THE EXPONENTIAL CONVERGENCE OF POSTERIOR PROBABILITIES WITH IMPLICATIONS FOR BAYES ESTIMATORS OF DENSITY FUNCTIONS

(Abbreviated title: Convergence of Bayes Estimators)

by

Andrew R. Barron

University of Illinois

at

Urbana-Champaign

Technical Report #7

April 1988

Department of Statistics
University of Illinois
725 South Wright Street
Champaign, IL 61820

The Exponential Convergence of Posterior Probabilities with Implications for Bayes Estimators of Density Functions

(Abbreviated title: Convergence of Bayes Estimators)

Andrew R. Barron

University of Illinois at Urbana-Champaign

Summary

Necessary and sufficient conditions are determined for sequences of posterior probabilities of parameter sets to converge to one at an exponential rate, assuming that the prior assigns positive probability to the relative entropy rate neighborhoods of the distribution of the process $\{X_n\}$. The result is applied to the case of independent random variables to determine conditions on the prior such that Bayes estimators $\hat{p}_n(x)$ of the probability density function $p(x)$ converge in the L^1 sense, i.e., $\int |\hat{p}_n - p|$ tends to zero, with probability one. Also, some useful bounds are obtained for all N for the expected value of the sum of relative entropies $\sum_{n \le N} \int p \log (p/\hat{p}_n)$. The proof uses a frequentist approximation to the Bayesian's joint law for the parameter and the data. The results are applied to a variety of interesting priors.

Submitted to the *Annals of Statistics*, November 1987.

[1] This work was supported in part by an Office of Naval Research grant N00014-86-K-0670 and by a National Science Foundation Postdoctoral Research Fellowship.

AMS 1980 subject classifications. Primary 62G05,62A15, secondary 62F15

Key words and phrases. Bayes estimators, posterior distribution, predictive density, density estimation, consistency, hypothesis testing, variation distance, relative entropy, large deviations.

1. Introduction

Some convergence results are established for posterior distributions and for Bayes estimators of probability density functions. Consider the case of (conditionally) independent and identically distributed random variables $X_1,X_2,...,X_n,...$ which take values in a measurable space $(\mathbf{X},\mathbf{B})$ and which have a coordinate distribution that is absolutely continuous with respect to a fixed sigma-finite measure $\lambda(dx)$. It is assumed that the space $(\mathbf{X},\mathbf{B})$ is separable (countably generated). Bayes procedures utilize a parameter space $(\Theta,\mathbf{B}_\Theta)$, a family of $\mathbf{B}\times\mathbf{B}_\Theta$ measurable probability density functions $q(x \mid \theta) = q_\theta(x)$, and a prior probability distribution $v(d\theta)$ to construct the posterior distribution which is given by Bayes rule

$$v_n(A \mid X_1,X_2,...,X_n) = \frac{m^n(X_1,...,X_n;\, A)}{m^n(X_1,...,X_n)} \text{ for } A \in \mathbf{B}_\Theta \tag{1}$$

where the numerator is $m^n(x_1,...,x_n;\, A) = \int_A \prod_{i=1}^n q(x_i \mid \theta) v(d\theta)$ and the denominator is $m^n(x_1,...,x_n) = m^n(x_1,...,x_n;\Theta)$ (if the denominator is zero, arbitrarily set the posterior probability to be $v(A)$). Then for any loss function $L(p,\hat{p})$ a Bayes estimator of the density is a function $\hat{p}_n(x) = \hat{p}_n(x;\, X_1,...,X_n)$ which minimizes the posterior loss $\int_\Theta L(q_\theta,\hat{p})\; v_n(d\theta \mid X_1,...,X_n)$.

The asymptotics of the posterior distribution and of Bayes estimators are examined under the assumption that the random variables X_i are independent with a probability density function p. In general this density p need not be a member of the family $\{q(\cdot \mid \theta)\}$; however, the density is assumed to be a limit point of this family in a sense made precise below (condition (A)). Fundamentally, we are adopting the non-Bayesian point of view, that p is the true but unknown density; nevertheless, the results should also be of interest to a Bayesian who asks what asymptotics obtain if he or she conditions on the event that the ergodic mode of the exchangeable process $\{X_n\}$ happens to be the distribution with coordinate density function p.

The principle aim of this paper is to determine reasonable and readily verifiable conditions such that Bayes density estimators are consistent in the sense that $\lim \int \mid \hat{p}_n(x) - p(x) \mid \lambda(dx) = 0$, with probability one. The key step in our development is the determination of conditions for sequences of posterior probabilities $v_n(A \mid X_1,...X_n)$ to convergence to one, with probability one. Of particular interest to us is the posterior probability of $A = \{\theta : \int \mid p - q_\theta \mid < \varepsilon\}$ for $\varepsilon > 0$. Some bounds are also obtained for the cumulative expected value of the relative entropy $\int p(x) \log(p(x)/\hat{p}_n(x))\; \lambda(dx)$. These bounds follow from an interesting large

deviation approximation to the Bayesian joint distribution of Θ and $X_1,\ldots,X_n$.

Let $D(p \,||\, q) = \int p \log (p/q)$ denote the relative entropy (or Kullback-Leibler informational divergence) for probability density functions. The following condition is required.

(A) The prior is *information dense* at p, in the sense that relative entropy neighborhoods are assigned positive prior probability, that is

$$\nu\{\theta : D(p \,||\, q_\theta) < \varepsilon\} > 0 \text{ for all } \varepsilon > 0.$$

The most useful priors are information dense for a large class of density functions. In the examples we present in section 8, condition (A) is satisfied for every probability density function p for which $D(p \,||\, q_0)$ is finite (where q_0 is a fixed density function). This includes all density functions for which the ratio $p(x)/q_0(x)$ is bounded.

Proposition 1: *Assume that the prior is information dense at p. Let A_n be any sequence of measurable parameter sets. For the sequence of posterior probabilities $\nu_n(A_n^c \mid X_1,X_2,\ldots,X_n)$ to be exponentially small, with P probability one, it is necessary and sufficient that there exist a sequence of measurable parameter sets B_n and C_n which satisfy three conditions:*

(i) $A_n \cup B_n \cup C_n = \Theta$,

(ii) $\nu(B_n)$ *is exponentially small, and*

(iii) *there is a uniformly consistent test of P versus $\{Q(\cdot \mid \theta) : \theta \in C_n\}$ with*

$P\{(X_1,\ldots,X_n) \in S_n \text{ infinitely often}\} = 0$ *and*

$\sup_{\theta \in C_n} Q_\theta\{(X_1,\ldots,X_n) \in S_n^c\}$ *exponentially small*

for some sequence of critical sets S_n in $\mathbf{B}^n$.

Here a sequence of nonnegative numbers a_n is said to be *exponentially small* if for some $r > 0$, $a_n < e^{-nr}$ for all large n or equivalently limsup $(1/n)\log a_n < 0$; a superscript c denotes the complement of a set in the relevant space; $(\mathbf{X}^n,\mathbf{B}^n)$ denotes the product space of n copies of $\mathbf{X}$ with the product sigma-field; and P and Q_θ are probability measures (on an underlying measurable space $(\Omega,\mathbf{B}_\Omega)$) for which the random variables $X_1,X_2,\ldots$ are independent with coordinate density functions p and q_θ respectively. (At the risk of abusing notation we also use P and Q_θ to denote the distribution of X_i on $\mathbf{X}$: the distinction should be clear from the context.) Note that a sufficient condition for $P\{(X_1,\ldots,X_n) \in S_n \text{ infinitely often}\} = 0$ is that the probability of error $P\{(X_1,\ldots,X_n) \in S_n\}$ be exponentially small. A sequence of tests for which the

probabilities of error for both sets of hypotheses are uniformly exponentially small is said to be *uniformly exponentially consistent* (UEC).

Proposition 1 is generalized in section 3 (Theorem 5) to permit nearly arbitrary sequences of dependent random variables $X_1,\ldots,X_n$, families $\{Q^n(\cdot \mid \theta) : \theta \in \Theta_n\}$ of distributions on $(\mathbf{X}^n,\mathbf{B}^n)$, and priors $\nu_n(d\theta)$, under an appropriate generalization of the notion of information denseness.

The conditions of Proposition 1 are an extension and adaptation of conditions determined by Schwartz (1965). Her sufficient (but not necessary) condition requires the existence of a uniformly exponentially consistent test for P versus $\{Q_\theta : \theta \in A^c\}$, whereas we show that such a test need only exist against a subset C_n of A^c for which the prior probability of the difference between these sets is small. An important use of Schwartz's result is to obtain weak convergence of Bayes estimators of the distribution by showing weak-star convergence of the posterior distribution (induced on the space of measures) to point mass at the distribution P. (Indeed, it is well known that for any finite partition T of $\mathbf{X}$ and any $\varepsilon > 0$, UEC tests exist against $\{Q : \sum_{B\in T} \mid P(B)-Q(B) \mid \geq \varepsilon\}$. If $\mathbf{X}$ is a separable metric space, then there are countably many T and $\varepsilon > 0$ such that *all* weak neighborhoods of P contain at least one of the sets $A = \{\theta : \sum_{B\in T} \mid P(B)-Q_\theta(B) \mid < \varepsilon\}$. Applying Proposition 1 or Schwartz's result to each of these sets, demonstrates that with probability one, the posterior distribution asymptotically concentrates inside every weak neighborhood of P.) For many finite dimensional parametric families, weak convergence of distributions in the family implies convergence of the densities; however, the class of densities for which convergence obtains may be severly restricted by this condition. To obtain convergence of density estimators without such restrictions, we use Proposition 1 to obtain stronger modes of convergence of the posterior.

First let's mention the case that the prior is discrete, assigning positive mass to a countable set of density functions. (For judicious choices of this countable set, the information limits may be a fairly rich class of densities.) Suppose that for some $0 < \alpha < 1$ the sum $\sum_\theta(\nu(\theta))^\alpha$ is finite; this condition guarantees that the prior has light tails (in particular the set $B_n = \{\theta : \nu(\theta) < e^{-n\varepsilon}\}$ has exponentially small prior probability for each $\varepsilon > 0$). For such light tailed priors the convergence of the posterior and of density estimates (using the L^1 distance) holds for every p for which the prior is information dense. This result for discrete priors is established in section 6.

Now in general, some recent results on the existence of uniformly exponentially consistent tests may be applied to obtain Bayes consistency results. Let

$d(p,q) = \int |p(x)-q(x)| \lambda(dx)$ be the L^1 distance between density functions (which is the same as the total variation distance between the corresponding distributions) and for measurable partitions T let $d_T(P,Q) = \sum_{A \in T} |P(A)-Q(A)|$ be the $T-$variation distance between probability distributions on $\mathbf{X}$. Unfortunately, there does not exist a uniformly consistent test of the hypothesis p against all densities q with $d(p,q) \geq \varepsilon$ (unless the dominating measure is discrete) so we cannot directly conclude that the posterior asymptotically concentrates on L^1 distance neighborhoods of p. Nevertheless, it is shown in Barron (1987b) that *for every* $\varepsilon > 0$ *there exists a uniformly exponentially consistent sequence of tests of the hypotheses* P *versus* $\{Q : d_{T_n}(P,Q) > \varepsilon/2\}$ *if and only if the sequence of partitions* T_n *has effective cardinality of order* n *with respect to* P. (This means that for every $\varepsilon > 0$, limsup $k_n(\varepsilon)/n < \infty$ where $k_n(\varepsilon)$ is the minimum number of sets in T_n for which the probability of the union is at least $1-\varepsilon$. When $(\mathbf{X},\mathbf{B})$ is the real line with the Borel sets and λ is the Lebesgue measure an important example of a sequence of partitions with effective cardinality of order n (with respect to any probability measure) is the sequence of uniform partitions $T_n = \{((i-1)/n, i/n] : i=\ldots,-1,0,1,\ldots\}$.)

Consequently, we may use Proposition 1 to conclude that the posterior asymptotically concentrates on the L^1 distance neighborhood $A = \{\theta : d(p,q_\theta) < \varepsilon\}$ provided that the set $B_n = \{\theta : d(p,q_\theta) \geq \varepsilon, d_{T_n}(P,Q_\theta) < \varepsilon/2\}$ or the larger set $\{\theta : d(p,q_\theta) - d_{T_n}(P,Q_\theta) > \varepsilon/2\}$ has exponentially small prior probability for some sequence of partitions with effective cardinality of order n. This amounts to requiring that the d and d_{T_n} distances from p be nearly the same for "most" of the Q_θ distributions on $\mathbf{X}$.

We shall see that for many priors, the condition is satisfied for any density function p. For any partition T and density function p define the density p^T which takes the form of a simple function or "theoretical histogram"

$$p^T(x) = \frac{\int_A p(y)\, \lambda(dy)}{\lambda(A)} \quad \text{for } x \in A \in T.$$

Here we take p^T to be zero wherever $\lambda(A) = 0$. Abou-Jaoude (1976) introduced a property of sequences of partitions T_n which is equivalent to lim $d(p,p^{T_n}) = 0$ for all probability density functions p with respect to λ. We call such a sequence of partitions *rich*. (In particular the sequence of uniform partitions of the line is rich when λ is Lebesgue measure.) If T_n is a rich sequence of partitions then for all large n, $d(p,p^{T_n}) < \varepsilon/4$ and hence by the triangle inequality the set B_n (from the preceding paragraph) is contained in $\{\theta : d(q_\theta, q_\theta^{T_n}) > \varepsilon/4\}$. Thus we are led to the following

condition on the prior probability of "wild" densities which ensures that the posterior asymptotically concentrates on L^1 distance neighborhoods.

(B) For every $\varepsilon > 0$, the prior probability $\nu\{\theta : d(q_\theta, q_\theta^{T_n}) > \varepsilon\}$ is exponentially small for some rich sequence of partitions T_n with effective cardinality of order n.

Condition (B) is trivially satisfied if the family of density functions $q_\theta(\cdot)$, $\theta \in \Theta$ is uniformly equicontinuous (since then for every $\varepsilon > 0$ there is a countable partition T such that $d(q_\theta, q_\theta^T) \le \varepsilon$ for all θ and any fixed countable partition has effective cardinality of order 1). By Markov's inequality, another sufficient condition for (B) is that for some positive and increasing function $f(u), u > 0$, the sequence of expected values $\int \exp\{nf(d(q_\theta, q_\theta^{T_n}))\}\, \nu(d\theta)$ is not exponentially large. In particular, this is true if for some constants $\alpha, \gamma > 0$, $n(d(q_\theta, q_\theta^{T_n}))^\gamma$ is bounded by a function $c(\theta)$ for which $\int e^{\alpha c(\theta)} \nu(d\theta)$ is finite. To verify this condition when $\mathbf{X}$ is the real line and the densities $q(x \mid \theta)$ are differentiable with respect to x, suppose there is an $h > 0$ for which $c(\theta) = \int c(x,\theta)dx$ is finite for each θ (and $\int e^{\alpha c(\theta)} \nu(d\theta)$ is finite) where $c(x,\theta)$ is a function which dominates $|(d/dy)q(y \mid \theta)|$ for $|y - x| < h$, and let T_n be the uniform partition with cells of width $1/n$, then for all $n > 1/h$ we have $nd(q_\theta, q_\theta^{T_n}) \le c(\theta)$ as desired. Although these sufficient conditions are useful, the examples in section 8 show that sometimes it is straightforward to verify condition (B) directly.

Let $\hat{p}_n$ be any of the following Bayes estimators of the density function: the posterior mean density estimator $\hat{p}_n(x; X_1, \ldots, X_n) = \int q(x \mid \theta)\, \nu_n(d\theta \mid X_1, \ldots, X_n)$; the Bayes estimator for any bounded loss function which is equivalent to the L^1 distance (i.e. $L(p, \hat{p}) \to 0$ iff $d(p, \hat{p}) \to 0$); or in the case of a discrete prior, the posterior mode density estimator $\hat{p}_n(\cdot) = q(\cdot \mid \hat{\theta}_n)$ where $\hat{\theta}_n$ is any global maximizer of the posterior likelihood $\nu(\{\theta\}) \prod_{i=1}^n q(X_i \mid \theta)$.

Proposition 2: *If conditions A and B are satisfied then the posterior probability of L^1 distance neighborhoods of p tends to one exponentially fast, i.e., for each $\varepsilon > 0$,*

$$\nu_n(\{\theta : d(p, q_\theta) \ge \varepsilon\} \mid X_1, \ldots, X_n) \text{ is exponentially small, with } P \text{ probability one,}$$

and consequently, the density estimator $\hat{p}_n$ converges to p in L^1, i.e.,

$$\lim_{n \to \infty} d(p, \hat{p}_n) = 0, \text{ with } P \text{ probability one.}$$

Some conclusions can still be made even if condition (B) is not satisfied. In the next two results only condition (A) is required. For proposition 3, the estimator $\hat{P}_n(\cdot; X^n)$ may either be the posterior mean $\int Q(\cdot \mid \theta)\, \nu_n(d\theta \mid X^n)$ or in the discrete parameter case, a posterior mode $Q(\cdot \mid \hat{\theta}_n)$ (where $\hat{\theta}_n$ is a global maximizer of the posterior likelihood). For sequences of partitions T_n, let $\hat{p}_n^{T_n}$ be the density estimator constructed from $\hat{P}_n$ and T_n as described above.

