29

Assouad, Fano, and Le Cam

Bin Yu[1]

ABSTRACT This note explores the connections and differences between three commonly used methods for constructing minimax lower bounds in nonparametric estimation problems: Le Cam's, Assouad's and Fano's. Two connections are established between Le Cam's and Assouad's and between Assouad's and Fano's. The three methods are then compared in the context of two estimation problems for a smooth class of densities on [0,1]. The two estimation problems are for the integrated squared first derivatives and for the density function itself.

29.1 Introduction

In nonparametric estimation problems, minimax is a commonly used risk criterion. An optimal minimax rate is often obtained by first deriving a minimax lower bound, often nonasymptotic, on the risk and then constructing an explicit estimator which achieves the rate in the lower bound. See for example, Has'minskii & Ibragimov (1978), Bretagnolle & Huber (1979), Stone (1984), Birgé (1986), Bickel & Ritov (1988), Donoho & Nussbaum (1990), Fan (1991), Donoho & Liu (1991), Birgé & Massart (1992), and Pollard (1993). Depending on the class considered and the function or functional estimated, techniques used to derive lower bounds differ. Three general methods have been widely employed: one formalizes some arguments of Le Cam (1973), and the other two are based on inequalities of Assouad (1983) and Fano (compare with Cover & Thomas 1991, p. 39). The first method will be referred to in this note as Le Cam's method. (Note that Le Cam (1986, Chapter 16) has developed a much more general theory.) It deals with two sets of hypotheses, while the Assouad and Fano methods deal with multiple hypotheses, indexed by the vertices of a hypercube and those of a simplex, respectively.

In this note, we explore the connections and differences between these three lemmas with the hope of shedding light on other more general problems. Section 2 contains two results (Lemma 2 and Lemma 5), which relate the three methods. It is known that Assouad's lemma (Lemma 2) gives

[1]University of California at Berkeley.

very effective lower bounds for many global estimation problems. One way to understand that lemma is through Le Cam's method: the global estimation problem can be decomposed into several sub-estimation problems, and Assouad's Lemma is obtained by applying Le Cam's method to the sub-problems. In Lemma 5 we use a simple packing number result to extract a subset of the vertices of a hypercube to which the Fano method is applied, thereby obtaining a lower bound similar to that of Assouad. Hence in this sense, Fano's method is stronger, as observed by Birgé (1986).

From the examples worked out in the literature, it appears that Le Cam's method often gives the optimal rate when a real functional is estimated, but it can be non-straightforward to find the appropriate two sets of hypotheses in some problems. On the other hand, the other two lemmas seem to be effective when the whole unknown function is being estimated, although Assouad's Lemma seems easier to use and therefore more popular than Fano's. In Section 3, we demonstrate this point in the context of a particular smooth class of densities on [0,1] and with two estimation problems, one for a real functional and one for the whole density. For the functional, Le Cam's method gives the optimal rate of convergence, while Assouad's and Fano's provide the optimal rate for the whole density.

After the completion of this note, closely related work by C. Huber was brought to my attention. In her article, which appears in this volume, she explores the connection between Assouad's and Fano's methods.

29.2 The three methods

Assume that $\mathcal{P}$ is a family of probability measures and $\theta(P)$ is the parameter of interest with values in a pseudo-metric space $(\mathcal{D}, d)$. (It would be inconvenient to require that $d(\theta, \theta') = 0$ implies that $\theta = \theta'$.) Let $\hat{\theta} = \hat{\theta}(X)$ be an estimator of $\theta(P)$ based on an X with distribution P, and denote by $co(\mathcal{P})$ the convex hull of $\mathcal{P}$.

Le Cam (1973) relates the testing problem of two sets of hypotheses to the L^1 distance of the convex hulls of the two hypothesis sets. Roughly speaking, if one wants to test between these two sets well, then their convex hulls have to be well separated. Since estimators also define tests between subsets of $\mathcal{D}$, Le Cam's testing bound also provides a lower bound for the accuracy of an estimator.

