Information-Theoretic Asymptotics of Bayes Methods

BERTRAND S. CLARKE AND ANDREW R. BARRON, MEMBER, IEEE

Abstract—**In the absence of knowledge of the true density function, Bayesian models take the joint density function for a sequence of n random variables to be an average of densities with respect to a prior. We examine the relative entropy distance D_n between the true density and the Bayesian density and show that the asymptotic distance is $(d/2)(\log n)+c$, where d is the dimension of the parameter vector. Therefore, the relative entropy rate D_n/n converges to zero at rate $(\log n)/n$. The constant c, which we explicitly identify, depends only on the prior density function and the Fisher information matrix evaluated at the true parameter value. Consequences are given for density estimation, universal data compression, composite hypothesis testing, and stock-market portfolio selection.**

I. Introduction

THE RELATIVE entropy is a mathematical expression that admits several different interpretations in information theory and statistics. These include the redundancy in source coding problems, the risk in statistical estimation, and the error exponents in hypothesis testing, among others. The general form of the relative entropy is an expectation of the logarithm of a density ratio that assesses how different the two densities are. Here, we will examine a particular form of the relative entropy, which arises in a Bayesian setting; we have a parametric family of random variables, and the parameter vector is assigned a prior distribution. We then ask how closely the Bayesian distribution for the data, which is also called the mixture of the distributions, approximates a member of the family we take as being true.

We characterize the asymptotic behavior of the relative entropy between the n-fold product of a given member of a parametrized family of distributions, say $P_{\theta_o}^n$, and a mixture of products of such distributions, which we denote by M_n. For smoothly parameterized families and for priors that assign positive mass to neighborhoods of θ_o, we show that the relative entropy increases in proportion to the logarithm of the sample size plus a constant, which we identify. We note that if the mixture excludes a neighborhood of the true density, then the behavior of the relative entropy is, asymptotically, of the order of the sample size; in addition, if the prior is discrete and assigns positive mass at θ_o, the relative entropy then asymptotically tends to a constant.

The relative entropy rate between the true distribution and the mixture of distributions has been examined by Barron [4]. It is shown that if the prior assigns positive mass to the relative entropy neighborhoods $\{\theta: D(P_{\theta_o}\|P_\theta) < \epsilon\}$, $\epsilon > 0$, then

$$\lim_{n\to\infty} \frac{1}{n} D\left(P_{\theta_o}^n \| M_n\right) = 0 \tag{1.1}$$

where D denotes the relative entropy. Thus, for large sample sizes, the Bayesian distribution M_n, which we know, is not far from the true distribution $P_{\theta_o}^n$, which is unknown. In [5], the condition on relative entropy neighborhoods is seen to be applicable even in some infinite dimensional settings. In the present paper, we use smoothness assumptions in the finite dimensional setting to assess the rate of convergence. To motivate the division by n in the relative entropy rate, note that for any two distinct product measures $(1/n)D(P_{\theta_o}^n\|P_\theta^n) = D(P_{\theta_o}\|P_\theta)$ remains fixed away from zero, which is in contrast to the behavior exhibited in (1.1).

Formally, we consider a parametrized family of distributions $\{P_\theta: \theta \in \Theta\}$ on a measurable space with $\Theta \subset \boldsymbol{R}^d$, and assume that $X_1, \cdots, X_n$ are independent and identically distributed random variables with respect to the distribution P_{θ_o}, where θ_o is a point in the interior of Θ. The probability measures P_θ are assumed to have probability density functions $p_\theta(x)$, with respect to a fixed sigma-finite measure $\lambda(dx)$. We denote the outcomes of X_n by x_n and a sequence of n random variables is denoted X^n with outcomes x^n.

Let $w(\theta)$ be the prior density for θ with respect to Lebesgue measure. The Bayesian marginal density function for X^n with respect to λ^n is the mixture of the conditional densities $p^n(x^n|\theta) = \prod_{i=1}^n p(x_i|\theta)$ obtained by integrating with respect to the prior, i.e.,

$$m_n(x^n) = \int_\Theta w(\theta) p^n(x^n|\theta)\, d\theta. \tag{1.2}$$

We denote the mixture distribution itself by M_n and use the notations $p_\theta(x)$ and $p(x|\theta)$ interchangeably. We omit

Manuscript received August 13, 1988; revised September 13, 1989. This work was supported in part by the Office of Naval Research under grant N00014-86-K-0670 and the Joint Services Electronics Program, contract N00014-84-C-0149.

B. S. Clarke was with the Department of Statistics, University of Illinois, Urbana-Champaign, IL. He is now with the Department of Statistics, Purdue University, West Lafayette, IN.

A. R. Barron is with the Department of Statistics, the Department of Electrical and Computer Engineering, the Coordinated Science Laboratory, and the Beckman Institute at the University of Illinois, Urb na, Champaign, IL.

IEEE Log Number 8933983.

0018-9448/90/0500-0453$01.00

n as a superscript or superscript on the joint densities p^n or m_n when the meaning is clear from the context, as it is in the expressions $p(x^n|\theta)$ and $m(x^n)$. Note that although X^n is a sample of n independent and identically distributed random variables under P_θ^n, under M_n, they are, in general, no longer independent.

The relative entropy, also called the Kullback–Leibler distance or informational divergence, is defined to be

$$D(P\|Q) = E_P \log \frac{p(X)}{q(X)}$$

for probability measures P, Q having densities p, q with respect to λ (see [37]). Except where specifically indicated otherwise, we let log denote the natural logarithm. Equivalently, the relative entropy can be denoted by $D(p\|q)$. We refer to D as a distance between probability densities even though it is not a metric. Similar measures of divergence have been proposed and studied (see Csiszár [20]); however, for the applications discussed here, D is the appropriate one. The focus of our interest is the relative entropy between $P_{\theta_o}^n$ and M_n:

$$D_n = D(P_{\theta_o}^n \| M_n). \tag{1.3}$$

We identify the asymptotic behavior of the relative entropy. Sufficient conditions are given such that

$$D(P_{\theta_o}^n \| M_n) = \frac{d}{2}\log\frac{n}{2\pi e} + \frac{1}{2}\log\det I(\theta_o) + \log\frac{1}{w(\theta_o)} + o(1) \tag{1.4}$$

where $I(\theta_o)$ is the Fisher information matrix. Therefore, the divergence of the Bayesian and frequentist distributions is precisely characterized. Although $D(P_{\theta_o}^n\|M_n)$ tends to infinity, the divergence per sample $(1/n)D(P_{\theta_o}^n\|M_n)$ tends to zero at rate $O((\log n)/n)$.

The form of the result (1.4) may be conjectured from the asymptotic normality of the posterior density. Indeed, we can motivate the result by writing the decomposition

$$D(P_{\theta_o}^n\|M_n) = E_{\theta_o}\log\frac{p(X^n|\hat\theta)}{m(X^n)} + E_{\theta_o}\log\frac{p(X^n|\theta_o)}{p(X^n|\hat\theta)} \tag{1.5}$$

where $\hat\theta$ is the maximum-likelihood estimator, and examining its terms.

The first term on the right side of (1.5) is the dominant term. It is the expected logarithm of a quantity related to the posterior density function for θ given X^n evaluated at $\hat\theta$, which is $w(\hat\theta)p(X^n|\hat\theta)/m(X^n)$. With high probability, the posterior density can be well approximated by a $N(\hat\theta, (nI(\hat\theta))^{-1})$ density under suitable technical conditions (see Walker [58], Le Cam [41], Bickel and Yahav [11], and Hartigan [25]). Since we want to approximate the expected value in (1.5) and not just show convergence in probability, new difficulties are introduced. Nevertheless, evaluation of the normal density suggests that the approximation to the first term on the right side of (1.5) should be

$$\log\left((2\pi)^{-d/2}\det(nI(\theta_o))^{1/2}/w(\theta_o)\right).$$

Pulling the factor n outside of the determinant, it is seen that the $(d/2)\log n$ behavior is captured by this approximation.

The second term on the right side of (1.5) is the expected value of the negative of the log-likelihood ratio test statistic $\log p(X^n|\hat\theta)/p(X^n|\theta_o)$, which by the theory of Wilks [59], Wald [56], and Chernoff [15] is known to have the asymptotic distribution of one-half a chi-square random variable with d degrees of freedom under suitable technical conditions. Thus, it is natural to conjecture that the expected value in the second term of (1.5) converges to $-d/2$. Combining the two terms leads to the approximation

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

To rigorously show that this expression is a valid approximation to D_n requires, by the previous method, that additional conditions be imposed to ensure the consistency of the maximum-likelihood estimator and to ensure that the limits can be taken in expectation as well as in probability. This can be done as in [19]. The method we give below is similar but avoids the use of the maximum-likelihood estimator to reduce the set of assumptions.

We remark that similar pointwise approximations to $\log p(X^n|\hat\theta)/m(X^n)$ are obtained by Leonard [43] as a consequence of theory in De Groot [23], by Tierney and Kadane [52], [53] as an application of Laplace's method of integration, by Rissanen [47] in the context of his stochastic complexity criterion, and by Schwarz [49] and Haughton [26] for parametric families of the general exponential or Koopman–Darmois form. An approximation to $D(P_{\theta_o}^n\|M_n)$ of order $(d/2)\log n$ is obtained in the context of universal source coding by Krichevsky and Trofimov [36] for the special case of Dirichlet mixtures of finite alphabet distributions. Rissanen [46] shows that for smooth families $\{P_\theta\}$ and for any distribution Q_n (not just those that are obtained as mixtures), $D(P_\theta^n\|Q_n)$ cannot be of smaller order than $(d/2)(1-o(1))\log n$ except for θ in a set of Lebesgue measure zero.

A different interpretation of (1.1) in the context of information and ergodic theory comes from the work of Kieffer [34], [35], who shows that the relative entropy rate between two stationary processes $\lim(1/n)D(P_{X^n}\|Q_{X^n})$ exists when the second measure Q is independent identically distributed (i.i.d.) or Markov, but by counterexample, the limit need not exist for certain non-Markov Q. The measures M that we examine provide examples of non-Markov measures (in fact exchangeable measures) for which the relative entropy rate does exist.

The proof of a key lemma in Section IV uses a technique for approximating integrals first introduced by Laplace in 1774 (published in Laplace [39] and translated by Stigler [50]). In the case first considered by Laplace, the integral $\int p_\theta(x^n)w(\theta)\,d\theta$ was approximated where p_θ is

the Bernoulli (θ) model and $w(\theta)$ is the uniform prior on $[0,1]$. Laplace's approximation for integrals is now a standard technique in analysis. Walker [58], and Tierney and Kadane [52], [53] provide two examples, and some general theory is presented by De Bruijn [22].

A byproduct of the analysis is the following asymptotics for the logarithm of the density ratio

$$\log\frac{p(X^n|\theta_o)}{m(X^n)} = \frac{d}{2}\log\frac{n}{2\pi} + \frac{1}{2}\log\det I(\theta_o) + \log\frac{1}{w(\theta_o)} - \frac{1}{2}S_n^T(I(\theta_o))^{-1}S_n + o(1) \quad (1.6)$$

where $o(1)\to 0$ in $L^1(P)$ as well as in probability as $n\to\infty$. Here, $S_n = (1/\sqrt{n})\nabla \log p(X^n|\theta_o)$ is the standardized score function for which $ES_nS_n^T = I(\theta_o)$ and $ES_n^T(I(\theta_o))^{-1}S_n = d$. Moreover, it is seen that

$$\log\frac{m(X_1,\cdots,X_n)}{p(X_1,\cdots,X_n|\theta_o)} + D(P_{\theta_o}^n\|M_n) \quad (1.7)$$

converges, in distribution, to $1/2(\chi_d^2 - d)$ where χ_d^2 has a chi-square distribution with d degrees of freedom.

In proving the main results, we assume that the posterior distribution concentrates on neighborhoods of the true value of the parameter at a fast enough rate. To permit easy verification of this hypothesis, we have found sufficient conditions that are natural in finite-dimensional families based on the work of Schwartz [48]. We do this by introducing a property that we call the soundness of the parametrization. By definition, a parametrization is sound if the convergence of a sequence of parameter values in the Euclidean norm is equivalent to the weak convergence of the distributions they index. It is shown that for smooth families, soundness implies convergence of the posterior distribution at the required rate.

In Section II, we state our main results and consider some examples. We discuss applications of the main result in Section III. It is seen that $D(P_\theta^n\|M_n)$ is a) the cumulative risk of Bayes' estimators of the density function, b) the redundancy of a source code based on M_n, c) the exponent of error probability for Bayes' tests of a simple versus composite hypothesis, and d) a bound on the financial loss in a stock-market portfolio selection problem. In Section IV, we formally prove our result. There are two key hypotheses: One is that the second derivative of the log likelihood is locally dominated by a function with finite expected square, and the other is that the posterior distribution concentrates on neighborhoods of the true value of the parameter at a fast enough rate. In a concluding section, we prove the consistency of the posterior distribution for soundly parametrized families.

II. Statement of Results

In the approximations we seek, the only quantities that appear are the information matrix and the prior density at the true value. This suggests that ideally, the only conditions that should be introduced are those that will control these quantities.

The behavior of the Fisher information can be controlled, for present purposes, by assuming that in a neighborhood of the true parameter value, the second derivative of the logarithm of the density exists, is dominated by a function with finite expected square, and that the second derivative is continuous at the true parameter value.

Condition 1: The density $p_\theta(x)$ is twice continuously differentiable at θ_o for almost every x, and there exists a $\delta > 0$ so that for each j and k from 1 to d

$$E\sup_{\{\|\theta-\theta_o\|<\delta\}}\left|\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log p(X|\theta)\right|^2 < \infty \quad (2.1)$$

and

$$E\left|\frac{\partial}{\partial\theta_j}\log p(X|\theta_o)\right|^2 < \infty.$$

We adopt the convention that, except where noted otherwise, E denotes expectation with respect to the true probability density p. For now, this density $p = p_{\theta_o}$ is assumed to be a member of the given family. In extensions of the theory, as will be developed later, p_{θ_o} is the density in the family closest to p in the relative entropy sense.

There are two information matrices that typically coincide and have a basic role in the analysis. These are the Fisher information

$$I(\theta_o) = E\left[\frac{\partial}{\partial\theta_j}\log p(X|\theta_o)\frac{\partial}{\partial\theta_k}\log p(X|\theta_o)\right]_{j,k=1\cdots d} \quad (2.2)$$

and the second derivative matrix for the informational divergence

$$J(\theta_o) = \left[\frac{\partial^2}{\partial\theta_j\partial\theta_k}D(p\|p_{\theta_o})\right]_{j,k=1\cdots d} \quad (2.3)$$

where in each case the derivatives are evaluated at $\theta = \theta_o$. When convenient, we also use the subscript notation I_{θ_o} and J_{θ_o}, respectively.

Since the desired expression involves the logarithm of a determinant of an information matrix and the logarithm of the prior density, it is natural to require positivity of the information matrix and the prior.

Condition 2: $D(p\|p_\theta)$ is twice continuously differentiable at θ_o, with $J(\theta_o)$ positive definite, and the prior $w(\theta)$ is continuous and positive at θ_o.

