On the construction of estimates for independent observations

By L. LeCam

University of California, Berkeley

1. Introduction. Let Θ and L be two sets. For each $\ell \in L$, let $\mathcal{A}_\ell$ be a σ-field carried by a set $\mathcal{X}_\ell$. Let $(\mathcal{X}, \mathcal{A})$ be the product of the spaces $(\mathcal{X}_\ell, \mathcal{A}_\ell)$. Suppose given for each $\theta \in \Theta$ and $\ell \in L$ a probability measure $p_{\theta,\ell}$ on $\mathcal{A}_\ell$ and let P_θ be the corresponding product measure on $\mathcal{A}$.

For each $\ell \in L$ and for every pair (s,t) of elements of Θ, let $h_\ell(s,t)$ be the Hellinger distance defined by

$$h_\ell^2(s,t) = \tfrac{1}{2} \int \left(\sqrt{dp_{s,\ell}} - \sqrt{dp_{t,\ell}} \right)^2 .$$

Let $H^2(s,t) = \Sigma_\ell\, h_\ell^2(s,t)$.

In a previous paper [1], this author considered the case where L is finite and where the systems $(\mathcal{X}_\ell, \mathcal{A}_\ell, p_{\theta,\ell})$ are replicates of each other. Using a definition of dimension of the Kolmogorov type, it was then claimed that there is a universal constant K such that, when Θ metrized by H

This research was partially supported by National Science Foundation, Grant No. GP-31091X2.

has dimension at most D, one can find estimates $\hat{\theta}$ for which

$$E_\theta H^2(\hat{\theta},\theta) \le K D .$$

In the present paper, we propose to show that an analogous result still holds for independent observations which are not necessarily identically distributed. However, in this more general case, we were able to obtain only the existence of estimates $\hat{\theta}$ for which

$$E_\theta H^2(\hat{\theta},\theta) \le K_1 D \log(1+D) .$$

The method of proof is generally the same as suggested by [1] and reproduced in [2], but with two important modifications. The "proof" given in [1] incorporates a flaw which would allow excessive freedom in the selection of successive tests, and retain probabilistic statements even when the sample points are fixed. As a result the computation of bounds on the probabilities of errors does not take into account all the possibilities. The intention was honorable, but the execution deficient, as will be spelled out in more detail below.

2. Description of the construction scheme.

In this section, it will be assumed that Θ is a finite set.

Let P_θ be the probability measure corresponding to θ and let W be a real valued function defined on $\Theta \times \Theta$ and such that $W(s,t) = W(t,s)$ for all pairs $(s,t) \in \Theta \times \Theta$. The usual applications will be to cases where W is a metric on Θ, but there are enough other interesting cases to warrant consideration of the more general situation.

Any subset $A \subset \Theta$ will be assigned a number diam (A) called its "diameter" and equal to

$$\text{diam}(A) = \sup\{W(s,t);\ s \in A,\ t \in A\} .$$

Consider a sequence $\{b_\nu ;\ \nu = 0, 1, 2, \cdots\}$ of positive numbers such that $b_\nu > b_{\nu+1}$ and such that b_0 is at least equal to the diameter of Θ. Let $\{a_\nu\}$ be another sequence such that $0 \leq a_\nu < b_\nu$. For each ν, let $\{A_{\nu,i};\ i \in I_\nu\}$ be a partition of Θ by sets $A_{\nu,i}$ whose diameter does not exceed a_ν.

<u>It will be assumed all the way through this section that the partition</u> $\{A_{\nu,i};\ i \in I_\nu\}$ <u>has minimum cardinality among all those subject to the restriction that</u> $\text{diam}\, A_{\nu,i} \leq a_\nu$.

Such minimal partitions certainly exist.

Say that a pair of elements of I_ν is a ν-distant pair if there are elements $s \in A_{\nu,i}$ and $t \in A_{\nu,j}$ such that $W(s,t) \geq b_\nu$. For such a pair, let $\varphi_{\nu,i,j}$ be a test of $A_{\nu,i}$ against $A_{\nu,j}$.

<u>It will be assumed that the tests are selected so that</u>

$$\varphi_{\nu,i,j} = \varphi^2_{\nu,i,j} = 1 - \varphi_{\nu,j,i} .$$

Construct sets B_ν , $\nu = 0, 1, 2, \cdots$ as follows. Start with $B_0 = \Theta$. If $B_{\nu-1}$ has been constructed, let $J_\nu \subset I_\nu$ be the set of indices $i \in I_\nu$ such that $A_{\nu,i}$ intersects $B_{\nu-1}$. Further, if $i \in J_\nu$, let $J_\nu(i)$ be the set of $j \in J_\nu$ which are ν-distant from i . Define $\psi_{\nu,i}$ by

$$\psi_{\nu,i} = \inf_j \{\varphi_{\nu,i,j} ; j \in J_\nu(i)\} .$$

For each sample point x , there is a certain set, say $S_\nu(x)$ of indices $i \in J_\nu$ for which $\psi_{\nu,i}(x) = 1$. Let B_ν be the set

$$B_\nu = \bigcup_i \{A_{\nu,i} \cap B_{\nu-1} ; i \in S_\nu(x)\} .$$

By construction one has $B_\nu \subset B_{\nu-1}$. Also, the diameter of B_ν satisfies the inequality diam $B_\nu < b_\nu$ since each $A_{\nu,i}$ has diameter at most $a_\nu < b_\nu$ and since for any ν-distant pair (i,j) one has $\psi_{\nu,i} \psi_{\nu,j} = 0$.

The construction just described is the one that we meant to use in [1]. However, the paper in question would allow fixing the covers $\{A_{\nu,i}\}$ after $B_{\nu-1}$ is already obtained.

This may lead to unfortunate consequences.

