perfect test between $(P_{0,n})$ and $(P_{1,n})$, for some $\alpha = \alpha(c_0)$.

This is done by using the following result of Meyer (1977a):

**Lemma (Meyer)** *If* $E_{P_0} L_n^2 < C < \infty$

$$\pi(P_{0,n}, P_{1,n}) > \alpha$$

*where* $\alpha$ *is the root of*

$$\frac{(1-\alpha)^2}{\alpha} = C$$

The steps after [F2] are identical to those of Stone's procedure. Thus, one ends up with a rate bound of the form $\delta_n = (T(F_{1,n}) - T(F_0))/2$ where there is no perfect test between $(P_{1,n})$ and $(P_{0,n})$. By theorem 6.1, $\delta_n = O(b(n^{-1/2}))$. The key difference between Stone's procedure and the modified-Farrell procedure is thus that Stone bounds $E\,|\log L_n|$ while Farrell bounds $E\,L_n^2$. Either bound rules out the existence of a perfect test; the choice between the two should be based on ease of verifying the condition.

Farrell's original procedure for bounding rates is much less transparent than either the Stone or modified-Farrell approaches. This has to do with the notion of rate bound he employs, which is basically equivalent, but leads to more complex arguments. To avoid cluttering up the exposition, we leave the proof for the original Farrell procedure to the appendix. However, we should say that the first two steps are exactly [F1] and [F2] given above: these contain the core inequality underlying the method.

### 6.4. Hasminskii's Procedure

Hasminskii (1979) has introduced what in the authors' opinion is the technically most ambitious (and potentially the most informative) method of obtaining lower bounds. It is based on parametric rather than testing ideas. Our main result is :

**Theorem 6.4.** *The bounding sequence $\{t_n\}$ in Hasminskii's method is of the form $t_n = T(F_{1,n}) - T(F_0)$, where $\{F_{1,n}\}$ satisfies the constraint*

$$\log L_n \rightarrow_D N(\frac{1}{2} c_0, c_0^2).$$

*This constraint implies there is no perfect test of $(P_{0,n})$ versus $(P_{1,n})$. Consequently, $t_n \leq b(c_1 n^{-1/2})$ for some $c_1 > 0$; if $b$ is regular then $t_n = O(b(n^{-1/2}))$.*

To prove this, we 'walk through' Hasminskii's procedure. The steps are:

[ H1 ] Introduce a sequence $\{F_{1,n}\}$ of distributions and define the parameter family $\{F_{\theta,n}\}$ with $\Theta = [0,1]$ via

$$F_{\theta,n} = (1 - \theta)F_0 + \theta F_{1,n} \tag{6.13}$$

(other choices of $F_{\theta,n}$ and $\Theta$ are possible).

[ H2 ] Check that the family satisfies the LeCam local asymptotic normality condition

$$\sum_{i=1}^{n} \log \frac{f_{\theta,n}(X_i)}{f_{0,n}(X_i)} = (c_0\theta)\, Z + \frac{1}{2}(c_0\theta)^2 + o_p(1)$$

where $Z \sim N(0,1)$, $c_0$ is a fixed positive number, and $o_p(1)$ is uniform in $\theta \in [0,1]$.

[ H3 ] Check that the functional is linearly related to the parameter $\theta$ via

$$T(F_{\theta,n}) - T(F_0) = t_n \theta \tag{6.14}$$

(It would suffice to have an approximate linearity $T(F_{\theta,n}) - T(F_0) = t_n(\theta + o(1))$).

[ H4 ] Invoke a Hajek-LeCam style asymptotic minimax theorem to get

$$\inf_{\hat{\theta}} \max_{\theta} P_\theta\{|\hat{\theta}_n - \theta| > \delta\} > C > 0 \tag{6.15}$$

for all sufficiently small $\delta$, and all $n > n_0$.

[ H5 ] Use [H3] to conclude that

$$\liminf_{n \rightarrow \infty} \inf_{T_n} \max_{\theta} P\{|T_n - T(F)| \geq t_n \delta\} > C > 0$$

for all sufficiently small $\delta$, and all $n > n_0$.

Hasminskii's method actually gives an explicit constant $C$.

To see why this method can at best imply the bound $b(n^{-1/2})$ on the rate of convergence, consider the implications of the L.A.N. condition for the choice of $\{F_{1,n}\}$. Step [H2] implies that under $P_{0,n}$ the

log likelihood ratio $L_n$ satisfies

$$\log L_n = c_0 Z - \frac{1}{2}c_0^2 + o_p(1) \tag{6.16}$$

where $Z \sim N(0,1)$. Thus $\log L_n$ is essentially normal with mean $\frac{1}{2}c_0$ and variance $c_0^2$. If it were exactly normal, then we would have

$$E_{P_{0,n}} |\log L_n| < c_1 < \infty \tag{6.17}$$

and

$$E_{P_{0,n}} L_n^2 < c_2 < \infty \tag{6.18}$$

where, for example, $c_1 < \sqrt{\frac{5}{4}}c_0$, and $c_2$ can be taken as the second moment of a lognormal distribution with parameters $(\frac{1}{2}c_0, c_0^2)$. It would then follow by either Stone's lemma or Meyer's lemma that there is no perfect test between $(P_{0,n})$ and $(P_{1,n})$.

To handle the approximate normality (6.16) we could apply a condition such as (6.17) or (6.18) after truncation of $\log L_n$. Instead we prefer to use a direct argument based on (6.16) which was shown to us by Lucien LeCam.

**Lemma (LeCam)** *If* $\log L_n \to N(\mu,\sigma)$

$$\pi(P_{0,n}, P_{1,n}) \to E(\max(1, e^{\mu + \sigma Z}))$$

where $Z$ is $N(0,1)$.

Combining this lemma with (6.14) and theorem 6.1, we again see that $t_n = O(b(n^{-1/2}))$.

## 6.5. Remarks

[a] In every case we have seen that any choice of $\{F_{1,n}\}$ consistent with the methods of Farrell, Stone, or Hasminskii cannot yield a rate bound that is stronger than $b(n^{-1/2})$. Of course, in any particular choice of $\{F_{1,n}\}$ the resulting bound can be much *weaker;* the user of these methods might not hit

on a combination that yields $b(n^{-1/2})$ to within constants.

[b] The key analytic tools of this section are the three lemmas of Stone, Meyer, and LeCam giving bounds on $\pi$ in terms of constraints on $L_n$. We remark that other lemmas of this form are possible. For example, Samarov (1977) has a lemma showing that $E \log^2 L_n < c$ implies $\pi(P_{0,n}, P_{1,n}) > \alpha$ for some $\alpha = \alpha(c)$.

[c] Lucien LeCam has shown us a general argument that can replace all such lemmas. In order for $\pi(P_{0,n}, P_{1,n}) \to 0$, he says, we must have $L_n \to_p \infty$ in $P_{0,n}$-probability. Therefore any constraint on $L_n$ that keeps

$$P\{L_n > c\} < 1 - \varepsilon$$

for some $c > 0$, for all $n$, implies the existence of $\alpha > 0$, $\alpha = \alpha(c, \varepsilon)$, such that $\pi(P_{0,n}, P_{1,n}) > \alpha$.

By Markov's inequality, constraints $E_{P_{0,n}} L_n^2 < c_0$, $E_{P_{0,n}} |\log L_n| < c_0$ and so forth all constrain $P\{L_n < c\}$ in this way; so LeCam remarks that without any calculation one can see that they rule out the existence of perfect tests.

## 7. Attainability of the geometric lower bound

In examples of section 3 and 4, we know that the rate of convergence $b(\varepsilon)$ is *attainable*: that there exist estimates $T_n$ with

$$\liminf_{C \to \infty} \inf_{N_\varepsilon(F_0)} P_F \{ |T_n - T(F)| \le C\ b(n^{-1/2}) \} = 1$$

In fact, they are attainable by quite simple estimates -- kernel estimates, for example. Adopting the terminology of Stone (1980) we say that if $b(n^{-1/2})$ is attainable, it is the optimal rate of convergence for $T$ (over the class F).

We do not have a general argument showing that $b(n^{-1/2})$ is always attainable. Of course, in every case where the Farrell procedure has been applied and found to give attainable rates of convergence, $b(n^{-1/2})$ is attainable, and therefore optimal. What we are able to do for estimates of functionals falls short of what Birgé (1983) was able to do for estimates of the entire density. Birgé showed that one could give geometric lower bounds on the risk of density estimates and one could have a geometrically-defined estimator which attained the lower bounds within a constant factor. Thus Birgé established that the "geometric" rates were in fact optimal.

Nevertheless, in the spirit of Birgé we can give a geometric procedure for getting *attainable rates*. The procedure is to compute the modulus in a "weak" topology. For example, let $|F - G|_{KS}$ denote the Kolmogorov-Smirnov distance

$$|F - G| = \sup_t |F(t) - G(t)|$$

Then define the modulus over KS distance:

$$b_{KS}(\varepsilon; \mathbf{F}, F_0) = \sup \{ |T(F) - T(F_0)| : |F - F_0| \le \varepsilon, F, F_0 \in \mathbf{F} \}$$

This is a geometric quantity in the same spirit as the Hellinger modulus, which, in this section and section 9, we denote $b_H(\varepsilon)$.

By the inequality

$$|F - G| \le L_1(F,G) \le H(F,G)\sqrt{4 - H^2} \le 2H(F,G)$$

we have the upper bound:

$$b_{KS}(\varepsilon;\mathbf{F},F_0) \ge b_H(2\varepsilon;\mathbf{F},F_0)$$

We say the modulus $b_{KS}(\varepsilon;F_0,\mathbf{F})$ holds uniformly in a neighborhood of $F_0$ if

$$\sup_{F \in N_\delta(F_0)} \frac{b_{KS}(\varepsilon;F,\mathbf{F})}{b_{KS}(\varepsilon;F_0,\mathbf{F})} \le C < \infty$$

for $\varepsilon < \varepsilon_0$.

**Theorem 7.1.** *If the modulus $b_{KS}$ holds uniformly in a neighborhood of $F_0$ and if $b_{KS}(\varepsilon) = \varepsilon^p$ for some $0 < p \le 1$ then the rate sequence*

$$b_{KS}(n^{-1/2})$$

*is attainable.*

**Proof.** We use a minimum distance procedure as in Donoho and Liu (1987). Let $\hat{F}_n$ be minimum distance estimate, i.e. any element of **F** such that

$$|F_n - \hat{F}_n| \le \min_{G \in \mathbf{F}} |F_n - G| + \frac{1}{n}$$

Then, by an application of the triangle inequality:

$$|\hat{F}_n - F| \le |F_n - \hat{F}_n| + |F_n - F|$$

Now by definition, $|F_n - \hat{F}_n| \le |F_n - F| + \frac{1}{n}$ so

$$|\hat{F}_n - F| \le 2|F_n - F| + \frac{1}{n} = \varepsilon_n,$$

say. Thus

$$|T(\hat{F}_n) - T(F)| \le b(\varepsilon_n;F)$$

Because $b(\varepsilon;F_0)$ holds uniformly,

$$|T(\hat{F}_n) - T(F)| \le C\ b(\varepsilon_n;F_0)$$

and because $\varepsilon_n = O_P(n^{-1/2})$

$$|T(\hat{F}_n) - T(F)| = O_P(b(n^{-1/2}))$$

**Corollary** *Suppose that the modulus $b_{KS}$ holds uniformly and that $0 < a < b_{KS}(n^{-1/2})/b_H(n^{-1/2}) < b < \infty$, for all $n > n_0$. Then $b_H(n^{-1/2})$ is the optimal rate of convergence for the functional $T$.*

Unfortunately, often $b_{KS}$ is much much larger than $b_H$. For example, consider the problem of estimating the mode in section 3.