Proposition 3: *If condition A is satisfied, then for any sequence of partitions T_n with effective cardinality of order n, the posterior probability of T_n−variation neighborhoods of P tends to one, indeed,*

$$\lim_{n\to\infty} \nu_n(\{\theta : d_{T_n}(P, Q_\theta) > \varepsilon\} \mid X_1,\ldots,X_n)$$ *is exponentially small, with P probability one*

and consequently, the estimator $\hat{P}_n$ converges to P in T_n variation, i.e.,

$$\lim_{n\to\infty} d_{T_n}(P, \hat{P}_n) = 0, \text{ with } P \text{ probability one.}$$

If also, T_n is a rich sequence of partitions, then the density estimator $\hat{p}_n^{T_n}$ converges in L^1,

$$\lim_{n\to\infty} d(p, \hat{p}_n^{T_n}) = 0, \text{ with } P \text{ probability one.}$$

The asymptotic concentration on T_n−variation neighborhoods is the strongest conclusion that can be made without imposing an additional condition (such as (B)). Indeed, in section 7, we demonstrate that if P is a continuous distribution on the line, then for any sequence of partitions T_n with effective cardinality of order greater than n with respect to P, there exists a parametric family and a prior which satisfies condition A, yet the posterior probability of the set $\{\theta : d_{T_n}(P, Q_\theta) < \varepsilon\}$ does not converge to one, in P probability, for some $\varepsilon > 0$.

The classic nonparametric estimators of density functions on the real line, such as the histogram and the kernel estimator, are based on smoothing the empirical distribution function. However, much more smoothing is required in order to obtain consistent density estimates from the empirical distribution than from Bayes estimators. In particular, the histogram requires that the sequence of partitions have effective cardinality of smaller order than n (i.e. bin widths $>> 1/n$ so that observations accumulate in the bins), whereas the Bayes estimators only need to be smoothed in bins of width δ/n (i.e. $T_n = \{(\delta(i-1)/n, \delta i/n] : i=\ldots,-1,0,1,\ldots\}$ where $\delta > 0$ is a small constant). Proposition 3 shows that Bayes estimators accurately estimate the probabilities of most small bins, even though most of these bins will be empty. This ability of Bayes estimators to learn the shape of the distribution in the areas between the

observations bodes well for applications in high dimensions where for moderate sample sizes most of the space appears to be empty.

Although L^1 consistent density estimates are important for certain applications (such as constructing discrimination functions which have nearly minimal average probability of error for a classification problem, see Glick 1972), there are also applications where Kullback-Leibler convergence is required (such universal data compression, see Davisson 1973, or stock market portfolio selection, see Barron and Cover 1988). Moreover, in view of condition (A) it would be most natural to obtain convergence results for $D(p \,||\, \hat{p})$. It is well known that D yields a stronger mode of convergence than the L^1 distance, indeed $d(p,q) \leq (2D(p \,||\, q))^{1/2}$, see Csiszár (1967).

When the relative entropy is taken as the loss function, it is readily seen that the Bayes estimator is the posterior mean density $\hat{p}_n(x) = \int q(x \mid \theta) \nu_n(d\theta \mid X_1,\ldots,X_n)$ which by Bayes rule is the conditional (or predictive) density $\hat{p}_n(x;X_1,\ldots,X_n) = m^{n+1}(X_1,\ldots,X_n,x)/m^n(X_1,\ldots,X_n)$ (recall that $m^n(x_1,\ldots,x_n) = \int (\prod^n q_\theta(x_i))\nu(d\theta)$). Moreover it is seen that the cumulative risk of the Bayes estimator is simply the relative entropy of the product density p^N with respect to the Bayesian density m^N, i.e.,

$$\sum_{n=0}^{N-1} E(D(p \,||\, \hat{p}_n)) = D(p^N \,||\, m^N). \tag{2}$$

Here the expectation is taken with respect to the distribution P which governs the random sample. This chain rule may be used to easily obtain the following proposition (see Barron 1987a).

Proposition 4: *If condition A is satisfied then the expected relative entropy $E(D(p \,||\, \hat{p}_n))$ converges to zero in the Cesàro sense, i.e.*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} E(D(p \,||\, \hat{p}_n)) = 0.$$

Consequently the risk $E(D(p \,||\, \hat{p}_n))$ is small for most large sample sizes, in the sense that given any $\varepsilon,\delta > 0$ for at least $(1-\delta)N$ of the sample sizes $n < N$, $E(D(p \,||\, \hat{p}_n) < \varepsilon$, for all sufficiently large N.

Since $D(p \,||\, \hat{p}_n) \geq (1/2)(d(p,\hat{p}_n))^2$ a similar Cesàro convergence holds for the expected square of the L^1 distance.

It is surprising that condition (A) is sufficient to obtain consistency properties of the Bayes density estimator while it is not sufficient to obtain consistency of the posterior probability of density neighborhoods. The non-negligible contribution to the posterior distribution from densities which lie outside a small L^1 distance ball does not

manage to disrupt the accuracy of the posterior mean, at least for most sample sizes.

As indicated above, the result of proposition 4 is equivalent to the convergence in information of p^n and m^n, i.e. $\lim (1/n)D(p^n \,||\, m^n) = 0$. A proof of this convergence is given in section 5 as a byproduct of a frequentist approximation to the Bayesian joint law for θ and X^n. The approximate law is shown to merge in information with the Bayesian law. A second byproduct is an approximation to the posterior distribution which agrees with the classic normal approximation (for small deviations of the parameter) in the case of smooth finite dimensional parametric families. However, unlike the normal approximation, the relative entropy based approximation that we develop is at least crudely accurate for large deviations and no smoothness or dimensionality assumptions are required.

Perhaps the most useful byproduct of the approximation is a bound on the Cesàro average of the risk which has the potential of yielding rate of convergence results as well as useful bounds for every finite sample size. Let $R_N = (1/N)\sum E(D(p \,||\, \hat{p}_n)$ denote this Cesàro risk which by (2) is the same as $(1/N)D(p^N \,||\, m^N)$. It is shown that for all n,

$$R_n \leq -\frac{1}{n} \log \int e^{-nD(\theta)} \, \nu(d\theta). \tag{3}$$

where $D(\theta) = D(p \,||\, q_\theta)$. Note that this bound depends only on the prior distribution of $D(\theta)$ and the sample size. As $n \to \infty$, the bound decreases to the ν-essential infimum of $D(\theta)$ which is zero if and only if condition (A) is satisfied. (The other proof of Proposition 4 in Barron 1987a does not lead to inequality (3)). A useful bound that follows from this inequality is

$$R_n \leq \varepsilon - \frac{1}{n} \log \nu(A_\varepsilon)$$

for any $\varepsilon > 0$ where $A_\varepsilon = \{\theta : D(\theta) < \varepsilon\}$. This directly relates the Cesàro risk of the Bayes estimator to the prior probability of relative entropy neighborhoods of the true density. Although the obtaining of rates of convergence is not the primary focus of the present paper, we illustrate the types of results which might be obtained. If $\nu(A_\varepsilon) \geq \exp\{-a(1/\varepsilon)^r\}$ for some $a, r > 0$ (as might be the case for certain priors if the logarithm of the density p has enough bounded derivatives) then setting $\varepsilon = n^{-1/(1+r)}$ we have the following bound on the Cesàro risk

$$R_n \leq (a+1)n^{-1/(1+r)} \quad \text{for all } n.$$

Faster rates of convergence such as $O(n^{-1}(\log n)^b)$ with $b \geq 1$ are possible if the prior probabilities $\nu(A_\varepsilon)$ are not as small (as might be the case for some priors if the

logarithm of the density is infinitely differentiable with exponentially decaying Fourier coefficients.) For smooth finite dimensional parametric families a convergence rate of $O(n^{-1}(\log n))$ is expected and this is indeed the case if $v(A_\varepsilon) > a\varepsilon^{d/2}$. (This bound holds for some a which depends on p whenever the follows conditions are satisfied: Θ is an open subset of $\mathbf{R}^d$, the prior has a positive density in a neighborhood of θ_0 where $p(x) = q(x \mid \theta_0)$, and the relative entropy neighborhood A_ε contains a Euclidean sphere of squared radius proportional to ε.) Typically, the rate $(\log n)/n$ cannot be improved, since it corresponds to the rate $1/n$ for the individual terms $E(D(p \,||\, \hat{p}_n))$ in the cumulative risk.

The outline of the remainder of this paper is as follows. A discussion of the history of Bayes consistency is given in section 2. Then the generalization and proof of the theorem giving necessary and sufficient conditions for the convergence of posterior probabilities is in section 3 followed by results for the merging of Bayes and frequentist distributions in section 4. Section 5 obtains convergence of Bayes estimators from convergence of the posterior. Implications for the case of a discrete prior are examined in section 6. Section 7 develops a counterexample showing inconsistency when the conditions are not satisfied. Finally, section 8 gives several examples where the conditions are satisfied and hence consistency obtains for a large class of density functions p.

2. History and Discussion

Various techniques for examining the asymptotics of Bayes estimators have had moderate successes in handling either finite dimensional or infinite dimensional (nonparametric) families of distributions. The important result due to Doob (1949) is that Bayes estimators are consistent for almost every parameter value with respect to the prior (technical refinements of Doob's theorem are in Breiman, LeCam, and Schwartz 1964 and in Diaconis and Freedman 1986, Corollary A.2). However, unless the parameter space is discrete (in which case almost everywhere becomes everywhere, see section 6), we are left not knowing for any particular distribution which might be realized whether or not it would be consistently estimated.

The classic approach to Bayes consistency as in LeCam (1953), Berk (1966,1970), and Strasser (1981a) is to adapt the Wald (1949) conditions for the consistency of maximum likelihood estimators (the key conditions require the integrability of certain suprema of the log likelihood function). Similarly, LeCam (1958), Bickel and Yahav (1967), and Johnson (1967,1970) use Wald type conditions as well as natural differentiability conditions to establish asymptotic normality of the posterior

distribution. Although successful for smooth finite dimensional families, the Wald conditions are not easily checked and often not satisfied for infinite dimensional or nonparametric problems.

Nonclassical approaches to Bayes consistency seek alternative conditions which are more readily checked (even in finite dimensional families) or more applicable to nonparametric problems. Both weak and strong convergence results are desired. It was once thought that to obtain weak-star convergence of the posterior (to point mass at the distribution P) it would be sufficient that P be in the weak-star support of the prior (that is, the prior assigns positive probability to the neighborhoods of P in the topology of weak convergence of measures). Examples of priors for which the weak-star support is known to include all distributions on the real line are the tailfree priors of Fabius (1964), the neutral priors of Doksum (1974), the Dirichlet process priors of Ferguson (1973) (which are a special case of both tailfree and neutral priors), and mixtures of Dirichlet process priors (Antoniak 1974); for a review of these methods see Ferguson (1974). Fabius (1964) establishs weak consistency for the particular case of tailfree priors. However, in general, weak-star support is *not* enough to prove weak-star convergence of posteriors as is demonstrated by the counterexamples of Freedman and Diaconis (1983,1986) (e.g. for mixtures of Dirichlet processes the posterior may be inconsistent). To obtain Bayes consistency in general it is necessary to impose stronger conditions on the prior.

Two commonly used conditions for the prior are that the relative entropy neighborhoods of the true distribution are assigned positive prior probability (condition (A)) or that total variation (or Hellinger) balls of radius ε/n have prior probability which is not exponentially small (this latter condition may be used in place of our condition (A) with the same effect, see section 4). Under either of these conditions, weak-star convergence of the posterior is guaranteed and stronger convergence results often hold. Early work in this direction is by Freedman (1963) and Schwartz (1965). Freedman (1963) shows that for a discrete (countable) sample space, condition (A) and an additional finite entropy assumption imply consistency in total variation. [1] Schwartz (1965, section 6) shows that either the relative entropy or the total variation condition implies that the posterior distribution asymptotically concentrates on any set for which there exists a uniformly consistent test against the complement of the set.

[1] Proposition 3 establishs Freedman's result without the additional finite entropy assumption. The proposition applies to the discrete case, since a fixed countable partition has effective cardinality of order (1).

Fortunately, there have been both general technical conditions for the existence of uniformly consistent tests (Kraft 1955, LeCam and Schwartz 1960, LeCam 1970) and some specific results on the existence of uniformly consistent tests for P versus $\{Q : L(P,Q) \geq \varepsilon\}$ for various popular loss functions $L(P,Q)$. For the Cramer - Von Mises, Kolmogorov - Smirnoff and Vapnik - Chervonenkis (1971) distances between distributions, the usual tests are known to be uniformly consistent by applying the results of Hoeffding and Wolfowitz (1958, p.705-706). On the other hand, if $L(P,Q)$ is any distance (or loss) function which dominates the total variation distance (such as the relative entropy, Chi-square, or Hellinger distance) then no uniformly consistent test exists (see Barron 1987b). (This surprising fact provides a contradiction to the "Corollary to Theorem 6.1" appended to the end of Schwartz's (1965) paper: a counterexample is given in section 7 of this paper. In defense of Schwartz, who met her untimely death in 1965, the incorrect result is not in her dissertation (1960) from which the rest of her paper is taken.)

There have been a few results which expand on Schwartz's technique. Strasser (1981b) generalizes her results to deal with infinite prior measures, recasts the condition on the existence of a uniformly consistent test as a continuity condition for the quantity to be estimated, and shows how the conditions may be checked for finite dimensional parametric families. LeCam (1973,1982,1986) uses an adaptation of Schwartz's technique to prove consistency in Hellinger distance. He covers the parameter space with many small Hellinger balls and uses the fact that uniformly consistent tests exist against each such ball; consistency then obtains under a condition on the Hellinger dimension of the parameter space. LeCam's approach might be workable in certain nonparametric cases, but I am not yet aware of specific examples. For finite dimensional parametric families, Ibragimov and Has'minskii (1973, 1981 Section I.5) obtain consistency for posterior distributions and for Bayes estimators under conditions similar to those used by LeCam. In particular, the Hellinger balls are assumed to contain and to be contained in Euclidean balls of appropriate radii. For many finite dimensional families these conditions are easier to verify than the classical Wald conditions.

Our generalization of Schwartz's technique simply involves breaking the complement of the parameter set of interest into two disjoint sets B_n and C_n. The set C_n corresponds to a composite set of distributions against which there is a uniformly exponentially consistent test and the set B_n has exponentially small prior probability as $n \to \infty$. In this way we can use the existence of UEC tests for a sequence of alternative sets which increases to the complement of the total variation ball. The necessity

and sufficiency of the existence of such sets B_n and C_n for the convergence of the posterior probabilities is demonstrated in section 3.

There is a fundamental distinction between our conditions for Bayes consistency and most of the other conditions for strong consistency which have been developed. In each case some sort of smoothness is assumed for the density or likelihood function $q(x \mid \theta)$. On one hand, the classical Wald conditions require it be a smooth function of the parameter θ for all X in a set of high probability (similarly, the LeCam or Ibragimov and Has'minskii type condition requires that Hellinger distances between densities q_θ behave smoothly as a function of θ). On the other hand, our condition (B) requires that the density $q(x \mid \theta)$ be a nearly a smooth function of the variable x (at least for θ in a set of high prior probability). One advantage of requiring smoothness in x (instead of in θ) is that this is more akin to the classic conditions for accurate nonparametric estimation of densities. Another advantage is that the validity of the assumptions does not depend on the parameterization. Indeed if the priors associated with two different parameterizations of the same set of probability measures induce the same prior distribution on the set of measures, then clearly conditions (A) and (B) and the Bayes estimators of the probability measures are unchanged. The practicality of our conditions is demonstrated by some "nonparametric" examples in section 8.

3. Consistency of Posterior Probabilities

We formalize our technique for proving convergence of posterior probabilities. In its most general setting the method does not require independent random variables. Moreover, all aspects of the model (i.e. the prior and the family of densities) may depend on the sample size. In this section we are *not* restricted to the case that $\mathbf{X}^n$ is the product space for random samples $X^n = (X_1, X_2, \ldots, X_n)$; however, remarks may allude to this most important setting.

