ON THE SPEED OF CONVERGENCE OF POSTERIOR DISTRIBUTIONS*

by

L. LeCam

University of California, Berkeley

I. Introduction. Let X_i ; $i=1,2,\cdots$ be a sequence of independent identically distributed random variables. Assume that the individual distributions of the X_j belong to a family $\{p_\theta ; \theta\epsilon\Theta\}$ indexed by some set Θ which carries a prior distribution μ . It is well known that under a variety of fairly mild conditions, the posterior distribution of θ given $X_1, X_2, \cdots, X_n$ concentrates itself around the "true value" θ_o of the parameter. It has also been shown by Lorraine Schwartz [1] that, under rather weak restrictions, the posterior probability of a neighborhood of θ_o tends to unity exponentially as n tends to infinity.

The present paper concerns itself with some refinements of this result of Lorraine Schwartz. The inequalities given here are probably not the best possible, but they already

*This paper was prepared with the partial support of the U.S. Army Research Office (Durham) contract DA-ARO-D-31-124-G1135.

yield various bounds which seem to answer agreeably certain questions raised by the study of sequential estimation problems.

The second section of the paper gives some preliminary definitions and propositions concerning the Hellinger transforms used as tools in the sequel. The third section elaborates two bounds on the posterior probabilities. The fourth section gives an application of these bounds to the independent identically distributed case under standard or weaker than standard regularity restrictions.

II. <u>The Hellinger transform.</u> Let $\mathcal{X}$ be a set carrying a σ-field $\mathcal{A}$. Let S be a finite set. For each $s \in S$ let P_s be a probability measure on $\mathcal{A}$. In the product space R^S, product of copies of the real line R let U be the simplex formed by element $\alpha = \{\alpha_s; s \in S\}$ such that $\alpha_s \geq 0$ and $\sum_s \alpha_s = 1$. Let $m = \sum_s P_s$ and let f_s be the density $f_s = dP_s/dm$. The Hellinger transform of the family $\{P_s; s \in S\}$ is by definition the function $\alpha \rightsquigarrow \varphi(\alpha)$ defined on U by

$$\varphi(\alpha) = \int [\prod_s f_s^{\alpha_s}] \, dm \quad .$$

This is a form of the Laplace transform of the joint distribution of logarithms of likelihood ratios.

If $\mathcal{B}$ is a sub-σ-field of $\mathcal{A}$ one can restrict the measures P_s and m to $\mathcal{B}$ and compute the Hellinger transform on $\mathcal{B}$ instead of $\mathcal{A}$. To indicate the σ-field used we shall employ the notation $\varphi[\alpha; \mathcal{A}]$, $\varphi[\alpha;\ ]$ and so forth. The following two lemmas are well known (See for instance C. Kraft [2]).

Lemma 1. The logarithm of $\varphi(\alpha)$ is a convex function of α. Furthermore if $\mathcal{B} \subset \mathcal{A}$ then

$$\varphi[\alpha; \mathcal{A}] \leq \varphi[\alpha;\ ] \ .$$

Proof. The first statement is the substance of Hölder's inequality. For the second statement, let m' be m restricted to $\mathcal{B}$ and let f'_s be the density of the restriction of P_s to $\mathcal{B}$ relatively to m'. Then f'_s is the conditional expectation $E\{f|\mathcal{B}\}$ taken for the measure m. Thus

$$\varphi(\alpha;\mathcal{B}) = \int \{\prod_s [E(f_s|\mathcal{B})]^{\alpha_s}\} dm \ .$$

However, since the function $(u_s;\ s \in S) \rightsquigarrow \Pi_s\, u_s^{\alpha_s}$ is concave we have $\Pi_s [E(f_s|\mathcal{B})]^{\alpha_s} \geq E\{(\Pi f_s^{\alpha_s})|\mathcal{B}\}$. Hence the result.

Taking "expectations" and conditional expectations for the measure m as above one obtains also the following result.

Lemma 2. Let $\{\mathcal{B}_\nu\}$ be an increasingly directed family of sub-σ-fields of $\mathcal{A}$. Assume that $\mathcal{A}$ is the smallest σ-field containing all the $\mathcal{B}_\nu$. Then $E[f_s|\mathcal{B}_\nu]$ converges in m-measure to f_s and $\varphi[\alpha;\mathcal{B}_\nu]$ decreases to $\varphi[\alpha;\mathcal{A}]$ for each $\alpha \in U$.

Proof. If the $\mathcal{B}_\nu$ were indexed increasingly by integers, this would follow for instance from the usual martingale theorems. However, even in the general case, the f_s are bounded by zero and unity, hence approximable in any L_p space by functions which are $\mathcal{B}_\nu$-measurable for a suitable $\mathcal{B}_\nu$. The result follows immediately.

