Asymptotic Normality of the Posterior in Relative Entropy

Bertrand S. Clarke, *Member, IEEE*

***Abstract*—We show that the relative entropy between a posterior density formed from a smooth likelihood and prior and a limiting normal form tends to zero in the independent and identically distributed case. The mode of convergence is in probability and in mean. Applications to codelengths in stochastic complexity and to sample size selection are briefly discussed.**

***Index Terms*—Asymptotic normality, posterior density, relative entropy.**

I. Introduction

INTEREST in the asymptotic behavior of the posterior density derives chiefly from parameter estimation because it is the fundamental quantity from which a Bayesian makes inferences. This has led to numerous results ensuring that the posterior density for a parameter approaches a limiting normal form as the sample size increases. These results are useful because it is easier to make inferences from a limiting normal than to obtain the posterior density itself. Posterior normality is analogous to the Central Limit Theorem (CLT) in classical statistics which guarantees that a sample mean will give asymptotically good confidence intervals for the population mean. As a parallel to [2], we establish the asymptotic normality of the posterior in relative entropy, a mode of convergence stronger than modes used before or incompatible with them.

Originally, use of the relative entropy as a mode of convergence was motivated by its role as the redundancy in source coding. While this remains important for density estimation and stochastic complexity as evidenced by [5], [26], [27], [4], and [3], the relative entropy can be regarded purely as a mathematical distance between two distributions and used more generally, see [13]–[15], [8], and [23], amongst many others.

Denote the relative entropy between two densities p, q with respect to a common dominating measure by $D(p\|q) = \int p \log(p/q)$, where $\log$ means the natural logarithm to base e. We give conditions under which $D(w(\cdot|X^n)\|N(\hat{\theta}, (nI(\theta_0))^{-1})$ tends to zero as the sample size n increases. By Bayes rule, the posterior density is

$$w(\theta|x^n) = \frac{w(\theta)p(x^n|\theta)}{m(x^n)}$$

where $X^n = (X_1, \cdots, X_n)$ has outcomes $x^n = (x_1, \cdots, x_n)$ drawn independently from the likelihood $p(\cdot|\theta_0) = p_{\theta_0}(\cdot)$, $\hat{\theta}$ is the maximum-likelihood estimate (MLE), and $w(\theta)$ is a smooth prior density with respect to Lebesgue measure for $\theta \in \Omega \subset R^d$. The parameter space Ω is an open set and the mixture density, or marginal for the data, is $m(x^n) = \int w(\theta)p(x^n|\theta)\, d\theta$ where

$$p(x^n|\theta) = \prod_{i=1}^{n} p(x_i|\theta).$$

The notation $\hat{N} = N(\hat{\theta}, (nI(\theta_0))^{-1})$ indicates the normal density

$$\begin{aligned}\varphi(\theta) &= \varphi(\theta; \hat{\theta}, I(\theta_0), n)\\ &= \frac{|nI(\theta_0)|^{1/2}}{(2\pi)^{d/2}} e^{-(n/2)(\theta-\hat{\theta})I(\theta_0)\,(\theta-\hat{\theta})}\end{aligned}$$

where $I(\theta_0)$ is the Fisher information matrix for p_{θ_0}. Note that we omit the transposes on vectors when no confusion will result. The location $\hat{\theta}$ need not be the MLE as we use below in Theorem 2.1. In Theorem 2.2, $\hat{\theta}$ is the pseudo-estimator used in [22].

The main results here are two senses in which $D(w(\cdot|X^n)\|\hat{N})$ tends to zero. We show

$$D(w(\cdot|X^n)\|\hat{N}) \overset{P_{\theta_0}, L^1(P_{\theta_0})}{\longrightarrow} 0. \tag{1.1}$$

Results similar to (1.1) have been obtained by many authors. Reference [21] proved (1.1) using the L^1-distance in place of D. Reference [28] proved the posterior probabilities of intervals in the parameter space converge to normal probabilities in P_{θ_0}-probability. Other contributions have been made by [6], [20], [18], [17], and [7].

