Probab. Th. Rel. Fields 71, 271–291 (1986)

Probability Theory and Related Fields

© Springer-Verlag 1986

On Estimating a Density Using Hellinger Distance and Some Other Strange Facts

Lucien Birgé*

U.E.R. de Sciences Econoniques, Université Paris X – Nanterre, 200 Av. de la République,
F-92001 Nanterre-Cedex, France, and
Mathematical Sciences Research Institute, Berkeley, California, USA

Résumé. On s'intéresse ici aux possibles vitesses d'estimation d'une densité à support compact dans $\mathbb{R}^m$ sous des hypothèses de régularité, lorsque la perte est mesurée par le carré de la distance de Hellinger (on regardera aussi le cas connu des normes $\mathbb{L}^q$ pour $1 \leqq q \leqq 2$) et le risque est le risque minimax sur la famille. On donne une méthode générale permettant de traiter les problèmes dans le cadre de la théorie de l'approximation sous des conditions concernant l'entropie métrique et l' ε-capacité des familles à estimer. Les rapports entre régularité et entropie métrique étant bien connus, nous pourrons aussi traiter les cas classiques et d'autres qui le sont moins. Sous des conditions de bornes inférieures les vitesses sont celles observées pour la norme $\mathbb{L}^1$ mais elles diffèrent dans le cas général. On montre aussi que les restrictions sur la compacité du support ou la régularité sont indispensables et que leur absence mène à l'impossibilité d'obtenir une estimation raisonnable en ce sens que n'importe quelle suite d'estimateurs sera arbitrairement mauvaise en un point au moins. Un résultat analogue est vrai sous des conditions de régularité.

I. Introduction

In recent years, numerous papers have been devoted to studying the problem of global estimation of a density function on $\mathbb{R}^m$, especially if this density is known to belong to a family of functions with a given smoothness, the loss being measured by a power of the distance associated to the $\mathbb{L}^q$ norm for $1 \leqq q \leqq +\infty$. Pioneering work was done by Bretagnolle and Huber [5] and a very extensive study may be found in Ibragimov and Khas'minskii [8] and [9]. The best obtainable rates of convergence are now well known for the classical smooth families. Some connected results appeared in slightly different directions: in [5], some results are given in particular cases using the Hellinger

[*] This work was carried out during a visit of the author at the Mathematical Sciences Research Institute at Berkeley

distance as a loss function. In [18] and [19] the classical results are extended to the estimation of derivatives of a density. In his recent paper [7] Devroye proved that estimation is impossible unless there exist uniform bounds for both the smoothness of the functions in the family and the compactness of their support. Actually a similar result was already known from Ibragimov and Khas'minskii (see [8] and [9]) but Devroye's theorem is stronger. Related facts are also found in [4]. The study of more general rates of convergence than the usual polynomial ones is carried out in [2] or [3] where it is proved that all "reasonable" rates may occur. Those special facts concerning the limits of the possibility of estimating a function without enough "a priori" restrictions, the fact that the speeds which are found when one measures the loss using Hellinger distance are different from those one finds when one works with $\mathbb{L}^1$ or $\mathbb{L}^2$ norms and the lack of unification in the theory suggested the following study with those three purposes:

i) Try to get a general way of finding the best rates of convergence, even for more general families than the usual ones, by putting the problem in the framework which is given in [2] or [3] and links the rates to the dimensional properties of the family. This allows us to express the assumptions in terms of approximation theory and then use all the well-known results from this theory to apply it to any particular case of density estimation. With these tools we shall be able to treat families with special moduli of continuity which are not found in the literature.

ii) Simultaneously give an extensive treatment for the case of Hellinger distance h.

iii) Indicate useful and general tools for finding lower bounds and use them in order to strengthen some known results in this direction and especially complement the results of Devroye [7].

We shall also recover a number of classical results from this theory, mainly in Sect. 3. Most of this section consists in a summary of previous results and is included here for sake of completeness as an illustration of the relations between the rates of convergence and the metric structure of the parameter space.

Actually, there is no fundamental difference in the treatment between $\mathbb{L}^1$ and Hellinger; the fact that different rates may occur when one uses Hellinger distance is related to this very simple remark: when you compute $h(f; f+g)$ where f is a density and g a small perturbation of order ε for the sup norm, if f is not small, this distance will be of order ε just like the $\mathbb{L}^1$ distance, but if f is also of order ε then h will be of order $\sqrt{\varepsilon}$, h and $\mathbb{L}^1$ being no longer comparable. This will lead us to distinguish in the sequel between two cases, α) and β): either α) our family Θ has a uniform lower bound and then h just behaves as the variation distance and so does not deserve a special treatment (that was the assumption used in [2] and [3]). Or β) there is no such lower bound, the densities in Θ may be arbitrarily small on some interval and we then get different estimates for the rates, such as the ones which were found in [5]. When dealing with the $\mathbb{L}^q$ norm, only case α) will be relevant.

In Sect. 2 we shall set up our assumptions using a dimensional point of view (see [3, 13] and [14]); they will allow us to give a simplified and unified presentation of the results. Translating smoothness assumptions into those is straightforward most of the time and is based on known results in approximation theory which we shall use freely. Section 3 will show how those assumptions give us upper bounds for the risk and Sect. 4 will be devoted to the tools necessary to get the corresponding lower bounds. Also a result in the spirit of [7] Theorem 1, ii) will be given for the usual smooth families. This will be Corollary 4.5 and will lead naturally to the problem of Sect. 5. In [7], Devroye proved that if we consider all density functions on [0; 1] which are bounded by 2, for any sequence $\{c_n\}$ converging to 0 and any sequence $\{\hat{f}_n\}$ of estimates there exists a density f such that

$$\limsup_n c_n^{-1}\, \mathbb{E}_f[\int |\hat{f}_n(x)-f(x)|\, dx] \geqq 1. \tag{1.3}$$

Notice the "*lim sup*" here which suggests the two questions: is this result a truly "*asymptotic*" one or not; what is to be expected about the "*lim inf*"? The proof of the analogous Corollary 4.5 suggests that it might be really "*asymptotic*" and, as L. Le Cam pointed out to me, in the finite dimensional case (ordinary parametric case), it is always possible to design a sequence of estimates for which the "*lim inf*" is uniformly small at each point. It is not difficult to see that in non-parametric cases with compact parameter spaces having finite ε-entropy for all positive ε, the same will be true. On the contrary, in the case studied by Devroye, the parameter space is not compact and contains ε-separated subsets of arbitrarily large cardinality. Using such subsets it is then possible to see that (1.3) also holds with "$\inf_n$" instead of "$\limsup_n$" provided that the sequence $\{c_n\}$ has a suitable upper bound. We shall also prove the analogous result for Hellinger distance. See [4] for similar non-asymptotic results for Hellinger balls.

Before proceeding, let us set up some notations and recall a few results. For two probability measures P and Q, with densities f and g with respect to μ, we may consider the following distances

$$h(P;Q)=\tfrac{1}{2}\int(\sqrt{f}-\sqrt{g})^2\, d\mu;$$
$$d_q(P;Q)=\|f-g\|_q=[\int |f-g|^q\, d\mu]^{1/q};$$
$$D(P;Q)=\tfrac{1}{2}\int |f-g|\, d\mu=\tfrac{1}{2}d_1(P;Q);$$

which are respectively Hellinger distance, $\mathbb{L}^q$-distance and total variation distance. Some other quantities are also of interest which are the testing affinity, the Hellinger affinity and the Kullback information:

$$\pi(P;Q)=\int(f\wedge g)\, d\mu; \qquad \rho(P;Q)=\int\sqrt{fg}\, d\mu; \qquad K(P;Q)=\int \log\frac{f}{g} f\, d\mu.$$

The following relations are well-known (see for example [6, 13] or [14]):

$$\pi=1-D; \qquad h^2=1-\rho; \qquad h^2 \leqq D \leqq h\sqrt{2}; \qquad \pi^2 \leqq \rho^2 \leqq \pi(2-\pi), \tag{1.1}$$

$$\rho(P^n;Q^n)=\rho^n(P,Q); \qquad K(P^n;Q^n)=nK(P,Q). \tag{1.2}$$

If Θ is our parameter space assumed to be a set of densities with respect to the measure μ, the points in Θ will be called alternatively θ, f_θ (density) or P_θ (probability) and by convention for any distance d on Θ $d(\theta;\theta')=d(f_\theta;f_{\theta'})=d(P_\theta;P_{\theta'})$. Given n i.i.d. observations following the unknown law P_θ, the minimax risk will be defined by

$$R_M(d^q,n)=\inf_{T_n}\sup_{\theta\in\Theta}\mathbb{E}_\theta[d^q(P_\theta;T_n)]$$

where T_n is any estimate (considered as a random probability) depending on n observations. Sometimes, when $d=d_q$ and $q>1$ we shall also use the normalized risk, which is

$$R'_M(d_q^q,n)=\inf_{T_n}\sup_{\theta\in\Theta}\mathbb{E}_\theta\left[\frac{d_q^q(P_\theta;T_n)}{\|f_\theta\|_q^q}\right].$$

We shall also need a few dimensional concepts concerning metric spaces. $\mathscr{B}(x,r)$ will denote the open ball of center x and radius r. If S is a subset of a metric space, N is called an ε-net for S if any point in S is within a distance ε of some point in N; S is said to be ε-separated if any two points in S are at a distance greater than ε. For any totally bounded subset Θ of a metric space we may define its ε-entropy N_ε. It is the logarithm (base 2) of the minimum cardinality of an ε-net. Its ε-capacity $\mathscr{C}_\varepsilon$ is the logarithm of the maximum cardinality of an ε-separated set. The definitions and properties of those quantities and the way of computing them in many cases are given in the remarkable paper [12] by Kolmogorov and Tikhomirov.

