Z. Wahrscheinlichkeitstheorie 4, 10—26 (1965)

On Bayes Procedures

By

LORRAINE SCHWARTZ †*

Summary

A result of DOOB regarding consistency of Bayes estimators is extended to a large class of Bayes decision procedures in which the loss functions are not necessarily convex. Rather weak conditions are given under which the Bayes procedures are consistent. One set involves restrictions on the a priori distribution and follows an example in which the choice of a priori distribution determines whether the Bayes estimators are consistent. Another example shows that the maximum likelihood estimators may be consistent when the Bayes estimators are not. However, the conditions given are of an essentially weaker nature than those established for consistency of maximum likelihood estimators.

1. Introduction

This is a contribution to the study of the asymptotic behavior of Bayes decision procedures. The main results include conditions which imply consistency of a class of Bayes procedures.

In 1949 DOOB [*3*] published a rather surprising and fundamental result regarding consistency of Bayes estimators. Roughly speaking, DOOB shows that under very weak measurability assumptions, for every a priori distribution λ on the parameter space Θ the Bayes estimators are consistent except possibly for a set of values θ in Θ having λ measure zero. DOOB's results are summarized and extended in several directions here in section 3. These results carry over to a large class of Bayes procedures in decision problems in which the loss functions are not necessarily convex. This is established in section 4.

In 1953 LECAM [*8*] (see also [*9*]) gave some conditions under which Bayes estimators are consistent, at least for suitable a priori distributions. However, these arose in connection with maximum likelihood estimators and are stronger than his conditions for consistency of the latter. In view of DOOB's results and the nature of maximum likelihood estimators it would seem reasonable to expect that conditions for consistency of Bayes estimators might be found which would be essentially weaker than those for maximum likelihood estimators. In fact conditions given in section 6 of the present paper, though not comparable with WALD's [*11*] or LECAM's conditions for consistency of maximum likelihood estimators, are of an essentially weaker nature.

FREEDMAN [*4*] very recently published results on the problem when the sample space is discrete and in this paper he proves that in the case of independent identically distributed variables taking on only finitely many values the Bayes estimators are consistent and asymptotically normal. However he gives an example in which the set of possible values of the random variables is countable and Doob's

* This paper was prepared with the partial support of the Office of Ordnance Research, U. S. Army, under Contıact DA-04-200-ORD-171. While the paper was in press, its author Dr. LORRAINE SCHWARTZ suddenly died. Sit ei terra levis! (L. LC., J. N.)

exceptional set is the complement of a set of the first category. Even in this case, then, conditions must be added to Doob's to ensure consistency.

In section 5 three examples are given which lead to the results in section 6. The first two are rather trivial. One is a case in which Doob's exceptional set has the power of the continuum and the other provides a situation in which the maximum likelihood estimators are consistent but the Bayes estimators are not. The third example, however, satisfies Wald's conditions as well as many other regularity conditions. Here the a priori measure used determines whether the Bayes estimators are consistent.

2. Basic assumptions and definitions

We define the Bayes procedures in the context of Wald's general decision theory. Let Θ be an arbitrary set, the "possible states of nature", and $\mathfrak{B}$ a σ-field of subsets of Θ. A set $\varDelta$ of available decisions together with a σ-field $\mathfrak{C}$ on $\varDelta$ is given. A real valued function w is defined on $\varDelta \times \Theta$. Assume w is $\mathfrak{C} \times \mathfrak{B}$-measurable. This is the loss function for the problem and a value $w(t, \theta)$ of it is interpreted as the amount lost when the statistician chooses $t \in \varDelta$ and $\theta \in \Theta$ is the "true state of nature". Let $\mathfrak{X}$ be given together with a σ-field $\mathfrak{A}$ on $\mathfrak{X}$ and let X be a random variable with range $\mathfrak{X}$. To each $\theta \in \Theta$ let correspond a probability measure P_θ on $\mathfrak{A}$ so that Θ is the index set for a given subset $\mathscr{D} = \{P_\theta, \theta \in \Theta\}$ of the family $\mathscr{P}$ of all probability measures on $\mathfrak{A}$. It will be convenient to give a structure to Θ which we shall assume in all sections except for sections 3 and 4; namely that Θ is homeomorphic to a subset of the infinite dimensional cube K with sides $J = [0, 1]$. Take for the Topology on K the one induced by the metric $\delta(x, y) = \sum_{k=1}^{\infty} \frac{1}{2^k} |x_k - y_k|$ where $x = (x_1, x_2, \ldots)$ and $y = (y_1, y_2, \ldots)$ belong to K. Since we will be concerned only with convergence properties we may without loss of generality appeal to a Slutsky type of argument and act as though Θ were in fact a subset of the cube. In this case it is no loss to assume that the loss function w is non-negative and bounded above by one. Note that under the assumption that $\Theta \subset K$, the a priori distributions automatically possess moments of all orders.

A space $\mathscr{T}$ of decision procedures is given, each element of which is a function associating with every $x \in \mathfrak{X}$ a probability distribution F_x on $\mathfrak{C}$. Thus if x is chosen,a solution to the decision problem will be given once an element $T \in \mathscr{T}$ is specified and $t \in \varDelta$ is chosen according to the distribution F_x given by $T(x)$. We assume $\mathscr{T}$ is the set of procedures for which $F_.(C)$ is $\mathfrak{A}$-measurable for each $C \in \mathfrak{C}$.

Let λ be a probability measure on $\mathfrak{B}$. For fixed λ the Bayes procedures are defined through functions W and R, on $\mathfrak{X} \times \Theta$ and $\mathscr{T}$ respectively, where

$$W(T(x), \theta) = \int_{\varDelta} w(t, \theta) F_x(dt),$$

$$R(T, \theta) = \int_{\mathfrak{X}} W(T(x), \theta) P_\theta(dx)$$

and

$$R(T) = \int_{\Theta} R(T, \theta) \lambda(d\theta).$$

$R(T, \theta)$, the expected loss when T is chosen and θ is "true", is called the risk function.

Definition 1. *A procedure $\beta \in \mathscr{T}$ is called Bayes for the problem specified by (w, λ) if $R(\beta) = \inf_{T \in \mathscr{T}} R(T)$.*

In most of what follows, we will be concerned with non-randomized procedures, those whose values are, for each $x \in \mathfrak{X}$, distributions assigning their total mass to a single point $t \in \Delta$. In this case it is convenient to equate the procedure T to the function on $\mathfrak{X}$ whose values are the corresponding points in Δ. Then we shall write $T(x) = t$.

To define consistency we shall need a sequence of decision problems. Let $\{\Theta, \mathfrak{B}\}$, $\{\Delta, \mathfrak{C}\}$, $\{\mathfrak{X}, \mathfrak{A}\}$ and w be fixed as above. Let $\{\mathfrak{A}_n\}$ be an increasing sequence of sub σ-fields of $\mathfrak{A}$, $\mathfrak{A}_n \subset \mathfrak{A}_{n+1}$, and assume that $\mathfrak{A}$ is generated by $\{\mathfrak{A}_n\}$. For each $n = 1, 2, 3, \ldots$ and for each P in the family $\mathscr{P}$ of probability measures on $\mathfrak{A}$, let P_n be the restriction of P to $\mathfrak{A}_n$. Also for each n, $\mathscr{T}_n$ is the space of decision procedures defined on $\mathfrak{X}$ and such that $F.(C)$ is $\mathfrak{A}_n$-measurable for each $C \in \mathfrak{C}$.

Let $\mathfrak{T}$ be an arbitrary family of subsets $F \subset \Theta$.

Definition 2. *We shall say that the sequence $\{T_n\}$ of decision procedures is weakly $\mathfrak{T}$-consistent if for every $F \in \mathfrak{T}$ and every $\varepsilon > 0$,*

$$\sup_{\theta \in F} P_\theta\{x \in \mathfrak{X} : W(T_n(x), \theta) - \inf_{t \in \Delta} w(t, \theta) > \varepsilon\} \to 0 \quad \text{as} \quad n \to \infty .$$

$\{T_n\}$ is strongly $\mathfrak{T}$-consistent if for every $F \in \mathfrak{T}$ and $\varepsilon > 0$,

$$\sup_{\theta \in F} P_\theta\{x \in \mathfrak{X} : \sup_{n \geqq N} [W(T_n(x), \theta) - \inf_{t \in \Delta} w(t, \theta)] > \varepsilon\} \to 0 \quad \textit{as} \quad N \to \infty .$$

We will be concerned mainly with the case in which each F contains a single point. In this case the definition reduces to the usual convergence "in P_θ probability" and "with P_θ probability 1", respectively, of $W(T_n, \theta)$ to $\inf_{t \in \Delta} w(t, \theta)$ for each θ in $\bigcup_{F \in \mathfrak{T}} F$. If in addition $\bigcup_{F \in \mathfrak{T}} F = \Theta$ then we say that $\{T_n\}$ is consistent.