To see the relationship between the two information matrices, we note that when Condition 1 is satisfied, the relative entropy in (2.3) is twice continuously differentiable, and $J(\theta_o)$ is seen to equal the matrix with entries $-E\partial^2(\log p(X|\theta_o))/\partial\theta_j\partial\theta_k$. This is the same as the Fisher information $I(\theta_o)$ when the true density is equal to p_{θ_o}, provided that derivatives with respect to θ of the equation $\int p_\theta(x) = 1$ can be brought inside the integral to yield

$\int \partial^2 p_{\theta_o}(x)/\partial\theta_j\partial\theta_k = 0$ (see, e.g., Lemma 2.6.1 in Lehmann [42]).

Note that both Condition 1 and Condition 2 are local assumptions depending only on the behavior of the family for θ near θ_o. A third condition is also required on the consistency of the posterior distribution. As is shown in Theorem 2.2, this posterior consistency is satisfied when the only P_θ's near P_{θ_o} are those for which θ is near θ_o.

Condition 3: The posterior distribution of θ given X^n asymptotically concentrates on neighborhoods of θ_o, except for X^n in a set of probability of order $o(1/\log n)$, i.e., $P^n\{W(N^c|X^n) > \delta\} = o(1/\log n)$ for every open set N containing θ_o and every $\delta > 0$, where $W(\cdot|X^n)$ is the posterior distribution of θ given X^n.

The main result is the following theorem. It is proved in Section IV.

Theorem 2.1: Suppose the parametric family $\{p_\theta\}$ and the prior $w(\theta)$ satisfy the smoothness Conditions 1 and 2 and the posterior consistency Condition 3 for θ_o in the interior of Θ. Then

$$\lim_{n\to\infty}\left(D(P^n_{\theta_o}\|M_n) - \frac{d}{2}\log\frac{n}{2\pi}\right) = \log\frac{1}{w(\theta_o)} + \frac{1}{2}\log\det J(\theta_0) - \frac{1}{2}\operatorname{tr}\left(I_{\theta_o}J^{-1}_{\theta_o}\right) \quad (2.4)$$

and moreover, the following limit holds in $L^1(P)$ and hence in probability

$$\lim_{n\to\infty}\left(\log\frac{p_{\theta_o}(X^n)}{m(X^n)} + \frac{1}{2}S_n^T J^{-1}_{\theta_o} S_n - \frac{d}{2}\log\frac{n}{2\pi}\right) = \log\frac{1}{w(\theta_o)} + \frac{1}{2}\log\det J(\theta_o) \quad (2.5)$$

where $S_n = (1/\sqrt{n})\nabla \log p(X^n|\theta_o)$. If $I(\theta_o) = J(\theta_o)$ as well, we have convergence of the expectation as in (1.4), convergence in $L^1(P)$ as in (1.6), and convergence in distribution as in (1.7). If Conditions 1 and 2, but not the posterior consistency, are satisfied, (2.4) and (2.5) are upper bounds on the limit superior of the respective sequences.

Next, we examine the consistency of the posterior distribution as required in Condition 3. The assumption of posterior consistency is more natural for the analysis of Bayesian methods than is the assumption of the consistency of the maximum-likelihood estimator, as was used in [19]. Moreover, in Theorem 2.2, we see that when Condition 2 is satisfied, Bayes' consistency holds under a hypothesis that is much weaker than the conditions for the consistency of the maximum-likelihood estimator due to Wald [57].

For the following definition, it is assumed that the Borel space X on which the probability distributions P_θ reside is a separable metric space.

Definition: A parametric family of distributions is sound if the convergence of a sequence of parameter values is equivalent to the weak convergence of the distributions they index, i.e.,

$$\theta \to \theta_o \Leftrightarrow P_\theta \to P_{\theta_o}.$$

Soundness is an identifiability condition that makes it impossible for a family to fold back on itself: no θ far from θ_o corresponds to a P_θ close to P_{θ_o}. When a parameterization is one to one (i.e., $\theta \neq \theta_o$ implies $P_\theta \neq P_{\theta_o}$) and continuous (i.e., $\theta \to \theta_o$ implies $P_\theta \to P_{\theta_o}$), then soundness is automatically satisfied on each compact subset of the parameter space (because one-to-one and continuous mappings on a compact set have a continuous inverse). Therefore, for one-to-one and continuous parameterizations, to check for soundness in noncompact cases, it is enough to check that for sequences θ that diverge from the parameter set, the measures P_θ do not converge to a member of the family. Continuity of the parameterization at θ_o is seen to be a consequence of continuity of $D(P_{\theta_o}\|P_\theta)$, which is assumed in Condition 2.

The following result, which is proved in Section VI, shows that soundness in conjunction with a local smoothness assumption is sufficient for the consistency of the posterior distribution.

Theorem 2.2: If a parametric family of distributions on a separable metric space is sound and if the smoothness Condition 2 is satisfied at θ_o, then the posterior consistency Condition 3 is satisfied. Moreover, for every neighborhood N of θ_o, there exists $r > 0$ such that

$$P^n\{W(N^c|X^n) > e^{-nr}\} \le O\left(\frac{1}{n}\right). \quad (2.6)$$

To provide a class of examples, we indicate that finite-dimensional exponential families satisfy the hypotheses of Theorems 2.1 and 2.2. Consider families of probability densities of the exponential form $e^{-\theta^T\phi(x)}g(x)/c(\theta)$ with the natural parameter space $\Theta = \{\theta \in \mathbf{R}^d: c(\theta) < \infty\}$, where $c(\theta) = \int e^{-\theta^T\phi(x)}g(x)\lambda(dx)$ is the normalizing constant. The function $g(x)$ and the dominating measure $\lambda(dx)$ are arbitrary. We assume that the vector-valued function $\phi(x)$ is such that $\theta^T\phi(x)$ is a nonconstant function, unless $\theta = 0$; therefore, the dimension of the family cannot be reduced. To verify Conditions 1 and 2, note that the derivatives

$$\partial^2(\log p(x|\theta))/\partial\theta_j\partial\theta_k = -\partial^2(\log c(\theta))/\partial\theta_j\partial\theta_k$$

are independent of x, and $\log c(\theta)$ is twice continuously differentiable and strictly convex (so the second derivative matrix is positive definite) at points in the interior of Θ, as is shown, for instance, in Brown [13]. The posterior consistency condition also holds: Berk [10] showed that $P^n\{W(N^c|X^n) > \delta\}$ converges at an exponential rate. In addition, the soundness condition can be verified as in Section VI. Indeed, a type of degeneracy occurs for any sequence of θ's that diverges from the family. In Clarke [18], the approximation to $D(P^n_{\theta_o}\|M_n)$ is worked out in detail for several specific examples.

Returning to the general context, the next theorem uses a more elaborate argument to obtain the limit supe-

rior half of the result in Theorem 2.1 under weaker conditions. It is enough that Condition 2 be satisfied and that $\log p_\theta(X)$ be mean-square differentiable at θ_o. In addition, Condition 2 alone is enough for the limit superior of $(D(P_{\theta_o}^n\|M_n)-(d/2)\log n)$ to be bounded by a constant, but the constant is somewhat larger than identified in Theorem 2.1.

We also obtain an extension of the upper bounds to the case that the true density p is not necessarily in the family $\{p_\theta\}$, and $D(p\|p_\theta)$ is assumed to be minimized at a point $\theta_o \in \text{int}(\Theta)$. In this case, $D(P^n\|M_n)$ typically grows at rate n.

We continue to assume that Condition 2 is satisfied. Note that the twice continuous differentiability of $D(p\|p_\theta)$ and the minimization of it at θ_o implies that the gradient is zero, and the second derivative matrix $J(\theta_o)$ is nonnegative definite at θ_o. Positive definiteness ensures that θ_o is an isolated minimum, i.e., there are no other minima in a neighborhood of θ_o.

In place of Condition 1, we use the following weaker hypothesis.

Condition 4: $\log p_\theta(X)$ is mean-square differentiable at θ_o, that is, there exists a vector-valued function $i_{\theta_o}(X)$ with $E\|i_{\theta_o}(X)\|^2 < \infty$ such that

$$\left(E\left(\log p_\theta(X)/p_{\theta_o}(X)-(\theta-\theta_o)^T i_{\theta_o}(X)\right)^2\right)^{1/2} = o(\|\theta-\theta_o\|). \tag{2.7}$$

The expectation is taken with respect to the true density p.

The mean-square derivative $i_{\theta_o}(X)$ is called the score function, and the Fisher information matrix I_{θ_o} is defined in this more general context as

$$I_{\theta_o} = E\left(i_{\theta_o}(X) i_{\theta_o}(X)^T\right). \tag{2.8}$$

Mean-square differentiability is weaker than (in particular, it is implied by) the assumption of pointwise differentiability with the norm square of the derivative that is locally dominated by a function of finite expectation.

For an example of a family satisfying the mean-square differentiability but not the pointwise differentiability, consider the two-sided exponential $p(x|\theta)=(1/2)e^{-|x-\theta|}$ like that in Pollard [44]. Take the true density to be in this family with, for convenience, $\theta_o = 0$. The pointwise derivative of $\log p(x|\theta)$ does not exist at $\theta = x$. Consequently, the differentiability at θ in a neighborhood of zero (required for Condition 1) does not hold for x near zero. Nevertheless, $\log p(x|\theta)$ is mean-square differentiable with $i_{\theta_o}(x) = -\text{sgn}(x-\theta_o)$ and $I_{\theta_o}=1$; in addition, $D(p\|p_\theta = e^{-|\theta|}+|\theta|-1$, which has the second-order expansion $(1/2)\theta^2 + o(\theta^2)$; therefore, $D(p\|p_\theta)$ is twice continuously differentiable with $J_{\theta_o}=1$, and Conditions 2 and 4 are verified.

We note that although the information matrices I_{θ_o} and J_{θ_o} are typically the same when the true density p is in the family (see the remark after the statement of Condition 2), they are generally not the same for p outside the family.

The relative entropy has the decomposition

$$D(P^n\|M_n) = nD(P\|P_{\theta_o}) + E\log\frac{p_{\theta_o}(X^n)}{m(X^n)}. \tag{2.9}$$

The following theorem provides a bound on the second term of order $\log n$. When $p = p_{\theta_o}$ is in the family,

$$E\log p_{\theta_o}(X^n)/m(X^n)$$

is the same as $D(P_{\theta_o}^n\|M_n)$, and the following provides weaker assumptions for asymptotic upper bounds of the sequences in Theorem 2.1.

Theorem 2.3: If Condition 2 is satisfied, then

$$\limsup_{n\to\infty}\left(E\log\frac{p_{\theta_o}(X^n)}{m(X^n)}-\frac{d}{2}\log\frac{n}{2\pi}\right) \le \log\frac{1}{w(\theta_o)}+\frac{1}{2}\log\det(J_{\theta_o}). \tag{2.10}$$

If $\log p_\theta(X)$ is also mean-square differentiable at θ_o (Condition 4) then

$$\log\frac{p_{\theta_o}(X^n)}{m(X^n)}-\frac{d}{2}\log\frac{n}{2\pi} \le \log\frac{1}{w(\theta_o)}+\frac{1}{2}\log\det(J_{\theta_o})-\frac{1}{2}S_n^T J_{\theta_o}^{-1}S_n+o(1) \tag{2.11}$$

where $o(1)$ tends to zero in $L^1(P)$ as $n\to\infty$, and consequently

$$\limsup_{n\to\infty}\left(E\log\frac{p_{\theta_o}(X^n)}{m(X^n)}-\frac{d}{2}\log\frac{n}{2\pi}\right) \le \log\frac{1}{w(\theta_o)}+\frac{1}{2}\log\det(J_{\theta_o})-\frac{1}{2}\text{tr}\left(I_{\theta_o}J_{\theta_o}^{-1}\right). \tag{2.12}$$

Here $S_n = (1/\sqrt{n})\sum_{i=1}^n i_{\theta_o}(X_i)$ is the standardized score function, which by the central limit theorem, is asymptotically distributed as $N(0, I_{\theta_o})$.

Note that the posterior consistency condition is not needed for the upper bounds in Theorems 2.1 and 2.3. This is because the mixture $m(X^n)=\int_\Theta p(X^n|\theta)w(\theta)\,d\theta$ is reduced to $m_\delta(X^n)=\int_{N_\delta} p(X^n|\theta)w(\theta)\,d\theta$ when obtaining these bounds, where N_δ is a neighborhood of θ_o. However, if the posterior distribution is not consistent, $E\log m_\delta(X^n)/m(X^n)$, which is the expected logarithm of the posterior probability of the neighborhood of θ_o, does not tend to zero for some $\delta>0$, and a nonzero gap exists in the limit of the difference $E\log p_{\theta_o}(X^n)/m_\delta(X^n) - E\log p_{\theta_o}(X^n)/m(X^n)$. In this way, it is seen that posterior consistency is necessary for the limit in Theorem 2.1.

Our final conclusion gives a strengthened form of the Bayesian central limit theorem on the asymptotic normality of the posterior distribution, which is shown to be equivalent to our L^1 convergence result in Theorem 2.1. Let $T=\sqrt{n}(\theta-\theta_o)$ for θ distributed according to w,

which has the posterior density function

$$w_T(t|X^n) = \frac{(1/\sqrt{n})^d w(\theta_o + t/\sqrt{n})\, p(X^n|\theta_o + t/\sqrt{n})}{m(X^n)} \tag{2.13}$$

and let $\phi_n(t)$ be the normal density with mean $J_{\theta_o}^{-1}S_n$ and covariance $J_{\theta_o}^{-1}$.

Theorem 2.4: Assume that Conditions 2 and 4 are satisfied. The convergence given in (2.5) is then equivalent to the following. For every t in $\boldsymbol{R}^d$, the difference in the logarithms of the posterior density and the normal density converges to zero in $L^1(P)$, i.e.

$$\lim_{n\to\infty} E|\log w_T(t|X^n) - \log \phi_n(t)| = 0. \tag{2.14}$$

In particular, this convergence (2.14) is implied by Conditions 1–3.

This shows that the posterior distribution of θ is approximately normal with mean $\hat{\theta} = \theta_o + (1/\sqrt{n})J_{\theta_o}^{-1}S_n$ and covariance $(nJ_{\theta_o})^{-1}$. Related results, showing convergence in probability rather than convergence of the logarithm in L^1 are given on p. 111 of Hartigan [25] and on p. 456 of Lehmann [42].

In the applications we develop below, it is the information-theoretic asymptotics, i.e., the asymptotics of $D(P_{\theta_o}^n\|M_n)$, that we use most directly, rather than the asymptotic normality of the posterior.

Further extensions of the theory, showing that the approximation to $D(P_\theta^n\|M_n)$ in Theorem 2.1 holds uniformly on compact subsets of the interior of Θ and giving conditions such that the approximation can be averaged with respect to the prior to yield an approximation to $\int_\Theta w(\theta)D(P_\theta^n\|M_n)\,d\theta$, are obtained in Clarke [18] (in the case that $I_\theta = J_\theta$) and will be developed in a subsequent paper by the authors. As shown in [18], a consequence of these extensions is that the prior $w(\theta)$, which is proportional to $\det(I_\theta)^{1/2}$, leads to a minimax value of the relative entropy.