In order to perform the indicated construction one must select the tests $\psi_{\nu,i,j}$. In this respect we shall use the following notation. For any two sets $A_i \subset \Theta$, let $\pi(A_1,A_2)$ be the number $\pi(A_1,A_2) = \inf_{\varphi} \sup_{s,t} \{ \int (1-\varphi)\, dP_s + \int \varphi\, dP_t \}$ where the supremum is taken for $s \in A_1$ and $t \in A_2$ and the infimum is taken over all measurable φ such that $0 \le \varphi \le 1$.

For a pair $(A_{\nu,i}, A_{\nu,j})$ there exist tests $\varphi_{\nu,i,j} = \varphi^2_{\nu,i,j}$ such that

$$\pi(A_{\nu,i}, A_{\nu,j}; \varphi_{\nu,i,j})$$

$$= \sup\{ \int (1-\varphi_{\nu,i,j})dP_s + \int \varphi_{\nu,i,j}dP_t \,;\, s \in A_{\nu,i}, t \in A_{\nu,j}\}$$

does not exceed $2\pi(A_{\nu,i}, A_{\nu,j})$.

For each $\theta \in \Theta$ and each ν , define $U_\nu(\theta)$ to be the set of points t such that there exists $s \in \Theta$ for which $W(\theta,s) < b_{\nu-1}$ and $W(s,t) \le a_\nu$. If W is a metric, this is contained in the open ball of radius $b_{\nu-1} + a_\nu$.

Assuming that the partitions $\{A_{\nu,i}\}$ and the tests $\varphi_{\nu,i,j}$ are selected as described, one can assert the following.

Proposition 1. Assume that there is a number $\beta(\nu)$ such that $\pi(A_{\nu,i}, A_{\nu,j}) \le \beta(\nu)$ for all pairs (i,j) which are ν-distant.

<u>Assume also that there is a number</u> $C(\nu)$ <u>such that every set of the form</u> $U_\nu(\theta)$ <u>can be covered by at most</u> $C(\nu)$ <u>sets of diameter</u> a_ν <u>or less. Then, for every</u> $k \geq 1$ <u>one has</u>

$$P_\theta\{\theta \notin B_k\} \leq 2 \sum_{\nu=1}^{k} \beta(\nu)\, C(\nu) .$$

<u>Proof</u>. Fix a particular $\theta \in \Theta$. For each ν, let J_ν^* be the set of indices $j \in I_\nu$ such that $A_{\nu,j}$ intersects the set $V_\nu(\theta) = \{t \in \Theta;\ W(\theta,t) < b_{\nu-1}\}$. If $i \in J_\nu^*(i)$ be the set of indices $j \in J_\nu^*$ for which (i,j) is a ν-distant pair. Denote $A_{\nu,o}$ that element of the partition $\{A_{\nu,i}\,;\, i \in I_\nu\}$ which contains θ and let

$$\psi_\nu^* = \inf_j \{\varphi_{\nu,0,j}\ ;\ j \in J_\nu^*(0)\} .$$

Whenever $\theta \in B_{\nu-1}$ one has $B_{\nu-1} \subset V_\nu(\theta)$ and therefore $\psi_\nu^* \leq \psi_{\nu,0}$. Thus, if χ_ν is the indicator of the set $\theta \in B_\nu$ one may write

$$\chi_{\nu-1}(1-\chi_\nu) \leq \chi_{\nu-1}(1-\psi_\nu^*) .$$

Since $B_i \supset B_{i+1}$ for all i, one has

$$(1-\chi_k) = (1-\chi_1) + \chi_1(1-\chi_2) + \cdots + \chi_{k-1}(1-\chi_k) .$$

This certainly implies

$$P_\theta\{\theta \notin B_k\} \leq \sum_{\nu=1}^{k} E_\theta(1-\psi_\nu^*) .$$

The expectation $E_\theta(1-\psi_\nu^*)$ does not exceed $2\,\beta(\nu)N_\nu$, where

N_ν is the number of sets $A_{\nu,i}$ which intersect the "ball" $V_\nu(\theta)$.

Suppose that $N_\nu \geq C(\nu) + 1$. Then one could cover the set $U_\nu(\theta)$ by a certain number $N'_\nu < N_\nu$ of sets $A'_{\nu,i}$ such that diam $A'_{\nu,i} \leq a_\nu$. Taking these, and the $A_{\nu,j}$ which do not intersect $V_\nu(\theta)$ would produce a new cover of cardinality strictly inferior to that of the partition $\{A_{\nu,i} ; i \in I_\nu\}$. The cover can then be converted to a partition. This would contradict the minimality property of $\{A_{\nu,i}\}$. The proof of the proposition is therefore complete.

The construction in question can be used to give estimates $\hat{\theta}$ of θ . For this, just select any integer m and a point $\hat{\theta} \in B_m$, or when B_m is empty, in the last B_j , $j \leq m$ which is nonempty.

With this definition one can also state the following corollary.

<u>Corollary</u>. <u>Let</u> g <u>be any monotone increasing function from</u> $[0,\infty)$ <u>to</u> $[0,\infty)$. <u>The estimate</u> $\hat{\theta}$ <u>defined here satisfies the inequality</u>

$$E_\theta \, g[W(\theta,\hat{\theta})] \leq g(b_m) + 2 \sum_{0 \leq \nu \leq m-1} g(b_\nu) C(\nu+1) \beta(\nu+1) .$$

<u>Proof</u>. Let $\alpha_k = P_\theta\{\theta \in B_k \cap B_{k+1}^c\}$.

Then

$$E_\theta\, g[W(\theta,\hat{\theta})] \leq g(b_m)\; P[\theta \in B_m] + \alpha_{m-1}\, g(b_{m-1}) + \cdots$$

$$+ \alpha_k\, g(b_k + \cdots + \alpha_0\, g(b_0) \; .$$

Since $$\alpha_0 + \cdots + \alpha_k \leq 2 \sum_{\nu=0}^{k} C(\nu+1)\; \beta(\nu+1)$$

and since the $g(b_k)$ form a decreasing sequence, the sum on the right side is smaller than

$$g(b_m) + 2 \sum_{0\leq\nu\leq m-1} g(b_\nu)\; C(\nu+1)\, \beta(\nu+1)$$ as claimed .