Let $\pi = \{A_1, A_2, \cdots, A_n\}$ be a finite partition of unity by sets $A_j \in \mathcal{A}$. Let $\varphi_\pi(\alpha)$ be defined by

$$\varphi_\pi(\alpha) = \sum_j \prod_s [P_s(A_j)]^{\alpha_s}$$

Corollary. The value $\varphi[\alpha; \mathcal{A}]$ is the infimum $\inf_\pi \varphi_\pi(\alpha)$ taken over all finite partitions of unity by elements of $\mathcal{A}$.

According to this corollary, the definition of $\varphi(\alpha; \mathcal{A})$ used here is the same as the original definition of Hellinger.

To obtain bounds on posterior probabilities we shall use a particular inequality on Hellinger transforms for product measures and averages of product measures as follows.

Let T be a Borel set in a complete separable metric space. Let $(\mathcal{X},\mathcal{A})$ and $(\mathcal{Y},\mathcal{B})$ be two measurable spaces. Let P be a probability measure on $\mathcal{A}$ and let Q be a probability measure on $\mathcal{B}$. For each $t \in T$, let F_t be a probability measure on $\mathcal{A}$ and let G_t be a probability measure on $\mathcal{B}$

Assume that for each $A \in \mathcal{A}$, the function $t \rightsquigarrow F_t(A)$ is measurable in t. Similarly assume $t \rightsquigarrow G_t(B)$ measurable. Denote product measures by the tensor symbol $F_t \otimes G_t$. Let μ and ν be Borel probability measures on T.

<u>Proposition</u> 1. <u>With the notation just described, let</u> $L = \int F_t \mu(dt)$, $M_\nu = \int G_t \nu(dt)$ <u>and</u> $H = \int (F_t \otimes G_t)\mu(dt)$. <u>Then, for any</u> $\alpha \in [0,1]$

$$\int (dH)^{\alpha}(dP\otimes Q)^{1-\alpha} \leq \{\int (dL)^{\alpha}(dP)^{1-\alpha})\}\gamma(\alpha) \quad ,$$

<u>where</u> $\gamma(\alpha) = \sup_\nu \int (dM_\nu)^{\alpha}(dQ)^{1-\alpha}$.

Proof. Consider the product space $\{\mathcal{X}\times\mathcal{Y}, \mathcal{A}\times\mathcal{B}\}$ carrying the measures H and $P\otimes Q$. Let $\pi' = \{A_1, A_2, \cdots, A_m\}$ be a partition of $\mathcal{X}$ by elements of $\mathcal{A}$ and let $\pi'' = \{B_1, B_2, \cdots, B_n\}$ be a partition of $\mathcal{Y}$ by elements of $\mathcal{B}$. Let $\pi = \pi' \times \pi''$ be the partition $\{A_i \times B_j \; ; \; i=1,2,\cdots,m \, , \; j=1,2,\cdots n\}$ of $\mathcal{X}\times\mathcal{Y}$. According to Lemmas 1 and 2 the Hellinger product $\varphi(\alpha) = \int (dH)^{\alpha}(dP\otimes Q)^{1-\alpha}$ is smaller than

$$\varphi_{\pi}(\alpha) = \sum_{i,j} [H(A_i\times B_j)]^{\alpha}[P(A_i)Q(B_j)]^{1-\alpha} \quad .$$

Now

$$H(A_i\times B_j) = \int F_t(A_i)G_t(B_j)\mu(dt)$$

$$= L(A_i) \int G_t(B_j)\nu_i(dt) \quad ,$$

with $L(A_i)\nu_i(dt) = F_t(A_i)\mu(dt)$. Thus, for any partition $\pi = \pi' \times \pi''$ one can write

$$\varphi(\alpha) \leq \varphi_{\pi}(\alpha) = \sum_{i,j} [L(A_i)]^{\alpha}[P(A_i)]^{1-\alpha}[\int G_t(B_j)\nu_i(dt)]^{\alpha}[Q(B_j)]^{1-\alpha}$$

In this double sum fix i and take a sum over j first. The corresponding factor is

$$\sum_j [\int G_t(B_j)\nu_i(dt)]^{\alpha}[Q(B_j)]^{1-\alpha} \quad .$$

Take an infimum over π'', leaving π' fixed. Applying Lemma 2 to the above factor we get

$$\varphi(\alpha) \leq \inf_{\pi''} \varphi_\pi(\alpha) = \sum_i [L(A_i)]^\alpha [P(A_i)]^{1-\alpha} \int (dM_{\nu_i})^\alpha (dQ)^{1-\alpha} .$$

$$\leq \sum_j [L(A_i)]^\alpha [P(A_i)]^{1-\alpha} \gamma(\alpha) .$$

Applying Lemma 2 again, but this time for refinement of π', this yields

$$\varphi(\alpha) \leq \gamma(\alpha) \int (dL)^\alpha (dP)^{1-\alpha} .$$

This is the desired result.

III. Two inequalities on posterior probabilities. Let T be a Borel set in a complete separable metrizable space. For each $t \in T$, let P_t be a probability measure on a space $\{\mathcal{X}, \mathcal{A}\}$. Assume that for each $A \in \mathcal{A}$ the function $t \leadsto P_t(A)$ is measurable. Let μ be a probability measure on T. For each Borel set $B \subset T$ let P_B be the probability measure defined by

$$\mu(B) P_B(A) = \int_B P_t(A) \mu(dt)$$

if $\mu(B) > 0$.