Results in [19], [16], and [12] imply asymptotic normality in a related mode of convergence. In particular, the Shannon mutual information between the parameter and a data set of size n

$$I(\Theta; X^n) = \int w(\theta)p(x^n|\theta)\log p(x^n|\theta)/m(x^n)\, d\theta\, dx^n$$

satisfies the identity

$$\begin{aligned}I(\Theta; X^n) = H(w) - \frac{1}{2}E_m \log(2\pi e)^d \det \mathrm{COV}\,(\Theta|X^n)\\ + E_m D(w(\cdot|X^n)\|N(E(\Theta|X^n), \mathrm{COV}(\Theta|X^n)))\end{aligned} \tag{1.2}$$

where $H(w) = \int w(\theta)\log 1/w(\theta)\, d\theta$ is the entropy of the prior w and E_m denotes expectation with respect to the

Manuscript received July 26, 1995; revised May 22, 1998. This work was supported in part by the NSERC under Operating Grant 5-54891.

The author is with the Department of Statistics, University of British Columbia, Vancouver, BC, Canada V6T 1Z2.

Communicated by A. R. Barron, Associate Editor for Nonparametric Estimation, Classification, and Neural Networks.

Publisher Item Identifier S 0018-9448(99)00073-5.

mixture. Using (1.2) one can obtain conditions under which its last term goes to zero. That is, one can derive

$$\int w(\theta)E_{\theta}D(w(\cdot|X^n)\|N(E(\Theta|X^n),\,\mathrm{COV}(\Theta|X^n)))\,d\theta\to 0$$

for a normal located at the posterior mean and having variance matrix the posterior covariance. Apart from a different, but asymptotically equivalent location and scale, the main difference between this result and (1.1) is that the mode of convergence is defined by m_n rather than $p_{\theta_0}^n$. The m_n mode was used by [1] and [24], where a CLT assumption is made. In fact, (1.1) is stronger since it corresponds to the expectation inside the integral. We note that the L^1 convergence in (1.1) is

$$\begin{aligned}E_{\theta_0}&D(w(\cdot|X^n)\|\hat{N})\\&=-E_{\theta_0}H(\Theta|X^n)-\frac{d}{2}\log\frac{n}{2\pi}-\frac{1}{2}\log|I(\theta_0)|\\&\quad+E_{\theta_0}\int w(\theta|X^n)\Big(\frac{n}{2}\Big)(\theta-\tilde{\theta})I(\theta_0)\,(\theta-\tilde{\theta})\,d\theta\\&=o(1).\end{aligned}$$

Here

$$E_{\theta_0}H(\Theta|X^n)=-E_{\theta_0}\int w(\theta|X^n)\log w(\theta|X^n)$$

is the expected conditional entropy. So, if

$$nE_{\theta_0}\int w(\theta|X^n)\,(\theta-\tilde{\theta})I(\theta_0)\,(\theta-\tilde{\theta})\,d\theta\to d$$

then we have

$$E_{\theta_0}H(\Theta|X^n)=-\frac{d}{2}\log\frac{n}{2\pi e}-\frac{1}{2}\log|I(\theta_0)|+o(1)$$

and conversely. Essentially, any two of these convergences implies the third. Related expansions occur in stochastic complexity, see [25]–[27], and [3].

Our results were also motivated by [5] who showed that

$$\sqrt{2h_n}n(D(f\|f_n)-E_fD(f\|f_n))\xrightarrow{\mathcal{L}}N(0,\sigma^2)\qquad(1.3)$$

where f_n is a nonparametric density estimator for the true density f, h_n is a decreasing sequence, and σ^2 can be identified. Here we have only shown that $D(w(\cdot|X^n)\|\hat{N})$ and $E_{\theta_0}D(w(\cdot|X^n)\|\hat{N})$ go to zero. Nevertheless, in view of (1.3) one might expect that there will be a sequence $\langle a_n\rangle|_{n=1}^{\infty}$ so that

