Entropy Risk and the Bayesian Central Limit Theorem

(Abbreviated Title: Entropy Risk and the Bayesian CLT)

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

Purdue University and The University of Illinois

Abstract

In a smooth finite dimensional parametric family of densities equipped with a prior, the expected Kullback-Leibler distance between the standardized posterior and the *Normal* (0, 1) is determined by the Shannon mutual information between the parameter and the data. This mutual information, $I(\Theta; X^n)$, can be recognized as the cumulative Bayes risk of the sequence of Bayes estimators under an entropy loss criterion. We give an asymptotic expansion for $I(\Theta; X^n)$. As a result, it is seen that the asymptotics of the cumulative Bayes risk are equivalent to a strengthened Bayesian central limit theorem. Consequences are given for parameter estimation and investment theory.

Submitted to the *Annals of Statistics*, September 1990. Withdrawn 1991

[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, Kullback-Leibler information, Fisher information, Shannon's mutual information, parametric density estimation, data compression.

1. Introduction. We examine large sample 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. The convergence of this 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.

Assume we have a parametric family $\{p_\theta: \theta \in \Omega\}$, $\Omega \subset \mathbf{R}^d$, of probability density functions $p_\theta(x) = p(x \mid \theta)$ with respect to a fixed dominating measure $\lambda(dx)$ on a measurable space X, and we have a prior distribution for Θ that has a probability density function $w(\theta)$ with respect to d-dimensional Lebesgue measure. Given θ, random variables $X_1, \ldots, X_n$ are assumed to be conditionally independent with density function $p_\theta^n(x^n) = \prod_{i=1}^n p_\theta(x_i)$, for $x^n \in \mathrm{X}^n$. Averaging out θ gives the marginal density function $m_n^w(x^n) = \int w(\theta) p_\theta(x^n)\, d\theta$ associated with the joint density $w(\theta)p_\theta^n(x^n)$ on $\Omega \times \mathrm{X}^n$.

We denote the relative entropy or Kullback-Leibler distance by $D(p \mid\mid q) = \int p \log p/q$, where it is assumed that p and q are probability densities with respect to the same dominating measure. The quantity of interest to us is the average relative entropy distance between p_θ^n and m_n^w, which we denote by

$$R_n(w) = \int_\Omega D(p_\theta^n \mid\mid m_n^w)\, w(\theta) d\theta. \tag{1.1}$$

This is the Bayes risk associated with the decision theory problem in which nature chooses the density p_θ^n and the statistician chooses a density q_n. In this case, the Bayes strategy is to choose the density $q_n = m_n^w$ because it achieves the minimal average loss $R_n(w) = \min_{q_n} \int D(p_\theta^n \mid\mid q_n)\, w(\theta) d\theta$, see Aitchison (1975). As is shown in Clarke and Barron (1990), there is also an interpretation of $R_n(w)$ as the cumulative Bayes risk of the sequence of Bayes estimators $\hat{p}_k(X_k) = m(X_k \mid X^{k-1})$ of the density function $p_\theta(X_k)$ based on the data $X_1,\ldots,X_{k-1}$ for $k = 1,2,\ldots,n$. Thus $R_n(w)/n = (1/n)\sum_{k=1}^n E_{\Theta,X^k} D(p_\theta \mid\mid \hat{p}_k)$ assesses the accuracy of Bayes estimators in a predictive context.

It is seen that this Bayes risk is also Shannon's mutual information between the parameter θ and the sample $X_1,\ldots,X_n$. 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 primary purpose of this paper is to derive an asymptotic expression for the Bayes risk $I(\Theta; X^n)$ that leads to an information-theoretic Bayesian central limit theorem. The asymptotics for $I(\Theta; X^n)$ also has direct implications for several applications in statistics and information theory.

In Theorem 2.1, we give, under suitable conditions, an asymptotic formula for the Bayes risk, expression (1.1). The asymptotic formula we obtain is

$$R_n(w) = \frac{d}{2}\log \frac{n}{2\pi e} + \frac{1}{2}\int_{\Omega} w(\theta) \log \det I(\theta)\, d\theta + H(w) + o(1), \tag{1.2}$$

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 $n \to \infty$. Thus $R_n(w)/n$ converges to zero at rate $((d/2)(\log\, n) + c + o(1))/n$ and the constant c is identified.

Theorem 2.2 is a new Bayesian central limit theorem that shows the posterior distribution is asymptotically normal in expected Kullback-Leibler distance, to wit,

$$E_m\, D(w(\cdot \mid X^n) \mid\mid \phi_n) \to 0, \tag{1.3}$$

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. Equivalently, the posterior distribution for the standardized parameter $T = \mathrm{COV}(\Theta \mid X^n)^{-1/2}\,(\Theta - E(\Theta \mid X^n))$ converges to the standard normal distribution in expected Kullback-Leibler distance. Asymptotic normality of the posterior in the L_1 sense is a well known classical result due to LeCam; see LeCam (1958, 1986), and most recently LeCam and Yang (1990). We prove an extension of the classical result for use in proving (1.3).

Theorem 2.2 demonstrates that, under reasonable conditions, the information-theoretic convergence in the Bayes CLT (1.3) is equivalent to the validity of the asymptotic expansion (1.2) for the Bayes risk $I(\Theta; X^n)$.

It is our goal that (1.1) and its associated expansion (1.2) be of interest to statisticians who concern themselves with Bayesian estimation and Bayesian central limit theory; as well as information theorists who concern themselves with universal data compression and channel capacity. The implications for the latter two topics will be discussed in detail elsewhere. Next we discuss how our work relates to some statistical literature.

First we note that the mutual information $I(\Theta; X^n)$ can be interpreted as the expected logarithm of the Bayes factor between a Bayesian who observes (Θ, X^n) and a Bayesian who observes only X^n. The logarithm of the Bayes factor occurs in model selection problems involving a Bayes criterion. Schwarz (1978), Leonard (1982) and Haughton (1988) have developed expansions similar to (1.2) for model selection problems. It is seen that

$(d/2)\log n + c$ is the penalty which must be paid for lack of knowledge of the parameters. A higher penalty is paid for higher dimensional families.

For a portfolio selection problem, X_i is interpreted as the vector of stock market returns for days $i = 1,...,n$ and Θ is a random variable representing side information that determines the distribution of the returns. It is shown in Barron and Cover (1988) that $I(\Theta; X^n)$ bounds the average difference in the logarithm of wealth gained over n days between an investor who knows Θ and invests optimally for the distribution P_θ and an investor who does not know Θ and invests instead according to the Bayes optimal strategy which is M_n. In this context, the asymptotics for $I(\Theta; X^n)$ shows that knowledge of Θ contributes only a polynomial growth factor to the wealth, which is already growing at an exponential rate.

Bernardo (1979) conjectured the form of the asymptotic expression for $I(\Theta; X^n)$ in order to identify the prior which maximized it. This is seen to be Jeffreys' prior, Jeffreys (1962), which is proportional to the square root of the determinant of the Fisher information matrix.

Ibragimov and Hasminskii (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 Hasminskii require (in Condition A.III) that the Fisher information be bounded and bounded away from zero. The approach developed for Theorem 2.1 below avoids these restrictions.

In the information theory context of universal data compression the quantities $R_n(\theta,w) = D(p_\theta^n \,||\, m_n)$ and $R_n(w) = \int w(\theta)D(p_\theta^n \,||\, m_n)\, d\theta$ can be interpreted as the redundancy and average 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 (1990) in an information theory setting are here extended to give the asymptotics for the average redundancies $R_n(w)$.

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 in this context will be examined in later work.

To obtain (1.2) we prove upper and lower bounds which are asymptotically identical. These two bounds require different techniques.

For the upper bound we set up an application of the dominated convergence theorem. To obtain the pointwise behavior we improve an earlier asymptotic formula due to Clarke and Barron (1990) by obtaining it under weaker hypotheses. The domination requires that we deal with points on or close to the boundary of the support of the prior. Our technique is to ensure that integrals about such points can be bounded by integrals about points within the interior.

For the lower bound we use a maximum entropy argument and assume that

$$n \ \mathrm{COV}(\Theta \mid X^n) \to I(\theta)^{-1} \tag{1.4}$$

in P_θ probability, for each θ in Ω. In a different result we give conditions which ensure (1.4) and are readily verifiable for many examples. It is apparent that (1.4) is an extension of the asymptotic normality of the posterior in the L_1 sense.

The outline for the remainder of this paper is as follows. In Section 2 we define our notation and state the main results. Then we show that the hypotheses of Theorem 2.1 are satisfied in two examples: the normal with a normal prior, in which is easy to evaluate $I(\Theta; X^n)$ directly, and the *Poisson*(θ) for $\theta \geq 1$ with any member of a class of priors, in which is difficult to evaluate $I(\Theta; X^n)$ without recourse to an asymptotic approximation. The proofs of the main results are subsequently given in sections 3, 4, and 5. Finally, in Section 6 we give some implications for parametric density estimation, and an application to stock market portfolio selection.