III. Applications

We consider implications of the asymptotics of $D(P_{\theta_o}^n\|M_n)$ for density estimation, universal data compression, tests of composite hypotheses, and stock-market portfolio selection.

For simplicity in these applications, we focus on the case that the true density is in the given parametric family and that the two information matrices coincide. Consequences may also be formulated in the more general context.

A. Implications for Density Estimation

The relative entropy has several mathematical properties that make it a natural choice as a loss function in a decision-theoretic framework for the estimation of a density function. Chief amongst these are the following: It induces convex neighborhoods; it satisfies Pythagorean relations even though it is not a metric; it satisfies a chain rule expansion for densities of jointly distributed random variables; in smooth parametric families, it locally approximates squared error loss; it is nonnegative and it equals zero only when its arguments are equal.

Suppose we are in the case described earlier. In particular, 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 to identify the true density, which is p_{θ_o}. One natural estimator of $p(x|\theta_o)$ at any given x is the mixture of the densities with respect to the posterior distribution

$$\hat{p}_n(x; X^n) = \int_\Theta p_\theta(x) w(\theta|X^n)\, d\theta \tag{3.1}$$

that is, the posterior mean of the density. Observe that this estimator is the predictive density

$$\hat{p}_n(x) = m(X_{n+1} = x|X^n)$$

where $m(X_{n+1}|X^n)$ is the conditional density of X_{n+1} given X^n according to the Bayesian model.

We use the relative entropy as the loss function for parametric density estimation and examine the behavior of the cumulative risk. Let δ_k for $k = 0, \cdots, n-1$ be a sequence of density estimators. Each δ_k estimates the density of X_{k+1}, given the data X^k. Here, δ_0 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

$$E_{\theta_o} D(p_{\theta_o}\|\delta_k).$$

We denote the cumulative risk of n uses of an estimator δ_k for $k = 0, \cdots, n-1$ 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 (relative entropy) risks plays an important role in the other applications as well (see, e.g., case D forthcoming). It is natural to expect that the individual risks $E_{\theta_o} D(p_{\theta_o}\|\delta_k)$ could be made to be of order $1/k$, and hence, the cumulative risk would be of order $\log n$. We obtain an order $\log n$ result for the cumulative risk of the Bayes' estimator.

Just as the posterior mean of θ is the Bayes' estimator under squared-error loss, it turns out that the posterior mean of $p(x|\theta)$ is the Bayes' estimator under relative entropy loss. We have the following result.

Proposition 3.A: For each n, the estimator $\hat{p}_n$ defined as in (3.1) is the Bayes' estimator of the density function. Moreover, the cumulative risk of this sequence of estimators is

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

under the convention that $\hat{p}_0(x) = m_1(x_1)$. Consequently,

under the conditions of Theorem 2.1, the cumulative risk is approximated by $(d/2)(\log n)+c$, and the average risk $(1/n)\sum E_{\theta_o} D(p_{\theta_o}\|\hat{p}_k)$ converges to zero at rate $(\log n)/n$.

Proof: The information inequality, $D(p\|q)\geq 0$, with equality if and only if $p=q$, implies that $\hat{p}_n$ is the Bayes' estimator because for any other density q, the posterior average of the risk is seen to equal

$$\int_\Theta D(p_\theta\|q)w(\theta|X^n)\,d\theta = \int_\Theta D(p_\theta\|\hat{p}_n)w(\theta|X^n)\,d\theta + D(\hat{p}_n\|q).$$

We see that the minimum is achieved when the second term is zero, i.e., when $q=\hat{p}_n$. (A similar characterization of the posterier mean density $\hat{p}_n$ is given in Aitchison [1].)

By Bayes' rule, $\hat{p}_n$ equals the predictive density, which is

$$m(X_{n+1}=x_{n+1}|X^n)=\frac{m_{n+1}(X^n,x_{n+1})}{m_n(X^n)}.$$

By the chain rule for the relative entropy, we have that

$$D(P_{\theta_o}^n\|M_n)=\sum_{k=0}^{n-1} ED(P_{\theta_o}\|\hat{P}_k) \tag{3.2}$$

where each summand is the risk in estimating the density using the Bayes estimate based on k observations. □

We remark that under the conditions of Theorem 2.1, the individual risk terms $ED(P_{\theta_o}\|\hat{P}_n)$ also converge to zero as $n\to\infty$. This follows from noting that

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

and applying Theorem 2.1 to each term on the right side. Thus, the predictive density is a consistent estimator of the true density in expected relative entropy.

We note that on p. 434 of Ĉencov [14], the author gives conditions such that for the maximum-likelihood density $P_{\hat{\theta}}$, the risk $ED(P_{\theta_o}\|P_{\hat{\theta}})$ is of order $d/(2n)+O(1/n)^{3/2}$ (moreover, he demonstrates the optimality of this rate). Summing these individual risks also yields a cumulative risk of order $(d/2)\log n$.

Parameter estimation can be regarded as a special case of density estimation in which the estimator of the density is restricted to be of the form $p(x|\hat{\theta})$. In the density estimation context that we consider, the estimated density is not restricted to be in the family. By enlarging the class of estimators in this way, the statistical risk can be reduced. In particular, the Bayes' risk in parametric density estimation lower bounds the Bayes' risk in parameter estimation, i.e., for every prior

$$\inf_{\delta'}\int E_\theta D(\theta\|\delta')w(\theta)\,d\theta \geq \inf_{\delta}\int E_\theta D(P_\theta\|\delta)w(\theta)\,d\theta \tag{3.3}$$

where the infimum on the left side is over parameter estimators δ' with loss function $D(\theta\|\delta')=D(P_\theta\|P_{\delta'})$, and the infimum on the right side is over density estimators δ.

B. Applications to Universal Source Coding

Suppose that X is a discrete random variable whose distribution is in the parametric family $\{P_\theta\colon \theta\in\Theta\}$, and we want to encode a block of data for transmission. It is known that a lower bound on the expected codeword length is the entropy of the distribution. Moreover, this entropy bound can be achieved, within one bit, when the distribution is known. Universal codes have expected length near the entropy no matter which member of the parametric family is true. The redundancy of a code is defined to be the difference between its expected length and the entropy.

Universal noiseless source coding for parametric families of distributions was introduced by Davisson [21]. Variable-length binary codes are assigned to blocks of data $X^n=(X_1,\cdots,X_n)$. Let $\{0,1\}^*$ denote the set of finite-length binary strings. Recall that by the Kraft–McMillan theorem (see, e.g., p. 50 of Blahut [12]) if

$$\phi\colon X^n\to\{0,1\}^*$$

is a uniquely decodable code, and $l(\phi(X^n))$ is the length of the codewords, then

$$Q_n(X^n)=2^{-l(\phi(X^n))}$$

defines a subprobability mass function on X^n. Moreover, for any subprobability mass function $Q_n(X^n)$ for which $-\log Q_n(X^n)$ takes integer values, a uniquely decodable code exists with those lengths. The redundancy is the difference between the expected value of the length of the codewords $\phi(X^n)$, and the expected value of the idealized length $\log 1/P_{\theta_o}(X^n)$, that is

$$\begin{aligned} R_n(\phi,P_{\theta_o}) &= E\left[l(\phi(X^n))-\log\left(\frac{1}{P_{\theta_o}(X^n)}\right)\right] \\ &= E\left[\log\left(\frac{1}{Q_n(X^n)}\right)-\log\left(\frac{1}{P_{\theta_o}(X^n)}\right)\right] \\ &= D(P_{\theta_o}^n\|Q_n) \end{aligned} \tag{3.4}$$

where the logarithm is taken base 2. Thus, the redundancy is the relative entropy. We want to choose the lengths to make the redundancy small for each P_θ without advance knowledge of the true distribution in the family. Among all subprobability mass functions Q, the one that minimizes the average of $D(P_\theta^n\|Q_n)$ with respect to a prior $w(\theta)$ is the mixture M_n. Thus, $D(P_{\theta_o}^n\|M_n)$ is referred to as the redundancy of the Bayes' code. The idealized lengths $\log 1/M_n(X^n)$ may violate the constraint of being integer valued. Nevertheless, for Shannon code based on M_n, i.e., the code with lengths

$$l(\phi(X^n))=\left\lceil \log\frac{1}{M_n(X^n)}\right\rceil,$$

the redundancy is within one bit of $D(P_{\theta_o}^n\|M_n)$.

The concepts of noiseless source coding of discrete data may also be applied to the case of continuous random variables that are arbitrarily finely quantized. In the

sense made clear by the following proposition, the relative entropy remains the redundancy for nondiscrete sources. If a noiseless code is specified for every finite quantization of a nondiscrete source, we define the redundancy of that source to be the supremum of the redundancies over all such quantizations. In the case that the random variables are discrete, this reduces to the usual definition of redundancy.

Proposition 3.B: For any source, the redundancy of the Shannon code based on M_n is $D(P_{\theta_o}^n \| M_n)$ to within one bit. Thus, the redundancy of the Bayes code is given asymptotically by

$$\frac{d}{2}\log\frac{n}{2\pi e} + \frac{1}{2}\log\det I(\theta_o) - \log w(\theta_o)$$

under the conditions of Theorem 2.1.

Proof: For any finite partition Π of X^n, we can specify a code $\phi = \phi_{n,\Pi}$, by use of the Shannon code based on the probability measure restricted to Π. For the Shannon code, we have an explicit formula for the length of the codewords

$$l(\phi_{n,\Pi}(A)) = \left\lceil \log\frac{1}{Q_n(A)} \right\rceil, \quad A \in \Pi$$

and the redundancy is

$$R_{n,\Pi}(\phi, P_{\theta_o}) = \sum_{A\in\Pi} P_\theta^n(A)\left(l(\phi_n(A)) - \log\frac{1}{P_\theta^n(A)}\right).$$

Now $l(\phi_{n,\Pi}(A))$ is within one bit of $\log 1/Q_n(A)$. Therefore, for each partition, the redundancy $R_{n,\Pi}(\phi, P_{\theta_o})$ is within one bit of the discrete divergence

$$\Sigma_{A\in\Pi} P_\theta^n(A)\log P_\theta^n(A)/Q_n(A).$$

Taking the supremum over all possible partitions gives $D(P_\theta^n \| Q_n)$, by using a well-known theorem (see pp. 6–7 of Kullback *et al.*, [38]). If Q_n is replaced by M_n, we then get the Bayes code, and the result is the asymptotic least upper bound on the redundancy. □

In Rissanen [46], it is shown that for any code, $(d/2)\cdot \log n - o(\log n)$ is an asymptotic lower bound on the redundancy for (Lebesgue) almost every θ in the family (assuming that the maximum-likelihood estimator is consistent and asymptotically normal). In addition, Rissanen [45] showed that for particular codes based on his minimum description length criterion, a redundancy of order $(d/2)\log n + c_\theta$ is achieved, although he did not attempt to optimize the constant. For a discussion of the best constants in Rissanen's framework of two-stage codes, see [8]. The optimum code according to the criteria of minimaxity or minimum average redundancy is not a two-stage code of the type considered in [45] or [8], rather it is a one-stage code based on a mixture M_n, where the choice of prior in the mixture is determined by the criterion. Rissanen [47] also considers codes based on mixtures and shows that pointwise, the codelength $-\log m(X^n)$ is approximated by $-\log p(X^n|\hat{\theta}) + (d/2)\log n + O(1)$ uniformly in X^n, provided the empirical Fisher information and the logarithm of the prior density evaluated at the maximum-likelihood estimator remain uniformly bounded. Previous results of this type, with the empirical Fisher information and the logarithm of the prior density incorporated in an almost-sure approximation, are given in Theorems 4.3 and 4.4 in [3], together with the coding interpretations.

C. An Application to Hypothesis Testing

It is well-known that the likelihood ratio test statistic converges in distribution to $1/2$ times a chi-square random variable with d degrees of freedom, i.e.,

$$\log\frac{p(X^n|\hat{\theta})}{p(X^n|\theta_o)} \to \frac{1}{2}\chi_d^2$$

in law, where χ_d^2 is a chi-square random variable with d degrees of freedom (see Wilks [59], Wald [56], and Chernoff [15]). It has been proved that the asymptotic expected value of the likelihood-ratio statistic is essentially $d/2$ (see [19]). An analogous result requiring fewer hypotheses can be proved for the statistic $\log m(X^n)/p(X^n|\theta_o)$. We consider a centered version of this statistic obtained by subtracting its mean under the distribution $P_{\theta_o}^n$. As stated in (1.7)

$$\log\frac{m(X^n)}{p(X^n|\theta_o)} + D(P_{\theta_o}^n \| M_n) \to \frac{1}{2}(\chi_d^2 - d)$$

in distribution. Conditions for the validity of this asymptotic distribution are given in Theorem 2.1.

We use this convergence to identify the critical value and the average power of a test for composite hypotheses. Consider testing H: P_{θ_o} versus K: P_θ, $\theta \neq \theta_o$. We constrain the probability of a type-I error to be less than $\alpha_1 \in (0,1)$ and examine the performance of tests in terms of the probability of a type-II error averaged with respect to a prior density $w(\theta)$ over the class of alternatives K. Let $c(\alpha)$ be the $1-\alpha$ quantile of a centered chi-square random variable with d degrees of freedom, i.e., $P(\chi_d^2 - E\chi_d^2 > c) = \alpha$. The Bayes' optimal test is defined to minimize the average probability of error. By a familiar argument, the problem is seen to reduce to a simple versus simple test for $P_{\theta_o}^n$ versus M_n; therefore, the optimal test compares the test statistic $\log m(X^n)/p(X^n|\theta_o)$ to a critical value $t = t_n(\alpha_1)$. The following proposition shows us how to select the critical value in practice. Specifically, Theorem 2.1 gives a convenient approximation to it. Moreover, the average power of the test is shown to be related to $D(P_{\theta_o}^n \| M_n)$.

Proposition 3.C: Fix α_1 in $(0,1)$. Under the assumptions of Theorem 2.1, the Bayes' test with critical value $t = D(P_{\theta_o}^n \| M_n) - 1/2c(\alpha_1)$ has asymptotic level α_1, and the optimal average probability of a type-II error is, to within a constant factor dependent only on α_1

$$\alpha_2 \doteq e^{-D(P_{\theta_o}^n \| M_n)} \doteq \frac{n^{-d/2}(2\pi e)^{d/2} w(\theta_o)}{\sqrt{\det I(\theta_o)}}.$$

Indeed, there exists a finite interval $[L(\alpha_1), U(\alpha_1)]$ such that every test with a type-I error less than or equal to α_1 satisfies

$$\liminf_{n\to\infty}\left[\log\alpha_2 + D\left(P_{\theta_o}^n\|M_n\right)\right] \geq L(\alpha_1). \qquad (3.5)$$

There also exists a test with a type-I error α_1 for which the following upper bound holds:

$$\limsup_{n\to\infty}\left[\log\alpha_2 + D\left(P_{\theta_o}^n\|M_n\right)\right] \leq U(\alpha_1). \qquad (3.6)$$

The functions L and U can be expressed in terms of $c(\alpha)$.