Let $\mathfrak{d}$ be a function defined on Θ and taking values in the cube K. Unless the contrary is explicitly specified we shall take as Bayes estimators of $\mathfrak{d}$ the sequence of conditional expectations $E(\mathfrak{d} \mid \mathfrak{A}_n)$, $n = 1, 2, \ldots$. This definition agrees with the one given above when the loss function w is of a suitable quadratic nature. The use of more general loss functions satisfying suitable regularity requirements introduces no essentially new difficulties as will be indicated in section 4.

We shall always assume that the sequence of problems under consideration corresponds to an increasing sequence of σ-fields $\{\mathfrak{A}_n\}$. However most of the results obtained in sections 5 and 6 are derived under the assumption that the random variables involved are independent and identically distributed. Under this assumption $\mathfrak{X}$ will be the infinite product of copies of a set $\mathfrak{X}$, on which a σ-field $\mathfrak{A}^1$ is given. X will be the vector $(Z_1, Z_2, \ldots)$ whose coordinates are completely independent and have the same distribution P^1 defined in terms of P as follows. Put $\mathfrak{A}_1 = \mathfrak{A}^1 \times \mathfrak{X}$ and for each positive integer n let $\mathfrak{X}^n$ be the product of n copies of $\mathfrak{X}_1$, $\mathfrak{A}^n$ the σ-field on $\mathfrak{X}^n$ generated by rectangles with sides in $\mathfrak{A}^1$. Then the $\{\mathfrak{A}_n\}$ will be defined as $\mathfrak{A}_n = \mathfrak{A}^n \times \mathfrak{X}$, $\mathfrak{A}$ as the σ-field on $\mathfrak{X}$ generated by the $\{\mathfrak{A}_n\}$ and the distribution P_1 of each Z_n will be given by $P^1(A) = P_1(A \times \mathfrak{X})$ for each $A \in \mathfrak{A}^1$ where, as before, P_1 is the restriction of P to $\mathfrak{A}_1$ for some P in $\mathscr{D}$. With regard to notation, when subscripts are involved it will be convenient to use P^n

to represent either the product measure on the n dimensional sets $\mathfrak{A}^n$ corresponding to P^1 on $\mathfrak{A}^1$ or the restriction P_n of the measure P to $\mathfrak{A}_n$. This should be clear from the context.

3. Doob's theorems on consistency of Bayes estimators

Assume throughout this section that λ is any probability measure on $\{\Theta, \mathfrak{B}\}$ which possesses finite first and second moments.

Roughly speaking Doob's main theorem, theorem 3.2 below, says that Bayes estimators are strongly consistent a. e. (λ); that is, there is a set B having λ measure zero such that the Bayes estimators are consistent for all $\theta \in \Theta$ not belonging to B.

Consider first the independent identically distributed case as defined in section 2. The assumptions of section 2 are to hold with the exception that Θ is as yet arbitrary. In what follows we shall make use of the assumptions:

A 1 *$\{\mathfrak{X}^1, \mathfrak{A}^1\}$ and $\{\Theta, \mathfrak{B}\}$ are both isomorphic to Borel sets in a complete separable metric space.*

A 2 *For every $A \in \mathfrak{A}_1$, $P_{\cdot}(A)$ is a $\mathfrak{B}$-measurable function.*

A 3 *If $\theta_1 \neq \theta_2$ there exists a set $A \in \mathfrak{A}_1$ for which $P_{\theta_1}(A) \neq P_{\theta_2}(A)$.*

A 4 *There exists an $\mathfrak{A}$-measurable function f on $\mathfrak{X}$ such that $f(x) = \theta$* a. e. (P_θ) *for each $\theta \in \Theta$.*

Theorem 3.1 (Doob): *Conditions* A 1, A 2 *and* A 3 *imply* A 4.

Theorem 3.2 (Doob): *If* A 1, A 2 *and* A 3 *hold then the Bayes estimators of the identity map $\mathfrak{d} : \Theta \to \Theta$ are strongly consistent a. e. (λ).*

Theorem 3.2 is an immediate consequence of Theorem 3.1 and the fact that the Bayes estimators form a martingale sequence. The argument is this. If $\Omega = \Theta \times \mathfrak{X}$, μ the measure on $\mathfrak{B} \times \mathfrak{A}$ determined by $\{P_\theta\}$ and λ then, writing $\mathfrak{d}(\omega) = \mathfrak{d}(\theta, x) = \theta$ and $\beta_n(\omega) = \beta_n(x) = E(\mathfrak{d} \mid z_1, \ldots, z_n)$, the $\{\beta_n\}$ form a martingale sequence, the martingale convergence theorem applies and

$\beta_n(\omega) \to E(\mathfrak{d} \mid z_1, z_2, \ldots) = E(\mathfrak{d} \mid x)$ as $n \to \infty$, a.e. (μ). Theorem 3.1 provides the final step, that $E(\mathfrak{d} \mid x) = \mathfrak{d}$ a.e. (μ) because by A4, $\int_C \mathfrak{d}\, d\mu = \int_C f\, d\mu$ for all $C \in \mathfrak{B} \times \mathfrak{A}$ so that $\mathfrak{d} = f$ a.e. (μ). $\mathfrak{d}$ is then equivalent to an $\mathfrak{A}$-measurable function so that $E(\mathfrak{d} \mid x) = \mathfrak{d}$ a.e. (μ). The theorem is proved because if $C = \{\omega : \beta_n(\omega) \to \mathfrak{d}(\omega)\}$ and $A_\theta = \{x : \beta_n(x) \to \theta\}$, then

$$1 = \mu(C) = \int_\Theta \int_{A_\theta} dP_\theta \lambda(d\theta) = \int_\Theta P_\theta(A_\theta) \lambda(d\theta)$$

and

$$P_\theta\{\beta_n \to \theta\} = 1 \text{ a.e. } (\lambda) .$$

A result similar to theorem 3.2 but valid for arbitrary systems $\{\mathfrak{X}, \mathfrak{A}\}$ may be obtained by using A4 as a condition, thus bypassing theorem 3.1. Such an extension may be of interest for applications to the theory of inference for stochastic processes. The same considerations as those for theorem 3.2 prove theorem 3.3.

Theorem 3.3: *If $\mathfrak{d}:\Theta \to K$ is a $\mathfrak{B}$-measurable map to the cube K and if* A2 *and* A4 *are satisfied, then the Bayes estimators of $\mathfrak{d}$ are strongly consistent a.e. (λ).*

Letting Θ be arbitrary but restricting the σ-field $\mathfrak{B}$ on Θ and continuing with the independent, identically distributed case, one can prove the following result as a consequence of results of LE CAM [9]. Assume that $\mathfrak{B}$ is the completion for λ of the σ-field generated by the functions $\{P_\theta^1(A), A \in \mathfrak{A}_1\}$ on Θ.

Theorem 3.4 (LE CAM): *If $\mathfrak{d}:\Theta \to K$ is $\mathfrak{B}$-measurable, then the Bayes estimators of $\mathfrak{d}$ are strongly consistent a.e. (λ).*

Proof: For all functions $f:\Theta \to R_1$ which are equivalent for λ to $\mathfrak{B}$-measurable and λ-integrable functions, define an index of approximation

$$\alpha_n(f) = \inf_{h_n \in \mathfrak{H}} \int\int |f(\theta) - h_n(x)|\, P_\theta(dx)\,\lambda(d\theta)$$

where $\mathfrak{H}$ is the space of $\mathfrak{A}_n$-measurable real valued functions. Then lemma 1 of [9] says that the space of "accessible" functions, i.e., functions f for which $\alpha_n(f) \to 0$ as $n \to \infty$, is the space of functions equivalent to λ-integrable functions which are measurable for some sub σ-field $\mathfrak{B}' \subset \mathfrak{B}$.

In particular, for each $A \in \mathfrak{A}'$, the function $P_\theta(A)$ is accessible since it is equivalent to the limit of $\mathfrak{A}_n$-measurable functions h_n defined by.

$h_n(x) = 1/n$ (number of coordinates $z_1, z_2, \ldots, z_n$ which are in A). Hence by lemma 1 of [9], for each accessible function f there exists an $\mathfrak{A}$-measurable function h such that

$$\int |f(\theta) - h(x)|\, P_\theta(dx)\,\lambda(d\theta) = 0\,. \tag{1}$$

By the martingale property of the sequence $\{E(f\,|\,\mathfrak{A}_m)\}$ and by (1), the result follows.