Throughout this paper we shall denote by A_i, C or C_i different constants but the A_i's will always have the same meaning (issued from our assumptions) while the C_i's will be generic constants (often depending on the former ones). For two sequences $\{u_n\}$ and $\{v_n\}$, $u_n \asymp v_n$ will mean that $C_1 u_n \leqq v_n \leqq C_2 u_n$ for some positive constants C_1, C_2. In this case we shall say that $\{u_n\}$ and $\{v_n\}$ are of the same order.

II. Smoothness and Dimension

As was mentioned in the introduction a smoothness assumption for a family of densities generally implies some dimensional properties. As a very simple illustration we may consider the case of all density functions on $[0;1]$ satisfying a uniform Lipschitz condition such as

$$|f(x)-f(y)|\leqq C|x-y| \qquad x,y\in[0;1].$$

Then the ε-entropy N_ε of this set is of order ε^{-1}. Using the techniques introduced in [2] and [3] we immediately find, by solving the equation $n\varepsilon^2=C_1\varepsilon^{-1}$ that an upper bound for the speed of estimation will be $C_2 n^{-1/3}$ when the risk is measured by d_1 for example. Analogous considerations using results concerning ε-capacity of this set give us the lower bounds. Since the relations between smoothness and dimension have been known for a long time by workers in approximation theory (see [12] and more recent results in [15] and [16] for example), we may express our assumptions in terms of dimensional properties of the density families. This will allow us to give a unified treatment of different problems. Our A1 reduces to an assumption concerning the metric ε-entropy of Θ, A2 is closely related to the ε-capacity of Θ (see [12] for details). We shall also need two other technical assumptions, ULB being used to distinguish between the two possible behaviors of Hellinger distance and CS to ensure the possibility of estimating (see counterexamples in [7] and Sect. 4).

CS. *All densities f in Θ have the same compact support.*

In this case we may always assume that μ is Lebesgue probability on the common support. This rescaling will not change the index of smoothness of the family but only the constants which are irrelevant here.

ULB. *There exists some constant $L>0$ such that $f \geqq L$ for f in Θ.*

The two following assume that k is a decreasing function on $[0;1]$.

A1(d,k). *The set Θ is metricized by some distance d and for each ε in $]0;1]$ there exists an ε-net $N(\varepsilon)$ for Θ the cardinal of which is smaller than $8^{A_0 k(\varepsilon)}$. Also $A_0 k(1) \geqq 1/2$.*

A2(k). *For any ε in $]0;1]$ there exists some function g_ε with support on I, $\mu(I) = \nu$ and $\int_I g_\varepsilon \, d\mu = 0$ with*

$$\|g_\varepsilon\|_\infty \leqq \varepsilon \quad \textit{and} \quad \mu\{|g_\varepsilon| \geqq A_1 \varepsilon\} \geqq A_2 \nu. \tag{2.1}$$

There exist r translates $I+x_i$ of the support of g_ε which do not intersect and are such that the family

$$\left\{ f + \sum_{i=1}^{r} \delta_i g_\varepsilon(x-x_i) \right\} \tag{2.2}$$

belongs to Θ with either

$$\text{i) } \delta_i = \pm 1 \qquad \text{ii) } \delta_i = 0 \textit{ or } 1,$$

f being constant and equal to c on $\bigcup_{i=1}^{r} I + x_i$. Moreover the following inequalities hold

$$\nu \leqq A_3 k^{-1}(\varepsilon); \qquad 1 \geqq r\nu \geqq A_4 > 0. \tag{2.3}$$

and either

$$\alpha)\ 2 \geqq c \geqq A_5 > 0 \quad \textit{or} \quad \beta)\ A_1 \varepsilon \leqq c \leqq A_5 \varepsilon. \tag{2.4}$$

It is clear that under ULB α) only is possible, β) is useful only in the case that $d=h$ and ULB does not hold.

Remark. 1) A1 is actually much stronger that what is needed in order to apply Theorem 3.1 but since it is easier to check and is satisfied in all subsequent applications, there is no need for the refined version given in [3].

2) In Sect. 3 we shall always assume that A1 holds with $d=d_\infty$. This implies that Θ is compact for d_∞ and therefore uniformly bounded. In this case all d_p's for $1<p\leqq 2$ are bounded by some multiple of h as noticed in [3].

In the simplest cases the function k will be a power function of ε which leads us to the following

Definition 2.1. If a family (Θ, d) satisfies A1 and A2 with $k(\varepsilon)=\varepsilon^{-\delta}$, $\delta>0$ it will be said to have an exponent of dimension δ.

This exponent is easily computed for the classical smooth families. For example assume that for f in Θ

$$|f^{(p)}(x)-f^{(p)}(y)|\leqq C|x-y|^\alpha \quad p\geqq 0,\ 0<\alpha\leqq 1,\ x, y\in[0;1]. \tag{2.5}$$

Then the family will be said to have an exponent of smoothness $s=p+\alpha$ and it is known that for such a family $\delta=s^{-1}$. For densities on some compact convex subset of dimension m in $\mathbb{R}^m$, if (2.5) holds ($f^{(p)}$ being now any partial derivative of order p) we shall find $\delta=\frac{m}{s}$. We may extend this to anisotropic smoothness on $\mathbb{R}^m$, which means that, as functions of coordinate x_i, $1\leqq i\leqq m$, the densities have smoothness s_i, but s_i depends on i. In this case we may define an exponent of global smoothness s by $s^{-1}=m^{-1}\sum_{i=1}^m s_i^{-1}$ and the following will hold:

Proposition 2.2. *If Θ is a family of densities on $\mathbb{R}^m$ with common compact convex support of dimension m and exponent of smoothness $s>0$, then* A1(d_∞, k) *and* A2(k) *hold with an exponent of dimension* $\delta=\frac{m}{s}$.

Much more general families can be used. In [2, 3] we consider families with a given concave modulus of continuity, i.e. the set of all functions satisfying

$$|f(x)-f(y)|\leqq C\omega(|x-y|) \tag{2.6}$$

ω being a concave non-decreasing function and C a constant such that $\omega(\varepsilon)\geqq\varepsilon$. The smooth families with $s\leqq 1$ are of this type. We may also consider generalizations of (2.5) of the form

$$|f^{(p)}(x)-f^{(p)}(y)|\leqq C\omega(|x-y|) \tag{2.7}$$

and multidimensional analogues (isotropic or not). For the general case on $\mathbb{R}^m$ with different conditions of type (2.7) for each coordinate, denoting by $\eta=\psi_i(\varepsilon)$ the reciprocal function of $\eta\mapsto\varepsilon=\eta^{p_i}\omega_i(\eta)$ we would find that

$$\operatorname{Log} k(\varepsilon)=-\sum_{i=1}^m \operatorname{Log}\psi_i(\varepsilon)$$

which is a simple extension of the ordinary case when $\omega_i(\eta)=\eta^{\alpha_i}$. We shall not insist on this because those properties are only related to approximation theory and may be checked by the arguments used in [12]. Actually other types of families could be dealt using assumptions A1 and A2 and more general functions k may occur (see [2] and [3] for illustrations of this fact). Properties of type A1 may be obtained directly for distances other than d_∞, depending upon the conditions imposed on the family (examples may be found in [16]). Here we shall always assume A1(d_∞, k) and then derive A1(d_q, k) from the fact that $d_q \leqq d_\infty$ when the dominating measure is a probability, which is assumed here if CS holds. For h, the problem is different. It is known from [2, 3] that under some ULB condition we have $h \leqq C d_\infty$; in this case there will not be any problem $\Big($just change h into $C^{-1}h$ which is not important for the type of result we look for, or change $k(\varepsilon)$ into $k\left(\frac{\varepsilon}{C}\right)\Big)$. But without ULB condition we only know that $2h^2 \leqq d_1 \leqq d_\infty$ so that A1(d_∞, k) implies A1(h, k_1) with $k_1(\varepsilon)=k(\varepsilon^2)$. As a consequence if a family has an exponent of dimension δ, the rates of convergence will be functions of δ for d_q or h under ULB and functions of 2δ for h without ULB, assuming that case β) of A2 holds. But before we come to this, let us give the following useful but simple lemma:

Lemma 2.3. *Assume that g_ε is a function with support on I, $\mu(I)=v$, that* (2.1) *holds and that f is constant and equal to $c \geqq \varepsilon$ on I. Then*

$$A_1^q A_2 \varepsilon^q v \leqq d_q^q(f; f+g) = 2^{-q} d_q^q(f-g; f+g) \leqq \varepsilon^q v, \tag{2.8}$$

$$\frac{A_1^2 A_2}{12}\,\frac{\varepsilon^2 v}{c} \leqq h^2(f; f \pm g) \leqq \frac{1}{2}\,\frac{\varepsilon^2 v}{c}, \tag{2.9}$$

$$\frac{A_1^2 A_2}{2}\,\frac{\varepsilon^2 v}{c} \leqq h^2(f+g; f-g) \leqq \frac{\varepsilon^2 v}{c}. \tag{2.10}$$

III. Upper Bounds for the Risk

The idea of relating the speed of convergence with the dimensional properties of the parameter space and of the construction of "universal" dimensional estimates goes back to Le Cam [13]. We shall use here the following result which may be found in [2, 3].

Theorem 3.1. *Suppose we observe n i.i.d. random variables from an unknown distribution P_θ in some parameter space (Θ, d) with metric d. Suppose also that the following properties are true for some constants B, D:*

i) $d(\theta; \theta') \leqq B h(P_\theta; P_{\theta'}) \qquad P_\theta, P_{\theta'} \in \Theta$,

ii) *for some ε with $n\varepsilon^2 \geqq D \geqq \frac{1}{2}$, there exists an ε-net N in (Θ, d) such that*

$$\operatorname{Card}[N \cap \mathscr{B}(\theta, 2^j \varepsilon)] \leqq 2^{jD} \qquad \forall\, \theta \in \Theta,\ j \geqq 3.$$

Then there exists an estimate $\hat{\theta}_n$ of θ such that

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta[d^t(\hat{\theta}_n;\theta)] \leqq C(B,t)\,\varepsilon^t. \tag{3.1}$$

Let us now apply this theorem to our particular case. If we restrict ourselves to d_1 or h, i) is certainly true because of (1.1). It is also true for d_q if $1<q\leqq 2$ and $\mathrm{A}1(d_\infty, k)$ holds because Θ is then uniformly bounded (see [3]). For $q>2$ this becomes false and the treatment of such cases does not naturally follow from this theory. That is why we shall restrict ourselves to the case $1\leqq q\leqq 2$ throughout this section.

Let us now assume that $\mathrm{A}1(d_\infty, k)$ holds. Then we may distinguish between two cases:

α) $d=d_q$, $1\leqq q\leqq 2$ or $d=Ch$ and ULB holds.

β) $d=h$ and ULB does not hold.

In case α) $A1(d,k)$ holds; in case β) $A1(d,k_1)$, $k_1(\varepsilon)=k(\varepsilon^2)$. But in order to apply Theorem 3.1 we only need to check that

$$n\varepsilon^2 \geqq D \geqq \tfrac{1}{2} \quad \text{with } \alpha)\ D=A_0 k(\varepsilon) \quad \text{or} \quad \beta)\ D=A_0 k(\varepsilon^2),$$

that is

$$\alpha)\ n\varepsilon^2 \geqq A_0\, k(\varepsilon) \qquad \beta)\ n\varepsilon^2 \geqq A_0\, k(\varepsilon^2). \tag{3.2}$$

If ε satisfies (3.2) we shall immediately get that

$$R_M(d^t, n) \leqq C\varepsilon^t \tag{3.3}$$

which leads to the following corollary using Proposition 2.2.

Corollary 3.2. *Assume that Θ is a family of densities for which $A1(d_\infty, \varepsilon^{-\delta})$ holds with a distance d satisfying either α) or β), then there exist estimates $\hat{\theta}_n$ of θ depending on n observations such that*

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta[d^t(\hat{\theta}_n;\theta)] \leqq \begin{cases} C_1\, n^{\frac{-t}{2+\delta}} & \textit{case } \alpha \\ C_2\, n^{\frac{-t}{2(1+\delta)}} & \textit{case } \beta \end{cases}.$$

If Θ is a family of smooth densities on some compact convex subset of $\mathbb{R}^m$ of dimension m with an exponent of global smoothness s then

$$R_M(d^t, n) \leqq C_3\, n^{\frac{-ts}{2s+m}} \qquad \textit{case } \alpha,$$

$$R_M(h^t, n) \leqq C_4\, n^{\frac{-ts}{2(s+m)}} \qquad \textit{case } \beta.$$

These are the classical results of [2, 3, 5, 8, 9, 18], but we may also deal with more special families using for example the moduli of continuity. Let us just consider two simple cases using Theorem 3.1.

Example 1. Let us look at densities on $[0;1]$ satisfying (2.6) with $\omega(x) = -x^\alpha \log x$; $0<\alpha\leqq 1$; $x\leqq\frac{1}{3}$. Then, following the same computations as in [2, 3], we easily find

Case α: $R_M(d^t, n) \leqq C_5\, n^{\frac{-t\alpha}{2\alpha+1}} \log n^{\frac{t}{2\alpha+1}}$,

Case β: $R_M(h^t, n) \leqq C_6\, n^{\frac{-t\alpha}{2\alpha+2}} \log n^{\frac{t}{2\alpha+2}}$.

Example 2. Same case as Example 1 but $\omega(x) = \frac{-1}{\log x}$ for $x \leqq 0.1$. Then

Case α: $R_M(d^t, n) \leqq C_7 (\log n)^{-t}$,

Case β: $R_M(h^t, n) \leqq C_8 (\log n)^{-t/2}$,

which shows that very slow rates may occur (those bounds being optimal as we shall see).

Other cases including classes of analytic functions could also be treated but to get lower bounds for those classes one generally needs assumptions that are more complex than A2. This would make this presentation more obscure. For examples of such cases see [2] or [3] and [8].

IV. Lower Bounds for the Risk

Let us begin with two general theorems which could be used in many circumstances to find lower bounds. They are Fano's Lemma and Assouad's Lemma. The first one has been applied very often in [11, 8, 9, 4] (see [10] or [3] for a proof) and 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. Assouad's Lemma is more specific but gives better constants and is also easier to apply. It is in the spirit of Proposition 2.1 of [5] or [18] and may be found in [1].

Fano's Lemma (see [3], Lemma 2.7). *Let S be a finite set of probabilities $S = \{P_1; \ldots; P_r\}$ such that*

$$K(P_i, P_j) \leqq K \qquad \forall P_i, P_j \in S.$$

Then for any estimate ϕ with values in $\{1; 2; \ldots; r\}$ we have

$$r^{-1} \sum_{i=1}^{r} P_i[\phi \neq i] \geqq \gamma \tag{4.1}$$

if

$$K + \log 2 \leqq (1-\gamma) \log (r-1). \tag{4.2}$$

In the case of n i.i.d. observations, we have $P_i = Q_i^n$ and $K(P_i, P_j) = nK(Q_i, Q_j)$ so that (4.2) is equivalent to

$$n[\sup_{i,j} K(Q_i, Q_j)] + \log 2 \leqq (1-\gamma) \log (r-1). \tag{4.3}$$

We shall now prove a suitable version of Assouad's Lemma following the original treatment of [1]. The hypercube $\mathscr{C}(r)$ of dimension r will be the set $\{0; 1\}^r$ with points $x = \{x_1; \ldots; x_r\}$, $x_i = 0$ or 1 and distance Δ:

$$\Delta(x, x') = \sum_{i=1}^{r} |x_i - x_i'|.$$

Assouad's Lemma. *Assume we are given n i.i.d. observations from some unknown distribution in some parameter space Θ. Assume that some finite subset $\Theta_{\mathscr{C}}$ of Θ is in one-to-one correspondence with $\mathscr{C}(r)$:*

$$x \in \mathscr{C}(r) \leftrightarrow P_x \in \Theta_{\mathscr{C}}$$

with the following properties

$$h^2(P_x; P_{x'}) \leqq \beta_i \leqq 1 \qquad \text{if } x_j = x'_j,\ j \neq i. \tag{4.4}$$