2. Conditions and Main Results. We adopt the notation that E means expectation with respect to p_θ unless denoted otherwise, 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 of Θ and X^n. A technicality is that X is a separable metric space, so that the set of probability measures on X is endowed with the topology of weak convergence. Also, we assume that the parameter space Ω has a nonvoid interior, and its boundary has d–dimensional Lebesgue measure zero. The prior probability density $w(\theta)$ is assumed to be given for θ in Ω and extended to R^d by setting it to be zero outside of Ω, it is assumed to be continuous except for θ in a set of measure zero, and we let Ω_w denote the set of points of positivity and continuity of w.

So as to facilitate the statements of hypotheses we give two conditions to which it will be convenient to refer. Let θ_o be a point in the interior of Ω_w.

Condition 1: *The parametric family is sound at θ_o, in the sense that the convergence of parameter values (in the Euclidean sense) is equivalent to the weak convergence of the distributions they index. That is,*

$$\theta \to \theta_o \text{ if and only if } P_\theta \to P_{\theta_o}.$$

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

Condition 2: *$p_\theta^{1/2}$ is differentiable in $L^2(\lambda)$ at $\theta = \theta_o$; the set $\{x : \nabla p_{\theta_o}^{1/2}(x) \neq 0, p_{\theta_o}^{1/2}(x) = 0\}$ has λ measure zero, and the Fisher information matrix,*

$$I(\theta_o) = E_{\theta_o} S(X)\, S(X)^T,$$

is positive definite, where

$$S(X) = 2(\nabla p_{\theta_o}^{1/2}(X))/p_{\theta_o}^{1/2}(X)$$

is the score function.

For a discussion of the soundness condition, see Clarke and Barron (1990a). Soundness is used there, in Proposition 6.2, to obtain the existence of a uniformly exponentially consistent test of the hypothesis $\theta = \theta_o$ *versus* $\{\theta: \mid \theta - \theta_o \mid > \delta\}$, for δ positive.

We note that differentiability in $L^2(\lambda)$ of $p_\theta^{1/2}$ is generally regarded as the natural smoothness assumption so as to invoke the theory of locally asymptotically normal experiments, see Ibragimov and Hasminskii (1981, Chapter 2) and LeCam (1986, Chapter 17). The requirement that $\lambda(\{\nabla p_{\theta_o}^{1/2} \neq 0, p_{\theta_o}^{1/2} = 0\}) = 0$ is a technicality which implies that

$$P_\theta(\{x: p_\theta(x) > 0, p_{\theta_o}(x) = 0\}) = o(\mid \theta - \theta_o \mid^2),$$

as $\theta \to \theta_o$, as required for the Hajek-LeCam theory. Furthermore, it ensures that $(\nabla p_{\theta_o}^{1/2}(x))$, the mean square derivative of $p_\theta^{1/2}$ in $L_2(\lambda)$, and $(\nabla p_{\theta_o}^{1/2}(x)/(p_{\theta_o}^{1/2}(x)))$, the mean square derivative of $p_\theta^{1/2}/p_{\theta_o}^{1/2}$ in $L_2(P_{\theta_o})$, yield the same value for the Fisher information, i.e., the two expressions $4\int (\nabla p_{\theta_o}^{1/2}(x))(\nabla p_{\theta_o}^{1/2}(x))^T\, \lambda(dx)$ and $4\int_{\{p_{\theta_o}(x) > 0\}} (\nabla p_{\theta_o}^{1/2}(x))(\nabla p_{\theta_o}^{1/2}(x))^T\, \lambda(dx)$ are identical.

If $p_\theta(x)$ is pointwise differentiable in θ with gradient $\nabla p_\theta(x)$ then the score function is more directly expressed as

$$S(X) = (\nabla p_{\theta_o}(X))/p_{\theta_o}(X) = \nabla \log p_{\theta_o}(X).$$

We denote the posterior density for θ given data $X^n = (X_1,\ldots,X_n)$ by

$$w_n(\theta) = w(\theta \mid X^n) = \frac{w(\theta) p_\theta(X^n)}{m(X^n)},$$

and the posterior mean by $\theta^* = E(\theta \mid X^n)$. Also, we will have occasion to use the pseudo-estimator

$$\theta' = \theta_o + (nI(\theta_o))^{-1} S_n$$
$$= \theta_o + (nI(\theta_o))^{-1/2} Z_n,$$

where $S_n = \sum_{i=1}^{n} S(X_i)$ is the total score and $Z_n = (nI(\theta_o))^{-1/2} S_n$ is the standardized score. Note that by the central limit theorem Z_n converges in distribution to a standard normal on $\mathbf{R}^d$ and $| Z_n |^2$ converges in distribution to a Chi-square with d degrees of freedom, when $X_1,...,X_n$ are i.i.d. p_{θ_o}.

Let $\phi_{\mu,\Sigma}(\theta)$ denote the normal density on $\mathbf{R}^d$ with mean μ and covariance Σ. For vectors θ in $\mathbf{R}^d$, we let $| \theta |_M$ denote the norm defined by the positive definite matrix M. When the Euclidean norm is meant we omit the subscript matrix. We choose to use the alternate norms because they arise in the second order Taylor expansion of $D(\theta \mid\mid \theta') = D(P_\theta \mid\mid P_{\theta'})$, which must be controlled in order to prove (1.2). For densities on $\mathbf{R}^d$ we let $| w - v | = \int | w(\theta) - v(\theta) | \, d\theta$ denote the L^1 distance.

In our proof of the lower bound part of Theorem 2.1, we use the fact that the Bayes risk $R_n(w)$, is the Shannon mutual information between the parameter and the data, which we denote by $I(\Theta;X^n)$. To derive the asymptotic lower bound for $I(\Theta;X^n)$ we make assumptions on the posterior mean $\theta^* = E(\Theta \mid X^n)$ and the posterior covariance $\mathrm{COV}(\Theta \mid X^n)$. Our assumptions on the posterior covariance hold under conditions given in Proposition 2.1.

Proposition 2.1 states asymptotic normality of the posterior in an L_1 mode of convergence and yields $n\ \mathrm{COV}(\Theta \mid X^n) \to I(\theta)^{-1}$ in P_θ probability, as is required in Theorem 2.1, expression (2.2). Similar results are in LeCam (1958, 1986, Theorem 17.7.1, p. 619), LeCam and Yang (1990), Walker (1967), Bickel and Yahav (1969), Ibragimov and Hasminskii (1981), Hartigan (1983), and Lehmann (1983).

Proposition 2.1. Asymptotics associated with the posterior distribution.

Part A. Assume that Conditions 1 and 2 are satisfied by the family of densities at a point $\theta_o \in \Omega_w$. When $X_1,X_2,...$ are independent with distribution P_{θ_o}, we have the following conclusions.

(1) *Asymptotics of the likelihood ratio:*

$$\lim_{n \to \infty} \frac{m(X^n)}{p_{\theta_o}(X^n)} n^{d/2} e^{-|Z_n|^2/2} = \frac{w(\theta_o)(2\pi)^{d/2}}{\det I(\theta_o)^{1/2}},$$

in P_{θ_o} probability.

(2) *Asymptotic normality of the posterior in* L_1:

$$\lim_{n \to \infty} | w_n - \phi_{\theta', (nI(\theta_o))^{-1}} | = 0,$$

in P_θ *probability.*

(3) *Asymptotics of posterior moments: For* $r > 0$,

$$\int_\Omega | \theta |^r \, w(\theta) \, d\theta$$

implies

$$\lim_{n \to \infty} \int_\Omega | \sqrt{n} \, (\theta - \theta') |^r \, | w(\theta | X^n) - \phi_{\theta', (nI(\theta_o))^{-1}}(\theta) | \, d\theta = 0,$$

in P_{θ_o} *probability. Consequently, for* $r = 1$, *we obtain* $\sqrt{n} \, (\theta^* - \theta') \to 0$, *in probability, and* $\sqrt{n} \, (E(\Theta | X^n) - \theta_o) \to N(0, I(\theta_o)^{-1})$ *in distribution. For* $r = 2$, *we obtain*

$$n \, \mathrm{COV}(\Theta | X^n) \to I(\theta_o)^{-1}.$$

Part B. Assume that Conditions 1 and 2 are satisfied by the family of densities at every point in Ω_w, *and that* $\int_\Omega | \theta | \, w(\theta) \, d\theta$ *is finite. For* $X_1, X_2, \ldots$ *governed by the Bayesian distribution* M_n, *we have the following conclusion. Here the normal approximation is centered at the posterior mean, and scaled with the posterior variance.*

(4) *Bayesian central limit theorem with convergence in* L_1:

$$\lim_{n \to \infty} | w_n - \phi_{E(\Theta | X^n), COV(\Theta | X^n)} | = 0.$$

The key step in the proof of the upper bound part of Theorem 2.1 requires that we identify an upper bound for the pointwise behavior in probability of the quantity $D(P_{\theta_o}^n \,||\, M_n)$. This is done in Proposition 2.2. It is similar Theorem 2.1 in Clarke and Barron (1990), but here it is obtained under the Conditions stated above which are weaker than those used in the earlier result.

Proposition 2.2. *Assume that conditions 1, and 2 are satisfied by the family of densities at a point* $\theta_o \in \Omega_w$. *Assume also that*

$$D(P_{\theta_o} \,||\, P_\theta) = \frac{1}{2} | \theta - \theta_o |^2_{I(\theta_o)} + o(| \theta - \theta_o |^2) \tag{2.1}$$

as $\theta \to \theta_o$.