Lemma 1 (Le Cam's method) *Let $\hat{\theta}$ be an estimator of $\theta(P)$ on $\mathcal{P}$ taking values in a metric space $(\mathcal{D}, d)$. Suppose that there are subsets D_1 and D_2 of $\mathcal{D}$ that are 2δ-separated, in the sense that, $d(s_1, s_2) \geq 2\delta$ for all $s_1 \in \mathcal{D}_1$ and $s_2 \in \mathcal{D}_2$,. Suppose also that $\mathcal{P}_1$ and $\mathcal{P}_2$ are subsets of $\mathcal{P}$ for which $\theta(P) \in \mathcal{D}_1$ for $P \in \mathcal{P}_1$ and $\theta(P) \in \mathcal{D}_2$ for $P \in \mathcal{P}_2$. Then*

$$\sup_{P \in \mathcal{P}} E_P d(\hat{\theta}, \theta(P)) \geq \delta \cdot \sup_{P_i \in co(\mathcal{P}_i)} \|P_1 \wedge P_2\|,$$

where the affinity $\|P_1 \wedge P_2\|$ is defined through

$$\|P_1 - P_2\|_1 = 2(1 - \|P_1 \wedge P_2\|).$$

Proof: For $P_i \in \mathcal{P}_i$,

$$\begin{aligned} M &:= 2 \sup_{P \in \mathcal{P}} E_P d(\hat{\theta}, \theta(P)) \\ &\geq E_{P_1} d(\hat{\theta}, \theta(P_1)) + E_{P_2} d(\hat{\theta}, \theta(P_2)) \\ &\geq E_{P_1} d(\hat{\theta}, \mathcal{D}_1) + E_{P_2} d(\hat{\theta}, \mathcal{D}_2), \end{aligned}$$

which implies

$$E_{P_1} d(\hat{\theta}, \mathcal{D}_1) + E_{P_2} d(\hat{\theta}, \mathcal{D}_2) \leq M \quad \text{for all } P_i \in co(\mathcal{P}_i).$$

Since $d(\hat{\theta}, \mathcal{D}_1) + d(\hat{\theta}, \mathcal{D}_2) \geq d(\mathcal{D}_1, \mathcal{D}_2) \geq 2\delta$, then for any $P_i \in co(\mathcal{P}_i)$,

$$M \geq 2\delta \inf_{f_i \geq 0; f_1 + f_2 = 1} (E_{P_1} f_1 + E_{P_2} f_2) = 2\delta \, \|P_1 \wedge P_2\|.$$

Hence

$$M := 2 \sup_{P \in \mathcal{P}} E_P d(\hat{\theta}, \theta(P)) \geq 2\delta \cdot \sup_{P_i \in co(\mathcal{P}_i)} \|P_1 \wedge P_2\|.$$

□

Remark

(i) If d is not a pseudo-metric, but a non-negative symmetric function wsatisfying the following "weak" triangle inequality, that is, for some constant $A \in (0, 1)$,

$$d(x, z) + d(z, y) \geq A d(x, y),$$

then the lower bound holds with an extra factor A. This observation is very useful in Example 2 in Section 1.3

(ii) In many cases, better lower bounds are obtained by considering the convex hulls of the $\mathcal{P}_i$, because the supremum of $\|P_1 \wedge P_2\|$ over the convex hulls can be much larger than the supremum over the $\mathcal{P}_i$ themselves.

Assouad's lemma gives a minimax lower bound over a class of 2^m hypotheses (probability measures) indexed by vertices of a m-dimensional hypercube. The following form of Assouad's lemma is reworded from Devroye (1987, p. 60) (compare with Le Cam 1986, p. 524) to emphasize the decomposability of the (pseudo) distance d into a sum of m (pseudo) distances, which correspond to m estimation subproblems. The proof is rewritten to make the point that each subproblem is like testing the hypotheses indexed by neighboring vertices on the hypercube along the direction determined by the particular subproblem and the argument used in Le Cam's method (Lemma 1) can be applied to each of the subproblems.

