BAYES AND MINIMAX ASYMPTOTICS OF ENTROPY RISK

by

Bertrand S. Clarke and Andrew R. Barron
Purdue University and University of Illinois

Technical Report #90-11

Department of Statistics
Purdue University

March 1990

◇

Bayes and Minimax Asymptotics of Entropy Risk

(Abbreviated Title: Bayes and Minimax Entropy Risk)

Bertrand S. Clarke [1] *and Andrew R. Barron* [2]

Purdue University and University of Illinois

Abstract

Decision theoretic asymptotics are examined for the Kullback-Leibler distance (relative entropy) between probability densities in smooth finite dimensional parametric families. Jeffreys' prior, which is based on the square root of the Fisher information, is characterized as the unique continuous prior that achieves the asymptotic minimax risk. Uniformly accurate expressions are obtained for the risk of Bayes strategies. Asymptotics are also derived for the Bayes risk which is recognized as a special case of Shannon's mutual information. These asymptotics for the Bayes risk are equivalent to a strengthened Bayesian central limit theorem.

For submission to the *Annals of Statistics*, February 1990.

[1] Supported, in part, by the Joint Services Electronics Program, contract N 00014-84-C-0149 while a student at the University of Illinois.

[2] Supported, in part, by the Office of Naval Research, contract N 00014-86-K-0670 and N 00014-89-J-1811.

AMS 1980 subject classifications. Primary 62C10, 62C20; secondary 62F12, 62F15.

Key words and phrases. Bayes risk, minimax risk, Kullback-Leibler information, Jeffreys' prior, Fisher information, Shannon's mutual information, parametric density estimation, data compression.

1. Introduction. We examine large sample decision theoretic properties associated with the relative entropy or Kullback-Leibler distance between probability density functions for independent and identically distributed random variables in smooth finite dimensional parametric families. We derive an asymptotic expression for the Bayes risk of the Bayes estimator and for the minimax risk. Among smooth priors, Jeffreys' prior uniquely achieves the asymptotic minimax value. Also, the convergence of the Bayes risk is shown to be equivalent to a strengthened Bayesian central limit theorem. Indeed, it is shown that the standardized posterior density converges to the normal density in the relative entropy sense.

Given a parametric family of probability density functions $\{p_\theta: \theta \in \Omega\}$, $\Omega \sum \mathbf{R}^d$, with respect to a fixed dominating measure $\lambda(dx)$ on a separable metric space X, with probability measures assumed to be defined on the Borel subsets of X, we denote the density of n independent outcomes $x^n = (x_1,...,x_n)$ by $p_\theta^n(x^n) = \prod_{i=1}^{n} p_\theta(x_i)$. A quantity of interest to us is the relative entropy $D(p_\theta^n \,||\, q_n)$, between the density functions $p_\theta^n(x^n)$ and an arbitrary joint density function $q_n(x^n)$, with respect to the same dominating measure $\lambda^n(x^n)$.

A game-theoretic interpretation is that one player, Nature, picks $\theta \in \Omega$ and assigns the joint density p_θ^n for each n, while a second player, the statistician, chooses q_n for each n. We let the relative entropy $D(p_\theta^n \,||\, q_n)$ be the risk to the statistician, or, in game-theoretic terminology, the 'payoff' to nature. For prior probability density functions $w(\theta)$, $\theta \in \Omega$ with respect to Lebesgue measure on $\mathbf{R}^d$, the Bayes strategy, which is to minimize $\int_\Omega w(\theta) D(p_\theta^n \,||\, q_n)\, d\theta$, is achieved by choosing $q_n(x^n) = m_n^w(x^n)$, where $m_n^w(x^n) = \int_\Omega p_\theta^n(x^n) w(\theta)\, d\theta$. Note that $m_n^w(X^n)$ is the marginal density for $X^n = (X_1,...,X_n)$ associated with the joint density $w(\theta) p_\theta^n(x^n)$ in which $p_\theta^n(x^n)$ is the conditional density function for X^n given θ and w is the prior.

Here we are interested in the asymptotics associated with the Bayes strategy. The quantities that we examine in this paper include the risk of the Bayes strategy,

$$R_n(\theta, w) = D(p_\theta^n \,||\, m_n^w), \tag{1.1}$$

its corresponding Bayes risk,

$$R_n(w) = \int_\Omega R_n(\theta, w) w(\theta) d\theta, \tag{1.2}$$

and the minimax value,

$$R_n = \inf_{q_n} \sup_{\theta \in \Omega} D(p_\theta^n \,||\, q_n). \tag{1.3}$$

Here $D(p \,||\, q)$ denotes the relative entropy or Kullback-Leibler distance which for densities p and q is defined to be

$$D(p \mid\mid q) = E_p \log \frac{p(X)}{q(X)}.$$

A statistical interpretation of $D(p_\theta^n \mid\mid m_n^w)$ is as the cumulative risk of a sequence of Bayes estimators. Indeed, let $\hat{p}_k(x)$ be the predictive density given by

$$\hat{p}_k(x) = m_n^w(X_k = x \mid X^{k-1}) = m_k^w(X^k)/m_{k-1}^w(X^k),$$

for $k=2,...,n$. For $n=1$, $\hat{p}_1(x) = m_1^w(x)$. Then, as in Clarke and Barron (1990a) or Aitchison (1975), it is seen that $\hat{p}_k$ is the Bayes estimator of the density of X_k based on X^{k-1}, under relative entropy loss. The cumulative risk of this sequence of estimators for $k=1,2,...,n$ is $D(p_\theta^n \mid\mid m_n^w)$. Indeed, by the chain rule for relative entropy ,

$$\begin{aligned}
\sum_{k=1}^{n} E_{p_\theta} D(p_\theta \mid\mid \hat{p}_k) &= \sum_{k=1}^{n} E_{p_\theta} \log \frac{p_\theta(X_k)}{m_n^w(X_k \mid X^{k-1})} \\
&= E_{p_\theta} \log \frac{p_\theta(X_1) \cdot \cdots \cdot p_\theta(X_n)}{m^w(X_1) \cdot \cdots \cdot m^w(X_n \mid X^{n-1})} \\
&= E_{p_\theta} \log \frac{p_\theta(X^n)}{m_n^w(X^n)} \\
&= D(p_\theta^n \mid\mid m_n^w).
\end{aligned}$$

Consequently, in a sequential estimation context, $R_n(\theta, w)$, $R_n(w)$, and R_n may be interpreted as the cumulative risk of the Bayes estimators, the cumulative Bayes risk, and the cumulative minimax risk, respectively.

In this paper we determine, under suitable conditions, asymptotic expressions for the quantities (1.1), (1.2), (1.3), and we find the asymptotically least favorable prior corresponding to the minimax risk.

The asymptotic risk of the Bayes estimator $R_n(\theta, w) = D(p_\theta^n \mid\mid m_n^w)$ is shown to satisfy

$$R_n(\theta, w) = \frac{d}{2} \log \frac{n}{2\pi e} + \log \det I(\theta) + \log \frac{1}{w(\theta)} + o(1), \tag{1.4}$$

in which the error, $o(1)$, tends to zero as $n \to \infty$, uniformly on compact sets in the support of the prior, where $I(\theta)$ is the Fisher information matrix.

The Bayes risk, $R_n(w) = \int R_n(\theta, w) w(\theta) d\theta$, is obtained by averaging the risk with respect to the prior w. It is seen that this Bayes risk is also Shannon's mutual information between the parameter θ and the sample $X_1,...,X_n$. That is,

$$R_n(w) = I(\Theta; X^n),$$

where, by definition, Shannon's mutual information $I(\Theta; X^n)$ is the relative entropy distance between the joint density $w(\theta)p_\theta(X^n)$ and the product of marginals $w(\theta)m_n^w(X^n)$. The asymptotic expression we obtain for the Bayes risk is

$$R_n(w) = \frac{d}{2}\log \frac{n}{2\pi e} + \frac{1}{2}\int_\Omega w(\theta) \log \det I(\theta)\, d\theta + H(w) + o(1), \tag{1.5}$$

where $H(w) = \int w(\theta)\log\,(\,1/w(\theta))\, d\theta$ is the entropy of the prior density w, and $o(1) \to 0$ as $n \to \infty$.

Moreover, we show that the asymptotically minimax risk is achieved by Jeffreys' prior,

$$w^*(\theta) = \frac{\sqrt{\det I(\theta)}}{c},$$

where $c = \int_\Omega \sqrt{\det I(\theta)}\, d\theta$ is the normalizing constant. The corresponding asymptotic minimax risk is

$$\inf_{q_n} \sup_{\theta \in \Omega} D(p_\theta^n \,||\, q_n) = \frac{d}{2}\log \frac{n}{2\pi e} + \log \int_\Omega \sqrt{\det I(\theta)}\, d\theta + o(1). \tag{1.6}$$

We also examine the asymptotic behavior of the posterior density for θ given X^n. We find that it is asymptotically normal in expected Kullback-Leibler distance, to wit,

$$E_m\, D(w(\cdot \mid X^n) \,||\, \phi_n) \to 0, \tag{1.7}$$

where ϕ_n is a normal density with mean $E(\Theta \mid X^n)$ and variance $\mathrm{cov}(\Theta \mid X^n)$, and $w(\theta \mid X^n) = w(\theta)p_\theta(X^n)/m_n^w(X^n)$ is the posterior density for θ given X^n. Note that the target normal has mean an variance dependent on the random variables X^n, so that it tracks the posterior but this does not mean that the posterior converges to a fixed normal. In Section 5, we demonstrate a surprising connection between asymptotic of the posterior, convergence of posterior covariances to the inverse of the Fisher information, and the asymptotics of $I(\Theta; X^n)$. Indeed, under conditions given in Theorem 4, we see that the bounds on the Bayes risk imply the convergence of the posterior to the normal.

It is our goal that the quantities (1.1), (1.2) and (1.3), and their associated asymptotic expansions (1.4), (1.5) and (1.6) be of interest to statisticians who concern themselves with minimax estimation, Bayesian estimation, choice of a non-informative prior, and Bayesian central limit theory; as well as to information-theorists who concern themselves with universal data compression and channel capacity. The implications for the latter two topics will be discussed in detail in Clarke and Barron (1990b). Next we discuss how our work relates to some statistical literature.

Schwarz (1978), Leonard (1982) and Haughton (1988) developed expansions similar to (1.4) for model selection problems using a Bayesian criterion. Indeed, if we have a list of

parametric families as hypotheses for the density of X^n, $H_k = \{p_\theta^k : \theta \in \mathbf{R}^{d_k}\}$, for $k = 1,2, \cdots$ and if prior densities $w_k(\theta_k)$ are assigned to each family, then a Bayesian criterion (minimal average probability of error) reduces the problem to a list of simple hypotheses $H_k : X^n \sim M_n^{w_k}$. Approximations to the expected value of $\log p_\theta(X^n)/m_n^{w_k}(X^n)$ which are of order $(d_k/2)\log n + O(1)$ reveal the role of the penalty term in the Schwarz criterion for model selection. The results of Schwarz and Haughton were restricted to families of exponential form.