$$a_n(D(w(\cdot|X^n)\|\hat{N})-E_{\theta_0}D(w(\cdot|X^n)\|\hat{N}))=O_P(1)\quad(1.4)$$

in P_{θ_0}-probability, and the left-hand side will converge in distribution to a recognizable limit. Recall that locally D behaves like the square of a distance. That is, on parametric families, $D(p_\theta\|p_{\theta'})$ is locally approximated by $(1/2)\,(\theta-\theta')I(\theta)\,(\theta-\theta')$. This suggests we anticipate the relative entropy between two densities based on independent data, when it goes to zero, should behave like the square of $O\,(1/\sqrt{n})$. In (1.4) we have two such relative entropies; each term in (1.4) may be conjectured to go to zero at rate $O\,(1/n)$. A difference between two relative entropies, each of which is conjectured to go to zero at rate $O\,(1/n)$, may go to zero faster. In (1.4), one of the two relative entropies is the mean of the other, so we expect an extra $O\,(1/\sqrt{n})$ factor, in parallel to the Central Limit Theorem. Now, we conjecture that $a_n=O(n^{3/2})$ is an appropriate choice. In fact, this choice gives (1.4) in simple examples such as the normal with a normal conjugate prior. More generally, exponential families with conjugate priors should satisfy (1.4) but we have not demonstrated this.

The structure of this paper is as follows. In the next section we formally state our main results. In Section III, we begin by presenting the simplest example and then briefly describe two application of our results. Section IV contains the proof of the main theorem. Convergence of the posterior to the normal in expected relative entropy is stated and proved in the Appendix.

II. Statement of Results

Formally, we suppose the density $p(x|\theta)$ admits at least two continuous derivatives in θ for almost every x and that the Fisher information $I(\theta)$ exists and is strictly positive for $\theta\in\Omega$, an open subset of R^d. In addition, we assume that for any $\theta\in\Omega$ there is a $\delta=\delta(\theta)>0$ so that

$$E_\theta\sup_{|\theta-\theta'|\le\delta}\left|\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p(X_1|\theta')\right|^2<\infty$$

for $i,j=1,\cdots,d$ and for $i=1,\cdots,d$

$$E_\theta\sup_{|\theta-\theta'|\le\delta}\left|\frac{\partial}{\partial\theta_i}\log p(X_1|\theta')\right|<\infty.$$

We assume that the parametric family p_θ is soundly parametrized in the sense that convergence of a sequence of parameter values in Euclidean distance is equivalent to weak convergence of the distributions they index. That is, $\|\theta'-\theta\|\to 0$ in Euclidean distance on R^d if and only if $F_{\theta'}(x)$, the distribution function for X_1 when θ' is true, converges to F_θ, the distribution function for X_1 when θ is true, for every x at which F_θ. This is a relatively weak identifiability criterion—it ensures that two distributions can only be close when their parameter values are close. Details on this assumption and verification that it is satisfied by parametric families in exponential form can be found in [11].

Expectations are with respect to a fixed but arbitrary p_{θ_0} with $\theta_0\in\Omega$, denoted by $E=E_{\theta_0}$ unless otherwise indicated. The prior density w is with respect to Lebesgue measure on R^d and has support contained in Ω. We assume $w(\theta)>0$ on a neighborhood of the $\theta_0\in\Omega$ taken as "true" and is bounded on Ω. Also, we suppose that w admits a second-order Taylor expansion at θ_0, and denote the prior distribution by $W(\cdot)$. The posterior distribution is $W(\cdot|x^n)$.

First, we establish Proposition 2.1, a large deviation principle for posterior probabilities for use in the proofs of the two theorems below. Essentially, Proposition 2.1 is a refinement of [11, Theorem 2.2] to give a tighter bound on the P_{θ_0}-probability that the posterior probability of a neighborhood excluding the true value is small. We show this probability is $O\,(e^{-\xi n})$ for some $\xi>0$ rather than $O\,(1/n)$. Let

$$B(\theta_0,\delta)=\{\theta|(\theta-\theta_0)I(\theta_0)\,(\theta-\theta_0)\le\delta^2\}$$

for given $\delta>0$. We have the following proposition.