Lemma 2 (Assouad's Lemma) *Let $m \geq 1$ be an integer and let $\mathcal{F}_m = \{P_\tau : \tau \in \{-1,1\}^m\}$ contain 2^m probability measures. Write $\tau \sim \tau'$ if τ and τ' differ in only one coordinate, and write $\tau \sim_j \tau'$ when that coordinate is the jth. Suppose that there are m pseudo-distances on $\mathcal{D}$ such that for any $x, y \in \mathcal{D}$*

$$d(x,y) = \sum_{j=1}^{m} d_j(x,y), \tag{1}$$

and further that, if $\tau \sim_j \tau'$,

$$d_j(\theta(P_\tau), \theta(P_{\tau'})) \geq \alpha_m. \tag{2}$$

Then

$$\max_{P_\tau \in \mathcal{F}_m} E_\tau d(\hat{\theta}, \theta(P_\tau)) \geq m \cdot \frac{\alpha_m}{2} \min\{\|P_\tau \wedge P_{\tau'}\| : \tau \sim \tau'\}.$$

Proof: For any given $\tau = (\tau_1, \ldots, \tau_m)$, let τ^j denote the m-tuple that differs from it in only the jth position. Then $d(\theta(P_\tau), \theta(P_{\tau^j})) \geq \alpha_m$.

$$\begin{aligned}
& \max_\tau E_\tau d(\theta(P_\tau), \hat{\theta}) \\
= \; & \max_\tau \sum_{j=1}^{m} E_\tau d_j(\theta(P_\tau), \hat{\theta}) \\
\geq \; & 2^{-m} \sum_\tau \sum_{j=1}^{m} E_\tau d_j(\theta(P_\tau), \hat{\theta}) \\
= \; & \sum_{j=1}^{m} 2^{-m} \sum_\tau E_\tau d_j(\theta(P_\tau), \hat{\theta}) \\
= \; & \sum_{j=1}^{m} 2^{-(m+1)} \sum_\tau \left(E_\tau d_j(\theta(P_\tau), \hat{\theta}) + E_{\tau^j} d_j(\theta(P_{\tau^j}), \hat{\theta}) \right)
\end{aligned}$$

For each fixed τ and j, we have a pair of hypotheses P_τ and P_{τ^j} sitting on the neighboring vertices of the hypercube along direction j. Therefore, as in Le Cam's method (Lemma 1), the average estimation error over these two hypotheses can be bounded from below by $\frac{1}{2}\alpha_m \|P_\tau \wedge P_{\tau^j}\|$. Thus,

$$\begin{aligned}
\max_\tau E_\tau d(\theta(P_\tau), \hat{\theta}) & \geq \sum_{j=1}^{m} 2^{-m} \sum_\tau \alpha_m \|P_\tau \wedge P_{\tau^j}\| \\
& \geq m \frac{\alpha_m}{2} \min\{\|P_\tau \wedge P_{\tau'}\| : \tau \sim \tau'\}
\end{aligned}$$

□

The relation $\tau \sim \tau'$ can also be written $W(\tau, \tau') = 1$, where W denotes the *Hamming distance*,

$$W(\tau, \tau') = \frac{1}{2} \sum_{j=1}^{m} |\tau_j - \tau'_j|,$$

the number of places where τ and τ' differ.

From Remark (i) after Lemma 1, Assouad's lower bound holds with an extra factor A if d_j are non-negative symmetric functions satisfying the weak triangle inequality with the same constant A.

Devroye (1987, p. 77) (compare with Le Cam 1986, p. 524) contains a generalized Fano's lemma in the case that $\theta(P)$ is the density of P and d is the L^1 norm. We present here a slightly stronger version whose proof is based on ideas from Han and Verdú (1994). We find their proof less involved than those in the statistics literature. It is based on information theory concepts and Fano's original inequality (compare with Cover & Thomas 1991, p. 39).

Lemma 3 (Generalized Fano method) *Let $r \geq 2$ be an integer and let $\mathcal{M}_r \subset \mathcal{P}$ contain r probability measures indexed by $j = 1, 2, \ldots, r$ such that for all $j \neq j'$*

$$d(\theta(P_j), \theta(P_{j'})) \geq \alpha_r,$$

and

$$K(P_j, P_{j'}) = \int \log(P_j / P_{j'}) dP_j \leq \beta_r.$$

Then

$$\max_j E_j d(\hat{\theta}, \theta(P_j)) \geq \frac{\alpha_r}{2} \left(1 - \frac{\beta_r + \log 2}{\log r}\right).$$