Ibragimov and Hasminsky (1973) interpreted $I(\Theta; X^n)$ as the information in a sample about a parameter. They established the same asymptotic formula for it under somewhat different hypotheses, stated only in the one-dimensional parameter case. One of their conditions A.IV (expression 4.1) requires that pairs of densities p_θ and $p_{\theta+s}$ asymptotically concentrate on disjoint sets for large s in the sense that the affinity $\int \sqrt{p_\theta(x)p_{\theta+s}(x)}\lambda(dx)$ tends to zero as $s \to \infty$, uniformly in θ. This rules out many common families such as the *Normal*$(0, \theta)$ and the *Poisson*(θ). Also, Ibragimov and Hasminsky require (in Condition A.III) that the Fisher information be bounded and bounded away from zero. The approach developed for Theorems 1 and 2 below avoids these restrictions and permits uniformly accurate expansions for the risk $D(p_\theta^n \,||\, m_n^w)$ as well as the Bayes risk $I(\Theta; X^n) = \int w(\theta) D(p_\theta^n \,||\, m_n^w)\, d\theta$.

The interpretation of Jeffreys' prior as the choice maximizing an asymptotic expression for $I(\Theta; X^n)$ was given by Bernardo (1979). Our analysis gives rigorous justification for Jeffreys' prior as the unique continuous prior for which the Bayes strategy achieves the asymptotically minimax relative entropy risk (Theorem 1). Bernardo's framework for identifying reference priors is extended to multidimensional problems with nuisance parameters in Berger and Bernardo (1989a, 1989b, and 1989c). In the absence of nuisance parameters the reference prior criterion results in the Jeffreys' prior.

Jeffreys' (1961, Sec. 3.10) observed that $(\det I(\theta))^{1/2}$ is the Jacobian of the transformation of the parameter space that makes Hellinger and relative entropy distances locally Euclidean and proposed $w^*(\theta) = (\det I(\theta))^{1/2}/c$ as a choice of prior which remains invariant under reparametrization.

In the information theory context of universal data compression the quantities $R_n(\theta,w)$, $R_n(w)$ and R_n can be interpreted as the redundancy, average redundancy, and minimax redundancy of universal codes, see Davisson (1973). Krichevsky and Trofimov (1981) studied minimax redundancy in the multinomial case, obtaining $R_n = (d/2)\log n + O(1)$ as its asymptotic expression. Rissanen (1986, 1987) showed that the redundancy $R_n(\theta, w)$ equals $(d/2)\log n + o(\log n)$ for smooth parametric families. The more exact asymptotics for $R_n(\theta, w)$ derived in Clarke and Barron (1990a) in an information theory setting are here

extended to give the asymptotics for the average and minimax redundancies $R_n(w)$ and R_n.

The characterization of $R_n(w)$ as a special case of Shannon's mutual information $I(\Theta; X^n)$ leads to implications for channel coding with one sender and many receivers. The applications of our work in this context will be examined in Clarke and Barron (1990b).

There are three main hypotheses which we use to prove (1.4). One is the finiteness of expectations involving the first and second derivatives of the log-likelihood. Another is that the relative entropy $D(p \mid\mid p_\theta)$ be twice continuously differentiable with positive definite second derivative. The third is that the posterior distribution concentrate on neighborhoods of the true value of the parameter at rate $o(1/\log n)$, regardless of which element of Ω is true. An alternative hypothesis involves a restriction on the type of parametrizations which are allowed. We call this concept the soundness of the parametric family and formally define it in Section 2.

The hypotheses used to prove (1.5) are somewhat different. For the upper bound we impose what amounts to a tail condition on the rate of decrease of the prior so that an information theoretic identity can be applied. For the lower bound we use a maximum entropy argument and assume that

$$n \ \mathrm{cov}(\theta \mid X^n) \to I(\theta_o)^{-1} \tag{1.8}$$

in P_{θ_o} probability, for each θ_o in Ω. In a follow up result we give conditions which will ensure (1.8) and are readily verifiable for many examples.

The outline for the remainder of this paper is as follows. Section 2 states the conditions and four main results which are subsequently proved in sections 3 through 5. Some implications for parametric density estimation and the merging of Bayesian beliefs is examined in Section 6.

2. Formal Statements of Conditions and Main Results. So as to facilitate the statements of upper and lower bounds, which typically have slightly different hypotheses, we give a list of conditions to which it will be convenient to refer.

Expectations, E, are taken with respect to p_θ unless denoted otherwise. In particular, E_m denotes expectation with respect to the mixture distribution with density $m_n = m_n^w$, and E_{Θ, X^n} is the expectation with respect to the joint distribution. We write $p(X \mid \theta)$ for $p_\theta(X)$ when convenient. Also, we assume that the parameter space Ω has nonvoid interior and that its boundary has d dimensional Lebesgue measure zero.

Condition 1: *The density* $p(x \mid \theta)$ *is twice continuously differentiable at* θ *for almost every* x, *and there is a* $\delta = \delta(\theta)$ *so that for each* j,k *from 1 to* d

$$E \sup_{\{||\theta' - \theta|| < \delta\}} \left| \frac{\partial^2}{\partial\theta'_j \partial\theta'_k} \log p(X \mid \theta') \right|^{2+\xi}$$

is finite and continuous on a neighborhood of θ *for some positive* ξ, *and for each* j *from 1 to* d

$$E \left| \frac{\partial}{\partial\theta_j} \log p(X \mid \theta) \right|^{2+\xi} < \infty$$

is finite and continuous on a neighborhood of θ.

There are two information matrices, which typically coincide, which we use here. One is the Fisher information which we take to be defined by

$$I(\theta) = E\left[\frac{\partial}{\partial\theta_j} \log p(X \mid \theta) \frac{\partial}{\partial\theta_k} \log p(X \mid \theta)\right]_{j,k=1\ldots d},$$

and the other is the second derivative of the relative entropy

$$J(\theta) = \left[\frac{\partial^2}{\partial\theta'_j \partial\theta'_k} D(p_\theta \mid\mid p_{\theta'}) \mid_{\theta' = \theta}\right]_{j,k=1\ldots d}.$$

When Condition 1 is satisfied the relative entropy is twice continuously differentiable and $J(\theta)$ is seen to equal the matrix with entries $-E_\theta\, \partial^2/\partial\theta_i \partial\theta_j \log p(X \mid \theta)$. The entries of $I(\theta)$ will be denoted by $i_{j,k}(\theta)$, and the entries of the empirical estimate $I^*(\theta)$ will be denoted

$$i^*_{j,k}(\theta) = \frac{1}{n} \sum_{k=1}^{n} \frac{\partial^2}{\partial\theta_j \partial\theta_k} \log p(x_k \mid \theta).$$

Condition 2: $I(\theta)$ *is positive definite and coincides with* $J(\theta)$.

Under Condition 1, condition 2 will be satisfied provided that $\int \partial^2/\partial\theta_i \partial\theta_j\, p(X \mid \theta)\, \lambda(dx) = 0$. See, e.g. Lehmann (1983), Lemma 2.6.1.

We next give a condition on the parametrization of the parametric family.

Condition 3: *The parametric family is sound at* θ, *in the sense that the convergence of parameter values is equivalent to the weak convergence of the distributions they index in the collection of all probabilities on* X. *That is,*

$$\theta' \to \theta \text{ if and only if } P_{\theta'} \xrightarrow{d} P_\theta.$$

where the parameter values converge in the Euclidean metric.

We say that the whole parametric family is sound if and only if it is sound for each value of the parameter.

The weakest condition that we can use in our argument is phrased in terms of posterior convergence. It is that the posterior probability of any open set N containing the true value θ is bounded away from unity with probability $o(1/\log n)$. More formally, we mean that for any open set N containing θ, given $\delta > 0$ we have that $P_\theta^n\{W(N^c \mid X^n) > \delta\} = o(1/\log n)$ where $W(\cdot \mid X^n)$ is the posterior distribution of θ given X^n. Condition 3 is stronger than posterior convergence at rate $o(1/\log n)$, indeed it implies posterior convergence at rate $O(1/n)$ as in Clarke and Barron (1990a).

We believe that soundness is an acceptable hypothesis; it appears to be fundamental in the sense that without some such property there is no connection between the parameter being estimated and the distribution of the random variable being observed. A discussion of soundness is in Clarke and Barron (1990a).

For future use we state a result from Clarke and Barron (1990a) which will be useful in some of the proofs of our results here.

Theorem 0: *Assume that conditions 1, 2, and 3 are satisfied for each θ in the interior of the support of the prior w. Then, for each such θ*

$$\log \frac{p_\theta(X^n)}{m(X^n)} + \frac{1}{2} S_n^T J(\theta)^{-1} S_n - \frac{d}{2} \log \frac{n}{2\pi e} \to \log \frac{1}{w(\theta)} + \frac{1}{2} \log \det J(\theta), \qquad (2.1)$$

in P_θ^n probability and in $L_1(P_\theta)$. Consequently,

$$R_n(\theta, w) = \frac{d}{2} \log \frac{n}{2\pi e} + \frac{1}{2} \log \det I(\theta) + \log \frac{1}{w(\theta)} + o(1).$$

Remark: These convergences do not require the full strength of condition 1. In fact, the convergence in probability holds if the $2 + \xi$ is replaced by 1, and the convergence in $L_1(P_\theta)$ holds if the $2 + \xi$ is replaced by 2.

The statement of $L_1(P_\theta)$ convergence implies that (1.4) is true pointwise in θ. The first theorem extends that result by giving the minimax asymptotics in the compact case.

Theorem 1: *Let w be any positive and continuous prior supported on a compact set K in the interior of Ω. Assume conditions 1, 2 and 3 are satisfied for each θ in K. Then the risk of the Bayes strategy satisfies the asymptotic upper bound*

$$\limsup_{n \to \infty} \sup_{\theta \in K} R_n(\theta, w) - \frac{d}{2} \log \frac{n}{2\pi e} - \log \frac{\sqrt{\det I(\theta)}}{w(\theta)} \le 0. \qquad (2.2)$$

Averaging with respect to w, the Bayes risk satisfies the lower bound

$$\liminf_{n \to \infty} \int_K [\, R_n(\theta, w) - \frac{d}{2} \log \frac{n}{2\pi e} - \log \frac{\sqrt{\det I(\theta)}}{w(\theta)}] d\theta \ge 0. \qquad (2.3)$$

The asymptotic minimax risk satisfies

$$\lim_{n \to \infty} [\, R_n - \frac{d}{2}\log\frac{n}{2\pi e} \,] = \log \int_K \sqrt{\det I(\theta)}\, d\theta \,. \tag{2.4}$$

Jeffreys prior, $w^*(\theta) = \sqrt{\det I(\theta)}/c$ *with* $c = \int_K \sqrt{\det I(\theta)}\, d\theta$, *is the unique continuous and positive prior on K for which the Bayes strategy achieves the asymptotic minimax value, that is*

$$\lim_{n \to \infty} [\, \sup_{\theta \in K} R_n(\theta, w^*) - \frac{d}{2}\log\frac{n}{2\pi e} \,] = \log \int \sqrt{\det I(\theta)} d\theta. \tag{2.5}$$

Remark 1: The upper bound part of Theorem 1 can be obtained under weaker assumptions. In particular, Condition 1 can be replaced by the requirement that $\log p_\theta(X)$ be mean square differentiable.

Remark 2: Uniform bounds on $R_n(\theta, w)$ for compact sets in Ω can be obtained when w is not compactly supported.