The construction described here can be carried out more directly. For each $A_{\nu,i}$ one can select a test $\psi_{\nu,i}$ of $A_{\nu,i}$ against the union of those $A_{\nu,j}$ which are ν-distant but intersect $B_{\nu-1}$. This is obviously possible, and will give better bounds in very many cases. However, in general, we do not know how to evaluate the probabilities of error.

3. Application to independent observations.

In this section we return to the situation described in the Introduction, with a system $(\mathcal{X}, \mathcal{A}, P_\theta)$ which is the direct product of components $(\mathcal{X}_\ell, \mathcal{A}_\ell, P_{\theta,\ell})$, $\ell \in L$.

Assume first that Θ is finite and that the sum H^2 is bounded on $\Theta \times \Theta$. Then, in order to evaluate the performance of the construction described in Section 2, with $W = H$ it will be sufficient to obtain

i) a bound on the number of sets of diameter a_ν necessary to cover a ball of radius $b_{\nu-1} + a_\nu$ and ii) a bound on the probabilities $\pi(A_{\nu,i}, A_{\nu,j})$. The latter will result from an application of inequalities of Bernstein type. For the former we shall use a dimensionality restriction as follows.

Assumption 1. *There is a number* $C = 2^D$ *such that every subset of* Θ *of* H-*diameter* $2d$, $d \geq 1/125$, *can be covered with no more than* C *sets of diameter at most* d.

For simplicity we shall also assume that $D \geq 1$.

The number $D = \log C$ is some sort of evaluation of the dimensionality of Θ. In R^k, with the maximum coordinate norm, cubes of diameter d can be covered with exactly 2^k cubes of diameter $d/2$. For other relations with dimension see Kolmogorov [3].

Suppose first that the numbers b_ν are given and take a number $c \in (1,\infty)$. Take two sets $A_{\nu,i}$ and $A_{\nu,j}$ and points $s \in A_{\nu,i}$, $t \in A_{\nu,j}$ such that $H^2(s,t) \geq b_\nu^2$.

Writing $h_j = h_j(s,t)$ for simplicity, let L_k, $k = 1, 2, \cdots n+1$ be a partition of L such that i) for each $k = 1, 2, \cdots, n$ one has $S_k^2 = \Sigma_{\ell \in L_k} h_\ell^2 \geq c$ and ii) the cardinality n is the maximum possible under this restriction.

For each k, let $P_{\theta,k} = \prod_{\ell \in L_k} p_{\theta,\ell}$. By construction, for $k \leq n$, the affinity between $P_{s,k}$ and $P_{t,k}$ is at most equal to $\exp\{-S_k^2\} \leq e^{-c}$. To proceed, let a be a number $a \in [0,1]$ and assume that

*) <u>for every</u> $i \in I_\nu$ <u>and every pair</u> (ξ,θ) <u>of elements of</u> $A_{\nu,i}$ <u>the Hellinger distance between</u> $P_{\xi,k}$ <u>and</u> $P_{\theta,k}$ <u>is at most equal to</u> a.

With all of this, one can state the following lemma.

<u>Lemma</u> 1. <u>In the situation described, if condition</u> $(*)$ <u>is always satisfied, then for any</u> ν-<u>distant pair</u> (i,j) <u>the sum of error probabilities</u> $\pi(A_{\nu,i},A_{\nu,j})$ <u>is at most equal to</u> $[\alpha(2-\alpha)]^{n/2}$ <u>with</u> $\alpha \leq e^{-c} + 2\,a(2-a^2)^{\frac{1}{2}}$ <u>and</u> n <u>equal to the integer part of</u> $b_\nu^2/c+1$.

Proof. To prove this note that for any pair (P, Q) of probability measures the L_1-norm and the Hellinger distance satisfy the inequality

$$\frac{1}{2}\, ||P - Q|| \leq h(P, Q)\, \{2 - h^2(P, Q)\}^{\frac{1}{2}} .$$

Thus, the condition (*) implies that
$\frac{1}{2}\, ||P_{s,k} - P_{\xi,k}|| \leq a\,(2-a^2)^{\frac{1}{2}}$ for every $\xi \in A_{\nu,i}$. A similar inequality holds for $A_{\nu,j}$. It is therefore possible to find tests based on the observations of indices $\ell \in L_k$ only, with sums of error probabilities not larger than $e^{-c} + 2\,a\,(2-a^2)^{\frac{1}{2}}$. The result for the product measures P_θ, $\theta \in A_{\nu,i}$ or $\theta \in A_{\nu,j}$ follows as usual.

At this point, it is convenient to write

$$e^{-2\gamma} = [e^{-c} + 2\,a\,(2-a^2)^{\frac{1}{2}}]\,[2 - e^{-c} - 2\,a(2-a^2)^{\frac{1}{2}}] .$$

The bound in the Lemma can then be replaced by the expression

$$\pi(A_{\nu,i}, A_{\nu,j}) \leq e^{\gamma n} \leq e^{\gamma} \exp\{-\frac{\gamma b_\nu^2}{c+1}\} ,$$

assuming of course that a and c are selected so that

$$e^{-c} + 2\,a\,(2-a^2)^{\frac{1}{2}} < 1 .$$

In the situation considered here, the $p_{\theta,\ell}$ may be rather arbitrary. Thus, there is no obvious way of enforcing the condition (*) except by insisting that the H-diameter of each $A_{\nu,i}$ is at most equal to a.

This leads to the following specification of the construction of Section 2. One selects two numbers a and c such that $w = e^{-c} + 2a(2-a^2)^{\frac{1}{2}} < 1$. One selects also a number $q > 1$ and let $b_{\nu-1} = q\, b_\nu$ with $b_0 \geq \text{diam } \Theta$.