Proof: Write θ_j for $\theta(P_j)$. Let Y be a random variable uniformly distributed on the hypothesis set $\{1, 2, \ldots, r\}$, and X be a random variable with the conditional distribution P_j given $Y = j$. Define Z as the value of j for which $d(\hat{\theta}(X), \theta_j)$ is a minimum. (It does not matter how we handle ties.) Because $d(\theta_j, \theta_{j'}) \geq \alpha_r$ for $j \neq j'$, we certainly have $Z = j$ when $d(\hat{\theta}(X), \theta_j) < \alpha_r/2$. It follows that

$$\begin{aligned}
\max_j E_j d(\hat{\theta}, \theta(P_j)) & \geq \frac{\alpha_r}{2} \max_j P(d(\hat{\theta}(X), \theta_j)) \geq \frac{\alpha_r}{2} \mid Y = j) \\
& \geq \frac{\alpha_r}{2r} \sum_{j=1}^{r} P(Z \neq j \mid Y = j) \\
& = \frac{\alpha_r}{2} P(Z \neq Y).
\end{aligned}$$

Let h be the entropy function with the natural log,

$$h(p) = -p\log p - (1-p)\log(1-p) \quad \text{for } p \in (0,1).$$

Then $0 \le h(\cdot) \le \log 2$. Denote by $I(Y;Z) = K(P_{(Y,Z)}, P_Y \times P_Z)$ the mutual information between Y and Z, and by $H(Y|Z)$ the equivocation or the average posterior entropy of Z given Y. Then

$$I(Y;Z) = H(Y) - H(Y|Z) = \log r - H(Y|Z).$$

Furthermore, by a property of mutual information and the convexity of the Kullback-Leibler divergence (Cover & Thomas 1991, pp. 30, 33), we have

$$\begin{aligned} I(Y;Z) = I(Y;\hat{\theta}(X)) \le I(Y;X) &= \frac{1}{r}\sum_{i=1}^{r} K\Big(P_i, \frac{1}{r}\sum_{j=1}^{r} P_j\Big) \\ &\le \frac{1}{r^2}\sum_{i,j} K(P_i, P_j). \end{aligned}$$

It follows from Fano's inequality (Cover & Thomas 1991, p. 39),

$$H(Y|Z) \le P(Z \ne Y)\log(r-1) + h(P(Z=Y)),$$

that

$$\begin{aligned} P(Z \ne Y)\log(r-1) &\ge H(Y|Z) - h(1/2) \\ &= H(Y) - I(Y;Z) - \log 2 \\ &\ge \log r - \frac{1}{r^2}\sum_{i,j} K(P_i, P_j) - \log 2. \end{aligned}$$

Increase the $\log(r-1)$ to $\log r$ and replace $K(P_i, P_j)$ by its upper bound β_r, then substitute the resulting lower bound for $P(Z \ne Y)$ into the minimax inequality to get the asserted bound. □

As remarked by Birgé (1986, p. 279), "[Fano's Lemma] is in a sense more general because it applies in more general situations. It could also replace Assouad's Lemma in almost any practical case ...". Indeed, Lemma 3 implies a result similar to Assouad's Lemma. The idea is to select the maximal subset of vertices from the m-dimensional hypercube which are $m/3$ apart in Hamming distance and apply Fano's Lemma to the selected set of vertices.

Lemma 4 *For a universal positive constant c_0, and each $m \ge 6$, there exists a subset A of $\{-1,+1\}^m$ consisting of at least $\exp(c_0 m)$ vertices, each pair greater than $m/3$ apart in Hamming distance.*

Proof: Let k be the integer part of $m/6$. Let A be a maximal set of vertices, each pair at least $2k+1$ apart in Hamming distance. The Hamming ball $B(\tau, 2k)$ of radius $2k$ and center τ contains $N(m,2k) = \sum_{r=0}^{2k} \frac{m!}{r!(m-r)!}$ vertices, corresponding to the subsets of $2k$ or fewer coordinates at which a vertex in the ball can differ from τ.