Remark 3: The (uniform) soundness on the compact set can be weakened to a uniform $o(1/\log n)$ rate of convergence of the posterior distribution. Specifically, we believe that sufficient conditions for uniform convergence of the posterior at rate $o(1/\log n)$ can be derived from uniform versions of hypotheses used by Wald (1949). We conjecture that a uniform version of those hypotheses will imply uniform consistency of the maximum likelihood estimator which will in turn imply that for some $\alpha > 0$ and some $\delta > 0$ we have

$$\sup_{\theta \in K} P_\theta(W(B^c(\theta,\alpha) \mid X^n) > \delta) = o(\frac{1}{\log n}).$$

Theorem 1 is seen to immediately give an asymptotic formula for the Bayes risk of the Bayes estimator under relative entropy loss when the parameter space is compact. Since the supremum in expression (2.2) tends to zero, so also does the average with respect to w. Thus we have

$$\int_K R_n(\theta, w)\, d\theta = \frac{d}{2}\log\frac{n}{2\pi e} + H(W) + \frac{1}{2}\int_K (\log \det I(\theta))\, w(\theta) d\theta + o(1),$$

a special case of (1.5).

The expansion for the Bayes risk can also be obtained for noncompact parameter spaces under suitable additional hypotheses. All we need is the pointwise result (Theorem 0) and conditions which allow us to take the limit of the integrals as an integral of the pointwise limits. We find that it is often the case that the Bayes risk is finite for reasonable choices of the prior, but the minimax value is infinite in noncompact parameter spaces due to the constant which normalizes the Jeffreys prior.

Let $\hat{\theta}$ denote the posterior mean, the coordinate-wise Bayes estimator of θ under squared error loss,

$$w_\varepsilon(\theta) = \inf_{\theta' \in B(\theta, \varepsilon)} w(\theta'),$$

$$\bar{I}_\varepsilon(\theta) = I(\theta) + M_\varepsilon(\theta),$$

where M_ε is a small positive definite matrix such that

$$D(\theta \,||\, \theta') = D(p_\theta \,||\, p_{\theta'}) \le (\theta - \theta')\bar{I}_\varepsilon(\theta)(\theta - \theta')$$

for all θ and θ' with $||\, \theta - \theta' \,|| < \varepsilon$. Also, let

$$T_{n,\varepsilon} = \{\, \theta\colon n\bar{I}_\varepsilon(\theta)\, \varepsilon^2 \ge 2\}.$$

We use the fact that the Bayes risk of the Bayes estimator, $\int R_n(\theta, w)w(\theta)\, d\theta$, is the Shannon mutual information between the parameter and the data which we denote by $I(\Theta;X^n)$. This allows us to prove the following for general parameter spaces.

Theorem 2: a)*Assume that the Bayes risk for the estimation of* θ *under squared error loss, is of order* $O(1/n)$*, that is, there exists a sequence of estimators* $\hat{\theta}$ *such that for each coordinate* θ_i *of* $\theta = (\theta_1,\ldots,\theta_d)^T$

$$\limsup_{n \to \infty} n\ E_{\Theta, X^n}(\theta_i - \hat{\theta}_i)^2 < \infty, \tag{2.6}$$

where the expectation is taken with respect to the joint distribution for Θ *and* X^n*, and that for each* θ *in the support of the prior,*

$$n\ \mathrm{cov}(\Theta \,|\, X^n) \to I^{-1}(\theta) \tag{2.7}$$

in P_{θ_o} *probability. Assume also that*

$$\int_\Omega |\, \log \det I(\theta)\, |\, w(\theta) d\theta < \infty.$$

Then we have the lower bound

$$\limsup_{n \to \infty} \left[\, I(\Theta, X^n) - \frac{d}{2}\log n\, \right]$$

$$\ge \frac{d}{2}\log \frac{1}{2\pi e} + \frac{1}{2}\int_\Omega \log \det I(\theta)\, w(\theta)\, d\theta + H(w). \tag{2.8}$$

Consequently, the minimax value satisfies the same bound,

$$\inf_{Q_n} \sup_{\theta \in \Omega} \left[\, R_n - \frac{d}{2}\log n\, \right] \ge \frac{d}{2}\log \frac{1}{2\pi e} + \log \int_\Omega w(\theta) \log \det I(\theta)\, d\theta\ . \tag{2.9}$$

b) *Suppose that (2.1) holds pointwise for* $\theta \in \Omega$*, in the* $L_1(P_\theta)$ *mode of convergence and that*

there is an $\varepsilon > 0$ *so that*

$$\int w(\theta) \log \frac{1}{w_\varepsilon(\theta)}\, d\theta < \infty, \tag{2.10}$$

and for some choice of $M_\varepsilon(\theta)$

$$\int w(\theta) \log \det \bar{I}_\varepsilon(\theta)\, d\theta < \infty, \tag{2.11}$$

and that

$$W(T^c_{n,\varepsilon}) = o(\frac{1}{\log n}). \tag{2.12}$$

Then, we have the upper bound

$$\liminf_{n \to \infty} [\, I(\Theta; X^n) - \frac{d}{2}\log n \,] \le \frac{d}{2} \log \frac{1}{2\pi e} + \frac{1}{2} \int_\Omega \log \det I(\theta)\, d\theta + H(w). \tag{2.13}$$

Together, (2.8) and (2.13) characterize the Bayes risk. There is no easy upper bound on the cumulative minimax risk in the noncompact case, because, as will be seen in the proof of Theorem 1, it requires a bound which is uniform. By contrast, the lower bound (2.9) only requires the average result (2.8).

The hypothesis (2.10) allows us to identify the entropy $H(\Theta)$ by providing an upper bound. Hypothesis (2.11) allows a Laplace integration argument to go through uniformly on the set $T_{n,\varepsilon}$. The "prior consistency" condition, (2.12), ensures that $T_{n,\varepsilon}$ increases fast enough.

In Section 4 it will be seen that some of the assumptions are used only to identify the constants so as to get the $o(1)$ convergence corresponding to (2.9) and (2.13). Weaker conditions will give $O(1)$ or coarser bounds using the techniques in Clarke and Barron (1990a, Section 5).

We note that if there is any estimator which has Bayes risk of order $O(1/n)$ then the Bayes estimator also has risk of order $O(1/n)$ since, by definition, the Bayes estimator has minimal Bayes risk.

The other key hypothesis for the lower bound was the convergence in probability of the posterior covariance. This may be difficult to verify in some cases. While use of the MLE and a result due to Bickel and Yahav (1969), will give one set of sufficient conditions as in Clarke (1989) a weaker list of assumptions can be found, as is given in Theorem 3 below. The proof is substantially due to Bickel, see Lehmann (1983), pages 454-465. There a result is proved which improves on Bickel and Yahav (1969) in that hypotheses to ensure the consistency of the MLE are only used to control the convergence of the posterior probability. Our result reduces the hypotheses further by dealing with the posterior probability directly.

The next two theorems treat asymptotic normality of the posterior. The first, Theorem 3, gives convergence in a strong enough mode to yield $n\,\mathrm{cov}(\theta \mid X^n) \to I(\theta)^{-1}$ in P_θ probability as required for Theorem 2. Let

$$S_n(\theta) = \frac{1}{n} \nabla \log p_\theta^n(X^n).$$

Theorem 3: *Assume that w is positive and continuous on Ω* and that for θ *in the interior of Ω Conditions 1, 2 and 3 are satisfied. Then,*

$$\int_\Omega \left| \frac{w(T_n + t/\sqrt{n} \mid X^n)}{n^{d/2}} - \phi_{I(\theta)}(t) \right| dt \to 0, \tag{2.14}$$

where $\phi_{I(\theta)}$ is the normal density with mean zero and covariance matrix $I(\theta)^{-1}$. Suppose that we also have

$$\int_\Omega \theta^T \theta\, w(\theta)\, d\theta < \infty. \tag{2.15}$$

Then, for $T_n = \theta + I^{-1}(\theta) S_n(\theta)$ we have that

$$\int t\, t^T \left| \frac{w(T_n + t/\sqrt{n} \mid X^n)}{n^{d/2}} - \phi_{I(\theta)}(t) \right| dt \to 0 \tag{2.16}$$

in P_θ probability. As a result we have

$$\int t\, t^T \left| \frac{w(\hat{\theta} + t/\sqrt{n} \mid X^n)}{n^{d/2}} - \phi_{I(\theta)}(t) \right| dt \to 0 \tag{2.17}$$

in P_θ probability, where $\hat{\theta} = E(\Theta \mid X^n)$.

Remark: In fact, in condition 1 we can choose $\xi = 0$. Also, the proof we give below will extend from the second moment so as to give any posterior moment provided the prior moment of that order exists. Indeed, the proof applies to a broad class of functions of the parameter, those which are bounded on compact sets, are integrable with respect to the target normal and with respect to the prior w, grow at a sub-exponential rate in probability, and allow the bounding argument after (5.21) to go through.

In part a) of Theorem 2, where we want to use (2.17), we note that condition 1 with $\xi = 0$ and condition 2 are already required since Theorem 0 has been assumed pointwise.

In their proof of (1.5) Ibragimov and Hasminsky (1973) used the asymptotic normality of a specially constructed density ratio. Also, when Bernardo conjectured the lower bound part of Theorem 2, he did so on the basis of asymptotic normality. We have also used a form of asymptotic normality by way of assumption (2.7). We show a converse result, that asymptotic normality can be obtained as a consequence of the validity of the asymptotic expansion for the

Bayes risk, expression (1.5). The mode of convergence is expected Kullback-Leibler distance, a mode which is either stronger than other modes which have been examined or is incomparable with them.

Theorem 4: *Assume that the conditions for the upper bound and lower bound in Theorem 2 are satisfied. Then, we have that*

$$\lim_{n \to \infty} E_M D(P_{\theta \mid X^n} \mid\mid N(E(\Theta \mid X^n), \operatorname{cov}(\Theta \mid X^n))) = 0 \tag{2.18}$$

If only the conditions of the upper bound are satisfied then the convergence in (2.18) is equivalent to the convergence

$$E_M \log \det n \operatorname{cov}(\Theta \mid X^n) \to \int w(\theta) \log \det I(\theta)^{-1} \, d\theta. \tag{2.19}$$

The expected Kullback-Leibler distance dominates both L^1 distance and Hellinger distance, see Csiszar (1967). Thus we have conditions which guarantee the asymptotic normality of the posterior in the sense that

$$\lim_{n \to \infty} E_{M_n} \mid w(\theta \mid X^n) - \phi_{E(\theta \mid X^n), \operatorname{cov}(\theta \mid X^n)}(\theta) \mid_1 = 0,$$

which means that, except for θ in a set of arbitrarily small measure, the same result holds with expectation defined by p_θ.

3. Proof of Theorem 1. In this section we prove Theorem 1 by showing how the proof of Theorem 0 can be extended so as to be uniformly valid on compact sets K in the parameter space. In Clarke and Barron (1990a) Section 4, three sets $A_n(\theta, \delta, \varepsilon)$, $B_n(\theta, \delta, \varepsilon)$, and $C_n(\theta, \delta)$ are introduced so as to prove Theorem 0. There, to control error terms, conditions were given which ensured that the probabilities of their complements decreased to zero fast enough. To prove Theorem 1 it is enough to demonstrate that those probabilities go to zero at the claimed rate uniformly for compact sets in Ω.