ℓ is a loss function defined on the space of probabilities and satisfying for all P:

$$\begin{gathered} \ell(P; P_x) \geqq \sum_{i=1}^{r} [x_i \ell_i(P) + (1-x_i)\, \ell'_i(P)]; \\ \ell_i(P), \ell'_i(P) \geqq 0; \quad \ell_i(P) + \ell'_i(P) \geqq \alpha > 0 \quad i = 1, \ldots, r. \end{gathered} \tag{4.5}$$

Let $\hat{P}_n$ be any estimate and let R_B be the Bayes risk of $\hat{P}_n$ with respect to the uniform probability on $\Theta_{\mathscr{C}}$:

$$R_B = 2^{-r} \sum_{x \in \mathscr{C}} \int \ell(\hat{P}_n, P_x)\, dP_x^n.$$

Then if $\beta = r^{-1} \sum_{i=1}^{r} \beta_i$,

$$R_B \geqq \frac{r\alpha}{2} \max[1 - (2n\beta)^{1/2}; \tfrac{1}{2}(1-\beta)^{2n}]. \tag{4.6}$$

Proof. Using (4.5) we may write

$$\begin{aligned} R_B &\geqq 2^{-r} \sum_{i=1}^{r} [\sum_{x|x_i=1} \int \ell_i(\hat{P}_n)\, dP_x^n + \sum_{x|x_i=0} \int \ell'_i(\hat{P}_n)\, dP_x^n] \\ &= \tfrac{1}{2} \sum_{i=1}^{r} [\int \ell_i(\hat{P}_n)\, d[2^{-r+1} \sum_{x|x_i=1} dP_x^n] + \int \ell'_i(\hat{P}_n)\, d[2^{-r+1} \sum_{x|x_i=0} dP_x^n]] \\ &\geqq \frac{\alpha}{2} \sum_{i=1}^{r} \pi[2^{-r+1} \sum_{x|x_i=1} P_x^n; 2^{-r+1} \sum_{x|x_i=0} P_x^n]. \end{aligned}$$

Inf(x, y) being a concave function of (x, y), Jensen's inequality entails that

$$\begin{gathered} R_B \geqq \frac{\alpha}{2} \sum_{i=1}^{r} [2^{1-r} \sum_{J_i} \pi(P_x^n, P_{x'}^n)], \\ J_i = \{(x, x') \,|\, x_i = 0,\ x'_i = 1,\ x_j = x'_j \text{ for } j \neq i\}. \end{gathered}$$

For the pairs $(P_x, P_{x'})$ such that (x, x') belongs to J_i, (4.4) holds. Using (1.1) and (1.2) this leads to $\rho^2(P_x^n, P_{x'}^n) \geqq (1-\beta_i)^{2n}$ and

$$\pi(P_x^n, P_{x'}^n) \geqq 1 - [1 - (1-\beta_i)^{2n}]^{1/2} \geqq \max[1 - (2n\beta_i)^{1/2}; \tfrac{1}{2}(1-\beta_i)^{2n}]$$

which implies

$$R_B \geqq \frac{r\alpha}{2} r^{-1} \sum_{i=1}^{r} \max[1-(2n\beta_i)^{1/2}; \tfrac{1}{2}(1-\beta_i)^{2n}];$$

the conclusion follows from a convexity argument. □

In the sequel we shall always use this lemma with all β_i's equal to β but this version may prove useful in other cases.

The use of (4.6) in conjunction with assumption A2 will then become fairly obvious:

Corollary 4.1. *Let $\Theta_{\mathscr{C}}$ be a subset of Θ of the following type*

$$\Theta_{\mathscr{C}} = \left\{ f + \sum_{i=1}^{r} \delta_i g_i \right\} \quad \textit{with i) } \delta_i = \pm 1 \textit{ or ii) } \delta_i = 0 \textit{ or } 1.$$

f is a density which is constant on the sum of the supports of the g_i's which supports are disjoint and the g_i's are translates of the same function g with $\int g(x)\,d\mu(x)=0$. Assume that $\ell(P,Q)=d_q^q(P,Q)$ or $h^2(P,Q)$ (in this case we take $q=2$ in (4.9)) and

$$h^2(f+g_1; f-g_1) \leqq \beta \;(\textit{case i}) \quad \textit{or} \quad h^2(f; f+g_1) \leqq \beta \;(\textit{case ii}) \tag{4.7}$$

$$\ell(f+g_1; f-g_1) \geqq \alpha \;(\textit{case i}) \quad \textit{or} \quad \ell(f; f+g_1) \geqq \alpha \;(\textit{case ii}). \tag{4.8}$$

Then the (uniform) Bayes risk on $\Theta_{\mathscr{C}}$ for any estimate depending on n i.i.d. variables satisfies

$$R_B \geqq 2^{-q} r\alpha \max[1-(2n\beta)^{1/2}; \tfrac{1}{2}(1-\beta)^{2n}]. \tag{4.9}$$

Proof. We shall just look at case i) for $\ell = d_q^q$ and apply Assouad's Lemma to $\Theta_{\mathscr{C}}$ which obviously satisfies (4.4) since all g_i's are translates of the same function with disjoint supports. Because of the definition of ℓ we easily see that

$$\ell\left(P, f + \sum_{i=1}^{r} \delta_i g_i\right) \geqq \sum_{i=1}^{r} \int_{I_i} \left| \frac{dP}{d\mu} - (f + \delta_i g_i) \right|^q d\mu$$

I_i being the support of g_i, which implies (4.5) with

$$\ell_i = \int_{I_i} \left| \frac{dP}{d\mu} - f - g_i \right|^q d\mu \qquad \ell_i' = \int_{I_i} \left| \frac{dP}{d\mu} - f + g_i \right|^q d\mu$$

and because $(a+b)^q \leqq 2^{q-1}(a^q + b^q)$

$$\ell_i + \ell_i' \geqq 2^{1-q} \int_{I_i} \left(\left| \frac{dP}{dx} - f - g_i \right| + \left| \frac{dP}{dx} - f + g_i \right| \right)^q d\mu$$
$$\geqq 2^{1-q} \int_{I_i} |2g_i|^q\, dx \geqq 2^{1-q} \alpha,$$

hence the result. □

Remark. The assumption that f is constant wasn't used in this case and could be removed but is useful for case ii) or when $\ell = h^2$; also in what follows, the constancy of f could be removed when $\delta_i = \pm 1$ and $\ell = d_q^q$.

Proposition 4.2. *Assume that* A2(k) *holds and* ε *satisfies* $n\varepsilon^2 \leqq A_6 k(\varepsilon^2)$ *if* $d=h$ *and* β) *holds,* $n\varepsilon^2 \leqq A_6 k(\varepsilon)$ *in all other cases. Then*

$$R_M(d^q, n) \geqq C\varepsilon^q. \tag{4.10}$$

Proof. Let us prove it when $d=h$ in case β, the other cases being analogous. Then A2(k) holds with (2.4) β) and (2.2) i) (say). The hypercube $\Theta_{\mathscr{C}}$ of Corollary 4.1 is then given by the family (2.2) with $g_i(x) = g_{\varepsilon^2}(x - x_i)$. Using (2.10) and (2.4) we may choose $\beta = \varepsilon^2 v A_1^{-1}$ in (4.7) and $\alpha = A_1^2 A_2 (2A_5)^{-1} \varepsilon^2 v$ in (4.8). Putting these values in (4.9) and using (2.3) gives the result. □

We may now use our whole set of assumptions to put together the upper and lower bounds using Theorem 3.1 and Proposition 4.2 to get

Proposition 4.3. *Assume that* A1(d_∞, k) *and* A2(k) *hold and that* ε_n *is a sequence such that*

$$n\varepsilon_n^2 \asymp k(\varepsilon_n) \text{ case } \alpha); \qquad n\varepsilon_n^2 \asymp k(\varepsilon_n^2) \text{ case } \beta)$$

then

$$R_M(d^q, n) \asymp \varepsilon_n^q.$$

Proof. We just consider case α. Then by assumption $n\varepsilon_n^2 \leqq Ck(\varepsilon_n)$; choosing $\varepsilon = [(A_0/3C)^{1/2} \vee 1]\, \varepsilon_n$, (3.2) will hold and (3.3) will give an upper bound of order ε^q which is also of order ε_n^q. The lower bound case is similar using the fact that (4.10) holds if $n\varepsilon^2 \leqq A_6 k(\varepsilon)$. □

As a consequence we find the well-known fact that the rates of convergence for smooth densities with exponent s in $\mathbb{R}^m$ given by Corollary 3.2 are the optimal ones. The same holds for the examples of Sect. 3.