Thus far we have been primarily concerned with the independent identically distributed case. If $\{\mathfrak{A}_n\}$ is an increasing sequence, measurability assumptions replacing A2 and A4 imply the conclusion of theorem 3.3.

Let $\{\mathfrak{X}, \mathfrak{A}\}$ be a measurable space. Let $\{\mathfrak{A}_n\}$ be an increasing sequence of sub σ-fields of $\mathfrak{A}; \mathfrak{A}_n \in \mathfrak{A}_{n+1}$ and $\mathfrak{A}_n \to \mathfrak{A}$. If f is a function to Θ, write $f^{-1}(\mathfrak{B})$ for the inverse image of $\mathfrak{B}$ under f.

Theorem 3.5: *If $P.(A)$ is $\mathfrak{B}$-measurable for every $A \in \mathfrak{A}$ and if there exists a function f on $\mathfrak{X}$ such that $\theta = f(x)$ a.e. (P_θ) and such that $f^{-1}(\mathfrak{B}) \subset \mathfrak{A}$, then the Bayes estimators of $\mathfrak{d} = f^{-1}$ are strongly consistent a.e. (λ).*

4. Extension of Doob's results to a class of procedures

By the definition in section 2, $\beta_n \in \mathscr{T}_n$ is a Bayes procedure for (λ, w) if for each $x \in \mathfrak{X}$ it prescribes a probability measure F_x on $\mathfrak{C}$ which minimizes the average risk

$$R(T) = \int\int\int w(t, \theta)\, F_x(dt)\, P_\theta^n(dx)\,\lambda(d\theta) \tag{2}$$

over all procedures T in $\mathscr{T}_n$, the integrals being taken over the whole range in each case. The inside integral is $\mathfrak{A}_n \times \mathfrak{B}$-measurable so that $R(T)$ may be written as

$$R(T) = \int\int\int w(t, \theta)\, Q_x^n(d\theta)\, F_x(dt)\, P_n(dx) \tag{3}$$

where Q_x^n is determined a.e. (P_n) by $Q_x^n(B) = E_n(\mathfrak{b} \mid x)$, $\mathfrak{b}$ = set indicator of B, for each $B \in \mathfrak{B}$ and $P_n(A) = \int P_\theta^n(A)\lambda(d\theta)$, $A \in \mathfrak{A}_n$, $n = 1, 2, \ldots$. Now if T_0 minimizes (2) then for all $T \in \mathcal{T}_n$ the inside integral in (3), namely

$$f(x, T) = \int\int w(t, \theta)\, Q_x^n(d\theta) F_x(dt)\,,$$

bears the relationship $f(x, T) \geqq f(x, T_0)$ a.e. (P_n) to $f(x, T_0)$. That is, suppose T_0 minimizes (2). Then $\int\limits_A f(x, T_0) P_n(dx) = \inf\limits_{T \in \mathcal{T}_n} \int\limits_A f(x, T) P_n(dx)$ for all $A \in \mathfrak{A}_n$ because otherwise there would be a set A with $P_n(A) > 0$ and $\int\limits_A f(x, T_0) P_n(dx) > \int\limits_A f(x, T') P_n(dx)$ for some $T' \in \mathcal{T}_n$. But then, since $\mathcal{T}_n$ is convex, $T'' = I_A T' + (1 - I_A) T_0$ would belong to $\mathcal{T}_n$ and would have $R(T_0) > R(T'')$.

Further, if A_T is the exceptional set on which $f(x, T) < f(x, T_0)$ and if $P_n(\bigcup\limits_{T \in \mathcal{T}_n} A_T) = 0$ then T_0 minimizes $f(x, T)$, the inside integral in (3), a.e. (P_n). Finally since $f(x, T_0)$ is an average over Δ, its integrand is for some $t \in \Delta$ less than or equal to $f(x, T_0)$. To summarize, provided $P_n(\bigcup\limits_{T \in \mathcal{T}_n} A_T) = 0$, T_0 minimizes (2) $\Longleftrightarrow$ T_0 minimizes $\int w(t, \theta) Q_x^n\ (d\theta)$a. e. (P_n). We shall take this as a condition in Lemma 4.1 and in what follows.

5. A $P_n(\bigcup\limits_{T \in \mathcal{T}_n} A_T) = 0$.

Lemma 4.1 states that these Bayes procedures are strongly consistent a.e. (λ) for a class of loss functions which include the usual ones in problems of testing and estimation.

Lemma 4.1: *Suppose*

(i) A 5 *and the conditions of theorem 3.2 hold,*

(ii) *for each $\theta \in \Theta$, $w(t, \theta)$ attains its minimum at $t(\theta) \in \Delta$, and*

(iii) *for every $\varepsilon > 0$ and each $\theta \in \Theta$ the sets $B(t, V_\theta, \varepsilon)$ defined by*

$$B(t, V_\theta, \varepsilon) = \{\theta' \in V_\theta : |w(t, \theta') - w(t, \theta)| < \varepsilon\}$$

where V_θ is any open neighborhood of θ, satisfy the two conditions

$$B(V_\theta, \varepsilon) = \bigcap_{t \in \Delta} B(t, V_\theta, \varepsilon) \quad \textit{belongs to } \mathfrak{B} \textit{ and } \quad \lambda(B(V_\theta, \varepsilon)) > 0\,.$$

Then the Bayes procedures for (λ, w) are strongly consistent a. e. (λ).

Proof: The discussion preceding lemma 4.1 establishes that the Bayes procedures β_n correspond to points $\{t_n(x)\}$ in Δ which minimize

$$g_n(t, x) = \int w(t, \theta) Q_x^n(d\theta)$$

a.e. (P_n). Recalling the definition 2 and taking the sets F to be those consisting of single points outside a set having λ measure zero, we need only show that

$w(t_n(x), \theta) \to w(t(\theta), \theta)$ a.e. (P), where $P(A) = \int P_\theta(A)\lambda(d\theta), A \in \mathfrak{A}$.

By condition (iii) and since $g_n(t_n(x), x) = \inf\limits_{t \in \Delta} g_n(t, x)$ we have for any $\varepsilon > 0$,

$$(w(t_n(x), \theta) - \varepsilon)\, Q_x^n(B(V_\theta, \varepsilon)) \leqq g_n(t(\theta), x) \quad \text{a. e.} \quad (P_n)\,. \tag{4}$$

By (i), $Q_x^n(B(V_\theta, \varepsilon)) \to 1$ a.e. (P). On the other hand since $w(t, .)$ is assumed in section 2 to be $\mathfrak{B}$-measurable for each t, (i) also says that the Bayes estimators of

$w\,t(\mathfrak{b})$ are strongly consistent a.e. (λ). That is, the right side of (4) converges to $w(t(\theta), \theta)$ a.e. (P_θ) for almost all $\theta(\lambda)$. Hence lim sup $w(t_n(x), \theta) \leqq w(t(\theta), \theta)$ a.e. (P). Further, since $w(t_n(x), \theta) \geqq w(t(\theta), \theta)$ for all n, so is the limit inferior of the sequence and this proves the lemma.

The next lemma allows us to restrict discussion to the a posteriori distributions $\{Q_x^n\}$ in the study of consistency of Bayes procedures for problems in which the loss functions satisfy the conditions of lemma 4.1.

Lemma 4.2: *If w satisfies the conditions of lemma 4.1; if the Bayes procedures for (λ, w) correspond to the sequence $\{t_n\}$ on $\mathfrak{X}$ and if the distributions $\{Q_x^n\}$ converge a.e. (P_θ) to the indicator of $\{\theta\}$, then $w(t_n, \theta) \to w(t(\theta), \theta)$ a.e. (P_θ). (That is, the* Bayes procedures for (λ, w) are strongly consistent for $\{\theta\}$).