Then,

$$\log \frac{p_{\theta_o}(X^n)}{m(X^n)} - \frac{d}{2} \log n + \frac{| Z_n |^2}{2} + \log [\, w(\theta_o)(2\pi)^{d/2} \det I(\theta_o)^{-1/2} \,] \leq o(1) \tag{2.2}$$

where o(1) tends to zero in $L_1(P_{\theta_o})$. *Consequently,*

$$\limsup_{n \to \infty} \left(D(P_{\theta_o}^n \,||\, M_n) - \frac{d}{2} \log n \right) \leq \log \left[w(\theta_o)(2\pi e)^{d/2} \det I(\theta_o)^{-1/2} \right]. \tag{2.3}$$

Now we state formally a set of conditions under which the asymptotic formula from (1.2) holds.

Theorem 2.1: a)*Assume that the Bayes risk for the estimation of* θ *under squared error loss, is of order* $O(1/n)$, *that is,*

$$\limsup_{n \to \infty} n \, E_{\Theta, X^n} | \theta - \theta^* |^2 < \infty, \tag{2.4}$$

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

$$n \, \mathrm{COV}(\Theta | X^n) \to I^{-1}(\theta), \tag{2.5}$$

in P_θ *probability. Then, we have the lower bound*

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

$$\geq \frac{d}{2} \log \frac{1}{2\pi e} + \frac{1}{2} \int_\Omega w(\theta) \log \det I(\theta) \, d\theta + H(w). \tag{2.6}$$

Consequently, the limit inferior of the minimax value, $R_n = \inf_{q_n} \sup_\theta D(p_\theta^n \,||\, q_n)$, *satisfies the bound,*

$$\liminf_{n \to \infty} \left[R_n - \frac{d}{2} \log n \right] \geq \frac{d}{2} \log \frac{1}{2\pi e} + \log \int_\Omega \sqrt{\det I(\theta)} \, d\theta \, . \tag{2.7}$$

b)*Assume the hypotheses of Proposition 2.2 hold. Suppose that there is an* $\varepsilon > 0$ *such that, for each* θ *in* Ω_w *there is a matrix* $I_\varepsilon(\theta)$ *which satisfies*

$$D(\theta \,||\, \theta') \leq \frac{1}{2} (\theta - \theta')^T I_\varepsilon(\theta) (\theta - \theta'), \tag{2.8}$$

on the set $\{\theta' \in \Omega_w : D(\theta \,||\, \theta') \leq 2\varepsilon\}$, *so that*

$$\int | \log \det I_\varepsilon(\theta) | \, w(\theta) \, d\theta < \infty. \tag{2.9}$$

Assume also that for each θ *in* Ω_w *there is a* θ'' *in* Ω_w *with* $| \theta - \theta'' |^2_{I_\varepsilon(\theta)} \leq \varepsilon$ *so that*

$$\{ \theta' : | \theta' - \theta'' |^2_{I_\varepsilon(\theta)} < \varepsilon \} \subset \mathrm{Interior}(\Omega_w). \tag{2.10}$$

Finally, assume that the prior w *is locally lower bounded in the sense that*

$$\int w(\theta) \log \frac{1}{\inf_{|\theta' - \theta''|^2_{I(\theta)} < \varepsilon} w(\theta')} \, d\theta < \infty. \tag{2.11}$$

Then, we have the upper bound

$$\limsup_{n \to \infty} \left[I(\Theta; X^n) - \frac{d}{2} \log n \right] \leq \frac{d}{2} \log \frac{1}{2\pi e} + \frac{1}{2} \int_{\Omega} \log \det I(\theta) \, d\theta + H(w). \tag{2.12}$$

Together, (2.6) and (2.12) determine the limit of $I(\Theta; X^n) - \frac{d}{2} \log n$. This verifies the expansion (1.2) of the Bayes risk.

The two restrictive hypotheses for the upper bound admit the following interpretation. The first, expressions (2.8) and (2.9), means that $D(\theta \,||\, \theta')$ is locally upper bounded by its second order Taylor expansion. The second, (2.10), is a generalization of convexity. It requires only that, in the position dependent norm defined by $I_\varepsilon(\theta)$, each point θ is contained in a neighborhood of radius $\sqrt{\varepsilon}$ which is in the interior of Ω_w. This condition enables us to get bounds near boundary points. For points away from the boundary, we can take θ'' to be θ itself. The hypothesis (2.11) is only slightly stronger than the finiteness of the prior entropy. In the one dimensional setting, cases where (2.11) is strictly stronger than finite entropy include those where the boundary of the support of the prior has infinitely many cluster points in a bounded set.

We note that if there is any estimator which has Bayes risk of order $O(1/n)$ then the Bayes estimator also has Bayes risk of order $O(1/n)$ since, by definition, the Bayes estimator has minimal Bayes risk. A standard approach for identifying $O(1/n)$ consistent estimators is based on the method of moments. In particular, let g and h be functions such that $\theta = h(E_{p_\theta} g(X))$. If the variance of $g(X)$ is integrable with respect to the prior (that is, if $E_{\Theta, X} \, | \, g(X) - E_{p_\theta} g(X) \, |^2$ is finite) and if h is a Lipschitz continuous function, then hypothesis (2.4) is satisfied by using the estimator $\hat{\theta} = h((1/n) \sum_{i=1}^n g(X_i))$.

The assumption (2.4), which is formulated in terms of a second moment, can be replaced with the weaker moment assumption that $E_{\Theta, X^n} (\sqrt{n} \, | \, \hat{\theta} - \theta \, |)^\alpha$ is bounded for some $\alpha > 0$, in the lower bound part of Theorem 2.1. However, in the end there is nothing to be gained from such a weakening. Indeed, the convergence of second moments of $\sqrt{n} \, (\Theta - \hat{\Theta})$ is necessary for the information-theoretic Bayesian central limit theorem in Theorem 2.2.

The other key hypothesis for the lower bound is (2.5). It is formulated as convergence in P_θ probability for each θ, since this is what is verified in Proposition 2.1, as a consequence of the asymptotic normality of the posterior. A Bayesian reformulation can be used in place of (2.5), requiring the convergence of $n \, \mathrm{COV}(\Theta \,|\, X^n) - I^{-1}(\Theta)$ to zero in probability with respect

to the joint distribution of Θ and X^n.

In Bernardo (1979) the minimax properties associated with relative entropy risk were considered. Our result gives the lower bound part, (2.7), rigorously which we state because the lower bound depends on the average with respect to the prior and so is easy. A formal verification for the upper bound part can be given but is a bit more difficult since it requires pointwise uniformity. This will be done in a subsequent paper.

Next we state an information-theoretic formulation of the Bayesian central limit theorem. Conditions are given such that the relative entropy distance between the posterior density and a normal density converges to zero. In the proof maximum entropy arguments with constraints on the first and second moment reveal the role of the normal distribution. In this way the connection between the Bayes risk problems formulated in terms of the information-theoretic framework and the more classical squared-error loss framework is revealed.

Under a subset of the conditions, the asymptotics for the mutual information $I(\Theta;X^n)$ and the relative entropy $D(w(\cdot|X^n)||\phi_{E(\Theta|X^n),\mathrm{COV}(\Theta|X^n)})$ are shown to be equivalent. It is not surprising that the asymptotic normality has implications for the mutual information $I(\Theta;X^n)$: indeed, the asymptotic normality is the basis of the proof of Ibragimov and Hasminskii (1973) and the conjecture of Bernardo (1979). It is the reverse implication that is somewhat surprising: Knowledge of the asymptotic Bayes risk $I(\Theta;X^n)$ for the estimation of the density of X^n given θ, determines the Gaussian shape of the posterior density of θ given X^n.

Theorem 2.2: *Assume that the conditions for the upper bound and lower bound in Theorem 2.1 are satisfied. Then, we have that*

$$\lim_{n\to\infty} E_{m_n} D\left(w(\cdot \mid X^n) \,||\, \phi_{E(\Theta|X^n),n\mathrm{COV}(\Theta|X^n)}\right) = 0. \tag{2.13}$$

If only (2.4) and the conditions of the upper bound are satisfied then the following are equivalent:

i) The limit in (2.13) exists and is zero;

ii) The following limit holds:

$$E_m \log\det n\,\mathrm{COV}(\Theta|X^n) \to \int w(\theta)\log\det I(\theta)^{-1}\,d\theta; \tag{2.14}$$

and

iii) The asymptotic expansion for the mutual information, (1.2) holds, i.e.,

$$I(\Theta;X^n) = \frac{d}{2}\log\frac{n}{2\pi e} + \frac{1}{2}\int w(\theta)\log\det I(\theta) + H(w) + o(1).$$

We illustrate Theorem 2.1 for two parametric families. The first is the *Normal*(θ,σ^2) with a *Normal*$(0,\sigma_p^2)$ prior where σ and σ_p are assumed known; the second is the *Poisson*(θ)

with any prior positive on $[1, \infty)$ and satisfying certain tail conditions.