The error terms which must be controlled here are the same. Those which occur in the lower bounding of $D(p_\theta^n \mid\mid m_n)$ are given in Clarke and Barron (1990a) by expressions (4.8), (4.9), and (4.10); and by expressions (4.12) and (4.13) for the upper bound. In order to be tight, both the upper and lower bounds require that the modulus of continuity be bounded on bounded sets and that the posterior probability in Condition 3 be uniformly $o(1/\log n)$. In addition, it is seen that if (4.10) in Clarke and Barron (1990a) is controlled then the other terms in the error for the lower bound pose no problem. To control (4.10) uniformly it is enough to prove

$$\lim_{n \to \infty} \sup_{\theta \in K} P_\theta(A_n^c(\theta)) \log n = 0 \tag{3.1}$$

and

$$\lim_{n \to \infty} \sup_{\theta \in K} P_\theta(B_n^c) \log n = 0. \tag{3.2}$$

It is also seen that (4.12) and (4.13) from Clarke and Barron (1990a) that the error term for the upper bound tends to zero uniformly if

$$\lim_{n \to \infty} \sup_{\theta \in K} P_\theta(B_n^c)\, n = 0, \tag{3.3}$$

$$\lim_{n \to \infty} \sup_{\theta \in K} P_\theta(C_n^c)\, n = 0, \tag{3.4}$$

and

$$\lim_{n \to \infty} \sup_{\theta \in K} \left| nE_\theta \mathbf{1}_{B_n \cap C_n} S_n^{T\prime}(\theta) I^{-1}(\theta) S_n'(\theta) - d \right| = 0. \tag{3.5}$$

Clearly, (3.3) implies (3.2).

Proving that conditions (3.1) through (3.5) are satisfied when K is a compact set in the interior of Ω amounts to extending a proof which has already been given. Therefore, we will not go over all the details, we will merely show that the terms which arise can be controlled.

The next lemma is easy to prove; it will be used so as to control probabilities in (3.3) and (3.4) and so as to identify constants as in (3.5).

Lemma 3.1: *Let $W_{\theta,n}$ be a sequence of sets for $\theta \in K$ so that*

$$\lim_{n \to \infty} \sup_{\theta \in K} P_\theta(W_{\theta,n}) = 0.$$

Let

$$Z_{n,\theta} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_{i,\theta}$$

be a sum of mean zero, i.i.d. random variables, satisfying

$$\sup_{\theta \in K} E_\theta X_{1,\theta}^{2+\varepsilon},$$

for some positive ε. Then,

$$\lim_{n \to \infty} \sup_{\theta \in K} E_\theta \mathbf{1}_{W_{\theta,n}} Z_{n,\theta}^{2+\varepsilon} = 0.$$

Proof: By the Holder inequality we have that

$$E_\theta \mathbf{1}_{W_{\theta,n}} Z_{n,\theta}^{2+\varepsilon} \le P_\theta(W_{\theta,n})^{\varepsilon/(1+\varepsilon)} \left[E_\theta Z_{\theta,n}^{2(1+\varepsilon)} \right]^{1/(1+\varepsilon)}.$$

By the second inequality in Ibragimov and Hasminsky (1981) page 186, we have that the right

hand member is bounded from above by

$$c(\varepsilon)P_\theta(W_{\theta,n})^{\varepsilon/(1+\varepsilon)}[\ E_\theta\ |\ X_{1,\theta}|^{2(1+\varepsilon)}\ ]^{1/(1+\varepsilon)},$$

where c is a constant dependent only on ε. This last expression tends to zero as n increases. □

Proof of Theorem 1: First we show how a uniform version of Theorem 0 implies the conclusions about the minimax value, expression (2.4), and about Jeffreys' prior, expression (2.5). We start with (2.4). The minimax risk is

$$R_n = R(n, K, \{P_\theta\}) = \inf_{Q_n} \sup_{\theta \in K} D(P_\theta^n \,||\, Q_n),$$

and we denote the maximin risk by

$$\begin{aligned} R_n^* = R^*(n, K, \{P_\theta\}) &= \sup_w \inf_{Q^n} D(P_\theta^n \,||\, Q^n) \\ &= \sup_w D(P_\theta^n \,||\, M_n). \end{aligned}$$

The minimax risk is realized by the minimax estimator. We can upper bound the minimax risk by replacing the minimax estimator with any other estimator, the Bayes estimator with respect to Jeffreys prior, for instance. So, we have that

$$R_n - \frac{d}{2}\log n \le \sup_{\theta \in K}\left[\, D(P_\theta^n \,||\, M_n) - \frac{d}{2}\log n\ \right],$$

in which case the right hand side is upper bounded by

$$\frac{d}{2}\log\frac{1}{2\pi e} + \log\int_K \sqrt{\det I(\theta)}\, d\theta + o(1)$$

by (2.2). The minimax risk is lower bounded by the maximin risk. In turn, it is lower bounded by replacing the least favorable prior with any other prior, the Jeffrey's prior for instance. So, we have that

$$\begin{aligned} R_n - \frac{d}{2}\log n &\ge R_n^* - \frac{d}{2}\log n \\ &\ge \int_K w_J(\theta) D(P_\theta \,||\, M_{n,w_J})\, d\theta - \frac{d}{2}\log n. \end{aligned}$$

By (2.3), the right hand side has

$$-(d/2)\log 2\pi e + \log\int_K \sqrt{\det I(\theta)}\, d\theta - o(1)$$

as a lower bound. Since the upper and lower bounds agree (2.4) is proved, the Jeffreys' prior is asymptotically least favorable, and the minimax estimator is the predictive density with respect to the Jeffreys' prior.

It remains to verify that Theorem 0 holds uniformly on compact sets. Fix $\theta \in K$. We show that there is an open set containing θ on which Theorem 0 holds uniformly. By the compactness of K this will be enough since we may cover K by finitely many such open sets.

Next, we note that (3.3) and (3.4) imply (3.5), by use of the moment assumptions in Condition 1 and Lemma 3.1. Proving (3.3) and (3.4) amounts to proving that the functions c_1 and c_2 in expressions (4.16) and (4.17) in Clarke and Barron (1990a) are continuous as functions of θ. This, too, can be done by Lemma 3.1, and the moment assumptions in Condition 1 by directly uniformizing the reasoning of Clarke and Barron (1990a) Section 4.

Now, it is enough to prove (3.1) for some open set containing θ.

We begin by noting that the assumption of soundness pointwise carries over to uniform soundness of the parametric family on compact sets. Specifically, consider the mapping

$$g: \theta \rightarrow P_\theta,$$

where the domain of g is a compact set K' which contains K in its interior and the range of g is the collection of all probability measures defined on X, endowed with the topology generated by weak convergence. By Condition 3, g is continuous. Therefore, the image of g is compact also. By soundness, g is one to one. So, g is a homeomorphism onto its image. By compactness, g and its inverse are both uniformly continuous.

Following Gray (1988), Section 8.2, we define the metric d_G on probabilities from the metric d_X on X. Let $G = \{F_1, F_2, \ldots\}$ be the countable field of sets generated by balls of the form $\{x: d_X(x, s_j) \leq 1/k\}$ for $j,k = 1, 2, \ldots$ where $s_1, s_2, \ldots$ is a countable dense sequence in X. We define d_G on probabilities P and Q by

$$d_G(P, Q) = \sum_{i=1}^{\infty} 2^{-1} | P(F_i) - Q(F_i) |.$$

Convergence in d_G is stronger than convergence in the Prohorov metric which metrizes the topology of weak convergence. However, we note that, when restricted to probabilities in the parametric family, the topology of weak convergence metrized by the Prohorov metric is equivalent to the topology metrized by d_G. This can be seen as follows. Under conditions 1 and 2, $\theta' \rightarrow \theta$ implies $D(P_{\theta'} \,||\, P_\theta) \rightarrow 0$, which is stronger than L_1 convergence. That implies the convergence of $P_{\theta'}$ to P_θ setwise, which gives convergence in d_G implying weak convergence, which is equivalent to convergence of the parameter values, by soundness.

Next we observe that Propositions 6.1, and 6.3 from Clarke and Barron (1990a) carry over to the present setting. By straightforward modifications of their proofs, Propositions (6.1) and (6.3) now give us the following two facts.

One: Since d_G satisfies

$$\sup_P P^n\{\, d_G(\hat{P}_n, P) > \varepsilon\} \le e^{-nr} \tag{3.6}$$

where $r > 0$, $\hat{P}_n$ is the empirical probability and the supremum is taken over all probability measures on X and g is uniformly continuous, there exists an $\eta_1 > 0$, $r > 0$ and a critical region C_n so that given any $\theta' \in B(\theta, \eta)$, and any $\delta > 0$, the hypothesis test $\theta = \theta'$ versus $\{\theta: |\theta - \theta'| > \delta\}$ has both probability of type one error and probability of type two error bounded above by e^{-nr}.

We remark that we need a topology which is metrizable by a metric d which satisfies (3.6). The metrizability cannot be avoided since the relative topology on the image of g is metrizable, and some form of (3.6) cannot be avoided since errors must be controlled. The form of the uniformity required so as to generalize Proposition 6.1 in Clarke and Barron (1990a), namely given $\delta > 0$ there exists an $\varepsilon > 0$ so that for any $\theta, \theta' \in \Omega$, $|\theta - \theta'| > \delta$ implies $d_G(P_\theta, P_{\theta'}) > \varepsilon$, forces us to use the topology generated by d_G rather than the weaker topology generated by weak convergence.

Two: Given $\delta > 0$, there exists $\eta_2 > 0$ and $r > 0$ so that we have

$$P_{\theta'}\Big(\int_{N(\theta', \delta)} w(\theta) p(x^n \mid \theta)\, d\theta < e^{nr} \int_{N(\theta', \delta)^c} w(\theta) p(x^n \mid \theta)\, d\theta\Big) = O\Big(\frac{1}{n}\Big),$$

uniformly for θ' in any closed set in $B(\theta, \eta_2)$ contained in K. The boundary points of K are no difficulty since any open set which contains one of them has positive prior probability.

From facts one and two we can conclude that there is an open set about θ, say $B(\theta, \eta)$, with compact closure inside K', so that for any $\theta' \in B(\theta, \eta)$ we have that the $P_{\theta'}$ probability of $A_n(\theta')^c$, the set on which we have posterior consistency fails, has probability decreasing at rate $O(1/n)$. Covering K with finitely many open sets $B(\theta_i, \eta_i)$, $i = 1,...,k$ we have the upper bound

$$\lim_{n \to \infty} \sup_{\theta \in K} P_\theta(A_n^c(\theta)) \le \max_{i=1}^{k} \lim_{n \to \infty} \sup_{\theta \in B(\theta_i, \eta_i)} P_\theta(A_n^c(\theta))$$

$$\le O(1) \lim_{n \to \infty} \frac{\log n}{n} = 0.$$

This proves (3.1).

We now have that Theorem 0 holds uniformly on K, which implies both (2.2) and (2.3).

□

4. Proof of Theorem 2. We start with the lower bound. We will be concerned with the behavior of various quantities under the mixture distribution for X^n, and under the joint distribution for Θ and $X_1,..., X_n,....$ It is here that we use the maximum entropy argument.