Proof: The inequality (4) follows from the conditions of lemma 4.1. Since $0 \leqq w \leqq 1$ by assumption (sec. 2).

$$\int\limits_{B^c(V_\theta,\, \varepsilon)} w(t(\theta), \theta)\, Q_x^n(d\theta) \leqq Q_x^n(B^c(V_\theta, \varepsilon)) .$$

Also,

$$\int\limits_{B(V_\theta,\, \varepsilon)} w(t(\theta), \theta)\, Q_x^n(d\theta) \leqq (w(t(\theta), \theta) + \varepsilon)\, Q_x^n(B(V_\theta, \varepsilon)) .$$

Add the left sides to get the right side of (4) and from these inequalities, whenever $Q_x^n(B(V_\theta, \varepsilon)) > 0$, (4) gives

$$w(t_n(x), \theta) \leqq 2\,\varepsilon + \frac{Q_x^n(B^c(V_\theta, \varepsilon))}{Q_x^n(B(V_\theta, \varepsilon))} + w(t(\theta), \theta) .$$

Since $w(t(\theta), \theta) \leqq w(t_n, \theta)$ for all n, and by assumption $Q_x^n(B(V_\theta, \varepsilon)) \to 1$ a. e. (P_θ), it follows that $\lim\limits_{n \to \infty} w(t_n(x), \theta) = w(t(\theta), \theta)$.

5. Examples

In this section several examples are given in which Bayes estimators are not consistent, though in each case consistent estimators do exist.

Example 1. Take Θ to be the real line. (Θ is homeomorphic to $(0, 1)$; $f(\theta) = 1/2$ $(\arctan \theta + 1)$, for example, defines a $1 - 1$ bicontinuous map). $\mathcal{D}$ the family of distributions whose restrictions P_θ^1 to $\mathfrak{A}_1$ correspond to the $N(\mathfrak{b}, 1)$ distributions on the line, where $\mathfrak{b}(\theta) = 1 - \theta$ if θ belongs to the Cantor set on $[0, 1]$ and $\mathfrak{b}(\theta) = \theta$ otherwise. The Bayes estimators of the identity map φ on Θ,

$$\beta_n(X) = \frac{1}{n+1} \sum_{i=1}^{n} Z_i ,$$

converge a. e. (P_θ) to $\mathfrak{b}(\theta)$ for every $\theta \in \Theta$. However, consistent estimators of φ do exist by a theorem in [6] since φ is of the 1 st Baire class for the distance $\| P - Q \|$ on $\mathcal{D}$.

$$\| P - Q \| = 2 \sup_{A \in \mathfrak{A}_1} | P(A) - Q(A) | . \tag{5}$$

Example 2. In example 1, the maximum likelihood estimators estimate the same function on Θ as the Bayes estimators do. Examples are easily found in which maximum likelihood estimators are not consistent but the Bayes estimators are. For instance, Bahadur's example 2 in [1] satisfies Doob's conditions and Θ is a countable set. The maximum likelihood estimators below are consistent though consistency in the case of the Bayes estimators depends on the choice of the a priori distribution; it fails for the uniform distribution on Θ.

In this example, $\Theta = [1, 2)$, λ is Lebesgue measure on the Borel sets of Θ. $X = (Z_1, Z_2, \ldots)$ as usual and the distribution of Z_1 has uniform density on $[0, 1)$ if $\theta = 1$ or on $[0, 2/\theta)$ if $1 < \theta < 2$. The maximum likelihood estimator for $\mathfrak{d}(\theta) = \theta$ is for each n, putting $Y_n = \max\limits_{i=1,\ldots,n} Z_i$,

$$\hat{\mathfrak{d}}_n(X) = \begin{cases} 1 & \quad Y_n \leqq 1 \\ & \text{when} \\ 2/Y_n & \quad Y_n > 1 \end{cases}$$

while the Bayes estimator is

$$\beta_n(X) = \begin{cases} \dfrac{n+1}{n+2} \cdot \dfrac{2^{n+2}-1}{2^{n+1}-1} & \qquad Y_n \leqq 1 \\ & \text{when} \\ \dfrac{\int_1^{2/Y_n} \theta^{n+1} d\theta}{\int_1^{2/Y_n} \theta^n d\theta} & \qquad Y_n > 1 \,. \end{cases}$$

So for $\theta = 1$, $\hat{\mathfrak{d}}_n$ is consistent while β_n is not.

It should be noted that $\{\beta_n\}$ would be consistent on Θ if instead of λ one took for the a priori distribution

$$\lambda_1 = \alpha\lambda + (1-\alpha)\lambda_2 \quad \text{where} \quad 0 < \alpha < 1 \quad \text{and} \quad \lambda_2(B) = 1 \quad \text{if}$$
$$B \cap \{1\} \neq \emptyset, \lambda_2(B) = 0 \quad \text{if} \quad B \cap \{1\} = \emptyset \,.$$

Example 3. Wald's conditions for consistency of maximum likelihood estimators in [11] are not satisfied in either of the above examples and it is reasonable to ask whether these conditions would imply consistency of Bayes estimators. The answer is no and example 3 substantiates this. Besides Wald's conditions the class of distributions considered here meet other regularity conditions. In particular, $\| P_\theta^1 - P_{\theta_0}^1 \| \to 0$ as $|\theta - \theta_0| \to 0$ and the densities are continuous. We shall go into some detail here because the arguments used to show the lack of consistency lead directly to the results in section 6. In this example the consistency or lack of it is determined by the a priori distribution.

Let $\Theta = [0, 1/2]$, P_θ is defined through the density f_θ of any one coordinate Z of X,

$$f_\theta(z) = e^{\frac{-1}{z+\theta^2}} I_{[0,\theta)}(z) + [a(\theta)z + b(\theta)]\, I_{[\theta,2\theta)}(z) + C(\theta)\, I_{[2\theta,1]}(z)$$

where I_A is the indicator of A, $C(\theta) = \dfrac{1 - \int_0^{2\theta} f(z)\, dz}{1 - 2\theta}$,

$$a(\theta) = \frac{C(\theta) - e^{-\frac{1}{\theta+\theta^2}}}{\theta} \quad \text{and} \quad b(\theta) = 2e^{-\frac{1}{\theta+\theta^2}} - C(\theta) \,.$$

For $\theta = 0$, $C(\theta)$ is defined by continuity to be 1 and f_θ becomes uniform on [0, 1]. The following properties of these distributions will be used. Put $\theta_0 = 0$. Then

(i) $\| P^1_\theta - P^1_{\theta_0} \|$ is an increasing function of θ and for

$$\theta < 1/4, \; \| P^1_\theta - P^1_{\theta_0} \| < 4\theta .$$

That is,

$$\| P^1_\theta - P^1_{\theta_0} \| = \int\limits_0^\theta \left(1 - e^{-\frac{1}{z+\theta^2}}\right) dz + \int\limits_\theta^{2\theta} | a(\theta) z + b(\theta) - 1 | \, dz + (1 - 2\theta)(C(\theta) - 1)$$

$$< \theta + \frac{\theta}{2}\left[C(\theta) - e^{-\frac{1}{\theta+\theta^2}}\right] + (1 - 2\theta)(C(\theta) - 1) .$$

For $\theta < \frac{1}{4}$, $C(\theta) < \frac{1}{1 - 2\theta}$ so that

$$\| P^1_\theta - P^1_{\theta_0} \| < \theta + \frac{\theta}{2(1 - 2\theta)} + 2\theta < 4\theta .$$

(ii) $H(\theta) = E_{\theta_0} \log \frac{f_\theta(z)}{f_{\theta_0}(z)}$ is an increasing function of θ in a neighborhood of θ_0.

In particular, when $\theta < .2$, $H'(\theta) > 1$. To see this compute

$$H(\theta) = \log \frac{\theta}{1+\theta} + \frac{1}{a(\theta)} \left(g(C(\theta)) - g\left(e^{-\frac{1}{\theta+\theta^2}}\right)\right) + (1 - 2\theta) \log C(\theta)$$

where $g(y) = y \log y - y$. Then

$$H'(\theta) = \frac{1}{\theta(1+\theta)} - \frac{a'(\theta)}{a^2(\theta)} \left(g(C(\theta)) - g\left(e^{-\frac{1}{\theta+\theta^2}}\right)\right) + \frac{1}{a(\theta)} (\log C(\theta) + \frac{1}{\theta+\theta^2} +$$
$$+ \frac{(1 - 2\theta) C'(\theta)}{C(\theta)} - 2 \log C(\theta) .$$

Since $\theta a(\theta) = C(\theta) - e^{-\frac{1}{\theta+\theta^2}}$, $a'(\theta) > -\frac{a(\theta)}{\theta}$ and since $\theta < .2$, $g(C(\theta)) - g\left(e^{-\frac{1}{\theta+\theta^2}}\right)$ is negative and greater than -1. Since $C(\theta)$ increases from 1, the third and fourth terms are positive and the last term is greater than -2. Finally, $a(\theta) > \frac{1-\theta}{\theta}$ so that

$$H'(\theta) > \frac{1}{\theta(1+\theta)} - \frac{\theta}{a(\theta)} - 2 > \frac{1}{\theta(1+\theta)} - \frac{\theta^2}{1-\theta} - 2 > 1 .$$

(iii) Apply LECAM's corollary 4.1 of [8] to get

$$\lim_{k \to \infty} \sup_{\theta \in \Theta} \left| \frac{1}{k} \sum_{i=1}^{k} \log \frac{f_\theta(z_i)}{f_{\theta_0}(z_i)} - H(\theta) \right| = 0 \quad \text{a.e.} \quad (P_{\theta_0}) .$$

Thus for any $\varrho > 0$ and for each x outside of a set having P_{θ_0} probability 0, there is an $N(\varrho, x)$ such that

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \log p^n_\theta(x) - H(\theta) \right| < \varrho$$

or

$$e^{n(H(\theta)) - \varrho)} < p^n_\theta(x) < e^{n(H(\theta) + \varrho)} \tag{6}$$

for $n > N(\varrho, x)$. Here we use the fact that $p^n_{\theta_0}(x) = 1$ on $\mathfrak{X}$.