We shall now prove an analogue of Theorem 1, ii) of [7] in the case of a given convergence to zero of the risk. Devroye's result could be proved in the same way but since we shall get an improved version in the next section we shall not consider his case but the following one, deliberately restricted for the sake of simplicity and brevity: Θ is assumed to be a family of densities with support on [0; 1], $d=h$ and ULB holds. It could easily be generalized to $\mathbb{R}^m$, d_q or case β. We shall need a slight modification of A2 allowing us to use simultaneously different g_ε's corresponding to different values of ε. It is practically no more restrictive than A2.

A′2(k). *Suppose that the* ε_i*'s,* $i \geqq 1$ *are small enough* ($\varepsilon_i \leqq \eta_i$*, say), that they form a decreasing sequence and that the* g_i*'s* ($g_i = g_{\varepsilon_i}$) *are chosen according to* A2 *with* (2.1). *Then for some numbers* r_i *the family*

$$\Theta_\infty = \left\{ 1 + \sum_{i \geqq 1} \sum_{j=1}^{r_i} \delta_{ij} g_i(x - x_{ij}) \right\}, \qquad \delta_{ij} = 0 \text{ or } 1,$$

is a subset of Θ, *the functions* $g_i(x - x_{ij})$ *have disjoint supports and*

$$A_4 2^{-i} \leqq r_i v_i \leqq 2^{-i}; \qquad v_i \leqq A_3 k^{-1}(\varepsilon_i). \tag{4.11}$$

Proposition 4.4. *Let us consider two decreasing sequences* $\{a_n\}$ *and* $\{c_n\}$, *converging to zero with*

$$C_1 n a_n^2 \leqq k(a_n) \leqq C_2 n a_n^2. \tag{4.12}$$

If A′2(k) *holds and* $\{\hat{f}_n\}$ *is a sequence of estimates, there exists at least one point* θ *in* Θ_∞ *such that*

$$\limsup_n c_n^{-1} a_n^{-2} \mathbb{E}_\theta[h^2(f_\theta; \hat{f}_n)] \geqq 1. \tag{4.13}$$

Notice that under assumptions A1 and A2 (4.12) implies that a_n^2 is the speed of convergence of $R_M(h^2, n)$. Statement (4.13) shows that it cannot be improved for a given sequence of estimates. For the smooth densities it implies the following

Corollary 4.5. *Θ being a family of smooth densities on* $[0; 1]$ *with exponent s and* $\{c_n\}$ *a sequence decreasing to zero, for any sequence* $\{\hat{f}_n\}$ *of estimates there exists a fixed f in Θ such that*

$$\limsup_n c_n^{-1} n^{\frac{2s}{2s+1}} \mathbb{E}_f[h^2(\hat{f}_n; f)] \geqq 1.$$

Proof of Proposition 4.4. Let us denote by δ_i the vector $\{\delta_{ij}\}_{j=1,\ldots,r_i}$ and call the δ_i's the coordinates of the points of Θ_∞. Θ_k is the subset of points with $\delta_i = 0$ for $i > k$. Choose a sequence $\{m_i\}$ such that $c_{m_i} \leqq C_3 C_4^2 2^{-i-2}$ and a sequence n_i satisfying

$$n_i \geqq m_i; \qquad C_4 a_{n_i} \leqq \eta_i, \tag{4.14}$$

$$a_{n_{i+1}} \leqq \frac{C_4 C_3}{5} n_i^{-1/2} 2^{-i/2} a_{n_i}^2, \tag{4.15}$$

with

$$C_3 = \frac{A_1^2 A_2 A_4}{100}; \qquad C_4 = \left(\frac{C_1}{4A_3}\right)^{1/2} \wedge 1.$$

This is always possible inductively since $a_n \to 0$ when $n \to +\infty$. Fix $\varepsilon_i = C_4 a_{n_i}$ and notice that if f_0 and f_1 in Θ_∞ just differ by one coordinate δ_i and only one δ_{ij} then by (2.9)

$$\frac{A_1^2 A_2 \nu_i \varepsilon_i^2}{12} \leqq h^2(f_0; f_1) \leqq \frac{\nu_i \varepsilon_i^2}{2}, \tag{4.16}$$

and if f_0 is in Θ_k and f_1 has the same k first coordinates

$$h^2(f_0; f_1) \leqq \sum_{i>k} r_i \frac{\nu_i \varepsilon_i^2}{2} \leqq \sum_{i>k} 2^{-i-1} \varepsilon_i^2 \leqq 2^{-k-1} \varepsilon_{k+1}^2.$$

This implies using (1.1), (1.2), (4.15) that

$$\rho^{n_k}(f_0; f_1) \geqq \exp\left[n_k \log\left(1 - \frac{\varepsilon_{k+1}^2}{2^{k+1}}\right)\right] \geqq 1 - n_k \frac{\varepsilon_{k+1}^2}{2^{k+1}},$$

$$D(P_0^{n_k}; P_1^{n_k}) \leqq n_k^{1/2} \varepsilon_{k+1} 2^{-k/2} \leqq \frac{C_3}{5} \varepsilon_k^2 2^{-k}, \tag{4.17}$$

and also for any probability P

$$h^2(P_1;P) \geqq \tfrac{1}{2}h^2(P_0;P) - h^2(P_0;P_1) \geqq \tfrac{1}{2}h^2(P_0;P) - \frac{C_3}{20}\,2^{-k}\,\varepsilon_k^2. \tag{4.18}$$

Now consider the subset of Θ_k formed by the points having their $k-1$ first coordinates fixed, δ_k being free and some estimate $\hat{P}$ depending on n_k observations. We are then in a situation to apply Corollary 4.1 which gives using (4.16) and (4.11)

$$R_B \geqq \tfrac{1}{4} r_k \frac{A_1^2 A_2}{12}\, v_k\, \varepsilon_k^2 [1-(n_k\, v_k\, \varepsilon_k^2)^{1/2}] \geqq 2^{-k}\, 2C_3\, \varepsilon_k^2 [1-(n_k\, A_3\, \varepsilon_k^2\, k^{-1}(\varepsilon_k))^{1/2}].$$

But since $\varepsilon_k \leqq a_{n_k}$ we have $n_k\, \varepsilon_k^2\, k^{-1}(\varepsilon_k) \leqq n_k\, C_4^2\, k^{-1}(a_{n_k})\, a_{n_k}^2 \leqq \frac{1}{4A_3}$ and then there exists one point P_0 in Θ_k with those first $k-1$ coordinates and

$$\mathbb{E}_{P_0^{n_k}}[h^2(\hat{P};P_0)] \geqq 2^{-k}\, C_3\, \varepsilon_k^2.$$

Because of (4.18), (4.17) for any P_1 having the same k first coordinates as P_0

$$\mathbb{E}_{P_1^{n_k}}[h^2(\hat{P};P_1)] \geqq \left[\frac{1}{2}2^{-k}\, C_3\, \varepsilon_k^2 - \frac{C_3}{20}\, 2^{-k}\, \varepsilon_k^2\right] - \frac{C_3}{5}\, \varepsilon_k^2\, 2^{-k} \geqq \frac{C_3}{4}\, 2^{-k}\, \varepsilon_k^2.$$

Now by induction on k, we shall find that for any k there exists $\delta_1, \ldots, \delta_k$ such that any point f_θ in Θ_∞ with those first coordinates will satisfy for all $i \leqq k$,

$$\mathbb{E}_\theta[h^2(f_\theta;\hat{f}_{n_i})] \geqq \frac{C_3}{4}\, 2^{-i}\, \varepsilon_i^2 \geqq \frac{C_3\, C_4^2}{4}\, 2^{-i}\, a_{n_i}^2 \geqq a_{n_i}^2\, c_{n_i}. \tag{4.19}$$

We deduce (4.13) by a simple limiting argument since Θ_∞ is compact and the θ's satisfying (4.19) up to order k form a decreasing sequence of closed subsets of Θ_∞. □

V. Bad Problems

We shall consider here what may happen when the families are not nice. Let us begin with the case considered by Devroye [7] of the densities bounded by 2 on $[0;1]^m$. Actually we may take $m=1$, the extension to $\mathbb{R}^m$ being obvious.