Lemma 4.1: *Suppose that for each* θ *we have that*

$$f_n(X^\infty) \to f(\theta),$$

in $P_{X^\infty | \theta}$. *Then the convergence holds in the joint measure:*

$$f_n(X^\infty) - f(\theta) \to 0,$$

in P_{Θ, X^∞}.

Proof: Let $\varepsilon > 0$ and note:

$$P_{\Theta, X^\infty}(\mid f_n(X^\infty) - f(\theta) \mid > \varepsilon) = \int_{\mathbf{R}^d} w(\theta) P_{X^\infty | \theta}(\mid f_n(X^\infty) - f(\theta) \mid > \varepsilon) d\theta,$$

which goes to zero by the dominated convergence theorem. □

The next idea we introduce so as to obtain a lower bound is a one sided version of uniform integrability. Following Chow and Teicher (1978) we say that a sequence of random variables Y_n is uniformly integrable from above if and only if its positive part is uniformly integrable. Equivalent to uniform integrability from above is the condition

$$\lim_{r \to \infty} \sup_n E\ Y_n\ \mathbf{1}_{\{Y_n > r\}} = 0.$$

We only use uniform integrability from above since obtaining a lower bound on $I(\Theta, X^n)$ will require us to upper bound the conditional entropy term which arises in its definition.

We next prove three lemmas. The first gives sufficient conditions which we will use to show that $\log \det n\,\mathrm{cov}(\theta \mid X^n)$ is uniformly integrable from above. It is modeled on the proof in Billingsley (1986), pg. 348.

Lemma 4.2: *If a sequence of positive random variables* Y_n *satisfies*

$$\sup_n E\ Y_n < \infty,$$

then $Z_n = \log Y_n$ *is uniformly integrable from above.*

Proof: Let $g(r) = e^r$. Then, for $r > 1$, the function re^{-r} is decreasing and consequently we have the inequalities

$$0 \le \sup_n E\ Z_n \mathbf{1}_{\{Z_n > r\}} = \sup_n E\ g(Z_n) \frac{Z_n \mathbf{1}_{\{Z_n > r\}}}{g(Z_n)}$$

$$\le \frac{r}{g(r)} \sup_n E\ g(Z_n).$$

By assumption the expectation on the right is finite and $r/g(r)$ converges to zero as $r \to \infty$, so the lemma is proved. □

The next lemma uses uniform integrability from above to identify how a limit of expectations is related to the expectation of the limit.

Lemma 4.3: *If Y_n is uniformly integrable from above and converges in probability to a random variable Z, then*

$$\limsup_{n \to \infty} E\ Y_n \leq E\ Z.$$

Proof: Write

$$EY_n = EY_n \mathbf{1}_{\{ Y_n \leq r\}} + EY_n \mathbf{1}_{\{ Y_n > r\}}.$$

For fixed r, the limit superior of the first term is bounded by $E\ Z\mathbf{1}_{\{Z \leq r\}}$ since the random variables $Y_n\ \mathbf{1}_{\{Y_n < r\}}$ are bounded above. For r large enough, the second term is finite by uniform integrability from above. As r increases, we have the result. □

Our fourth lemma is an identity which relates the Bayes risk of the Bayes estimator to two other terms which are easier to analyze. We will see that one tends to the constant $d/2$, pointwise in θ, as n increases. The other term has a form to which Laplace integration can be applied readily. The asymptotically constant term involves an approximation to the posterior which we denote

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

where c_n is a normalizing constant.

Lemma 4.4 *Assume Condition 1 is satisfied and that w has a density with respect to Lebesgue measure. Then, we have*

$$R_n(\theta, w) + E_\theta D(w^* \,||\, w(\cdot \,|\, X^n)) = -\log \int_{\mathbf{R}^n} e^{-nD(\theta \,||\, \theta')} w(\theta')\, d\theta'. \tag{4.1}$$

Proof: Since Condition 1 is satisfied ν is well defined for each n. Proving (4.1) is a calculation. We note that $E_\theta D(\nu \,||\, w(\cdot \,|\, X^n))$ can be written in terms of ν and the posterior density for θ given X^n. Using the definition of ν and Bayes rule on the posterior gives an expression which can be rearranged to yield (4.1).□

Now we use the four preceding lemmas to give a proof of part a) of Theorem 2. Here, the operator E by itself means expectation with respect to the joint distribution.

Proof of Theorem 2, lower bound part: The Bayes risk $R_n(w) = \int_\Omega D(p_\theta^n \,||\, m_n) w(\theta)\, d\theta$ is equal to Shannon's mutual information which we expand as the difference between the entropy

of the prior $H(w) = H(\Theta)$ and its conditional entropy

$$H(w \mid X^n) = \int_{X^n} \int_{\Omega} w(\theta \mid x^n) \log \frac{1}{w(\theta \mid x^n)} \, d\theta \, m(x^n) \, dx^n,$$

which we also denote by $H(\Theta \mid X^n)$. Therefore the Bayes risk is

$$\begin{aligned} I(\Theta;X^n) &= H(\Theta) - H(\Theta \mid X^n) \\ &= H(\Theta) - \int_{X^n} H(\Theta \mid X^n = x^n) m(x^n) \lambda(dx^n) \\ &= H(\Theta) - \int_{X^n} H(\Theta - \hat{\theta} \mid X^n = x^n) m(x^n) \lambda(dx^n) \\ &\geq H(\Theta) - \frac{1}{2} \int_{X^n} m(x^n) \log [(2\pi e)^d \det E_{w(\cdot \mid x^n)} n(\Theta - \hat{\theta})(\Theta - \hat{\theta})^t \,] \lambda(dx^n) \\ &= H(\Theta) + \frac{d}{2} \log \frac{n}{2\pi e} \\ &\quad - \frac{1}{2} \int_{X^n} m(x^n) \log \det E_{w(\cdot \mid x^n)} \sqrt{n} (\Theta - \hat{\theta}) \sqrt{n} (\Theta - \hat{\theta})^t \, \lambda(dx^n), \end{aligned} \tag{4.2}$$

where the inequality comes from the fact that the normal achieves the maximal entropy under a covariance constraint.

We will show that $\log \det n \operatorname{cov}(\theta \mid X^n)$ is uniformly integrable from above with respect to the mixture by bounding it with a sum of functions each of which is uniformly integrable from above. By Hadamard's inequality we have the following bounds:

$$\log \det [\, n \, \operatorname{cov}(\Theta \mid X^n) \,] \leq \sum_{i=1}^{d} \log [\, n \, Var(\Theta_i \mid X^n) \,] \tag{4.3}$$

By assumption,

$$\sup_{n,i} E_{M_n} E_{\Theta \mid X^n} n(\Theta_i - \hat{\theta}_i)^2 < \infty,$$

so, by Lemma 4.2 we have that each

$$\log E_{\Theta \mid X^n} n(\Theta_i - \hat{\theta}_i)^2$$

is uniformly integrable from above. Thus, the right hand member of (4.3) is uniformly integrable from above. This implies that

$$\log \det [\, n \, \operatorname{cov}(\theta \mid X^n) \,]$$

is uniformly integrable from above, and therefore so is

$$\log \det [\, n \, \operatorname{cov}(\theta \mid X^n) \,] + \log \det I(\theta).$$

By assumption we have that

$$\log\det n\ [\ \mathrm{cov}(\theta \mid X^n)\ ] + \log\det I(\theta) \to 0,$$

in $P_{X^n \mid \theta}$ probability, for each θ in the support of w, and therefore, by Lemma 4.1, in the joint probability of (Θ, X^∞). Now, by Lemma 4.3,

$$\limsup_{n\to\infty} E_m\ [\ \log\det n\,\mathrm{cov}(\theta \mid X^n)\ ] \le -\int_\Omega \log\det I^{-1}(\theta) w(\theta)\, d\theta.$$

Finally, from inequality (4.2), we have that

$$\liminf_{n\to\infty} [\ I(\Theta; X^n) - H(\Theta) - \frac{d}{2}\log\frac{n}{2\pi e}\ ] \ge -\limsup_{n\to\infty} E_{M_n}\ [\ \log\det n\,\mathrm{cov}(\theta \mid X^n)\ ]$$
$$= \int_\Omega w(\theta)\log\det I(\theta)\, d\theta,$$

which proves part a) of the theorem.

Proof of Theorem 2, upper bound part: By application of Theorem 0 to the second term on the left in equation (4.1) in Lemma 4.4, and Laplace integration on the right hand member we know that the left hand member tends to $d/2$ in P_θ probability for each θ. Integrating (4.1) with respect to w we have that

$$\int_\Omega w(\theta) D(P_\theta^n \mid\mid M_n)\, d\theta$$
$$= -\int_\Omega w(\theta) E_\theta\, D(\nu \mid\mid w(\cdot \mid X^n))\, d\theta - \int_\Omega w(\theta)\log\int_\Omega e^{-nD(\theta \mid\mid \theta')} w(\theta')\, d\theta'. \qquad (4.4)$$

Since $E_\theta\, D(\nu \mid\mid w(\cdot \mid X^n)$ is positive we can apply Fatou's lemma to see that the limit inferior of its integral with respect to w is bounded below by $d/2$ also. So, the first term on the right in (4.4) is upper bounded by $-d/2$.

We upper bound the second term in (4.4) by

$$-\int_\Omega w(\theta)\log \int_{|\theta-\theta'|<\varepsilon} e^{-nD(\theta \mid\mid \theta')} w(\theta')\, d\theta'\ d\theta$$
$$\le -\int_\Omega w(\theta)\log \int_{|\theta-\theta'|<\varepsilon} e^{-(n/2)(\theta-\theta')^T \bar{I}_\varepsilon(\theta)(\theta-\theta')} w(\theta')\, d\theta'\ d\theta$$
$$\le -\int_\Omega w(\theta)\log w_\varepsilon(\theta)\, d\theta \qquad (4.5)$$
$$-\int_{T_{n,\varepsilon}} w(\theta)\log \int_{|\theta-\theta'|<\varepsilon} e^{-(n/2)(\theta-\theta')^T \bar{I}_\varepsilon(\theta)(\theta-\theta')} w(\theta')\, d\theta'\ d\theta \qquad (4.6)$$
$$-\int_{T_{n,\varepsilon}^c} w(\theta)\log \int_{|\theta-\theta'|<\varepsilon} e^{-(n/2)(\theta-\theta')^T \bar{I}_\varepsilon(\theta)(\theta-\theta')} w(\theta')\, d\theta'\ d\theta \qquad (4.7)$$

Term (4.5) is finite for ε small enough, by assumption. For given ε sufficiently small we can show that term (4.7) is upper bounded by a quantity which tends to zero as n increases. Indeed, by the definition of $T_{n,\varepsilon}$ we have that (4.7) is upper bounded by

$$- \int_{T_{n,\varepsilon}^c} w(\theta) \log \int_{|\theta - \theta'| < \varepsilon} e^{-2}\, d\theta'\, d\theta = - W(T_{n,\varepsilon}^c) \log e^{-2} V_\varepsilon, \tag{4.8}$$

where V_ε is the volume of the ball in d dimensions of radius ε. Now (4.8) tends to zero since the prior probability of $T_{n,\varepsilon}^c$ does.