Remark: This extends Stein's lemma (see [16], [2], or [12]) for simple versus simple hypotheses, say P_{θ_o} versus P_θ with $\theta \neq \theta_o$, which asserts that to first order in the exponent

$$\alpha_2 \doteq e^{-D(P_{\theta_o}^n\|P_\theta^n)}.$$

Proof: We first prove the lower bound statement (3.5). Let $\tilde{C}_n$ be any critical region with $P_{\theta_o}(\tilde{C}_n) \leq \alpha_1$, and let A_n be the "typical set"

$$A_n = \left\{x^n : \log\frac{p(x^n|\theta_o)}{m(x^n)} \leq D\left(P_{\theta_o}^n\|M_n\right) - \frac{1}{2}c(\alpha)\right\}$$

where $\alpha > \alpha_1$. Observe that

$$\lim_{n\to\infty} P_{\theta_o}^n(A_n) = \alpha.$$

Then, the average probability of a type-II error satisfies

$$\begin{aligned}\alpha_2 = M_n\left(\tilde{C}_n^c\right) &\geq M_n\left(\tilde{C}_n^c \cap A_n\right)\\ &\geq e^{-D(P_{\theta_o}^n\|M_n)+(1/2)c(\alpha)}P_{\theta_o}^n\left(\tilde{C}_n^c \cap A_n\right)\\ &\geq e^{-D(P_{\theta_o}^n\|M_n)+(1/2)c(\alpha)}\left[P_{\theta_o}^n\left(\tilde{C}_n^c\right) - P_{\theta_o}^n\left(A_n^c\right)\right].\end{aligned}$$

Since

$$\lim_{n\to\infty}\left[P_{\theta_o}^n\left(\tilde{C}_n^c\right) - P_{\theta_o}^n\left(A_n^c\right)\right] = \alpha - \alpha_1 > 0$$

we take logarithms to obtain

$$\liminf_{n\to\infty}\left[\log\alpha_2 + D\left(P_{\theta_o}^n\|M_n\right)\right] \geq \frac{1}{2}c(\alpha) + \log(\alpha - \alpha_1)$$

where $\alpha \in (\alpha_1, 1)$. Note that c is strictly decreasing in α and ranges from $-E\chi_d^2$ to ∞, and $\log(\alpha - \alpha_1)$ is strictly increasing. It is possible to get an implicit algebraic relation that must be satisfied by the α that maximizes the right side, or we may choose $\alpha = (\alpha_1 + 1)/2$ to get a lower bound of the form (3.5).

Now we prove the upper bound (3.6). The Bayes' optimal test is of the following form: Reject H if and only if $(X_1, \cdots, X_n) \in C_n$, where C_n is the critical set

$$C_n = \left\{x^n : \log\frac{p(x^n|\theta_o)}{m(x^n)} \leq t\right\}.$$

Choosing

$$t = D\left(P_{\theta_o}^n\|M_n\right) - \frac{c(\alpha_1)}{2}$$

we have that

$$-2\left[\log\frac{p(X^n|\theta_o)}{m(X^n)} - D\left(P_{\theta_o}^n\|M_n\right)\right]$$

converges weakly to a chi-square random variable with d degrees of freedom. Therefore, the limiting probability of a type-I error is

$$\lim_{n\to\infty} P_{\theta_o}(C_n) = \alpha_1.$$

By Markov's inequality, the average probability of a type-II error satisfies

$$\alpha_2 = M_n\left(C_n^c\right) \leq e^{-t} = e^{-D(P_{\theta_o}^n\|M_n)+(1/2)c(\alpha_1)}.$$

Taking logarithms and rearranging gives

$$\limsup_{n\to\infty}\left[\log\alpha_2 + D\left(P_{\theta_o}^n\|M_n\right)\right] \leq \frac{1}{2}c(\alpha_1)$$

so that $c(\alpha_1)/2$ is an upper bound on the limit superior of the left side. Thus, (3.6) is proved. □

D. Application to Portfolio Selection Theory

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n, \cdots$ be a sequence of independent stock-market return vectors, where the coordinates X_{ij} denote the multiplicative factor by which dollars invested in stock j, $j = 1, \cdots, k$ are increased during the ith investment period. At the beginning of each investment period, stocks are bought or sold to result in a portfolio of stock proportions $\boldsymbol{b} = (b_1, \cdots, b_k)$, $b_j \geq 0$, $\sum_{j=1}^k b_j = 1$. Beginning with one unit of wealth, the wealth at the end of n investment periods is $S_n = \prod_{i=1}^n \boldsymbol{b}_i^T \boldsymbol{X}_i$, where $\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots$ is the sequence of portfolio vectors. If the true distribution P_{θ_o} were known, the portfolio $\boldsymbol{b}^* = \boldsymbol{b}(P_{\theta_o})$ would then be chosen to achieve

$$W^* = \max_{\boldsymbol{b}} E\log\boldsymbol{b}^T\boldsymbol{X}$$

in order to achieve maximum possible exponential growth rate of wealth (see Kelly [32]). Not knowing the true distribution, we may base our portfolio $b_n = b_n(\hat{P}_{n-1})$ for the nth investment period on an estimate $\hat{P}_n$ of the true distribution. In [7], it is shown that the resulting decrement in the exponential growth of wealth is bounded by

$$\frac{1}{n}\sum_{i=1}^n ED\left(P_{\theta_o}\|\hat{P}_{i-1}\right).$$

In particular, if we use the predictive density estimator $\hat{p}_n(x) = m(\boldsymbol{X}_{n+1} = x|\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)$, the bound on the decrement is then precisely $(1/n)D(P_{\theta_o}^n\|M_n)$, which is the very quantity approximated by our theorem. The Bayes' sequential investment strategy, which uses the predictive density to select the portfolio, is optimal with respect to M_n. If P_{θ_o} were known, the resulting optimal wealth is

$$S_n^* = e^{n(W^* + o(1))}$$

where $o(1) \to 0$ in probability. We can lower bound the wealth of the Bayes' strategy in terms of the optimal wealth.

Proposition 3.D: The Bayes' strategy, investing based on M_n, achieves wealth at least

$$S_n \geq S_n^* e^{-D(P_{\theta_o}^n \| M_n)} \doteq S_n^* \left(\frac{2\pi e}{n}\right)^{d/2} \frac{w(\theta_o)}{\sqrt{\det I(\theta_o)}}$$

where the last expression holds asymptotically under the conditions of Theorem 2.1. Indeed, for any $\alpha \in (0,1)$ and any $\tau > 0$

$$S_n \geq S_n^* e^{-D(P_{\theta_o}^n \| M_n) - (1/2)c(\alpha) - \tau}$$

except on a set with probability asymptotically less than or equal to $\alpha + e^{-\tau}$, as $n \to \infty$, where $c(\alpha)$ is the same as in the last proposition.

Proof: By Markov's inequality, the wealth satisfies

$$S_n \geq S_n^* \frac{m(X^n)}{p(X^n|\theta_o)}$$

except on a set of probability

$$P_{\theta_o}^n\left(\left\{\frac{S_n^*}{S_n}\frac{m(X^n)}{p(X^n|\theta_o)} \geq e^{\tau}\right\}\right) \leq e^{-\tau} E_{\theta_o} \frac{S_n^*}{S_n}\frac{m(X^n)}{p(X^n|\theta_o)}$$
$$\leq e^{-\tau} E_{m_n} \frac{S_n^*}{S_n}$$
$$\leq e^{-\tau}$$

where the inequality $E_m S_n^*/S_n \leq 1$ follows from the conditions for the optimality of S_n for the distribution M_n (see [7]). The result then follows as it does in the proof of the proposition on hypothesis testing from the fact that twice $\log m(X^n)/p_{\theta_o}(X^n) + D(P_{\theta_o}^n \| M_n)$, asymptotically, has a centered chi-square distribution with d degrees of freedom. □

IV. Proof of the Main Theorem

To prove Theorem 2.1, we will use a lemma that gives upper and lower bounds on the integrand of $D(P_{\theta_o}^n \| M_n)$ on certain sets that have high probability.

We introduce the following notation. Let

$$N_\delta = \{\theta : |\theta - \theta_o| \leq \delta\}$$

where, for convenience, the norm of vectors in $\boldsymbol{R}^d$ is taken to be $|\xi| = |\xi|_{J_{\theta_o}}$ defined by

$$|\xi|_{J_{\theta_o}}^2 = \xi^T J_{\theta_o} \xi.$$

For $0 < \epsilon < 1$ and $\delta > 0$, define the events

$$A_n(\delta, \epsilon) = \left\{\int_{N_\delta^c} p(X^n|\theta) w(\theta)\, d\theta \leq \epsilon \int_{N_\delta} p(X^n|\theta) w(\theta)\, d\theta\right\} \quad (4.1)$$

and

$$B_n(\delta, \epsilon) = \Big\{(1-\epsilon)(\theta-\theta_o)^T J_{\theta_o}(\theta-\theta_o) \leq (\theta-\theta_o)^T J_{\hat{\theta}}^*(\theta-\theta_o) \leq (1+\epsilon)(\theta-\theta_o)^T J_{\theta_o}(\theta-\theta_o), \text{ for all } \theta, \tilde{\theta} \in N_\delta\Big\} \quad (4.2)$$

where

$$J_{\hat{\theta}}^* = J^*(\hat{\theta}) = -\left[(1/n)\partial^2(\log p(X^n|\theta))/\partial\theta_j \partial\theta_k\right]_{j,k=1\cdots d}$$

is the empirical information matrix. In addition, let

$$C_n(\delta) = \left\{l_n'(\theta_o)^T J_{\theta_o}^{-1} l_n'(\theta_o) \leq \delta^2\right\} \quad (4.3)$$

where we have denoted the average score function by

$$l_n'(\theta_o) = \frac{1}{n}\nabla \log p(X^n|\theta_o).$$

A closely related quantity is the standardized score $S_n = (1/\sqrt{n})\nabla \log p(X^n|\theta_o)$.

The set A_n contains those points x^n for which the posterior probability of the neighborhood N is at least $1/(1+\epsilon)$; the set B_n allows us to bound an empirical estimate of the Fisher information by its true value; the set C_n is the set where a norm of the average score is near zero. We bound the behavior of the prior by the modulus of continuity of its logarithm on a neighborhood of the true value:

$$\rho(\delta, \theta_o) = \sup_{\theta \in N_\delta} \left|\log \frac{w(\theta)}{w(\theta_o)}\right|.$$

In the motivation for the result outlined in Section I we used the maximum-likelihood estimator. We find that weaker hypotheses can be stated if a different estimator is used. In what follows, we will use an analog to the maximum-likelihood estimator, which we denote by $\hat{\theta}$. Its definition is

$$\hat{\theta} = \theta_o + J_{\theta_o}^{-1} l_n'(\theta_o).$$

It amounts to a stochastic perturbation about the true value of the parameter. Note that $\hat{\theta}$ is not really an estimator since it depends on the estimand. The quantity $\hat{\theta}$ is used on p. 11i of Hartigan [25] and on p. 456 of Lehmann [42], in proofs of the asymptotic normality of the posterior density.

We next state and prove tight upper and lower bounds on the density ratio. In accordance with Laplace's method, the proof will use a second-order Taylor expansion about θ_o to lead to an approximation by a normal integral.

Lemma 4.1: Suppose Condition 2 is satisfied so that $w(\theta)$ is continuous and positive at θ_o, and J_{θ_o} is positive definite. On the set $A_n \cap B_n$, we have the upper bound

$$\frac{m(x^n)}{p(x^n|\theta_o)} \leq (1+\epsilon) w(\theta_o) e^{\rho(\delta,\theta_o)} \cdot e^{(n/2(1-\epsilon))l_n'(\theta_o)^T J_{\theta_o}^{-1} l_n'(\theta_o)} (2\pi)^{d/2} \cdot \det\left(n(1-\epsilon)J_{\theta_o}\right)^{-1/2}. \quad (4.4)$$

On $B_n \cap C_n$, we have the lower bound

$$\frac{m(x^n)}{p(x^n|\theta_o)} \geq w(\theta_o) e^{-\rho(\delta,\theta_o)} e^{(n/2(1+\epsilon))l_n'(\theta_o)^T J_{\theta_o}^{-1} l_n'(\theta_o)} (2\pi)^{d/2} \cdot (1 - 2^{d/2} e^{-\epsilon^2 n\delta^2/8}) \det\left(n(1+\epsilon) J_{\theta_o}\right)^{-1/2}. \quad (4.5)$$

Proof of Lemma 4.1: In both cases we apply Laplace integration to the mixture density. For the upper bound (4.4), we have, by restriction to A_n and then to B_n, that

$$\begin{aligned}
\frac{m(x^n)}{p(x^n|\theta_o)} &\leq (1+\epsilon) \int_{N_\delta} \frac{p(x^n|\theta)}{p(x^n|\theta_o)} w(\theta)\, d\theta \\
&= (1+\epsilon) \int_{N_\delta} e^{n(\theta-\theta_o)^T l_n'(\theta_o) - (n/2)(\theta-\theta_o)^T J_{\tilde{\theta}}^*(\theta-\theta_o)} w(\theta)\, d\theta \\
&\leq (1+\epsilon) w(\theta_o) e^{\rho(\delta,\theta_o)} \cdot \int_{N_\delta} e^{n(\theta-\theta_o)^T l_n'(\theta_o) - (n/2)(1-\epsilon)(\theta-\theta_o)^T J_{\theta_o}(\theta-\theta_o)}\, d\theta \\
&= (1+\epsilon) w(\theta_o) e^{\rho(\delta,\theta_o)} e^{(n/2)(1-\epsilon) l_n'(\theta_o)^T J_{\theta_o}^{-1} l_n'(\theta_o)} \cdot \int_{N_\delta} e^{-(n/2)(1-\epsilon)(\theta-u)^T J_{\theta_o}(\theta-u)}\, d\theta \\
&\leq (1+\epsilon) w(\theta_o) e^{\rho(\delta,\theta_o)} e^{(n/2)(1-\epsilon) l_n'(\theta_o)^T J_{\theta_o}^{-1} l_n'(\theta_o)} \cdot (2\pi)^{d/2} \det\left(n(1-\epsilon) J_{\theta_o}\right)^{-1/2}
\end{aligned}$$

where we have used a second-order Taylor expansion (with $\tilde{\theta}$ a point in N_δ that depends on θ and x^n), $u = \theta_o + (1/(1-\epsilon))(\hat{\theta} - \theta_o)$, and we have used the following identity, which may be verified by completing the square

$$\begin{aligned}
(\theta-\theta_o)^T l_n'(\theta_o) - \frac{1}{2}(1-\epsilon)(\theta-\theta_o)^T J_{\theta_o}(\theta-\theta_o) &= -\frac{(1-\epsilon)}{2}(\theta-u)^T J_{\theta_o}(\theta-u) \\
&\quad + \frac{1}{2(1-\epsilon)} l_n'(\theta_o)^T J_{\theta_o}^{-1} l_n'(\theta_o). \quad (4.6)
\end{aligned}$$