With this one performs the operation described in Section 2 with sets $A_{\nu,i}$ of diameter at most $a \geq 1/125$. (These sets can then be taken independently of ν if one so wishes.) The operation stops at a certain integer m.

It will be convenient to refer to this as Procedure $[c,a,q,m]$.

The performance of the procedure is partially described by the following result in which $e^{-2\gamma} = w(2-w)$ as before.

Proposition 2. For the product situation described and for any finite set Θ metrized by H there is a choice of the integer m and of (c,a,q) such that Procedure $[c,a,q,m]$ gives an estimate $\hat{\theta}$ which satisfies the inequality

$$E_\theta H^2(\hat{\theta},\theta) \leq 54D + (4.03)D \log D + (23.1).$$

Proof. As a first step, let us evaluate the covering numbers $C(\nu)$. By Assumption 1 any set of diameter $d > a$ can be covered by at most C^N sets of diameter a or less, N being the smallest integer at least as large as $\log_2 \frac{d}{a}$. Thus

$$C^N \leq C^{1+\log_2 \frac{d}{a}} = \left(2 \frac{d}{a} \right)^D .$$

For a ball of radius $b_{\nu-1} + a$ the number of sets needed to cover is therefore at most

$$C(\nu) \leq \left[4 \left(\frac{b_{\nu-1}}{a} + 1 \right) \right]^D .$$

Since, typically, b_{m-1} is going to be rather large compared to a we shall replace this by the bound

$$C(\nu) \leq \left(r \frac{b_{\nu-1}}{a} \right)^D ,$$

with the requirement that $r \geq 4 + [a/b_{\nu-1}]$.

One can then substitute this in the formula of Proposition 1 and its Corollary obtaining

$$E_\theta H^2(\hat{\theta},\theta) \leq b_m^2 + 2 e^\gamma T_m ,$$

with

$$T_m = \sum_{0 \leq k \leq m-1} b_k^2 \left(\frac{r b_k}{a} \right)^D \exp\{ -\frac{\gamma}{c+1} b_{k+1}^2 \} .$$

Introducing the point

$$u = \frac{D(c+1)q^2}{2\gamma} ,$$

this can be written

$$T_m = \left(\frac{r}{a} \right)^D \sum_{\nu=1}^{m} (b_m q^\nu)^{2+D} \exp\{ - \frac{D}{2u} b_m^2 q^{2\nu} \} .$$

In this form it resembles the integral

$$I_m = \left(\frac{r}{a} \right)^D \int_0^\infty (b_m q^2)^{2+D} \exp\{- \frac{D}{2u} b_m^2 q^{2\nu} \} d\nu \ .$$

Take as new variable the expression

$x = \frac{1}{u} b_m^2 q^{2\nu}$ and change variables.

The new expression of I_m may be written

$$I_m(z) = \left(\frac{r}{a} \right)^D \frac{u^{1+\frac{D}{2}}}{2 \log q} \int_z^\infty x^{D/2} \exp\{- \frac{D}{2} x \} dx$$

with $z = b_m^2 u^{-1}$.

The approximation to the bound on $E_\theta H^2$ so obtained is then

$$K(z) = uz + \left(\frac{r}{a} \right)^{2s} \frac{e^\gamma}{\log q} u^{1+s} \int_z^\infty x^s \exp\{- sx\} dx$$

where we have written $D = 2s$ for simplicity.

Note that

$$u^s = s^s \left(\frac{c+1}{\gamma} \right)^s q^{2s} .$$

Thus letting

$$A_0 = \left(\frac{r}{a} \right)^{2s} e^\gamma s^s \left(\frac{c+1}{\gamma} \right)^s \frac{q^{2s}}{\log q} ,$$

one can write $K(z) = u f(z)$ with

$$f(z) = z + A_0 \int_z^\infty x^s \exp\{ - sx \} dx \ .$$

The derivative of f is given by

$$f'(z) = 1 - A_0 z^s \exp\{-sz\} .$$

Also $f''(z)$ is proportional to

$$z^{s-1} \exp\{-sz\} (z-1) .$$

It follows that f' achieves its minimum at the point $z = 1$. This minimum is $f'(1) = 1 - A_0 \exp\{-s\}$. In the expression of A_0, the term $q^{2s} (\log q)^{-1}$ is at least equal to $2se$. The product γa^2 cannot be large. Since $r \geq h$ and $D = 2s \geq 1$, the term $A_0 \exp\{-s\}$ will be larger than unity. Thus $f'(z)$ vanishes at two points and 1 is between these points. It follows that the function f achieves its minimum at a value $z_0 \geq 1$ satisfying the relation $z_0 = \log z_0 + A$ with

$$\begin{aligned} A &= \frac{1}{s} \log A_0 \\ &= \log\left(\frac{c+1}{\gamma} \frac{r^2}{a^2} \right) + \log s \\ &\quad + \frac{1}{s} \left[\gamma + \log\left(\frac{q^{2s}}{\log q} \right) \right] \end{aligned}$$

The main terms in this expression are the first two, especially when s is large.

In fact it will be convenient to have a lower bound for A. One can obtain a crude one as follows. By construction

$e^{-2\gamma} \geq e^{-c}(2-e^{-c})$. Thus $2\gamma \leq c$. Also $a\,2\sqrt{2}$ cannot be much larger than unity. We shall assume $a\,2\sqrt{2} \leq 1$. In this case one has certainly $c+1/\gamma > 2$ and $a^{-2} \geq 8$. The first term in the expression of A is then larger than $11 \log 2 \geq 7.6$. Thus, even for $D = 1$, the number A must be larger than 6.9 . Returning to the root z_0 one sees that $z_0 \geq z_1 = A + \log A$.