To be precise, for each $n \geq 1$, let $(\mathbf{X}^n, \mathbf{B}^n)$ be a measurable space, let $\{Q^n(\cdot \mid \theta) : \theta \in \Theta_n\}$ be a family of probability measures on $\mathbf{X}^n$ with density functions $q^n(\mathbf{x}^n \mid \theta)$ with respect to a sigma-finite measure $\lambda^n(d\mathbf{x}^n)$, let $\mathbf{B}_{\Theta_n}$ be a sigma-field of subsets of the parameter space Θ_n, and let ν_n be a prior measure with $\nu_n(\Theta_n) \leq 1$. It is assumed that for every n, the density functions $q^n(\mathbf{x}^n \mid \theta)$ may be chosen to be $\mathbf{B}^n \times \mathbf{B}_{\Theta_n}$ measurable and, in particular, versions are chosen which are $\mathbf{B}_{\Theta_n}$ measurable for every $\mathbf{x}^n$.

For parameter sets A_n in Θ_n, define the ($\mathbf{B}^n$ measurable) posterior probability by

$$\nu_n(A_n \mid \mathbf{x}^n) = \frac{\int_{A_n} q^n(\mathbf{x}^n \mid \theta)\, \nu_n(d\theta)}{\int_{\Theta_n} q^n(\mathbf{x}^n \mid \theta)\, \nu_n(d\theta)} \tag{4}$$

for all $\mathbf{x}^n$ for which the denominator is positive and finite (for any other $\mathbf{x}^n$ set $\nu_n(A_n \mid \mathbf{x}^n) = \nu_n(A_n)$, say). For the analysis of posterior probabilities, the (mixture) measures $M^n(\cdot, A_n) = \int_{A_n} Q^n(\cdot \mid \theta)\, \nu_n(d\theta)$ and $M^n(\cdot) = M^n(\cdot, \Theta_n)$ are useful. These measures have density functions $m^n(\mathbf{x}^n, A_n) = \int_{A_n} q^n(\mathbf{x}^n \mid \theta)\, \nu_n(d\theta)$ and $m^n(\mathbf{x}^n) = m^n(\mathbf{x}^n, \Theta_n)$ which are respectively, the numerator and the denominator of the posterior probability.

It is assumed that there is an underlying probability space $(\Omega, \mathbf{B}_\Omega, P)$ and for each n, there is a $\mathbf{B}^n/\Omega$ measurable random variable (sample) X^n. It is assumed that the induced distribution P^n on $\mathbf{X}^n$ is absolutely continuous with respect to λ^n with probability density function $p^n(\mathbf{x}^n)$.

If it is clear from the context, we omit some of the superscripts when writing the density functions $q(\mathbf{x}^n \mid \theta)$, $m(\mathbf{x}^n)$, $p(\mathbf{x}^n)$, etc.

Definition 1: We say that P^n and M^n *merge in probability* if for every $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left\{\frac{m(X^n)}{p(X^n)} > e^{-n\varepsilon}\right\} = 1.$$

We say that they *merge with probability one* if for every $\varepsilon > 0$

$$P\left\{\frac{m(X^n)}{p(X^n)} > e^{-n\varepsilon} \text{ for all large } n\right\} = 1.$$

It turns out that this definition of merging is equivalent to $\lim(1/n) \log (p(X^n)/m(X^n)) = 0$ in probability or with probability one, respectively. (This follows from the fact that $m(X^n)/p(X^n)$ is not greater than $e^{n\varepsilon}$ except in a set of probability less than $e^{-n\varepsilon}$ by application of Markov's inequality.) Conditions which ensure merging are developed in section 4. In particular, in the stationary independent case with a fixed prior ν, it is enough that relative entropy neighborhoods of P have positive prior probability.

Definition 2: A sequence of $\mathbf{B}^n$ measurable functions $0 \le \phi_n(\mathbf{x}^n) \le 1$ is a *uniformly consistent test* of P^n against a set K^n of probability measures on $\mathbf{X}^n$, if both $\int \phi_n dP^n$ and $\sup_{Q^n \in K^n} \int (1-\phi_n) dQ^n$ tend to zero as $n \to \infty$. The test is *uniformly exponentially*

consistent (UEC) if these probabilities of error are uniformly less than e^{-nr} for all large n, for some $r > 0$.

We remark that for any critical function ϕ_n there is a critical region $S_n = \{\mathbf{x}^n : \phi_n(\mathbf{x}^n) > 1/2\}$ for which the probabilities of error $P^n(S_n)$ and $Q^n(S_n^c)$ are no more than twice $\int \phi_n dP^n$ and $\int (1-\phi_n) dQ^n$, respectively. (This follows from Markov's inequality.) Thus if attention is restricted to nonrandomized tests, the property of existence or nonexistence of uniformly consistent tests remains the same.

The following Theorem is our principle tool for establishing Bayes consistency. Assuming merging properties we obtain necessary and sufficient conditions for the posterior probability $v_n(A_n \mid X^n)$ to converge to one, at an exponential rate, in probability or with probability one.

Theorem 5: Consistency of posterior probabilities.
(1) If P^n and M^n merge in probability, then for any sequence of sets $A_n \in \mathbf{B}_{\Theta_n}$, there exists $r > 0$ such that

$$\lim_{n\to\infty} P\{ v_n(A_n^c \mid X^n) < e^{-nr} \} = 1 \tag{5}$$

if and only if there exist measurable parameter sets B_n and C_n and constants $r_1, r_2 > 0$ *such that*

(i) $A_n \cup B_n \cup C_n = \Theta_n$,

(ii) *B_n has negligible prior probability, i.e., $v_n(B_n) \le e^{-nr_1}$, and*

(iii) *there is a uniformly consistent test of P^n versus $\{ Q^n(\cdot \mid \theta) : \theta \in C_n \}$ with*

$$\lim_{n\to\infty} P^n(S_n) = 0 \text{ and } \sup_{\theta \in C_n} Q^n(S_n^c \mid \theta) \le e^{-nr_2} \text{ for some } S_n \in \mathbf{B}^n.$$

(2) If P^n and M^n merge with probability one, then there exists $r > 0$ such that

$$P\{ v_n(A_n^c \mid X^n) \ge e^{-nr} \text{ infinitely often} \} = 0$$

if and only if there exist parameter sets B_n and C_n such that (i), (ii), and (iii) are satisfied with $P\{ X^n \in S_n \text{ infinitely often} \} = 0$.

Remarks: When checking for consistency it is sufficient that the exponential bounds $v_n(B_n) \le e^{-nr_1}$ and $\sup Q^n(S_n \mid \theta) \le e^{-nr_2}$ hold for all *large n*. Nevertheless, the necessity proof shows that it can be arranged that the exponential bounds hold for *all n*.

Clearly, Theorem 5 is a generalization of Proposition 1 in the introduction. As discussed there, the first conclusions of Propositions 2 and 3 readily follow from Theorem 5 by using the fact (from Barron 1987b) that in the stationary independent case there exists a uniformly exponentially consistent test against all Q with $d_{T_n}(Q,P) > \varepsilon$ if and only if T_n has effective cardinality of order n with respect to P.

We note that in some contexts the existence of a uniformly consistent test is equivalent to the existence of a uniformly *exponentially* consistent test. This surprising fact is shown in Schwartz (1965, Lemma 6.1) or LeCam (1973, Lemma 4) for the problem of testing a distribution P against a *fixed* set K of alternative distributions on $(\mathbf{X},\mathbf{B})$ using independent random variables.

These facts about uniformly exponentially consistent tests help motivate the insistence on exponential bounds in this theorem. Another motivation is the comparative ease of establishing the merging property $\lim P\{ m(X^n)/p(X^n) \geq a_n \} = 1$ for $a_n = e^{-n\varepsilon}$ than for sequences a_n which tend more slowly to zero. Nevertheless, the theorem may be modified to allow other rates of convergence.

We remark that whenever consistency holds as in equation (5), the merging property is equivalent to the condition that $\lim P\{ m(X^n,A_n)/p(X^n) \geq e^{-n\varepsilon}\} = 1$ for all $\varepsilon > 0$. In this case merging is seen to be a local property, depending only on the mixture of the distributions $Q^n(\cdot \mid \theta)$ for θ in A_n.

The proof of Theorem 5 follows readily from the next two lemmas. These lemmas utilize the following conditions for sequences of parameter sets A_n, B_n, C_n and constants a_n, b_n, c_n.

(a) Merging of P^n and M^n:

$$\lim_{n\to\infty} P\{ m(X^n)/p(X^n) \geq a_n \} = 1.$$

(b) Prior negligibility of B_n: $\nu_n(B_n) \leq b_n$.

(c) Existence of a uniformly consistent test against C_n:

$$\lim_{n\to\infty} P^n(S_n) = 0 \text{ and } \sup_{\theta\in C_n} Q^n(S_n^c \mid \theta) \leq c_n \text{ for some } S_n \in \mathbf{B}^n.$$

(d) Totality: $A_n \cup B_n \cup C_n = \Theta_n$.

Let conditions $(a)'$ and $(c)'$ be the same as conditions (a) and (c) except that $P\{ m(X^n)/p(X^n) < a_n \text{ infinitely often}\} = 0$ and $P\{ X^n \in S_n \text{ infinitely often}\} = 0$.

Lemma 6: *Suppose conditions (a), (b), (c), and (d), are satisfied with*

$\lim b_n = \lim c_n = 0$ *and let* $r_n = (b_n + c_n)/a_n$, *then for all* $\delta > 0$,

$$\limsup_{n\to\infty} P\{ v_n(A_n^c \mid X^n) > r_n/\delta\} \le \delta \tag{6}$$

If also $(a)'$ *and* $(c)'$ *are satisfied, then for any summable sequence* $\delta_n > 0$,

$$P\{ v_n(A_n^c \mid X^n) > r_n/\delta_n \text{ infinitely often}\} = 0. \tag{7}$$

Remarks: From (6) it is seen that if $r_n \to 0$, then $v_n(A_n^c \mid X^n)$ converges to zero in probability with rate arbitrarily close to r_n. Indeed, for any r'_n with $r_n = o(r'_n)$, equation (6) shows that $\limsup P\{ v_n(A_n^c \mid X^n) > r'_n\} \le \delta$, so letting $\delta \to 0$ we have $\lim P\{ v_n(A_n^c \mid X^n) > r'_n\} = 0$. Equation (7) is only useful to us if r_n tends tends to zero sufficiently fast that when divided by a summable sequence δ_n, the result still tends to zero.

To prove the sufficiency of the conditions in Theorem 5 we use Lemma 6 with $b_n = e^{-nr_1}$, $c_n = e^{-nr_2}$, $a_n = e^{-n\varepsilon}$, and $\delta_n = e^{-n\Delta}$ with $\varepsilon, \Delta > 0$ and $\varepsilon + \Delta < \min\{r_1, r_2\}$. Then r_n and $r'_n = r_n/\delta_n$ tend to zero exponentially fast.

Proof of Lemma 6: With P^∞ probability one, the densities $p(X^n)$ are greater than zero for all n and the posterior probability satisfies

$$v_n(A_n^c \mid X^n) = \frac{m(X^n, A_n^c)}{m(X^n)} = \frac{m(X^n, A_n^c)/p(X^n)}{m(X^n)/p(X^n)}$$

Consider the numerator: Let E_n be the event that $m(X^n, A_n^c)/p(X^n)$ is greater than $(b_n + c_n)/\delta$. For any sequence of measurable sets S_n in $\mathbf{B}^n$ we have $P^n(E_n) \le P^n(E_n \cap S_n^c) + P^n(S_n)$ and $P\{X^n \in E_n \text{ i.o.}\} \le P\{X^n \in (E_n \cap S_n^c) \text{ i.o.}\} + P\{X^n \in S_n \text{ i.o.}\}$ where *i.o.* is the abbreviation for infinitely often. In particular, take S_n to be critical sets which satisfy condition (c) and $B_n \in \mathbf{B}_{\Theta_n}$ to be parameter sets which satisfy condition (b). Then by Markov's inequality, the Fubini theorem for nonnegative integrands, and the inclusion of A_n^c in $B_n \cup C_n$, we have

$$\begin{aligned}
P^n(E_n \cap S_n^c) &\le \frac{\delta}{b_n + c_n} \int_{S_n^c} (m(\mathbf{x}^n, A_n^c)/p^n(\mathbf{x}^n))\, P^n(d\mathbf{x}^n) \\
&\le \frac{\delta}{b_n + c_n} \int_{A_n^c} Q^n(S_n^c \mid \theta)\, v_n(d\theta) \\
&\le \frac{\delta}{b_n + c_n} \left(\int_{B_n} v_n(d\theta) + \int_{C_n} Q^n(S_n^c \mid \theta)\, v_n(d\theta) \right) \\
&\le \frac{\delta}{b_n + c_n} (b_n + c_n) = \delta.
\end{aligned}$$

Consequently $P^n(E_n) \leq \delta + P^n(S_n)$ and limsup $P^n(E_n) \leq \delta$. If δ_n is chosen to be summable, then by the Borel-Cantelli Lemma $P\{X^n \in (E_n \cap S_n^c)\ i.o.\} = 0$; if also $P\{X^n \in S_n\ i.o.\} = 0$ then $P\{X^n \in E_n\ i.o.\} = 0$.

Finally consider the denominator: By condition (a) (respectively (a)'), the event that $m(X^n)/p(X^n)$ is less than a_n (infinitely often) has probability which tends to zero (equals zero). The results for the numerator and denominator are combined using the union of events bound. This completes the proof of Lemma 6.

Lemma 7: *If* $\lim P\{\nu_n(A_n^c \mid X^n) \leq r_n\} = 1$ *for some sequence of constants* r_n, *then for any* b_n *and* c_n *with* $b_n c_n \geq r_n$, *there exist sets* B_n *and* C_n *such that conditions (b), (c), and (d) are satisfied. Moreover, if* $P\{\nu_n(A_n^c \mid X^n) > r_n$ *infinitely often* $\} = 0$, *then conditions (b), (c)', and (d) are satisfied.*

Remark: This lemma demonstrates the necessity of the conditions in Theorem 5. Given $r_n = e^{-nr}$ as in the statement of the Theorem, simply set $b_n = e^{-nr_1}$ and $c_n = e^{-nr_2}$ for any $r_1, r_2 > 0$ with $r_1 + r_2 \leq r$, then Lemma 7 guarantees the existence of sets B_n and C_n which satisfy the properties required to complete the proof of Theorem 5.

Proof of Lemma 7: Set $S_n = \{\mathbf{x}^n : \nu_n(A_n^c \mid \mathbf{x}^n) > r_n\}$. The assumption of the Lemma is that $\lim P^n(S_n) = 0$, or moreover, that $P\{X^n \in S_n\ i.o.\} = 0$. We will use the fact that for all $\mathbf{x}^n$ in S_n^c, $m(\mathbf{x}^n, A_n^c) \leq r_n m(\mathbf{x}^n)$. (This inequality is trivially true whenever $m(\mathbf{x}^n)$ is zero, otherwise it is the same as $\nu_n(A_n^c \mid \mathbf{x}^n) \leq r_n$.)

Let $C_n = \{\theta : Q^n(S_n^c \mid \theta) \leq c_n\}$ and $B_n = \{\theta \in A_n^c : Q^n(S_n^c \mid \theta) > c_n\}$. Then conditions (c) and (d) are clearly satisfied. Moreover, by Markov's inequality and Fubini's Theorem

$$\begin{aligned}
\nu_n(B_n) &\leq \frac{1}{c_n} \int_{A_n^c} Q^n(S_n^c \mid \theta)\ \nu_n(d\theta) \\
&= \frac{1}{c_n} \int_{S_n^c} m(\mathbf{x}^n, A_n^c)\ \lambda^n(d\mathbf{x}^n) \\
&\leq \frac{r_n}{c_n} \int_{S_n^c} m(\mathbf{x}^n)\ \lambda^n(d\mathbf{x}^n) \\
&\leq \frac{r_n}{c_n}
\end{aligned}$$

$$\leq b_n.$$

Thus condition (b) is also satisfied. This completes the proof of Lemma 7.

4. Merging of Bayes and Frequentist Distributions

Continuing in the general framework adopted in the preceding section, we determine conditions on the sequence of priors for the merging of the distributions P^n and M^n in probability. Recall that the definition of merging given in section 3 is equivalent to the convergence of $(1/n)\log(p(X^n)/m(X^n))$ to zero in P probability. The conditions we use in fact ensure that $(1/n)\log(p(X^n)/m(X^n))$ converges to zero in $L^1(P)$. It is seen that this convergence is equivalent to convergence in information, i.e. $\lim (1/n)\, D(p^n \| m^n) = 0$. This convergence is established as a byproduct of an approximation to the Bayesian joint distribution of θ and X^n. After developing this approximation, we will also discuss some other conditions for the merging of P^n and M^n in probability and with probability one. In the stationary independent case these conditions are given implicitly in Schwartz (1965).