For the lower bound (4.5), we have

$$\begin{aligned}
\frac{m(x^n)}{p(x^n|\theta_o)} &\geq \int_{N_\delta} \frac{p(x^n|\theta)}{p(x^n|\theta_o)} w(\theta)\, d\theta \\
&= \int_{N_\delta} e^{n(\theta-\theta_o)^T l_n'(\theta_0) - n/2(\theta-\theta_o) J_{\tilde{\theta}}^*(\theta-\theta_o)} w(\theta)\, d\theta
\end{aligned}$$

where $\tilde{\theta} \in N_\delta$. Again, we use the identity previously stated (4.6), but we now replace $(1-\epsilon)$ with $(1+\epsilon)$ and let $u = \theta_o + (1/(1+\epsilon))(\hat{\theta} - \theta_o)$. Because of the restriction to B_n, we can continue the inequality

$$\begin{aligned}
&\geq w(\theta_o) e^{-\rho(\delta,\theta_o)} \int_{N_\delta} e^{n(\theta-\theta_o)^T l_n'(\theta_o) - (n/2)(1+\epsilon)(\theta-\theta_o)^T J_{\theta_o}(\theta-\theta_o)}\, d\theta \\
&= w(\theta_o) e^{-\rho(\delta,\theta_o)} e^{(n/2(1+\epsilon))l_n'(\theta_o)^T J_{\theta_o}^{-1} l_n'(\theta_o)} \cdot \int_{N_\delta} e^{-(1+\epsilon)(n/2)(\theta-u)^T J_{\theta_o}(\theta-u)}\, d\theta \\
&= w(\theta_o) e^{-\rho(\delta,\theta_o)} e^{(n/2(1+\epsilon))l_n'(\theta_o)^T J_{\theta_o}^{-1} l_n'(\theta_o)} \cdot \left[\int_{\boldsymbol{R}^d} e^{-(1+\epsilon)(n/2)(\theta-u)^T J_{\theta_o}(\theta-u)}\, d\theta \right. \\
&\quad \left. - \int_{N_\delta^c} e^{-(1+\epsilon)(n/2)(\theta-u)^T J_{\theta_o}(\theta-u)}\, d\theta \right]. \quad (4.7)
\end{aligned}$$

Since we have restricted to C_n and the norm is with respect to J_{θ_o}, we have, by writing $\alpha = 1/(1+\epsilon)$ and using the definition of u and $\hat{\theta}$, that for $\theta \in N_\delta^c$

$$\begin{aligned}
|\theta - u| &= \left|\theta - \theta_o - \alpha(\hat{\theta} - \theta_o)\right| \\
&= \left|\theta - \theta_o - \alpha J_{\theta_o}^{-1} l_n'(\theta_o)\right| \\
&\geq |\theta - \theta_o| - \alpha \left| J_{\theta_o}^{-1} l_n'(\theta_o) \right| \\
&\geq (1-\alpha)\delta = \frac{\epsilon\delta}{(1+\epsilon)}.
\end{aligned}$$

Consequently, in the second integral of (4.7), the integrand is not greater than

$$e^{-n\epsilon^2\delta^2/4(1+\epsilon)} \cdot e^{-(1+\epsilon)n|\theta-u|^2/4}.$$

Therefore, expanding the domain of integration in the second integral of (4.7) and rearranging, we have the lower bound (4.5). This completes the proof of Lemma 4.1. □

We now have some control over the logarithm of the mixture density over the true density. The integrand of the relative entropy approximated by the theorem uses the reciprocal of that density ratio. Thus, when obtaining upper bounds on it, we will be concerned with the probability of $B_n \cap C_n$, and when obtaining lower bounds, the probability of $A_n \cap B_n$ will be important. It will be demonstrated that the probability of the complements of those sets converges at a fast enough rate to suit our needs. The expected supremum Condition 1 is used to control the probability of B_n^c and of C_n^c. The posterior consistency Condition 3 is used to control the probability of A_n.

Proof of Theorem 2.1: Denote the error in the approximation we seek by

$$R_n = \log \frac{p_{\theta_o}(X^n)}{m(X^n)} - \left(\frac{d}{2} \log \frac{n}{2\pi} + \log \frac{1}{w(\theta_o)} + \frac{1}{2} \log \det\left(J_{\theta_o}\right) - \frac{1}{2} S_n^T J_{\theta_o}^{-1} S_n \right).$$

where $S_n = \sqrt{n}\, l_n'(\theta_o)$. Our task is to show that $\lim E(R_n) = 0$ and, moreover, that $\lim E|R_n| = 0$. This we do by upper bounding R_n by a positive quantity, which tends to

zero in L^1, and lower bounding it by a negative quantity, which also tends to zero in L^1.

First, we obtain the lower bound. From Lemma 4.1 and the positivity of $S_n^T J_{\theta_o}^{-1} S_n$, we have

$$R_n \geq -\left(\frac{\epsilon}{2(1-\epsilon)} S_n^T J_{\theta_o}^{-1} S_n + \frac{d}{2} \log \frac{1}{(1-\epsilon)} + \log(1+\epsilon) + \rho(\delta, \theta_o) \right) \tag{4.8}$$

$$- 1_{(A_n \cap B_n)^c} \left(\log \frac{p_{\theta_o}(X^n)}{m(X^n)} \right)^- \tag{4.9}$$

$$- 1_{(A_n \cap B_n)^c} \left| \frac{d}{2} \log \frac{n}{2\pi} + \log \frac{1}{w(\theta_o)} + \frac{1}{2} \log \det \left(J_{\theta_o} \right) \right|. \tag{4.10}$$

Since each term is negative, this also provides a bound on the negative part of R_n, denoted $(R_n)^- = \max\{0, -R_n\}$. First, we examine the term in (4.9). Define the event $G_n = (A_n \cap B_n)^c \cap \{m(X^n) \geq p_{\theta_o}(X^n)\}$ and note that $\lim P_{\theta_o}^n(G_n) = 0$ if the probabilities of A_n^c and B_n^c tend to zero. The expected value of this term is bounded by use of the concavity of the logarithm and Jensen's inequality to obtain

$$\begin{aligned} E 1_{(A_n \cap B_n)^c} \left(\log \frac{p_{\theta_o}(X^n)}{m(X^n)} \right)^- &= E 1_{G_n} \left(\log \frac{m(X^n)}{p_{\theta_o}(X^n)} \right) \\ &\leq P_{\theta_o}^n(G_n) \log \frac{M_n(G_n)}{P_{\theta_o}^n(G_n)} \\ &\leq P_{\theta_o}^n(G_n) \log \frac{1}{P_{\theta_o}^n(G_n)}, \end{aligned} \tag{4.11}$$

which tends to zero as $P_{\theta_o}^n(G_n) \to 0$.

Next, we note that the expected value of (4.10) tends to zero if $P(A_n^c)$ and $P(B_n^c)$ are both $o(1/\log n)$. Now A_n^c is the event that the posterior probability of N_δ is less than $1/(1+\epsilon)$, and so, $P(A_n^c) = o(1/\log n)$ by Condition 3. It will be seen by an application of Chebyshev's inequality that Condition 1 implies $P(B_n^c) = O(1/n)$.

For the first term in (4.8), we use $E(S_n^T J_{\theta_o}^{-1} S_n) = tr(J_{\theta_o}^{-1} E S_n S_n^T) = tr(J_{\theta_o}^{-1} I_{\theta_o})$. Therefore, collecting these bounds, we have that for every $0 < \epsilon < 1$ and $\delta > 0$

$$\liminf_{n \to \infty} E(R_n) \geq -\left(\frac{\epsilon}{2(1-\epsilon)} tr\left(J_{\theta_o}^{-1} I_{\theta_o} \right) + \frac{d}{2} \log \frac{1}{(1-\epsilon)} + \log(1+\epsilon) + \rho(\delta, \theta_o) \right).$$

Letting ϵ and δ tend to zero shows that $\liminf E(R_n) \geq 0$. In the same way it is seen that, moreover, $\lim E(R_n)^- = 0$.

Next we obtain the upper bound. We use Lemma 4.1, as before, to get

$$R_n \leq \frac{\epsilon}{2(1+\epsilon)} S_n^T J_{\theta_o}^{-1} S_n + \frac{d}{2} \log \frac{1}{(1+\epsilon)} + \rho(\delta, \theta_o) - \log\left(1 - 2^{d/2} e^{-\epsilon^2 n \delta^2/8}\right) + 1_{(B_n \cap C_n)^c} \left(\log \frac{p_{\theta_o}(X^n)}{m(X^n)} \right)^+ \tag{4.12}$$

$$+ 1_{(B_n \cap C_n)^c} \left| \frac{d}{2} \log \frac{n}{2\pi} + \log \frac{1}{w(\theta_o)} + \frac{1}{2} \log \det \left(J_{\theta_o} \right) \right| + S_n^T J_{\theta_o}^{-1} S_n 1_{(B_n \cap C_n)^c}. \tag{4.13}$$

Now, (4.12) and (4.13) are the terms that are examined by methods different from those used in the lower bound.

For (4.13), note that by the central limit theorem, S_n converges in distribution to $Z \sim \text{Normal}(0, I_{\theta_o}^{-1})$. Therefore, $S_n^T J_{\theta_o}^{-1} S_n$ is uniformly integrable since it converges in distribution (to the distribution of the random variable $Z^T J_{\theta_o}^{-1} Z$) and it has convergent, indeed constant, expected absolute value (see p. 100 of Chung [17]). By uniform integrability, $\lim E(S_n^T J_{\theta_o}^{-1} S_n 1_{(B_n \cap D_n)^c}) = 0$ if the probability of B_n^c and C_n^c tend to zero.

Now, for the term (4.12), an adequate upper bound may be obtained by restricting the integral in the definition of m_n to a neighborhood of θ_o and using a first-order Taylor expansion to get

$$\log \frac{p_{\theta_o}(X^n)}{m(X^n)} \leq \sum_{i=1}^n \sup_{\theta, \tilde{\theta} \in N_\delta} (\theta - \theta_o)^T \nabla \log p(X_i | \tilde{\theta}) - \log W(N_\delta)$$

from which we have

$$\begin{aligned} &E 1_{(B_n \cap C_n)^c} \left(\log \frac{p_{\theta_o}(X^n)}{m(X^n)} \right)^+ \\ &\leq E 1_{(B_n \cap C_n)^c} \left(\sum_{i=1}^n \sup_{\theta, \tilde{\theta} \in N_\delta} (\theta - \theta_o)^T \nabla \log p(X_i | \tilde{\theta}) \right) \\ &\quad + P\left((B_n \cap C_n)^c \right) |\log W(N_\delta)|. \end{aligned} \tag{4.14}$$

Adding and subtracting the expected value of the supremum from each term in the sum and then using the Cauchy–Schwarz inequality yields

$$\begin{aligned} &E 1_{(B_n \cap C_n)^c} \left(\sum_{i=1}^n \sup_{\theta, \tilde{\theta} \in N_\delta} (\theta - \theta_o)^T \nabla \log p(X_i | \tilde{\theta}) \right) \\ &\leq n P\left((B_n \cap C_n)^c \right) E \sup_{\theta, \tilde{\theta} \in N_\delta} (\theta - \theta_o)^T \nabla \log p(X | \tilde{\theta}) \\ &\quad + \left(P\left((B_n \cap C_n)^c \right) \right)^{1/2} \\ &\quad \cdot \left(n \operatorname{var} \left(\sup_{\theta, \tilde{\theta} \in N_\delta} (\theta - \theta_o)^T \nabla \log p(X | \tilde{\theta}) \right) \right)^{1/2}. \end{aligned} \tag{4.15}$$

Now, sufficiently small δ, the expected suprema are finite by application of Condition 1. (Condition 1 assumes that the second-order derivatives are locally dominated by square integrable functions; then, by application of Taylor's expansion, lower order derivatives are also locally dominated.) Consequently, these upper bounds tend to zero as $n \to 0$, provided $P(B_n^c)$ and $P(C_n^c)$ are $o(1/n)$.

In this case, incorporating these bounds in (4.12) and (4.13), we obtain $\limsup E(R_n) \le 0$ and, moreover, $\lim E(R_n)^+ = 0$. We remark that a different proof of the same upper bound on the limit superior, by a somewhat fancier argument, is given in Section V.

Combining with the limit inferior result, we have established our main result (2.4), which is equivalent to the convergence to zero of the expected value of R_n. Moreover, as a byproduct of the analysis, we also have convergence in L^1, that is, $\lim E|R_n| = 0$.

The probabilities of B_n and C_n must still be examined. Bounds on $P(B_n^c)$ are needed for both the limit inferior and the limit superior in this proof. Bounds on $P(C_n^c)$ are used for the limit superior.

In controlling the probability of B_n, we will use the fact, based on Chebyshev's inequality, that for the sample average $\bar{Y}$ of i.i.d. outcomes of a random variable Y with finite variance

$$P(|\bar{Y} - EY| > \epsilon) \le \frac{1}{n\epsilon^2} En(\bar{Y} - EY)^2 1_{\{|\bar{Y}-EY|>\epsilon\}} = \frac{1}{n\epsilon^2} c(n,\epsilon)$$

where $c(n,\epsilon) = En(\bar{Y} - EY)^2 1_{\{|\bar{Y}-EY|>\epsilon\}}$ tends to zero as $n \to \infty$ for any fixed $\epsilon > 0$. To see that $c(n,\epsilon) \to 0$, note that $n(\bar{Y} - EY)^2$ is uniformly integrable since it converges in distribution and has convergent indeed constant, expected absolute value.

From Chebyshev's inequality, we will obtain bounds of the form

$$P(B_n^c(\delta,\epsilon)) \le \frac{c_1(\theta_o, n, \epsilon, \delta)}{n\epsilon^2} \tag{4.16}$$

where the function c_1 tends to zero as n increases for any fixed δ and ϵ. By Markov's inequality, we will show that we have

$$P_{\theta_o}(C_n^c(\delta)) \le \frac{c_2(\theta_o, n, \delta)}{n\delta^2} \tag{4.17}$$

where c_2 tends to zero as n increases for any fixed δ. We proceed with proving that c_1 and c_2 exist as we want.