Let us start with some $\mathscr{C}^\infty$ function $\tilde{g}$ in $[0;1]$ with the following properties

$$\begin{gathered} \tilde{g}(x+\tfrac{1}{2}) = -\tilde{g}(\tfrac{1}{2}-x) \quad 0 \leqq x \leqq \tfrac{1}{2} \qquad \text{so that } \int_0^1 \tilde{g}(x)\,dx = 0 \\ |\tilde{g}| \leqq 1 \quad \text{and} \quad \tilde{g}=1 \text{ on } \left[\frac{\varepsilon}{4}; \frac{1}{2}-\frac{\varepsilon}{4}\right]; \qquad \tilde{g}^{(p)}(0)=0 \ \forall p. \end{gathered} \tag{5.1}$$

Now fix an integer r and put $g(x) = \tilde{g}(xr)$; then it is immediate to check that

$$h^2(1+g;1-g) = \alpha, \qquad \frac{1-\varepsilon}{r} \leqq \alpha \leqq \frac{1}{r}; \qquad d_q^q(1+g;1-g) \geqq 2^q\,\frac{1-\varepsilon}{r}.$$

Let us consider the family $\Theta_{\mathscr{C}}=\left\{1+\sum_{i=1}^{r}\delta_i g\left(x-\frac{i-1}{r}\right)\right\}$, $\delta_i=\pm 1$. Applying Assouad's Lemma we get from Corollary 4.1,

$$R_M(h^2,n)\geqq\frac{1-\varepsilon}{4}(1-(2nr^{-1})^{1/2});\qquad R_M(d_q^q,n)\geqq(1-\varepsilon)(1-(2nr^{-1})^{1/2}).$$

Letting $r\to+\infty$ and $\varepsilon\to 0$ we find that for the set Θ of $\mathscr{C}^\infty$ functions on $[0;1]$ bounded by two

$$R_M(h^2,n)\geqq\tfrac{1}{4}\quad\text{and}\quad R_M(d_q^q,n)\geqq 1.$$

To get Theorem 1, i) of [7] we just need a proper normalization by the $\mathbb{L}^q$ norm of the true density. But in $\Theta_{\mathscr{C}}$, all elements have the same $\mathbb{L}^q$ norm, which is approximately 2^{q-1}; dividing by this quantity we find for the normalized risk $R'_M(d_q^q)\geqq 2^{-q+1}$.

Let us now come back to any smooth family but just assume that its elements may have arbitrary compact support (i.e. CS is no more true) and consider the following transformation Λ from $\mathscr{C}^\infty[0;1]$ to $\mathscr{C}^\infty(\mathbb{R})$ given by $\Lambda f(x)=\lambda f(\lambda x)$, $\lambda>0$. It transforms a density into a density and has the following property that

$$h^2(\Lambda f,\Lambda g)=h^2(f;g);\qquad d_q^q(\Lambda f;\Lambda g)=\lambda^{q-1}d_q^q(f;g);\qquad \frac{d_q^q(\Lambda f;\Lambda g)}{\|\Lambda f\|_q^q}=\frac{d_q^q(f;g)}{\|f\|_q^q}.$$

Because of this property the risks $R_M(h^2,n)$ or $R'_M(d_q^q,h)$ will remain constant if we transform $\Theta_{\mathscr{C}}$ using Λ. But for given r, even very large, there exists some $\lambda\leqq 1$ such that $\Lambda(\Theta_{\mathscr{C}})$ will become arbitrarily smooth with compact support because all elements of $\Theta_{\mathscr{C}}$ have their p^{th} derivative uniformly bounded. That means that any family of arbitrary smooth densities with compact (not the same) support contains some $\Lambda\Theta_{\mathscr{C}}$ and that for those families

$$R_M(h^2,n)\geqq\tfrac{1}{4};\qquad R'_M(d_q^q,n)\geqq 2^{-q+1}.$$

Remark. This is the analogue of Devroye's result for the family $\mathscr{L}(g)$. We could get exactly the same result as his using a modified version of the above proof and applying case ii) of Corollary 4.1.

We shall now improve the second part of Devroye's theorem. We shall need some preliminary considerations and first recall this known lemma:

Lemma 5.1. *Assume that $\Theta_{\mathscr{C}}$ is in one-to-one correspondence with some hypercube of dimension r and that d is a distance such that*

$$d^q(P_x;P_{x'})\geqq\alpha\Delta(x,x').$$

For any λ with $0<\lambda<1$ there exists a subset S of $\Theta_{\mathscr{C}}$ such that

$$\operatorname{Card}S\geqq\exp\left[\frac{r\lambda^2}{2}\right];\qquad d^q(P;Q)>\alpha r\frac{1-\lambda}{2}\qquad\forall P,Q\in S,\ P\neq Q.$$

Remark. In conjunction with Fano's Lemma this leads to an equivalent of Assouad's Lemma which was used in [2, 3] or [8] to get lower bounds and could have been used here, but the constants would not have been the same and Assouad's Lemma is simpler.

First we must describe the families of densities in $\mathscr{C}^\infty[0;1]$ that we shall use. Let us come back to the function $\tilde{g}$ of (5.1) and consider two sequences $\{a_i\}$, $\{r_i\}$ for $i \geqq 1$ with

$$a_i > 0; \quad \sum_{i=1}^{+\infty} a_i = 1; \quad r_i \to +\infty \quad \text{when} \quad i \to +\infty,\ r_i \in \mathbb{N}.$$

Define

$$A_0 = 0; \quad A_i = \sum_{j=1}^{i} a_j \quad \text{if } i \geqq 1; \quad \lambda_i = r_i^{-1} a_i; \quad I_i = [A_{i-1}; A_i];$$

$$f_{ij}(x) = (1-\varepsilon^2)^{1/2} \left[\tilde{g}\left(\frac{x - A_{i-1} - (j-1)\lambda_i}{\lambda_i} \right) \right] \quad i \geqq 1;\ j = 1, 2, \ldots, r_i$$

and consider

$$\Theta'_1 = \left\{ 1 + \sum_{i=1}^{+\infty} \left[\sum_{j=1}^{r_i} \delta_{ij} f_{ij}(x) \right] \right\}, \qquad \delta_{ij} = \pm 1.$$

Θ'_1 is a family of $\mathscr{C}^\infty$ densities on $[0;1]$ bounded by 2. Easy computations with $K = 2\log(2\varepsilon^{-1})$ give

$$K(1+\delta_{ij} f_{ij}; 1-\delta_{ij} f_{ij}) \leqq K\lambda_i; \qquad h^2(1+f_{ij}; 1-f_{ij}) \leqq \lambda_i, \tag{5.2}$$

$$h^2(1+f_{ij}; 1-f_{ij}) \geqq \lambda_i (1-\varepsilon)^2; \qquad d_q^q(1+f_{ij}; 1-f_{ij}) \geqq 2^q \lambda_i (1-\varepsilon)(1-\varepsilon^2)^{q/2}. \tag{5.3}$$

From now on we shall restrict our discussion to the case of h, the other one being similar up to multiplication by 2^q which cancels in (5.11). Θ'_1 may be considered as a product of hypercubes of dimension r_i: $\Theta'_1 = \prod_{i=1}^{\infty} \mathscr{C}'_i$. Applying Lemma 5.1 to $\mathscr{C}'_i$ with $\lambda = 10^{-1/2}$ and d replaced by h restricted to the set I_i, we shall find a subset $\mathscr{C}_i$ of $\mathscr{C}'_i$, with cardinal larger than $\exp\left[\frac{r_i}{20}\right]$, having the following property: for any two points x, x' in $\mathscr{C}_i$, (using (5.3)) and suitable choice of ε,

$$h^2(P_x 1_{I_i}; P_{x'} 1_{I_i}) \geqq \lambda_i r_i (1-\varepsilon)^2 \, \frac{1-10^{-1/2}}{2} \geqq \frac{a_i}{3}. \tag{5.4}$$

Since the set $\mathscr{C}_i$ will only depend on the part of the measures supported by I_i, i.e. on coordinate i, the subset $\Theta_1 = \prod_{i=1}^{+\infty} \mathscr{C}_i$ will have the following property that if θ and θ' are two points in Θ_1, $\theta = \{\theta_i\}_{i \geqq 1}$, $\theta_i \in \mathscr{C}_i$ and $\Delta(\theta_i, \theta'_i) = 0$ if $\theta_i = \theta'_i$, 1 in the other case:

$$h^2(\theta, \theta') = \sum_{i=1}^{+\infty} \Delta(\theta_i; \theta'_i)\, h_i^2; \qquad \frac{a_i}{3} \leqq h_i^2 \leqq a_i, \tag{5.5}$$

$$K(\theta, \theta') \leqq K \sum_{i=1}^{+\infty} \Delta(\theta_i; \theta'_i)\, a_i \tag{5.6}$$

because of (5.2) and (5.4). Moreover we may assume that

$$\exp\left(\frac{r_i}{20}\right) \leqq \operatorname{Card} \mathscr{C}_i = \mathrm{U}_i \tag{5.7}$$

and notice that Θ_1 is totally bounded and closed, hence compact by (5.5).