Define a probability measure on $\mathcal{X} \times T$ by the integral

$\int_B P_t(A)\mu(dt)$ and let $x \rightsquigarrow F_x$ be a choice of the conditional distribution of t given x for this joint distribution. A simple computation shows that one must have almost everywhere the equality

$$F_x(B) = \mu(B) \frac{dP_B}{dP_T}(x) \quad ,$$

where the right side is the product of $\mu(B)$ by the evaluation at x of the appropriate Radon-Nikodym derivative.

In this expression, the existence and essential uniqueness of the measures F_x is guaranteed by the assumption made on the Borelian character of T. To derive from the equality some useful inequalities one can proceed as follows.

Let P be another probability measure on $\{\mathcal{X},\mathcal{A}\}$. This measure need not be one of the P_t. Take a number $\alpha \in [0,1]$ and two sets V and C which are Borel subsets of T. To use the inequalities given below, one attempts to choose V so that P_V be much closer to P than P_C.

Consider the number $\beta(\alpha) = \int (dP_C)^{\alpha}(dP)^{1-\alpha}$ and the corresponding probability measure $dH = \beta(\alpha)^{-1} (dP_C)^{\alpha}(dP)^{1-\alpha}$. This is well defined if $\beta(\alpha) > 0$, that is if P and P_C are not disjoint. Let φ be a test of H against P

P chosen so that $\int \varphi dH$ and $\int (1-\varphi)dP$ are "small".

<u>Proposition</u> 2. <u>With the notations and assumptions just described, let</u> a <u>be a positive number and let</u> A <u>be the set of values of</u> x <u>where</u> $\frac{dP_V}{dP} \geq e^{-a}$. <u>Then</u>

$$\int_A [F_x(C)]^\alpha P(dx) \leq \int_A (1-\varphi)dP + e^{\alpha a}\beta(\alpha)\left[\frac{\mu(C)}{\mu(V)}\right]^\alpha \int_A \varphi dH .$$

<u>Proof</u>. Let f by the function $x \rightsquigarrow f(x) = [F_x(C)]^\alpha$. By construction $0 \leq f(x) \leq 1$. Also, on the set A the measure P is absolutely continuous with respect to P_T and $dP_T \geq \mu(V)dP_V$. Thus, on this set A one can write

$$f(\lambda) \leq \left[\frac{\mu(C)}{\mu(V)}\right]^\alpha \left[\frac{dP_C}{dP}\right]^\alpha \left[\frac{dP_V}{dP}\right]^{-\alpha}$$

$$= \left[\frac{\mu(C)}{\mu(V)}\right]^\alpha \beta(\alpha) \frac{dH}{dP} \left[\frac{dP_V}{dP}\right]^{-\alpha}$$

this gives

$$\int_A fdP = \int_A (1-\varphi)fdP + \int_A \varphi fdP$$

$$\leq \int_A (1-\varphi)dP + \beta(\alpha)e^{\alpha a}\left[\frac{\mu(C)}{\mu(V)}\right]^\alpha \int_A \varphi \frac{dH}{dP} dP$$

$$\leq \int_A (1-\varphi)dP + \beta(\alpha)e^{\alpha a}\left[\frac{\mu(C)}{\mu(V)}\right]^\alpha \int_A \varphi dH .$$

this is the desired result.

Corollary 1. Under the assumptions of Proposition 2 one has for each $\varepsilon > 0$

$$P\{F_x(C) \geq \varepsilon\} \leq P(A^c) + \varepsilon^{-\alpha} \beta(\alpha) e^{\alpha a} \left[\frac{\mu(C)}{\mu(V)}\right]^{\alpha} .$$

This is an immediate consequence of Markov's inequality applied to the inequality of Proposition 2 with φ taken equal to unity.

Corollary 2. Under the same assumptions

$$P[F_x(C) \geq \varepsilon] \leq P(A^c) + \int_A (1-\varphi)\, dP + \varepsilon^{-1} e^a \frac{\mu(C)}{\mu(V)} \int \varphi\, dP_C$$

for any test φ of P_C against P.

This is obtained from Proposition 2 by taking $\alpha = 1$. The two inequalities given by the above corollaries are really very closely related since $\beta(\alpha)$ will be small when the corresponding terms in Corollary 2 are small and conversely. However actual use will depend on how easily the appropriate terms can be evaluated.