**Theorem 7.2.** *Let $F_0$ be as in section 3. The* ε-least favorable distribution within $\{G : |G - F_0| \le \varepsilon\} \cap U$ *has the density*

$$f_\varepsilon(x) = \begin{cases} f(x) & x \le w_1, \text{ or } x > w_4 \\ f(w_2) & w_1 < x \le b(\varepsilon) \\ f(w_4) & b(\varepsilon) < x \le w_4 \end{cases}$$

*such that*

$$w_4 > b(\varepsilon) > w_3 > w_2 > w_1$$
$$-P_f([w_1,w_2]) + (w_2 - w_1)f(w_2) = \varepsilon$$
$$-(w_3 - w_2)f(w_2) + P_f([w_2,w_3]) = 2\varepsilon$$
$$-P_f([w_3,b(\varepsilon)]) + (b(\varepsilon) - w_3)f(w_2) = 2\varepsilon$$
$$P_f([b(\varepsilon),w_4]) - (w_4 - b(\varepsilon))f(w_4) = \varepsilon$$

*The modulus*

$$b_{KS}(\varepsilon) = A\varepsilon^{1/3} + o(\varepsilon^{1/3})$$

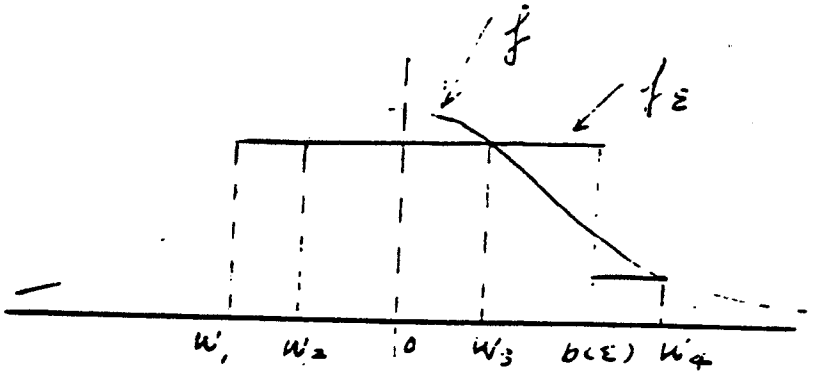


Figure 5.

Thus, our Hellinger lower bound says the optimal rate of convergence is not better than $n^{-1/5}$, but the Kolmogorov upper bound says it is not worse than $n^{-1/6}$. The optimal rate is $n^{-1/5}$.

Consider now the problem of section 4.1 estimating the density $f(\frac{1}{2})$ over the class $\mathbf{F} = \mathbf{D}$ of decreasing densities.

**Theorem 7.3.** *Let* $F_0$ be as in section 4.1. The $\varepsilon$-least favorable distribution within $\{G : |G - F_0| \le \varepsilon\} \cap \mathbf{D}$ *has the density*

$$f_\varepsilon(x) = \begin{cases} f(x) & x \le w_1, \text{ or } x > w_4 \\ f(w_2) & w_1 < x \le 1/2 \\ f(w_3) & 1/2 < x \le w_4 \end{cases}$$

*such that*

$$w_4 > w_3 > 1/2 > w_2 > w_1$$
$$P_f([w_1,w_2]) - (w_2 - w_1)f(w_2) = \varepsilon$$
$$(1/2 - w_2)f(w_2) - P_f([w_2,1/2]) = 2\varepsilon$$
$$P_f([1/2,w_3]) - (w_3 - 1/2)f(w_3) = 2\varepsilon$$
$$(w_4 - w_3)f(w_3) - P_f([w_3,w_4]) = \varepsilon$$

*The modulus of* $T(F) = f(\frac{1}{2})$ *is*

$$b_{KS}(\varepsilon) = \sqrt{-4f'(1/2)}\varepsilon^{1/2} + o(\varepsilon^{1/2})$$

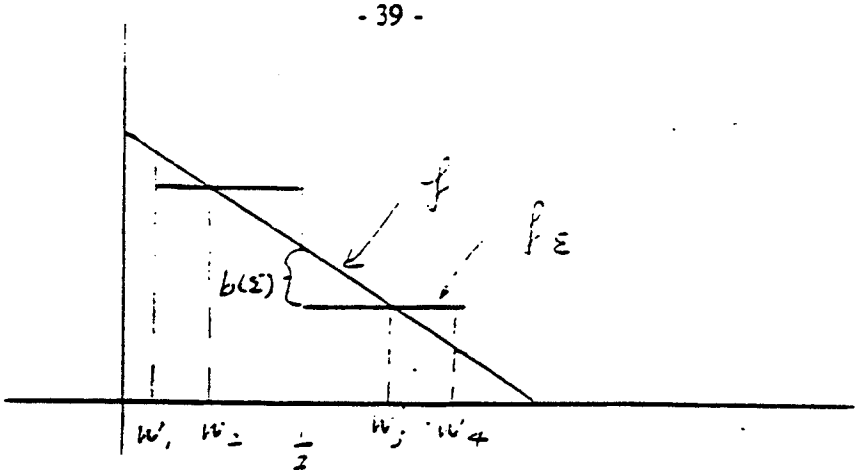


Figure 6.

Thus although our Hellinger lower bound says the optimal rate of convergence is not better than $n^{-1/3}$, the Kolmogorov upper bound says the optimal rate of convergence is not worse than $n^{-1/4}$. The optimal rate is $n^{-1/3}$.

These two examples point out that there is an "attainability gap" in our geometric viewpoint. The Hellinger lower bound is quite often the optimal rate, but we have no geometric way of showing this rate to be attainable.

A more detailed study of the "attainability gap" is possible if we consider specific functionals. Let $T(F) = \int f^2$. Then by section 5, $b_H(\varepsilon) \le \sqrt{C_1}\varepsilon$ if $F = L_3(C_1) = \{f : \int f^3 \le C_1\}$ while, by Donoho (1985), $b_{KS}(\varepsilon) \le C_2\varepsilon$ if $F = TV(C_2) = \{f : Variation\ (f) \le C_2\}$. Thus if $F = L_3(C_1) \cap TV(C_2)$, the modulus $b_{KS}$ holds uniformly,

$$\frac{b_{KS}(\varepsilon)}{b_H(\varepsilon)} < \frac{C_2}{\sqrt{C_1}} < \infty \tag{7.1}$$

and we can conclude that $n^{-1/2}$ is the optimal rate of convergence over F.

On the other hand, it appears that the $n^{-1/2}$ rate is attainable over bigger classes. Ibragimov and Hasminskii (1978, Theorem 2) show that if F admits uniformly $n^{-1/4-\delta}$-consistent estimates of a density (global $L_2$ error), then one can estimate $T$ at a $n^{1/2}$-rate, as follows. Let $\hat{f}_n$ be an estimate of the density using the initial $n/\log(n)$ observations, and let $F_n^*$ be the empirical distribution of the remaining

$n(1 - \frac{1}{\log(n)})$ observations. Estimate the functional $\int f^2 = \int f \, dF$ by

$$T_n(X_1, \ldots, X_n) = \int \hat{f}_n \, dF_n^*$$

This $T_n$ is not just uniformly $n^{1/2}$ consistent for $T$ throughout F; Ibragimov and Hasminskii show it is even efficient in the L.A.M. sense.

Now the assumption that F admit $n^{-1/4-\delta}$-consistent estimates of a density essentially means that F consists of functions with "1/2" derivative. This is somewhat better than our requirement, as $f \in TV(C_2)$ requires "1"-derivative. Thus one can do somewhat better than our purely geometric approach indicates.

Nevertheless, neither by geometric means nor by hard analysis has it been established that $b(n^{-1/2})$ is the optimal rate of convergence for $\int f^2$ over $L_3(C)$; this would require showing that $\int f^2$ can be estimated at a root-$n$ rate only assuming $\int f^3 < \infty$ i.e. with no smoothness assumptions at all. We wonder if this is an example of a true "attainability gap", where the geometric Hellinger procedure is over-optimistic.

If so, it is interesting to speculate on the reasons why the geometric bound is not attainable in functional estimation problems. Birgé (1983) considered the estimation of the entire density and showed that geometry gives optimal rates; but in a sense the entire density is a straightforward functional of the density! When we have to estimate an arbitrary nonlinear functional of the density, (mode, rate of tail decay, .... ) it perhaps is surprising that the geometry should yield the optimal rates, as it often does!

## 8. Parametric and Semi-parametric Cases

In this section, we briefly discuss the applications of $b(n^{-1/2})$ to classical models. We do not aim for the best possible results; only for a convincing sketch.

### 8.1. $b(\varepsilon)$, Fisher Information, Cramèr-Rao

Let $\mathbf{F} = \{F_\theta\}$ be a parametric family and let $T(F_\theta) = \theta$ be the functional of interest. We say that $\{F_\theta\}$ is *well-parametrized* (at $\theta_0$), if on an open interval $(-t,t)$ containing $\theta_0$, $H(F_\theta,F_0)$ is decreasing in $\theta$ for $\theta < \theta_0$ and increasing in $\theta$ for $\theta > \theta_0$, and if for $|\theta| > t$, $H(F_\theta,F_0) > \max\{H(F_{-t},F_0),H(F_t,F_0)\}$. A well parametrized family does not curve back on itself; we assume the family $\{F_\theta\}$ is well-parametrized.

The classical formula for Fisher's information assumes that $F_\theta$ has a density at $f_\theta$ and that $\sqrt{f_\theta(x)}$ has a $\theta$-derivative a.e. Then

$$I_{Fisher} = 4\int(\frac{\partial}{\partial\theta}\sqrt{f_\theta})^2 . \qquad (8.1)$$

Using *only the modulus* $b(\varepsilon)$ we can also define an information number; let the *Geometric Information* be defined by

$$(\frac{1}{4} I_{Geometric})^{-1/2} = \limsup_{\varepsilon \to 0} \frac{b(\varepsilon)}{\varepsilon} . \qquad (8.2)$$

Note that, unlike Fisher information, Geometric information can always be defined (it may be $\infty$). In nice enough situations the two are equal. We say that $\{F_\theta\}$ is differentiable in quadratic mean (DQM) at $\theta = 0$ if $F_\theta$ has a density $f_\theta$ and if there is a function $\eta_0 \in L_2(\mathbf{R})$ with

$$\int(f_\theta^{1/2} - f_0^{1/2} - \theta\,\eta_0)^2 = o(\theta^2) .$$

**Theorem 8.1** *If $\{F_\theta\}$ is well-parametrized and differentiable in quadratic mean at $\theta = 0$,*

$$I_{Geometric} = I_{Fisher}$$

*at* $\theta = 0$.

On the other hand, in less nice situations, the two notions of information differ.

**Theorem 8.2** *If* $\{F_\theta\}$ *is well-parametrized and* $I_{Fisher}$ *can be defined*

$$I_{Geometric} \geq I_{Fisher} .$$

Which is the "right" notion of information? Work of Pitman (1979) shows that $I_{Geometric}$ is the right quantity for use in the information inequality, and that $I_{Fisher}$ has meaning only in the case where $I_{Geometric} = I_{Fisher}$. Actually, Pitman did not define $I_{Geometric}$; rather, he work with a quantity he called sensitivity, defined by

$$S_0 = \lim_{\theta \to \theta_0} \inf \frac{H(F_\theta, F_{\theta_0})}{|\theta - \theta_0|} .$$

This is the rate of change of $F_\theta$ at $\theta = \theta_0$ as measured in Hellinger distance. Now note that, for a well parametrized family, and for all $\delta$ small enough

$$\inf H(F_\theta, F_{\theta_0}) \quad \textit{subject to} \quad |\theta - \theta_0| = \delta$$

is just

$$\inf \varepsilon \quad \textit{subject to} \quad b(\varepsilon) \geq \delta .$$

Therefore

$$S_0 = \lim_{\varepsilon \to 0} \inf \frac{\varepsilon}{b(\varepsilon^+)} . \tag{8.3}$$

From which it follows, comparing with (8.2)

$$4 S_0^2 \equiv I_{Geometric} \tag{8.4}$$

Now Pitman proved a number of interesting properties of $S_0$; however because of his discursive style, few mathematical statisticians seem to have regarded this work as other than a textbook. Nevertheless, reading chapter 3 of his book carefully, and piecing together discussion in sections 2, 3 and 9 of that chapter, one gets (via 8.4) a proof for theorems 8.1 and 8.2 above. More interestingly, Pitman sketches

a new derivation of Cramèr-Rao in which $I_{Geometric}$, rather than $I_{Fisher}$, plays the key role. Underlying this is an apparently new inequality, non-infinitesimal in nature, that Pitman derives in passing. We use it several places below and think it derives special notice.