Let v_n^* be the distribution on the parameter space given by

$$v_n^*(d\theta) = \frac{e^{-nD_n(\theta)}\, v_n(d\theta)}{c_n}$$

where $D_n(\theta) = (1/n)D(p^n \| q^n(\cdot \mid \theta))$ and $c_n = \int e^{-nD_n(\theta)}\, v_n(d\theta)$ and let L_n^* be the joint product distribution for θ and X^n defined by

$$L_n^*(d\theta, d\mathbf{x}^n) = v_n^*(d\theta)P^n(d\mathbf{x}^n).$$

We consider this distribution as an approximation to the Bayesian joint law for θ and X^n. The Bayesian law is

$$L_n^{Bayes}(d\theta, d\mathbf{x}^n) = v_n(d\theta \mid \mathbf{x}^n)M^n(d\mathbf{x}^n).$$

For the next result we use the following generalization of condition (A) from section 1.

(A′) The sequence of relative entropy neighborhoods of P^n has prior probability which is not exponentially small, i.e. for every $\varepsilon, r > 0$ there is an N such that for all $n > N$

$$v_n\{\theta : D_n(\theta) < \varepsilon\} \geq e^{-nr}.$$

Note that this reduces to condition (A) in the stationary independent case with a fixed parametric family and prior $\nu_n = \nu$, because in that case $D_n(\theta) = D_1(\theta)$ is the same for all n.

Lemma 8: *If condition (A′) is satisfied then the Bayesian law L_n^{Bayes} and the frequentist approximation L_n^*, merge in information, in the sense that*

$$\lim_{n\to\infty} \frac{1}{n} D(L_n^* \,||\, L_n^{Bayes}) = 0.$$

Consequently, P^n and M^n merge in information (and hence in probability), i.e.

$$\lim_{n\to\infty} \frac{1}{n} D(p^n \,||\, m^n) = 0.$$

Also

$$\lim_{n\to\infty} \frac{1}{n} E(D(\nu_n^* \,||\, \nu_n(\cdot \,|\, X^n))) = 0.$$

Here the expectation E is with respect to the distribution P^n for X^n.

Remarks: The distribution ν_n^* is an interesting approximation to the posterior distribution $\nu_n(\cdot \,|\, X_n)$. The convergence of ν_n^* and $\nu_n(\cdot \,|\, X_n)$ in information demonstrates the necessity of the exponential factor $e^{-nD(\theta)}$. With any other exponent, the convergence would not hold. (However, because of the division by n, the nonexponential factors could be changed arbitrarily.) It is worthwhile to note that this large deviation approximation ν_n^* also matches the usual (small deviation) Gaussian approximation. Indeed, consider the stationary independent case and suppose that $\{q(\cdot \,|\, \theta) : \theta \in \Theta\}$ is a smooth finite dimensional parametric family ($\Theta \subset R^m$) with continuous Fisher information matrix I_θ and that the prior measure $\nu(d\theta)$ has a continuous density function $\nu'(\theta)$ with respect to Lebesgue measure. Suppose also that the true density function is $p(x) = q(x \,|\, \theta_0)$ for some θ_0 with $\nu'(\theta_0) > 0$ and $\det(I_{\theta_0}) > 0$. Our approximation to the posterior density function is

$$\frac{e^{-nD(\theta)}\, \nu'(\theta)}{c_n}.$$

It is well known that under regularity conditions the relative entropy $D(\theta) = D(q_{\theta_0} \,||\, q_\theta))$ is approximated by its second order Taylor expansion

$$D(\theta) = \frac{1}{2}(\theta - \theta_0)' I_{\theta_0}(\theta - \theta_0) + o(||\theta - \theta_0||^2) \quad \text{as} \quad ||\theta - \theta_0|| \to 0.$$

Note that this reduces to condition (A) in the stationary independent case with a fixed parametric family and prior $\nu_n = \nu$, because in that case $D_n(\theta) = D_1(\theta)$ is the same for all n.

Lemma 8: *If condition (A′) is satisfied then the Bayesian law L_n^{Bayes} and the frequentist approximation L_n^*, merge in information, in the sense that*

$$\lim_{n\to\infty} \frac{1}{n} D(L_n^* \,||\, L_n^{Bayes}) = 0.$$

Consequently, P^n and M^n merge in information (and hence in probability), i.e.

$$\lim_{n\to\infty} \frac{1}{n} D(p^n \,||\, m^n) = 0.$$

Also

$$\lim_{n\to\infty} \frac{1}{n} E(D(\nu_n^* \,||\, \nu_n(\cdot \,|\, X^n))) = 0.$$

Here the expectation E is with respect to the distribution P^n for X^n.

Remarks: The distribution ν_n^* is an interesting approximation to the posterior distribution $\nu_n(\cdot \,|\, X_n)$. The convergence of ν_n^* and $\nu_n(\cdot \,|\, X_n)$ in information demonstrates the necessity of the exponential factor $e^{-nD(\theta)}$. With any other exponent, the convergence would not hold. (However, because of the division by n, the nonexponential factors could be changed arbitrarily.) It is worthwhile to note that this large deviation approximation ν_n^* also matches the usual (small deviation) Gaussian approximation. Indeed, consider the stationary independent case and suppose that $\{q(\cdot \,|\, \theta) : \theta \in \Theta\}$ is a smooth finite dimensional parametric family ($\Theta \subset R^m$) with continuous Fisher information matrix I_θ and that the prior measure $\nu(d\theta)$ has a continuous density function $\nu'(\theta)$ with respect to Lebesgue measure. Suppose also that the true density function is $p(x) = q(x \,|\, \theta_0)$ for some θ_0 with $\nu'(\theta_0) > 0$ and $\det(I_{\theta_0}) > 0$. Our approximation to the posterior density function is

$$\frac{e^{-nD(\theta)}\, \nu'(\theta)}{c_n}.$$

It is well known that under regularity conditions the relative entropy $D(\theta) = D(q_{\theta_0} \,||\, q_\theta))$ is approximated by its second order Taylor expansion

$$D(\theta) = \frac{1}{2}(\theta - \theta_0)' I_{\theta_0}(\theta - \theta_0) + o(||\theta - \theta_0||^2) \quad \text{as} \quad ||\theta - \theta_0|| \to 0.$$

Thus for θ near θ_0 our approximation to the posterior density function is nearly the same as the joint Gaussian density function with mean θ_0 and covariance matrix $(nI_\theta)^{-1}$. (Indeed the approximate posterior density for $\xi = (\theta-\theta_0)\sqrt{n}$ converges to the Normal $(0,(I_\theta)^{-1})$ density as $n\to\infty$.) The point of these remarks is the that the relative entropy based approximation $e^{-nD(\theta)}\,\nu(d\theta)/c_n$ to the posterior distribution may be preferable to the Gaussian approximation in that while they are equally accurate for small deviations only the relative entropy based approximation is at least crudely accurate for large deviations.

Proof of Lemma 8: The Bayesian law $L_n^{Bayes}(dx^n,d\theta) = Q^n(dx^n \mid \theta)\nu_n(d\theta)$ has a joint density function $q(x^n \mid \theta)$ with respect to $\nu_n \times \lambda^n$, whereas the approximation L_n^* has a density function $p(x^n)e^{-nD_n(\theta)}/c_n$ with respect to $\nu_n \times \lambda^n$. Therefore,

$$D(L_n^* \,||\, L_n^{Bayes}) = E \log \frac{p(X^n)e^{-nD_n(\theta)}/c_n}{q(X^n \mid \theta)}$$

where E denotes expectation with respect to L_n^*. Taking an iterated expectation first with respect to P^n and then with respect to ν_n^*, it is seen that the result of the inner expectation does not depend on θ. In this manner the expression easily simplifies to

$$D(L_n^* \,||\, L_n^{Bayes}) = \log 1/c_n = -\log \int e^{-nD_n(\theta)}\,\nu_n(d\theta) \tag{8}$$

Now for any $\varepsilon > 0$ we obtain

$$\frac{1}{n}D(L_n^* \,||\, L_n^{Bayes}) \le \varepsilon - \frac{1}{n}\log \nu_n\{\theta : D_n(\theta) < \varepsilon\}$$

Condition (A') implies that as $n\to\infty$ the second term tends to zero for any $\varepsilon > 0$. Thus $\limsup(1/n)D(L_n^* \,||\, L_n^{Bayes}) \le \varepsilon$ and the limit is zero since ε may be chosen arbitrarily small. Now the relative entropy satisfies the chain rule

$$\frac{1}{n}D(L_n^* \,||\, L_n^{Bayes}) = \frac{1}{n}D(p^n \,||\, m^n) + \frac{1}{n}E(D(\nu_n^* \,||\, \nu_n(\cdot \mid X^n))) \tag{9}$$

so both of these terms must tend to zero. By a known inequality for the Kullback-Leibler number: $E \mid \log p(X^n)/m(X^n) \mid \le D(p^n \,||\, m^n) + 2/e$. (This follows from observing that the negative part of $(p/m)\log(p/m)$ is less than $1/e$, then taking the expectation with respect to M.) Consequently $E \mid (1/n)\log p(X^n)/m(X^n) \mid$ tends to zero. It follows that P^n and M^n merge in probability. This completes the proof of Lemma 8.

Remark: Note that the identities (8) and (9) yield the bound

product sigma-field $\mathbf{B}^\infty$ and assume that the random samples are defined by $X^n(\omega) = (x_1,\ldots,x_n)$ for $\omega = (x_1,x_2,\ldots)$ in $\mathbf{X}^\infty$. Suppose that P is a stationary and ergodic probability measure on $\mathbf{X}^\infty$ and for each n, the distribution P^n for X^n is dominated by λ^n (which is here assumed to be the $n-$fold product of a sigma-finite measure λ on $\mathbf{X}$). The parameter space Θ is fixed. For each θ and for $n > k(\theta)$, the probability measures $Q^n(\cdot \mid \theta)$ are assumed to be Markov with a stationary transition measure $Q(\cdot \mid x_1,\ldots,x_k,\theta)$ having a conditional density function $q(x_{k+1} \mid x_1,\ldots,x_k,\theta)$ with respect to λ. The Markov order $k = k(\theta) \geq 0$ may depend on θ and may be arbitrarily large.

Lemma 10: *In this stationary and ergodic case, if condition* $(\tilde{A})$ *is satisfied, then* P^n *and* M^n *merge with probability one, that is*

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{p(X^n)}{m(X^n)} = 0, \text{ with } P \text{ probability one.}$$

Proof: Given any $\varepsilon > 0$, let $A_\varepsilon = \{\theta : \bar{D}(\theta) < \varepsilon/2\}$. For each θ in A_ε and each n, P^n is absolutely continuous with respect to $Q^n(\cdot \mid \theta)$ with density ratio $p(X^n)/q(X^n \mid \theta)$. Set $D_n(X^n,\theta) = (1/n) \log (p(X^n)/q(X^n \mid \theta))$. By the ergodic theorem for densities (Barron 1985a),

$$\lim_{n\to\infty} D_n(X^n,\theta) = \bar{D}(\theta) \tag{10}$$

with P probability one, for each θ in A_ε. It follows by Fubini's Theorem that there is a set G in $\mathbf{B}^\infty$ such that for each sequence in G, the limit (10) holds for almost every θ in A_ε. Let $\rho_n(\theta) = e^{n\varepsilon/2}(d\nu_n/d\tilde{\nu})(\theta)$. Now

$$e^{n\varepsilon} \frac{m(X^n)}{p(X^n)} \geq e^{n\varepsilon/2} \int_{A_\varepsilon} \frac{q(X^n \mid \theta)}{p(X^n)} \rho_n(\theta)\, \tilde{\nu}(d\theta)$$

$$= \int_{A_\varepsilon} e^{n(\varepsilon/2 - D_n(X^n,\theta))} \rho_n(\theta)\, \nu(d\theta).$$

The integrand in this expression is positive and tends infinity for ν almost every θ in A_ε, for sequences in a set of probability one. Also $\nu(A_\varepsilon) > 0$ by condition $(\tilde{A})$, so by Fatou's Lemma

$$\liminf_{n\to\infty} e^{n\varepsilon} \frac{m(X^n)}{p(X^n)} = \infty \text{ with P probability one.} \tag{11}$$

Hence $e^{n\varepsilon} m(X^n)/p(X^n) \geq 1$ for all large n, with probability one. Thus P^n and M^n merge with probability one according to the definition in section 3. As indicated there, Markov's inequality and the Borel-Cantelli lemma show that

$e^{-n\varepsilon}m(X^n)/p(X^n) \le 1$ for all large n, with probability one, and together these facts yield $\lim (1/n) \log m(X^n)/p(X^n) = 0$. This completes the proof of Lemma 10.

Remarks: It is shown in Barron (1985a) that if P is stationary and Q_θ is Markov with stationary transitions, then when the relative entropy rate $\overline{D}(\theta)$ is finite, it may be expressed as

$$\overline{D}(\theta) = E \log \frac{p(X_{k+1} | X_1,\ldots,X_k)}{q(X_{k+1} | X_1,\ldots,X_k,\theta)} + I_k(P) \qquad (12)$$

for any $k \ge k(\theta)$. The first term in (12) is simply the expected relative entropy between the k^{th} order conditional densities, i.e. $E(D(p(\cdot | X^k) \,||\, q(\cdot | X^k,\theta)))$, where E denotes expectation with respect to P. The term $I_k(P)$ does not depend on θ and decreases to zero as $k \to \infty$. (This I_k is the Shannon mutual information between X_{k+1} and the infinite past given $X_1,\ldots,X_k$, see Barron 1985a). Consequently, condition $(\tilde{A})$ is satisfied in this context if for every $\varepsilon > 0$

$$\tilde{\nu}\left\{\theta : E(D(p(\cdot | X^k) \,||\, q(\cdot | X^k,\theta))) < \frac{\varepsilon}{2},\ k(\theta) \le k\right\} > 0.$$

where k is sufficiently large that $I_k(P) \le \varepsilon/2$. It should be possible to devise families of Markov processes $\{Q(\cdot | \theta)\}$ for which this condition is satisfied for a large class of stationary ergodic distributions P.

For completeness we give another condition which implies merging in probability which is used in the work of Schwartz (1965) and LeCam (1973,1982,1986). This condition uses the total variation distance instead of the relative entropy. As in condition (A') it is required that a sequence of parameter sets have prior probability that is not exponentially small. Unfortunately, with the total variation condition the sequence of parameter sets must shrink rapidly to zero, (whereas the relative entropy condition involves sets which do not change with n). The examples given in section 8 suggest that the relative entropy condition is usually easier to verify.

Define $M^n(\cdot | A) = M^n(\cdot, A)/\nu_n(A) = \int_A Q^n(\cdot | \theta)\nu_n(d\theta)/\nu_n(A)$ for parameter sets A in $\mathcal{B}_{\Theta_n}$. This is the Bayesian conditional distribution for X^n given that θ is in A. We are not restricted to the stationary case and the family and prior may change with n.

Lemma 11: *For P^n and M^n to merge in probability it is sufficient that* $\lim d(P^n, M^n(\cdot | A_n)) = 0$ *for some sequence of parameter sets A_n for which the prior probability is not exponentially small (i.e.* $\lim 1/n \log \nu_n(A_n) = 0$*). In particular it is sufficient that for every $\varepsilon > 0$ the following prior probability is not exponentially small*

$$\nu_n\{\theta : d(P^n, Q_\theta^n) < \varepsilon\}. \tag{13}$$

Remark: In the stationary independent case, $d(P^n,Q^n) \le n\, d(P^1,Q^1)$ whence it is sufficient that the probability $\nu_n\{\theta : d(P^1,Q_\theta^1) < \varepsilon/n\}$ is not exponentially small. If $\nu_n = \nu$ does not depend on n, then using an argument similar to Strasser (1981b, Lemma 3.10) it is seen that an interesting sufficient condition for this property is *diametric regularity:* $\nu(B_{2a}) < c\,\nu(B_a)$ for all $a > 0$ for some $c > 1$, where $B_a = \{\theta : d(P^1,Q_\theta^1) < a\}$. Diametric regularity implies that (with $\alpha = \log_2 c$) the probabilities satisfy $\nu(B_{\varepsilon/n}) \ge (1/c)(\varepsilon/n)^\alpha\, \nu(B_1)$ which is clearly not exponentially small.

Proof of Lemma 11: Given A_n in $\mathbf{B}_{\Theta_n}$, set $r_n = -(1/n)\log \nu_n(A_n)$. If A_n satisfies the conditions of the lemma then $\lim r_n = 0$. To show that P^n and M^n merge in probability it is enough that the set $\{x^n : m(x^n) \le p(x^n)e^{-nr_n}/2\}$ has P^n probability which tends to zero. This set is the same as $\{x^n : m(x^n)/\nu_n(A_n) \le p(x^n)/2\}$ which is contained in $S_n = \{x^n : m(x^n \mid A_n) \le p(x^n)/2\}$. (Here $m(x^n \mid A_n)$ is the density function for $M^n(\cdot \mid A_n)$ with respect to λ^n.) Clearly, $M^n(S_n \mid A_n) \le P^n(S_n)/2$. Now the variation distance satisfies $d(P^n, M^n(\cdot \mid A_n)) \ge 2(P^n(S_n) - M^n(S_n \mid A_n))$ which is not less than $P^n(S_n)$. Thus $\lim P^n(S_n) = 0$ and so P^n and M^n merge in probability. This completes the proof of the first claim.