In typical cases, to force $P(A^c)$ to be small one will have to take for V a very small neighborhood of a parameter point θ (if any) such that $P_\theta = P$. Thus, the term $[\mu(C)]^{\alpha} [\mu(V)]^{-\alpha}$ will be large. The inequality will be effective only if either $\beta(\alpha)$ or $\int \varphi dP_C$ can be made much smaller than $[\mu(V)]^{\alpha}$. We shall show in the next section that this is often the case. For the present, let us note that it is often convenient to evaluate $P(A^c)$ through the use of Hellinger transforms as follows. Take $\alpha \leq 1/2$. Let $\varphi(\alpha) = \int (dP_V)^{\alpha}(dP)^{1-\alpha}$. Assume $\varphi(\alpha) > 0$ and let M be the probability measure $dM = [\varphi(\alpha)]^{-1}(dP_V)^{\alpha}(dP)^{1-\alpha}$. This construction gives $\varphi(\alpha)\,\frac{dM}{dP} = \left(\frac{dP_V}{dP}\right)^{\alpha}$. Consider the affinity $\int [dMdP]^{1/2} = \varphi(\frac{\alpha}{2})[\varphi(\alpha)]^{1/2}$. Since the logarithm of φ is convex one can write $\log \varphi(\frac{\alpha}{2}) \geq \frac{1-\alpha/2}{1-\alpha} \log \varphi(\alpha)$. Since we limit ourselves to $\alpha \in (0,1/2]$ this gives $\varphi(\frac{\alpha}{2}) \geq [\varphi(\alpha)]^{3/2}$ hence $\int [dMdP]^{1/2} \geq \varphi(\alpha)$. It follows that $||P-M|| \leq 2[1-\varphi^2(\alpha)]^{1/2}$. Finally this gives

$$P[1-\varepsilon \leq \frac{dM}{dP} \leq 1+\varepsilon] \geq 1 - 2\varepsilon^{-1}[1-\varphi^2(\alpha)]^{1/2}$$

Direct application of the inequality of Corollary 1 of Proposition 2 yields the following result.

Lemma 3. Let P, P_V and P_C be defined as above. Let $\varphi(\alpha) = \int (dP_V)^{\alpha}(dP)^{1-\alpha}$ and let a be a non negative number. Then, assuming $\alpha \in (0,1/2]$ one has

$$P[F_x(C) \geq \varepsilon] \leq 2\,\frac{[1-\varphi^2(\alpha)]^{1/2}}{1-e^{-\alpha a}} + \varepsilon^{-\alpha}\beta(\alpha)e^{\alpha a}\left[\frac{\mu(C)}{\mu(V)}\right]^{\alpha}$$

Finally let us note that the function $x \rightsquigarrow \varphi(x) = F_x(C)$ is a test function. The power of this test satisfies the relation $\int \varphi dP_T = \mu(C)$. Hence $\mu(C) = \mu(C) \int \varphi dP_C + \mu(B) \int \varphi dP_B$ and if $\mu(C)$ is close to unity

$$\int \varphi dP_C = 1 - \frac{1-\mu(C)}{\mu(C)} \int \varphi dP_B$$

$$\geq \frac{2\mu(C) - 1}{\mu(C)} \quad ,$$

is also close to unity. Thus if P is a measure such that $\int F_x(C)\,dP$ is small, the test φ provides a test of P against P_C whose level at P is small but power at P_C or P_T is large.

IV. Independent, identically distributed observations.

Let Θ be a Borel subset of a k-dimensional Euclidean space R^k . For each $\theta \in \Theta$ let p_θ be a probability measure carried by a σ-field $\mathcal{A}$ on a set $\mathcal{X}$. For each integer n , let $\{\mathcal{X}^n, \mathcal{A}^n\}$ be the corresponding product space and let $P_{\theta,n}$ be the product measure, product of n copies of p_θ . Let μ be a probability measure on the Borel subsets of Θ . Denote the conditional distribution of θ given $x = (x_1, x_2, \cdots, x_n)$ by the symbol $F_{x,n}$.

Let θ_o be a particular element of Θ and let $p = p_{\theta_o}$ and $P_n = P_{\theta_o,n}$ for simplicity. We shall concern ourselves with the limiting behavior under P_n of the posterior distribution of θ in the above set-up.

Following the notation of section 3, assume that for each $A \in \mathcal{a}$ the function $t \rightsquigarrow p_t(A)$ is measurable and let $P_{C,n}$ be the measure defined by $P_{C,n} = \frac{1}{\mu(C)} \int_C P_{\theta,n}\mu(d\theta)$.

Let $\beta_n(\alpha)$ be the number $\beta_n(\alpha) = \int (dP_{C,n})^{\alpha}(dP_n)^{1-\alpha}$. The following proposition is analogous to Lemma 6.1 in Schwartz [1].

Proposition 3. Assume that there is an integer m and a test function φ defined on $\{\mathcal{X}^n, \mathcal{a}^n\}$ such that $\int \varphi dP_n \leq y_1$ and $\int \varphi dP_{t,n} \geq 1-y_2$ for every $t \in C$. Then for every integer $n \geq m$ and every $\alpha \in (0,1/2]$ one can write

$$\beta_n(\alpha) \leq [s(2-s)]^{\alpha[n/m]}$$

with $s = y_1 + y_2$ and with $[n/m]$ equal to the integer part of n/m .