Another iteration shows that

$$z_0 \leq z_1 [1 + (z_1 - 1)^{-1} \log (\frac{z_1}{A})] .$$

To compute the value of $f(z_0)$, note that one may write

$$\begin{aligned} f(z_0) &= z_0 + \int_{z_0}^{\infty} (\frac{x}{z_0})^{s} \exp\{-s(x-z_0)\}\, dx \\ &= z_0 \{ 1 + \int_0^{\infty} (1+\xi)^{s} \exp\{ - s z_0 \xi \}\, d\xi \} \\ &\leq z_0 [1 + \int_0^{\infty} \exp\{ - s(z_0-1)\xi \}\, d\xi] \\ &= z_0 [1 + \frac{1}{s(z_0-1)}] . \end{aligned}$$

Assembling all terms one obtains the bound

$$K(z_0) \leq s(\frac{c+1}{\gamma})\, q^2\, z_0 \{ 1 + \frac{1}{s(z_0-1)} \} .$$

The root z_0 is an increasing function of the quantity A used above. The <u>bound</u> obtained for $f(z_0)$, that is

$z_0 \{ 1 + [s(z_0-1)]^{-1} \}$, is also an increasing function of z_0 whenever $z_0 \geq 1 + (1/\sqrt{s})$. Since $z_0 \geq A \geq 6.9$ this last condition is certainly satisfied. Thus it will be possible to substitute an upper bound for A and still obtain a bound for $K(z_0)$. It should be emphasized however that $K(z_0)$ is a bound for an integral approximation to the sum which bounds $E_\theta H^2$.

We claim that this same $K(z_0)$ also bounds the sum itself.

To show this consider the function

$$\varphi(\nu) = q^{2\nu(1+s)} \exp\{ - sz_0 q^{2\nu} \}$$

in the interval $0 \leq \nu \leq m$. This φ will be a decreasing function of ν in that interval whenever $z_0 \geq 1 + s^{-1}$. However $z_0 \geq A \geq 6.9$ and $2s \geq 1$, so that the condition is duly satisfied. It follows then that

$$\sum_{\nu=1}^{m} \varphi(\nu) \leq \int_0^\infty \varphi(\nu) d\nu ,$$

proving our claim.

To bound the quantity A , let us recall that $\gamma = -\frac{1}{2} \log w(2-w)$ with $w = e^{-c} + 2a\,(2-a^2)^{\frac{1}{2}}$.

For small a , the term a^2 does not play a major role here. Thus we shall replace w by the larger $w_1 = e^{-c} + a\,2\sqrt{2}$ and arbitrarily decide to put e^{-c} equal to

$a\, 2\sqrt{2}$. With this substitution $w_1 = 2\, e^{-c}$ and $w_1(2-w_1) = 4\, e^{-c}(1-e^{-c})$.

Here again we shall select c arbitrarily and set $c = 3.69$. Then one can write

$$.024971 \leq e^{-c} \leq .025$$

and

$$(77.9)\, 10^{-6} \leq a^2 \leq (78.1)\, 10^{-6} .$$

In particular $\sqrt{1-(a^2/2)} \geq .9998$ and $.097385 \leq w(2-w) \leq w_1(2-w_1) \leq .0975$ and, therefore, $1.1639 \leq \gamma \leq 1.1648$.

From this we also deduce that

$$\frac{c+1}{\gamma} \leq 4.03 ,$$

and

$$-\log a^2 \leq 9.46 .$$

With this particular choice of constants we certainly have $A \geq 12.9$ hence

$$u\, z_0 = \frac{D}{2} \frac{c+1}{\gamma} q^2\, z_0 \geq 25 .$$

From this one sees that $a/b_m \leq 4\; 10^{-4}$ and therefore $r \leq 4.002$. Finally

$$\log \left(\frac{c+1}{\gamma} \frac{r^2}{a^2} \right) \leq 13.627 .$$

To finish the specification, we have to select a value of q . The above formulas seem to indicate that q should be selected as a decreasing function of D having a form somewhat like $q = 1 + K_1/(K_2+D)$. For simplicity we shall take $q = 1.2$. With this choice

$$\gamma - \log \log q \leq 2.868 .$$

Therefore

$$A \leq 14 + \log \frac{D}{2} + \frac{5.736}{D} .$$

Using this expression, it is possible to bound the root z_0 of the equation $z_0 = \log z_0 + A$. Indeed a short computation shows that when $A \geq 12.9$ one has $z_0 \leq (1.22)A$. Thus

$$z_0 \leq 17.08 + (1.22) \log \frac{D}{2} + \frac{7}{D} .$$

In addition

$$u = \frac{D}{2} \frac{c+1}{\gamma} q^2 \leq (2.902) D .$$

Finally

$$E_\theta H^2(\hat{\theta},\theta) \leq u z_0 \{ 1 + \frac{2}{D} \frac{1}{(z_0-1)} \}$$

$$\leq (2.902) D \{17.08 + (1.22) \log \frac{D}{2} + \frac{7}{D} \}$$

$$\times \{ 1 + \frac{(.132)}{D} \}$$

$$\leq (2.902)(1.132) D \{16.235 + (1.22) \log D + \frac{7}{D} \}$$

$$\leq (53.5) D + (4.02) D \log D + (23.1) .$$

This is smaller than the bound given in the Proposition.

4. The identically distributed case.

In this Section we shall keep all the assumptions made in Section 3 and add the further requirement that the spaces $(\mathcal{X}_\ell, \mathcal{A}_\ell, p_{\theta,\ell})$ are replicates of each other.

One can proceed exactly as in Section 3. However, in the present case it is possible to assume in addition that the pieces L_k $k = 1, 2, \cdots, n$ of the partition used on the index set L have a cardinality independent of k. Thus they are replicates of one another.

When this condition is satisfied, the requirement (*) of Section 3 can hold for sets $A_{\nu,i}$ which have a diameter $\leq a_\nu$ depending on ν.