In the first case, we know that (Θ, X^n) is jointly normally distributed, and $\bar{X}$ is sufficient. Thus,

$$I(\Theta; X^n) = I(\Theta; \bar{X}) = \frac{1}{2}\log\left(1 + \frac{n\sigma_p^2}{\sigma^2}\right).$$

This is asymptotically the same as the formula from Theorem 2.1 which we find is (1/2) log $n\sigma_p^2/\sigma^2$, since the Fisher information is $1/\sigma^2$, and the entropy of the *Normal* $(0, \sigma_p^2)$ is (1/2)log $2\pi e\,\sigma_p^2$.

The hypotheses of Theorem 2.1 are easily satisfied. Conditions 1 and 2 are obviously satisfied. Hypothesis (2.4) is satisfied by the mean and so by the Bayes estimator. So, the lower bound part of Theorem 2.1 applies. The upper bound part is similarly easy. We may choose $I_\varepsilon(\theta) = I(\theta) = 1$, so that (2.8) and (2.9) are obviously satisfied. There are no boundary points for the support of the normal prior so (2.10) is vacuous. Finally, (2.11) is satisfied for any positive ε.

The *Poisson* (θ) is a nontrivial example. The theorem due to Ibragimov and Hasminskii does not apply. This is so because the hypothesis that the Hellinger distance between two distributions does not go to zero when the Euclidean distance between their parameters is bounded away from zero is not satisfied. We see this by noting that Hellinger distance is dominated by the Kullback-Leibler distance, and for the Poisson, $D(\theta_n \,||\, \theta'_n)$ goes to zero for the choice of sequences $\theta_n = n$ and $\theta'_n = n + 1$ whose difference is always unity.

Our hypotheses are satisfied. It is easy to show $D(\theta \,||\, \theta') \le \theta(\theta - \theta')^2$ for $\theta \ge 1$ and $\theta' \ge 1/2$: both sides agree when $\theta' = \theta$, for $\theta' > \theta$ the derivative of the right side is greater than the derivative of the left, and for $\theta' < \theta$ the derivative of the right side is less than the derivative of the left. This means that hypotheses (2.8) is satisfied for any $I_\varepsilon(\theta)$ less than or equal to 2θ. Note that $I_\varepsilon(\theta)$ is necessarily greater than or equal to $1/\theta$, the Fisher information.

We assume that the prior is positive on $[1, \infty)$. The left hand endpoint is the only boundary point, and away from it (2.10) is satisfied. As θ tends to one, the point θ'', as a function of θ, may be chosen to be the right hand endpoint of the interval $\{\theta'': |\theta - \theta''|^2_{I_\varepsilon(\theta)} < \varepsilon\}$, so that (2.10) remains satisfied.

Now we impose two conditions on the prior. The first is that $w(\theta)$ be bounded away from zero at the boundary point. The other is that for θ large enough the prior density is dominated by a function which is $O(1/\theta^{1+\eta})$ for some positive η. The second condition ensures (2.9). Together the two conditions imply (2.11).

Now, from Theorem 2.1, we have that

$$I(\Theta; X^n) = \frac{1}{2} \log \frac{n}{2\pi e} - \frac{1}{2} \int_1^\infty w(\theta) \log \theta \, d\theta + H(w) + o(1),$$

which is difficult to derive any other way.

3. Asymptotics of $I(\Theta; X^n)$. We start with the lower bound. It is here that we use the maximum entropy argument. 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

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

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

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

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.

The lemma below 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 3.1: *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

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

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

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. □

Now we give a proof of part a) of Theorem 2.1.

Proof of Theorem 2.1, lower bound part: The Bayes risk $R_n(w) = \int_\Omega D(p_\theta^n \mid\mid m_n) w(\theta)\, d\theta$ 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

$$\begin{aligned} I(\Theta;X^n) &= H(\Theta) - H(\Theta \mid X^n) \\ &= H(\Theta) - \int_{X^n} H(\Theta \mid X^n = x^n) m(x^n) \lambda(dx^n) \\ &= H(\Theta) - \int_{X^n} H(\Theta - \theta^* \mid X^n = x^n) m(x^n) \lambda(dx^n) \\ &\geq H(\Theta) - \frac{1}{2} \int_{X^n} m(x^n) \log \left[(2\pi e)^d \det E_{w(\cdot \mid x^n)} n(\Theta - \theta^*)(\Theta - \theta^*)^t \right] \lambda(dx^n) \qquad (3.1) \\ &= H(\Theta) + \frac{d}{2} \log \frac{n}{2\pi e} \\ &\quad - \frac{1}{2} \int_{X^n} m(x^n) \log \det E_{w(\cdot \mid x^n)} \sqrt{n}(\Theta - \theta^*) \sqrt{n}(\Theta - \theta^*)^T \lambda(dx^n), \qquad (3.2) \end{aligned}$$

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

We will show that $\log \det n\,\mathrm{COV}(\Theta \mid X^n)$ 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:

$$\log \det \left[n\ \mathrm{COV}(\Theta \mid X^n) \right] \leq \sum_{i=1}^{d} \log \left[n\ Var(\Theta_i \mid X^n) \right] \qquad (3.3)$$

By assumption,

$$\sup_{n,i} E_m\, E_{\Theta \mid X^n}\, n(\Theta_i - \theta_i^*)^2 < \infty,$$

so, by Lemma 3.1 we have that each

$$\log E_{\Theta \mid X^n} n(\Theta_i - \theta_i^*)^2 \qquad (3.4)$$

is uniformly integrable from above. Thus, the right hand member of (3.3) is uniformly integrable from above. This implies that

$$\log \det \left[n\ \mathrm{COV}(\theta \mid X^n) \right] \qquad (3.5)$$

is uniformly integrable from above, and therefore so is

$$\log \det [\, n \ \mathrm{COV}(\theta \mid X^n) \,] + \log \det I(\theta). \tag{3.6}$$

By assumption we have that

$$\log \det n \ [\, \mathrm{COV}(\theta \mid X^n) \,] + \log \det I(\theta) \to 0, \tag{3.7}$$

in $P_{X^n \mid \theta}$ probability, for each θ in the support of w, and therefore in the joint probability of (Θ, X^∞). Now, by uniform integrability from above we have

$$\limsup_{n \to \infty} E_m \ [\, \log \det n \mathrm{COV}(\theta \mid X^n) \,] \leq - \int_\Omega \log \det I^{-1}(\theta) w(\theta) \, d\theta. \tag{3.8}$$

Finally, from inequality (3.2), we have that

$$\begin{aligned} \liminf_{n \to \infty} [\, I(\Theta; X^n) - H(\Theta) - \frac{d}{2} \log \frac{n}{2\pi e} \,] &\geq - \limsup_{n \to \infty} E_m \ [\, \log \det n \mathrm{COV}(\theta \mid X^n) \,] \\ &= \int_\Omega w(\theta) \log \det I(\theta) \, d\theta, \end{aligned}$$

which proves part (2.6) of the theorem. To finish, we examine the minimax value

$$R_n = \inf_{Q_n} \sup_{\theta \in \Omega} R_n(Q_n, \theta)$$

in which Q_n is a subprobability used to estimate of the density P_θ, and

$$R_n(Q_n, \theta) = D(P_\theta^n \mid\mid Q_n).$$

Since an average lower bounds a supremum, we have that the minimax value can be lower bounded by the Bayes risk of the Bayes estimator, which is $R_n(w) = I(\Theta; X^n)$. Now (2.6) implies (2.7), finishing the proof of part a) of the theorem.

Before embarking on the proof of the upper bound part, we give the proof of Proposition 2.2.

Proof of Proposition 2.2: Let

$$R_n = \log \frac{p_{\theta_o}(X^n)}{m(X^n)} - \frac{d}{2} \log n + \frac{\mid Z_n \mid^2}{2} + \log [\, w(\theta_o)(2\pi)^{d/2} \det I(\theta_o)^{-1/2} \,]. \tag{3.9}$$