Proof. Let ν be a probability measure carried by C. Consider the test between P_m and $Q = \int P_{t,m} \nu(dt)$ which minimizes the sum s of the probabilities of error. According to the assumption made $s = ||P_m \wedge Q|| \leq y_1 + y_2$. However $||P_m - Q|| = 2[1-||P_m \wedge Q||] \leq 2\sqrt{1-\rho^2}$ where $\rho = \int \sqrt{dP_m dQ}$. This can be written in the form $\rho^2 \leq s(2-s)$. The logarithmic convexity of the Hellinger transform implies then that

$$\int (dP_m)^{1-\alpha}(dQ)^{\alpha} \leq \rho^{2\alpha} \leq [s(2-s)]^{\alpha}$$

for every $\alpha \in [0,1/2]$.

Suppose that $n \geq m$ is of the form $nk + n'$ for an integer k and an integer $n' \leq m$. An application of Proposition 1 gives

$$\beta_n(\alpha) \leq [s(2-s)]^{\alpha k} \quad .$$

This is the desired result.

The above proposition will allow us to derive exponential bounds for certain posterior probabilities of fixed sets C. One can also use other bounds which can be substituted in Corollary 2 of Proposition 2. One possibility is the following.

Let χ be a measurable map from $\mathcal{X}, \mathcal{A}$ to the r-dimensional Euclidean space. Let H_t be the cumulative distribution of y for p_t. For each integer n let H_n be the empirical cumulative distribution of y. Let $||H_n - H|| = \sup_y (H_n(y) - (H(y)|$.

For $\tau > 0$ let $\psi(\tau)$ be the number $\psi(\tau) = \sup_{n,H} P\{n^{1/2}||H_n - H|| \geq \tau\}$ where the supremum extends to all cumulative distribution functions on R^r. According to a result of J. Kiefer [3] for each $\varepsilon > 0$ there is a constant $c(\varepsilon, r)$ such that $\psi(\tau) \leq c(\varepsilon, r) \exp[-(2-\varepsilon)\tau^2]$.

__Lemma__ 4. _Let_ W _be a neighborhood of_ Θ_o _in_ Θ _and let_ λ _be the_ k-_dimensional Lebesgue measure on_ R^k. _Assume that there is a number_ $b > 0$ _such that for all_ $t \in W$ _one has_ $||H_t - H_{\Theta_o}|| \geq 2b|t - \Theta_o|$. _For each integer_ n _let_ v_n _be a positive number and let_ $\delta_n = n^{-1/2}$. _Let_ φ_n _be the test function equal to unity if_ $||H_n - H_{\Theta_o}|| \leq b v_n \delta_n$ _and to zero otherwise. Let_ $C_n = \{t \in W;\ |t - \Theta_o| \geq v_n \delta_n\}$. _Then, for all_ $t \in C_n$ _and all_ n _one has_ $\int \varphi_n dP_{t,n} \leq \psi[b|t - \Theta_o|\sqrt{n}\,]$ _and_ $\int (1 - \varphi_n) dP_n \leq \psi(v_n)$. _Also_

$$\int_{C_n} [\int \varphi_n dP_{t,n}] \lambda(dt) \leq \delta_n^k \int_{|\tau| \geq v_n} \psi(b\tau) \lambda(d\tau) \quad .$$

Proof. Let S_n be the set of values of $x=(x_1,x_2,\cdots,x_n)$ such that $||H_n-H_{\Theta_o}|| \leq bv_n\delta_n$. If $t \in C_n$ and $x \in S_n$ one can write

$$||H_n-H_t|| \geq ||H_t-H_{\Theta_o}|| - ||H_n-H_{\Theta_o}||$$

$$\geq b|t-\Theta_o| + b[|t-\Theta_o| - v_n\delta_n]$$

$$\geq b|t-\Theta_o| \quad .$$

This implies $P_{t,n}(S_n) \leq \psi[b|t-\Theta_o|\sqrt{n}\,]$ by definition of ψ . The last inequality follows by integration.

In many problems the existence of functions y whose cumulative satifies the conditions of Lemma 4 is obvious. Existence of such transformations can also be proved under a variety of regularity restrictions. A system of conditions insuring the existence of the transformations was given in LeCam [4] (see page 184). They are similar to assumptions which will be stated below.

In order to apply the inequalities of Proposition 2 we need, in addition to the above upper bounds estimates of lower bounds for the densities (dP_V/dP) . The following argument is often usable.

Let α be a number, $\alpha \in (0,1)$ and let

$\omega(\alpha,t) = \int (dp_t)^{\alpha}(dp_{\theta_o})^{1-\alpha}$. Since we have assumed only the measurability of the maps $t \rightsquigarrow p_t(A)$ it is conceivable that $t \rightsquigarrow \omega(\alpha,t)$ might not be measurable. We do not know what the situation is in general. However $t \rightsquigarrow \omega(\alpha,t)$ is certainly measurable if either $\mathcal{Q}$ is countably generated or if the maps $t \rightsquigarrow p_t$ are strongly measurable.