**Pitman's Inequality.** Let $T_n$ be a statistic with finite variance under $F_1$ and $F_0$. Then

$$\frac{(E_{F_1}T_n - E_{F_0}T_n)^2}{2(Var_{F_1}T_n + Var_{F_0}T_n)} \leq \frac{H^2(P_{0,n}, P_{1,n})}{1 - H^2(P_{0,n}, P_{1,n})} . \tag{8.5}$$

The inequality says that if $H^2(P_{0,n}, P_{1,n})$ is small, then either $E_{F_1}T_n$ is close to $E_{F_0}T_n$, or $Var_{F_1}T_n + Var_{F_0}T_n$ is large. It is analogous in some respects to the Hammersley-Chapman-Robbins inequality (see Lehmann, 1983). Using this inequality one gets the following

**Geometric form of Cramèr-Rao.** If $Var_{F_\theta}T_n$ is continuous at $\theta = 0$ and if the bias $B(\theta) = \theta - E_{F_\theta}T_n$ is differentiable at $\theta = 0$, then

$$\frac{(1 + B'(\theta))^2}{n\, I_{Geometric}} \leq Var_{F_0}T_n . \tag{8.6}$$

Usual treatments of Cramèr-Rao involve $I_{Fisher}$; but they require additional assumptions on the family $\{F_\theta\}$ - assumptions strong enough to insure differentiability in quadratic mean and hence $I_{Geometric} = I_{Fisher}$. In short, the Cramèr-Rao inequality is "really" about $I_{Geometric}$ - *which is a quantity defined by the modulus alone.*

Our proof of (8.6) goes as follows, without loss of generality, assume there is a subfamily $\{F_\varepsilon\}$ of F, parametrized so that $H(F_\varepsilon, F_0) = \varepsilon$, $|T(F_\varepsilon) - T(F_0)| = b(\varepsilon)$. Put $\theta_\varepsilon = T(F_\varepsilon)$ for short, and rearrange (8.5) getting

$$(\theta_\varepsilon - \theta_0 + B(\theta_\varepsilon) - B(\theta_0))^2 \frac{1 - H_n^2}{4\, H_n^2} \leq \frac{1}{2}(Var_{F_\varepsilon}T_n + Var_{F_0}T_n)$$

Now

$$(\theta_\varepsilon - \theta_0 + B(\theta_\varepsilon) - B(\theta_0))^2 = (\theta_\varepsilon - \theta_0)^2 (1 + B'(\theta_0) + o(1))^2$$

as $\varepsilon \to 0$, by differentiability of $B$. Also

$$\lim_{\varepsilon \to 0} \frac{1}{2} \left( Var_{F_\varepsilon} T_n + Var_{F_0} T_n \right) = Var_{F_0} T_n$$

by continuity of $Var_{F_0} T_n$. So if we can show

$$\limsup_{\varepsilon \to 0} (\theta_\varepsilon - \theta_0)^2 \frac{1 - H_n^2}{4 H_n^2} = \frac{1}{n I_{Geometric}} \tag{8.7}$$

then, combining the last three displays, we will have proved (8.6). Now $H^2(P_{0,n}, P_{\varepsilon,n}) = 2 - 2 (1 - \frac{1}{2} H^2(F_0, F_\varepsilon))^n$, so that, by the binomial formula

$$H^2(P_{0,n}, P_{\varepsilon,n}) = n \varepsilon^2 + o(n \varepsilon^2) .$$

Now recalling how $\theta_\varepsilon$ was defined, and using (8.2)

$$\limsup_{\varepsilon \to 0} \frac{(\theta_\varepsilon - \theta_0)^2}{4 \varepsilon^2} = \limsup_{\varepsilon \to 0} \frac{b^2(\varepsilon)}{4 \varepsilon^2} = \frac{1}{I_{Geometric}} ,$$

so that

$$\frac{(\theta_\varepsilon - \theta_0)^2}{4 \varepsilon^2} \frac{1 - H_n^2}{H_n^2 / \varepsilon^2} \to \frac{1}{I_{Geometric}} \frac{1}{n} .$$

which completes the proof of (8.7), and so also of (8.6).

### 8.2. Misbehavior of Fisher Information

Classically, there are two problems with Fisher information. (1) What to say if it evaluates to 0? (2) What to say if it evaluates to $\infty$?

When $I_{Fisher}$ is zero, several things could happen. First, the parameter may not be identifiable. Second, the parameter may not actually be estimable at a root-$n$ rate. In both cases, $b(\varepsilon)$ explains the situation clearly. If the parameter is not identifiable, then $b(\varepsilon) \not\to 0$ as $\varepsilon \to 0$. If a root-$n$ rate is not possible, then $b(\varepsilon)/\varepsilon \to \infty$ as $\varepsilon \to 0$. There is a third possibility, where $I_{Fisher} = 0$ but a *faster* than root-$n$ rate is possible. Let $F_\theta$ be uniform on $[0, 1 + \theta]$. By the formula we have given above, $I_{Fisher} = 0$. However, by calculation, $b(\varepsilon) = \varepsilon^2$. While $I_{Fisher} = 0$ suggest that a root-$n$ rate is impossible, $b(n^{-1/2}) = n^{-1}$, which suggests that a $n^{-1}$ rate is possible. And, in fact the $1/n$-rate *is* achievable,

using the sample maximum $T_n = \max(X_1,.....,X_n)$. We note that in this case $I_{Geometric} = \infty$ suggesting, as it should, that a faster than $n^{-1/2}$ rate is possible.

When $I_{Fisher}$ is infinite, $I_{Geometric}$ is also infinite, and so $b(\varepsilon) = o(\varepsilon)$. This suggests that a faster than root-$n$ rate is possible. While the information bound $Var\ T_n \geq \frac{1}{n\ I} = 0$ is not incorrect, it is not very informative either. It does not indicate whether the right rate might be $n^{-1}$ or something else entirely. On the other hand, the $b(n^{-1/2})$ bound is still valid, and suggests what the optimal rate might be.

In short, in cases where the Fisher information gives a confusing or incorrect picture, the Geometric information more accurate; and the modulus is generally much more informative than either notion of information.

Our conclusion was foreshadowed by LeCam (1973), who showed that Hellinger distance is the key to understanding rates of convergence in both regular and non-regular cases. We quote

> "It is a familiar phenomenon that, when $\Theta$ is the real line, a number of well-worn regularity restrictions imply the existence of estimates $\hat{\theta}_n$ such that $n^{-1/2}(\hat{\theta}_n - \theta)$ stays bounded. Another familiar phenomenon occurs if $P_\theta$ is the uniform distribution on $[0,\theta]$. There, the usual estimates are such that $n^{-1}(\hat{\theta}_n - \theta)$ stays bounded.
>
> "In both examples the factors $n^{-1/2}$ or $n$ correspond to a certain natural rate of separation of the measures $\{P_{\theta,n}\}$ which can be described in terms of Hellinger distance.
>
> "Letting $h(s,t) = H(p_s,p_t)$, the two factors $n^{-1/2}$ and $n$ correspond now to the same rate. In both cases the statement is that $n^{-1/2}h(\hat{\theta}_n,\theta)$ stays bounded in probability."

LeCam is saying here that Hellinger distance provides a sort of universal scale on which to measure estimation errors - a scale on which all parameter estimation problems have the same rate structure. An unpublished result of LeCam's, mentioned in Birgé (1983), talks about lower bounds for errors and again shows that Hellinger distance provides a universal scale.

**Definition.** The parameter family $\{F_\theta\}$ is $(a,b)$-continuously parametrized if for each $a,b > 0$, for each $\theta_0$, and each $\varepsilon > 0$, there exists $\theta$, with $a\ \varepsilon < H(F_{\theta_0},F_\theta) < b\ \varepsilon$.

The purpose of the definition is to rule out essentially discrete parameter problems.

**Theorem (LeCam)** *Let $\{F_\theta\}$ be $(a,b)$-continuously parametrized. Then*

$$\inf_{T_n} \sup_{\theta} n\ E_{F_\theta} H^2(F_{T_n},F_\theta) > C(a,b)$$

*where $C = C(a,b)$ depends only on $a$ and $b$ and not on the problem.*

The theorem shows that for all continuously parametrized problems there is a universal lower bound on risk on the Hellinger scale -- independent of the particular problem. LeCam (1975) has also essentially shown that in finite dimensional parameter sets, there exist estimators with $n\ E\ H^2(F_{T_n},F_\theta) < \infty$.

Against this background, $b(n^{-1/2})$ may be viewed as a way of converting statements in the universal, Hellinger scale into statements in the scale of the original parameter set. This is why $b(n^{-1/2})$ is effective when the optimal rate of convergence is other than $n^{-1/2}$.

### 8.3. $b(\varepsilon)$ and the least informative families of Levit-Stein

An important notion underlying Levit's work defining information for estimation of functionals is that of a *least informative* family of distributions. The idea also arises in much recent work on semiparametric estimation, see Bickel (1982), and Begun, Hall, Huang, and Wellner (1984). In both areas, workers trace the ideas back to Stein (1956). One is interested in $T(F)$, and has *a priori* knowledge that $F \in \mathbf{F}$; but $\mathbf{F}$ is infinite-dimensional. How to calculate the information about $T(F)$? Stein's idea was that the information about $T$ at $F_0$ should be the least information about $T$ in any 1-dimensional parametric family passing through $F_0$. The family solving the minimum problem is called *least informative.*

Levit (1974,1975) was apparently first to formalize Stein's notion of semi-parametric information. He gave this definition

$$I_{Levit} = \inf \{ I_{Fisher}(F_0) : \{F_\theta\} \subset \mathbf{F}, F_\theta|_{\theta=0} = F_0, T(F_\theta) = T(F_0) + \theta, \{F_\theta\} \ DQM \ at \ \theta = 0 \} .$$

Note the condition that $F_\theta$ be differentiable in quadratic mean.

This quantity is generally the right one in cases where $n^{-1/2}$-consistent estimation is possible. Indeed, Levit (1974,1975) gave L.A.M. lower bounds using $(n\ I_{Levit})^{-1/2}$ (compare (1.4) in section 1.3). Thus the minimax risk is not generally better than $(n\ I)^{-1/2}$; there are constructions showing that often this risk may be attained.

$I_{Levit}$ can be understood from the point of view of $b(\varepsilon)$; in fact in calculating $b(\varepsilon)$ in a regular semiparametric case, one is calculating $I_{Levit}$ in a hidden fashion. To begin with, recall the definition of Geometric information

$$(I_{Geometric})^{-1/2} = \lim_{\varepsilon \to 0} \sup \frac{b(\varepsilon)}{2\varepsilon}$$

this applies equally well in the semiparametric case, where now $\mathbf{F}$ is an infinite dimensional set of probabilities, and $T$ is a functional of interest. This emphasizes the underlying unity of two situations that are commonly considered quite different. Our first fact about this situation is elementary.

**Theorem 8.3**

$$I_{Geometric} \le I_{Levit} .$$

Our second fact emphasizes the utility of Geometric information. We say that a sequence $\{T_n\}$ of estimators is *asymptotically unbiased* up to order $n^{-1/2}$ if the local bias

$$\beta_n = \sup_{N_\delta(F_0)} |E_F T_n - T(F)|$$

has $\beta_n = o(n^{-1/2})$. We say that a sequence $\{T_n\}$ of estimators has *stable variances* at $F_0$ if

$$v_n = \sup_{N_\delta(F_0)} |Var_F T_n - Var_{F_0} T_n|$$

has $v_n \to 0$ as $n \to \infty$. These two conditions impose some stability on the asymptotic distribution of

$T_n$.