Our task is to upper bound R_n by a quantity that goes to zero in $L_1(P_{\theta_o})$. By conclusion (1) of Proposition 2.1, we have that R_n goes to zero in probability. Therefore we must show that it is uniformly integrable from above. Since $\mid Z_n \mid^2$ is convergent in distribution and has constant expectation, it is uniformly integrable. It remains to upper bound $\log \frac{p_{\theta_o}(X^n)}{m(X^n)} - \frac{d}{2} \log n$ by a uniformly integrable sequence of random variables. First, note that

$$m(X^n) \geq \int_B w(\theta) p_\theta(X^n)\, d\theta.$$

So,

$$\log \frac{p_{\theta_o}(X^n)}{m(X^n)} - \frac{d}{2}\log n \leq -\log\left[\frac{\int_B w(\theta)p_\theta(X^n)/p_{\theta_o}(X^n)d\theta}{n^{-d/2}}\right]$$

$$\leq -\log\left[\frac{\int_B p_\theta(X^n)/p_{\theta_o}(X^n)\,d\theta}{Vol(B)}\right] + \log\frac{Vol(B)}{n^{-d/2}} - \log \underline{w}(\theta_o), \quad (3.10)$$

where $B = \{\theta: \mid \theta - \theta_o \mid_{I(\theta_o)} \leq K/\sqrt{n}\}$, $\underline{w}(\theta_o) = \inf_{\theta \in B} w(\theta)$ which goes to $w(\theta_o)$ as $n \to \infty$, and $Vol(B)/n^{-d/2} = Vol\{u: \mid u \mid \leq K\} \det I(\theta_o)^{1/2}$ is constant. Now we use Jensen's inequality to get

$$-\log\frac{\int_B p_\theta(X^n)/p_{\theta_o}(X^n)\,d\theta}{Vol(B)} \leq \frac{1}{Vol(B)}\int_B \log p_{\theta_o}(X^n)/p_\theta(X^n)\,d\theta$$

$$= \frac{1}{Vol(C)}\int_{\mid u \mid \leq K} \log p_{\theta_o}(X^n)/P_{\theta(u)}(X^n)\,du, \quad (3.11)$$

where $\theta(u) = \theta_o + (nI(\theta_o))^{-1/2}u$ and $C = \{u: \mid u \mid \leq K\}$. This is the upper bound which we show to be uniformly integrable. Since $\int_{\mid u \mid \leq K} u\, du = 0$, it is equivalent to show that

$$\int_{\mid u \mid \leq K} [\log p_{\theta_o}(X^n)/p_{\theta(u)}(X^n) - u^T Z_n]\, du \quad (3.12)$$

is uniformly integrable. Now Condition 2 implies that the integrand $\log p_{\theta_o}(X^n)/p_{\theta(u)}(X^n) - u^T Z_n$ converges to $(1/2)\mid u \mid^2$ in P_{θ_o} probability for each fixed u in C. This follows from LeCam's asymptotic normality of experiments, see LeCam (1986). Consequently, by application of Fubini's theorem, the convergence also holds in probability with respect to $\nu \times P_{\theta_o}$ where $\nu(du)$ is the uniform distribution for u in C. Then,

$$\lim_{n \to \infty} E_{P_{\theta_o}} \int_C \mid \log p_{\theta_o}(X^n)/p_{\theta(u)}(X^n) - u^T Z_n - \frac{1}{2}\mid u \mid^2 \mid du = 0, \quad (3.13)$$

provided that $\log p_{\theta_o}(X^n)/p_{\theta(u)}(X^n) - u^T Z_n$ is uniformly integrable in $L_1(\nu \times P_{\theta_o})$.

To show that $\log p_{\theta_o}(X^n)/p_{\theta(u)}(X^n) - u^T Z_n$ is uniformly integrable we note that it is the sum of a positive quantity (3.14) and a uniformly integrable quantity (3.15). Namely:

$\log p_{\theta_o}(X^n)/p_{\theta(u)}(X^n) - u^T Z_n$

$$= 2(-\log[p_{\theta(u)}(X^n)/p_{\theta_o}(X^n)]^{1/2} + [p_{\theta(u)}(X^n)/p_{\theta_o}(X^n)]^{1/2} - 1) \quad (3.14)$$

$$+ 2(1 - [p_{\theta(u)}(X^n)/p_{\theta_o}(X^n)]^{1/2}) - u^T Z_n. \quad (3.15)$$

Here (3.14) is positive (since $\log \rho \le \rho - 1$ for $\rho \ge 0$) and (3.15) is uniformly integrable in $L_1(\nu \times P_{\theta_o})$. Indeed, because they have bounded expected square, $(p_{\theta(u)}(X^n)/p_{\theta_o}(X^n))^{1/2} - 1$ and $u^T Z_n$ are uniformly integrable. Thus, to show that $\log p_{\theta_o}(X^n)/p_{\theta(u)}(X^n) - u^T Z_n$ is uniformly integrable in $L_1(\nu \times P_{\theta_o})$ it is enough to show that the sequence of expected values converges to the expected value of the limit. Using expression (2.1), the limit of the expected values of (3.14) plus (3.15) with respect to $L_1(\nu \times P_{\theta_o})$ is

$$\begin{aligned}
&\lim_{n \to \infty} \frac{1}{Vol(C)} \int_C E_{p_{\theta_o}} \log p_{\theta_o}(X^n)/p_{\theta(u)}(X^n) - u^T Z_n \, du \\
&\quad = \lim_{n \to \infty} \frac{1}{Vol(C)} \int_C D(P^n_{\theta_o} \,||\, P^n_{\theta(u)}) \, du \\
&\quad = \lim_{n \to \infty} \frac{1}{Vol(C)} \int_C \frac{n}{2} \mid \theta(u) - \theta_o \mid^2_{I(\theta_o)} du \\
&\quad = \frac{1}{Vol(C)} \int_C \frac{1}{2} \mid u \mid^2 du, \qquad (3.16)
\end{aligned}$$

which is the expected value of the limit in $\nu \times P_{\theta_o}$, as desired.

From (3.13) we conclude that

$$\lim_{n \to \infty} E_{p_{\theta_o}} \mid \int_C (\log p_{\theta_o}(X^n)/p_{\theta(u)}(X^n) - u^T Z_n \,) \, du - \frac{1}{2} \int_C \mid u \mid^2 du \mid = 0. \qquad (3.17)$$

So, the integral in (3.12), or equivalently (3.11), is uniformly integrable in $L_1(P_{\theta_o})$. This completes the proof of Proposition 2.2. □

Proof of Theorem 2.1, upper bound part: We set up an application of the dominated convergence theorem. Let

$$\psi_n(\theta) = D(p^n_\theta \,||\, m_n) - \frac{d}{2} \log n,$$

An upper bound for the pointwise behavior of ψ_n is given by Proposition 2.2, as being

$$\frac{d}{2} \log \frac{1}{2\pi e} + \frac{1}{2} \log \det I(\theta) + \log \frac{1}{w(\theta)}.$$

To set up the domination, note that by an inequality due to Barron (1987) we have that

$$\psi_n(\theta) \le 2\varepsilon n - \frac{d}{2} \log n - \log W\{\theta' : D(\theta \,||\, \theta') < 2\varepsilon\}, \qquad (3.18)$$

where W is the prior probability, which we must lower bound. By assumptions (2.8) and (2.10) we have that

$$D(\theta \,||\, \theta') \le \frac{1}{2} \mid \theta - \theta' \mid^2_{I_\varepsilon(\theta)}$$

$$\leq |\theta - \theta''|^2_{I_\varepsilon(\theta)} + |\theta' - \theta''|^2_{I_\varepsilon(\theta)}$$

$$\leq \varepsilon + |\theta' - \theta''|^2_{I_\varepsilon(\theta)}. \tag{3.19}$$

Now we have that

$$W\{\theta': D(\theta \| \theta') < 2\varepsilon\} \geq W\{\theta': |\theta' - \theta''|^2_{I_\varepsilon(\theta)} < \varepsilon\}$$

$$\geq \inf_{|\theta' - \theta''|^2_{I_\varepsilon(\theta)} < \varepsilon} w(\theta') \int_{\{\theta': |\theta' - \theta''|^2_{I_\varepsilon(\theta)} < \varepsilon\}} d\theta$$

$$= \inf_{|\theta' - \theta''|^2_{I_\varepsilon(\theta)} < \varepsilon} w(\theta') \det\left(\frac{I_\varepsilon(\theta)}{\varepsilon}\right)^{-1/2} \mathrm{Vol}(B(0, 1)), \tag{3.20}$$

where the volume of the unit ball in d dimensions is a constant that does not affect the calculations. We choose $\varepsilon = \varepsilon_n = 1/n$ and obtain the upper bound

$$\psi_n(\theta) \leq -\log \inf_{|\theta' - \theta''|^2_{I_\varepsilon(\theta)} < \varepsilon} w(\theta') + \frac{1}{2} \log \det I_\varepsilon(\theta) + c, \tag{3.21}$$

where c is a constant. Since there is an $\varepsilon' > 0$ so that the integral of both terms is finite, we can apply the dominated convergence theorem to the sequence ψ_n for $n > 1/\varepsilon'$. Now the upper bound part of Theorem 2.1 follows. □

4. Asymptotic normality of the posterior in information. In this section we prove Theorem 2.2. This amounts to noting that when moments match, a Kullback-Leibler number can be written as a difference of conditional entropies.

Proof of Theorem 2.2: Let $Z = Z_{\Theta | X^n}$ denote a random variable for which the conditional distribution of Z given X^n is normal with mean $E(\Theta | X^n)$ and variance matrix $\mathrm{COV}(\Theta | X^n)$. 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 $N(E(\Theta | X^n), \mathrm{COV}(\Theta | X^n))$. By the definition of the mutual information

$$I(\Theta; X^n) = H(w) - H(w | X^n)$$

$$= H(w) - H(Z | X^n) + [H(Z | X^n) - H(w | X^n)]$$

$$= H(w) - \frac{1}{2} E_m \log (2\pi e)^d \det \mathrm{COV}(\Theta | X^n)$$

$$+ E_m D(w(\cdot | X^n) \| N(E(\Theta | X^n), \mathrm{COV}(\Theta | X^n))), \tag{4.1}$$

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

$$E_m D(w(\cdot \mid X^n) \mid\mid N(E_{w(\cdot \mid X^n)}\Theta, \mathrm{COV}_{w(\cdot \mid X^n)}\Theta))$$

$$= I(\theta; X^n) - H(w) - \frac{d}{2}\log\frac{n}{2\pi e} - \frac{1}{2}\int_\Omega \log\det I(\theta)\, w(\theta)\, d\theta \tag{4.2}$$

$$+ \frac{1}{2}\left(E_m \log\det n\,\mathrm{COV}(\Theta \mid X^n) - \int_\Omega \log\det I(\theta)^{-1}\, w(\theta)\, d\theta\right). \tag{4.3}$$

Expression (4.2) is controlled by the assumed upper bound on $I(\Theta; X^n)$. In view of (3.8), (4.3) tends to zero also. Thus, we obtain (2.13).