Lemma 5. For every neighborhood V of θ_o such that $\mu(V) \geq 0$ and every $\alpha \in (0,1)$ for which $t \rightsquigarrow \omega(\alpha,t)$ is measurable on V one has

$$\varphi_n(\alpha) \geq \frac{1}{\mu(V)} \int [\omega(\alpha,t)]^n \mu(dt)$$

$$\geq \{\frac{1}{\mu(V)} \int \omega(\alpha,t)\mu(dt)\}^n \quad .$$

Proof. Let ν be the probability measure defined by $\mu(V)\nu(dt) = \mu(dt)$ on V. According to Lemma 1

$$\int (dP_{V,n})^{\alpha}(dP_n)^{1-\alpha} \geq \int \{\int (dP_{t,n})^{\alpha}(dP_n)^{1-\alpha}\}\nu(dt) \quad .$$

In addition

$$\int (dP_{t,n})^{\alpha}(dP_n)^{1-\alpha} = [\omega(\alpha,t)]^n \quad .$$

The result follows by application of Hölder's inequality.

Let $J(t)$ be the "information number"

$$J(t) = \int [\log \frac{dp_t}{dp_{\theta_o}}] dp_{\theta_0} \quad .$$

This is equal to the derivative $\frac{d\omega(\alpha,t)}{d\alpha}$ at $\alpha = 0$ if this derivative exists. Thus the convexity of $\log \omega(\alpha,t)$ in α implies $\omega(\alpha,t) \geq \exp[\alpha J(t)]$. One can then deduce from this and from Lemma 5 that if $J(t) \geq -b$ for all $t \in V$ then $\varphi_n(\alpha) \geq \exp[-n\alpha b]$. This result can be used to reconstruct the proof given by Schwartz in [1] (Theorem 6.1). However, in many respects the bound $J(t) \geq -b$ is an unnatural requirement since it is possible to obtain the same type of exponential bounds in stiuations where $J(t)$ is always $-\infty$.

Let us consider now a more restricted situation describable as follows.

(A1) <u>The point</u> θ_o <u>is an interior point of</u> Θ

(A2) <u>the prior measure</u> μ <u>has, with respect to the k-dimensional Lebesgue measure</u> λ , <u>a continuous</u> strictly positive density.

(A3) <u>Let</u> $h(s,t)$ <u>be the Hellinger distance defined by</u> $h^2(s,t) = \int | \sqrt{dp_s} - \sqrt{dp_t} |^2$. <u>Then for</u> $|\theta - \theta_o| \leq \varepsilon_0$

$$\limsup_{|t|\to 0} \frac{1}{|t|} \; h(\theta+t,t) = \sigma^2(\theta) \le \sigma_2^2 < \infty \quad .$$

Theorem 1. Let (A1), (A2), and (A3) be satisfied. Let W be a neighborhood of the point θ_o and let C be its complement. Assume that there is a uniformly consistent test of θ_o against C. Then there is a neighborhood W_1 of θ_0 a constant K and a number $\gamma < 1$ such that

$$P_{\theta,n}\{|F_x(C)| \ge \gamma^n\} \le K\gamma^n$$

for every $\theta \in W_1$.

Proof. Let m be so large that there is a test φ defined on $\{\mathcal{X}^m, \mathcal{A}^m\}$ such that $16 \int (1-\varphi)\, dP_{\theta_{o,m}} \le 1$ and $8 \int \varphi dP_{t,m} \le 1$ for all $t \in C$. Let $p(t,\theta)$ be the affinity $\rho(t,\theta) = \int [dp_\theta dp_t]^{1/2}$. Select an open set W_1 such that $\theta_o \in W_1$ and such that $\rho^{2m}(t,\theta_o) \ge 1 - \frac{1}{1024}$ for all $t \in W$. The bound chosen on $\rho(t,\theta_o)$ is such that $||P_{m,t} - P_{m,\theta_o}|| \le 1/16$. Therefore, we have also $8 \int (1-\varphi)\, dP_{t,m} \le 1$ for all $t \in W_1$.

Let $\beta_n(\alpha,\theta)$ be the number

$$\beta_n(\alpha,\theta) = \int (dP_{C,n})^\alpha (dP_{\theta,n})^{1-\alpha}$$

for $\theta \in W_1$ and $\alpha \in [0,1/2)$. According to Proposition 3 one can write

$$\beta_n(\alpha,\theta) \leq K_1\gamma_1^{2n\alpha}$$

for $\gamma_1 = 2^{-1/2m}$ and $K_1 = 2^{1/m}$ and all $\theta \in W_1$.