Let $V = [0, \varepsilon)$, $\varepsilon < .1$, and write the Bayes estimates $\{\beta_n\}$ as

$$\beta_n(x) = b_n^*(x)\, Q_x^n(V) + \bar{b}_n(x)\, Q_x^n(V^c)$$

where b_n^* and $\bar{b}_n$ are the averages over V and V^c, resp. Then if $Q_x^n(V) \to 0$, $\liminf_{n\to\infty} \beta_n(x) \geqq \varepsilon$ and $\{\beta_n(x)\}$ does not converge to 0. We shall show that this happens a. e. (P_{θ_0}).

By (iii) and (ii), for almost all $x(P_{\theta_0})$ and for $n > N(\varrho, x)$,

$$\int_{V^c} p_\theta^n \lambda(d\theta) > \int_{[2\varepsilon, .2]} e^{n(H(\theta)-\varrho)} \lambda(d\theta) > e^{n[H(2\varepsilon)-\varrho]} \lambda([2\varepsilon, .2]) . \tag{7}$$

Let $V_n = \left[0, \frac{1}{n^3}\right]$. For $n > 2$, $\| P_\theta^n - P_{\theta_0}^n \| \leqq n \| P_\theta^1 - P_{\theta_0}^1 \| < 4n\theta$ by (i) so that for any $\delta > 0$,

$$P_{\theta_0}\left\{\left| \frac{1}{\lambda(V_n)} \int_{V_n} \frac{p_\theta^n}{p_{\theta_0}^n} \lambda(d\theta) - 1 \right| > \delta\right\}$$

$$< \frac{1}{\delta\lambda(V_n)} \int_{V_n} \int_{\mathfrak{X}} \left| \frac{p_\theta^n}{p_{\theta_0}^n} - 1 \right| dP_{\theta_0}^n\, \lambda(d\theta)$$

$$< \frac{4n}{\delta\lambda(V_n)} \int_{V_n} \theta\, \lambda(d\theta) < \frac{4}{\delta n^2} .$$

For A_n the set in brackets,

$$P_{\theta_0}\left(\bigcup_{k=N}^{\infty} A_k\right) \leqq \frac{4}{\delta} \sum_{k=N}^{\infty} \frac{1}{k^2} \to 0 \quad \text{as} \quad N \to \infty .$$

This gives convergence to 1 a.e. (P_{θ_0}) but also for

$$x \in \left(\bigcap_{N=1}^{\infty} \bigcup_{k=N}^{\infty} A_k\right)^c, \quad \text{and for some } N_1(\delta, x),$$

$$\lambda(V_n)(1-\delta) < \int_{V_n} p_\theta^n \lambda(d\theta) < (1+\delta)\lambda(V_n) \tag{8}$$

for $n > N_1(\delta, x)$.

Finally, using (iii) again, this time on $V - V_n$,

$$\int_{V-V_n} p_\theta^n \lambda(d\theta) < \int_{V-V_n} e^{n(H(\theta)+\varrho)} \lambda(d\theta) < e^{n(H(\varepsilon)+\varrho)} \lambda(V) . \tag{9}$$

Putting (7), (8) and (9) together, we have

$$Q_x^n(V) = \frac{\int_V p_\theta^n(x)\, \lambda(d\theta)}{\int_\Theta p_\theta^n(x)\, \lambda(d\theta)} < \frac{\int_V p_\theta^n(x)\, \lambda(d\theta)}{\int_{V^c} p_\theta^n(x)\, \lambda(d\theta)} < C_1 \lambda(V_n) e^{n(\varrho - H(2\varepsilon))} + C_2 e^{n[H(\varepsilon) - H(2\varepsilon) + 2\varrho]}$$

where $C_1 = (1+\delta)[\lambda[2\varepsilon, .2])]^{-1}$ and $C_2 = \lambda(V)[\lambda([2\varepsilon, .2])]^{-1}$.

So far ϱ has been arbitrary and so has λ. If $\varrho < \frac{H(2\varepsilon) - H(\varepsilon)}{2}$, then the second term $\to 0$ a. e. (P_{θ_0}). Also, if $(\lambda(V_n))^{1/n} \varepsilon^{\varrho - H(2\varepsilon)} < 1$ then the first term does the same thing. For example, if λ has the density $C_3 e^{-1/(\theta)}$ on $[0, 1/2]$, then $\lambda(V_n) < C_3 e^{-n^3}$ and the first term is less then $C_1 C_3 e^{-(n^2 - \varrho + H(2\varepsilon))}$.

However, if λ were chosen to be uniform on $[0, 1/2]$ then the $\{\beta_n\}$ would be consistent at $\theta_0 = 0$.

6. Conditions for consistency of Bayes procedures

In this section we assume that the conditions of lemma 4.1 are satisfied so that lemma 4.2 applies and we may continue to restrict discussion to the a posteriori distributions Q^n.

If δ is a norm on Θ, Θ compact and $\varphi: P_\theta^1 \to \theta$, then example 3 shows that DOOB's conditions together with the condition that φ and its inverse be continuous for the distances δ and ϱ, $\varrho(P, Q) = \| P - Q \|$, are not sufficient for consistency of Bayes estimators. Neither is the stronger condition of continuity of density functions. We remark here that WALD's conditions imply the existence of uniformly consistent tests of the hypothesis that Z has the distribution $P_{\theta_0}^1$ against the alternative that the distribution is P_θ^1 for some θ in the complement of any open neighborhood of θ_0. The existence of such a test will be one of the conditions in each set of conditions for consistency given in this section. A useful result in this connection is a necessary and sufficient condition due to KRAFT [7]. This and an inequality which we state as lemma 6.1 will be used to establish the theorems which follow.

Let $\{P_\theta, \theta \in \Theta_1\}$ and $\{Q_\theta, \theta \in \Theta_2\}$ be two families of probability measures on $\mathfrak{A}$. On the space of probability measures on $\mathfrak{A}$ define the inner product $\varrho(P, Q) = \int \sqrt{pq}\, d\mu$ where μ is any σ-finite measure with respect to which P and Q are both absolutely continuous and where p and q are the corresponding densities. If there is a set $A \in \mathfrak{A}$ such that $P(A) = 0$ and $Q(A) = 1$ then P and Q are orthogonal. Then ϱ has the following properties:

$$\begin{aligned} &\text{(i)}\ 0 \leqq \varrho \leqq 1 \\ \text{(10)}\quad &\text{(ii)}\ \varrho(P, Q) = 0 \Leftrightarrow P \text{ and } Q \text{ are orthogonal and } \varrho\,(P, Q) = 1 \Leftrightarrow P = Q\,. \\ &\text{(iii)}\ 2(1 - \varrho(P, Q)) \leqq \| P - Q \| \leqq 2\sqrt{1 - \varrho_p^2(P, Q).} \end{aligned}$$

Let $\varrho_n(P, Q) = \varrho(P_n, Q_n) = \int_{\mathfrak{X}} \sqrt{p_n q_n}\, d\mu_n$ where P_n, Q_n, μ_n are measures on $\mathfrak{A}_n \subset \mathfrak{A}$. Let $\mathfrak{M}_1$ and $\mathfrak{M}_2$ be the spaces of all probability measures on Θ_1 and Θ_2, respectively.

Any sequence $\{\varphi_n\}$ of $\mathfrak{A}_n$-measurable functions on $\mathfrak{X}$ with $0 \leqq \varphi_n \leqq 1$, $n = 1, 2, \ldots$ is a *test* of a hypothesis that a probability measure on $\mathfrak{A}$ belongs to a given set against the hypothesis that it belongs to an alternative set. $\{\varphi_n\}$ is *consistent* for the hypothesis $P \in \{P_\theta, \theta \in \Theta_1\}$ against the alternative $P \in \{Q_\theta, \theta \in \Theta_2\}$ if $E_\theta(\varphi_n) \to I_{\Theta_2}(\theta)$, $\theta \in \Theta_1 \cup \Theta_2$. $\{\varphi_n\}$ is *uniformly consistent* if the convergence is uniform on $\Theta_1 \cup \Theta_2$.