**Theorem 8.4** *If $\{T_n\}$ is asymptotically unbiased to order $n^{-1/2}$ and has stable variances, then*

$$\liminf_{n \to \infty} \sup_{F \in N_\delta(F_0)} n \, E_F(T_n - T(F))^2 \geq \frac{1}{I_{Geometric}} .$$

This is a sort of information inequality for functionals and semiparametric models. It indicates that $I_{Geometric}$ measures information rigorously; it is proved via Pitman's inequality.

When are $I_{Levit}$ and $I_{Geometric}$ equal? The following condition seems natural and geometric. Note that in examples of sections 3 and 4 we were able to identify precisely the *least-favorable* distributions $F_\varepsilon$ satisfying $|T(F_\varepsilon) - T(F_0)| = b(\varepsilon)$ and $H(F_\varepsilon, F_0) = \varepsilon$. So suppose we are in a case where such least-favorable distributions exist, and suppose in addition that $b(\varepsilon)$ is a continuous strictly increasing function of $\varepsilon$ (for $\varepsilon$ small enough). (We have seen many examples where these conditions hold.) When they hold, $b(\varepsilon)$ has a continuous inverse $b^{-1}$, and we can re-parametrize, producing $\tilde{F}_\theta = F_{b^{-1}(\theta)}$. This family satisfies $T(\tilde{F}_\theta) = T(F_0) + \theta$, for $\theta > 0$.

We claim that in regular cases this *least-favorable* family $\{F_\theta\}$ is exactly Stein's *least-informative* family. That is, suppose that the family $\tilde{F}_\theta$ (recall we have only defined it for $\theta > 0$ so far) can be "continued" to negative $\theta$ so that $\tilde{F}_\theta$ is (two sidedly) differentiable in quadratic mean.

**Theorem 8.5** *Under this assumption,*

$$I_{Geometric} = I_{Levit} ,$$

*and $\tilde{F}_\theta$ is Stein's least-informative family for $T$.*

In short, computation of $b(\varepsilon)$ and the associated least-favorable family $F_\varepsilon$ contains the Stein-Levit calculation in a hidden fashion, in regular cases.

An example may help. Let $T(F) = F(0)$, the distribution function evaluated at 0. Let F = {*all*

*distributions*}; $b(\varepsilon)$ can be calculated explicitly (Liu, 1987). It turns out that $b(\varepsilon) = T(F_\varepsilon^+$ where $F_\varepsilon^+$ is the solution to

$$\max T(F) \qquad \textit{subject to} \quad H(F, F_0) \le \varepsilon .$$

The least-favorable family $F_\varepsilon^+$ has a density given by

$$f_\varepsilon^+(x) = \begin{cases} (1+\alpha)^2 f_0(x) & x \le 0 \\ (1-\beta)^2 f_0(x) & x > 0 \end{cases}$$

with $\alpha$ and $\beta$ smooth functions of $\varepsilon$.

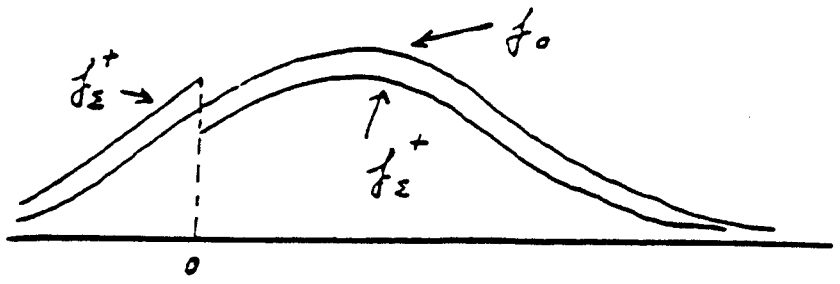


Figure 7

Now $T(F_\varepsilon^+) = [(1+\alpha)^2 - 1]\, p$, where $p = T(F_0) = F_0(0)$, and one has therefore

$$b'(\varepsilon)\,|_{\varepsilon=0} = \sqrt{4p(1-p)}$$

so that

$$I_{\text{Geometric}} = (p(1-p))^{-1} .$$

One can also solve the optimization problem

$$\min T(F) \qquad \textit{subject to} \quad H(F, F_0) \le \varepsilon ;$$

call the solution $F_\varepsilon^-$. Now

$$f_\varepsilon^-(x) = \begin{cases} (1-\gamma)^2 f_0(x) & x \le 0 \\ (1+\delta)^2 f_0(x) & x > 0 \end{cases}$$

where $\gamma$ and $\delta$ are smooth functions of $\varepsilon$.

Define now the family $G_\varepsilon$ by $G_\varepsilon = F_\varepsilon^+$ if $\varepsilon > 0$ and $= F_\varepsilon^-$ if $\varepsilon < 0$. One can check that $-\gamma(0) = \alpha'(0)$ and $-\delta'(0) = \beta'(0)$; it follows straight forwardly that $G_\varepsilon$ is (two-sidedly) differentiable in quadratic mean at $\varepsilon = 0$. Now $T(G_\varepsilon) = [(1+\alpha)^2 - 1]\,p$ for $\varepsilon > 0$ and $= [(1-\gamma)^2 - 1]\,p$ for $\varepsilon < 0$; from smoothness and monotonicity of $\alpha$ and $\gamma$ and the matching derivatives at 0, we can reparametrize $G_\varepsilon$, getting $\{\tilde{F}_\theta\}$ defined for $\theta$ in an interval about zero and satisfying $T(\tilde{F}_\theta) = p + \theta$. $\tilde{F}_\theta$ is a least-informative family for this problem. Indeed, one simply checks that $I_{Fisher} = (p(1-p))^{-1}$; then from the inequality $I_{Geometric} \le I_{Levit} \le I_{Fisher}$ and the formula $I_{Geometric} = (p(1-p))^{-1}$, one knows one has found a DQM family with $I_{Fisher} = I_{Levit}$.

This family arose from extending $\{F_\varepsilon\}$ and reparametrizing it; so the least favorable family in this case generates the least informative family.

By theorem 8.4 we may conclude that no essentially unbiased estimator of the c.d.f. $F$ at the point zero can have variance essentially smaller than $(n\,p(1-p))^{-1}$, uniformly over a Hellinger neighborhood of $F_0$. Of course, the usual estimator $F_n(0)$ attains this and is in this sense efficient.

This example is rather homely; it probably can be replaced by one using a semiparametric model of current interest; we leave this to the reader. The calculation of $b(\varepsilon)$ in the symmetric location model (Donoho and Liu, 1985) may be of interest here.

### 8.4. Finite Sample Results.

We have seen that the modulus $b(\varepsilon)$, through the quantity $I_{Geometric}$, gives information about the difficulty of classical estimation problems. Is there other information "encoded" in the modulus, about higher order terms? For example, is $b(n^{-1/2})/2$ a "better" bound than $(n\,I_{Geometric})^{-1/2}$, to which it is first order equivalent?

If so, we would expect a result saying that for all unbiased estimates $T_n$

$$\sqrt{Var_{F_0}} \ge b(n^{-1/2})/2\,. \tag{8.7}$$

Such a result cannot be hold in general. At the $N(\theta,1)$ model, exact calculation gives

$$b(\varepsilon) = \sqrt{-8 \log (1 - \varepsilon^2/2)}$$

so that

$$b(n^{-1/2})/2 = n^{-1/2} + \frac{n^{-3/2}}{8} + o(n^{-3/2}) .$$

Thus for large $n$, $b(n^{-1/2})/2 > 1/\sqrt{n}$; since at the $N(\theta,1)$ model $Var\ \bar{X} = 1/n$, (8.7) cannot hold for $T_n = \bar{X}$.

A geometric picture helps explain why $b(n^{-1/2})/2$ is incorrect at third and higher order terms.

Let $\{F_\theta\}$ be a DQM family at every $\theta$, and let $\alpha(\theta_0,\theta_1;\{P_\theta\})$ be the *arc length* between $\theta_0$ and $\theta_1$ along the curve $\sqrt{f_\theta}$ in $L_2$.

$$\alpha(\theta_0,\theta_1;\{P_\theta\}) = \int_{\theta_0}^{\theta_1} \left\| \frac{\partial}{\partial \theta} \sqrt{f_\theta} \right\| d\ \theta \tag{8.8}$$

Note that $\alpha$ is a metric on $F_\theta$, and that

$$\lim_{\theta_1 \to \theta_0} \frac{\alpha(\theta_0,\theta_1)}{H(F_{\theta_0},F_{\theta_1})} = 1 \tag{8.9}$$

for smooth families. Now define, in analogy to $b(\varepsilon)$,

$$a(\varepsilon) = \sup \{ |\theta_1 - \theta_0| : \alpha(\theta_0,\theta_1) \le \varepsilon \}$$

Because of (8.9), $\frac{a(\varepsilon)}{b(\varepsilon)} \to 1$ as $\varepsilon \to 0$, so the two are first order equivalent. On the other hand,

$$\alpha(\theta_0,\theta_1) \ge H(F_{\theta_0},F_{\theta_1}) \tag{8.10}$$

because $\alpha$ is the distance from $\sqrt{f_{\theta_0}}$ to $\sqrt{f_{\theta_1}}$ following a path in F, while $H$ is the distance from $\sqrt{f_{\theta_0}}$ to $\sqrt{f_{\theta_1}}$ along a straight line segment — $t\sqrt{f_{\theta_0}} + (1-t)\sqrt{f_{\theta_1}}$ — a geodesic of $L_2(\mathbf{R})$. As probability distributions must lie on the sphere $\int (\sqrt{f_\theta})^2 = 1$, a path in F must always be longer than a line segment, hence (8.10).

A consequence of (8.10) is that we always have

$$a(\varepsilon) \le b(\varepsilon) ,$$

so that $a(n^{-1/2})/2$ is a strictly smaller lower bound than $b(n^{-1/2})/2$.

On the other hand, we have the formula

$$a\,(n^{-1/2})/2 = (n\ I)^{-1/2}$$

when $F_\theta$ is a location family. For in this case $\| \frac{\partial}{\partial \theta} \sqrt{f_\theta} \| = constant = \sqrt{4\ I}$, and so

$$\alpha(\theta_0,\theta_1) = (\theta_0 - \theta_1)\ 2\ \sqrt{I}\ .$$

Consequently, in the location family case

$$b\,(n^{-1/2})/2 > (n\ I)^{-1/2}$$

for the purely geometric reason that paths restricted to the unit sphere in $L_2$ are longer than geodesics. A detailed analysis shows that even if the family $\sqrt{f_\theta}$ is part of a geodesic on the sphere, $b\,(n^{-1/2})/2$ and $(n\ I)^{-1/2}$ differ by a term $c\ n^{-3/2}$, where the coefficient is due to curvature of the sphere in which $\sqrt{f_\theta}$ lies.

We thus see that curvature effects associated with the Hellinger viewpoint prevent $b\,(n^{-1/2})/2$ from giving information about third and higher order asymptotics. Nevertheless, we find it remarkable that the formula $b\,(n^{-1/2})/2$ gives information accurate to first order in such a wide variety of cases.

## 9. Discussion

### 9.1. Computing the modulus

We should not leave the impression that the Hellinger modulus is always easy to compute. The examples where we have had success all had the following structure: $T$ was a nice functional of $f^{1/2}$ and the regularity class F was also nicely expressed in terms of $f^{1/2}$. The reason for this, naturally, is that Hellinger distance is just the $L_2$ - distance between square roots of densities. Thus if everything can be nicely written in terms of $f^{1/2}$, one may end up with an optimization problem in $L_2(\mathbf{R})$, and so all the tools of classical analysis become available. In the proofs of theorems 4.5 and 4.6, for example, the extremal functions are solutions of an ordinary differential equation, linear and with constant coefficients; the solutions are quite easy using Fourier analysis.

Unfortunately, when the functional of interest is highly nonlinear with respect to the square root of the density, one can not expect to obtain more than *bounds* on the modulus. Of course, lower bounds are always available. A specific family $\{F_\varepsilon\}$ of distributions in F indexed by $\varepsilon$ with $H(F_\varepsilon,F_0) = \varepsilon$ gives a lower bound:

$$b(\varepsilon) \geq |T(F_\varepsilon) - T(F_0)|$$

by definition.