Next, we prove the equivalence of *i*), *ii*) and *iii*). First, *i*) implies *ii*): The assumed upper bound deals with (4.2), and the existence of the limit from *i*) gives that expression (4.3) tends to zero, which is the same as (2.14) in *ii*).

We have that *ii*) implies *iii*): All we require is a tight lower bound on $I(\Theta; X^n)$. Since the Kullback-Leibler number is positive, use of *ii*) in (4.3) gives that

$$I(\theta; X^n) - H(w) - \frac{d}{2}\log\frac{n}{2\pi e} - \frac{1}{2}\int_\Omega \log\det I(\theta)\, w(\theta)\, d\theta \geq 0. \tag{4.4}$$

Last, *iii*) implies *i*): To control (2.13) note that, by *iii*), (4.2) goes to zero, and the assumption (2.4) allows the argument in the proof of the lower bound part of Theorem 2.1 to hold so that (3.3) is valid. This gives an upper bound of zero for the positive quantity in the limit of (2.13).□

5. L_1 Asymptotic normality of the posterior. In this section we prove Proposition 2.1 so that we can replace the assumption (2.5) in Theorem 2.1 with better hypotheses.

Proof of Proposition 2.1 We first demonstrate the conclusions (1) and (2). In place of $w_n(\theta)$ we deal with $w_n^*(\theta) = C_n\, w_n(\theta)$ given by

$$w_n^*(\theta) = \frac{w(\theta)p_\theta(X^n)}{w(\theta_o)p_{\theta_o}(X^n)\det(nI(\theta_o))^{-1/2}(2\pi)^{d/2}e^{\mid Z_n\mid^2/2}}, \tag{5.1}$$

which has the advantage of agreeing with the normal approximation at $\theta = \theta_o$. Here,

$$C_n = \frac{m(X^n)}{w(\theta_o)p_{\theta_o}(X^n)\det(nI(\theta_o))^{-1/2}(2\pi)^{d/2}e^{\mid Z_n\mid^2/2}}, \tag{5.2}$$

and we note that conclusion (1) is equivalent to $C_n \to 1$ in probability.

Letting $\phi_n(\theta) = \phi_{\theta', (nI(\theta_o))^{-1}}(\theta)$, we have that

$$\mid C_n - 1\mid = \mid \int (w_n^* - \phi_n)\mid \leq \int \mid w_n^* - \phi_n\mid, \tag{5.3}$$

and

$$\begin{aligned}\int | w_n - \phi_n | &= \int | w_n - C_n w_n + C_n w_n - \phi_n | \\ &\le | C_n - 1 | + \int | w_n^* - \phi_n | \\ &\le 2 \int | w_n^* - \phi_n |. \end{aligned} \tag{5.4}$$

Consequently, to demonstrate conclusions (1) and (2), it is enough to show that $\int | w_n^* - \phi_n | \to 0$, in probability.

We decompose the L_1 distance $\int | w_n^* - \phi_n |$ into three pieces, each of which is shown to have asymptotically negligible contributions in P_{θ_o} probability. Let $B = \{\theta: | \theta - \theta_o |_{I(\theta_o)} \le K/\sqrt{n}\}$. We have

$$\int | w_n^*(\theta) - \phi_n(\theta) | \, d\theta \le \int_B | w_n^* - \phi_n | \, d\theta + \int_{B^c} w_n^* \, d\theta + \int_{B^c} \phi_n \, d\theta. \tag{5.5}$$

First we handle the last term in (5.5). The event B^c is contained in the event $B'^c = \{\theta: | \theta - \theta' |_{I(\theta_o)} \ge K/\sqrt{n} - | Z_n | /\sqrt{n}\}$, and a change of variables to $t = (nI(\theta_o))^{-1/2}(\theta - \theta')$ yields $\int_{B'^c} \phi_n(\theta) \, d\theta = \int_{| t | \ge K - | Z_n |} \phi(t) \, dt$ where ϕ is the standard normal density on $\mathbf{R}^d$, which is negligible for large K, as can be seen by two applications of Chebyshev's inequality. One gives that $| Z_n | \le K/2$ except on a set of probability $4d/K^2$; the other gives that when $| Z_n | \le K/2$, we have that $\int_{B^c} \phi_n \le 4d/K^2$.

Next we handle the middle term on the right in (5.5). Conditions (1) and (2) imply that there exists a test of θ_o versus B^c with acceptance region A_n such that $P_{\theta_o}(A_n^c) \le e^{-a_0 K^2}$, $P_\theta(A_n) \le e^{-n a_1 | \theta - \theta_o |^2/2}$ for $K/\sqrt{n} \le | \theta - \theta_o | \le \varepsilon$, and $P_\theta(A_n) \le e^{-n a_2 \varepsilon^2}$ uniformly for $| \theta - \theta_o | > \varepsilon$ and all n, where a_0, a_1, a_2 are positive constants. This follows from LeCam (1986, pp. 619-621), see also LeCam and Yang (1990), page 155 and 161, steps 4,5, and 6. Then, by Markov's inequality and Fubini's theorem,

$$\begin{aligned} P_{\theta_o}\Big(\int_{B^c} w_n^* \, d\theta \ge 1/K\Big) &\le P_{\theta_o}\Big(A_n \cap \int_{B^c} w_n^* \, d\theta \ge 1/K\Big) + P_{\theta_o}(A_n^c) \\ &\le \int_{B^c} E_{\theta_o} \mathbf{1}_{A_n} w_n^*(\theta) \, d\theta + e^{-a_0 K^2}. \end{aligned} \tag{5.6}$$

Now, by definition (5.1) of $w_n^*(\theta)$ and $e^{-| Z_n |^2/2} \le 1$ we have

$$\begin{aligned} E_{\theta_o} \mathbf{1}_{A_n} w_n^*(\theta) &\le \frac{w(\theta) E_{\theta_o} \mathbf{1}_{A_n} (p_\theta(X^n)/p_{\theta_o}(X^n))}{w(\theta_o)(2\pi)^{d/2} \det(nI(\theta_o))^{-1/2}} \\ &\le \frac{w(\theta) P_\theta(A_n)}{w(\theta_o)(2\pi)^{d/2} \det(nI(\theta_o))^{-1/2}}. \end{aligned} \tag{5.7}$$

By the continuity of the prior at θ_o, we may choose $\varepsilon > 0$ such that $w(\theta) \le 2w(\theta_o)$ for $|\theta - \theta_o| < \varepsilon$ and integrate the bounds to get

$$\begin{aligned}\int_{B^c} w(\theta)P_\theta(A_n)\,d\theta &\le \int_{\varepsilon > |\theta-\theta_o| \ge K/\sqrt{n}} w(\theta)P_\theta(A_n)\,d\theta + \int_{|\theta-\theta_o|>\varepsilon} w(\theta)P_\theta(A_n)\,d\theta \\ &\le 2w(\theta_o)\int_{\varepsilon > |\theta-\theta_o| \ge K/\sqrt{n}} e^{-na_1|\theta-\theta_o|^2/2}\,d\theta + e^{-na_2\varepsilon^2} \\ &\le \frac{2w(\theta_o)}{n^{d/2}}\int_{|u|>K} e^{-a_1|u|^2/2}\,du + e^{-na_2\varepsilon^2} \\ &\le \frac{2w(\theta_o)(2\pi)^{d/2}}{(a_1 n)^{d/2}}\frac{d}{a_1K^2} + e^{-na_2\varepsilon^2}. \end{aligned} \tag{5.8}$$

Combining (5.6), (5.7), and (5.8) yields

$$P_{\theta_o}\Big(\int_{B^c} w_n^*\,d\theta \ge 1/K\Big) \le e^{-a_0K^2} + Ke^{-na_2\varepsilon^2} + \frac{2d}{a_1^{d/2}K\det I(\theta_o)^{-1/2}}. \tag{5.9}$$

Letting $n \to \infty$, we see that the contribution from the term $\int_{B^c} w^*(\theta)\,d\theta$ is negligible for large K.