Indeed, let h^2 be the square diameter of $A_{\nu.i}$ for the sum of square Hellinger distances taken on the index set L_k. This must be inferior to a^2. However there are $n+1$ pieces L_k so that the diameter on the whole sum has the form

$$a_\nu^2 = n\, h^2 + v^2$$

with $0 \leq v^2 < c$. Condition (*) will be satisfied if $a_\nu^2 - v^2 \leq n\, a^2$. The number n is at least equal to the integer part of $b_\nu^2/(c+1)$. The condition (*) is then certainly satisfied whenever a_ν^2 is such that

$$a_\nu^2 \leq \left(\frac{b_\nu^2}{c+1} - 1 \right) a^2 ,$$

or, equivalently

$$\frac{b_\nu^2}{a_\nu^2} \geq \frac{1}{a^2} \left[1 + \frac{(c+1)\, a^2}{a_\nu^2} \right] .$$

Assuming, as in Section 3, that $b_{\nu-1} = q\, b_\nu$, the ratio $2[b_{\nu-1} + a_\nu]\, a_\nu^{-1}$ is equal to $2[q\, b_\nu + a_\nu]\, a_\nu^{-1}$. The above argument shows that one may assume

$$[q\, b_\nu + a_\nu]\, a_\nu^{-1} \leq 1 + \frac{q}{a} \left[1 + \frac{(c+1)\, a^2}{2\, a_\nu^2} \right] .$$

The construction procedure stops at some diameter b_m which will be determined later on. However the integer m to be selected is bound to be such that a_m remains substantially larger than a. This is why we have kept separately the term which involves $a\, a_\nu^{-1}$.

In any event, $[q\, b_\nu + a_\nu]\, a_\nu^{-1}$ remains bounded by some number K independent of ν for $0 \leq \nu \leq m$. The number $C(\nu)$ of sets of diameter a_ν needed to cover the ball of radius $b_{\nu-1} + a_\nu$ is therefore bounded by an inequality of the type

$$C(\nu) \leq 4^D K^D .$$

The evaluation of probabilities of error of Section 3 remains valid. This, with the Corollary of Proposition 1, yields the bound

$$E_\theta\, H^2(\hat{\theta},\theta) \leq b_m^2 + e^\gamma\, (4K)^D \sum_{0\leq\nu\leq m-1} b_\nu^2 \exp\{-\frac{\gamma}{c+1}\, b_{\nu+1}^2\}$$

$$= b_m^2 + e^\gamma (4K)^D \sum_{1\leq k\leq m} b_m^2\, q^{2k} \exp\{-\frac{\gamma\, b_m^2}{(c+1)q^2}\, q^{2k}\} .$$

Let z be the ratio

$$z = \frac{\gamma}{(c+1)q^2}\, b_m^2 .$$

A simple computation shows that the function of k which is summed here is a decreasing function as long as $z \geq 1$. Therefore the sum can be bounded by an integral giving

$$E_\theta\, H^2(\hat{\theta},\theta) \leq B(z) \qquad \text{with}$$

$$B(z) = \frac{(c+1)q^2}{\gamma}\, \{ z + e^\gamma (rK)^D\, I(z) \} ,$$

$$I(z) = \frac{\gamma}{(c+1)q^2}\, b_m^2 \int_0^\infty q^{2k} \exp\{-\frac{\gamma}{(c+1)q^2}\, q^{2k}\}\, dk$$

$$= \frac{1}{2\log q} \int_z^\infty e^{-x}\, dx$$

$$= \frac{1}{2\log q}\, e^{-z} .$$

This yields

$$B(z) = \frac{(c+1)q^2}{\gamma}\, \{z + \frac{e^\gamma (4D)^K}{2\log q}\, e^{-z}\} .$$

This function of z reaches a minimum

$$B(z_0) = \frac{(c+1)q^2}{\gamma} (1+z_0)$$

at the point z_0 such that

$$z_0 = D(\log 4K) + \gamma - \log \log q^2 .$$

Take $c = 3.69$ and $q = 1.2$ as in Section 3. Then $\gamma - \log \log q^2 \leq 2.175$. Also, using a crude lower bound for K one can see that $z_0 \geq 8.42$. Therefore the replacement of the sum by the integral is fully justifiable. In addition this lower bound for z_0 will imply that $a_m^2 \geq 85\, a^2$ and therefore

$$K \leq 1 + \frac{1.2}{a}\left(1 + \frac{4.69}{170}\right)$$

$$\leq 1 + \frac{1.24}{a} \leq 141 .$$

Finally, $\log 4K \leq 6.34$ and $z_0 \leq (2.175) + (6.34)D$, yielding

$$B(z_0) \leq 18.50 + (36.8)D .$$

To summarize, we have proved the following result.

<u>Proposition</u> 3. <u>Assume that the experiment</u> $\{\mathcal{X}, \mathcal{A}, P_\theta ; \theta \in \Theta\}$ <u>satisfies the conditions of Section</u> 3 <u>with a dimension</u> D. <u>Assume in addition that the components</u> $(\mathcal{X}_\ell, \mathcal{A}_\ell, P_{\theta,\ell})$ <u>are replicates of one another</u>.

<u>Then, there are estimates</u> $\hat{\theta}$ <u>such that</u>

$$E_\theta H^2(\hat{\theta},\theta) \leq (18.50) + (36.8)D .$$

The bound obtained here can be compared with the one given by the (incorrect) argument of [1]. This was $16 \log 6C$, where C is the number of sets of diameter ε needed to cover a set of diameter $\varepsilon\, 2^{11/2}$. The $\log C$ of this previous assertion is therefore to be compared to the present $(5.5)\, D$, from which one can presume that the present bound is in fact smaller than the one proposed in [1]. The reason is that the choice of constants a, c and q made here is better than the previous one.

Section 5. Extension to unbounded parameter sets.

For the proof of Propositions 1, 2 and 3 we have assumed that i) the set Θ is finite and that ii) the function H is finite on $\Theta \times \Theta$. The present section is intended to show that both restrictions are removable.