To show the existence of c_1, it is enough to examine the probability of sets of the form

$$\left\{ \sup_{\theta \in N_\delta} |J_{j,k}^*(\theta) - J_{j,k}(\theta_o)| > \epsilon' \right\}$$

where $J_{j,k}$ and $J_{j,k}^*$ denote the entries of the information matrix $J(\theta_o)$ and the empirical information matrix $J^*(\theta)$, respectively. This is suggested by noting that $B_n(\delta,\epsilon)$ can be written as

$$B_n(\delta,\epsilon) = \left\{ \left| \frac{\xi^T (J^*(\tilde{\theta}) - J(\theta_o))\xi}{\xi^T \xi} \right| < \epsilon, \quad \text{for all } \theta, \tilde{\theta} \in N_\delta \right\}$$

where $\xi = \theta - \theta_o$. By the Cauchy–Schwarz inequality, the quotient in this set is not greater than the square root of the sum of the squares of the entries in the matrix $J^*(\tilde{\theta}) - J(\theta_o)$, which, in turn, is less than d times the maximum absolute value. Then, by the union of events bound, setting $\epsilon' = \epsilon/d$

$$P(B_n^c) \le \sum_{j,k} P\left\{ \sup_{\theta \in N_\delta} |J_{j,k}^*(\theta) - J_{j,k}(\theta_o)| > \epsilon' \right\}.$$

For each of the finitely many terms in this sum, the probability is upper-bounded by adding and subtracting $J_{j,k}^*(\theta_o)$ to get

$$P\left(\sup_{|\theta_o - \theta| < \delta} |J_{j,k}^*(\theta) - J_{j,k}^*(\theta_o)| > \frac{\epsilon'}{2} \right) + P\left(|J_{j,k}^*(\theta_o) - J_{j,k}(\theta_o)| > \frac{\epsilon'}{2} \right). \tag{4.18}$$

By Chebyshev, the second term in (4.18) is upper bounded by

$$\frac{4}{n(\epsilon')^2} E_{\theta_o} 1_{\{|J_{j,k}^*(\theta_o) - J_{j,k}(\theta_o)| > \epsilon'/2\}} \left(\sqrt{n} \left(J_{j,k}^*(\theta_o) - J_{j,k}(\theta_o) \right) \right)^2. \tag{4.19}$$

For the first term in (4.18), we choose δ so small that

$$E \sup_{|\theta_o - \theta| < \delta} \left| \frac{\partial^2}{\partial\theta_j \partial\theta_k} \log p(X|\theta) - \frac{\partial^2}{\partial\theta_j \partial\theta_k} \log p(X|\theta_o) \right| < \frac{\epsilon'}{4}$$

and set up another application of Chebyshev's inequality. Let

$$Y_i = \sup_{|\theta_o - \theta| < \delta} \left| \frac{\partial^2}{\partial\theta_j \partial\theta_k} \log p(X_i|\theta) - \frac{\partial^2}{\partial\theta_j \partial\theta_k} \log p(X_i|\theta_o) \right|.$$

Now, the first term in (4.18) is upper-bounded by

$$P_{\theta_o}(|\bar{Y} - EY| > \epsilon'/4) \le \frac{16}{n(\epsilon')^2} E\left(\left(\sqrt{n} (\bar{Y} - EY_1) \right)^2 \right) 1_{\{|\bar{Y}-EY|>\epsilon'/4\}}. \tag{4.20}$$

Adding the bounds (4.19) and (4.20) for the terms of (4.18), we see that we have an expression for c_1 of the form desired for (4.16).

Similarly from Markov's inequality, we can identify an expression for c_2 for use in (4.17):

$$P_{\theta_o}(C_n^c) \le \frac{1}{n\delta} E 1_{C_n^c} n l_n'(\theta_o)^T J_{\theta_o}^{-1} l_n'(\theta_o) = \frac{1}{n\delta} E 1_{C_n^c} S_n^T J_{\theta_o}^{-1} S_n.$$

Again, the expectation goes to zero as $n \to \infty$ by the uniform integrability of $S_n^T J_{\theta_o}^{-1} S_n$. Thus, $P(B_n^c)$ and $P(C_n^c)$ are of order $o(1/n)$. This completes the proof of Theorem 2.1. □

V. Upper Bounds Under Weaker Conditions and Asymptotic Normality

In this section, we prove Theorems 2.3 and 2.4. Condition 2 on the twice continuous differentiability of the relative entropy and Condition 4 on the mean square differentiability of the logarithm of the density are the two key assumptions used here.

Proof of Theorem 2.3: First, we show that Condition 2 is enough for $\limsup (E \log p_{\theta_o}(X^n)/m(X^n) - (d/2)\log n) \le c$ but with a constant c somewhat larger than identified in Theorem 2.1.

Fix $K > d$, set $\delta_n = \sqrt{K/n}$, and let the norm $|\theta - \theta_o| = |\theta - \theta_o|_{J_{\theta_o}}$ be taken with respect to J_{θ_o} as in Section IV. We start by reexamining $\log(w(\theta_o)p_{\theta_o}(X^n)/m(X^n))$, restricting the integral to a sequence of neighborhoods of the form $N_{\delta_n} = \{\theta\colon |\theta - \theta_o| \le \delta_n\}$, multiplying and dividing by a truncated normal density $\phi_n(\theta|N_\delta) = (1/c_n)e^{-(n/2)|\theta-\theta_o|^2}1_{\{\theta \in N_\delta\}}$ where $c_n = \int_{N_\delta} e^{-(n/2)|\theta-\theta_o|^2}\,d\theta$, and applying Jensen's inequality as follows:

$$\begin{aligned}
&\log \frac{w(\theta_o)p_{\theta_o}(X^n)}{m(X^n)} \\
&\quad = -\log \int \frac{p_\theta(X^n)}{p_{\theta_o}(X^n)} \frac{w(\theta)}{w(\theta_o)}\,d\theta \\
&\quad \le -\log \int_{N_{\delta_n}} \frac{p_\theta(X^n)}{p_{\theta_o}(X^n)} \frac{w(\theta)}{w(\theta_o)}\,d\theta \\
&\quad = -\log \int_{N_{\delta_n}} \frac{p_\theta(X^n)}{p_{\theta_o}(X^n)} \frac{e^{(n/2)|\theta-\theta_o|^2}}{c_n} \frac{w(\theta)}{w(\theta_o)} \phi_n(\theta|N_{\delta_n})\,d\theta \\
&\quad \le \int_{N_{\delta_n}} \left(\log \frac{p_{\theta_o}(X^n)}{p_\theta(X^n)} - \frac{n}{2}|\theta-\theta_o|^2 - \log \frac{w(\theta)}{w(\theta_o)} \right) \\
&\qquad \cdot \phi_n(\theta|N_{\delta_n})\,d\theta - \log c_n. \qquad (5.1)
\end{aligned}$$

Then, taking an expected value inside the integral, which is valid by an application of Fubini's theorem, and using the second-order Taylor approximation $E \log p_{\theta_o}(X)/p_\theta(X) = (1/2)|\theta - \theta_o|^2 + o(|\theta - \theta_o|^2)$ and the positivity and continuity of $w(\theta)$ at θ_o, which are valid under Condition 2, we obtain

$$E \log \frac{w(\theta_o)p_{\theta_o}(X^n)}{m(X^n)} \le -\log c_n + o(1). \qquad (5.2)$$

By comparison with a multivariate normal integral, we bound $c_n = \int_{N_{\delta_n}} e^{-(n/2)|\theta-\theta_o|^2}\,d\theta$ as follows

$$\begin{aligned}
(1 - d/K)(2\pi)^{d/2} n^{-d/2} \det J_{\theta_o}^{-1/2} \\
< c_n < (2\pi)^{d/2} n^{-d/2} \det J_{\theta_o}^{-1/2} \qquad (5.3)
\end{aligned}$$

where the d/K term follows from the use of Chebyshev's inequality to bound the normal$(\theta_o, (nJ_{\theta_o})^{-1})$ probability of the event that $|\theta - \theta_o| > \delta_n$. From (5.2) and (5.3), we have

$$\begin{aligned}
\limsup &\left(E \log \frac{p_{\theta_o}(X^n)}{m(X^n)} - \frac{d}{2}\log \frac{n}{2\pi} \right) \\
&\le \log \frac{1}{w(\theta_o)} + \frac{1}{2}\log \det J_{\theta_o} - \log\left(1 - \frac{d}{K}\right).
\end{aligned}$$

Let $K \to \infty$ to complete the proof of inequality (2.10) in Theorem 2.3.

Next, we use mean-square differentiability (Condition 4) as well as Condition 2 in a refinement of the previous proof to obtain a bound on the limit superior, which is the same as in Theorem 2.1.

Here, given $K > d$, let $\hat{N}_{\delta_n} = \{\theta\colon |\theta - \hat{\theta}| \le \delta_n\}$ where $\delta_n = \sqrt{K/n}$, and

$$\hat{\theta} = \theta_o + \frac{1}{\sqrt{n}} J_{\theta_o}^{-1} S_n 1_{\{S_n^T J_{\theta_o}^{-1} S_n \le K\}}.$$

Let $\hat{\phi}_n(\theta) = (1/c_n)e^{-(n/2)|\theta-\hat{\theta}|^2}1_{\{\theta \in \hat{N}_\delta\}}$ be the truncated normal centered at $\hat{\theta}$ instead of θ_o. The normalizing constant c_n is the same as previously stated. Note that for θ in $\hat{N}_\delta$, we have $|\theta - \theta_o| \le |\theta - \hat{\theta}| + |\hat{\theta} - \theta_o| \le 2\sqrt{K/n} = 2\delta_n$. Then, by the same reasoning as in (5.1), we have

$$\begin{aligned}
\log \frac{w(\theta_o)p_{\theta_o}(X^n)}{m(X^n)} &\le -\log c_n + \rho(2\delta_n, \theta_o) \\
&+ \int_{\hat{N}_{\delta_n}} \left(\log \frac{p_{\theta_o}(X^n)}{p_\theta(X^n)} - \frac{n}{2}|\theta - \hat{\theta}|^2 \right) \hat{\phi}_n(\theta)\,d\theta \qquad (5.4)
\end{aligned}$$

where $\rho(2\delta_n, \theta_o)$, obtained from the modulus of continuity of $\log w(\theta)$ at θ_o, tends to zero as $\delta_n \to 0$. Now expanding the square, we have for θ in $\hat{N}_{\delta_n}$

$$\begin{aligned}
\frac{n}{2}|\theta - \hat{\theta}|^2 &= \frac{n}{2}|\theta - \theta_o|^2 + \sqrt{n}\,(\theta_o - \theta)^T S_n + \frac{1}{2} S_n^T J_{\theta_o}^{-1} S_n \\
&\quad - \left(\sqrt{n}\,(\theta_o - \theta)^T S_n + \frac{1}{2} S_n^T J_{\theta_o}^{-1} S_n \right) 1_{\{S_n^T J_{\theta_o}^{-1} S_n > K\}} \\
&\ge \frac{n}{2}|\theta - \theta_o|^2 + \sqrt{n}\,(\theta_o - \theta)^T S_n + \frac{1}{2} S_n^T J_{\theta_o}^{-1} S_n \\
&\quad - \frac{3}{2} S_n^T J_{\theta_o}^{-1} S_n 1_{\{S_n^T J_{\theta_o}^{-1} S_n > K\}}. \qquad (5.5)
\end{aligned}$$

The last inequality follows from observing that in the event $\{S_n^T J_{\theta_o}^{-1} S_n > K\}$, we have $\hat{\theta} = \theta_o$, and by the Cauchy–Schwarz inequality $\sqrt{n}\,|(\theta_o - \theta)^T S_n| \le \sqrt{n}\,|\hat{\theta} - \theta|(S_n^T J_{\theta_o}^{-1} S_n)^{1/2}$, which for θ in $\hat{N}_{\delta_n}$, is not greater than $K^{1/2}(S_n^T J_{\theta_o}^{-1} S_n)^{1/2} \le S_n^T J_{\theta_o}^{-1} S_n$. Incorporating (5.5) into

(5.4), we have

$$\log\frac{w(\theta_o)p_{\theta_o}(X^n)}{m(X^n)} \le -\log c_n - \frac{1}{2}S_n^T J_{\theta_o}^{-1}S_n + \rho(2\delta_n,\theta_o) + \int_{\hat{N}_{\delta_n}}\left(\log\frac{p_{\theta_o}(X^n)}{p_\theta(X^n)} - \sqrt{n}(\theta_o-\theta)^T S_n - \frac{n}{2}|\theta_o-\theta|^2\right)\cdot\hat{\phi}_n(\theta)\,d\theta + \frac{3}{2}S_n^T J_{\theta_o}^{-1}S_n 1_{\{S_n^T J_{\theta_o}^{-1}S_n > K\}}. \quad (5.6)$$

We now show that the integral in the second line of (5.6) tends to zero in $L^1(P)$. Toward this end, we bound its absolute value by

$$\frac{e^{2K}}{1-d/K}\int_{N_{2\delta_n}}\left|\log\frac{p_{\theta_o}(X^n)}{p_\theta(X^n)} - \sqrt{n}(\theta_o-\theta)^T S_n - \frac{n}{2}|\theta_o-\theta|^2\right|\cdot\phi_n^*(\theta)\,d\theta \quad (5.7)$$

where $\phi_n^*(\theta)$ is the Normal$(\theta_o,(nJ_{\theta_o})^{-1})$ density. Here, we have used the bound

$$\hat{\phi}_n(\theta) \le (1/c_n)e^{2K}e^{-(n/2)|\theta-\theta_o|^2}1_{\{|\theta-\theta_o|<2\delta_n\}} \le (1-d/K)^{-1}e^{2K}\phi_n^*(\theta),$$

which follows from the bounds on c_n as in (5.3) and from the inequality $(n/2)|\theta-\hat{\theta}|^2 \ge (n/2)|\theta-\theta_o|^2 - 2K$ from θ in $\hat{N}_{\delta n}$. Thus, the density $\hat{\phi}_n(\theta)$, which is centered at the random point $\hat{\theta}$, is bounded in terms of the density $\phi_n^*(\theta)$ centered at the nonrandom θ_o.

Taking the expected value of the integral in (5.7), we obtain

$$\frac{e^{2K}}{1-d/K}\int_{N_{2\delta_n}}\phi_n^*(\theta)E\left|\log\frac{p_{\theta_o}(X^n)}{p_\theta(X^n)} - \sqrt{n}(\theta_o-\theta)^T S_n - \frac{n}{2}|\theta_o-\theta|^2\right|d\theta \quad (5.8)$$

where the exchange of integral and expectation is valid by the Fubini–Tonelli theorem for nonnegative functions.

Thus, it is enough to show that the following expectation converges to zero, uniformly for $\theta \in N_{2\delta_n}$

$$E\left|\log\frac{p_{\theta_o}(X^n)}{p_\theta(X^n)} - \sqrt{n}(\theta_o-\theta)^T S_n - \frac{n}{2}|\theta_o-\theta|^2\right|. \quad (5.9)$$

This we bound by

$$E\left|\log\frac{p_{\theta_o}(X^n)}{p_\theta(X^n)} - \sqrt{n}(\theta_o-\theta)^T S_n - nE\log\frac{p_{\theta_o}(X^n)}{p_\theta(X^n)}\right| \quad (5.10)$$

$$+ n\left|E\log\frac{p_{\theta_o}(X)}{p_\theta(X)} - \frac{1}{2}|\theta_o-\theta|^2\right|. \quad (5.11)$$

Now, given any $\epsilon > 0$, (5.10) is bounded by

$$\left(E\left(\log\frac{p_{\theta_o}(X^n)}{p_\theta(X^n)} - \sqrt{n}(\theta_o-\theta)^T S_n - nE\log\frac{p_{\theta_o}(X^n)}{p_\theta(X^n)}\right)^2\right)^{1/2} \le \left(nE\left(\log\frac{p_{\theta_o}(X)}{p_\theta(X)} - (\theta_o-\theta)^T i_{\theta_o}(X)\right)^2\right)^{1/2} \quad (5.12)$$

$$\le \sqrt{n}\,\epsilon|\theta_o-\theta| \quad (5.13)$$

for all θ near θ_o, by the mean-square differentiability of $\log p_\theta(X)$ at θ_o. Here, we have used the fact that the expected square in (5.12) is the variance of the sum of independent copies of the random variables $\log p_{\theta_o}(X)/p_\theta(X) - (\theta_o-\theta)^T i_{\theta_o}(X)$. (Note that $S_n = (1/\sqrt{n})\sum i_{\theta_o}(X_i)$ has mean zero since as a consequence of the mean-square differentiability, $Ei_{\theta_o}(X)$ must be the gradient of $E\log p_{\theta_o}(X)/p_\theta(X)$ at θ_o, which is zero under Condition 2.)