Now consider the total variation neighborhoods as in expression (13). If $\lim (1/n)\log \nu_n\{\theta : d(P^n, Q^n(\cdot \mid \theta)) < \varepsilon\} = 0$ for every $\varepsilon > 0$, then there exists a sequence ε_n which tends to zero sufficiently slowly that when ε_n is substituted for ε the limit is still zero. Let $A_n = \{\theta : d(P^n, Q^n(\cdot \mid \theta)) < \varepsilon_n\}$. Now $M^n(\cdot \mid A_n)$ is the average of $Q^n(\cdot \mid \theta)$ for θ in A_n, so by the convexity of the total variation distance we have

$$d(P^n, M^n(\cdot \mid A_n)) \le \frac{1}{\nu_n(A_n)} \int_{A_n} d(P^n, Q^n(\cdot \mid \theta))\, \nu_n(d\theta) \le \varepsilon_n.$$

So it is verified that $\lim d(P^n, M^n(\cdot \mid A_n)) = 0$ and $\lim (1/n)\log \nu_n(A_n) = 0$, consequently P^n and M^n merge in probability. This completes the proof of Lemma 11.

5. Consistency of Estimators

For a non-Bayesian such as the author, perhaps the most important reason to examine posterior probabilities is to obtain results on the accuracy of estimators. Convergence results for Bayes estimators can be obtained from convergence results

for posterior probabilities. In this section we explore methods which show how the loss of an estimate is bounded in terms of posterior probabilities. One method assumes that the loss function satisfies the triangle inequality or that it is topologically equivalent to a loss function which satisfies the triangle inequality. Another method does not require the triangle inequality but does assume that the loss is a convex function and that the estimator is the posterior mean density. Finally, a separate argument is used to handle the maximizer of the posterior probability in the case of discrete prior.

The results are presented in a somewhat general decision-theoretic setting which suits our needs. Let $(\Omega, \mathbf{B}_\Omega)$ be a measurable space, let Ξ be a set of probability measures on this space, and let P be a fixed probability measure in this set. For each $n \geq 1$, let X^n be a random sample taking values in a measurable space $(\mathbf{X}^n, \mathbf{B}^n)$, with induced distribution P^n. Let Φ be a set and suppose that to each Q in Ξ there is assigned a $\phi_Q \in \Phi$. Here ϕ is usually an attribute of the distribution (such as the probability density function for the X_i) which is to be estimated from observations $\mathbf{x}^n$ and Φ is the set of possible decisions.

Let $L_n(\phi, \hat{\phi})$ and $d_n(\phi, \hat{\phi})$ denote sequences of non-negative loss functions on $\Phi \times \Phi$. We assume that d_n is a pseudo-metric on $\Phi \times \Phi$ (i.e. $d_n(\phi, \hat{\phi})$ is non-negative, equal to zero if $\phi = \hat{\phi}$, symmetric $d_n(\phi, \hat{\phi}) = d_n(\hat{\phi}, \phi)$, and satisfies the triangle inequality $d_n(\phi, \hat{\phi}) \leq d_n(\bar{\phi}, \phi) + d_n(\bar{\phi}, \hat{\phi})$ for all ϕ, $\hat{\phi}$, and $\bar{\phi}$ in Φ).

For each $n \geq 1$, let $\{Q_\theta : \theta \in \Theta_n\}$ be a family of probability measures in Ξ for which the induced distributions of the random variable X^n are Q_θ^n and let $\nu_n(d\theta)$ be prior distributions on the parameter space $(\Theta_n, \mathbf{B}_{\Theta_n})$. Assume that there exists a regular posterior distribution $\nu_n(d\theta \mid \mathbf{x}^n)$ given that $X^n = \mathbf{x}^n$. (This means that $\nu_n(\cdot \mid \mathbf{x}^n)$ is a probability measure for each $\mathbf{x}^n$, and $\nu_n(A \mid \mathbf{x}^n)$ is an $\mathbf{B}^n$ measurable function for each A in $\mathbf{B}_{\Theta_n}$; these properties are guaranteed if the family of distributions $Q^n(\cdot \mid \theta)$ is dominated by a sigma-finite measure λ^n and the density $q(\mathbf{x}^n \mid \theta)$ may be chosen to be a $\mathbf{B}^n \times \mathbf{B}_{\Theta_n}$ measurable function as was assumed in the previous sections.) Let ϕ_θ be an abbreviated notation for ϕ_{Q_θ}. We require that $d_n(\phi_\theta, \phi)$ and $L_n(\phi_\theta, \phi)$ be $\mathbf{B}_{\Theta_n}$ measurable functions for each $\phi \in \Phi$. (For the loss functions we work with, this measurability follows from the joint measurability of the densities $q(\mathbf{x}^n \mid \theta)$.)

A decision $\hat{\phi}_n = \hat{\phi}(\cdot\,; \mathbf{x}^n)$ in Φ is a *Bayes estimate* of ϕ (for the loss function L_n, prior ν_n, and data $\mathbf{x}^n$) if it achieves

$$\min_{\hat{\phi} \in \Phi} \int L_n(\phi_\theta, \hat{\phi})\, \nu_n(d\theta \mid \mathbf{x}^n).$$

In some cases a minimum does not exist. To handle such cases we say that $\hat{\phi}(\cdot\,;\mathbf{x}^n)$ is an *approximate* Bayes estimate if it achieves within δ_n of the infimum, where $\lim \delta_n = 0$. Of course, approximate Bayes estimators are also useful when Bayes estimators cannot be computed exactly.

For the next result we require that the loss function L_n be topologically equivalence to the pseudo-metric d_n. Specifically assume that there exist strictly increasing functions $f(u)$ and $g(u)$, $u \geq 0$, continuous at 0, with $f(0) = g(0) = 0$, such that $L_n \leq f(d_n)$ and $d_n \leq g(L_n)$ for every pair of decisions in Φ_n and all n. This means that $\lim L_n(\phi_n,\hat{\phi}_n) = 0$ if and only if $\lim d_n(\phi_n,\hat{\phi}_n) = 0$.

Set $\phi = \phi_P$ in Φ and assume that there exists a bound $\bar{L} < \infty$ such that $L_n(\phi_\theta,\phi) \leq \bar{L}$ for all θ in Θ_n and all n. For $\varepsilon > 0$ let

$$A_{\varepsilon,n} = \{\theta : d_n(\phi,\phi_\theta) \leq \varepsilon\}.$$

Lemma 12: *For any* $\mathbf{x}^n$, let $\hat{\phi}(\cdot\,;\mathbf{x}^n)$ *be a Bayes estimate (or an approximate Bayes estimate) for a bounded loss function* L_n *which is equivalent to a pseudo-metric* d_n, *with* $f, g, \bar{L}, \delta_n$ *as above, then for every* $\varepsilon > 0$

$$d_n(\phi,\hat{\phi}_n) \leq \varepsilon + g\left[\frac{\varepsilon + \bar{L}\, \nu_n(A^c_{\varepsilon',n} \mid \mathbf{x}^n) + \delta_n}{1 - \nu_n(A^c_{\varepsilon,n} \mid \mathbf{x}^n)}\right] \quad . \tag{15}$$

where $\varepsilon' = f^{-1}(\varepsilon)$. *Consequently, for any sequence* $\mathbf{x}^n$ *for which* $\lim \nu_n(A^c_{\varepsilon,n} \mid \mathbf{x}^n) = 0$ *for all* $\varepsilon > 0$ *we have*

$$\lim_{n\to\infty} d_n(\phi,\hat{\phi}_n) = 0.$$

Consequently, if $\lim \nu(A^c_{\varepsilon,n} \mid X^n) = 0$ in probability or with probability one for every $\varepsilon > 0$, then $\lim d_n(\phi,\hat{\phi}_n) = 0$ in probability or with probability one, respectively. (Observe that the measurability of $d_n(\phi,\hat{\phi}(\cdot;\mathbf{x}^n))$ need is not be required since (15) provides a measurable upper bound for establishing the convergence. Thus setting $S_n = \{\mathbf{x}^n : L_n(\phi,\hat{\phi}) \geq \varepsilon\}$, for convergence in probability it is meant that the *outer* P^n probability of S_n tends to zero and for almost sure convergence it is meant that the *outer* P probability of $\{X_n \in S_n \ i.o.\}$ is zero for every $\varepsilon > 0$.)

Proof of Lemma 12: If $\hat{\phi}(\cdot\,;\mathbf{x}^n)$ is an approximate Bayes estimate for the loss function L_n, then

$$\begin{aligned}\int L_n(\hat{\phi}_n,\phi_\theta)\, \nu_n(d\theta \mid \mathbf{x}^n) &\leq \int L_n(\phi,\phi_\theta)\, \nu_n(d\theta \mid \mathbf{x}^n) + \delta_n \\ &\leq \varepsilon\, \nu(B \mid \mathbf{x}^n) + \bar{L}\, \nu(B^c \mid \mathbf{x}^n) + \delta_n\end{aligned}$$

$$\leq \varepsilon + \bar{L}\, \nu(A^c_{\varepsilon',n} \mid \mathbf{x}^n) + \delta_n$$

where $B = \{\theta : L_n(\phi,\phi_\theta) \leq \varepsilon\}$ contains $\{\theta : f(d_n(\phi,\phi_\theta)) \leq \varepsilon\} = A_{\varepsilon',n}$.

On the other hand, by the triangle inequality we have for θ in $A_{\varepsilon,n}$ that $d_n(\phi,\hat{\phi}_n) \leq \varepsilon + d_n(\phi_\theta,\hat{\phi}_n)$ which is not greater than $\varepsilon + g(L_n(\hat{\phi}_n,\phi_\theta))$. Consequently

$$\begin{aligned}\int L_n(\hat{\phi}_n,\phi_\theta)\, \nu_n(d\theta \mid \mathbf{x}^n) &\geq \int_{A_{\varepsilon,n}} L_n(\hat{\phi}_n,\phi_\theta)\, \nu_n(d\theta \mid \mathbf{x}^n) \\ &\geq \int_{A_{\varepsilon,n}} g^{-1}(d_n(\phi,\hat{\phi}_n) - \varepsilon)\, \nu_n(d\theta \mid \mathbf{x}^n) \\ &= g^{-1}(d_n(\phi,\hat{\phi}_n) - \varepsilon)\, \nu_n(A_{\varepsilon,n} \mid \mathbf{x}^n).\end{aligned}$$

Combining these bounds yields

$$d_n(\phi,\hat{\phi}_n) \leq \varepsilon + g\left[\frac{\varepsilon + \bar{L}\, \nu_n(A^c_{\varepsilon',n} \mid \mathbf{x}^n) + \delta_n}{1 - \nu_n(A^c_{\varepsilon,n} \mid \mathbf{x}^n)}\right]$$

which is the desired inequality. This completes the proof of Lemma 12.

Remarks: Consider the estimation of probability density functions using independent random variables as discussed in section 1. (In this case $(\mathbf{X}^n,\mathbf{B}^n)$ is the n-fold product of the space $(\mathbf{X},\mathbf{B})$ and Ξ is the set of probability measures on the underlying space $(\Omega,\mathbf{B}_\Omega)$ for which the distribution of $X_1,X_2,\ldots,X_n$ is independent with a coordinate density function which is absolutely continuous with respect to $\lambda(dx)$.) Let $d(p,q) = \int |p(x)-q(x)|\, \lambda(dx)$ be the L^1 distance between density functions. Conditions have been given which imply that $\lim \nu_n(A^c_\varepsilon \mid X^n) = 0$ with probability one. Consequently, if $\hat{p}_n$ is a Bayes estimator with bounded loss function L which is equivalent to d, then the same conditions imply that $\lim d(p,\hat{p}_n) = 0$ with probability one. This shows how Proposition 2 is obtained for such estimators.

For example, if L is the squared Hellinger distance $\int (\sqrt{p}-\sqrt{q})^2$, the Bayes estimator $\hat{p}_n(x)$ is the square of the mean of the posterior distribution of $\sqrt{q(x \mid \theta)}$, assuming that Φ consists of all non-negative measurable functions $\hat{p}$ for which $\int \hat{p}(x)\lambda(dx) \leq 1$. It is known that the squared Hellinger distance is equivalent to the L^1 distance on $\Phi\times\Phi$ with $\int (\sqrt{p}-\sqrt{q})^2 \leq \int |p-q| \leq 2(\int (\sqrt{p}-\sqrt{q})^2)^{1/2}$. As another example, if $L=d$ is the L^1 distance, the Bayes estimator is a median of the posterior distribution of $q(x \mid \theta)$ (assuming that a version may be chosen which is a measurable function of x). In this latter case we must take Φ to be all non-negative functions which are integrable with respect to λ. Proposition 2 clearly applies to either of these estimators.

For the particular case of the posterior mean estimator of a probability distribution, we use a specialized argument to relate the asymptotics of posterior probabilities to the consistency of the estimator. Let Y be a random variable: a measurable mapping from $(\Omega,\mathbf{B}_\Omega)$ into a measurable space $(\mathbf{Y},\mathbf{B}_\mathbf{Y})$. Let Φ be the set of probability measures on $(\mathbf{Y},\mathbf{B}_\mathbf{Y})$ and let ϕ_Q be the probability distribution for Y induced by any Q in Ξ. Although we risk abusing notation, we write P and Q_θ respectively for ϕ_P and ϕ_{Q_θ} in what follows. The posterior mean estimator based on a prior ν_n and data $\mathbf{x}^n$ is $\hat{P}_n(B;\mathbf{x}^n) = \int Q_\theta(B)\,\nu_n(d\theta\,|\,\mathbf{x}^n)$ for all B in $\mathbf{B}_\mathbf{Y}$. The accuracy of the estimate is measured by a sequence of loss functions $L_n(P,\hat{P}_n)$ on $\Phi\times\Phi$. Although both P and $\hat{P}_n$ are proper probability measures we consider loss functions L_n for which the domain of definition extends in a useful way to $\Phi\times\tilde{\Phi}$ where $\tilde{\Phi}$ is the set of *sub*-probability measures on $(\mathbf{Y},\mathbf{B}_\mathbf{Y})$. We assume that $L_n(P,Q_\theta)$ is $\mathbf{B}_{\Theta_n}$ measurable function for each $0 < a \le 1$ and each P in Φ_1.

The properties that we require of loss functions L on $\Phi\times\tilde{\Phi}$ are the following.

(i) Monotonicity: $Q_2 \ge Q_1$ implies $L(P,Q_2) \le L(P,Q_1)$

(ii) Convexity: $L(P,\alpha_1Q_1 + \alpha_2Q_2) \le \alpha_1L(P,Q_1) + \alpha_2L(P,Q_2)$

(iii) Scaling: $L(P,aQ) \le L(P,Q) + \rho(a)$ where $\lim_{a\to1}\rho(a) = 0$.

Here Q, Q_1, Q_2 are arbitrary subprobabilities in $\tilde{\Phi}$, α_1,α_2 are nonnegative with $\alpha_1 + \alpha_2 = 1$, $0 \le a \le 1$ and the function ρ does not depend on n. To handle a degenerate case we set $\rho(0) = \infty$. For subprobabilities, $Q_2 \ge Q_1$ means that $Q_2(B) \ge Q_1(B)$ for all B in $\mathbf{B}_\mathbf{Y}$.

Examples of loss functions $L(P,Q)$ which satisfy these properties include $\int p \log p/q$ (which is the relative entropy), $2\int (p-q)^+$ (which reduces to the L^1 distance when restricted to proper probabilities) and $2(1-\int \sqrt{p}\sqrt{q})$ (which reduces to the squared Hellinger distance when restricted to probabilities). Here p and q are the density functions with respect to any measure which dominates both P and Q. In these examples $\rho(a)$ for $0 < a \le 1$ is given by $-\log a$, $2(1-a)^+$ and $2(1-\sqrt{a})$, respectively.

Lemma 13: *Let L_n be a sequence of non-negative loss functions with extensions to $\Phi\times\tilde{\Phi}$ which satisfy the three properties (i),(ii), and (iii). Then for any $\mathbf{x}^n$ and any P in Φ, the posterior mean estimator $\hat{P}_n(\cdot\,;\mathbf{x}^n)$ satisfies*

$$L_n(P,\hat{P}_n) \le \varepsilon + \rho(\nu_n(A_{\varepsilon,n}\,|\,\mathbf{x}^n)) \text{ for all } \varepsilon > 0$$

where $A_{\varepsilon,n} = \{\theta : L_n(P,Q_\theta) < \varepsilon\}$. *Thus for any sequence* $x^n, n=1,2,...$

$$\nu_n(A_{\varepsilon,n} \mid x^n) \to 0 \text{ for all } \varepsilon > 0 \quad \text{implies} \quad L_n(P,\hat{P}_n) \to 0.$$

Remarks: In particular for the density estimation problem in the context of section 1, take Y to be an independent copy of X_1, and let d be the L^1 distance between probability densities (which may be extended to a loss L on $\Phi\times\tilde{\Phi}$ as indicated above), then conditions (A) and (B) imply that $\lim \nu_n(A_\varepsilon \mid X^n) = 0$ and hence $\lim d(p,\hat{p}_n) = 0$ with probability one. This show how Proposition 2 is obtained for the posterior mean estimator of the density function.