It is term (4.6) which provides the dependence on n. By recognizing the normal form of the integral we see that it is

$$- \int_{T_{n,\varepsilon}} w(\theta) \log \sqrt{2\pi}\, |\, \bar{I}_\varepsilon(\theta)|^{-1/2} w(\theta)\, d\theta - \int_{T_{n,\varepsilon}} w(\theta) \log P(\, |\, Z_\theta| < \varepsilon)\, d\theta, \tag{4.9}$$

where $Z_\theta \sim N(0, (n\bar{I}_\varepsilon(\theta))^{-1})$. The first term in (4.9) gives us what we want since $W(T_{n,\varepsilon}) \geq 1 - o(1/\log n)$. The second term in (4.9) is asymptotically upper bounded by an arbitrarily small number. Indeed, it is

$$\begin{aligned} &- \int_{T_{n,\varepsilon}} w(\theta) \log (1 - P(|\, Z_\theta| > \varepsilon)\,)\, d\theta \\ &\qquad \leq -\varepsilon \int_{T_{n,\varepsilon}} w(\theta) \log (1 - (n\bar{I}_\varepsilon(\theta))^{-1}/\varepsilon^2\,)\, d\theta. \end{aligned} \tag{4.10}$$

By the definition of $T_{n,\varepsilon}$ we have that $1/(n\bar{I}_\varepsilon(\theta)\varepsilon^2) \leq 1/2$, so the logarithm in (4.10) is well defined. Thus the integrand converges to zero pointwise, as a function of θ and is bounded. By the dominated convergence theorem the integral goes to zero as n increases.

Now, we have that for any $\eta > 0$, any ε small enough, and n large,

$$\begin{aligned} \int_\Omega R_n(\theta, w)\, w(\theta)\, d\theta - \frac{d}{2} \log \frac{n}{2\pi} \leq &-(\frac{d}{2} - \eta) - \int_\Omega w(\theta) \log w_\varepsilon(\theta)\, d\theta \\ &+ \int w(\theta) \log \det \bar{I}_\varepsilon(\theta)\, d\theta + \eta. \end{aligned}$$

Next, we take the limit superior as n increases. Letting η go to zero and then letting ε go to zero finishes the proof of the upper bound. □

5. Proofs of Theorems 3 and 4. In this section we prove our two theorems concerning posterior convergence. First, we give a proof of Theorem 3 since its conclusion is used as a hypothesis for Theorem 2. Our proof is modeled on that due to Bickel, see Lehmann (1983), but where Bickel's proof essentially used the consistency of the MLE to prove that the integral outside of an open set around θ was negligible, we use posterior consistency as guaranteed by

Condition 3.

One more point bears mention. In Bickel's proof the standardized parameter is used. Here we use the Jacobian transformation so that the result is phrased in terms of the actual parameter. Indeed, the relation between the two forms is

$$\pi(t \mid X^n)dt = \frac{w(\theta(X^n) + t/\sqrt{n} \mid X^n)}{n^{d/2}}\, dt,$$

where π is the posterior for $t = \sqrt{n}\,(\theta - \theta(X^n))$ and $\theta(X^n)$ is an estimator or θ, here either the Bayes estimator $\hat{\theta}$ under squared error loss, or the pseudo - estimator of Bickel, T_n. In writing $\pi(t \mid X^n)$ for the posterior we assume that w is defined on all of $\mathbf{R}^d$. This can be done by letting w be zero off of its domain of definition.

Proof of Theorem 3: We may write the posterior as

$$\pi(t \mid X^n) = \frac{w(T_n + t/\sqrt{n})e^{\log p(X^n \mid T_n + t/\sqrt{n})}}{\int w(T_n + t/\sqrt{n})e^{\log p(X^n \mid T_n + t/\sqrt{n})}\, dt},$$

which may be written as

$$\pi(t \mid X^n) = \frac{e^{u(t)}w(T_n + t/\sqrt{n})}{C_n},$$

where

$$u(t) = \log p(X^n \mid T_n + t/\sqrt{n}) - \log p(X^n \mid \theta_o)$$
$$- \frac{1}{2n} \nabla p(X^n \mid \theta_o)^T I(\theta_o)^{-1} \nabla p(X^n \mid \theta_o),$$

and

$$C_n = \int_{\mathbf{R}^d} e^{u(s)}w(T_n + s/\sqrt{n})\, ds.$$

We assume that some θ_o has been fixed so that $T_n = T_n(\theta_o)$. First we prove (2.14), that is, we show

$$\int_{\mathbf{R}^d} \mid \pi(t \mid X^n) - \sqrt{\det I(\theta_o)}\phi(I^{1/2}(\theta_o)t) \mid\, dt \to 0,$$

in P_{θ_o} probability, where ϕ is a normal density with mean zero and covariance matrix the $d \times d$ identity matrix. The key step in the proof is proving that

$$J = \int_{\mathbf{R}^d} \mid e^{u(t)}w(T_n + t/\sqrt{n}) - e^{-t^T I(\theta_o)t/2}w(\theta_o) \mid\, dt \to 0, \tag{5.1}$$

in P_{θ_o} probability. The convergence in (5.1) implies that

$$C_n \rightarrow w(\theta_o)\sqrt{\det 2\pi I(\theta_o)},$$

in P_{θ_o} probability, which we shall require for our result. Showing that J tends to zero under our assumptions can be done as follows. Given $\delta > 0$, we decompose the parameter space into three parts: $| t | \leq M$ for some large M, $M \leq | t | \leq \delta \sqrt{n}$, and $| T | > \delta \sqrt{n}$. Next we bound J by taking bounds on the integrals over those regions. This gives

$$J \leq \sup_{|t| \leq M} | e^{u(t)} w(T_n + t/\sqrt{n}) - e^{-t^T I(\theta_o)t/2} w(\theta_o) | \tag{5.2}$$

$$+ \int_{M \leq |t| \leq \delta\sqrt{n}} | e^{u(t)} w(T_n + t/\sqrt{n}) - e^{-t^T I(\theta_o)t/2} w(\theta_o) | \, dt \tag{5.3}$$

$$+ \int_{|t| > \delta\sqrt{n}} | e^{u(t)} w(T_n + t/\sqrt{n}) - e^{-t^T I(\theta_o)t/2} w(\theta_o) | \, dt \tag{5.4}$$

where M can be chosen by us. We show that a) with the exception of the "MLE consistency" condition, B3 in Lehmann (1983) p. 455, Bickel's conditions are implied by ours and that b) the "MLE consistency" condition can be replaced by soundness, our Condition 3.

Since B3 was not used in showing that expressions (5.2) and (5.3) go to zero in P_{θ_o} probability, Bickel's reasoning can be applied to those terms if a) is proved. Reviewing Bickel's proof we see that the same technical assumptions have been made here. So, to prove a), it is enough to prove that for each positive ε there is a positive δ so that

$$P_{\theta_o}\left(\sup_{\{\theta: | \theta - \theta_o | \leq \delta\}} | \frac{1}{n} E_n(\theta) | > \varepsilon \right) \rightarrow 0, \tag{5.5}$$

where E_n is defined by the Taylor expansion

$$\log p(X^n | \theta) - \log p(X^n | \theta_o)$$

$$= (\theta - \theta_o) \nabla p(X^n | \theta_o) - \frac{1}{2}(\theta - \theta_o)^T [nI(\theta_o) + E_n(\theta)](\theta - \theta_o).$$

By adding and subtracting we upper bound the event in (5.5)

$$\sup_{\{\theta: | \theta - \theta_o | \leq \delta\}} | \frac{1}{n} E_n(\theta) |$$

$$\leq \sup_{\{\theta: | \theta - \theta_o | \leq \delta\}} | -D(P_{\theta_o} || P_\theta) + \frac{1}{2}(\theta - \theta_o)^T I(\theta_o)(\theta - \theta_o) |.$$

$$+ \sup_{\{\theta: | \theta - \theta_o | \leq \delta\}} | \frac{1}{n} \sum_{i=1}^{n} \log \frac{p(X_i | \theta)}{p(X_i | \theta_o)} + D(P_{\theta_o} || P_\theta) |$$

$$+ \delta | S_n | \tag{5.6}$$

So, it is enough to show that each term on the right of (5.6) tends to zero in probability. The third does by Condition 1 by use of Chebyshev's inequality. The first is made less than $\varepsilon/2$ by choice of δ sufficiently small by use of the Taylor expansion which exists by Conditions 1 and 2. The second is upper bounded as follows.

We have that term 2 is upper bounded by

$$\frac{1}{n}\sum_{i=1}^{n} \sup_{\{\theta:\ |\theta-\theta_o|\le\delta\}} \left| \log \frac{p(X_i \mid \theta)}{p(X_i \mid \theta_o)} + D(P_{\theta_o} \,||\, P_\theta)\right|, \tag{5.7}$$

so adding and subtracting

$$E_{\theta_o} \sup_{\{\theta:\ |\theta-\theta_o|<\delta\}} \left| \log \frac{p(X_i \mid \theta)}{p(X_i \mid \theta_o)} + D(\theta_o \,||\, \theta)\right| \tag{5.8}$$

which tends to zero as δ tends to zero, allows us to apply Chebyshev's inequality to the difference between (5.7) and (5.8), provided that

$$E_{\theta_o} \sup_{\{\theta:\ |\theta-\theta_o|<\delta\}} \left| \log \frac{p(X_i \mid \theta)}{p(X_i \mid \theta_o)}\right|^2 \tag{5.9}$$

is finite. We show that the finiteness of (5.9) is implied by Condition 1.

By the Taylor expansion we have that (5.9) is bounded by a the sum of

$$E_{\theta_o} \sup_{\theta:\ |\theta-\theta_o|<\delta} \left| \frac{\partial}{\partial\theta_i} \log p(X \mid \theta)\right|^2. \tag{5.10}$$

over i. Adding and subtracting $\frac{\partial}{\partial\theta_i}\log p(X \mid \theta_o)$ within the absolute values, expanding the square and using a first order Taylor expansion on the difference of first derivatives gives the upper bound which is the sum of

$$E_{\theta_o} \sup_{\theta:\ |\theta-\theta_o|<\delta} \left| \frac{\partial^2}{\partial\theta_i\partial\theta_j} \log \frac{p(X \mid \theta)}{p(X \mid \theta_o)}\right|^2 + E_{\theta_o} \left| \frac{\partial}{\partial\theta_i} \log p(X \mid \theta_o)\right|^2$$

$$+ E_{\theta_o} \left| \frac{\partial}{\partial\theta_i} \log p(X \mid \theta_o)\right| \sup_{\theta:\ |\theta-\theta_o|<\delta} \left| \frac{\partial^2}{\partial\theta_i\partial\theta_j} \log \frac{p(X \mid \theta)}{p(X \mid \theta_o)}\right|. \tag{5.11}$$

over j. The first two terms in (5.11) are finite by condition 1. By applying the Cauchy-Schwarz inequality, we see that the third is finite by Condition 1 also. Now, by Bickel's argument we have that terms (5.2) and (5.3) go to zero as n increases.