Given a subset T_j of Θ_1 of the following form

$$T_j = \{\theta_1\} \times \ldots \times \{\theta_{j-1}\} \times S_j \times \{\theta_{j+1}) \times \ldots, \qquad S_j \subset \mathscr{C}_j \tag{5.8}$$

with $\operatorname{Card} T_j = \operatorname{Card} S_j = V_j \leqq U_j$ and an estimate $\hat{\theta}_j$ of θ_j, with values in S_j, depending on n observations, we may apply Fano's Lemma and using (5.6) find that if

$$n K a_j + \log 2 \leqq \tfrac{1}{13} \log(V_j - 1) \tag{5.9}$$

then $V_j^{-1} \sum_{\theta_j \in S_j} P_{\theta_j}[\hat{\theta}_j \neq \theta_j] \geqq \frac{12}{13}$, which implies because of (5.5)

$$V_j^{-1} \sum_{\theta_j \in S_j} \mathbb{E}_{\theta_j}[h^2(P_{\theta_j}; P_{\hat{\theta}_j})] \geqq \frac{4a_j}{13}.$$

By (5.5) $h^2(P_{\theta_j}; P_{\hat{\theta}_j})$ is smaller that a_j; then there exists a subset S'_j of S_j with $\operatorname{Card} S'_j \geqq \frac{2}{11} V_j$ and

$$\inf_{\theta_j \in S'_j} \mathbb{E}_{\theta_j}[h^2(P_{\theta_j}; P_{\hat{\theta}_j})] \geqq \frac{2a_j}{13}. \tag{5.10}$$

Given an arbitrary probability P let us call $\tilde{\theta}_j$ any point in S_j which minimizes $h(P 1_{I_j}; P_{\theta_j} 1_{I_j})$. It is easy to see that we always have (with θ, $\tilde{\theta}$ in T_j)

$$h^2(P; P_\theta) \geqq \tfrac{1}{4} h^2(P_{\tilde{\theta}}; P_\theta).$$

This remark and (5.10) give the following lemma:

Lemma 5.2. *Given a subset T_j of Θ_1 defined by* (5.8) *with* $\operatorname{Card} S_j = V_j$ *satisfying* (5.9) *and an estimate $\hat{P}$ depending on n observations, there exists a subset T'_j of T_j with*

$$\inf_{\theta \in T'_j} \mathbb{E}_\theta[h^2(P_\theta; \hat{P})] \geqq \frac{a_j}{26}; \qquad \operatorname{Card} T'_j \geqq \frac{2}{11} V_j. \tag{5.11}$$

$$\left(\mathit{or}\ \inf_{\theta \in T'_j} \mathbb{E}_\theta[d_q^q(P_\theta; \hat{P})] \geqq \frac{2a_j}{13}\right).$$

We are now in a position to prove our theorem. Let any sequence $\{c_i\}_{i \geqq 1}$ converging to zero be given such that $\frac{1}{78} < \sup_i c_i < \frac{1}{26}$. Fix $a_1 = 26 \sup_i c_i$ and $a_j = 2^{1-j}(1 - a_1)$ for $j \geqq 2$. Then $\{a_j\}_{j \geqq 1}$ is a strictly decreasing sequence. Define $m_0 = 0$ and

$$m_j = \left[m_{j-1} \vee \sup\left\{ i \mid c_i > \frac{a_j}{26} \right\} \right] + 1, \qquad j \geqq 1.$$

Then $m_1=1$ and $\{m_j\}_{j\geqq 1}$ is a strictly increasing sequence such that $i\geqq m_j$ implies $c_i\leqq\frac{a_j}{26}$. For convenience let us set the following notations:

$$M_j=m_{j+1}-m_j;\quad m_j'=m_{j+1}-1;\quad B_j=(2/11)^{M_j};$$
$$C_j=(2/11)^{-m_j'+1};\quad D_j=1+2^{13}\exp[13Km_j'a_j].$$

Notice that $C_1=B_1^{-1}$ and $C_{j+1}=C_jB_{j+1}^{-1}$. We may now choose the r_j's inductively in order that the U_i satisfy (5.7) together with

$$U_1\geqq C_1D_1;\quad U_j\geqq C_jD_j[\prod_{i\leqq j-1}U_i^{j-i-1}],\quad j\geqq 2; \tag{5.12}$$

(in the case of d_q just replace in those formulas 26 by 6.5). We shall first prove the following

Lemma 5.3. *For any T_j of the form* (5.8) *such that*

$$m_j'Ka_j+\log 2\leqq\tfrac{1}{13}\log[B_jV_j-1] \tag{5.13}$$

and any sequence of M_j estimates $\{\hat{P}_i\,|\,i=m_j;\,m_j+1;\,\ldots;\,m_j'\}$ such that $\hat{P}_i$ is a function of i variables there exists a subset T_j'' of T_j with

$$\inf_{\theta\in T_j''}\ \inf_{i=m_j,\ldots,m_j'}\ c_i^{-1}\,\mathbb{E}_\theta[h^2(\hat{P}_i;P_\theta)]\geqq 1;\qquad \operatorname{Card}T_j''\geqq B_jV_j.$$

Proof. Assuming that (5.9) is true with $n=m_j$ we may use Lemma 5.2 to get some T_j', $\operatorname{Card}T_j'\geqq\frac{2}{11}V_j$ and satisfying (5.11) which gives us since $\frac{a_j}{26}\geqq c_i$ for $i\geqq m_j$

$$\inf_{\theta\in T_j'}c_{m_j}^{-1}\,\mathbb{E}_\theta[h^2(\hat{P}_{m_j};P_\theta)]\geqq 1.$$

But we may now apply the same lemma to T_j' if (5.9) is still valid with $n=m_j+1$ and V_j is replaced by $\frac{2}{11}V_j$, and so on. The cardinal of the last set T_j'' will be at least B_jV_j. Since during this process the right member of (5.9) is decreasing and the left one is increasing, all conditions will be true if the last one is true which amounts to (5.13). □

Suppose we are given a sequence $\{\hat{P}_n\}_{n\geqq 1}$ of estimates, $\hat{P}_n$ depending on n variables. In order to apply Lemma 5.3, we may notice that (5.13) is equivalent to $B_jV_j\geqq D_j$ which will always be satisfied because of (5.12) if we have

$$V_1=U_1;\quad V_j\geqq C_{j-1}^{-1}[\prod_{i\leqq j-1}U_i^{j-i-1}]^{-1}U_j,\quad j\geqq 2. \tag{5.14}$$

We start the construction with T_1 given by (5.8), and $S_1=\mathscr{C}_1$ so that $V_1=U_1$ and we may apply Lemma 5.3 getting some T_1'' of the form

$$T_1''=S_1''\times\{\theta_2\}\times\ldots;\qquad \operatorname{Card}S_1''\geqq B_1V_1,$$

where S_1'' obviously depends on $\theta_2,\theta_3,\ldots$. Keeping $\theta_3,\theta_4,\ldots$ fixed we shall now vary θ_2 and write $S_1''(\theta_2)$. When θ_2 varies in $\mathscr{C}_2$ we shall consider

$\sum_{\theta_2 \in \mathscr{C}_2} \operatorname{Card} S_1''(\theta_2)$. Since $S_1''(\theta_2)$ is a subset of $\mathscr{C}_1$ we may see that if each θ_1 in $\mathscr{C}_1$ belongs to no more than V_2 distinct $S_1''(\theta_2)$ then

$$V_2 U_1 \geqq \sum_{\theta_2 \in \mathscr{C}_2} \operatorname{Card} S_1''(\theta_2) \geqq U_2 B_1 U_1 \tag{5.15}$$

which implies that $V_2 \geqq B_1 U_2$ and proves the existence of some S_2 with $\operatorname{Card} S_2 = V_2 \geqq B_1 U_2 = U_2 C_1^{-1}$ and some $\theta_1 \in \bigcap_{\theta_2 \in S_2} S_1''(\theta_2)$ such that with $T_2 = \{\theta_1\} \times S_2 \times \{\theta_3\} \times \dots$,

$$\inf_{\theta \in T_2} \inf_{i=1,\dots,m_1'} c_i^{-1} \mathbb{E}_\theta[h^2(\hat{P}_i; P_\theta)] \geqq 1.$$

Let us now proceed by induction. Assume that for any sequence $\theta_{j+1}, \theta_{j+2,\dots}$ there exists $\theta_1, \theta_2, \dots, \theta_{j-1}$ and some subset S_j of $\mathscr{C}_j$ such that (with T_j as in (5.8))

$$\inf_{\theta \in T_j} \inf_{i=1,\dots,m_{j-1}'} c_i^{-1} \mathbb{E}_\theta[h^2(\hat{P}_i; P_\theta)] \geqq 1,$$

$$\operatorname{Card} S_j = V_j \geqq U_j C_{j-1}^{-1} [\prod_{i \leqq j-1} U_i^{j-i-1}]^{-1}.$$