To establish Proposition 3 use the same reasoning with $d_{T_n}(P,Q) = \sum_{A\in T_n} |P(A)-Q(A)|$ (for proper probabilities) which extends to $L_n(P,Q) = 2\sum(P(A)-Q(A))^+$ on $\Phi\times\tilde{\Phi}$. Condition (A) alone implies that for this sequence of loss functions $\lim \nu_n(A_{\varepsilon,n} \mid X^n) = 0$ and hence $\lim d_{T_n}(P,\hat{P}_n) = 0$ with probability one.

Proof of Lemma 13: Given $\varepsilon > 0$ and x^n, set $a = \nu_n(A_{\varepsilon,n} \mid x^n)$. If $a = 0$ the inequality is trivially satisfied. Suppose $a > 0$ and let $\nu_n(d\theta \mid x^n, A_{\varepsilon,n})$ be the distribution obtained by conditioning on $\theta \in A_{\varepsilon,n}$. Successive application of monotonicity, convexity, scaling and the definition of A_ε yields

$$\begin{aligned}
L^n(P,\hat{P}_n) &\le L_n(P, \textstyle\int_{A_{\varepsilon,n}} Q_\theta \, \nu_n(d\theta \mid x^n)) \\
&= L_n(P, \textstyle\int_{A_{\varepsilon,n}} (a\, Q_\theta) \, \nu_n(d\theta \mid x^n, A_{\varepsilon,n})) \\
&\le \textstyle\int_{A_{\varepsilon,n}} L_n(P, a\, Q_\theta) \, \nu_n(d\theta \mid x^n, A_{\varepsilon,n}) \\
&\le \textstyle\int_{A_{\varepsilon,n}} (L_n(P,Q_\theta) + \rho(a)) \, \nu_n(d\theta \mid x^n, A_{\varepsilon,n}) \\
&\le \varepsilon + \rho(a).
\end{aligned}$$

Which is the desired result. This completes the proof of Lemma 14.

6. Discrete Priors

Next we demonstrate the consistency of maximum posterior likelihood estimators in the case of a countable parameter space Θ.

Let $\nu_n(\theta)$ be a sequence of priors (discrete mass functions) on Θ with $\sum_\theta \nu_n(\theta) \le 1$. It is assumed that the sequence of prior probabilities of each θ is not exponentially small (i.e. $\liminf_n e^{nr}\nu_n(\theta) \ge 1$ for every $r > 0$, for every $\theta \in \Theta$). In which case, the information denseness condition ($\tilde{A}$) reduces to an assumption on the

countable family: namely, that $\inf \{ \overline{D}(\theta) : \theta \in \Theta \} = 0$. (The measure $\tilde{\nu}$ required by definition 4 in section 4 may be taken to be any strictly positive probability mass function on Θ. Here $\overline{D}(\theta)$ reduces to $D(p \,||\, q_\theta)$ in the stationary independent case.) Thus the set of densities p for which the sequence of priors is information dense consists of all information limits of the countable family $\{q_\theta : \theta \in \Theta\}$.

The posterior likelihood function for θ given data $\mathbf{x}^n$ is proportional to the joint likelihood $\nu_n(\theta) q(\mathbf{x}^n \mid \theta)$. Recall that an estimator $\hat{\theta}_n = \hat{\theta}_n(\mathbf{x}^n)$ taking values in Θ is said to be an approximate maximum posterior likelihood estimator if

$$\nu_n(\hat{\theta}_n) q(\mathbf{x}^n \mid \hat{\theta}_n) > \sup_\theta \nu_n(\theta) q(\mathbf{x}^n \mid \theta)\, e^{-n\delta_n}$$

for all $\mathbf{x}^n$, where $\lim \delta_n = 0$. We require that $\hat{\theta}_n(\mathbf{x}^n)$ be a measurable function of $\mathbf{x}^n$. The corresponding density estimator is $\hat{p}_n(\cdot) = q(\cdot \mid \hat{\theta}_n)$. For Lemma 14 and Theorem 15 we assume a stationary independent model, whence $q(\mathbf{x}^n \mid \theta) = \prod_{i=1}^n q(x_i \mid \theta)$.

The following Lemma is basic to examining asymptotics for maximum posterior likelihood estimators.

Lemma 14: *Let $\hat{\theta}_n$ be an approximate maximum posterior likelihood estimator and let A_n be a sequence of measurable parameter sets. If the sequence of priors is information dense at p, and if $\nu_n(A_n^c \mid X^n)$ is exponentially small with probability one, then $\hat{\theta}_n \in A_n$ for all large n, with probability one.*

Remark: The lemma applies to the fixed parameter set $A_n = \{\theta : d(p, q_\theta) < \varepsilon\}$ (for each $\varepsilon > 0$), under the assumption that conditions (A) and (B) are satisfied, to obtain that for any approximate maximum posterior likelihood estimator $\lim d(p, \hat{p}_n) = 0$ with probability one. Similarly, using $A_n = \{\theta : d_{T_n}(P, Q_\theta) < \varepsilon\}$ one obtains that if only condition (A) is satisfied then $\lim d_{T_n}(P, \hat{P}_n) = 0$. This shows how the second conclusions of Propositions 2 and 3 are obtained in this context.

Proof: It is enough to show that with P probability one

$$\sup_{\theta \in A_n^c} \nu_n(\theta) q(X^n \mid \theta) < \sup_\theta \nu_n(\theta) q(X^n \mid \theta)\, e^{-n\delta_n} \tag{16}$$

for all large n, for then $\nu(\hat{\theta}_n) q(X^n \mid \hat{\theta}_n)$ is strictly larger than the posterior likelihood for all $\theta \in A_n^c$ and hence $\hat{\theta}_n \in A_n$.

Now when P^n and M^n merge with probability one, the exponential convergence of $\nu_n(A_n^c \mid X^n)$ is equivalent to $\sum_{\theta \in A_n^c} \nu_n(\theta) q(X^n \mid \theta) < p(X^n) e^{-nr}$ for all large n, with probability one, for some $r > 0$. Choose θ^* in Θ for which $D(p \,||\, q_{\theta^*}) < r/4$.

Then by the strong law of large numbers (and the fact that $v_n(\theta^*)$ is not exponentially small) it is seen that $p(X^n) < v_n(\theta^*)q(X^n \mid \theta^*)\, e^{nr/2}$ for all large n with P probability one. Combining these bounds yields

$$\begin{aligned}\sup_{\theta \in A_n^c} v_n(\theta)q(X^n \mid \theta) &\leq \sum_{\theta \in A_n^c} v_n(\theta)q(X^n \mid \theta) \\ &< p(X^n)\, e^{-nr} \\ &< v_n(\theta^*)q(X^n \mid \theta^*)\, e^{-nr/2} \\ &< \sup_{\theta} v_n(\theta)q(X^n \mid \theta)\, e^{-n\delta_n}\end{aligned}$$

for all large n, with P probability one. So (16) is satisfied and this completes the proof of Lemma 14.

The following result gives perhaps the simplest conditions which guarantee Bayes consistency in the case of a countable prior.

Theorem 15: *Suppose that for each* θ, *the sequence of prior probabilities* $v_n(\theta)$ *is not exponentially small. Also suppose that for some* $0 < \alpha < 1$, *the sum* $c_n = \sum_\theta (v_n(\theta))^\alpha$ *is not exponentially large, i.e.* $\lim (1/n)\log c_n = 0$. *Then for any density function* p *which is an information limit of the family* $\{q_\theta\}$, *we have for every* $\varepsilon > 0$

$$v_n(\{\theta : d(p,q_\theta) \geq \varepsilon\} \mid X_1,\ldots,X_n) \text{ is exponentially small, with } P \text{ probability one,}$$

and consequently, any approximate maximum posterior likelihood density estimator $\hat{p}_n$ *converges to* p *in* L^1, *i.e.,*

$$\lim_{n\to\infty} d(p,\hat{p}_n) = 0, \text{ with } P \text{ probability one.}$$

Remark: For a fixed prior $v(\theta)$ the summability condition is simply that $c = \sum_\theta (v(\theta))^\alpha$ is finite for α in a neighborhood of 1. This is equivalent to asserting that $\log 1/v(\theta)$ has a finite moment generating function in a neighborhood of zero. Similar moment generating function assumptions are necessary in other large deviation contexts to obtain exponential bounds. It is an interesting open question whether the condition $\sum (v(\theta))^\alpha < \infty$ is *necessary* as well as sufficient for the conclusions of Theorem 15 to hold for all p for which the prior is information dense.

Proof of Theorem 15: Let $B_n = \{\theta : v_n(\theta) < \exp\{-n\varepsilon/4\}\}$, then $v_n(B_n) < c_n \exp\{-n(1-\alpha)\varepsilon/4\}$. Thus $v_n(B_n)$ is exponentially small. Now let $A = \{\theta : d_H(p,q_\theta) \leq \varepsilon\}$ where $d_H(p,q) = \int (\sqrt{p} - \sqrt{q})^2$ is the squared Hellinger distance. Let $C_n = \{\theta \in A_n^c : v_n(\theta) \geq e^{-n\varepsilon/4}\}$ and note that since $\sum_{\theta \in C_n} v_n(\theta) \leq 1$, the number of points in C_n is less than $e^{n\varepsilon/4}$. We show that there exists a uniformly

defined by $\nu_n(\theta) = w(C(\theta,2^{-b_n}))$, where $C(\theta,2^{-b})$ is the cube which has lower corner at θ and sides of length 2^{-b}. It is seen that if $\lim b_n = \infty$ and $\lim b_n/n = 0$, then for each θ in Θ, $\nu_n(\theta)$ is not exponentially small and $c_n = \sum_\theta(\nu_n(\theta))^\alpha$ is not exponentially large. Since Θ is dense in $\mathbf{R}^d$, the condition that the relative entropy is continuous implies that the information denseness condition is satisfied for all $p \in \{q_\theta : \theta \in \mathbf{R}^d\}$. Theorem 15 applies to show that the sequence of maximum posterior likelihood estimators of the density is consistent in the L^1 sense for all such p.

Selecting a family: Suppose a statistician has a countable list of favorite parametric families (each of which is discretized as above), then a maximum posterior likelihood estimator amounts to an automatic selection of a model as well as an estimator of the parameter values within the chosen family. Although it is not known in advance which of the families contains the true density, nevertheless the density is consistently estimated.

The following method may be regarded as an idealization of the procedure by which a family for the density is chosen. It is also an extension of the nonparametric example given above. Note that an essential but often unmentioned requirement of practical estimators that the probabilities be computable.

The most likely simple density (Cover 1972): Let $\{q_k, k=1,2,...\}$ be an enumeration of the density functions on the real line for which the corresponding distribution functions are computable. (A cumulative distribution function Q is said to be computable if the set $\{Q^b(x) : b \in \{1,2,...\}, x \text{ rational}\}$ is recursively enumerable, where $Q^b(x)$ has a b bit binary representation and $|Q(x)-Q^b(x)| \le 2^{-b}$.) Cover's density estimator sets $\hat{p}_n$ to be the density $q_{\hat{k}}$ which maximizes the likelihood for $k \le \tau_n$ where τ_n tends to infinity, but not exponentially fast. This is the maximum posterior likelihood estimator with uniform prior $\nu_n(k)$ on $1 \le k \le \tau_n$. The conditions of Theorem 15 are satisfied for this sequence ν_n. Moreover, it follows from Lemma 17 that the information limits consist of all densities p for which $D(p \,||\, q_k)$ is finite for some k (see Barron 1985b). Therefore, Cover's density estimator is consistent for every such density p. Note that the only densities for which convergence is not obtained are those which are infinitely far away from every computable density.

Complexity minimization: A refinement of Cover's estimator is examined in some detail in Barron (1985b). Estimating the density is related to finding short descriptions for finely discretized data X^n. A natural prior is $\nu(k) = 2^{-L(k)}$ where $L(k)$ is the length of the shortest binary program for Q_k on a fixed universal computer with a domain which satisfies the prefix condition, (which implies that $\sum_k 2^{-L(k)} \le 1$);

because for this prior the maximum posterior likelihood estimator corresponds to finding the shortest length program for X^n among the programs with length $L(k) + [\log 1/Q_k(X^n)]$. (Here $[\log 1/Q_k(X^n)]$ is the length of Shannon's code for X^n based on Q_k, which *must* be prefaced by a description of Q_k, whence the $L(k)$ term.) The set of distributions P for which the family of probability measures is information dense not only determines distributions which can be consistently estimated but also determines distributions for which the length of the shortest program has asymptotically negligible redundancy. Again, the computable measures are information dense for every P for which $D(P \| Q_k) < \infty$ for some Q_k. Unfortunately, this complexity based prior does not satisfy the root summability condition of Theorem 15; nevertheless, Proposition 3 applies to obtain a useful convergence result (convergence in T_n variation). To force convergence in total variation, the prior $\nu(k) = 2^{-2L(k)}$ is preferred when selecting the maximum posterior likelihood estimate $\hat{k}_n$. (Although this maximization no longer corresponds exactly to the minimization of the description length, it may be shown that the resulting description length $L(\hat{k}_n) + [\log 1/Q_{\hat{k}_n}(X^n)]$ is still nearly minimal). Similar convergence results also obtain if $\nu_n(k) = 2^{-L_n(k)}$ where $L_n(k)$ is the length of the shortest program for Q_k which runs in time not exceeding τ_n where $\lim \tau_n = \infty$.

Perhaps the most surprising result for countable priors concerns the case that the true distribution P happens to equal one of the distributions Q_θ for which $\nu(\theta) > 0$. Let $\hat{P}_n = Q_{\hat{\theta}_n}$ be the maximum posterior likelihood estimator based on $X_1,\ldots,X_n$. Assume that with respect to each Q_θ, the process is stationary and ergodic. We require that for each $\theta \in \Theta$ the sequence of priors $\nu_n(\theta)$ increases to $\nu(\theta)$ and that $\sum_\theta \nu(\theta) \leq 1$.

Theorem 16: *If P is a member of the countable family $\{Q_\theta : \nu(\theta) > 0, \theta \in \Theta\}$, then the maximum posterior likelihood estimator satisfies*

$$\hat{P}_n \equiv P \text{ for all large } n \text{ with } P \text{ probability one.}$$

Remark: Thus if a random process is governed by a computable law, then eventually this law will be discovered and thereafter never refuted.

Proof: This result is essentially a specialization of Doob's (1949) result to the countable parameter case. Another simple proof is this: Let P^∞ and Q_θ^∞ be the induced distributions for $X_1,X_2,\ldots$ on $(\mathbf{X}^\infty,\mathbf{B}^\infty)$. Distinct stationary and ergodic distributions must be mutually singular on $\mathbf{X}^\infty$ (by applications of the ergodic theorem to the relative frequences of any event for which the distributions differ). Thus the measures P^∞ and $\sum_{\theta \in C} \nu(\theta) Q_\theta^\infty$ are mutually singular, where C is the set of all θ for which

Q_θ^∞ is not equal to P^∞. It follows that the density ratio $\sum_C \nu(\theta)q_\theta(X^n)/p(X^n)$ must converge to zero, with P probability one. But then $\max_C \nu_n(\theta)q_\theta(X^n) < (1/2)\nu(\theta^*)q_{\theta^*}(X^n)$, for all large n, with probability one, for any fixed θ^* in the set $A = \{\theta : Q_\theta^\infty \equiv P^\infty\}$. Now for all large n, $\nu_n(\theta^*) \geq (1/2)\nu(\theta^*)$ and hence the (joint) likelihood $\nu_n(\theta^*)q_{\theta^*}(X^n)$ is greater than the likelihoods for all θ in $A^c = C$. Whence $\hat{\theta}_n$ is in A for all large n, with probability one. This completes the proof of Theorem 16.

7. Counterexample

In this section we show that the posterior probability of total variation or relative entropy neighborhoods of P do not necessarily converge to one, even if the prior is information dense at P. What is more, we show that posterior probabilities of T_n variation neighborhoods ($A_n = \{\theta : d_{T_n}(P,Q_\theta) < \delta\}$) do not necessarily converge to one, if T_n is any sequence of partitions for which the effective cardinality is of order greater than n. By Theorem 5 this amounts to showing that there are priors for which there do not exist decompositions of the parameter space into sets A_n, B_n, C_n satisfying the indicated properties.