To complete the proof of parts (1) and (2), it remains to show that the integral over B is negligible. By the continuity and positivity of w at θ_o we remove the dependence on the prior. The triangle inequality gives

$$\begin{aligned}\int_B |w_n^*(\theta) - \phi_n(\theta)|\,d\theta &= \int_B \Big|\frac{w(\theta)}{w(\theta_o)}v_n(\theta) - \phi_n(\theta)\Big|\,d\theta \\ &\le \int_B \Big|\frac{w(\theta)}{w(\theta_o)} - 1\Big|\,v_n(\theta)\,d\theta + \int_B |v_n(\theta) - \phi_n(\theta)|\,d\theta \\ &\le \rho\int_B v_n(\theta)\,d\theta + \int_B |v_n(\theta) - \phi_n(\theta)|\,d\theta \\ &\le \rho + (1+\rho)\int_B |v_n(\theta) - \phi_n(\theta)|\,d\theta, \end{aligned} \tag{5.10}$$

where $\rho = \sup_{\theta \in B} |w(\theta)/w(\theta_o) - 1|$ which tends to zero as n increases. Here we have let

$$v_n(\theta) = \frac{p_\theta(X^n)}{p_{\theta_o}(X^n)(2\pi)^{d/2}e^{|Z_n|^2/2}\det(nI(\theta_o))^{-1/2}}. \tag{5.11}$$

Now it is enough to show that the integral in (5.10) goes to zero in probability. Changing variables to $u = (nI(\theta_o))^{1/2}(\theta - \theta_o)$ the integral simplifies to

$$\int_B |v_n(\theta) - \phi_n(\theta)|\,d\theta = \frac{1}{(2\pi)^{d/2}}\int_{|u|\le K} \Big|\frac{p_{\theta(u)}(X^n)}{p_{\theta_o}(X^n)e^{|Z_n|^2/2}} - e^{-|u-Z_n|^2/2}\Big|\,du, \tag{5.12}$$

where $\theta(u) = \theta_o + (nI(\theta_o))^{-1/2}u$. We show that the expected value with respect to P_{θ_o} converges to zero, from which convergence in probability follows. By Fubini's theorem the expected value is

$$E_{\theta_o} \int_B | v_n(\theta) - \phi_n(\theta) | \, d\theta$$

$$= \frac{1}{(2\pi)^{d/2}} \int_{|u| \le K} E_{\theta_o} \left| \frac{p_{\theta(u)}(X^n)}{p_{\theta_o}(X^n) e^{|Z_n|^2/2}} - e^{-|u - Z_n|^2/2} \right| du. \tag{5.13}$$

To show that the integral tends to zero, it follows from the bounded convergence theorem that it is enough to show that for each fixed u the following expectation converges to zero,

$$E_{\theta_o} \left| \frac{p_{\theta(u)}(X^n)}{p_{\theta_o}(X^n) e^{|Z_n|^2/2}} - e^{-|u - Z_n|^2/2} \right| . \tag{5.14}$$

Now, condition (2) implies that the quantity inside the expectation tends to zero in probability. This is the Hajek-LeCam theorem for the local asymptotic normality of the family of experiments $P_{\theta(u)}$; see Ibragimov and Hasminskii (1981).

To get the desired $L_1(P_{\theta_o})$ convergence, it remains to show that the quantity inside the expectation is uniformly integrable in $L_1(P_{\theta_o})$. Since the two exponentials in (5.14) are bounded by one, it suffices to show that $p_{\theta(u)}(X^n)/p_{\theta_o}(X^n)$ is uniformly integrable in $L_1(P_{\theta_o})$. Now, $p_{\theta(u)}(X^n)/p_{\theta_o}(X^n)$ is a positive random variable that converges in distribution to $e^{u^T Z - |u|^2/2}$ where Z is a standard normal random vector on $\mathbf{R}^d$, again by the Hajek-LeCam theorem. So, for uniform integrability, it suffices to note that the expectations $E_{\theta_o} p_{\theta(u)}(X^n)/p_{\theta_o}(X^n)$ ~~are bounded by~~ converges to one, which is the expectation of the limit, $E\, e^{u^T Z - |u|^2/2} = 1$. This last calculation follows by recalling that the moment generating function of Z is $E\,(e^{u^T Z}) = e^{|u|^2/2}$. We have now completed the proof of conclusions (1) and (2).

To prove conclusion (3), let $r > 0$ be given. Assume $\int_\Omega |\theta|^r \, d\theta$, is finite and hence that $\int_\Omega |\theta - \theta_o|^r_{I(\theta_o)} \, d\theta$ is finite also. Now, we let $B' = \{\theta: |\theta - \theta'|_{I(\theta_o)} \le K/\sqrt{n}\}$. We bound the contributions in the following decomposition.

$$\int_\Omega |\sqrt{n}(\theta - \theta')|^r_{I(\theta_o)} \, |w_n(\theta) - \phi_n(\theta)| \, d\theta$$

$$\le K^r \int_{B'} |w_n(\theta) - \phi_n(\theta)| \, d\theta + \int_{B'^c} |\sqrt{n}(\theta - \theta')|^r_{I(\theta_o)} \, \phi_n(\theta) | \, d\theta$$

$$+ \int_{B'^c} |\sqrt{n}(\theta - \theta')|^r_{I(\theta_o)} \, w_n(\theta)| \, d\theta. \tag{5.15}$$

The first term is less than $K^r \int |w_n - \phi_n|$ which tends to zero in probability by conclusion 2.

By change of variables the second term is $\int_{|t|>K} |t|\,\phi(t)\,dt$ where ϕ is the standard normal density on $\mathbf{R}^d$. For fixed $r > 0$, it tends to zero as K increases. It remains to bound the last term. When $|Z_n| \le K/2$, we have that on B'^c,

$$K \le \sqrt{n}\,|\theta-\theta'|_{I(\theta_o)} \le \sqrt{n}\,|\theta-\theta_o|_{I(\theta_o)} + |Z_n|$$
$$\le \sqrt{n}\,|\theta-\theta_o|_{I(\theta_o)} + K/2 \tag{5.16}$$

and consequently, $K/2 \le \sqrt{n}\,|\theta-\theta_o|_{I(\theta_o)}$ so continuing the inequality in (5.16) we get

$$|\theta-\theta'|_{I(\theta_o)} \le 2\sqrt{n}\,|\theta-\theta_o|_{I(\theta_o)}. \tag{5.17}$$

So, when $|Z_n| \le K/2$,the remaining term in (5.15) is bounded by

$$\int_{B'^c} |\sqrt{n}(\theta-\theta')|^r_{I(\theta_o)}\, w_n(\theta)|\,d\theta \le \int_{B^c} 2|\sqrt{n}(\theta-\theta_o)|^r_{I(\theta_o)}\, w_n(\theta)|\,d\theta, \tag{5.18}$$

where here $B^c = \{2\sqrt{n}\ |(\theta-\theta_o)|_{I(\theta_o)} \ge K\}$. Expression (5.18) equals

$$\frac{1}{C_n}\int_{B^c} 2|\sqrt{n}(\theta-\theta_o)|^r_{I(\theta_o)}\, w_n^*(\theta)|\,d\theta. \tag{5.19}$$

Here, C_n is defined as in (5.2), and by conclusion (1) it converges to one in probability. Thus it remains to control

$$\int_{B^c} 2|\sqrt{n}(\theta-\theta_o)|^r_{I(\theta_o)}\, w_n^*(\theta)|\,d\theta. \tag{5.20}$$

This can be done in the same manner as in (5.6), (5.7) and (5.8). The analogue of (5.6) and (5.7) becomes

$$P^n_{\theta_o}\{\int_{B^c} 2|\sqrt{n}(\theta-\theta_o)|^r_{I(\theta_o)}\, w_n^*(\theta)|\,d\theta \ge 1/K\}$$
$$\le K\,\frac{\int_{B^c} 2|\sqrt{n}(\theta-\theta_o)|^r_{I(\theta_o)}\, w(\theta)P_\theta(A_n)|\,d\theta}{w(\theta_o)n^{-d/2}(2\pi)^{d/2}\det I(\theta_o)^{-1/2}} + e^{-a_0K^2/4}. \tag{5.21}$$

The analogue of (5.8) becomes

$$\int_{B^c} 2|\sqrt{n}(\theta-\theta_o)|^r_{I(\theta_o)}\, w(\theta)P_\theta(A_n)|\,d\theta$$
$$\le \frac{2w(\theta_o)}{n^{d/2}}\int_{2|u|\ge K} 2|u|^r\, e^{-a_1|u|^2/2}\,du$$
$$+ 2^r n^{r/2} e^{-na_2\varepsilon^2}\int_\Omega |(\theta-\theta_o)|^{r_{I(\theta_o)}}\, w(\theta)|\,d\theta. \tag{5.22}$$

By Markov's inequality,

$$\int_{2|u|\ge K} 2|u|^r\, e^{-a_1|u|^2/2}\,du \le (\frac{1}{K})^{r+1}\int \tilde{2}|u|^{r+1}\, e^{-a_1|u|^2/2}\,du. \tag{5.23}$$

Combining (5.22) with (5.21) yields

$$P^n_{\theta_o}\{\int_{B^c} 2 \mid \sqrt{n}\,(\theta - \theta_o)\mid^r_{I(\theta_o)} w^*_n(\theta) \mid d\theta \geq 1/K\}$$

$$\leq O(1/K^r) + O(Kn^{(d+r)/2}e^{-na_2\varepsilon^2}) + O(e^{-a_0K^2}) \tag{5.24}$$

which tends to zero as $n \to \infty$ and then $K \to \infty$. This completes the proof of convergence of r^{th} moments in conclusion (3).