Of course we shall keep the main dimensionality Assumption 1, according to which any set of diameter 2d can be covered by no more than 2^D sets of diameter d, at least if $125\,d \geq 1$.

In the preceding sections, the assumption that Θ is finite was not used in any essential manner. Its role was simply to avoid at that point arguments about the existence of finite coverings by sets of diameter $\leq a_\nu$.

When Assumption 1 is satisfied and when H is a bounded metric for Θ such finite partitions $\{A_{\nu,i}; i \in I_\nu\}$ will certainly exist as long as $a_\nu \geq a \geq 1/125$. Therefore, the construction procedure can be carried out exactly in the same manner. Thus Propositions 2 and 3 remain entirely valid if the assumption that Θ is finite is replaced by the assumption that H is bounded on $\Theta \times \Theta$.

In the situation covered by these Propositions the space $\{\mathcal{X}, \mathcal{A}, P_\theta\}$ is a direct product of components $\{\mathcal{X}_\ell, \mathcal{A}_\ell, P_{\theta,\ell}\}$, $\ell \in L$. If the cardinality of L is a finite number card $L = N$, then $H^2(s,t) \leq N$ for all pairs (s,t) and the boundedness assumption is automatically satisfied.

If on the contrary L is infinite, the sum H^2 may well be unbounded or even assume actually infinite values on $\Theta \times \Theta$.

In such a situation, one can avoid all difficulties by redefining what is meant by "estimate" , as explained in [2].

Specifically, let $\mathcal{E}$ be the experiment $\mathcal{E} = \{P_\theta ; \theta \in \Theta\}$. It defines a certain abstract L-space $L(\mathcal{E})$. On Θ itself, Let Γ be a uniform lattice of real valued functions which separates the points of Θ and is such that, for each $\theta \in \Theta$ the function $t \rightsquigarrow H(t,\theta)$ is a pointwise supremun of the set $\{\gamma : \gamma \in \Gamma , \gamma(t) \leq H(t,\theta)$ for all $t\}$.

Define "estimates" as transitions T from $L(\mathcal{E})$ to the dual Γ^* of Γ and define the risk of T by the formula:

$$R(T,\theta) = \sup \{ \gamma T P_\theta \ ; \ \gamma \leq H(\circ,\theta) \} .$$

With these definitions one can assert the following.

<u>Theorem</u> 1. <u>Assume that</u> $\{\mathcal{X}, \mathcal{A}, P_\theta\}$ <u>is a direct product, as described. Suppose that the dimensionality Assumption</u> 1 <u>is satisfied. Then, without any other restrictions, there always exists a transition</u> T <u>from</u> $L(\mathcal{E})$ <u>to</u> Γ^* <u>such that</u>

$$R(T,\theta) \leq B = 54D + (4.02)D \log D + 24$$

for all $\theta \in \Theta$.

If in addition the family $\{P_\theta ; \theta \in \Theta\}$ *is of* Σ-*finite type, there exists a randomized estimate* $\hat{\theta}$ *in the usual sense such that*

$$E_\theta \, H^2(\hat{\theta},\theta) \leq 2B + \frac{2}{(125)^2} .$$

Proof. To prove the first assertion let us first note that for any finite subset S of the index set L, the expression $\{\sum_{\ell \in S} h_\ell^2 (s,t)\}^{\frac{1}{2}}$ still defines a pseudo-metric on Θ . It results immediately from this that Θ can be partitioned in classes $\{\Theta_j ; j \in J\}$ such that

i) for $(s,t) \in \Theta_j \times \Theta_k$; the value $H(s,t)$ is finite

ii) for $(s,t) \in \Theta_j \times \Theta_k$, $j \neq k$, the value $H(s,t)$ is infinite .

The second property just mentioned shows if $(s,t) \in \Theta_j \times \Theta_k$, $j \neq k$, then $\int \sqrt{dP_s dP_t} \leq \exp\{- H^2(s,t)\} = 0$, so that P_s and P_t are disjoint.

Consider then a finite subset $F \subset \Theta$. Let $J(F) \subset J$ be the set of indices j such that $F_j = \Theta_j \cap F$ is nonempty. Since $J(F)$ is finite there is a measurable partition $\{\mathcal{X}_j ; j \in J(F)\}$ of $\mathcal{X}$ such that $P_\theta(\mathcal{X}_j) = 1$ if $\theta \in F_j$.

For each F_j and each space $\mathcal{X}_j$ one can apply Proposition 2. This obviously gives an estimate $\hat{\theta}$ such that $E_\theta H^2(\hat{\theta},\theta) \leq B$ for all $\theta \in F$.

Now, with the replacement of "estimates" by transitions, the minimax theorem holds in the form

$$\min_T \sup_\theta R(T,\theta) = \sup_\mu \inf_T \int R(T,\Theta)\mu(d\Theta) ,$$

the supremum being taken over the set of probability measures μ which have <u>finite support</u> on Θ. This implies the first statement of the theorem.

For the second statement consider one of the elements Θ_j of the partition of Θ. Select a point $\theta_j \in \Theta_j$ and let $U(j,m)$ be the ball of radius 2^m centered at θ_j. According to Assumption 1, this ball can be covered by a finite partition whose elements have diameter at most $(1/125)$. Select a point in each set of the partition of $U(j,m)$. Proceed for increasing m, selecting new points only in $U(j,m+1) \setminus U(j,m)$. This yields a countable subset S_j of Θ_j such that i) for each $\theta \in \Theta_j$ there is an $s \in S_j$ for which $H(\theta,s) \leq 1/125$ and ii) the set S_j is discrete.