The theorem in [7] then says that a uniformly consistent test exists if and only if

$$\sup_{\lambda_1 \in \mathfrak{M}_1, \lambda_2 \in \mathfrak{M}_2} \varrho_n(E_{\lambda_1}(P_{\mathfrak{d}}), E_{\lambda_2}(Q_{\mathfrak{d}})) \to 0 \quad \text{as} \quad n \to \infty, \mathfrak{d}(\theta) = \theta\,.$$

Lemma 6.1 is known in perhaps a variety of different forms. Its proof makes use of inequalities which may be found, for example, in a paper by CHERNOFF [2].

The independent, identically distributed case will be assumed in what follows and the lifting of this restriction will be discussed following the proof of theorem 6.4.

Let $\mathscr{P}_1 = \{P_\theta, \theta \in \Theta_1\}$, $\mathscr{P}_2 = \{P_\theta, \theta \in \Theta_2\}$, $\Theta_1 = \{\theta_0\}$, $\Theta_2 = V_{\theta_0}^c$ where V_{θ_0} is any open neighborhood of θ_0.

Lemma 6.1: *If there is a uniformly consistent test of the hypothesis $P_\theta \in \mathscr{P}_1$ against the alternative $P_\theta \in \mathscr{P}_2$ then there exists a real number $r > 0$ and a positive integer k such that $\| P_{\theta_0}^n - P^n \| \geqq 2(1 - 2e^{-mr})$ where $mk \leqq n < (m+1)k$ and $P^n = E_{\lambda_2}(P_{\mathfrak{b}}^n)$.*

Proof: Let $\{\varphi_n\}$ be the uniformly consistent test assumed to exist. Then there exist $k > 0$ such that $E_{\theta_0}(\varphi_n) < 1/8$ and $E_\theta(\varphi_n) > 1 - 1/8$ for all $\theta \in V_{\theta_0}^c$ and $n \geqq k$. For $j = 1, 2, \ldots$ let $\mathfrak{X}_{k,j} = \{(z_{(j-1)k+1'}, z_{(j-1)k+2'}, \ldots, z_{jk})\}$. On $\mathfrak{X}_{k\ j}$ define random variables $\varphi_{k,j} = \varphi_k(Z_{(j-1)k+1'}, Z_{(j-1)k+2'}, \ldots, Z_{jk})$, $j = 1, 2, \ldots$. Then $Y_m = 1/m \sum_{j=1}^{m} \varphi_{k,j}$ is a sum of independent identically distributed random variables with expectation $E_\theta(Y_m) = E_\theta(\varphi_k)$ and by the strong law of large numbers

$$Y_m \to \begin{cases} E_{\theta_0}(\varphi_k) < \frac{1}{8} & \theta = \theta_0 \\ & \text{for} \\ E_\theta(\varphi_k) > \frac{7}{8} & \theta \in V_{\theta_0}^c \end{cases}$$

a.e. (P_θ).

The argument to prove the lemma depends on the fact that $P_\theta^{mk}\{Y_m \leqq 1/4\}$ decreases to 0 exponentially, uniformly on $V_{\theta_0}^c$. Write $U = \varphi_k - 1/4$ and suppose $t \leqq 0$, t real. Then

$$P_\theta^{mk}\left\{m\left(Y_m - \frac{1}{4}\right) \leqq 0\right\} \leqq E_\theta(e^{tm(Y_m - 1/4)}) = (E_\theta e^{tU})^m .$$

We shall show that for some t and C, $E_\theta e^{tU} \leqq C < 1$. Now e^{tU} is bounded for t in any neighborhood of the origin and $E_\theta e^{tU}$ is continuous in t. By looking at the slope of the curve, $E_\theta U e^{tU}$, it will be seen that in some interval containing 0 $E_\theta e^{tU}$ is strictly increasing to 1 as t increases to 0, whatever be $\theta \in V_{\theta_0}^c$; in fact $E_\theta U e^{tU} > 1/2$ on $V_{\theta_0}^c$ for $0 > t >$ some t_0. That is, at $t = 0$ the slope is positive and the value is 1. For $t < 0$,

$$|E_\theta U e^{tU} - E_\theta U| < E_\theta[|U|\,|e^{tU} - 1|] < E_\theta|e^{tU} - 1| .$$

Also, $|e^{tU} - 1| < \max(1 - e^t, e^{-t} - 1) \leqq \max(|t|, e^{-t} - 1)$ so that for some $t_0 < 0$, $\max(|t_0|, e^{-t_0} - 1) < \varepsilon$ and $|E_\theta U e^{tU} - E_\theta U| < \varepsilon$ if $t_0 \leqq t \leqq 0$. In particular, for $\varepsilon > 1/8$, $E_\theta U e^{tU} > E_\theta U - \varepsilon > 1/2$ for all $\theta \in V_{\theta_0}^c$. It follows that for some $r_1 > 0$, $E_\theta e^{t_0 U} = e^{-r_1} < 1$ and that

$$P^n\left\{Y_m \leqq \frac{1}{4}\right\} = \int_{V_{\theta_0}^c} P_\theta^n\left\{Y_m \leqq \frac{1}{4}\right\} \lambda_2(d\theta) \leqq e^{-mr_1} \quad \text{for} \quad n \geqq mk .$$

By using a similar argument with $t \geqq 0$ an $r_2 > 0$ may be found such that $P_{\theta_0}^n\{Y_m \geqq 1/4\} \leqq e^{-mr_2}$ for $n \geqq mk$. For $r = \min(r_1, r_2)$ it follows that

$$\| P_{\theta_0}^n - P^n \| = 2 \sup_{A \in \mathfrak{A}_n} | P_{\theta_0}^n(A) - P^n(A) | \geqq 2(1 - 2e^{-mr})$$

for $mk \leqq n < (m+1)k$ by considering $A = \{Y_m \leqq 1/4\}$.

For theorems 6.1 through 6.4 we assume the independent identically distributed case and also the existence of a measure with respect to which all the P_θ^n admit densities p_θ^n. As before H is defined by $H(\theta) = E_{\theta_0}\left(\log \frac{p_\theta(z)}{p_{\theta_0}(z)}\right)$.

Theorem 6.1: *Suppose that* (i) *the densities may be chosen* $\mathfrak{B} \times \mathfrak{A}_1$*-measurable.* (ii) $V \in \mathfrak{B}$ *is a neighborhood of* θ_0 *and there is a uniformly consistent test of the hypothesis* $P_\theta = P_{\theta_0}$ *against the alternative* $P_\theta \in \{P_\theta, \theta \in V^c\}$, *and* (iii) *for every* $\varepsilon > 0$ V *contains a subset* W *such that* $\lambda(W) > 0$ *and* $H(\theta) > -\varepsilon$ *on* W. *Then* $Q_x^n(V^c) \to 0$ *a. e.* (P_{θ_0}).

Proof. Let $p_n = \frac{1}{\lambda(V^c)} \int\limits_{V^c} p_\theta^n \lambda(d\theta)$, $q_n = \frac{1}{\lambda(W)} \int\limits_W p_\theta^n \lambda(d\theta)$ and P_n defined by $P_n(A) = \int P_\theta^n(A)\lambda(d\theta), A \in \mathfrak{A}_n$. Then for $\int p_\theta^n(x)\lambda(d\theta) > 0$, $Q_x^n(V^c) \leqq \frac{\lambda(V^c)p_n(x)}{\lambda(W)q_n(x)}$. It will be shown that $p_n(x)/q_n(x) \to 0$ exponentially a. e. (P_{θ_0}).