Using $r = 1$, we obtain

$$\sqrt{n} \mid E(\Theta \mid X^n) - \theta' \mid = \mid \int \sqrt{n}\,(\theta - \theta')(w_n(\theta) - \phi_n(\theta)\, d\theta \mid$$

$$\leq \int \sqrt{n} \mid \theta - \theta' \mid \mid w_n(\theta) - \phi_n(\theta) \mid d\theta \tag{5.25}$$

which goes to zero in probability. Thus, $\sqrt{n}\,(E(\Theta \mid X^n) - \theta_o) - I(\theta_o)^{-1/2}Z_n$ goes to zero in probability. Since $I(\theta_o)^{-1/2}Z_n$ converges in distribution to $N(0, I(\theta_o)^{-1})$, by the central limit theorem, it follows that $\sqrt{n}\,(E(\Theta \mid X^n) - \theta_o)$ also converges in distribution to the same limit.

Taking $r = 2$ we find in the same manner that

$$\mid nE[\,(\theta - \theta')(\theta - \theta')^T \mid X^n] - I(\theta_o)^{-1} \mid$$

$$= \mid \int n\ (\theta - \theta')(\theta - \theta')^T\ (w(_n(\theta) - \phi_n(\theta))\, d\theta \mid \tag{5.26}$$

and the right hand side of (5.26) goes to zero in probability.

Finally,

$$n\,\mathrm{COV}(\Theta \mid X^n) = nE[\,(\Theta - \theta')(\Theta - \theta')^T \mid X^n) - n(E(\Theta \mid X^n) - \theta')(E(\Theta \mid X^n) - \theta')^T \tag{5.27}$$

and the last term tends to zero in probability by (5.25). Consequently,

$$\mid n\,\mathrm{COV}(\Theta \mid X^n) - I(\theta_o)^{-1} \mid \to 0, \tag{5.28}$$

in probability. This completes the proof of part A.

For part B, note that as a consequence of (5.25) and (5.28), if $\int \mid \theta \mid^2 w(\theta)\, d\theta$ is finite, we may replace θ' and $(nI(\theta_o))^{-1}$ in the normal density approximation to conclude that

$$\int \mid w_n - \phi_{E(\Theta \mid X^n),\ \mathrm{COV}(\Theta \mid X^n)} \mid \to 0, \tag{5.29}$$

in p_{θ_o} probability, and hence also in P_{θ_o} expectation since it is bounded by the constant 2. Assuming that Conditions 1 and 2 are satisfied for every θ in a set of prior probability one, it follows by Fubini's theorem and the bounded convergence theorem that

$$\lim_{n \to \infty} E_{M_n} \int \mid w(\cdot \mid X^n) - \phi_{E(\Theta \mid X^n),\ \mathrm{COV}(\Theta \mid X^n)} \mid$$

$$= \lim_{n \to \infty} \int w(\theta) E_\theta \int \mid w(\cdot \mid X^n) - \phi_{E(\Theta \mid X^n),\ \mathrm{COV}(\Theta \mid X^n)} \mid = 0.$$

This completes the proof of the theorem. □

6. Applications. In this section we give two applications of our results. The first is to parametric density estimation. We show how the quantity we have examined lower bounds the risk in parametric estimation. The second is to investment theory. We show that the wealth achieved by the Bayes optimal strategy for investment differs by only a polynomial factor from the wealth achieved by the strategy one would employ if one had full market knowledge.

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:

$$\inf_{\delta} E_w E_\theta D(\theta \,||\, \delta) \geq \inf_{Q} E_w E_\theta D(P_\theta \,||\, Q), \tag{6.1}$$

where δ is an estimator of the parameter, Q is an estimator of the density and $D(\theta \,||\, \delta) = D(P_\theta \,||\, P_\delta)$ is the relative entropy loss for parameter estimation. The quantity in Theorem 2.1 gives an asymptotic lower bound on the Bayes risk of parameter estimation.

This lower bound is achieved by the predictive distribution $\hat{P}_n$, which is the Bayes estimator of $p(x_{n+1} \mid \theta)$ based on X^n, as in Clarke and Barron (1990), see also Aitchison (1975). Under the conditions of Theorem 2.1 in the former, the individual risk terms $E_{\theta_o} D(P_{\theta_o} \,||\, \hat{P}_n)$ also converge to zero as $n \to \infty$. This follows from noting that

$$E_{\theta_o} D(P_{\theta_o} \,||\, \hat{P}_n) = D(P^n_{\theta_o} \,||\, M_n) - D(P^{n-1}_{\theta_o} \,||\, M_{n-1}), \tag{6.2}$$

and applying Theorem 2.1 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

$$\int_{\Omega} E_\theta D(P_\theta \,||\, \hat{P}_k) w(\theta) d\theta = o(1), \tag{6.3}$$

from Theorem 2.1 here.

In the investment theory context, we generalize a result due to Barron and Cover (1988). Assume that each $X_i = (X_{i,1},\ldots,X_{i,k})$ represents the k multiplicative factors by which dollars invested in stock j for $j = 1,\ldots,k$, are increased or decreased during the i^{th} investment period. At the beginning of each investment period, stocks are bought or sold so as to result in a

portfolio $b_i = (b_{i,1},...,b_{i,k})$ of stock proportions with each $b_{i,j}$ positive, and $\sum_{j=1}^{k} b_{i,j} = 1$. If one dollar is invested, at the end of n investment periods we have wealth

$$W_n = \prod_{i=1}^{n} b_i^T X_i = e^{\sum_{i=1}^{n} \log b_i^T X_i}.$$

The sequence of b_i's is our investment strategy.

If we knew θ, then the optimal strategy would be to choose $b_i = b(\theta)$ so as to achieve

$$Q(\theta) = \sup_b E_{p_\theta} \log b^T X_1,$$

with resulting wealth denoted W^*. On average, this strategy performs better than any other one.

We compare W^* to the wealth achieved by the Bayes strategy which is optimal, at each time step, with respect to the predictive distribution, rather than the true distribution. That means we choose $b_i = b(X^{i-1})$ so as to achieve

$$Q_B = \sup_b E_{m(\cdot|\, X^{i-1})} \log b^T X_i,$$

with resulting wealth denoted W_B.

Both expressions for the wealth achieved, W_B and W^*, grow at an exponential rate. The following result shows that best strategy, based on information we can never have, outperforms the Bayes strategy by a polynomial factor, at best.

Proposition 6.1: *As $n \to \infty$, W^* and W_B differ only by a polynomial factor, with high joint (Θ, X^n) probability.*

Proof: Let $C > 0$ be large. We have that

$$P_{\Theta, X^n}(\log \frac{W^*}{W_B} > cI(\Theta; X^n)) \leq \frac{E_{\Theta, X^n} [\log \frac{W^*}{W_B}]^+}{CI(\Theta; X^n)}. \tag{6.4}$$

Next, by adding and subtracting the negative part, we note that

$$E [\log \frac{W^*}{W_B}]^+ = E \log \frac{W^*}{W_B} + E [\log \frac{W_B}{W^*}]^+ \tag{6.5}$$

By Theorem 3 in Barron and Cover (1988), the first term on the right of (6.5) is bounded above by $I(\Theta; X^n)$. The second term on the right of (6.5) is bounded above by unity. This follows from the Kuhn-Tucker conditions for the optimality of $b(\theta)$, see Bell and Cover (1980).

Using (6.5) in (6.4), and dividing through on the right by $I(\Theta; X^n)$ gives

$$P_{\Theta, X^n}(\log \frac{W^*}{W_B} > cI(\Theta; X^n)) \leq \frac{1}{C} + \frac{1}{C \log n}, \tag{6.6}$$

which implies that

$$W_B \geq W^* \ e^{-C\, I(\Theta; X^n)} \tag{6.7}$$

holds with high joint (Θ, X^n) probability for large C. The expansion for $I(\Theta; X^n)$ given in Theorem 2.1 ensures that scaling up W_B by a polynomial factor makes the Bayes strategy perform at least as well as the optimal strategy. □

REFERENCES

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

Barron, A. R. (1987). The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Under revision for *The Annals of Statistics*.

Barron, A. R. and Cover, T. M. (1988). A bound on the financial value of information. *IEEE Trans. Inform. Theory* **34** 1097-1100.

Bell, R. and Cover, T. (1988). Competitive optimality of logarithmic investment. *Math. Operations Res.* **5** 161-166.

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. and Barron, A. (1990). Information theoretic asymptotics of Bayes methods. *IEEE Trans. Inform. Theory* **36** 453-471.

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

Hartigan, J. A. (1983). *Bayes Theory*. 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 Hasminskii, R. Z. (1973). On the information in a sample about a parameter. *Second International Symposium on Information Theory* 295-309 Akademiai, Kiado, Budapest. Ibragimov, I. A. and Hasminskii, R. Z. (1981). 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.

LeCam, L. (1958). Les proprietes asymptotique des solutions de Bayes. *Publ. Inst. Statist. Univ. Paris.* **7** 18-35.

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

LeCam, L. and Yang, G. (1990). *Asymptotics in Statistics.* Springer-Verlag, New York.

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.

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.

Walker, A. M. (1967). On the asymptotic behaviour of posterior distributions. *J. Roy. Statist. Soc. Ser. B* **31** 80-88.