Proposition 2.1: Assume that the Renyi relative entropy of order $1+\lambda$

$$\frac{1}{\lambda}\log\int p(x|\theta_o)\left(\frac{p(x|\theta_o)}{p(x|\theta)}\right)^{\lambda}dx$$

is bounded in θ for some neighborhood of θ_o for some $\lambda > 0$. Then for any open neighborhood N of θ_0, there is an $r > 0$ and a $\rho > 0$ so that

$$P_{\theta_0}\left(\int_N w(\theta)p(X^n|\theta) < e^{nr}\int_{N^c} w(\theta)p(X^n|\theta)\,d\theta\right) = O(e^{-n\rho}). \quad (2.1)$$

Proof: The technique of proof is to set up an application of Jensen's inequality and then recognize a large deviation. We want an exponential bound $e^{-n\rho}$ on

$$P_{\theta_o}\left(\frac{1}{n}\log\frac{p(X^n|\theta_o)}{m(X^n|B(\theta_0,\delta))} > r'\right)$$

where $m(X^n|B(\theta_0,\delta))$ is the mixture of likelihoods with respect to the normalised restriction of w to $B(\theta_o,\delta)$, i.e.,

$$w(\theta|B(\theta_o,\delta)) = w(\theta)/W(B(\theta_o,\delta)).$$

Applying Jensen's inequality on the logarithm inside gives the upper bound

$$P_{\theta_o}\left(\frac{1}{n}\int\log\frac{p(X^n|\theta_o)}{p(X^n|\theta)}w(\theta|B(\theta_0,\delta))\,d\theta > r'\right) = P_{\theta_o}\left(\frac{1}{n}\sum_{i=1}^{n}g(X_i) > r'\right). \quad (2.2)$$

Expression (2.2) is the tail probability for a sample average of random variables

$$g(X_i) = \int\log\frac{p(X_i|\theta_o)}{p(X_i|\theta)}w(\theta|B(\theta_0,\delta))\,d\theta$$

that have expectation $\int D(p_{\theta_o}\|p_\theta)w(\theta|B(\theta_0,\delta))\,d\theta$. This expectation is less than any fixed r' for δ sufficiently small. An exponential bound on (2.2) is a large deviation inequality. So, from the standard Cramer–Chernoff large deviations theory it is enough to show that

$$\int p(x|\theta_o)e^{\lambda g(x)}\mu(dx)$$

the moment generating function for $g(X_1)$ is finite for some $\lambda > 0$. We show that this follows from the Renyi entropy assumption.

Applying Jensen's inequality to $g(x)$ gives

$$e^{\lambda g(x)} \leq \int (p(x|\theta_o)/p(x|\theta))^{\lambda}w(\theta|B(\theta_0,\delta))\,d\theta.$$

This means the moment generating function of g is bounded by

$$\int\left(\int p_{\theta_o}(p_{\theta_o}/p_\theta)^{\lambda}w(\theta|B(\theta_0,\delta))\,d\theta\right)dx$$

which in turn is bounded using the Renyi relative entropy for θ in a neighborhood of θ_o. □

This Proposition also controls expectations of posterior probabilities which may arise in L^1 convergence. Indeed, Proposition 2.1 implies that for any open neighborhood N of θ_0

$$P_{\theta_0}(W(N^c|X^n) \geq e^{-nr}) = O(e^{-n\rho}) \quad (2.3)$$

for some choice of $r,\rho > 0$. Consequently, we have that for any $\beta > 0$

$$\begin{aligned}E_{\theta_0}W(N^c|X^n)^{\beta} &= E_{\theta_0}W(N^c|X^n)^{\beta}\chi_{\{W(N^c|X^n) < e^{-nr}\}}\\ &\quad + E_{\theta_0}W(N^c|X^n)^{\beta}\chi_{\{W(N^c|X^n)\geq e^{-nr}\}}\\ &\leq O(e^{-nr\beta}) + O(e^{-n\rho}) \quad (2.4)\end{aligned}$$