Letting all θ_i fixed except for θ_j, we may apply Lemma 5.3 to T_j since (5.14) is satisfied and for some T_j'' and S_j'' with $\operatorname{Card} S_j'' = V_j' \geqq B_j V_j$

$$\inf_{\theta \in T_j''} \inf_{i=1,\dots,m_j'} c_i^{-1} \mathbb{E}_\theta[h^2(\hat{P}_i; P_\theta)] \geqq 1.$$

To make the induction work we keep $\theta_{j+2}, \dots$ fixed and let θ_{j+1} change; then $\theta_1, \dots, \theta_{j-1}$ and S_j'' are functions of θ_{j+1}. But the possible choices for $\theta_1, \dots, \theta_{j-1}$ are at most $\prod_{i=1}^{j-1} U_i$. Since because of (5.12) $U_{j+1} [\prod_{i \leqq j-1} U_i]^{-1} \geqq 1$, we may restrict θ_{j+1} to vary within a subset S_{j+1}' of $\mathscr{C}_{j+1}$ of cardinal $U_{j+1}' \geqq U_{j+1} [\prod_{i \leqq j-1} U_i]^{-1}$ and keep $\theta_1, \dots, \theta_{j-1}$ fixed for all the θ_{j+1} in S_{j+1}'. Now S_j'' is still depending on θ_{j+1} but we may also fix it using the same argument that we used for (5.15) and get some fixed θ_j for a subset S_{j+1} of S_{j+1}' of cardinal V_{j+1} and

$$V_{j+1} U_j \geqq \sum_{\theta_{j+1} \in S_{j+1}'} \operatorname{Card} S_j''(\theta_{j+1}) \geqq U_{j+1}' V_j'$$

which implies

$$V_{j+1} \geqq \frac{U_{j+1}' V_j'}{U_j} \geqq B_j \frac{V_j}{U_j} [\prod_{i \leqq j-1} U_i]^{-1} U_{j+1}. \tag{5.16}$$

Using these $\theta_1, \dots, \theta_{j-1}, \theta_j, S_{j+1}$, to define T_{j+1}, we get

$$\inf_{\theta \in T_{j+1}} \inf_{i=1,\dots,m_j'} c_i^{-1} \mathbb{E}_\theta[h^2(\hat{P}_i; P_\theta)] \geqq 1$$

and from (5.12) and (5.16)

$$V_{j+1} \geqq \frac{U_{j+1}}{C_j} [\prod_{i \leqq j} U_i^{j-i}]^{-1},$$

which is our induction assumption. We have then proved the following theorem in case i) for $n < +\infty$. Case ii) is absolutely analogous because all points in Θ_1 have the same $\mathbb{L}^q$ norm and this norm is approximately 2^{q-1} when ε is small. The case $n = +\infty$ follows from the fact that the subset of the θ's satisfying (5.17) for some n is closed then compact, non-void and decreasing with n.

Theorem 5.4. *For any given sequence $\{c_n\}_{n \geqq 1}$ converging to zero such that*

i) $\frac{1}{78} < \sup_i c_i < \frac{1}{26}$ *if* $d = h$, ii) $\frac{2}{39} < \sup_i c_i < \frac{2}{13}$ *if* $d = d_q$

there exists a compact subset Θ of densities in $\mathscr{C}^\infty[0; 1]$ bounded by two and for any sequence $\{\hat{P}_n\}_{n \geqq 1}$ of estimates depending on n observations there exist infinite subsets Θ'_n with

$$\inf_{\theta \in \Theta'_n} \inf_{k < n} c_k^{-1} \mathbb{E}_\theta[h^2(\hat{P}_k; P_\theta)] \geqq 1,$$

or

$$\inf_{\theta \in \Theta'_n} \inf_{k < n} c_k^{-1} \mathbb{E}_\theta[d_q^q(\hat{P}_k; P_\theta)] \geqq 1, \tag{5.17}$$

or

$$\inf_{\theta \in \Theta'_n} \inf_{k < n} c_k^{-1} \mathbb{E}_\theta\left[\frac{d_q^q(\hat{P}_k; P_\theta)}{\|P_\theta\|_q^q}\right] \geqq 2^{1-q}$$

and there is at least one point in Θ satisfying (5.17) *with* $n = +\infty$.

Remarks. 1) The constants which we used here ($\frac{1}{26}$, etc. ...) are not optimal at all and were chosen for convenience. The same proof with longer computations could lead to $\frac{1}{8}$ which is still not optimal.

2) The same method also gives a similar result for the second family $\mathscr{L}(g)$ considered by Devroye for d_1 or h (not for d_q if $q > 1$). We just need minor modifications and the transformation Λ mentioned at the beginning of the section. We should use a family $\{\Lambda_i\}_{i \geqq 1}$ of such applications with different values of the λ_i's: $\lambda_i = r_i^{-1} a_i$ and apply Λ_i to the parts of the measures which have a support on I_i. This will transform I_i into an interval of length r_i and each f_{ij} will have a support of length 1. The details are easy. It does not work for $q > 1$ just as in Devroye's paper because there is no longer a suitable normalization by the $\mathbb{L}^q$ norm.

Je remercie tout particulièrement Lucien Le Cam et Charles Stone pour leurs nombreux commentaires et suggestions.

References

1. Assouad, P.: Deux remarques sur l'estimation. C.R. Acad. Sci. Paris **296**, Sér. I, 1021–1024 (1983)
2. Birgé, L.: Thèse, 3^e-partie. Université Paris VII (1980)
3. Birgé, L.: Approximation dans les espaces métriques et théorie de l'estimation. Z. Wahrscheinlichkeitstheor. Verw. Geb. **65**, 181–237 (1983)
4. Birgé, L.: Non-asymptotic minimax risk for Hellinger balls. Probability and Math. Statistics. To appear

5. Bretagnolle, J., Huber, C.: Estimation des densités: risque minimax. Z. Wahrscheinlichkeitstheor. Verw. Geb. **47**, 119–137 (1979)
6. Dacunha-Castelle, D.: École d'Eté de Probabilités de Saint-Flour VII. Lecture Notes in Mathematics no. **678**. Berlin-Heidelberg-New York: Springer 1977
7. Devroye, L.: On arbitrarily slow rates of convergence in density estimation. Z. Wahrscheinlichkeitstheor. Verw. Gebiete **62**, 475–483 (1983)
8. Ibragimov, I.A., Khas'minskii, R.Z.: Estimation of distribution density. Zap. Nauchn. Semin. LOMI **98**, 61–85 (1980) in russian, and J. Sov. Math. **21**, 40–57 (1983)
9. Ibragimov, I.A., Khas'minskii, R.Z.: On the non-parametric density function. Zap. Nauchn. Semin. LOMI **108**, 73–89, in russian (1981)
10. Ibragimov, I.A., Khas'minskii, R.Z.: Statistical Estimation, Asymptotic Theory. Berlin-Heidelberg-New York: Springer 1981
11. Khas'minskii, R.Z.: A lower bound on the risks of non-parametric estimates of densities in the uniform metric. Theory Probab. Appl. **23**, 794–796 (1978)
12. Kolmogorov, A.N., Tikhomirov, V.M.: ε-entropy and ε-capacity of sets in function spaces. Am. Math. Soc. Transl. (2) **17**, 277–364 (1961)
13. Le Cam, L.: Convergence of estimates under dimensionality restrictions. Ann. Statist. **1**, 38–53 (1973)
14. Le Cam, L.: Asymptotic methods in statistical decision theory. To be published
15. Lorentz, G.G.: Metric entropy and approximation. Bull. Amer. Math. Soc. **72**, 903–937 (1966)
16. Lorentz, G.G.: Approximation of Functions. New York: Holt, Rinehart, Winston 1966
17. Müller, H.G., Gasser, T.: Optimal convergence properties of kernel estimates of derivatives of a density function. In: Smoothing Techniques for Curve Estimation (T. Gasser and M. Rosenblatt eds.). Lecture Notes in Mathematics no. **757**, pp. 144–154. Berlin-Heidelberg-New York: Springer 1979
18. Stone, C.J.: Optimal rates of convergence for nonparametric estimators. Ann. Statist. **8**, 1348–1360 (1980)
19. Stone, C.J.: Optimal uniform rate of convergence for nonparametric estimators of a density function or its derivatives. University of California Berkeley: preprint (1983)

Received June 15, 1983; in revised form May 28, 1985