Suppose that W_1 has been chosen such that if $\theta \in W_1$ then $2|\theta-\theta_o| < \varepsilon_0$ so that assumption 3 applies to all points within $\varepsilon_0/2$ of W_1 . For each $\theta \in W_1$ and each n select a small neighborhood V_n of θ . For a given $\varepsilon > 0$, the second term in the inequality of Proposition 2 can be evaluated at $\alpha = 1/2$ giving a bound of the type

$$\frac{K_1}{\sqrt{\varepsilon}}\,\gamma_1^n\, e^{a/2}[\mu(V_n)]^{-1/2}$$

where a is an arbitrary positive number, say $a = 2\log 2$. For this term we can obtain a bound decreasing exponentially provided that $\mu(V_n)$ be larger than a term of the type $K_2\gamma_2^{2n}$ with $\gamma_2 > \gamma_1$. For this purpose take for V_n a cube centered at θ with sides equal to $(\gamma_2^{2n})^{1/k}$. Then $\mu(V_n)$ is larger than $K_2^2\,\gamma_2^{2n}$ with a coefficient K_2 depending only on the lower bound of the density of μ on the set $\{\theta: |\theta-\theta_o| < \varepsilon_0\}$ which contains W_1 .

It remains to see that, with this choice of a and

V_n the probability of the set $A_n = \{x: \frac{dP_{V_{n,n}}}{dP_{\Theta,n}} \leq \frac{1}{2}\}$ decreases exponentially.

For this purpose let $\rho(t,\Theta)$ be the affinity defined above. Assumption A3 implies that $h(t,\Theta) \leq \sigma_2|t-\Theta|$ or equivalently $\rho^2(t,\Theta) = 1 - \frac{1}{2}\sigma_2^2|t-\Theta|^2$ for $\Theta \in W_1$ and $|t-\Theta| \leq \varepsilon_0/2$. According to the definition of V_n this gives

$$\int (dP_{Vn,n})^{1/2}(dP_\Theta)^{1/2} \geq [1-K_3\ \gamma_3^n]^n$$

with $K_3 = \frac{1}{2}\sigma_2^2\ 2^{1/k}$ and $\gamma_3 = \gamma_2^{2/k}$. Thus, for n large enough to insure that $K_3\gamma_3^n < 1/2$ this gives

$$\varphi_n = \int [dP_{V_{n,n}}\, dP_\Theta]^{1/2} \geq \exp[-\ n\ K_3\gamma_3^n(1-k_3\gamma_3^n)]$$

$$\geq \exp[-\ \tfrac{1}{2}\ n\ K_3\gamma_3^n]\quad .$$

It follows that $1-\varphi_n^2 \leq 1 - \exp(-\ n\ K_3\gamma_3^n) \leq n\ K_3\gamma_3^n$.

Take a value γ_4 such that $\gamma_3 < \gamma_4 < 1$. Then $n(\gamma_3/\gamma_4)^n$ satys bounded and therefore $(1-\varphi_n^2) \leq K_4\gamma_4^n$. This gives, according to Lemma 3, the inequality

$$P_{\Theta,n}[F_x(C) \geq (\frac{\gamma_1}{\gamma_2})^{2n/3}] \leq 4\ K_4\ \gamma_4{}^{n/2} + 2\,\frac{K_1}{K_2}\,(\frac{\gamma_1}{\gamma_2})^{2n/3}\quad .$$

Ths choice of γ_2 being unhampered except for the restrictions $\gamma_2 > \gamma_1$ and $\gamma_2 < \gamma_4^{k/2} < 1$ one can choose γ_2 such that $\gamma_2^{2(2k+1)} = \gamma_1^{k}$ and then γ_4 slightly larger than $\gamma_1^{\frac{1}{4(2k+1)}}$. With this choice

$$P_{\theta,n}[F_x(C) \geq \gamma^n] \leq K \gamma^n$$

for any γ such that $\gamma^{4(2k+1)} > \gamma_1$.

This concludes the proof of the theorem.

Remark 1. It may be worth noting that, for a given prior μ, the rate of convergence obtained above depends only on the bound of Assumption A3 and on the integer m necessary to achieve a suitable test of $P_{\theta_{o,n}}$ against $\{P_{t,n}, t \in C\}$. The neighborhood W_1 depends only on that integer m , the bound of Assumption 3 and the range in which this bound is valid.

Remark 2. One could also obtain a similar result using Lemma 5 and instead of (A3) the following assumption.

(A_4') There is an $\epsilon_0 > 0$ such that if $|\theta - \theta_o| < \epsilon_0$ the quantities

$$J(t,\theta) = \int [\log \frac{dp_t}{dp_\theta}] dp_\theta$$

tend to zero uniformly as $t \to \theta$.

A proof can be carried out following the lines of the proof of Theorem 6.1 in [1].

Note that (A3) and (A_4') are very different. Neither one implies the other. It is possible to obtain the result of Theorem 1 under a much weaker assumption which is very much weaker than either (A3) or (A_4').

A corollary relative to the behavior of Bayes estimates can be phrased as follows.