since $W(N^c|X^n) \leq 1$. Thus there is a $\xi > 0$ so that

$$E_{\theta_0}W(N^c|X^n)^{\beta} = O(e^{-n\xi}).$$

Theorem 2.1: Assume the conclusion of Proposition 2.1 and the regularity conditions at the beginning of this section. Also, assume the [29] conditions for the consistency of the maximum-likelihood estimator $\hat{\theta} = \text{MLE}(\theta)$ and that $|\theta|^2$ has a finite mean under w. Then

$$D(w(\cdot|X^n)\|N(\hat{\theta},(nI(\theta_0))^{-1})) \xrightarrow{P_{\theta_0}} 0. \quad (2.5)$$

Proof: See Section IV. □

We remark that [29] consistency hypotheses for the MLE can be expressed as follows. For each θ_0 assume there is a $\rho = \rho(\theta_0)$ so large that

$$E_{\theta_0}\sup_{|\theta-\theta_0|>\rho}\log\frac{p(X|\theta)}{p(X|\theta_0)} < \infty \quad (2.6a)$$

and for each θ there is a $\delta = \delta(\theta) > 0$ so small that

$$E_{\theta_0}\sup_{\theta':|\theta-\theta'|<\delta}\log\frac{p(X|\theta')}{p(X|\theta_0)} < \infty \quad (2.6b)$$

and that for any x, $p(x|\theta) \to 0$ as $\|x\| \to \infty$.

In the Appendix, Theorem A.1 gives sufficient conditions for the L^1 convergence to zero of the relative entropy between the posterior and the limiting normal. For measures of distance weaker than the relative entropy, L^1 convergence may follow from the convergence in probability in (2.5). This is the case if the L^1 distance itself is used in place of the relative entropy. Here, Theorem A.1 has much stronger hypotheses than Theorem 2.1 and so is relegated to the Appendix.

The proofs of both Theorems 2.1 and A.1 below rely on the same decomposition of $D(w(\cdot|X^n)\|\hat{N})$ into three terms. Write

$$\begin{aligned}D(w(\cdot|X^n)\|\hat{N}) &= \int_{B(\theta_0,\varepsilon)} w(\theta|X^n)\log\frac{w(\theta|X^n)}{\varphi(\theta)}d\theta \quad (2.7)\\ &\quad + \int_{B^c(\theta_0,\varepsilon)} w(\theta|X^n)\log w(\theta|X^n)\,d\theta \quad (2.8)\\ &\quad - \int_{B^c(\theta_0,\varepsilon)} w(\theta|X^n)\log\varphi(\theta)\,d\theta. \quad (2.9)\end{aligned}$$

To prove the theorems it is enough to show that the three terms go to zero in probability and in mean. The hypotheses arise from bounding each term by something which goes to zero, in the appropriate mode as n increases.

For Theorem 2.1, in addition to other techniques, we use Proposition 2.1 to control the error terms (2.8) and (2.9). To control the local behavior (2.7) we use a Laplace's method argument, necessitating the expected-supremum hypotheses at the beginning of this section, and the consistency of the MLE.

III. An Example and Two Applications

Consider the simplest example of an independent and identically distributed (i.i.d.) $N(\mu, 1)$ likelihood with an $N(0, 1)$ prior on μ. It is straightforward to verify that the hypotheses of Theorems 2.1 and A.1 are satisfied. Now $\hat{N} = N(\overline{x}, 1/n)$ and the posterior is

$$w(\theta|x^n) = N((n/n+1)\overline{x}, 1/(n+1)).$$

So, the relative entropy between them is

$$\begin{aligned} &D(w(\cdot|x^n)\|\hat{N}) \\ &\quad = \int \sqrt{\frac{n+1}{2\pi}} e^{-((n+1/2))\,((n/n+1)\overline{x}-\mu)^2} \\ &\qquad \cdot \log\sqrt{\frac{n+1}{n}} e^{-((n+1/2))\,((n/n+1)\overline{x}-\mu)^2+(n/2)\,(\overline{x}-\mu)^2}\, d\mu \\ &\quad = \frac{1}{2}\log\left(1+\frac{1}{n}\right)+\int\sqrt{\frac{n+1}{2\pi}} e^{-((n+1/2))\,((n/n+1)\overline{x}-\mu)^2} \\ &\qquad \cdot \left(\left(\frac{n}{2}\right)(\overline{x}-\mu)^2-\left(\frac{n+1}{2}\right)\left(\frac{n}{n+1}\overline{x}-\mu)^2\right)\, d\mu. \end{aligned}$$