It may turn out that it is much easier to work with (say) $L_1$ or $L_2$ distance than with Hellinger distance. In such cases, the modulus in one of those other distances may furnish a bound on the Hellinger modulus. Thus, suppose one can compute the $L_1$-modulus of $T$. By the standard inequality

$$H^2(F,G) \leq L_1(F,G) \leq 2\,H(F,G) \tag{9.1}$$

(LeCam, 1986 page 48) one has the bounds

$$b_{L_1}(\varepsilon) \geq b_H(\varepsilon^2), \quad b_H(2\varepsilon) \geq b_{L_1}(\varepsilon) \tag{9.2}$$

Examples in Liu's thesis (Liu, 1987), show that by using the remarks of the last two paragraphs one can get useful information even in problems where the Hellinger modulus is not computable. The examples fall in the area of "ill posed" estimation problems. For example, let $G$ be a scale mixture of exponentials

$$G(t) = \int E(t/s)\, dF(s) \tag{9.3}$$

where $E(t) = 1 - e^{-t}$, $t \geq 0$, and $F$ is the mixing distribution. Suppose we have data $Y_1, \ldots, Y_n$ i.i.d. $G$ and want to recover the mixing distribution. The problem is discussed in Jewell (1982).

Let $T(G) = F(1)$. Let $\mathbf{F} = \{G : G \text{ is a scale mixture of exponentials}\}$. The problem of recovering $F$ is so highly ill-posed that in fact $b_H(\varepsilon)$ goes to zero slower than any power of $\varepsilon$:

$$\frac{b(\varepsilon)}{\varepsilon^r} \to \infty \qquad as \quad \varepsilon \to 0$$

for each $r > 0$. It follows that no ordinary rate of convergence $n^{-r}$ is possible in this problem, if "rate" is interpreted in the local minimax sense. Thus recovery of a mixing measure $F$ without a-priori-information on $F$ is essentially the hardest possible estimation problem.

We remark that $T$ is a nonlinear functional of $g^{1/2}$; we have no way of computing the modulus in Hellinger metric exactly. We *can* exhibit, for each $r > 0$, a sequence $\{F_n^{(r)}\} \in \mathbf{F}$ which shows that the $L_1$-modulus of $T$ is not $O(\varepsilon^r)$. It then follows from (9.2) that the Hellinger modulus is not $O(\varepsilon^r)$, either.

## 9.2. Using Other Metrics

Are there other metrics than the Hellinger in which it would be interesting to compute the modulus? We have just seen an example applying the $L_1$-modulus. One might suppose, in view of the importance of $\pi$ to the lower bound procedures and the relation $L_1(F,G) = 2 - 2\pi(F,G)$, that $b_{L_1}$ would be just as useful as $b_H$. Indeed $b_{L_1}(n^{-1/2})$ does bound the rate of convergence (just use (9.2) and

Corollary 2.3). However, from the inequalities (9.1) and (9.2) one can also see that $b_{L_1}$ may be much smaller than $b_H$. This may be connected with the fact that $L_1(P_{0,n}, P_{1,n})$ bears no simple relation with $L_1(F_{0,n}, F_{1,n})$, or at least nothing as simple as what holds between the Hellinger distance $H(P_{0,n}, P_{1,n})$ and $H(F_{0,n}, F_{1,n})$.

In this respect one thinks of the Kullback-Leibler number $K(F_1;F_0) = \int \log \frac{f_0}{f_1} f_0$, which satisfies

$$K(P_{1,n}, P_{0,n}) = n\ K(F_{1,n}, F_{0,n})\ . \tag{9.4}$$

Define the Kullback-Leibler modulus

$$b_{KL}(\varepsilon) = \sup\{\, |T(F) - T(F_0)| \,:\, K(F;F_0) \geq \varepsilon^2 \}\ .$$

We claim that if $b_{KL}(\varepsilon)$ is regularly increasing then $b_{KL}(n^{-1/2})$ bounds the rate of convergence of estimates $T_n$ to $T$. Indeed, consider what we call "Samarov's method" for bounding rates of convergence: one proceeds exactly as in the Stone or modified-Farrell methods, only constraining $E_{P_{0,n}} \log L_n$ rather than $E_{P_{0,n}} |\log L_n|$ or $E_{P_{0,n}} L_n^2$ as Stone and Farrell do. This method gives a sequence $\{\delta_n\}$ bounding the rate of convergence; suppose that the constraint in place when defining $\{\delta_n\}$ was $E_{P_{0,n}} \log L_n < c_0$. Now notice that

$$E_{P_{0,n}} \log L_n = -K(P_{1,n}, P_{0,n})$$

and so from (9.4) we see that

$$K(F_{1,n}, F_0) \geq -c_0/n\ .$$

Thus

$$b_{KL}(\sqrt{c_0/n})/2 = \sup\{\delta_n : \delta_n = (T(F_{1,n}) - T(F_0))/2 \quad \textit{and} \quad E_{P_{0,n}} \log L_n \leq c_0\} \tag{9.5}$$

Thus $b_{KL}(n^{-1/2})$ does bound the rate of convergence (assuming every sequence $\{\delta_n\}$ does; but this is essentially equivalent to $b_{KL}$ being regularly increasing). And $b_{KL}$ shows exactly what rate bounds are possible with Samarov method.

For another example of this kind, consider the $\chi^2$-discrepancy

$$\chi^2(F_1, F_0) = \int \frac{(f_1 - f_0)^2}{f_0} .$$

Define the $\chi^2$-modulus

$$b_{\chi^2}(\varepsilon) = \sup \{ |T(F_1) - T(F_0)| : \chi^2(F_1, F_0) \le \varepsilon^2 \} .$$

This modulus is closely connected with the modified-Farrell's method. First, there is the identity

$$1 + \chi^2(P_{1,n}, P_{0,n}) = E_{P_{0,n}} L_n^2 .$$

Second, there is the approximate relation

$$\chi^2(P_{1,n}, P_{0,n}) = n \ \chi^2(F_{1,n}, F_{0,n}) \tag{9.6}$$

which can be derived by calculating formally and dropping "high order terms". We do not go into details, but we conjecture that the constraint $E_{P_{0,n}} L_n^2 < c_0$ can be shown to imply that $\chi^2(F_{1,n}, F_0) \le \sqrt{c_0(1+\varepsilon)/n}$, $n > n_0(\varepsilon)$. (This would certainly be true, of course, if precise equality held in (9.6)). If the conjecture is true then one has results for $b_{\chi^2}$ analogous to those of $b_{KL}$: $b_{\chi^2}(n^{-1/2})$ provides a bound on the rate of convergence, and $b_{\chi^2}(\sqrt{c_0(1+\varepsilon)/n})/2$ would be bigger than any $\delta_n$ attained from the modified-Farrell's method with constraint $E_{P_{0,n}} L_n^2 < c_0$.

These remarks suggest that the user of the Samarov or modified-Farrell method who makes the cleverest choice of family $\{F_{1,n}\}$ is basically computing $b_{KL}$ or $b_{\chi^2}$ up to constants. This observation could be useful in case $b_{KL}$ or $b_{\chi^2}$ happen to be easy to compute in a problem where $b_H$ is hard to compute.

Because of the ease of many calculations for the Hilbert space $L_2(\mathbf{R})$, one might suppose that the $L_2$-modulus could be of use in calculating rates of convergence. This would particularly be true if the functional of interest is linear in terms of $f$ rather than in terms of $f^{1/2}$.

As an example, suppose we are interested in the functional $\int (f^{(k)})^2$. Analysis of this functional in terms of $f^{1/2}$ leads rather quickly to computations, owning to the fact that $f^{(k)}$ is nonlinear in $f^{1/2}$. On the other hand viewed as a functional of $f$, $\int (f^{(k)})^2$ is strictly convex and has support functional $\lambda_{f_0}(h) = 2 \int f^{(k)} h^{(k)}$. If $f$ has $2k$ derivatives in $L_2$ then integration by parts gives $\lambda_{f_0}(h) = (-1)^k \; 2 \int f^{(k)} h$. Thus by the same sort of argument as in Theorem 5.3, on the set $\mathbf{F} = L_2(C)$ of densities with $2k$ derivatives in $L_2$, $\int (f^{(2k)})^2 \leq C$, we have $b_{L_2}(\varepsilon) \leq \sqrt{C}\ \varepsilon$.

To make use of this in establishing a rate bound, we need inequalities between Hellinger and $L_2$ distance. In general, no such inequalities exist, but (as for example Birgè (1984) has noted) at a specific $f_0$, and over a compact set, we can get such inequalities.

For example, let $\mathbf{F} = L^0_{2,k}(C)$ the set of densities supported on [0,1] with $2k$ derivatives in $L_2$, $\int (f^{(2k)})^2 \leq C$, and boundary conditions $f^{(k+l)}(0) = f^{(k+l)}(1) = 0 \;\; (l = 0,\ldots\ldots k)$. At $F_0 \in \mathbf{F}$ that is bounded away from zero and infinity, we can find constants $A(F_0), B(F_0)$, with, for any $F_0 \in \mathbf{F}$

$$A \; H(F,F_0) \leq L_2(F,F_0) \leq B \; H(F,F_0) .$$

It then follows that

$$b_{L_2}(\frac{\varepsilon}{B}) \leq b_H(\varepsilon) \leq b_{L_2}(\frac{\varepsilon}{A}) . \tag{9.7}$$

This comment is dready interest for the functional $\int (f^{(k)})^2$, as it shows that $b_H(\varepsilon) = \varepsilon$ when $\mathbf{F} = L^0_{2,k}(C_1)$. This suggests that $\int (f^{(k)})^2$ might be estimable at root-$n$ rate for $f \in \mathbf{F}$. Actually, the attainability argument

of section 7 shows that if $\mathbf{F} = \{f : f^{(2k)} \in BV(C_2)\}$ then $b_{KS}(\varepsilon) \leq C_2 \varepsilon$ and so a root-$n$ rate is attainable if $f$ has $f^{(2k)} \in BV$, without any assumption of compact support or boundary condition at all.

It follows that $n^{-1/2}$ rate is optimal on $BV(C_1) \cap L_{2,k}(C_2)$.

The $L_2$ modulus is also of interest for determining case where the optimal rate is not $n^{-1/2}$. Indeed, suppose that we are interested in the functional $\int (f^{(k)})^2$, but now we know only that $f$ has $k < m < 2k$ derivatives in $L_2$, $\int (f^{(k)})^2 \leq C$. To bound the modulus, one ends up with the optimization

problem

$$\sup 2 \int f^{(k)} h^{(k)}$$

subject to

$$\begin{cases} \int h^2 \le \varepsilon^2 \\ \int (h^{(m)})^2 \le C \\ \int h = 0 \end{cases}$$

This is a problem of maximizing a functional subject to two quadratic constraints and a linear equality constraint. By passing to the Fourier domain, one gets the problem

$$\sup 2 i^k \int \omega^{2k} \hat{f}\, \overline{\hat{h}}$$

subject to

$$\begin{cases} \int \hat{h}^2 \le \varepsilon^2 \\ \int \omega^{2m} \hat{h}^2 \le C \\ \hat{h}(0) = 0 \end{cases}$$

From this point on, as in our proof of theorem 4.5, 4.6, one applies the method of Lagrange Multipliers to calculate

$$b_{L_2} = \varepsilon^{\frac{m-k}{k}}.$$

Establishing inequality (9.7) over $L^0_{2,m}(C)$ at an appropriate $f_0$ then gives

$$b_H(\varepsilon) = \varepsilon^{\frac{m-k}{k}}$$

as well.

## 9.3. Relation to Robustness

There are still other kinds of discrepancies to use in computing the modulus. Using certain ones, we get insights about the *robustness* of $T$ rather than its stochastic properties.