Let P be a probability measure on $(\mathbf{X},\mathbf{B})$, let λ be any sigma-finite measure which dominates P, and let T_n be a sequence of partitions. The proof of Barron (1987b, Theorem 2) shows that if the effective cardinality of T_n with respect to P is not of order n, then there exists a constant $\delta > 0$ and a sequence of constants $r_n > 0$, parameter spaces Θ_n^1, probability measures Q_θ for $\theta \in \Theta_n^1$, and proper priors ν_n^1, such that liminf $r_n = 0$, $d_{T_n}(P,Q_\theta) \geq \delta$, and $Q_\theta << P$ for all θ with

$$\frac{\int_{A_n^c} q^n(\mathbf{x}^n \mid \theta)\nu_n^1(d\theta)}{p^n(\mathbf{x}^n)} \geq e^{-nr_n} \tag{17}$$

for all $\mathbf{x}^n$, where $q^n(\cdot \mid \theta)$ and $p^n(\cdot)$ are the product density functions with respect to the λ^n. (In fact Barron 1987 takes ν_n^1 to be a discrete uniform on large finite sets Θ_n^1, and Q_θ to have density ratio $q(x \mid \theta)/p(x)$ which typically oscillates irregularly between values near 0 and 2.)

The above prior is not yet information dense at P. To give the desired counterexample we modify it by mixing with another prior ν^2 (for another family $q(\cdot \mid \theta)$, $\theta \in \Theta^2$) which is information dense at P and yet satisfies

$$\frac{\int q^n(X^n \mid \theta)\nu^2(d\theta)}{p^n(X^n)} < e^{-nr_n} \text{ infinitely often with P probability one.} \tag{18}$$

Take the overall prior ν_n to be a (1/2,1/2) mix of ν_n^1 and ν^2 on the parameter space Θ_n which is a disjoint union of Θ_n^1 and Θ^2. Then taking the ratio of (17) and (18) shows that $m(X^n, A_n^c)/m(X^n, A_n) > 1$ and hence $\nu_n(A_n \mid X^n) < 1/2$ infinitely often with P^∞ probability one. In which case, the posterior probability of the neighborhoods $A_n = \{\theta : d_{T_n}(P, Q_\theta) < \delta\}$ does not converge to one.

It remains to show that an information dense prior can be chosen for which (18) is satisfied. Suppose we take P to be the standard Normal distribution, $\Theta^2 = (0,1)$ to be the unit interval, and Q_θ to be the Normal$(\sqrt{2\theta}, 1)$ family of distributions for $0 < \theta < 1$. Note that $q(x \mid \theta)/p(x) = \exp\{-\theta + x\sqrt{2\theta}\}$ and $D(P \mid\mid Q_\theta) = \theta$. So if we chose the prior to have a strictly positive density function (with respect to Lebesgue measure) in an interval adjoining $\theta = 0$, then the prior is information dense at P. Now since liminf $r_n = 0$ there is a strictly decreasing function $f(u)$ for $u \geq 0$ such that $f(n) = r_n$ infinitely often, $\lim f(n) = 0$, and f has an inverse function $g(\theta), \theta > 0$. Then infinitely often, $g(\theta) + n\theta \geq nr_n$ for all $0 < \theta < 1$. Set the prior to be $\nu^2(d\theta) = (1/c)\exp\{-g(\theta)\}d\theta$ for $0 < \theta < 1$ where $c = \int_0^1 \exp\{-g(\theta)\}d\theta$. In this case

$$\frac{\int q^n(X^n \mid \theta)\nu^2(d\theta)}{p^n(X^n)} = \int_0^1 \exp\{-g(\theta) - n\theta + S_n\sqrt{2\theta}\}d\theta$$

$$\leq \exp\{-nr_n + (S_n)^+\sqrt{2}\} \text{ infinitely often}$$

$$< \exp\{-nr_n\} \text{ infinitely often, with } P^\infty \text{ probability one}$$

where $S_n = \sum_{i=1}^n X_i$ which is negative infinitely often, P almost surely, by the law of the iterated logarithm. Thus (18) is satisfied. Consequently, for this example the posterior probability of $A_n = \{\theta : d_{T_n}(P, Q_\theta) < \delta\}$ does not converge to one. By using the probability inverse transformation this example could be modified to allow P to be any continuous distribution on the line.

8. Examples

The following four classes of models provide a glimpse of the range of applicability of the results. Some other examples involving discrete priors were given in section 6. Here the priors are typically not discrete.

Example 1: The prior makes a random density function on the line by using the

exponential of a Gaussian process. We start with a family of probability density functions on the unit interval

$$g(t \mid Z) = \frac{e^{Z(t)}}{\int_0^1 e^{Z(s)}ds}, \quad 0 \le t \le 1. \tag{18}$$

Here the parameter space consists of bounded continuous functions $Z(t), 0 \le t \le 1$. The prior ν is chosen such that Z is a mean zero Gaussian process. In particular we assume that it is either a Wiener process (with covariance $E_\nu Z(s)Z(t) = \sigma^2 \min\{s,t\}$) or a stationary Gaussian Markov process (with covariance $\sigma^2 e^{-\beta|s-t|}$) where σ and β are fixed positive constants. A smoother model for the density function is to take Z to be a Gaussian process with kth derivative equal to either a Wiener process or a stationary Gaussian Markov process.

To parameterize density functions on the real line we use the model

$$q(x \mid \theta) = \frac{e^{\theta(x)}}{c(\theta)} q_0(x) \tag{19}$$

where $c(\theta) = \int e^{\theta(x)} q_0(x) \lambda(dx)$ and θ is a bounded continuous function. Here the function $q_0(x)$ is an "initial guess" of the probability density. The prior is chosen to make $\theta(x) = Z(Q_0(x))$ where $Q_0(x)$ is the cumulative distribution function corresponding to the density q_0 and Z is one of the above mentioned Gaussian processes. (Thus θ is a mean zero Gaussian process with covariance $E_\nu \theta(x)\theta(y) = R(Q_0(x), Q_0(y))$ where $R(s,t)$ is the covariance of Z.) The models (18) and (19) are simply related by the transformation $T = Q_0(X)$.

These Gaussian exponent priors are information dense, i.e. condition (A) is satisfied, for every probability density function p for which the relative entropy $D(p \,||\, q_0)$ is finite. Also it is shown that the smoothness condition (B) is satisfied. Therefore the Bayes density estimates corresponding to these priors converge in L^1 to any such p with probability one.

Verification of condition (A): The relative entropy is invariant under monotone transformation of the random variables, so it is enough to check the condition for the first model. Indeed, the prior probability of $\{D(p \,||\, q_\theta) < \varepsilon\}$ is the same as the prior probability of $\{D(f \,||\, g(\cdot \mid Z)) < \varepsilon\}$ where $f(t)$ is the true probability density function for the transformed random variable T. The assumption that $D(p \,||\, q_0) < \infty$ becomes $D(f \,||\, u) < \infty$ where $u(t)$ is the uniform density function on $[0,1]$.

If $D(f \,||\, u)$ is finite, then for every $\varepsilon > 0$ there is a density $\tilde{f}$ such that $\psi(t) = \log \tilde{f}(t)$ is bounded function with a bounded continuous derivative and

$D(f \,||\, \tilde{f}) < \varepsilon/2$ (see Lemma 17, Appendix). Now for any density g the relative entropy $D(f \,||\, g)$ satisfies

$$D(f \,||\, g) = \int f \log \frac{f}{g}$$

$$= \int f \log \frac{f}{\tilde{f}} + \int f \log \frac{\tilde{f}}{g}$$

$$< \varepsilon/2 + \sup_{0 \le t \le 1} \log \frac{\tilde{f}(t)}{g(t)}$$

For $g(t \mid Z) = e^{Z(t)}/(\int e^{Z(s)}ds)$ we have

$$\sup_{0 \le t \le 1} \log \frac{\tilde{f}(t)}{g(t)} = \sup(\psi(t) - Z(t)) + \log \int e^{Z(t)}dt$$

$$= \sup(\psi(t) - Z(t)) + \log \int e^{Z(t)-\psi(t)} \tilde{f}(t)dt$$

$$\le 2 \sup |Z(t) - \psi(t)|$$

Thus $D(f \,||\, g(\cdot \mid Z) < \varepsilon/2 + \sup |Z(t) - \psi(t)|$. Consequently, for condition (A) it is enough that the following probability is positive

$$\nu\{\sup_{0 \le t \le 1} |Z(t) - \psi(t)| \le \varepsilon/4\} \tag{20}$$

Now a Gaussian process Z which has mean zero and covariance function R is equivalent (mutually absolutely continuous) to the Gaussian process with non-zero mean function $\psi(t)$ (and the same covariance) if and only if $||\psi||_R < \infty$, where $||\psi||_R$ is the RKHS norm of ψ corresponding to the covariance R of the Gaussian process (see Parzen 1970). In particular, for the Wiener process $||\psi||_R^2 = \int_0^1 (\psi'(t))^2 dt$, for the stationary Gaussian Markov process $||\psi||_R^2 = (1/2)\int_0^1 ((\psi(t))^2 + (\psi'(t))^2)dt + (1/2)((\psi(0))^2 + (\psi(1))^2)$. These norms are finite by the choice of ψ. Consequently, the probability in (20) is positive if and only if

$$\nu\{\sup_{0 \le t \le 1} |Z(t))| < \varepsilon/4\} > 0. \tag{21}$$

This probability is known to be positive for all $\varepsilon > 0$ when $Z = W$ is the standard Weiner process (see Siegmund 1985, p.56). Now any mean zero Gauss Markov process Z is equal in distribution to a scaled and time shifted Wiener process, i.e. $Z(t) = a(t)W(G(t))$ where G is non-decreasing (in particular the stationary Markov case obtains with $G(t) = e^{2\beta t}$ and $a(t) = e^{-\beta t}$). In which case $\sup |Z(t)| \le \bar{a} \sup\{|W(x)| : G(0) \le x \le G(1)\}$ where $\bar{a} = \sup |a(t)|$). Consequently if $\bar{a}$ and $G(1) - G(0)$ are finite, then the positivity of the probability in (21)

for all $\varepsilon > 0$ follows from the positivity for the Weiner process case.

The case that the kth derivative of Z is either a Weiner process or a stationary Gaussian Markov process is handled in the same way except that ψ should be chosen to have $k + 1$ bounded continuous derivatives.

Verification of condition (B): As before, it is enough to check the condition for the first model ($g(t \mid Z)$). Indeed, let π_n be a partition of the line into n quantiles of the distribution Q_0 and let T_n be a uniform partition of $[0,1]$ into cells of width $1/n$, then the sets $\{d(q_\theta, q_\theta^{\pi_n}) > \varepsilon\}$ and $\{d(g(\cdot \mid Z), g^{T_n}(\cdot \mid Z) > \varepsilon\}$ have the same prior probability. The aim is to show that this prior probability is exponentially small. Take the case that $Z(t)$, $0 \le t \le 1$ is a Wiener process (with respect to the prior). Set $\delta = (1/2)\varepsilon^2$. Since $d \le \sqrt{2D}$, it is enough to show that the following probability is exponentially small

$$\nu\{D(g(\cdot \mid Z) \,||\, g^{T_n}(\cdot \mid Z)) > \delta\}. \tag{22}$$

Now the relative entropy in this expression is an average logarithm of the density ratio which is less than the maximum for $1 \le i \le n$ of the following terms

$$\begin{aligned} \sup_{t \in [(i-1)/n, i/n]} \log \frac{g(t \mid Z)}{g^{T_n}(t \mid Z)} &\le \sup_{s,t \in [(i-1)/n, i/n]} \log \frac{g(t \mid Z)}{g(s \mid Z)} \\ &= \sup_{s,t \in [(i-1)/n, i/n]} (Z(t) - Z(s)) \\ &\le 2 \sup_{t \in [(i-1)/n, i/n]} \left| Z(t) - Z\left(\frac{i-1}{n}\right) \right| \end{aligned}$$

where the first inequality is by the mean value theorem. These terms have identical distributions with respect to the prior. Consequently, by the union of events bound, the probability in (22) is less than or equal to

$$n\, \nu\{ \sup_{0 \le t \le 1/n} |Z(t)| > \delta/2\} \tag{23}$$

The supremum of $|Z(t)|$ in this expression is less than or equal to the maximum of the two random variables $\sup_{[0,1/n]} Z(t)$ and $\sup_{[0,1/n]} (-Z(t))$ each of which is known to have the same distribution as $|Z(1/n)|$. Thus we may bound (23) by

$$2n\, \nu\{ |Z(1/n)| > \delta/2\} \le 4n\, e^{-n\delta^2/8\sigma^2} \tag{24}$$

since $Z(1/n)$ is a Normal random variable with mean zero and variance σ^2/n. Since this probability is exponentially small, condition (B) is verified. Other Gauss-Markov priors for Z are handled in a similar way using the representation as time scaled Wiener processes (The partition T_n may be chosen as the "quantiles" of the increasing

function $G(t)$).

Example 2: The density function is modeled using an infinite dimensional exponential family. Let $\phi_0(x)=1$, $\phi_1(x)$, $\phi_2(x)$,... be linearly independent measurable functions on the real line. We assume that these functions are bounded $|\phi_k(x)| \leq b$ for all k and that linear combinations of the functions ϕ_k are dense in $L^2(Q_0)$, where Q_0 is a fixed distribution function with density $q_0(x)$. It is convenient to assume that $\phi_k(x) = r_k(Q_0(x))$ where the r_k are functions on $[0,1]$. In particular we may use algebraic polynomials $r_k(t) = t^k$ or trigonometric polynomials $r_{2k}(t) = \cos(2\pi kt)$, $r_{2k+1}(t) = \sin(2\pi kt)$. Consider probability density functions which may be represented as

$$q(x \mid \theta) = q_0(x)e^{\sum_{k=0}^{\infty}\theta_k\phi_k(x)} = q_0(x)\frac{e^{\sum_{k=1}^{\infty}\theta_k\phi_k(x)}}{c(\theta)}$$

where $e^{-\theta_0} = c(\theta) = \int e^{\sum_{k=1}^{\infty}\theta_k\phi_k(x)}Q_0(dx)$.

If the prior is chosen such that the θ_k are independent Gaussian random variables with mean zero and variance λ_k for $k \geq 1$, then the exponent in the second expression for the density is a mean zero Gaussian process with covariance function $R(x,y) = \sum_{k=1}^{\infty}\lambda_k\phi_k(x)\phi_k(y)$. (Indeed, the models in example 1 may represented in this way for appropriate choices of ϕ_k.)

We assume that the prior is such that $E|\theta_k| \leq a_k$ for some summable sequence a_k, then (by Fubini's Theorem) $\sum_k |\theta_k|$ has a finite expectation with respect to the prior. Consequently, $\sum_{k=1}^{\infty}\theta_k\phi_k(x)$ is absolutely convergent for every x and bounded (as a function of x) with prior probability one. In which case $c(\theta)$ is finite and the densities $q(x \mid \theta)$ are well defined with prior probability one.

Suppose each $r_k(t)$ is differentiable with derivative bounded by b_k (for algebraic polynomials $b_k = k$ for trigonometric polynomials $b_k = 2\pi k$). If $e^{\alpha\sum_1^{\infty}|\theta_k|b_k}$ has finite expectation with respect to the prior for some $\alpha > 0$, then condition (B) is satisfied. In particular if $\theta_1,\theta_2,...$ are independent with respect to the prior it is enough to check that $E(e^{\alpha b_k|\theta_k|}) \leq e^{a_k}$ for some summable sequence a_k.

If the prior distribution of $\theta_1,\theta_2,...,\theta_m$ has support equal to all of R^m for every $m \geq 1$, then condition (A) is satisfied for every density $p(x)$ for which $D(p \,||\, q_0)$ is finite. In which case Bayes estimates of the density function are consistent.

In the next example the model may also be put into the form of an infinite dimensional exponential family. However, the functions ϕ_k are indicator functions

(which are not differentable) and the parameters have simplex constraints (which do not correspond to all of R^m). Nevertheless, conditions (A) and (B) are directly verifiable.

Example 3: The random density function is defined using a refining sequence (tree) of partitions T_k which generates the measurable space. The family of probability measures is represented as

$$Q(A \mid \theta) = \prod_{k=1}^{s} \theta_{k,A_k}$$

for any A in T_s where A_k denotes to the set in T_k which contains A. Each parameter θ_{k,A_k} is interpreted as the conditional probability that X is in A_k given that X is in A_{k-1}. Thus these parameters are required to satisfy $\theta_{k,A} \geq 0$ and $\sum_{A \in T_{k,B}} \theta_{k,A} = 1$ for each B in T_{k-1}, where $T_{k,B}$ is the collection of sets in T_k which are subsets of B. Let Q_0 be a fixed probability measure with density function $q_0(x)$. The probability measures which are absolutely continuous with respect to Q_0 have probability density functions

$$q(x \mid \theta) = q_0(x) \prod_{k=1}^{\infty} \frac{\theta_{k,A_k}}{\alpha_{k,A_k}} \tag{2}$$

where $\alpha_{k,A_k} = Q_0(A_k \mid A_{k-1})$ denotes the conditional probabilities for Q_0 and $A_k = A_k(x)$ is the set in T_k which contains x. (The limit in this expression for the density exists for almost every x, by application of the Lebesgue density theorem.)