Because A is maximal, no vertex in the cube $\{-1,+1\}^m$ can lie further than $2k$ from A; the whole cube is covered by a union of balls $B(\tau, 2k)$, with τ ranging over A. This union contains at most $|A|N(m,2k)$ vertices, which is therefore an upper bound for 2^m.

It remains to calculate an upper bound for $N(m,2k)$ by means of the usual generating function argument. Let Z denote a random variable with a $\mathrm{Bin}(m,1/2)$ distribution. Put $s = (m-2k)/2k$, which is greater than 1. Then

$$N(m,2k) = 2^m P(Z \le 2k) \le 2^m E s^{2k-Z} = 2^m s^{2k}\left(\frac{1}{2} + \frac{1}{2s}\right)^m.$$

The bound simplifies to $\exp(mh(2k/m))$, which leads to the asserted lower bound for A if we take $k = \lfloor m/6 \rfloor$. □

Lemma 5 *Let $m \ge 1$ be an integer and let $\mathcal{F}_m = \{P_\tau : \tau \in \{-1,1\}^m\}$ contain 2^m probability measures, and let W be the Hamming distance. Suppose that there are constants α_m and γ_m such that*

$$d(\theta(P_\tau), \theta(P_{\tau'})) \ge \alpha_m W(\tau, \tau') \tag{3}$$

$$K(P_\tau, P_{\tau'}) \le m\gamma_m. \tag{4}$$

Then

$$\max_{P_\tau \in \mathcal{F}_m} E_\tau d(\hat{\theta}_n, \theta(P_\tau)) \ge m \cdot \frac{\alpha_m}{6}\left(1 - \frac{1}{c_0}(\gamma_m + \log 2/m)\right).$$

Proof: Apply Lemma 3 to the set A of $r = \exp(c_0 m)$ vertices given by Lemma 4. From (3) and the $m/3$ separation of vertices in A, we have $d(\theta(P_\tau), \theta(P_{\tau'})) \ge \alpha_m W(\tau,\tau') \ge m\alpha_m/3$ for distinct vertices in A. □

Let us now compare the conditions in Assouad's lemma and those in Lemma 5. The first two conditions (1) and (2) in Assouad's Lemma do not quite imply condition (3) in Lemma 5, but condition (2) together with the following stronger condition

$$\min_j \{d_j(\theta(P_\tau), \theta(P_{\tau'}) : \tau_j \ne \tau'_j\} \ge \alpha_m.$$

would imply condition (3). Note that this new condition is satisfied by hypercube classes constructed through perturbations of a fixed density over a partition, as in the next section. Moreover, note that condition (4) implies a lower bound on the affinity $\|P_\tau \wedge P_{\tau'}\|$ through the Kullback-Csiszár-Kemperman inequality (Devroye 1987, p. 10):

$$\|P_\tau - P_{\tau'}\|_1 \le \sqrt{2K(P_\tau, P_{\tau'})}.$$

Since $\|P_\tau - P_{\tau'}\|_1 = 2(1 - \|P_\tau \wedge P_{\tau'}\|)$, $K(P_\tau, P_{\tau'}) \le \gamma_m$ implies

$$\|P_\tau \wedge P_{\tau'}\| \ge 1 - \sqrt{\gamma_m/2}.$$

This seems to suggest that condition (4) in Lemma 5 is stronger than the affinity condition in Assouad's lemma. However, as is the case in many, if not all, hypercube classes one actually constructs, the probability measures in the hypercube class often have densities bounded from below by $c > 0$. In this case, a lower bound β_m on the affinity implies an upper bound on the Kullback divergence:

$$\begin{aligned} K(P_\tau, P_{\tau'}) &\le c^{-1}\|P_\tau - P_{\tau'}\|_1 \\ &= 2c^{-1}(1 - \|P_\tau \wedge P_{\tau'}\|) \\ &\le 2c^{-1}(1 - \beta_m) \end{aligned}$$

provided that $\|P_\tau \wedge P_{\tau'}\| \ge \beta_m$.

So far we have connected Le Cam's method with Assouad's in Lemma 2 and Fano's with Assouad's in Lemma 5. Comparing Le Cam's with Fano's would complete the circle. In a way, they are similar in that they both deal with hypothesis testing: Le Cam's for two sets of hypotheses, and Fano's for multiple hypotheses. Fano's method, however, does not cover the case of testing two simple hypotheses since the lower bound it gives when $r = 2$ is non-positive.