Suppose we use the Prohorov metric $Proh(F,G) = \inf \{\delta : F[S^\delta] \le G[S] + \delta \text{ for all } S\}$ where $S^\delta$ is the set of points at most $\delta$ away from some point in $S$. Let $\mathbf{F} = \{\textit{all distributions}\}$, and compute $b_{Proh}(\varepsilon)$. If $b_{Proh}(\varepsilon) \to 0$ as $\varepsilon \to 0$, Hampel (1968) says that $T$ is *qualitatively robust*, in the sense that changes in the underlying distribution which are small in Prohorov distance cause small changes in $T$. More generally, Donoho and Liu (1985) say that $T$ is *qualitatively robust with respect to* $\mu$-*perturbations* if the modulus of $b$ in $\mu$-discrepancy, $b_\mu(\varepsilon)$, goes to zero as $\varepsilon \to 0$. Thus functionals $T$ with $b_H(\varepsilon) \to 0$ as $\varepsilon \to 0$ are qualitatively robust with respect to Hellinger perturbations. Suppose we use the Gross-Errors discrepancy

$$GrossErrors(F_1,F_0) = \inf \{\varepsilon : (1-\varepsilon)F_0[S] \le F_1[S] \quad \text{for all } S\}$$

If $b_{GrossErrors}(\varepsilon)/\varepsilon \to \gamma^*$ as $\varepsilon \to 0$, and if $\gamma^* < \infty$, Hampel (1968) says that $T$ has a finite *gross-errors sensitivity* $\gamma^*$; $\varepsilon$-perturbations of $F_0$ cause only a correspondingly small change in $T$:

$$|T((1-\varepsilon)F_0 + \varepsilon H) - T(F_0)| \le (\gamma^* + o(1))\varepsilon .$$

More generally, Donoho and Liu (1985) say that $T$ has a finite $\mu$-sensitivity of $F_0$ if it is locally $\mu$-Lipshitz at $F_0$, or, what is the same thing

$$b_\mu(\varepsilon)/\varepsilon \to \gamma^* \qquad as\ \varepsilon \to 0 .$$

Of course, this use of the modulus differs from that of previous sections because we are taking $\mathbf{F} = \{\textit{all distributions}\}$. Most of the functionals we have been interested in are not qualitatively robust in this sense - arbitrarily small perturbations of $F$ (even in Hellinger distance) can make them quite large. Only for "regular functionals" do these notions make sense. On the other hand, in regular cases there are some interesting connections: the Hellinger sensitivity of a regular functional is $\gamma^* = (\frac{1}{4}I)^{-1/2}$, where $I = I_{Geometric}$.

In short, the modulus is involved in understanding rates of convergence *and* robustness. It arose in robustness first; we have borrowed the name "$b(\varepsilon)$" from robustness where it originates in the work

of Hampel (1968). Interestingly, the first work we know of computing $b(\varepsilon)$ was in Huber's (1964) work where he computes the modulus of the median over Gross Errors neighborhoods and shows that the median is bias minimax; however, he does not define the modulus explicitly.

### 9.4. On "Best Constants", II

Section 8 shows that $b(n^{-1/2})/2$ holds as a lower bound on root-mean-square-error in the classical cases; one might wonder if $b(n^{-1/2})/2$ plays a similar role in general. We have as yet no result to this effect. However, by applying Pitman's inequality, we get

**Theorem 9.1** *Let $b(\varepsilon)$ be Hölderian with exponent $q \in (0,1]$. Then for any sequence of estimators $\{T_n\}$ we have for $n > n_0(b,q)$ the inequality*

$$\sup_{N_\varepsilon(F_0)} \sqrt{E_F(T_n - T(F))^2} \geq b(n^{-1/2})\,\xi(q)$$

*where $\xi(q)$ is effectively computable. Also, $\xi(q) > 1/4$ for $q \in (0,1]$.*

The range $0 \leq q \leq 1$ covers the $n^{-1/5}$, $n^{-1/4}$, $n^{-1/3}$, $n^{-2/5}$, and $n^{-1/2}$ rates of convergence which often occur.

## 10. Proofs:

*"There are good reasons why the theorems should all be easy and definitions hard. ..... a single principle can masquerade as several difficult results; the proofs of many theorems involve merely stripping away the disguise. The definitions, on the other hand, serve a twofold purpose: they are rigorous replacements for vague notions, and machinery for elegant proofs"*

Michael Spivak

**Proofs of Theorem 2.1.**

*Proof of (2.2)*

Without loss of generality, let $F_n$ be a distribution such that $d_H(F_0,F_n) = n^{-1/2}$ and $b(n^{-1/2}) = |T(F_n) - T(F_0)|$. Then, for any estimator $T_n$ for estimating $T(F)$, let's consider a testing statistic $Z_n$ for testing the hypothesis $H_0$: $F_0$ against the hypothesis $H_{1,n}$: $F_n$ such that

$$Z_n = |T_n - T(F_0)|$$

and $H_0$ will be rejected when $Z_n > b(n^{-1/2})/2$.

From lemma 2.2 we know that

$$Type\ I\ error + Type\ II\ error \geq (1 - \frac{1}{n})^n$$

which implies

$$Type\ I\ error \geq (1 - \frac{1}{n})^n/2$$

or

$$Type\ II\ error \geq (1 - \frac{1}{n})^n/2$$

equivalently, they are saying

$$P_{F_0}\{|T_n - T(F_0)| > b(n^{-1/2})/2\} \geq (1 - \frac{1}{n})^n/2 \qquad (2.1)$$

or

$$P_{F_n}\{|T_n - T(F_0)| \le b(n^{-1/2})/2\} \ge (1 - \frac{1}{n})^n/2 \tag{2.2}$$

Since

$$P_{F_n}\{|T_n - T(F_0)| \le b(n^{-1/2})/2\} \le P_{F_n}\{|T_n - T(F_n)| > b(n^{-1/2})/2\} \tag{2.3}$$

Therefore, from (2.1), (2.2), (2.3), it is clear that

$$\inf_{T_n} \sup_{F \in N_\varepsilon(F_0)} P_F\{|T_n - T(F)| > b(n^{-1/2})/2\} \ge (1 - \frac{1}{n})^n/2$$

and consequently,

$$\liminf_{n \to \infty} \inf_{T_n} \sup_{F \in N_\varepsilon(F_0)} P_F\{|T_n - T(F)| > b(n^{-1/2})/2\} \ge e^{-1}/2$$

*Proof of (2.3)*

We can use the similar argument as in the previous proof, but using a different testing statistic (for each $\delta$)

$$Z_{\delta,n} = |T_n - T(F_0)|$$

and $H_0$ will be rejected when $Z_{\delta,n} > b(\frac{\delta}{n^{1/2}})/2$.

Therefore, from lemma 2.2 and from the similar argument as in the proof of (2.2), we can see that, when $\delta$ is small,

$$\liminf_{n \to \infty} \inf_{T_n} \sup_{F \in N_\varepsilon(F_0)} P_F\{|T_n - T(F)| > b(\frac{\delta}{n^{1/2}})/2\} \ge e^{-\delta^2}/2.$$

As $\delta$ tends to zero, we shall have

$$\lim_{\delta \to 0} \liminf_{n \to \infty} \inf_{T_n} \sup_{F \in N_\varepsilon(F_0)} P_F\{|T_n - T(F)| > b(\frac{\delta}{n^{1/2}})/2\} \ge 1/2.$$

*Proof of (2.4)*

For each $c$, use the testing statistic

$$Z_{c,n} = |T_n - T(F_0)|$$

and reject $H_0$ when $Z_{c,n} > b(c\ n^{-1/2})/2$. Then, follow the same argument as in the proof of (2.3). It follows

$$\lim_{n\to\infty} \inf_{T_n} \sup_{F \in N_\varepsilon(F_0)} P_F\{|T_n - T(F)| > b(\frac{c}{n^{1/2}})/2\} \geq e^{-c^2}/2 > 0 .$$

Which gives (2.4).

**Proof of Theorem 3.1**

see Liu (1987).

**Proof of Theorem 3.2**

see Liu (1987)

**Proof of Theorem 4.1**

see Liu (1987)

**Proof of Theorem 4.2**

see Liu (1987)

**Proofs of Theorem 4.5 and 4.6**

Let $r = f^{1/2}$, $r_0 = f_0^{1/2}$ be square root of desities. The optimization problem of finding the extremal function $f_\varepsilon$ is to

by defn of $b(\varepsilon)$

$$\max\ f(0) - f_0(0) \qquad s.t. \quad \begin{cases} I(f) \leq C \\ \int r^2 = 1 \\ d_H(f, f_0) \leq \varepsilon \end{cases}$$

Since for any desity $f$ with $0 < d_H(f, f_0) = \delta \leq \varepsilon$ we have the following representation

$$r = f^{1/2} = A\, r_0 + B\, h$$

where $A = (1 - \frac{\delta^2}{2})$, $B = \sqrt{\delta^2 - \frac{\varepsilon^2}{4}}$, $\int h\, r_0 = 0$, and $\int h^2 = 1$.

Using this representation, we can see that

$$\begin{aligned} & f(0) - f_0(0) \\ &= (A^2 - 1)\, r_0^2(0) + 2\, A\, B\, r_0(0)\, h(0) + B^2\, h^2(0) . \end{aligned}$$

1. Since $A$, $B$ are positive. Therefore, to maximize $(f(0) - f_0(0))$ is equivalent as to maximize $h(0)$.

Moreover, since

$$\begin{aligned} \int(\dot{r})^2 &= \int \{A^2 \dot{r}_0^2 + 2\, A\, B\, \dot{r}_0 \dot{h} + B^2 \dot{h}^2\} \\ &= \frac{A^2}{4} I_0 + 2\, A\, B \int \dot{r}_0 \dot{h} + B^2 \int \dot{h}^2 \\ &= \frac{A^2}{4} I_0 - 2\, A\, B \int \ddot{r}_0 h + B^2 \int \dot{h}^2 . \end{aligned}$$

Which implies

2.

$$\begin{aligned} \int \dot{h}^2 &= \frac{1}{B^2} \{\int \dot{r}^2 - \frac{1}{4} A^2 I_0 + 2\, A\, B \int \ddot{r}_0 h\} \\ &= \frac{1}{B^2} \{\frac{1}{4} C - \frac{1}{4} A^2 I_0 + 2\, A\, B \sqrt{\int \ddot{r}_0^2}\} \\ &= C_2(\delta) . \end{aligned}$$

Thus, the problem now becomes to

$$\begin{aligned} \max \quad & h(0) \ (= h_\delta(0)) \qquad \text{s.t.} \\ & \int \dot{h}^2 \le C_2(\delta) \\ & \int h^2 = 1 \\ & \int h\, r_0 = 0 \\ & r \ge 0 . \end{aligned}$$

After solving this problem, and we will show that $h_\delta(0)$ increases as $\delta$ increases. Thus, $b(\varepsilon)$ will be obtained in this way.