In addition, from the second-order Taylor expansion of $E\log p_{\theta_o}(X)/p_\theta(X)$, which is valid under Condition 2, the expression in (5.11) is bounded by $n\epsilon|\theta_o-\theta|^2$ for all θ near θ_o. Together with (5.13), this shows that the expectation in (5.9) is bounded for all large n, by $(4n\delta_n^2 + 2\sqrt{n}\,\delta_n)\epsilon \le (4K+2\sqrt{K})\epsilon$, uniformly for θ in $N_{2\delta_n}$.

We can now finish the proof of Theorem 2.3. Denote the error in the approximation we seek as

$$R_n = \log\frac{w(\theta_o)p_{\theta_o}(X^n)}{m(X^n)} - \frac{d}{2}\log\frac{n}{2\pi} - \frac{1}{2}\log\det(J_{\theta_o}) - \frac{1}{2}S_n^T J_{\theta_o}^{-1}S_n.$$

Using (5.6), collecting the bounds on the terms from (5.3) and (5.8), and setting $\epsilon = e^{-3K}$, we have that for all large n

$$E(R_n)^+ \le -\log(1-d/K) + \rho(2\delta_n,\theta_o) + \frac{e^{-K}}{(1-d/K)}(4K+2\sqrt{K}) + \frac{3}{2}ES_n^T J_{\theta_o}^{-1}S_n 1_{\{S_n^T J_{\theta_o}^{-1}S_n > K\}}. \quad (5.14)$$

Now $S_n^T J_{\theta_o}^{-1}S_n$ is uniformly integrable; therefore, we may let $n\to\infty$, and then $K\to\infty$ in (5.14) to conclude that

$$\lim_{n\to\infty} E(R_n)^+ = 0.$$

Thus we have bounded R_n above by a function that converges to zero in $L^1(P)$. Consequently, $\limsup E(R_n) \le 0$. This completes the proof of Theorem 2.3. □

Proof of Theorem 2.4: Here, we show that if Conditions 2 and 4 are satisfied, then the result of Theorem 2.1, (2.5) is equivalent to the L^1 convergence of the difference of the logarithms of the posterior density of $T = \sqrt{n}(\theta-\theta_o)$, denoted $w_T(t|X^n)$, and the Normal$(J_{\theta_o}^{-1}S_n, J_{\theta_o}^{-1})$ density, denoted $\phi_n(t)$.

For $t=0$, we evaluate $\log w_T(t|X^n)-\log \phi_n(t)$ and see that it is

$$\log\left((1/\sqrt{n})^d w(\theta_o)\, p(X^n|\theta_o)/m(X^n)\right) - \log\left((2\pi)^{-d/2}\det J_{\theta_o}^{1/2}\right)+(1/2)S_n^T J_{\theta_o}^{-1}S_n,$$

which tends to zero in $L^1(P)$ if and only if (2.5) holds.

For any fixed $t\neq 0$, we have that $\log w(\theta_o + t/\sqrt{n})/w(\theta_o)$ tends to zero by the continuity and positivity of the prior at θ_o. Note that $\log(\phi_m(t)/\phi_n(0)) = t^T S_n - (1/2)t^T J_{\theta_o} t$. Now, $\log(p(X^n|\theta_o + t/\sqrt{n})/p(X^n|\theta_o)) - t^T S_n + (1/2)t^T J_{\theta_o} t$ converges to zero in $L^1(P)$, as is shown for (5.9), assuming Conditions 2 and 4. In this case, $\lim E|\log w_T(t|X^n)-\log\phi_n(t)| = 0$ if and only if this limit obtains at $t=0$, which is equivalent to (2.5). This completes the proof of Theorem 2.4. □

VI. Posterior Consistency

The posterior distribution is consistent if it converges to a degenerate distribution at the true parameter value. Posterior consistency is traditionally used as a key step in showing the convergence of Bayes' estimators. The study of asymptotics for the posterior distribution began with Laplace and has been subsequently examined by many including Le Cam [40], [41], Schwartz [48], Von Mises [55], Walker [58], Berk [9], [10], Johnson [30], [31], Bickel and Yahav [11], Ibragimov and Hasminskii [29], and Hartigan [25]. We use here the techniques of Schwartz [48], based on the existence of uniformly consistent tests of hypotheses, to derive the sufficiency of the soundness condition for posterior consistency at a given rate.

Formally, by posterior consistency, we mean that when θ_o is taken to be true, then for every neighborhood N of θ_o, the posterior probability $W(N|X^n)$ converges to one in probability, i.e., for every $\alpha>0$

$$\lim_{n\to\infty} P_{\theta_o}\{W(N|X^n)>\alpha\} = 0.$$

By posterior consistency at rate $O(f(n))$ we mean that for each neighborhood N and $\alpha>0$, there exists c such that

$$P_{\theta_o}\{W(N|X^n)>\alpha\} \le cf(n)$$

where $f(n)\to 0$ as $n\to\infty$. Posterior consistency with rate $o(f(n))$ is defined similarly.

As is defined in Section II, the main condition we use is the soundness of the parametric family. Other conditions may be used for posterior consistency. For instance, the conditions of Wald [57] are sufficient. In particular, the conclusion of Wolfowitz [60] readily yields a uniformly consistent test (see also Strasser [51] and Le Cam [40]). In some cases, however, Wald's conditions (especially the condition that $E_{\theta_o}\sup_{|\theta|>r}\log p(X|\theta)<\infty$ for some $r>0$) are not satisfied or they are hard to verify. We find the soundness condition to be more fundamental and in some cases easier to verify.

A test of composite hypotheses is said to be uniformly exponentially consistent (UEC) if the type-I and type-II error are uniformly upper bounded by e^{-nr} for some positive r (see [6]).

The next three propositions amount to a proof of Theorem 2.2. Proposition 6.1 shows that analogs of the soundness condition for certain metrics on probability measures imply the existence of UEC tests. Proposition 6.2 shows that metrics with the desired consistency property exist. In Proposition 6.3, we use the existence of a UEC test to guarantee the consistency of the posterior distribution at the desired rate.

Consider metrics $d(P,Q)$ on the space of probability measures on X with the property that for any $\epsilon>0$, there exists $r>0$ such that

$$P^n\{d(\hat{P}_n,P)>\epsilon\} \le e^{-nr} \tag{6.1}$$

uniformly over all probability measures P, where $\hat{P}_n$ is the empirical distribution. Examples of metrics that satisfy (6.1) include the Kolmogorov–Smirnov distance, as is shown by Kiefer and Wolfowitz [33], the distances of Vapnik and Chervonenkis [54], and as shown below in Proposition 6.2, certain metrics constructed to imply weak convergence. The idea for the following proposition is from Hoeffding and Wolfowitz [28].

Proposition 6.1: Suppose d is a metric satisfying (6.1) and

$$d(P_\theta, P_{\theta_o})\to 0 \quad \text{implies} \quad \theta\to\theta_o. \tag{6.2}$$

Then, for any $\delta>0$, there exists a UEC hypothesis test of $\theta=\theta_o$ versus $\{\theta: |\theta-\theta_o|>\delta\}$.

Proof: From (6.2), given $\delta>0$, there exists an $\epsilon'>0$ such that $|\theta-\theta_o|>\delta$ implies $d(P_\theta,P_{\theta_o})>\epsilon'$. If we have a UEC test of

$$H: P=P_{\theta_o} \text{ versus } K: P\in\{Q: d(Q,P_{\theta_o})>\epsilon'\}$$

then we have a UEC test of

$$H:\theta=\theta_o \text{ versus } K:\theta\in\{\theta': |\theta'-\theta_o|>\delta\}.$$

The identification of a UEC test for the nonparametric class of alternatives remains. Let $\hat{P}_n$ denote the empirical distribution, choose $\epsilon=\epsilon'/2$, and let

$$C_n=\{x^n: d(\hat{P}_n,P_o)>\epsilon\}$$

be the critical region. By (6.1), we have that the probability of a type-I error satisfies

$$P_{\theta_o}(C_n)\le e^{-nr}$$

and for any choice Q in the set of alternatives, we want to show that the probability of a type-II error

$$Q(C_n^c)=Q\{d(\hat{P}_n,P_o)\le\epsilon\}$$

is uniformly exponentially small. From the triangle inequality, we have that for X^n in C_n^c

$$2\epsilon \le d(P_{\theta_o},Q)\le d(P_{\theta_o},\hat{P}_n)+d(\hat{P}_n,Q)$$
$$\le \epsilon + d(\hat{P}_n,Q).$$

Therefore, again by (6.1)

$$Q(C_n^c)\le Q(d(\hat{P},Q)\ge\epsilon)$$
$$\le e^{-nr}$$

uniformly for Q in K. □

For the following proposition, the probability measures are assumed to be defined on the Borel subsets of a separable metric space $\boldsymbol{X}$.

Proposition 6.2: For probability measures on a separable metric space, there exists a metric $d_G(P,Q)$ that satisfies (6.1) and such that convergence in d_G implies weak convergence of the measures. Therefore, in particular, for measures in a parametric family

$$d_G(P_\theta, P_{\theta_o}) \to 0 \text{ implies } P_\theta \to P_{\theta_o}.$$

Consequently, if the parametric family is soundly parameterized, there exists a UEC test of $\theta = \theta_o$ versus $\{\theta: |\theta - \theta_o| > \delta\}$.

Proof: Let $G = \{F_1, F_2, \cdots\}$ be the countable field of sets generated by balls of the form $\{x: d_X(x, s_j) \le 1/k\}$ for $j, k = 1, 2, \cdots$, where here, d_X denotes the metric for the space $\boldsymbol{X}$, and $s_1, s_2, \cdots$ is a countable dense sequence of points in $\boldsymbol{X}$. Define

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

Here, d_G is a metric on the space of probability measures with the property that if $d_G(P_n, P_o) \to 0$, then P_n converges weakly to P_o (see pp. 251–253 of Gray [24]).

Now, for any $\epsilon > 0$

$$\begin{aligned} d_G(\hat{P}_n, P_o) &\le \sum_{i=1}^{k} 2^{-i} |\hat{P}_n(F_i) - P_o(F_i)| + \sum_{i=k+1}^{\infty} 2^{-i} \\ &\le \max_{1 \le i \le k} |\hat{P}_n(F_i) - P_o(F_i)| + \epsilon/2 \end{aligned}$$

for $k \ge 1 + \log 2/\epsilon$. Then

$$\begin{aligned} P_o\{d_G(\hat{P}_n, P_o) \ge \epsilon\} &\le P_o\left\{ \max_{1 \le i \le k} |\hat{P}_n(F_i) - P_o(F_i)| > \epsilon/2 \right\} \\ &\le \sum_{i \le k} P_o\{|\hat{P}_n(F_i) - P_o(F_i)| > \epsilon/2\} \\ &\le 2ke^{-2n(\epsilon/2)^2} \\ &= 2ke^{-n\epsilon^2/2} \end{aligned}$$

by Hoeffding's inequality [27]. This verifies (6.1) and completes the proof. □

The third proposition uses the conclusion of the preceding proposition as its hypothesis and obtains a posterior consistency result as is required for Theorem 2.2.

Proposition 6.3: Suppose that Condition 2 is satisfied by the family, that is, $D(P_{\theta_o} \| P_\theta)$ is twice continuously differentiable at $\theta = \theta_o$, with $J(\theta_o)$ positive definite, and the prior $w(\theta)$ is continuous and positive at θ_o. For any neighborhood N of θ_o, if there exists a UEC test of $\theta = \theta_o$ versus $\theta \in N^c$, then there is an $r > 0$ such that

$$P_{\theta_o}\left(\int_N w(\theta) p(x^n|\theta)\, d\theta < e^{nr} \int_{N^c} w(\theta) p(x^n|\theta)\, d\theta \right) = O\left(\frac{1}{n}\right)$$

and consequently

$$P_{\theta_o}(W(N^c|X^n) > 2e^{-nr}) = O\left(\frac{1}{n}\right).$$

Proof: To make use of the existence of a UEC test, we will first want to show that for any given $r' > 0$, the probability of the event

$$\tilde{U}_n^c = \left\{ \int_N w(\theta) p(X^n|\theta)\, d\theta < e^{-nr'} p(X^n|\theta_o) \right\}$$

is $O(1/n)$. Set $N_\delta = \{\theta: |\theta - \theta_o| < \delta\}$, where the norm is taken with respect to J_{θ_o} as in Section IV. Since N_δ is contained in N for all small δ, it is enough to show that the following event has probability of the desired rate

$$U_n^c = \left\{ \int_{N_\delta} w(\theta) p(X^n|\theta)\, d\theta < e^{-nr'} p(X^n|\theta_o) \right\}.$$

This set may be rewritten as

$$U_n^c = \left\{ \log \frac{p(X^n|\theta_o)}{m(X^n|N_\delta)} > nr_n' \right\}$$

where $m(x^n|N_\delta) = \int_{N_\delta} w(\theta) p(x^n|\theta)\, d\theta / W(N_\delta)$ is the mixture of distributions with respect to the prior conditioned on $\theta \in N_\delta$, and

$$r_n' = r' - \frac{1}{n} \log W(N_\delta).$$

By Condition 2 and the second-order Taylor expansion of $D(P_{\theta_o}|P_\theta)$, we see that for all small $\delta > 0$, the points in the set N_δ satisfy $D(P_{\theta_o}\|P_\theta) \le \delta^2$, and $w(\theta) \ge w(\theta_o)/2$. Then, by evaluation of the volume of the ellipsoid N_δ and setting $\delta = 1/\sqrt{n}$, we have for all large n

$$\begin{aligned} W(N_\delta) &= \int_{N_\delta} w(\theta)\, d\theta \\ &\ge \frac{w(\theta_o)}{2} c_d \det(J_{\theta_o})^{-1/2} \left(\frac{1}{n}\right)^{d/2} \end{aligned}$$

where c_d is the volume of the unit ball in R^d. Consequently, $(1/n)|\log W(N_\delta)| = O((\log n)/n)$, which tends to zero; therefore, r_n' converges to r', and hence, $r_n' \ge r'/2$ for all large n.