Let $S = \bigcup_j S_j$ be topologized by the discrete topology, and let Γ_0 be the restriction of Γ to S. Take a T, transition from $L(\mathcal{E})$ to Γ^*, such that $R(T,\Theta) \leq B$ for all θ . Then, there is a T_0, transition from $L(\mathcal{E})$ to Γ_0^* such that $R(T,\theta) \leq 2\{B + (125)^{-2}\}$. Indeed, to prove this it is enough to prove it when Θ is finite. In this case T can be approximated by a simple Markov kernel which gives for each $x \in \mathcal{X}$ a probability measure T_x carried by a finite subset of Θ (see Theorem 1 Section 3 of [2]) . One replaces each point of the support of that measure by the element of S closest to it. The inequality results then from the Hilbertian structure of the metric H . The statement remains valid for arbitrary Θ by the usual compactness argument.

Note that if S is discrete then for each θ , and each real number α , the set $\{s : H^2(\theta,s) \leq \alpha \ , \ s \in S\}$ is finite and therefore compact. Therefore, according to a standard argument of decision theory, if $R(T_0,\theta) < \infty$ the image $T_0 P_\theta$ of P_θ must admit an extension to a Radon measure on S. If in addition the system $\{\mathcal{X},\mathcal{A},P_\theta\}$ is Σ-finite, the transition T_0 is automatically representable by a Markov kernel. Hence the statement of the theorem.

<u>Note</u> 1. The same argument proves also that if Θ itself can be topologized in such a way that for each θ and $\alpha \in R$ the set $\{s : H^2(\theta,s) \leq \alpha\}$ is a compact Hausdorff space then the conclusion of the theorem holds in the form $E_\theta H^2(\hat{\theta},\theta) \leq B$.

Note 2. Proposition 3 can be translated in the same manner. The necessary modifications are obvious.

To conclude, let us note that the dimensionality Assumption 1 has been stated as imposing a limit on the number of sets of diameter d necessary to cover a set of diameter $2d$ but with the restriction that $125\, d \geq 1$.

This is reasonable in the sense that there is no point in looking at sets which are very much smaller for the purposes of Propositions 2 and 3. However, in many cases, one can strengthen the Assumption by removing the lower bound on d. In this case a set of finite diameter is necessarily precompact for the distance H . Nothing essential is changed if one replaces it by its compactification. If so, the second statement of Theorem 2 holds automatically, in the form $E_\theta\, H^2(\hat{\theta},\theta) \leq B$ provided that $\{P_\theta\,;\, \theta \in \Theta\}$ be Σ-finite.

This strengthening of Assumption 1 is appealing in the identically distributed case since then the dimension does not depend on the number of observations, as long as it remains finite.

6. Supplementary remarks on the independent case.

The results of Sections 3 and 4 have been stated in terms of the distance H defined by the sum $H^2(s,t) = \Sigma_\ell h_\ell^2(s,t)$. Although this is often acceptable, it is clearly not appropriate in some circumstances.

For instance, it may happen that $H^2(s,t) \leq 2$ but that P_s and P_t are disjoint. Thus the problem arises of deriving inequalities for loss functions W which afford a sharper discrimination.

One possibility is the loss function W defined on $\Theta \times \Theta$ by the relation $W(s,t) = -\log \rho(s,t)$, where $\rho(s,t)$ is the affinity $\int \sqrt{dP_s \, dP_t}$.

Writing $\rho_\ell(s,t) = 1 - h_\ell^2(s,t)$ and $W_\ell(s,t) = -\log \rho_\ell(s,t)$ one sees that $W(s,t)$ is the sum $\Sigma_\ell W_\ell(s,t)$. It is therefore possible to begin an argument exactly similar to the one carried out in Section 3. However the argument breaks down, as we shall now indicate. We do not know whether appropriate modifications could salvage the proofs, even though, as will be indicated elsewhere, one can carry out a partial argument which still yields some information. To be specific, suppose that Θ is a finite set. Then, one can first carry out tests between pairs (P_s, P_t) which are disjoint and reduce the problem to a case in which W remains finite.

If one partitions the sums $\Sigma_\ell W_\ell(s,t)$ into blocks $\sum_{\ell \in L_k} W_\ell(s,t) \geq c$, the bounds $\exp\{- \sum_{\ell \in L_k} h_\ell^2\}$ of Section 3 can be replaced by the better bounds

$$\exp\{- \sum_{\ell \in L_k} W_\ell(s,t)\} = \prod_{\ell \in L_k} \rho_\ell(s,t) .$$

Similarly, the restriction that the H-diameter of sets $A_{\nu,i}$ be inferior to a certain value a can be replaced by the somewhat stronger requirement that for every pair (ξ_1, ξ_2) of elements of $A_{\nu,i}$ one have $\sup\{- W(\xi_1, \xi_2)\} \geq 1 - a^2$.

In other words the procedure works exactly as before, <u>except that it is no longer possible to obtain a usable lower bound on the number of blocks</u> L_k <u>of the partition of</u> L .

This difficulty can be remedied if one assumes that each component $(\mathcal{X}_\ell, \mathcal{A}_\ell, p_{\theta,\ell})$ provides very little information by itself. For instance one may assume that there is a number $\varepsilon_0 \leq 1/2$ such that $h_\ell^2(s,t) \leq \varepsilon_0$ for all pairs (s,t) . However, under such an assumption, the result of Section 3 is already of the same general nature as the one that we sought here, since we can write

$$H^2 \leq W \leq [-\frac{1}{\varepsilon_0}\log(1-\varepsilon_0)]\,H^2 \leq (2 \log 2)\,H^2 .$$

The only other available improvement is that, when the restriction $h_\ell^2 \leq \varepsilon_0$ is satisfied one can obtain better coefficients. For this it is enough to replace $(c+1)$ by $(c+\varepsilon_0)$ at all approximate places.

References

[1] L. LeCam - "On the information contained in additional observations."
To appear - Ann. Stat. August 1974

[2] L. LeCam - Notes on asymptotic methods in statistical decision theory.
Centre de Recherches Mathématiques
Université de Montréal. 1974
XVI + 270 pages

[3] A. N. Kolmogorov - On some asymptotic characteristics of totally bounded metric spaces.
Doklady Akad Nauk SSSR 108 (3) (1956)