We deal with (5.4) differently. Note that the second term in the integrand can be neglected so that we are left with

$$\int_{|t|\ge\delta\sqrt{n}} e^{u(t)} w(T_n + t/\sqrt{n})\, dt. \tag{5.12}$$

Up to a factor which converges to a nonzero number, expression (5.12) is a posterior probability. Indeed, it is

$$C_n\ \pi(|t| \geq \delta\sqrt{n} \mid X^n) = C_n W(|\ \theta - T_n(\theta_o)| \geq \delta |\ X^n)$$
$$\leq C_n W(|\ \theta - \theta_o\ | \geq \delta - |\theta_o - T_n(\theta_o)|\ X^n). \tag{5.13}$$

We will prove that

$$W(|\ \theta - \theta_o\ | \geq \delta - |\theta_o - T_n(\theta_o)|\ \mid X^n)$$

goes to zero in P_{θ_o} probability and that

$$P_{\theta_o}(C_n > K) \leq a/(\log K + b)$$

for some choice of a and b. This will imply that expression (5.13) goes to zero in P_{θ_o} probability. We have that for each K

$$P_{\theta_o}(C_n W(|\ \theta - \theta_o\ | \geq \delta - |\theta_o - T_n(\theta_o)|\ \mid X^n) > \varepsilon)$$
$$\leq P_{\theta_o}(KW(|\ \theta - \theta_o\ | \geq \delta - |\theta_o - T_n(\theta_o)|\ \mid X^n) > \varepsilon) + P_{\theta_o}(C_n > K)$$

in which the first term in the upper bound goes to zero since the posterior probability converges to zero, and the second term can be made arbitrarily small since C_n is bounded in probability.

That $W(|\ \theta - \theta_o\ | \geq \delta - |\theta_o - T_n(\theta_o)|\ \mid X^n)$ goes to zero in P_{θ_o} probability follows from soundness. For, given $\eta > 0$, we have that

$$P_{\theta_o}(W(|\theta - \theta_o\ | \geq \delta - |\theta_o - T_n\ |\ \mid X^n) > \eta)$$
$$\leq P_{\theta_o}(W(|\ \theta - \theta_o\ | > \delta/2) + P_{\theta_o}(|\ \theta_o - T_n\ | > \delta/2). \tag{5.14}$$

By soundness the first term in (5.14) goes to zero. The second term is

$$P_{\theta_o}(|\ I^{-1}(\theta_o)S_n(\theta_o)| > \delta/2),$$

which for $\gamma = \gamma(\delta)$ small enough is upper bounded by

$$P_{\theta_o}(\frac{1}{n}|\ \nabla \log p\,(X^n \mid \theta_o)| > \gamma) \leq \sum_{i=1}^{d} P_{\theta_o}(\frac{1}{n}(\frac{\partial}{\partial\theta_i}\log p\,(X^n \mid \theta_o))^2 > \gamma)$$
$$\leq \frac{1}{\gamma}\sum_{i=1}^{d} Var_{\theta_o}(\frac{1}{n}\frac{\partial}{\partial\theta_i}\log p\,(X^n \mid \theta_o))$$

which is $O(1/n)$. So, (5.14) tends to zero.

We show that C_n is bounded in probability. It can be written as

$$C_n = n^{d/2}\,\frac{m(X^n)}{p\,(X^n \mid \theta_o)}\,e^{-\frac{n}{2}S_n(\theta_o)^T I^{-1}(\theta_o)S_n(\theta_o)}$$

in which we recognize that the density ratio is the same as in Clarke and Barron (1990a), expression (4.4). For given K large we have

$$P_{\theta_o}(C_n > K) = P_{\theta_o}\left(\frac{m(X^n)}{p(X^n \mid \theta_o)} \geq Kn^{d/2} e^{\frac{n}{2} S_n(\theta_o)^T I^{-1}(\theta_o) S_n(\theta_o)} \right) \tag{5.15}$$

to which we can apply (4.4). Expression (5.15) is upper bounded by intersecting with the set on which (4.4) holds and with the complement of that set. This gives

$$P_{\theta_o}((4.4) \text{ violated}) + P_{\theta_o}\left(c(\varepsilon, \delta, \theta_o) e^{\frac{n}{2(1-\varepsilon)} S_n^T I(\theta_o)^{-1} S_n} \geq K \, e^{\frac{n}{2} S_n^T I(\theta_o)^{-1} S_n}\right) \tag{5.16}$$

where c is a function of the arguments listed. The first term in (5.16) goes to zero under our assumptions as in Clarke and Barron (1990a) and the second term can be upper bounded by Markov's inequality. Taking logarithms inside the probability and rearranging gives

$$P_{\theta_o}\left(\frac{n}{2} S_n^T I^{-1}(\theta_o) S_n \geq \frac{\log K - \log c}{1/(1-\varepsilon) - 1} \right)$$

which is upper bounded by $d/2(\frac{\log k - \log c}{1/(1-\varepsilon) - 1})^{-1}$. This means that we have proved that (5.4) goes to zero in P_{θ_o} probability. Thus we have that J goes to zero and that C_n goes to the constant $w(\theta_o)\sqrt{2\pi/\det I(\theta_o)}$ and the proof of (2.14) is complete.

Next, we use (2.14) to prove that the posterior variance converges to that of the normal. That follows if we prove that for each i, j from 1 to d that

$$J_1 = \int_{\mathbf{R}^d} (1 + \mid t_i t_j \mid) \mid \pi(t \mid X^n) - \sqrt{\det I(\theta_o)}\, \phi(t\sqrt{\det I(\theta_o)}) \mid dt$$

goes to zero in P_{θ_o} probability. We use the same sort of decomposition as before:

$$J_1 \leq \sup_{\mid t \mid \leq M} (1 + \mid t_i t_j \mid) \mid \pi(t \mid X^n) - \sqrt{\det I(\theta_o)}\, \phi(t\sqrt{\det I(\theta_o)}) \mid \tag{5.17}$$

$$+ \int_{M \leq \mid t \mid \leq \delta\sqrt{n}} (1 + \mid t_i t_j \mid) \mid \pi(t \mid X^n) - \sqrt{\det I(\theta_o)}\phi(t\sqrt{\det I(\theta_o)}) \mid dt \tag{5.18}$$

$$+ \int_{\mid t \mid \geq \delta\sqrt{n}} (1 + \mid t_i t_j \mid) \mid \pi(t \mid X^n) - \sqrt{\det I(\theta_o)}\phi(t\sqrt{\det I(\theta_o)}) \mid dt. \tag{5.19}$$

Since $\mid t_i t_j \mid$ is bounded on bounded sets, Bickel's reasoning for (5.2) applies to (5.17). Analogously to (5.3), we can show that (5.18) tends to zero by showing that its integrand is bounded above by an integrable function with P_{θ_o} probability greater than $1 - \varepsilon$. Then (5.18) can be made small in probability by choosing M large. It is enough to show that for given $\varepsilon > 0$ there is a $\delta > 0$ and a finite C so that for n large

$$P_{\theta_o}((1 + \mid t_i t_j \mid) e^{u(t)} w(T_n + t/\sqrt{n}) \leq C e^{-t^T I(\theta_o)^{-1} t/4} \text{ for } \mid t \mid < \delta/\sqrt{n}) \geq 1 - \varepsilon. \tag{5.20}$$

By (2.15), the prior moment condition, we can define a new prior

$$w(\theta) = \frac{(1 + | \theta_i \theta_j |)w(\theta)}{k_{i,j}}$$

where $k_{i,j}$ is a normalizing constant. If we divide by $k_{i,j}$ inside the probability in (5.20) then we can observe that Bickel's proof continues to apply to (5.20) so that (5.18) goes to zero in P_{θ_o} probability. Similarly, our asymptotically zero upper bound on (5.4) applies to (5.19) by using the new prior $w(\theta)$. So, the posterior variances converge as in (2.16).

To derive (2.17) we must relocate the posterior density at the Bayes estimator $\hat{\theta}$ rather than at T_n. We have that

$$\begin{aligned}
\sqrt{n}\,(\hat{\theta} - T_n) &= E(\sqrt{n}\,(\theta - T_n) \mid X^n) \\
&= \int_{\Omega} \sqrt{n}\,(\theta - T_n) w(\theta \mid X^n) d\theta \\
&= \int_{\mathbf{R}^d} u\; w(T_n + u/\sqrt{n} \mid X^n) \frac{du}{\sqrt{n}} \\
&= \int_{\mathbf{R}^d} u\; \pi(u \mid X^n) du \qquad (5.21)
\end{aligned}$$

which tends to zero in P_{θ_o} probability since, under our hypotheses the posterior expectations converge to those of the normal. Now, writing θ_i for the i^{th} component of θ_o we have

$$\begin{aligned}
&\int_{\mathbf{R}^d} t_i t_j \mid \frac{w(T_n + t/\sqrt{n} \mid X^n)}{n^{d/2}} - \phi_{I(\theta_o)^{-1}}(t) \mid dt \\
&= \int_{\Omega} (\sqrt{n}\,(\theta_i - \hat{\theta}_i) + \sqrt{n}\,(\hat{\theta}_i - T_{n,i}))(\sqrt{n}\,(\theta_j - \hat{\theta}_j) + \sqrt{n}\,(\hat{\theta}_j - T_{n,j})) \\
&\qquad \times \mid w(\theta \mid X^n) - \phi_{I(\theta_o)}(\sqrt{n}\,(\theta - \hat{\theta}) + \sqrt{n}\,(\hat{\theta} - T_n) \mid d\theta \qquad (5.22)
\end{aligned}$$

in which we use (5.21). This completes the proof of Theorem 3. □

To conclude this section we prove Theorem 4. This amounts to noting that when moments match, a Kullback-Leibler number can be written as a difference of conditional entropies.

Proof of theorem 4: Let $Z = Z_{\theta \mid X^n}$ denote a random variable for which the conditional distribution of Z given X^n is normal with mean $E(\Theta \mid X^n)$ and variance matrix $\mathrm{cov}(\Theta \mid X^n)$. Such a random variable can be defined by Bayes rule: use m_n as the marginal for X^n and choose the conditional density for θ to be $N(E(\Theta \mid X^n), \mathrm{cov}(\Theta \mid X^n))$. By the definition of the mutual information

$$I(\Theta; X^n) = H(w) - H(Z \mid X^n) + [\, H(Z \mid X^n) - H(w \mid X^n)]$$

$$= H(w) - \frac{1}{2} E_M \log (2\pi e)^d \det \operatorname{cov}(\Theta \mid X^n)$$

$$+ E_M D(w(\cdot \mid X^n) \mid\mid N(E(\Theta \mid X^n), \operatorname{cov}(\Theta \mid X^n))),$$

since $(Z \mid X^n)$ and $(\Theta \mid X^n)$ have the same first two moments. By rearranging the expression we find that

$$E_M D(w(\cdot \mid X^n) \mid\mid N(E_{w(\cdot \mid X^n)}\Theta, \operatorname{cov}_{w(\cdot \mid X^n)}\Theta))$$

$$= I(\theta; X^n) - H(w) - \frac{d}{2} \log \frac{n}{2\pi e} - \frac{1}{2} \int_\Omega \log \det I(\theta)\, w(\theta)\, d\theta$$

$$+ \frac{1}{2} \left(E_M \log (2\pi e)^d \det \operatorname{cov}(\Theta \mid X^n) - \int_\Omega \log \det I(\theta)^{-1}\, w(\theta)\, d\theta \right). \tag{5.23}$$

From (5.23) the conclusions of the Proposition follow.□

The identity (5.23) shows that the convergence of the posterior to the normal in expected Kullback-Leibler distance is equivalent to the validity of the asymptotic expansion for the mutual information. The two terms on the left in expression (5.23) represent the upper and lower bounds respectively.