By lemma 6.1 there exist numbers k and r such that for $mk \leqq n < (m+1)k$, $\| P_{\theta_0}^n - P_n \| \geqq 2(1 - 2e^{-mr})$. Thus by (iii) of (10),

$$P_{\theta_0}\left\{\frac{p_n}{p_{\theta_0}^n} > \varepsilon_n\right\} \leqq \frac{1}{\varepsilon_n} \varrho_n(P_{\theta_0}^1, P_1) \leqq \frac{1}{\varepsilon_m} \sqrt{1 - (1 - 2e^{-mr})^2} = \frac{2e^{-\frac{mr}{2}}}{\varepsilon_n} \sqrt{1 - e^{-mr}}.$$

For $\varepsilon_m = e^{-\frac{mr}{4}}$

$$P_{\theta_0}\left\{\frac{p_n}{p_{\theta_0}^n} > e^{-\frac{mr}{4}}\right\} \leqq 2e^{-\frac{mr}{4}}. \tag{11}$$

For $A_n = \left\{\frac{p_n}{p_{\theta_0}^n} > e^{-\frac{rn}{2k}}\right\}$. A_n is contained in the set on the left side of (11), $2e^{r/2}e^{-\frac{rn}{2k}}$ is greater than the right side and it follows from (11) that $P_{\theta_0}\left(\bigcap\limits_{N=1}^{\infty} \bigcup\limits_{n \geqq N} A_n\right) = 0$. That is, for almost all $x(P_{\theta_0})$ there is an integer $N_1(x)$ such that

$$\frac{p_n(x)}{p_{\theta_0}^n(x)} \leqq e^{-\frac{rn}{2k}} \quad \text{for each } n > N_1(x). \tag{12}$$

To find a bound for $q_n/p_{\theta_0}^n$, define averages $\varphi_n(x, \theta) = \frac{1}{n} \sum\limits_{i=1}^{n} (\log p_\theta^1(z_i) - \log p_{\theta_0}^1(z_i))$. For each θ, $\varphi_n(\cdot, \theta) \to H(\theta)$ a. e. (P_{θ_0}) by the strong law of large numbers. But also by the $\mathfrak{B} \times \mathfrak{A}_n$-measurability of φ_n and by Fubini's theorem there is an x set with P_{θ_0} measure zero such that for all x in its complement $\varphi_n(x, \cdot) \to H$ a. e. (ν), $\nu(B) = \frac{1}{\lambda(W)} \lambda(W \cap B)$, $B \in \mathfrak{B}$. For fixed $\varepsilon > 0$ and W given by condition (iii), an application of Fatou's lemma and a Hölder inequality gives

$$\liminf_{n \to \infty} \int e^{\varphi_n(x,\theta)} \nu(d\theta) \geqq \int e^{H(\theta)} \nu(d\theta) > e^{-\varepsilon}$$

so that for some $N_2(x)$ and each $n > N_2(x)$

$$\int e^{n\varphi_n(x,\theta)}\nu(d\theta) = \frac{q_n(x)}{p_{\theta_0}^n(x)} \geqq e^{-n\varepsilon} \text{ a. e. } (P_{\theta_0}). \tag{13}$$

By (12) and (13)

$Q_x^n(V^c) \leqq \frac{\lambda(V^c)}{\lambda(W)} e^{-n\left(\frac{r}{2k}-\varepsilon\right)}$ for all $n > \max(N_1(x), N_2(x))$ a. e. (P_{θ_0}). The result follows by choosing $\varepsilon < \frac{r}{2k}$.

The second set of conditions for consistency of Bayes procedures for (λ, w), w satisfying the conditions of lemma 4.1, also involves H and the existence of a uniformly consistent test and it follows almost immediately from theorem 6.1.

Theorem 6.2*: *Suppose* (i) *the densities p_θ may be chosen $\mathfrak{B}\times\mathfrak{A}_1$-measurable,* (ii) $H(\theta)\to 0$ *as* $\theta\to\theta_0$, (iii) *for every neighborhood $V\in\mathfrak{B}$ of θ_0 there exists a uniformly consistent test of the hypothesis that $P_\theta = P_{\theta_0}$ against the alternative that $P_\theta\in\{P_\theta,\theta\in V^c\}$. Then for every λ which assigns positive probability to the open sets in Θ the Bayes estimators $\{\beta_n\}$ converge to θ_0 a. e. (P_{θ_0}).*

Proof. Write $\beta_n(x) = b_n^*(x)\,Q_x^n(V) + \bar{b}_n(x)\,Q_x^n(V^c)$ where

$$b_n^* = \frac{1}{Q_\cdot^n(V)} E(\mathfrak{d}\, I_v \,|\, \mathfrak{A}_n)$$

and $\bar{b}_n$ is a similar average over V^c. By assumption, for every $\varepsilon > 0$ there is a neighborhood $W_\varepsilon\subset V$ of θ_0 such that $\lambda(W_\varepsilon) > 0$ and $H(\theta) > -\varepsilon$ on W_ε. By theorem 6.1, $\delta(\beta_n, b_n^*)\to 0$ a. e. (P_{θ_0}). Also V was any neighborhood of θ_0 and $\delta(b_n^*, \theta_0) \leqq \sup\{\delta(\theta,\theta') : \theta\,\varepsilon\, V,\ \theta'\,\varepsilon\, V\}$. This establishes the theorem.

The next two results depend on the local behavior of the a priori measure λ. Let $W_\varepsilon = \{\theta : \| P_\theta' - P_{\theta_0}'\| < \varepsilon\}$.

Theorem 6.3: *Suppose* (i) p_θ^1 *is* $\mathfrak{B}\times\mathfrak{A}_1$-measurable, (ii) *for each neighborhood $V\in\mathfrak{B}$ of θ_0 there exists a uniformly consistent test of the hypothesis $P_\theta = P_{\theta_0}$ against the alternative $P_\theta\in\{P_\theta,\theta\in V^c\}$ and* (iii) *for each V there is a sequence $\{\varepsilon_n\}$ of positive numbers such that $n\varepsilon_n\to 0$ and* $\liminf_{n\to\infty}[\lambda(V\cap W_{\varepsilon_n})]^{1/n} = 1$. *Then $\beta_n\to\theta_0$ in P_{θ_0} probability.*

Further, if $\varepsilon_n \leqq \frac{1}{n^{2+\delta}}$ for some $\delta > 0$ then $\beta_n\to\theta_0$ a. e. (P_{θ_0}).

Proof. The idea of the proof for theorem 6.1 may be used to prove this. Compare the average densities on V^c with those on neighborhoods of θ_0. The second condition implies that the average $p_n = \frac{1}{\lambda(V^c)}\int_{V^c}\frac{p_\theta^n}{p_{\theta_0}^n}\lambda(d\theta)$ tends to zero exponentially a. e. (P_{θ_0}). It also implies that for neighborhoods $V_n\subset V$ of θ_0 chosen so that $\| P_\theta^n - P_{\theta_0}^n\|\to 0$ rapidly enough uniformly on V_n, the average $q_n = \frac{1}{\lambda(V_n)}\int_{V_n} p_\theta^n/p_{\theta_0}^n\,\lambda(d\theta)$ tends to 1 either in P_{θ_0} probability or a. s. Condition (iii) then insures that $Q_\theta^n(V^c) \leqq \frac{\lambda(V^c)\,p_n(x)}{\lambda(V_n)\,q_n(x)}\to 0$ either in P_{θ_0} probability or a. s. according to the possible choices for $\{\varepsilon_n\}$ in (iii).

* Theorem 6.2 and a result similar to theorem 6.3 have been announced in page 48 of the July issue (1964) of the Proceedings of the National Academy of Sciences.

Specifically, conditions (i) and (ii) imply inequality (11) and (12) so that for some integer $k \geqq 1$ and real $r > 0, p_n \leqq e^{-\frac{r}{2k}n}$ for all sufficiently large n, a. e. (P_{θ_0})· On the other hand, taking $V_n = V \cap W_{\varepsilon_n}$,

$$(14) \quad P_{\theta_0}\{|q_n - 1| > \delta\} \leqq \frac{1}{\delta} \int |q_n - 1| P_{\theta_0}^n(dx) \leqq \frac{1}{\delta\lambda(V_n)} \int_{V_n} \| P_\theta^n - P_{\theta_0}^n \| \lambda(d\theta)$$

for any $\delta > 0$. Since on V_n the integrand is less than $n\varepsilon_n$, $q_n \to 1$ in P_{θ_0} probability. By condition (iii), since

$$(15) \qquad Q_x^n(V^c) \leqq \frac{\lambda(V^c)\, p_n(x)}{\lambda(V_n)\, q_n(x)} \leqq \left(\frac{e^{-\frac{r}{2k}}}{[\lambda(V_n)]^{1/n}}\right)^n \frac{\lambda(V^c)}{q_n(x)},$$

the first conclusion will follow.