A reasonable choice for the prior is to make $\{\theta_{k,A} : A \in T_{k,B}\}$ have independent Dirichlet distributions with parameters $\{\beta_{k,A} : A \in T_{k,B}\}$ for each B in T_{k-1} and each $k \geq 1$. If $\beta_{k,A} = bQ_0(A)$ for some constant b, then this is the Dirichlet process prior with parameter measure bQ_0 (see Fergeson 1973,1974). However, the resulting random measures $Q(\cdot \mid \theta)$ are discrete with probability one. On the other hand, Kraft (1964) and Métevier (1971) give conditions on the choice of the prior such that the measures are absolutely continuous with probability one. We recommend setting $\beta_{k,A} = b_k \alpha_{k,A}$ with a sequence b_k which tends to infinity as $k \to \infty$ (so that the multiplicands in equation (2) concentrate near one for large k with high probability).

If the sequence b_k tends to infinity sufficiently rapidly and if the partitions T_k are chosen such that the $\alpha_{k,A}$ stay bounded away from zero, then the measures $Q(\cdot \mid \theta)$ are absolutely continuous with probability one and these measures are sufficiently smooth that condition (B) is satisfied. Moreover, condition (A) is satisfied for every

density p for which $D(p \,||\, q_0)$ is finite. Therefore, Bayes estimators are consistent in L^1 for any such density.

An advantage of this prior is that the posterior distribution is readily characterized in terms of the same tree of partitions (see Fabius 1964, Freedman 1974). Consequently, it is possible to readily compute the posterior mean estimate of the probability of any set A in T_k.

Example 4: Some regression problems may also be addressed using the results of this paper. Let response variables Y_i be conditionally independent given inputs X_i for $i=1,2,...,n$. The conditional density function $p(y \mid x)$ is assumed to be Normal with unknown mean $\theta^*(x)$ and known variance σ^2. By specifying a prior distribution for the regression function $\theta(x)$, we obtain Bayes estimators for this conditional density. To assess the convergence of the estimators using the techniques of this paper, it is necessary to assume a distribution P_X for the input variables.

Suppose a prior distribution is chosen which makes $\{\theta(x) : x \in \mathbf{X}\}$ a mean zero Gaussian process with covariance function $R(x,y) = E(\theta(x)\theta(y))$. It is a standard fact that for each x, the posterior distribution for $\theta(x)$ given X^n, Y^n is Gaussian with conditional mean $\hat{\theta}_n(x) = \sum_{i=1}^n w_i(x) Y_i$ and conditional variance $a_n(x)$. (We avoid all the details, but do remark that the vector of weights $w_i(x)$ may be expressed as $\underline{w}(x) = (R + \sigma^2 I)^{-1} \underline{r}(x)$ where R is the $n \times n$ matrix with entries $R(X_i, X_j)$ and $\underline{r}(x)$ is the vector with entries $R(X_i, x)$ for $i,j=1,2,...,n$).

It is conjectured that for reasonable choices of the Gaussian process prior, condition (A) will be satisfied for all conditionally Normal distributions for which the regression function $\theta^*(x)$ is in $L^2(P_X)$.

For each x, the Bayes estimate of the conditional density function $p(y \mid x)$ (with relative entropy loss) based on X^n, Y^n is seen to be Normal with mean $\hat{\theta}_n(x)$ and variance $\sigma_n^2(x) = \sigma^2 + a_n(x)$. Consequently the relative entropy loss of this estimator is

$$D(p(\cdot \mid x) \,||\, \hat{p}_n(\cdot \mid x)) = \frac{1}{2} \frac{(\theta^*(x) - \hat{\theta}_n(x))^2}{\sigma_n^2(x)} + \frac{1}{2}\left(\frac{\sigma^2}{\sigma_n^2(x)} - 1\right) - \frac{1}{2} \log \frac{\sigma^2}{\sigma_n^2(x)}$$

Note that in essence this is the sum of a squared error loss function for the regression function and a separate loss function $L(\sigma^2, \hat{\sigma}^2) = 1/2(\sigma^2/\hat{\sigma}^2 - 1 - \log \sigma^2/\hat{\sigma}^2)$ for the variance. Integrating this loss with respect to the distribution of X, we obtain the relative entropy between the distributions $P_{X,Y}$ and $\hat{P}_{X,Y}$. (Here $\hat{P}_{X,Y}$ has conditional density $\hat{p}_n(y \mid x)$ and marginal distribution P_X). Taking the expected value with

respect to the joint distribution of Y^n and X^n yields

$$E(D(p \,||\, \hat{p}_n)) = \frac{1}{2} E \frac{(\theta^*(X) - \hat{\theta}_n(X))^2}{\sigma_n^2} + \frac{1}{2} E\, L(\sigma^2, \sigma_n^2(X))$$

Proposition 4 shows that whenever condition (A) is satisfied the risk $E(D(p \,||\, \hat{p}_n))$ converges to zero in the Cesàro sense as $n \to \infty$. In which case, if $\sigma_n^2(x)$ is bounded, the mean squared error $E\ (\theta^*(X) - \hat{\theta}_n(X))^2$ converges to zero in the Cesàro sense.

Appendix

In this appendix we give an approximation Lemma for the relative entropy. It is used to show that for reasonable priors condition (A) holds for all distributions P which have finite relative entropy (with respect to a fixed reference measure). Let $(\mathbf{X}, \mathbf{B}, Q)$ be a probability space and let $L(Q,b)$ be the set of all measurable functions $f : \mathbf{X} \to \mathbf{R}$ for which the Q-essential supremum of $|f(x)|$ is less than or equal to b. (Here we will use the topology of convergence in $L^1(Q)$; of course for such uniformly bounded functions, $L^1(Q)$ convergence is equivalent to $L^p(Q)$ for all $p > 0$). Similarly, let C be a class of measurable functions and let $C(Q,b)$ be the set of functions in C with essential supremum of the absolute value less than or equal to b. In many cases $C(Q,b)$ is dense in $L(Q,b)$. When $\mathbf{X}$ is the unit interval with the Borel set and Q is Lebesgue measure (the uniform distribution), familiar examples include the class of bounded continuous functions, the class of functions with bounded kth derivative, the class of linear combinations of trigonometric functions, the class of polynomial functions, etc. From any such class C with the uniform distribution, a class with arbitrary distribution function Q on the real line can be obtained by composition $\tilde{C} = \{f(Q(x)) : f \in C\}$.

Fix Q and C and suppose that for some $r \geq 1$, $C(Q, rb)$ is dense in $L(Q,b)$ for all large b.

Lemma 17: *If $D(P \,||\, Q)$ is finite, then for any $\varepsilon > 0$ there is a bounded function ψ in C such that*

$$D(P \,||\, P_\psi) < \varepsilon$$

where P_ψ is a probability measure equivalent to Q with

$$\frac{dP_\psi}{dQ}(x) = \frac{e^{\psi(x)}}{\int e^{\psi} dQ}$$

Thus the class of such tilted measures P_ψ for ψ in C is dense in the set of all probability measures for which $D(P \| Q)$ is finite.

Proof: The relative entropy is $D(P \| Q) = E \log dP/dQ$ where E denotes expectation with respect to P. Define the bounded function ρ to equal $\log dP/dQ$ on the set where $|\log dP/dQ| < b$ and to equal b times the sign of $\log dP/dQ$ elsewhere. Given an arbitrarily small ε in $(0,1)$, choose b large enough that

$$E\,|\log \frac{dP}{dQ}|\,1_{\{|\log dP/dQ| \geq b\}} < \varepsilon.$$

This choice of b ensures that

$$E\,|\log \frac{dP}{dQ} - \rho| < \varepsilon$$

and

$$P\{|\log \frac{dP}{dQ}| > b\} < \frac{\varepsilon}{b}.$$

Let ψ in $C(Q,rb)$ be such that $\int |\rho-\psi|\,dQ < \varepsilon^2 e^{-rb}$. (For the moment we just need that this is less than εe^{-b}). It follows that $\int |\rho-\psi|\,dP$ is also small. Indeed,

$$\begin{aligned}
\int |\rho-\psi|\,dP &= \int |\rho-\psi| \frac{dP}{dQ}\,dQ \\
&\leq a \int |\rho-\psi|\,dQ + (b+rb)\,P\{\frac{dP}{dQ} > a\} \\
&< \varepsilon + (r+1)b\,P\{\log \frac{dP}{dQ} > b\} \\
&< \varepsilon + (r+1)\varepsilon
\end{aligned}$$

where $a = e^b$. Now let $dP_\psi = (e^\psi dQ)/c$ where $c = \int e^\psi dQ$. The relative entropy of P with respect to P_ψ is

$$\begin{aligned}
D(P \| P_\psi) &= E \log \frac{dP}{(e^\psi dQ)/c} \\
&= E(\log \frac{dP}{dQ} - \psi) + \log c \\
&\leq E\,|\log \frac{dP}{dQ} - \rho| + E\,|\rho-\psi| + \log c \\
&< (r+3)\varepsilon + \log c.
\end{aligned}$$

It remains to bound $\log c$. We show that $c = \int e^\psi dQ$ is near one by using the fact

that e^{Ψ}, e^{ρ}, and dP/dQ are close to each other with high probability:

$$\int e^{\Psi} dQ \leq \int e^{\rho+\varepsilon} dQ + e^{rb} Q\{\psi-\rho > \varepsilon\}$$

$$\leq (1 + e^{-b})e^{\varepsilon} + \frac{e^{rb}}{\varepsilon}\int |\psi-\rho| dQ$$

$$< (1 + e^{-b})e^{\varepsilon} + \varepsilon$$

$$< 1 + e^{-(b-1)} + e\varepsilon.$$

Thus $\log c$ is less than $e^{-(b-1)} + e\varepsilon$ which is less than 3ε if we require that b be chosen sufficiently large. Consequently,

$$D(P \| P_{\psi}) < (r+6)\varepsilon.$$

Thus there exist essentially bounded functions ψ in C for which $D(P \| P_{\psi})$ is arbitrarily small. This completes the proof of Lemma 17.

References

Abou-Jaoude S. (1976). Conditions nécessaires et suffisantes de convergence L^1 en probabilité de l'histogramme pour une densité. *Ann. Inst. Henri Poincaré*, B **12** 213-231.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian non-parametric problems. *Ann. Statist.* **2** 1152-1174.

Barron, A. R. (1985a). The strong ergodic theorem for densities: generalized Shannon-McMillan-Breiman theorem. *Ann Probab.* **13** 1292-1303.

Barron, A. R. (1985b). *Logically smooth density estimation.* PhD dissertation, Stanford University.

Barron, A. R. (1986). Discussion on the consistency of Bayes estimates. *Ann. Statist.* **14** 26-30.

Barron, A. R. (1987a). Are Bayes rules consistent in information? In *Open Problems in Communications and Computation* (T.M. Cover and B. Gopinath, editors) 85-91. Springer-Verlag, New York.

Barron, A. R. (1987b). Uniformly powerful goodness of fit tests. To appear in the *Annals of Statistics.*

Barron, A. R. and Cover, T. M. (1988) A bound on the financial value of information. To appear in *IEEE Trans. Inform. Theory.*

Berk, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Ann. Math. Statist.* **37** 51-58.

Berk, R. H. (1970). Consistency a posteriori. *Ann. Math. Statist.* **41** 894-906.

Bickel, P. J. and Yahav, Y. A. (1969). Some contributions to the asymptotic theory of Bayes solutions. *Z. Wahrsch. verw. Geb.* **11** 257-276.

Breiman, L., LeCam, L., and Schwartz, L. (1964). Consistent estimates and zero-one sets. *Ann. Math. Statist.* **35** 157-161.

Cover, T. M. (1972). A hierarchy of probability density function estimates, in

Frontiers in Pattern Recognition Academic Press, New York, 83-98.
Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.* **2** 299-318.
Davisson, L. D. (1973). Universal noiseless coding, *IEEE Trans. Inform. Theory* **19** 783-795.
Diaconis, P. and Freedman, D. (1986a). On the consistency of Bayes estimates. *Ann. Statist.* **14** 1-26.
Diaconis, P. and Freedman, D. (1986b). On inconsistent Bayes estimates of location. *Ann. Statist.* **14** 68-87.
Doksum, K. (1974). Tail-free and neutral random probabilities and their posterior distributions. *Ann. Probab.* **2** 183-201.
Doob, J.L. (1949). Application of the theory of martingales. In *Le Calcul de Probabilities et ses Applications.* Colloques Internationaux du Centre National de la Researche Scientifique, Paris 23-27.
Fabius, J. (1964). Asymptotic behavior of Bayes estimates. *Ann. Math. Statist.* **35** 846-856.
Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209-230.
Ferguson, T. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2** 615-629.
Freedman, D. (1963). On the asymptotic behavior of Bayes estimates in the discrete case. *Ann. Math. Statist.* **34** 1386-1403.
Freedman, D. and Diaconis, P. (1983). On inconsistent Bayes estimates in the discrete case. *Ann. Math. Statist.* **11** 1109-1118.
Glick N. (1972). Sample-based classification procedures derived from density estimators, *J. American Statist. Assoc.* **67** 116-122.
Hoeffding, W. and Wolfowitz, J. (1958). Distinguishability of sets of distributions. *Ann. Math. Statist.* **29** 700-718.
Ibragimov, I. A. and Has'minskii, R. Z. (1973). On moments of generalized Bayesian estimators and maximum likelihood estimators. *Theory Probab. Appl.* **18** 508-520.
Ibragimov, I. A. and Has'minskii, R. Z. (1981). *Statistical Estimation: Asymptotic Theory* Springer-Verlag, New York.
Johnson, R. A. (1967). An asymptotic expansion for posterior distributions. *Ann. Math. Statist.* **38** 1899-1906.
Johnson, R. A. (1970). Asymptotic expansions associated with posterior distributions. *Ann. Math. Statist.* **41** 851-864.
Kraft, C. H. (1955). Some conditions for consistency and uniform consistency of statistical procedures. *Univ. California Publ. Statist.* **2** 125-142.
Kraft, C. H. (1964). A class of distribution function processes which have derivatives. *J. Appl. Probab.* **1** 385-388.
Kullback, S. (1967). A lower bound for discrimination information in terms of variation. *IEEE Trans. Inform. Theory* **IT-13** 126-127.
LeCam, L. (1953). On some asymptotic properties of the maximum likelihood estimates and related Bayes estimates. *Univ. California Publ. Statist.* **1** 277-330.
LeCam, L. (1958). Les propriétés asymptotiques des solutions de Bayes. *Publ. Inst. Statist. Univ. Paris* **7** 17-35.
LeCam, L. (1970). On the weak convergence of probability measures. *Ann. Math. Statist.* **41** 621-625.
LeCam, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 38-53.

LeCam, L. (1982). On the risk of Bayes estimates. In *Statistical Decision Theory and Related Topics III* **2** (S. Gupta and J. Berger, editors) 121-138. Academic Press.

LeCam, L. (1986). Discussion on the consistency of Bayes estimates. *Ann. Statist.* **14** 59-60.

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

LeCam, L. and Schwartz, L. (1960) A necessary and sufficient condition for the existence of consistent estimates. *Ann. Math. Statist.* **31** 140-150.

Métevier, M. (1971). Sur la construction de mesures aléatoires presque sûrement absolument continues par rapport à une mesure donnée. *Z. Wahrsch. verw. Geb.* **20** 332-344.

Parzen, E. (1970). Statistical inference on time series by RKHS methods. In *Proc. 12th Bienniel Seminar Canadian Mathematical Congress* (R. Pyke editor) 1-27. Canadian Mathematical Congress, Montreal.

Schwartz, L. (1960). Consistency of Bayes' procedures. Doctoral dissertation. Univ. California, Berkeley.

Schwartz, L. (1965). On Bayes' Procedures. *Z. Wahrsch. verw. Geb.* **4** 10-26.

Siegmund, D. (1985). *Sequentail Analysis* Springer-Verlag, New York.

Strasser, H. (1981a). Consistency of maximum likelihood and Bayes estimates. *Ann. Statist.* **9** 1107-1113.

Strasser, H. (1981b). Convergence of estimates. *J. Multivariate Analysis* **11** 127-151.

Vapnik, V.N. and Chervonenkis, A.Y. (1971). On the uniform convergence of relative frequences of events to their probabilities. *Theory Probab. Appl.* **16** 264-280.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimates. *Ann. Math. Statist.* **20** 595-601.

Departments of Statistics and
Electrical and Computer Engineering
University of Illinois
725 S. Wright Street
Champaign, IL 61820