Now by Markov's inequality, we note that

$$\begin{aligned} P_{\theta_o}\left((U_n^c) \le P_{\theta_o}\left(\left| \log \frac{p(X^n|\theta_o)}{m(X^n|N_\delta)} \right| > nr'/2 \right) \right. \\ \le \frac{2}{nr'} E_{\theta_o} \left| \log \frac{p(X^n|\theta_o)}{m(X^n|N_\delta)} \right| \\ \le \frac{2}{nr'} \left(D(P_{\theta_o}^n \| M^n(\cdot|N_\delta)) + 2e^{-1} \right) \end{aligned} \tag{6.3}$$

where we have used the fact that the negative part of the integrand in the relative entropy is always bounded below by e^{-1} since $x \log x \ge -e^{-1}$. It is enough to bound the relative entropy in (6.3). Noting that $M_n(\cdot|N_\delta)$ is an average of measures P_θ^n for θ in N_δ, we have, by convex-

ity and the choice of $\delta = 1/\sqrt{n}$, that

$$D\left(P_{\theta_o}^n | M_n(\cdot|N_\delta)\right) \le \int_{N_\delta} nD\left(P_{\theta_o} \| P_{\theta_o}\right) d\theta \le n\delta^2 = 1. \quad (6.4)$$

Consequently, for all large n

$$P_{\theta_o}(U_n^c) \le \frac{2}{nr'}(1+2e^{-1}) \quad (6.5)$$

which is clearly $O(1/n)$.

At last, we use the hypothesis on the existence of a UEC test. By an argument due to Schwartz [48], in the proof of Theorem 6.1, the existence of a UEC test implies the existence of $r_o, r_1 > 0$ so that

$$P_{\theta_o}\left\{ p(X^n|\theta_o) \le e^{nr_o} \int_{N^c} w(\theta) p(X^n|\theta)\, d\theta \right\} \le e^{-nr_1}. \quad (6.6)$$

We can now obtain a bound on the probability of concern. Let $r \in (0, r_o)$ and set $r' = r_o - r$. Then, by use of U_n to set up (6.5) and (6.6), we have that

$$\begin{aligned} P_{\theta_o}&\left\{ \int_N w(\theta) p(X^n|\theta)\, d\theta \le e^{nr} \int_{N^c} w(\theta) p(X^n|\theta)\, d\theta \right\} \\ &\le P_{\theta_o}\left\{ U_n \cap \left\{ \int_N w(\theta) p(X^n|\theta)\, d\theta \right.\right. \\ &\qquad \left.\left. < e^{nr} \int_{N^c} w(\theta) p(X^n|\theta)\, d\theta \right\} \right\} + P_{\theta_o}(U_n^c) \\ &\le P_{\theta_o}\left\{ p(X^n|\theta_o) < e^{n(r+r')} \int_{N^c} w(\theta) p(X^n|\theta)\, d\theta \right\} \\ &\quad + P_{\theta_o}(U_n^c) \\ &\le e^{-nr_1} + O\left(\frac{1}{n}\right) \end{aligned}$$

which gives the desired result. □

These previous two propositions use mild hypotheses to guarantee posterior consistency at a good rate. Here, the key assumption was soundness. We conclude this section with a brief demonstration of soundness of exponential families.

Soundness of Exponential Families: As in Section II, consider families of probability densities of the exponential form $e^{-\theta^T \phi(x)} g(x)/c(\theta)$ with the natural parameter space $\Theta = \{\theta \in \boldsymbol{R}^d : c(\theta) < \infty\}$, where $c(\theta)$ is the normalizing constant. Setting $z = \phi(x)$ and choosing the appropriate dominating measure $\nu(dz)$, the family is expressed in standard exponential form as $p(z|\theta) = e^{-\theta^T z - \psi(\theta)}$, for $\theta \in \Theta$, with $\psi(\theta) = \log c(\theta)$ and $c(\theta) = \int e^{-\theta^T z} \nu(dz)$. The assumption that $\theta^T \phi(x)$ is nonconstant, unless $\theta = 0$, means that in the terminology of Brown [13], the standard exponential family is minimal.

Let θ_o be a point in the interior of Θ. By direct evaluation, the relative entropy between densities in the family is seen to equal $D(p_{\theta_o} \| p_\theta) = \psi(\theta) - \psi(\theta_o) + (\theta - \theta_o)^T E_{\theta_o} Z$ and $\nabla\psi(\theta_o) = -E_{\theta_o} Z$. From the continuity and convexity of ψ, it follows that $D(p_{\theta_o} \| p_\theta) \to 0$ if and only if $\theta \to \theta_o$. Now, $D(p_{\theta_o} \| p_\theta) \to 0$ implies $P_\theta \to P_{\theta_o}$. It remains to be shown that in exponential families, the reverse implication is also true.

Here, we assume that $E_{\theta_o} Z$ is in the interior of the support of Z, as is the case, in particular, if this support is convex. (In general, $E_{\theta_o} Z$ is in the interior of the convex hull of the support of Z, see Theorem 3.6 of Brown [13].) To prove soundness, we show that given any sequence of θ's that stays bounded away from θ_o, the sequence P_θ does not converge to P_{θ_o}. Given such a sequence of θ's, fix an orthant occupied infinitely often by $\theta - \theta_o$, and let A be the event that $Z - E_{\theta_o} Z$ is in that orthant. Then, restricting to the subsequence in the orthant, we have that $(\theta - \theta_o)^T (z - E_{\theta_o} Z)$ is positive for z in A, from which it follows that $p(z|\theta) \le p(z|\theta_o) e^{-D(p_{\theta_o} \| p_\theta)}$ for z in A. Consequently, $P_\theta(A) \le P_{\theta_o}(A) e^{-D(p_{\theta_o} \| p_\theta)}$. Since $P_{\theta_o}(A)$ is positive, it follows that if θ does not converge to θ_o, then P_θ cannot converge weakly to P_{θ_o}. This demonstrates the soundness condition.

We note also that since $D(p_{\theta_o} \| p_\theta)$ is a strictly convex function of θ, which is minimized at θ_o, the sets $\{\theta: D(p_{\theta_o} \| p_\theta) < r\}$ are compact. Therefore, if θ diverges from the family Θ, i.e., if the sequence of θ's is eventually outside any compact subset of Θ, then $D(p_{\theta_o} \| p_\theta)$ tends to infinity. By this reasoning, the weak limits of such sequences must assign zero measure to the set $\{Z - E_\theta Z \in A\}$ for some orthant A, whereas the P_{θ_o} measure of the set is nonzero. As mentioned in Section II, this degeneracy of divergent sequences provides another demonstration of the soundness condition.

REFERENCES

[1] J. Aitchison, "Goodness of prediction fit," *Biometrika*, vol. 62, no. 3, pp. 547–554, Dec. 1975.

[2] R. R. Bahadur, "Some limit theorems in statistics," in *Proc. Regional Conf. Series Appl. Mathematics*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1971.

[3] A. R. Barron, "Logically smooth density estimation," Ph.D. dissertation, Dept. Elec. Eng., Stanford Univ., Stanford, CA, Aug. 1985.

[4] ____, "Are Bayes rules consistent in information?" in *Problems in Communications and Computation* (T. M. Cover and B. Gopinath, Eds.). New York: Springer-Verlag, 1987, pp. 85–91.

[5] ____, "The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions," Tech. Report 7, Univ. of Illinois, Dept. of Statist., Apr. 1988.

[6] ____, "Uniformly powerful goodness of fit tests," *Ann. Statist.*, vol. 17, no. 1, pp. 107–124, Mar. 1989.

[7] A. R. Barron and T. M. Cover, "A bound on the financial value of information," *IEEE Trans. Inform. Theory*, vol. 34, pp. 1097–1100, Sept. 1988.

[8] ____, "Minimum complexity density estimation," to appear in *IEEE Trans. Inform. Theory*.

[9] R. H. Berk, "Limiting behavior of posterior distributions when the model is incorrect," *Ann. Statist.*, vol. 37, no. 1, pp. 51–58, Feb. 1966.

[10] ____, "Consistency *a posteriori*," *Ann. Statist.*, vol. 41, no. 3, pp. 894–906, June 1970.

[11] P. Bickel and J. Yahav, "Some contributions to the asymptotic

theory of Bayes' solutions," *Z. Wahrsch. verw. Gebiete*, vol. 11, pp. 257–276, 1969.
[12] R. E. Blahut, *Principles and Practice of Information Theory*. Reading, MA: Addison-Wesley, 1987.
[13] L. Brown, *Fundamentals of Statistical Exponential Families*. Hayward, CA: Institute of Mathematical Statistics, 1986.
[14] N. N. Čencov, *Statistical Decision Rules and Optimal Inference*. Providence: American Mathematical Society, 1981.
[15] H. Chernoff, "On the distribution of the likelihood ratio," *Ann. Math. Statist.*, vol. 25, no. 3, pp. 573–578, Sept. 1954.
[16] ____, "Large sample theory: parametric case," *Ann. Math. Statist.*, vol. 27, no. 1, pp. 1–22, Mar. 1956.
[17] K. L. Chung, *A Course in Probability Theory*. New York: Academic, 1974.
[18] B. Clarke, "Asymptotic cumulative risk and Bayes' risk under entropy loss, with applications," Ph.D. dissertation, Dept. Statistics, Univ. of Illinois, Urbana-Champaign, 1989.
[19] B. Clarke and A. R. Barron, "Information theoretic asymptotics of Bayes' methods." Univ. of Illinois, Department of Statistics Tech. Rep. 26, July 1989.
[20] I. Csiszár, "Information-type measures of difference of probability distributions and individual observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.
[21] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 783–795, Nov. 1973.
[22] N. G. De Bruijn, *Asymptotic Methods in Analysis*. New York: Dover, 1958.
[23] M. De Groot, *Optimal Statistical Decisions*. New York: McGraw-Hill, 1970.
[24] R. Gray, *Probability, Random Processes, and Ergodic Properties*. New York: Springer-Verlag, 1988.
[25] J. Hartigan, *Bayes' Theory*. New York: Springer-Verlag, 1983.
[26] D. Haughton, "On the choice of a model to fit data from an exponential family," *Ann. Statist.*, vol. 16, no. 1, pp. 342–355, Mar. 1988.
[27] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.*, vol. 58, pp. 13–30, Mar. 1963.
[28] W. Hoeffding and J. Wolfowitz, "Distinguishability of sets of distributions," *Ann. Math. Statist.*, vol. 29, no. 3, pp. 700–718, Sept. 1958.
[29] I. Ibragimov and R. Hasminskii, *Statistical Estimation: Asymptotic Theory*. New York: Springer-Verlag, 1980.
[30] R. A. Johnson, "Asymptotic expansions associated with the nth power of a density," *Ann. Statist.*, vol. 38, no. 4, pp. 1266–1272, Aug. 1967.
[31] R. A. Johnson, "Asymptotic expansions associated with posterior distributions," *Ann. Statist.*, vol. 41, no. 3, pp. 851–864, June 1970.
[32] J. Kelly, "New interpretation of information rate," *Bell Syst. Tech. J.*, vol. 35, pp. 917–926, July 1956.
[33] J. C. Kieffer and J. Wolfowitz, "On the deviations of the empiric distribution function of vector chance variables," *Trans. Amer. Math. Soc.*, vol. 87, pp. 173–186, 1958.
[34] ____, "A counterexample to Perez's generalization of the Shannon-McMillan theorem, *Ann. Probab.*, vol. 1, no. 2, pp. 362–364, Apr. 1973; "Correction," vol. 4, no. 1, pp. 153–154, Feb. 1976.
[35] ____, "A simple proof of the Moy-Perez generalization of the Shannon-McMillan theorem," *Pacific J. Math.*, vol. 51, pp. 203–206, 1974.
[36] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 199–207, Mar. 1981.
[37] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.
[38] S. Kullback, J. C. Keegel, and J. H. Kullback, *Topics in Statistical Information Theory*. Berlin: Springer-Verlag, 1980.
[39] P. S. Laplace, *Oevres*. Paris: Imprimerie Royale, 1847, vol. 7.
[40] L. Le Cam, "On some asymptotic properties of maximum likelihood and related Bayes' estimates," in *University of California Publications in Statistics* (J. Neyman, M. Loeve, and O. Struve, Eds.). London: Cambridge University Press, 1953, pp. 277–329, vol. 1.
[41] ____, "Les proprietes asymptotiques des solutions de Bayes," *Publ. Inst. Statist. Univ. Paris*, vol. 7, pp. 18–35, 1958.
[42] E. L. Lehmann, *Theory of Point Estimation*. New York: Wiley, 1983.
[43] T. Leonard, "Comment on 'A simple predictive density function,'" *J. Amer. Statist. Assoc.*, vol. 77, pp. 657–658, 1982.
[44] D. Pollard, "New ways to prove central limit theorems," *Econometric Theory*, vol. 1, pp. 295–314, 1985.
[45] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 629–636, July 1984.
[46] ____, "Stochastic complexity and modeling," *Ann. Statist.*, vol. 14, pp. 1080–1100, Sept. 1986.
[47] ____, "Stochastic complexity," *J. Royal Statist. Soc., Ser. B*, vol. 49, no. 3, pp. 223–239, 1987.
[48] L. Schwartz, "On Bayes' consistency," *Z. Wahrsch. verw. Gebiete*, vol. 4, pp. 10–26, 1965.
[49] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, Mar. 1978.
[50] S. M. Stigler, "Laplace's 1774 memoir on inverse probability," *Statistical Sci.*, vol. 1, no. 3, pp. 359–378, Aug. 1986.
[51] H. Strasser, "Consistency of maximum likelihood and Bayes' estimates," *Ann. Statist.*, vol. 9, no. 5, pp. 1107–1113, Sept. 1981.
[52] L. Tierney and J. Kadane, "Accurate approximations for posterior moments and marginal densities," Univ. of Minnesota Tech. Rep. 431, 1984.
[53] L. Tierney and J. Kadane, "Accurate approximations for posterior moments and marginal densities," *J. Amer. Statist. Assoc.*, vol. 81, pp. 82–86, Mar. 1986.
[54] V. Vapnik and A. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory Probab. Appl.*, vol. 16, pp. 264–280, 1971.
[55] R. Von Mises, *Mathematical Theory of Probability and Statistics*. (H. Geiringer, Ed.). New York: Academic, 1964.
[56] A. Wald, "Tests of statistical hypotheses concerning several parameters when the number of observations is large," *Trans. Amer. Math. Soc.*, vol. 54, pp. 426–482, Nov. 1943.
[57] ____, "Note on the consistency of the maximum likelihood estimate," *Ann. Math. Statist.*, vol. 20, no. 4, pp. 595–601, Dec. 1949.
[58] A. M. Walker, "On the asymptotic behaviour of posterior distributions," *J. Roy. Statist. Soc., Ser. B*, vol. 31, pp. 80–88, 1967.
[59] S. S. Wilks, *Mathematical Statistics*. New York: Wiley, 1962.
[60] J. Wolfowitz, "On Wald's proof of the consistency of the maximum likelihood estimate," *Ann. Math. Statist.*, vol. 20, no. 4, pp. 601–602, Dec. 1949.