To prove the second statement, suppose (iii) holds with $\varepsilon_n \leqq \frac{1}{n^{2+\delta}}$. Take the union of the sets in the left side of (14) and sum the right side over all $n \geqq N$. Then

$$P_{\theta_0}(\bigcup_{n \geqq N}\{|q_n - 1| > \delta\}) \leqq \frac{1}{\delta} \sum_{n=N}^{\infty} \frac{1}{\lambda(V_n)} \int_{V_n} \| P_\theta^n - P_{\theta_0}^n \| \lambda(d\theta) \leqq \frac{1}{\delta} \sum_{n=N}^{\infty} n\varepsilon_n .$$

Let $N \to \infty$ and this gives $q_n \to 1$ a. e. (P_{θ_0}). By (15), $Q_x^n(V^c) \to 0$ a. e. (P_{θ_0}).

Since V was arbitrary, the theorem follows from the form of $\beta_n(x)$ in the first line of the proof of theorem 6.2.

This provides a convenient verification of the fact that the Bayes estimators in example 3 are consistent when λ is uniform on $[0, 1/2]$.

Theorems 6.1, 6.2 and 6.3 do not depend on the structure of Θ. For the next result, assume that Θ is a locally compact metric space. Let δ be the distance on Θ.

Theorem 6.4: *Suppose* (i) *there exists a compact neighborhood V of θ_0 and a uniformly consistent test of $P_\theta = P_{\theta_0}$ against $P_\theta \in \{P_\theta, \theta \in V^c\}$,* (ii) *for $\theta' \in V$, $\| P_\theta - P_{\theta'} \| \to 0$ when $\delta(\theta, \theta') \to 0$,* (iii) *for $\{\varepsilon_n\}$ and $\{W_{\varepsilon_n}\}$ as in theorem* 6.3, $\liminf_{n\to\infty}(\lambda(V \cap W_{\varepsilon_n}))^{1/n} = 1$, (iv) *$p_\theta^1$ is $\mathfrak{B} \times \mathfrak{A}_1$-measurable. Then $\beta_n \to \theta_0$ either in P_{θ_0} probability or a. e. (P_{θ_0}) according to wheter ε_n may be chosen so that $n\varepsilon_n \to 0$ or $n^2\varepsilon_n \to 0$.*

Proof. Since V is compact so is $V \cap W^c$ where W is any open neighborhood of θ_0 with $W \subset V$. Conditions (i) and (ii) imply the existence of a uniformly consistent test of $P_\theta = P_{\theta_0}$ against $P_\theta \in \{P_\theta, \theta \in W^c\}$. The result then follows from theorem 6.3.

The assumptions of these theorems are often easily verified. For example, the conditions of theorem 6.1 are satisfied in the example of KIEFER and WOLFOWITZ [*6*] in which Θ is the upper half-plane $\{-\infty < \mu < +\infty, 0 < \sigma < \infty\}$ and the underlying family of distributions are those given by the densities

$$p(z) = \frac{1}{2\sqrt{2\pi}}\left[\exp\left\{-\frac{(z-\mu)^2}{2}\right\} + \frac{1}{\sigma}\exp\left\{-\frac{(z-\mu)^2}{2\sigma^2}\right\}\right], \quad -\infty < z < \infty .$$

It may be noted that when the family of distributions $\mathcal{D}$ consists of distributions on the real line, if Θ is compact and $\| P - Q \|$ is equivalent to the Kolmogorov-

Smirnov distance and the conditions of theorem 6.1, for example, are usually easily verified.

Although the results have been stated for the independent identically distributed case, similar results follow when there is some sort of weak dependence. For example the conditions of independence and identity of distributions came into the proof of theorem 6.1 in the use of a strong law of large numbers and the use of lemma 6.1 in conjunction with the fact that $\varrho(P^n, Q^n) = \varrho_n(P, Q) \leqq \varrho_1^n(P, Q)$ in this case. Similar strong laws of large numbers are known when certain types of dependence are involved; e. g., Katz and Thomasian [5] in the case of discrete Markov processes satisfying Doeblin's condition. Here the H of the theorems would be replaced by the expected value of the log of the ratio of densities taken with respect to the stationary measure π on $\mathfrak{A}$ which corresponds to the θ_0 of the theorem. When the conditions $\varrho(P_1^n\ Q^n) \leqq \varrho_1^n(P, Q)$ and $\| P_{\theta_0}^n - P_n \| \geqq 2(1 - 2e^{-cn})$, $c > 0$ are met (again P_n is the average with respect to λ of the distributions P_θ on $\mathfrak{A}_n$, $\theta \in V^c$), the arguments in the proof of theorem 6.1 apply to obtain the same conclusion. The Bayes estimators $\{\beta_n\}$ are still consistent for each $\theta_0 \in \Theta$ for which the conditions hold for all V with $\lambda(V) > 0$.

Further the existence of a uniformly consistent test in theorem 6.1 may be replaced by a condition such as that $H(\theta) < -\eta$ on V^c for some $\eta > \varepsilon$ to obtain a corollary to theorem 6.1.

Corollary to theorem 6.1: *If* (i) *the densities* p^1 *may be chosen* $\mathfrak{B} \times \mathfrak{A}_1$*-measurable and* (ii) $\lambda(V) > 0$, $H(\theta) < -\eta$ *on* V^c *for some* $\eta > 0$ *and* $H(\theta) > -\varepsilon$ *on some set* $W_\varepsilon \subset V$ *with* $\lambda(W_\varepsilon) > 0$ *for each* ε *with* $0 < \varepsilon < \eta$, *then* $Q_x^n(V^c) \to 0$ *a. e.* (P_{θ_0}).

This follows from theorem 6.1 since

$$\frac{1}{n}\sum_{i=1}^{n}[\log p_\theta^1(z_i) - \log p_{\theta_0}^1(z_i)] = \varphi_n(x, \theta)$$

may be used as a test statistic to construct a uniformly consistent test of the hypothesis $P_\theta = P_{\theta_0}$ against the alternative $P_\theta \in \{P_\theta, \theta \in V^c\}$. Alternatively, a similar argument to that for (13) gives

$$\limsup_{n \to \infty} \frac{1}{\lambda(V^c)} \int_{V^c} e^{\varphi_n(x,\theta)} \lambda(d\theta) \leqq \frac{1}{\lambda(V^c)} \int_{V^c} e^{H(\theta)} \lambda(d\theta) < e^{-\eta}$$

a. e. (P_{θ_0}), so that a. e. (P_θ) for all sufficiently large n,

$$\frac{1}{\lambda(V^c)} \int_{V^c} e^{n\varphi_n(x,\theta)} \lambda(d\theta) = \frac{p_n}{p_{\theta_0}^n} \leqq e^{-n\eta}.$$

This together with (13) gives $Q_x^n(V^c) \leqq ce^{-n(\eta - \varepsilon)}$ and the result. Similar variations may be obtained for theorem 6.3.

Extensions in other directions are possible; for example the class of loss functions considered could be increased.

I am deeply indebted to Professor L. Le Cam, under whose guidance I obtained the results established in my thesis and given in this paper.

References

[1] BAHADUR, R. R.: Examples of the inconsistency of the maximum likelihood estimate, Sankhya, **20**, 207—210 (1958).

[2] CHERNOFF, H.: Sequential design of experiments. Ann. math. Statistics, **30**, 755—770 (1959).

[3] DOOB, J. L.: Application of the theory of martingales. Colloque international Centre nat. Rech. Sci, Paris, 22—28 (1949).

[4] FREEDMAN, D. A.: On the asymptotic behaviour of Bayes estimates in the discrete case, Ann. math. Statistics, **34**, 1386—1403 (1963).

[5] KATZ, M., and A. J. THOMASIAN: A bound for the law of large numbers for discrete Markov processes, Ann. math. Statistics **32**, 336—337 (1961).

[6] KIEFER, J., and J. WOLFOWITZ: Consistency of the maximum likelihood estimate in the presence of infinitely many incidental parameters, Ann. math. Statistics, **27**, 887—906 (1956).

[7] KRAFT, C.: Some conditions for consistency and uniform consistency of statistical procedures, Univ. California, Publ. Statist. **2**, 125—142 (1955).

[8] LECAM, L.: On some asymptotic properties of the maximum likelihood estimates and related Bayes estimates. Univ. California Pub. Statist., **1**, 277—330 (1953).

[9] — Les Proprietes Asymptotiques des Solutions de Bayes. Publ. Inst. Statist. Univ. Paris. **7**, 17—35 (1958).

[10] —, and L. SCHWARTZ: A necessary and sufficient condition for the existence of consistent estimates, Ann. math. Statistics **31**, 140—150 (1960).

[11] WALD, A.: Note on the consistency of the maximum likelihood estimates, Ann. math. Statistics **20**, 595—601 (1949).

Dept. of Mathematics
University of British Columbia
Vancouver 8, B. C., Canada

(Received March 16, 1964)