Assume that (A1)(A2) are satisfied and that either (A3) or (A_4') are also satisfied. Let δ be a positive number and let C_δ be the set of points which are at distance at least 2δ from W. Let ℓ be the loss function obtained by taking $\ell(\theta, t) = 1$ if $|\theta - t| \geq \delta$ and $\ell(\theta, t) = 0$ otherwise. Let $\tilde{\theta}_n$ be a Bayes estimate for this system. Then there are coefficients K and γ, $\gamma \in (0,1)$ such that $P_{t,n}[\overset{\infty}{\theta} \in C_\delta] \leq K \gamma^n$ for all $t \in W_1$.

One can obtain further refinements of the above results if further assumptions are imposed on the family $\{p_\theta;\ \theta \in \Theta\}$. For simplicity let us add the following assumption.

(A4) <u>All the</u> p_t , $t \in W$ <u>are mutually absolutely continuous.</u>

Let then $X_\Theta(t)$ be the process obtained by substituting in $(dp_t/dp_\Theta)^{1/2}$ a variable whose distribution is p_Θ .

(A5) <u>The process</u> $t \rightsquigarrow X_\Theta(t)$ <u>is differentiable in quadratic mean at</u> $t=\Theta$ <u>and the derivative</u> Y_Θ <u>has a covariance matrix</u> $A(\Theta) = E_\Theta Y_\Theta Y_\Theta^T$ <u>which is non singular</u>

If (A1) to (A5) are satisfied, one may assume without loss of generality that $\{\mathcal{X},\mathcal{A}\}$ where the variable x takes its values is the interval [0,1] itself. Also, letting H_t be the cumulative distribution of x for p_t there is for each $\Theta \in W$ an interval $|t-\Theta| \le \varepsilon(\Theta)$ and a number $b(\Theta)$ for which the assumption $||H_t-H_\Theta|| \ge 2b(\Theta)|t-\Theta|$ of Lemma 4 is satisfied. (See [4]).

According to Lemma 4 for $\Theta \in W$ and for $b = b(\Theta)$ there is a test φ_n of Θ against the set $C_n = \{|t-\Theta| \ge v_n\delta_n\}$; $\delta_n\sqrt{n} = 1$ such that $\int(1-\varphi_n)\,dP_{\Theta,n} \le K \exp[-b^2v_n^2]$ and

$$\int \varphi_n dP_{Cn,n} \le K\, \delta_n^k \frac{1}{v_n^{\,k-2}} \exp[-b^2v_n^2]$$

Take $v_n^2 = q^2 \log n$ for some constant q . The second quantity written above takes the form

$$K[q^2 \log n]^{-\frac{(k-2)}{2}} \; n^{-[\frac{k}{2} + b^2 q^2]} .$$

Thus if we take for V_n a set of the type $V_n = \{t; |t-\Theta| \leq \frac{\varepsilon}{n}\}$ and apply the same argument as in Theorem 1 we obtain the following result.

Theorem 2. Let conditions (A1) to (A5) be satisfied. Let Θ be an element of W and let C_n be the set C_n be the set $C_n = \{|t-\Theta| \geq q \frac{\log n}{\sqrt{n}}\} \cap W$.

For each $\Theta \in W$, each integer r there is a number q and a constant K such that

$$P_{\Theta,n}[F_x(C_n) > \frac{1}{n^r}] \leq K \frac{1}{n^r}$$

It is sufficient for this purpose to take q so that b^2q^2 is slightly larger than $(r-\frac{1}{2})k+1$.

Unfortunately the technique of proof uses the lemma of [4] which does not seem to lend itself easily to a strengthening which would be uniform on an open subset of W , although this is probably possible to obtain.

Of course, under assumptions (A1) to (A5) the usual

Bernstein-Von Mises theorem is still available. The posterior distribution of Θ, differs little, in L_1-norm, from a suitable Gaussian distribution. The rate of convergence in this approximation seems to depend mostly on two factors.

One of them is the rate of decreases in L_1-mean for μ of the remainder term in the first order Taylor expansion of the process X_Θ. That is, of the rate of convergence to zero of

$$\int_{-\varepsilon}^{\varepsilon} \frac{1}{|\xi|} \, ||X_\Theta + \xi) - X_\Theta(\Theta) - \xi Y_\Theta|| d\xi$$

as $\varepsilon \to 0$. The other factor is the rapidity of convergence to normality of the distribution of $\frac{1}{\sqrt{n}} \sum_j^n Y_\Theta(x_j)$. We hope to give results of this nature in a further paper.

References

[1] Lorraine Schwartz, "On Bayes procedures," Zeitschrift für Wahrscheinlichkeitstheorie u. v. G. Vol. 4, pp. 10-26 (1965).

[2] C. Kraft, "Some conditions for consistency and uniform consistency of statistical procedures," Univ. of Calif. Publ. Statistics, Vol. 2, pp. 125-142 (1955).

[3] J. Kiefer, "On large deviations of the empiric d.f. of vector chance variables and a law of the iterated logarithm," Pacific Journal of Math., Vol. 11, pp. 649-660 (1961).

[4] L. LeCam, "Likelihood function for large numbers of independent observations," Festschrift for J. Neyman, F. N. David editor, Wiley and Sons, N. Y. 1966, pp. 167-187.