(I) **Upper bound** To simplify the computation, let's relex the othogonal constraint, positivity constraint, and the unitary constraint for a while in order to get an upper bound for the above optimization problem. Now we want to solve

$$\max \; h(0) \qquad s.t.$$
$$\int \dot{h}^2 \le C_2(\delta)$$
$$\int h^2 \le 1 .$$

Furthermore, since for any solution $h_1$ of problem (I), $h = (h_1 + h_2)/2$ where $h_2(x) = h_1(-x)$ is also a sloution of problem (I). So there is no loss of generality to assume the solution of problem (I) is symmetric. Now, if we transform the above problem into frequency domain by Fourier transformation, then the problem becomes to

$$\max \int g \qquad s.t.$$
$$\int \omega^2 \, |g|^2 \le C_2(\delta)$$
$$\int |g|^2 \le 1$$
$$g \;\; is \; real \; .$$

By considering the Lagrange multipliers, we have

$$J(g) = -\int g + \int (\lambda + \mu \, \omega^2) \, g^2 - (\lambda + \mu \, C_2(\delta)) .$$

Let $s$ be any real valued function with $|s|_{L^2} = 1$. Since

$$\frac{dJ(g + t\, s)}{dt} \Big|_{t=0} = \int s \, ((\lambda + \mu \, \omega^2) \, 2 \, g - 1) . \qquad (4.1)$$

Thus, by setting (4.1) to zero, we must have

$$g = \frac{1}{2 \, (\lambda + \mu \, \omega^2)}$$
$$= \frac{1}{A + B \, \omega^2} .$$

Using inverse Fourier transform, we can see that the solution for problem (I) is

$$h = C_0 \, e^{-C_1 |x|} \qquad for \; some \; positive \; numbers \; C_0, C_1 .$$

By setting

$$\int h^2 = 1 \qquad and$$

$$\int h^2 = C_2(\delta) ,$$

we have

$$\begin{cases} C_0 = C_2^{1/4}(\delta) \\ C_1 = C_2^{1/2}(\delta) \end{cases}.$$

Plugging in the formula $r = A\ r_0 + B\ h$, we have

$$\begin{aligned} (\Delta) &\equiv f(0) - f_0(0) \\ &= (A^2 - 1) r_0^2(0) + 2\,A\ B\ r_0(0)\ h(0) + B^2\ h^2(0) . \end{aligned} \tag{4.2}$$

From Vaxima (a symbolic calculation package developed by MIT), we get

$$(\Delta) = \sqrt{2\ f_0(0)}\ (4\ C' - I_0)^{1/4}\ \delta^{1/2} + [\sqrt{4\ C' - I_0}/2]\ \delta + O(\delta^{3/2}) . \tag{4.3}$$

The first order term gives the information that $(\Delta)$ increases as $\delta$ increases. Hence, we get an upper bound for $b(\varepsilon)$ when $\delta = \varepsilon$

**(II) Lower bound** Let

$$r_\varepsilon = \frac{(A(\varepsilon)\ r_0 + B(\varepsilon)\ h)}{|A(\varepsilon)\ r_0 + B(\varepsilon)\ h|} .$$

Then, we can get a lower bound for $b(\varepsilon)$ which is

$$\begin{aligned} &[(\frac{A}{G})\ r_0(0) + (\frac{B}{G})\ h(0)] - f_0(0) \\ &= \frac{1}{G^2}\ [((A^2 - 1)\ f_0(0) + 2\,A\ B\ r_0(0)\ h(0) + B^2\ h^2(0))] \\ &+ (\frac{1}{G^2} - 1)\ f_0(0) . \end{aligned} \tag{4.4}$$

Now, since

$$\begin{aligned} &\frac{1}{(A+B)^2} \\ &= \frac{1}{A^2 + B^2 + 2\,A\ B} \\ &\le \frac{1}{A^2 + B^2 + 2\,A\ B \int r_0\ h} \\ &\equiv \frac{1}{G^2} \\ &\le \frac{1}{A^2 + B^2} , \end{aligned}$$

where $G = |A(\varepsilon)\, r_0 + B(\varepsilon)\, h|$; and since

$$(A + B)^2 = 1 - 2\,\varepsilon + 4\,\varepsilon^2 + o(\varepsilon^2)$$
$$A^2 + B^2 = 1 + o(\varepsilon^2)\,.$$

Therefore, it is clear that

$$\frac{1}{G^2} = 1 + o(\varepsilon^q) \qquad \textit{with} \quad q \geq 1/2\,. \tag{4.5}$$

By combining (4.2) - (4.5) in (I) and (II), we have that

$$b(\varepsilon) = \sqrt{2\, f_0(0)}\,(4\, C' - I_0)^{1/4}\, \varepsilon^{1/2} + O(\varepsilon)\,,$$

where $C' = \frac{1}{4}C$, and $I(f) \leq C$; with the approximate extremal function as specified in the theorem.

**Proofs of Theorem 5.1 and Theorem 5.2**

It is straight forward to show that the conditions of Theorem 5.3 are satisfied. Then, by applying Theorem 5.3 the results follow.

**Proofs of Theorem 7.2 and 7.3**

See Liu (1987).

**Proof of Theorem 8.1**

We first formally state the conclusion reached in the text near (8.****)

**Lemma** *If $\{F_\theta\}$ is well parametrized*

$$\limsup_{\varepsilon \to 0} \frac{b(\varepsilon)}{\varepsilon} = \limsup_{\theta \to \theta_0} \frac{|\theta - \theta_0|}{H(F_\theta, F_0)}\,.$$

To prove Theorem, recall that if $\{F_\theta\}$ is differentiable in quadratic mean then

$$\int (\sqrt{f_\theta} - \sqrt{f_{\theta_0}} - (\theta - \theta_0)\, \eta_{\theta_0})^2 = o((\theta - \theta_0)^2)$$

so that by two applications of the triangle inequality

$$\left|\sqrt{\int (\sqrt{f_\theta} - \sqrt{f_{\theta_0}})^2} - \sqrt{\int ((\theta - \theta_0)\,\eta_{\theta_0})^2}\right| = o(|\theta - \theta_0|)\,.$$

Thus

$$H(F_\theta, F_{\theta_0}) = (\theta - \theta_0)\sqrt{(\eta_{\theta_0})^2} + o(|\theta - \theta_0|)\,.$$

Now $\eta_\theta = \frac{\partial}{\partial \theta}\sqrt{f_\theta}$ almost everywhere, so that we have

$$H(F_\theta, F_{\theta_0}) = (\theta - \theta_0)\,\frac{i}{2}\,\sqrt{I} + o(|\theta - \theta_0|)$$

and so

$$\limsup_{\theta \to \theta_0} \frac{|\theta - \theta_0|}{H(F_\theta, F_0)} = \frac{2}{\sqrt{I_{Fisher}}}\,.$$

Because the family is well parametrized, the lemma and the definition of $I_{Geometric}$ together imply that the LHS is $\frac{2}{\sqrt{I_{Geometric}}}$.

**Proof of Theorem 8.2**

By hypothesis, $I_{Fisher}$ can be defined, so

$$\eta_{\theta_0} = \frac{f_\theta^{1/2} - f_{\theta_0}^{1/2}}{\theta - \theta_0} \to_{a.e.} \frac{\partial}{\partial \theta}\sqrt{f_\theta}\,|_{\theta = \theta_0}$$

By Fatou's lemma

$$\liminf_{\theta \to \theta_0} \int \eta_\theta^2 \geq \int (\frac{\partial}{\partial \theta}\sqrt{f_\theta})^2 = \frac{1}{4}\,I_{Fisher}\,.$$

But $H^2(F_\theta, F_{\theta_0})/(\theta - \theta_0)^2 = \int g_\theta^2$, so we conclude

$$\limsup_{\theta \to \theta_0} \frac{(\theta - \theta_0)^2}{H^2(F_\theta, F_{\theta_0})} \leq \frac{4}{I_{Fisher}}\,.$$

As the family is well-parametrized, the lemma above implies that the LHS of this display is $\frac{4}{I_{Geometric}}$.

**Proof of Theorem 8.3**

By the definition of $I_{Geometric}$ terms of a limit superior, there exrots a sequence $F_k$ in $\mathbf{F}$, converging to $F_0$, with the property

$$\frac{T(F_k) - T(F_0)}{H(F_k, F_0)} \to \limsup_{\varepsilon \to 0} \frac{b(\varepsilon)}{\varepsilon} = \frac{2}{\sqrt{I_{Geometric}}}$$

Put $\theta_k = T(F_k)$, $\theta_0 = T(F_0)$, $\varepsilon_k = H(F_k, F_0)$, for short.

Using Pitman's inequality,

$$(E_{F_k} T_n - E_{F_0} T_n)^2 \frac{1 - H_{n,k}^2}{H_{n,k}^2} \leq \frac{1}{2}(Var_{F_k} T_n - Var_{F_0} T_n) \tag{11.A}$$

where we put $H_{n,k}^2 = H^2(P_{0,n}, P_{1,n})$ for short.

Now the left hand side of this inequality is larger than

$$(1 - \frac{2\beta_n}{(\theta_k - \theta_0)})^2 \frac{(\theta_k - \theta_0)^2}{4\varepsilon_k^2} \frac{1 - H_{n,k}^2}{H_{n,k}^2/\varepsilon_k^2} . \tag{11.B}$$

Now $\beta_n = o(n^{-1/2})$ by hypothesis; thus $\delta_n = n^{1/2}\beta_n \to 0$. Pick a sequence $g_n \to 0$ such that $\delta_n/g_n \to 0$ (e.g. $g_n = \sqrt{\delta_n}$ will do). Extract subsequences $\{k_m\}$, $\{n_m\}$ so that

$$100\, n_m^{-1/2} g_{n_m} \geq \varepsilon_{k_m} \geq n_m^{-1/2} g_{n_m} ,$$

say. Then

$$\varepsilon_{k_m}^{-1} \beta_{n_m} \to 0$$
$$\varepsilon_{k_m}^2 n_m \to 0 .$$

Now by the product formula $H_{n,k}^2 = 2 - 2(1 - \frac{\varepsilon_k^2}{2})^n$, and so $H_{n,k}^2 = n\varepsilon_k^2 + o(n\varepsilon_k^2)$. By the properties of the sequences $\{k_m\}$, $\{n_m\}$,

$$\frac{1 - H_{n_m,k_m}^2}{H_{n_m,k_m}^2/\varepsilon_{k_m}^2} = (n_m + o(n_m))^{-1}$$

as $m \to \infty$.

Now

$$(1 - \frac{2\beta_n}{(\theta_k - \theta_0)})^2 = (1 - (\frac{\varepsilon_k^{-1}\, 4}{\sqrt{I_{Geometric}}} + o(\varepsilon_k^{-1}))\, \beta_n)^2$$

By the properties of $\{k_m\}$, $\{n_m\}$,

$$\varepsilon_k \beta_n \to 0$$

and so

$$(1 - \frac{2\beta_{n_m}}{(\theta_{k_m} - \theta_0)})^2 \to 1 .$$

Finally, by construction of $\theta_k$ and $\varepsilon_k$,

$$\frac{(\theta_k - \theta_0)^2}{4\,\varepsilon_k^2} \to \frac{1}{I_{Geometric}} .$$

Putting these pieces together in equations (11.A), (11.B)

$$(1 + o(1))^2 \frac{1}{I_{Geometric}} \frac{1}{n_m + o(n_m)} \le \frac{1}{2}(Var_{F_{k_m}} T_{n_m} + Var_{F_0} T_{n_m}) .$$

Now $Var_{F_{k_m}} T_{n_m} \ge Var_{F_0} T_{n_m} - \nu_{n_m}/n$, and so

$$\frac{1}{2}(Var_{F_{k_m}} T_{n_m} + Var_{F_0} T_{n_m}) \le Var_{F_0} T_{n_m} - \nu_{n_m}/2n .$$

Of course, $MSE_{F_0} T_{n_m} \ge \frac{1 + o(1)}{I_{Geometric}} + \nu_{n_m}/2$. $\nu_{n_m} \to 0$ and so we have established the Theorem.

**Proof of Theorem 8.4**

Fix $\delta > 0$. Let the family $\{F_\theta\}$ be a family used in the definition of Levit information and nearly attaining the infimum there, so that $T(F_t) = T(F_0) + t$, $\{F_\theta\}$ DQM, and $I_{Fisher}(\theta = 0, \{F_\theta\}) \le I_{Levit} + \delta$. As $\{F_\theta\}$ is DQM, we have by theorem 8.1

$$\limsup_{\varepsilon \to 0} \frac{b(\varepsilon, T, \{F_\theta\})}{\varepsilon} = \frac{2}{\sqrt{I_{Fisher}}} .$$

Now for any family $\{F_\theta\} \subset \mathbf{F}$, $b(\varepsilon, T, \mathbf{F}) \ge b(\varepsilon, T, \{F_\theta\})$. Therefore

$$\limsup_{\varepsilon \to 0} \frac{b(\varepsilon, T, \mathbf{F})}{\varepsilon} \ge \limsup_{\varepsilon \to 0} \frac{b(\varepsilon, T, \{F_\theta\})}{\varepsilon} .$$

But the LHS is just $2/\sqrt{I_{Geometric}}$ by definition. Combining this fact with the last two display shows that $I_{Geometric} \le I_{Levit} + \delta$). As $\delta > 0$ was arbitrary, this completes the proof.