In the next section, we apply the above lemmas to a concrete class where Le Cam's method provides the optimal rate of convergence for a quadratic functional estimation problem and both Fano's and Assouad's lemmas provide the optimal rate for a global density estimation problem. These results are known. See for example Bickel & Ritov (1988) and Devroye (1987) respectively. The common feature of these two lower bound problems is that they both rely on the same hypercube class.

29.3 Two examples

Let $\mathcal{M}$ denote the class of smooth densities f's on $[0,1]$ for which

$$0 < c_0 \le f(x) \le c_1 < \infty, \ \ |f^{(2)}(x)| \le c_2 < \infty, \ \ \int_0^1 f(x)dx = 1.$$

Let us apply the results from Section 2 to derive lower bounds for minimax rates in two cases: a quadratic functional of f, with errors measured by the usual Euclidean distance; and the whole density f, with errors measured by Hellinger distance.

In both cases assume the estimators are based on a sample of n independent observations from some f in $\mathcal{M}$. Write f^n for the joint density, and $\mathcal{P}$ for the corresponding class of product measures, with f in $\mathcal{M}$.

The lemmas will be applied to small perturbation of the uniform density, u, on $[0,1]$. Take g a fixed twice differentiable function on $[0,1]$ for which

$$\int_0^1 g(x)dx = 0, \int_0^1 g^2(x)dx = a > 0 \text{ and } \int_0^1 (g'(x))^2 dx = b > 0.$$

Divide $[0, 1]$ into m disjoint intervals of size $1/m$ and denote their centers by $x_1, \ldots, x_m$. For $j = 1, 2, \ldots, m$, let

$$g_j(x) = cm^{-2} g(mx - x_j)$$

with c small enough so that $|g_j| < 1$. Let

$$\mathcal{M}_m = \{f_\tau = 1 + \sum_{j=1}^m \tau_j g_j(x) : \ \tau = (\tau_1, \ldots, \tau_m) \in \{-1, +1\}^m\},$$

and define the *hypercube class*

$$\mathcal{F}_m = \mathcal{M}_m^n = \{f_\tau^n : f_\tau \in \mathcal{M}_m\}.$$

Note that $\mathcal{M}_m$ is simply the class of perturbed uniform densities with a rescaled g as the perturbation.

Example 1 Consider the quadratic functional

$$T(f) = \int_0^1 (f'(x))^2 dx$$

on $\mathcal{F}_m$. That is, $\theta(f^n) = \theta(f) = T(f)$, which takes values in the real line equipped with its metric $d(\theta, \theta') = |\theta - \theta'|$.

To obtain a minimax lower bound, we might try to use Assouad's lemma or Fano's method. Unfortunately, the functional $\theta(f) = T(f)$ takes the same value on the vertices of the hypercube and therefore the two results give only the trivial lower bound zero. However $\theta(u) = 0$, which differs from $\theta(f_\tau)$ for every τ, which lets us apply Le Cam's method to u^n and f_τ^n. For any fixed τ on the hypercube, it is easy to check that

$$\begin{aligned} H^2(u, f_\tau) &= O(\sum_j \int g_j^2) = O(m \cdot c^2 \cdot a \cdot m^{-5}) = O(m^{-4}), \\ \|u^n - f_\tau^n\|_1 &\le 2H^2(u^n, f_\tau^n) \\ &= 2(1 - (1 - 2^{-1}H^2(u, f_\tau))^n \\ &= 2(1 - (1 - O(m^{-4}))^n), \end{aligned}$$

and

$$T(u) = 0 \ne T(f_\tau) = \sum_j (\int g_j')^2 = c^2 b m^{-2}.$$

If we choose $m = O(n^{1/4})$, then

$$\|u^n \wedge f_\tau^n\| = 1 - \|u^n - f_\tau^n\|_1/2 \geq (1 - O(m^{-4}))^n) > 0,$$

and by Le Cam's method (Lemma 1), a lower bound on the minimax estimation rate is