6. Applications. In this section we give two applications of the theorems we have proved. The first is for parametric density estimation we show that the quantity we have examined lower bounds the risk in parametric estimation. In the second we determine how much influence the choice of prior can have on the estimator asymptotically.

In the parametric density estimation context, suppose we are given a parametric family indexed by θ and that θ_o is the true value of the parameter. However, suppose that it is not the parameter 'per se' that interests us. Rather, we are using the parametric family so as to identify the true density p_{θ_o}. One natural estimator of $p(x \mid \theta_o)$ at any given x is the predictive density $\hat{p}_n(\cdot)$, which is the posterior mean of $p(x \mid \Theta)$.

We use the Kullback-Leibler number as the loss function for parametric density estimation and examine the behavior of the cumulative risk. Let δ_k for $k = 0,\ldots,n-1$ be a sequence of density estimators. Each δ_k estimates the density of X_{k+1}, given the data X^k. Here, δ_o is a fixed density function not dependent on the data. When θ_o is true, the risk associated with $\delta_k = \delta_k(X^k))$ is

$$E_{\theta_o} D(P_{\theta_o} \mid\mid \delta_k),$$

and we denote the cumulative risk of n uses of an estimator δ_k for $k = 0,...,n-1$ by $C(n, \theta_o, \delta)$. It is the sum of the individual risks:

$$C(n, \theta_o, \delta) = \sum_{k=0}^{n-1} E_{\theta_o} D(P_{\theta_o} \,||\, \delta_k).$$

The sum of the Kullback-Leibler risks plays an important role in universal coding theory, sequential estimation, hypothesis testing and portfolio selection theory, see Clarke and Barron (1990a).

Just as the posterior mean of Θ is the Bayes estimator under squared error loss it turns out that the posterior mean of $p(x \mid \Theta)$ is the Bayes estimator under relative entropy loss, see also Aitchison (1975). Indeed, we have the following.

Proposition 6.1: *$\hat{p}_n$ is the Bayes estimator of the density function. The cumulative risk of this estimator is*

$$C(n, \theta, \hat{p}_n) = \sum_{k=0}^{n-1} E_{\theta_o} D(p_{\theta_o} \,||\, \hat{p}_k) = D(P^n_{\theta_o} \,||\, M_n),$$

under the convention that $\hat{p}_0(x) = m_1(x_1)$. Its cumulative Bayes risk is

$$\int w(\theta) D(P^n_\theta \,||\, M_n)\, d\theta$$

and if the parameter space is compact the minimax risk is realized by choosing w to be the Jeffreys prior which is least favorable. Consequently, under the conditions of Theorems 1 and 2, the cumulative risks are asymptotically approximated by $(d/2)\log n + c$, and the average risk $(1/n)\sum E_{\theta_o} D(p_{\theta_o} \,||\, \hat{p}_k)$ and $(1/n)\sum E_{\theta_o} D(p_{\theta_o} \,||\, \hat{p}_k)$ converge to zero at rate $O(\log n)/n$.

Proof: The characterization of the cumulative risk is as in Clarke and Barron (1990a). The conclusions then follow by Theorems 1 and 2.□

We remark that under the conditions of Theorem 1 the individual risk terms $E_{\theta_o} D(P_{\theta_o} \,||\, \hat{P}_n)$ also converge to zero as $n \to 0$. This follows from noting that

$$E_{\theta_o} D(P_{\theta_o} \,||\, \hat{P}_n) = D(P^n_{\theta_o} \,||\, M_n) - D(P^{n-1}_{\theta_o} \,||\, M_{n-1}),$$

and applying Theorem 1 to each term on the right hand side. Thus, the predictive density is consistent for the true density in expected Kullback- Leibler distance. In a similar fashion we have that

$$\int_\Omega E_\theta D(P_\theta \,||\, \hat{P}_k) w(\theta) d\theta = o(1).$$

We next note that our results for Bayes parametric density estimation have implications for Bayes parameter estimation. This is so because parameter estimation can be regarded as a

special case of density estimation in which we restrict the estimator of the density to be of the form $p(x \mid \theta(X^n))$. In the present context we have used the parametric family as a tool to generate an estimator, relinquishing information from the family about what the true value of the parameter is. By enlarging the class of estimators we see that in terms of global optimality properties, the Bayes risk in parametric density estimation lower-bounds the Bayes risk in parametric estimation:

$$\inf_{\delta} E_w E_\theta D(\theta \mid\mid \delta) \geq \inf_{Q} E_w E_\theta D(P_\theta \mid\mid Q).$$

Similarly, for the maximin risk we have

$$\sup_{w} \inf_{\delta} \int w(\theta) E_\theta D(\theta \mid\mid \delta) d\theta \geq \sup_{w} \inf_{Q} \int w(\theta) E_\theta D(P_\theta \mid\mid Q) d\theta,$$

and for the minimax risk we have

$$\inf_{\delta} \sup_{\theta} E_\theta D(\theta \mid\mid \delta) \geq \inf_{Q} \sup_{\theta} E_\theta D(P_\theta \mid\mid Q),$$

where δ is an estimator of the parameter, Q is an estimator of the density and $D(\theta \mid\mid \delta) = D(P_\theta \mid\mid P_\delta)$ is the relative entropy loss for parameter estimation. The quantity in Theorem 1 gives a lower bound on the minimax and maximin risk of parameter estimation; the quantity in Theorem 2 gives an asymptotic lower bound on the Bayes risk of parameter estimation.

Finally, we consider some other convergences which were studied by McCulloch (1986). In particular we will see that the difference between two predictive densities, with respect to different priors, converges to zero. First suppose that the true distribution is M_n, a mixture of independent and identical distributions and that we estimate by another mixture, N_n based on the prior v which has the same support as w. If that support is compact then the Kullback-Leibler distance between the true distribution and the estimator is

$$\begin{aligned} D(M_n \mid\mid N_n) &= \int_K D(P_\theta \mid\mid N_n) w(\theta)\, d\theta - \int_K D(P_\theta \mid\mid M_n) w(\theta)\, d\theta \qquad (6.1) \\ &= D(w \mid\mid v) + o(1), \end{aligned}$$

as n increases by applying Theorem 1 to each term. If the predictive distribution based on v is denoted by $\hat{Q}_k$ then by direct calculation we have that

$$E_M D(\hat{P}_k \mid\mid \hat{Q}_k) = D(M_{k+1} \mid\mid N_{k+1}) - D(M_k \mid\mid N_k). \qquad (6.2)$$

So, as $k \to \infty$ we see that $E_M D(\hat{P}_k \mid\mid \hat{Q}_k)$ tends to zero, which means that except for θ in a set of arbitrarily small prior measure, we have that $E_\theta D(\hat{P}_k \mid\mid \hat{Q}_k)$ tends to zero in P_θ probability. In this sense predictive densities based on different priors are asymptotically

indistinguishable.

We obtain similar behavior for the posteriors:

$$E_M D(w(\cdot \mid X^n) \mid\mid v(\cdot \mid X^n)) = D(w \mid\mid v) - D(M_n \mid\mid N_n) = o(1),$$

so, we have that

$$E_\theta D(w(\cdot \mid X^n) \mid\mid v(\cdot \mid X^n)) \to 0,$$

in the joint probability for X^n and θ. Also, we have that

$$D(w(\cdot \mid X^n) \mid\mid v(\cdot \mid X^n)) \to 0,$$

in the joint probability for X^n and θ. From the recursion relation (6.2) we see that

$$D(w \mid\mid v) = \sum_{l=0}^{\infty} E_M D(\hat{P}_l \mid\mid \hat{Q}_l),$$

where, under our convention $E_M D(\hat{P}_0 \mid\mid \hat{Q}_0) = D(M_1 \mid\mid N_1)$, so the number of times $E_M D(\hat{P}_n \mid\mid \hat{Q}_n)$ exceeds $1/n$ must have negligible cumulative effect.

The formula we have proved in Theorem 1 assumes that P_{θ_o} is the true density. If the mixture is the true density then estimating with an element of the parametric family is a poor strategy. We see that if the prior v is unitmass at a point θ in the support of w, then (6.1) shows that

$$D(M_n \mid\mid P_{\theta_o}^n) = n \int D(\theta \mid\mid \theta_o) w(\theta) d\theta - I(\Theta; X^n),$$

that is, the risk increases at rate n no matter what estimator we use, since we know the second term on the right hand side behaves like $(d/2) \log n$.

REFERENCES

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika* **62** 547-554.

Berger, J. O. and Bernardo, J. M. (1989a). Estimating a product of means: Bayesian analysis with reference priors. *J. Amer. Statist. Assoc.* **84** 200-207.

Berger, J. O. and Bernardo, J. M. (1989b). Ordered group reference priors with applications to multinomial and variance components problems. Technical Report, Purdue University, Department of Statistics.

Berger, J. O. and Bernardo, J. M. (1989c). Reference priors in a variance components problem. Technical report, Purdue University, Department of Statistics.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. Ser. B* **41** 113-147.

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

Billingsley, P. (1986). *Probability and Measure*. Wiley, New York.

Chow, Y. S. and Teicher, H. (1978). *Probability Theory, Independence, interchangeability* and Martingales. Springer-Verlag, New York.

Clarke, B. (1989). Asymptotic cumulative risk and Bayes risk under entropy loss, with applications. Ph. D. dissertation, University of Illinois.

Clarke, B. and Barron, A. (1989). Information theoretic asymptotics of Bayes methods. Technical Report #26, Department of Statistics, University of Illinois.

Clarke, B. and Barron, A. (1990a). Information theoretic asymptotics of Bayes methods. To appear in *IEEE Trans. Inform. Theory*.

Clarke, B. and Barron, A. (1990b). On redundancy and capacity. In preparation for *IEEE Trans. Inform. Theory*.

Csiszar, I. (1967). Information-type measures of difference of probability distributions and individual observations. *Studia Sci. Math. Hungar.* **2** 299-318.

Davisson, L. (1973). Universal noiseless coding. *IEEE Trans. Inform. Theory* **19** 783-795.

Gray, R. M. (1988). *Probability, Random Processes, and Ergodic Properties*. Springer-Verlag, New York.

Haughton, D. M. A. (1988). On the choice of a model to fit data from an exponential family. *Ann. Statist.* **16** 342-355.

Ibragimov, I. A. and Hasminsky, R. Z. (1973). On the information in a sample about a parameter. *Second International Symposium on Information Theory* 295-309 Akademiai, Kiado, Budapest.

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, New York.

Krichevsky, R. E. and Trofimov, V. K. (1981). The performance of universal encoding.

IEEE Trans. Inform. Theory **27** 199-207.

Lehmann, E. L. (1983). Theory of Point Estimation. *Wiley*, New York.

Leonard, T. (1982). Comment on "A simple predictive density function. *J. Amer. Statist.* Assoc. **77** 657-658.

McCulloch, R. E. (1986). Information asymptotics and inequalities for posterior and predictive distributions. Submitted to *Can. J. Statist.*

Rissanen, J. (1986). Stochastic complexity and modeling. *Ann. Statist.* **14** 1080-1100.

Rissanen, J. (1987). Stochastic complexity. *J. Roy. Statist. Soc. Ser. B* **49** 223-239. *IEEE Trans. Inform. Theory* **30** 629-636.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461-464.

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