**Proof of Theorem 8.5**

Under the assumption of the theorem, there exists a parameter family $\{F_\theta\}$ satisfies $b(\varepsilon,T,F) = b(\varepsilon,T,\{F_\theta\})$. It follows that

$$\limsup_{\varepsilon \to 0} \frac{b(\varepsilon,T,F)}{\varepsilon} = \limsup_{\varepsilon \to 0} \frac{b(\varepsilon,T,\{F_\theta\})}{\varepsilon} .$$

The left hand side is just $2/\sqrt{I_{Geometric}}$. The right hand side, because $\{F_\theta\}$ is assumed DQM, is just $2/\sqrt{I_{Fisher}}$. Hence $I_{Geometric} = I_{Fisher}$. But as $I_{Levit} \le I_{Fisher}$ $((\theta = 0,\{F_\theta\})$ we have $I_{Levit} \le I_{Geometric}$. In this case, since by Theorem 8.3 we know $I_{Levit} \ge I_{Geometric}$ always, $I_{Levit} = I_{Geometric}$.

**Proof of Theorem 9.1**

For ease of exposition, assume that there is, for each $\varepsilon$ we require, a distribution $F_\varepsilon$ with $H(F_0,F_\varepsilon) = \varepsilon$ and $b(\varepsilon) = T(F_\varepsilon) - T(F_0)$. Let the statistic $T_n$ be given; we need its fractional bias $\gamma$, defined by

$$Ave\ \{|E_F(T_n) - T(F)|\}/b(\varepsilon) = \gamma .$$

where here and below averages are over the two-point set $F \in \{F_0,F_\varepsilon\}$. It follows from the arithmetic-geometric mean inequality that if $(a + b)/2 = c$ then $\frac{1}{2}(a^2 + b^2) \ge c^2$. This implies

$$Ave\ \{(E_F(T_n) - T(F))^2\} \ge \gamma^2 b^2(\varepsilon) , \tag{9.1}$$

$$(E_{F_0}(T_n) - E_{F_\varepsilon}(T_n))^2 \ge b^2(\varepsilon)\,(1 - 2\gamma)^2 . \tag{9.2}$$

By Pitman's inequality

$$Ave\ \{Var_F T_n\} \ge \frac{1}{4} \frac{1 - H_n^2}{H_n^2} (E_{F_0}(T_n) - E_{F_\varepsilon}(T_n))^2 . \tag{9.3}$$

Now clearly

$$Ave\ MSE \ge Ave\ Bias^2 + Ave\ Variance ,$$

so using (9.1) and (9.2)

$$Ave\ MSE \ge \gamma^2\, 2b^2(\varepsilon) + \frac{1}{4} \frac{1 - H_n^2}{H_n^2}\, b^2(\varepsilon)\,(1 - 2\gamma)^2 \tag{9.4}$$

Put $e = \frac{1}{4}\,\frac{1-H_n^2}{H_n^2}$; the right-hand side of (9.4) is a function of $\gamma, \varepsilon, e$, and $b(\varepsilon)$; call it AMSE. Then

$$AMSE = b^2(\varepsilon)\,\{\gamma^2 + e\,(1-2\gamma)^2\}\,. \tag{9.5}$$

If $\gamma \le \frac{1}{2}$, then by factoring the $\gamma$-polynomial $\gamma^2 + e\,(1-2\gamma)^2\}$ we have

$$AMSE = b^2(\varepsilon)\,\{(1+4e)\,((\gamma - \frac{2e}{1+4e})^2 + \frac{e}{(1+4e)^2})\} \tag{9.6}$$

which implies that

$$\min_{\gamma < 1/2} AMSE = \frac{e}{1+4e}\,b^2(\varepsilon)\,; \tag{9.7}$$

the minimum being attained at $\gamma = \frac{2e}{1+4e}$. On the other hand, for $\gamma \ge 1/2$, we have the bound $AMSE \ge b^2(\varepsilon)/4$. Combining (9.4) and (9.7) we have

$$\begin{aligned} Ave\ MSE &\ge \frac{e}{1+4e}\,b^2(\varepsilon) \\ &= \frac{1}{4}\,(1-H_n^2)\,b^2(\varepsilon) \end{aligned}$$

By the product rule for Hellinger distance $H_n$,

$$H_n^2 \equiv H^2(P_{0,n}, P_{\varepsilon,n}) = 2 - 2\,(1-\varepsilon^2/2)^n\,,$$

and we have

$$Ave\ MSE \ge \frac{1}{4}\,(-1 + 2\,(1-\varepsilon^2/2)^n)\,b^2(\varepsilon)\,.$$

Now let $a = .62$, and let $\varepsilon = \sqrt{a/n}$. Suppose $b(\varepsilon) = A\,\varepsilon^q + o(\varepsilon^q)$. Then

$$b^2(\sqrt{a/n}) = a^q\,b^2(n^{-1/2}) + o(n^{-q})\,.$$

Define

$$\xi_n^2(q) = \{-1 + 2\,(1-\varepsilon^2/2)^n\}\,a^q$$

and note that $\xi_n^2(q) \to \xi(q) = \{2\,e^{-a/2} - 1\}\,a^q$. Now the lemma below shows that with $a = .62$, $\xi^2(q) \ge \xi^2(1) = .28$. So for $n > n_0(q,\delta)$, $\xi_n(q) > .279\,\delta$, and the $o(n^{-q})$ term in (11.*****) is smaller than $(\frac{.029}{4})\,b^2(\sqrt{a/n})$; consequently

$$Ave\ MSE \geq \frac{1}{16}\, b^2(\sqrt{a/n}) .$$

Now for fixed $\delta$, and all $n > n_1(\delta)$, $\varepsilon = \sqrt{a/n} < \delta$, so that $F_\varepsilon \in N_\delta(F_0)$, and so for $n > \max(n_0, n_1)$

$$\sup_{N_\delta(F_0)} MSE \geq Ave\ MSE \geq \frac{1}{16}\, b^2(\sqrt{a/n})$$

which completes the proof.

## References

Ahmad, I. A. (1976) "On asymptotic properties of an estimate of a functional of a probability density." *Scand. Actuarial J.*, 1, 176-181.

Begun, J. M., Hall, W. J., Huang, W. M. & Wellner, J. A. "Information and asymptotic efficiency in parametric - nonparametric models." *Ann. Stat.*, 11, 432-452.

Beran, R. J. (1977a) "Minimum distance estimation for parametric models." *Ann. Stat.*, 5, 445-463.

Beran, R. J. (1977b) "Robust location estimates." *Ann. Stat.*, 5, 431-444.

Bertero, M., Bocacci, P., & Pike, E. R. (1982) "On the recovery and resolution of exponential relexation rates from exponential data : a singular-value analysis of the Laplace transform inversion in the presence of noise. I." *Proc. Roy. Soc., London*, 383, 15-29.

Bickel, P. J. (1982) "On adaptive estimation." *Ann. Stat.*, 10, 647-671.

Birgé, L. (1983) "Approximation dans les espaces métriques et théorie de l'estimation." *Zeit.Wahr.*, 65, 181-237.

Boyd, D. W. & Steele, J. M. (1978) "Lower bounds for nonparametric density estimation rates." *Ann. Stat.*, 6, 932-934.

Bretagnolle, J. & Huber, C. (1979) "Estimation des densités: risque minimax." *Zeit. Wahr.*, 47, 119-137.

Centsov, N. (1962) "Evaluation of an unknown distribution density from observations." *Soviet Math.*, 3,1, 15-30.

Devroye L. & Györfi L. (1985) *Nonparametric Density Estimation : The $L^1$ View*. J. Wiley.

Donoho, D. L. (1985) "Nonparametric inference about functionals of a density." (to appear *Ann. Stat.*).

Donoho, D. L. & Liu, R. C. (1985) "Automatic robustness of minimum distance functionals." (to appear *Ann. Stat.*).

Eddy, W. F. (1980) "Optimum kernel estimate of the mode." *Ann. Stat.*, 8,4, 870-882.

Farrell, R. H. (1967) "On the lack of uniformity consistent sequence of estimators of a density function in certain cases." *Ann. Math. Stat.*, 38, 471-474.

Farrell, R. H. (1972) "On the best obtainable asymptotic rates of convergence in estimation of a density function at a point." *Ann. Math. Stat.*, 43, #1, 170-180.

Farrell, R. H. (1980) "On the efficiency of density function estimators." (unpublished manuscript)

Hall, P. & Marron, S. (1987) "On the amount of noise inherent in bandwidth selection." *Ann. Stat.*, (to appear).

Hall, P. & Welsh, A. H. (1984) "Best attainable rates of convergence for parameters of regular variation." *Ann. Stat.*, 12, 3, 1079-1084.

Hasminskii, R. Z. (1978) "A lower bound for risks of nonparametrical estimates of density in the uniform metrics." *Teor. Veroyathost i Primenen*, 824-828.

Hasminskii, R. Z. (1979) "Lower bound for the risks of nonparametric estimates of the mode." in *Contribution to Statistics (J. Hajek memorial volume)*, Academia, Prague 1979, 91-97.

Hampel, F. R. (1968) *Contributions to the Theory of Robust Estimation.* Thesis, University of Calif., Berkeley.

Huber, P. J. (1964) "Robust estimation of a location parameter." *Ann. Math. Stat.*

Ibragimov, I. A. and Hasminskii, R. Z. (1978) "On the nonparametric estimation of functionals." *Symposium in Asymptotic Statistics*, Prague, 41-52.

Ibragimov, I. A. and Hasminskii, R. Z. (1981) *Statistical Estimation: Asymptotic Theory.* Springer Verlag, New York. 237-240.

Jewell N. (1982) "Mixtures of exponential distributions." *Ann. Stat.* 10, 479-484.

Kiefer, J. (1982) "Optimum rates for nonparametric density and regression estimates, under order restrictions." *Stat. and Prob. : Essays in Honor of C.R. Rao*, North Holland Publishing Co., 419-428.

Khoshevnik, Y. A. & Levit, B. Y. (1976) "On a non-parametric analog of the information matrix." *Theory of Prob. and Appl.*, 21, 738-753.

LeCam, L. (1970) "On the assumptions used to prove asymptotic normality of estimates." *Ann. Math. Stat.*, 41, 802-828.

LeCam, L. (1973) "Convergence of estimates under dimensionality restrictions." *Ann. Stat.*, 1, 38-53.

LeCam, L. (1975) "On local and global properties in the theory of asymptotic normality of experiments." in *Stochastic Processes and Related Topics*, Vol. 1.

LeCam, L. (1986) *Asymptotic Methods in Statistical Decision Theory*. Springer Verlag.

Levit, B. Y. (1975) "On the efficiency of a class of nonparametric estimates." *Theory of Prob. and Appl.*, 20, 723-740.

Liu, R. C. (1987) *Geometry in Robustness and Nonparametrics*. Thesis, University of Calif., Berkeley.

Meyer, T. G. (1977a) "On fixed or scaled radii confidence sets: the fixed sample size case." *Ann. Stat.*, 5,1, 65-78.

Meyer, T. G. (1977b) "Bounds for estimation of density functions and their derivatives." *Ann. Stat.*, 5, 136-142.

Millar, P. W. (1981) "Robust estimation via minimum distance methods." *Zeit. Wahr.*

Pitman E. J. G. (1979) *Some Basic Theory for Statistical Inference*. Halsted press.

Ritov, Y. (1986) (Manuscript).

Samarov, A. M. (1976) "Minimax bound on the risk of nonparametric density estimates." *Problems of Information Transmision*, 12, 242-244.

Samarov, A. M. (1977) "Lower bounds for integral risk of density estimates." *Prob. of Const. of Systems for Information Transmision*, (E. Bloch, Ed), (Ph.D. Thesis).

Stone, C. J. (1980) "Optimal rates of convergence for nonparametric estimators." *Ann. Stat.*, 8,6, 1348-1360.

Stone, C. J. (1983) "Optimal uniform rate of convergence for nonparametric estimators of a density function or its derivatives." *Ann. Stat.*

Venter, J. H. (1967) "On estimation of the mode." *Ann. Math. Stat.*, **38,5,** 1446-1455.

Wahba, G. (1975) "Optimal convergence properties of variable knot, kernel, and orthogonal series: methods for density estimation." *Ann. Stat.*, 3,1, 15-29.