$$|T(u) - T(f_\tau)| = |T(f_\tau)| = O(n^{-2/4}) = O(n^{-1/2}).$$

Unfortunately, this rate is not optimal, but the minimax optimal rate can be obtained (Bickel & Ritov 1988, Birgé & Massart 1992, Pollard 1993), by Le Cam's method applied to $\mathcal{P}_1 = \{u^n\}$ and $\mathcal{P}_2 = \mathcal{F}_m$. To be precise, an upper bound on the L^1 distance is obtained between u^n and the mixture of the product measures of the densities indexed by the vertices of the hypercube. Hence we can derive a lower bound on $\sup_{P_i \in co(\mathcal{P}_i)} \|P_1 \wedge P_2\|$. Denote by $h_n(x^n) = 2^{-m} \sum_\tau \prod_{i=1}^n f_\tau(x_i)$ the mixture of the product measures. Then the L^1 distance between u^n and h_n can be bounded, for example, by Pollard (1993) or by Birgé & Massart (1992) as

$$\|u^n - h_n\|_1^2 \leq \exp(2^{-1}n^2 \sum_j (\int g_j^2)^2) - 1. \tag{5}$$

Note that

$$n^2 \sum_j (\int g_j^2)^2 = c^4 a^2 n^2 m^{-9},$$

and if we choose $m = O(n^{2/9})$ and c small, then there is an $\epsilon > 0$ such that

$$\|u^n - h_n\|_1^2 \leq \exp(2^{-1}n^2 \sum_j (\int g_j^2)^2) - 1 < (2(1 - 2\epsilon))^2.$$

Hence

$$\|u^n \wedge h_n\| = 1 - \|u^n - h_n\|_1/2 \geq 1 - (2 - 2\epsilon)/2 = \epsilon > 0.$$

Thus by Le Cam's method (Lemma 1) and because

$$T(u) = 0, \;\; T(f_\tau) = c^2 b m^{-2} = O(n^{-4/9}),$$

we have a lower bound decreasing at the slower $n^{-4/9}$ rate, which turns out to be the achievable rate (Bickel & Ritov 1988).

Example 2 Consider estimation of the whole density $\theta(f^n) = \theta(f) = f$ as an element of the space $\mathcal{D} = \{$ densities on [0,1] $\}$ equipped with the Hellinger metric, defined by

$$H^2(f,g) = \int_0^1 \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx.$$

That is, $d(f,g) = H(f,g)$.

Denote by A_j the jth subinterval of size $1/m$ of [0,1] and let

$$d_j(f,g) = \int_{A_j} (\sqrt{f} - \sqrt{g})^2, \;\; \text{then } d(f,g) = \sum_j d_j(f,g).$$

Note that d_j are not pseudo-metrics, but non-negative symmetric functions satisfying the weaker triangle inequality with a universal constant $A = 1/2$. Therefore Assouad's method (Lemma 2) applies with an extra factor $1/2$ in the lower bound.

Since on A_j

$$(\sqrt{f_\tau} + \sqrt{f_{\tau^j}})^2 = f_\tau + f_{\tau^j} + 2\sqrt{f_\tau f_{\tau^j}} \leq 2(f_\tau + f_{\tau^j}) = 4.$$

$$\begin{aligned} d_j(f_\tau, f_{\tau^j}) &= \int_{A_j} (\sqrt{f_\tau} - \sqrt{f_{\tau^j}})^2 \geq 4^{-1} \int_{A_j} (2g_j(x))^2 dx \\ &= c^2 a m^{-5} \equiv \alpha_m. \end{aligned}$$

Note that for $\tau \sim \tau'$, $\|P_\tau \wedge P_{\tau'}\| = \|f_\tau^n \wedge f_{\tau'}^n \geq O((1 - O(m^{-5}))^n)$. Plugging this and α_m into the expression in Assouad's Lemma and maximizing the lower bound by choosing $m = O(n^{1/5})$, we obtain a lower bound of order $O(n^{-4/5})$, which is achieved by a kernel estimator with binwidth $O(n^{-1/5})$; hence the rate is optimal.

Since all the densities in $\mathcal{M}_m$ are bounded from below by $1 - c_g$ for $c_g = c \cdot \sup_x |g(x)|$, one can bound the K-L divergence from above as follows

$$\begin{aligned} K(f_\tau^n, f_{\tau'}^n) = nK(f_\tau, f_{\tau'}) &\leq n \int_0^1 \frac{(\sqrt{f_\tau} - \sqrt{f_{\tau'}})^2}{f_\tau} dx \\ &\leq n(1 - c_g)^{-1} \int_0^1 (\sqrt{f_\tau} - \sqrt{f_{\tau'}})^2 dx \\ &\leq 2n(1 - c_g)^{-1} m \alpha_m \equiv m\gamma_m \end{aligned}$$

where $\gamma_m = 2(1 - c_g)^{-1} n \alpha_m$. Note also that

$$d(\theta(f_\tau), \theta(f_{\tau'})) = H^2(f_\tau, f_{\tau'}) = \sum_j \int_{A_j} (\sqrt{f_\tau} - \sqrt{f_{\tau'}})^2 dx \geq \alpha_m W(\tau, \tau').$$

Recalling Lemma 4, and choosing $m = O(n^{1/5})$ we obtain a lower bound of the optimal order $O(n^{-4/5})$. Therefore, both Fano's method and Assouad's lemma give the optimal rate lower bound for this problem.

Acknowledgments: This work began when I was visiting Yale University in the spring of 1993. I would like to thank members of the Statistics Department for a friendly working environment and Professor David Pollard in

particular for many stimulating discussions on related topics and for many helpful comments on the draft. Thanks are also due to Professor Sergio Verdú for commenting on the draft.

Research supported in part by ARO Grant DAAL03-91-G-007.

29.4 References

Assouad, P. (1983), 'Deux remarques sur l'estimation', *Comptes Rendus de l'Academie des Sciences, Paris, Ser. I Math* **296**, 1021–1024.

Bickel, P. J. & Ritov, Y. (1988), 'Estimating integrated squared density derivatives: sharp best order of convergence estimates', *Sankhyā: The Indian Journal of Statistics, Series A* **50**, 381–393.

Birgé, L. (1986), 'On estimating a density using Hellinger distance and some other strange facts', *Probability Theory and Related Fields* **71**, 271–291.

Birgé, L. & Massart, P. (1992), Estimation of integral functionals of a density, Technical Report 024-92, Mathematical Sciences Research Institute, Berkeley.

Bretagnolle, J. & Huber, C. (1979), 'Estimation des densites: risque minimax', *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **47**, 119–137.

Cover, T. M. & Thomas, J. A. (1991), *Elements of Information Theory*, Wiley, New York.

Devroye, L. (1987), *A Course in Density Estimation*, Birkhäuser, Boston.

Donoho, D. L. & Liu, R. C. (1991), 'Geometrizing rates of convergence, II', *Annals of Statistics* **19**, 633–667.

Donoho, D. L. & Nussbaum, M. (1990), 'Minimax quadratic estimation of a quadratic functional', *Journal of Complexity* **6**, 290–323.

Fan, J. (1991), 'On the estimation of quadratic functionals', *Annals of Statistics* **19**, 1273–1294.

Gilbert, E. N. (1952), 'A comparison of signaling alphabets', *Bell System Technical Journal* **31**, 504–522.

Han, T. S. & Verdú, S. (1994), 'Generalizing the Fano inequality', *IEEE Transactions on Information Theory* **40**, 1247–1251.

Has'minskii, R. & Ibragimov, I. (1978), On the non-parametric estimation of functionals, *in* P. Mandl & M. Hušková, eds, 'Prague Symposium on Asymptotic Statistics', North Holland, Amsterdam, pp. 41–52.

Ibragimov, I. A. & Has'minskii, R. Z. (1981), *Statistical Estimation: Asymptotic Theory*, Springer-Verlag, New York.

Le Cam, L. (1973), 'Convergence of estimates under dimensionality restrictions', *Annals of Statistics* **1**, 38–53.

Le Cam, L. (1986), *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York.

Pollard, D. (1993), Hypercubes and minimax rates of convergence, Technical report, Yale University.

Stone, C. (1984), 'An asymptotically optimal window selection rule for kernel density estimates', *Annals of Statistics* **12**, 1285–1297.