Recognizing that the second term in the difference gives a variance, we see

$$\begin{aligned} D(w(\cdot|x^n)\|\hat{N}) = &\frac{1}{2}\log\left(1+\frac{1}{n}\right)-\left(\frac{n+1}{2}\right)\left(\frac{1}{n+1}\right) \\ &+\left(\frac{n}{2}\right)\int\sqrt{\frac{n+1}{2\pi}} e^{-((n+1/2))((n/n+1)\overline{x}-\mu)^2} \\ &\cdot(\overline{x}-\mu)^2\, d\mu. \end{aligned}$$

Writing

$$\begin{aligned} (\overline{x}-\mu)^2 = &\left(\frac{n}{n+1}\overline{x}-\mu\right)^2+\frac{\overline{x}^2}{(n+1)^2} \\ &+2\left(\frac{n}{n+1}\overline{x}-\mu\right)\left(\overline{x}-\frac{n}{n+1}\overline{x}\right) \end{aligned}$$

and noting that the expectation of the first term is a variance and the expectation of the last term is zero, we get

$$\begin{aligned} &D(w(\cdot|x^n)\|\hat{N}) \\ &\quad = \frac{1}{2}\log\left(1+\frac{1}{n}\right)-1/2+\left(\frac{n}{2}\right)\left(\frac{1}{n+1}+\frac{\overline{x}^2}{(n+1)^2}\right) \\ &\quad = \frac{1}{2}\log\left(1+\frac{1}{n}\right)-\frac{1}{2(n+1)}+\frac{n\overline{x}^2}{2(n+1)^2}. \end{aligned}$$

From the last expression it is seen that $2nD(w(\cdot|x^n)\|\hat{N})$ is asymptotically equivalent to $\overline{x}^2$, which converges to μ^2 in probability.

If μ_0 is regarded as the true value then

$$E_{\mu_0}\overline{X}^2 = (1/n)\,(1+n\mu_0^2).$$

So,

$$\begin{aligned} &E_{\mu_0}D(w(\cdot|X^n)\|\hat{N}) \\ &\quad = \frac{1}{2}\log\left(1+\frac{1}{n}\right)-\frac{1}{2(n+1)}+\frac{(1+\mu_0^2 n)}{2(n+1)^2}. \end{aligned}$$

Thus it is seen that $2nE_{\mu_0}D(w(\cdot|X^n)\|\hat{N})$ converges to μ_0^2. Also, one has that

$$\begin{aligned} &D(w(\cdot|X^n)\|\hat{N})-E_{\mu_0}D(w(\cdot|X^n)\|\hat{N}) \\ &\qquad = \frac{n(\overline{X}^2-\mu_0^2)}{2(n+1)^2}-\frac{1}{2(n+1)^2}. \end{aligned}$$

Since $\overline{x}^2 = (\overline{x}-\mu_0)^2+\mu_0^2+2\mu_0(\overline{x}-\mu_0)$ we have

$$\begin{aligned} &n^{3/2}(D(w(\cdot|X^n)\|\hat{N})-E_{\mu_0}D(w(\cdot|X^n)\|\hat{N})) \\ &\quad = \frac{-n^{3/2}}{2(n+1)^2}+\frac{n^2}{2(n+1)^2} \\ &\qquad \cdot(\sqrt{n}(\overline{X}-\mu_0)^2+2\mu_0\sqrt{n}(\overline{X}-\mu_0)). \end{aligned}$$

The first term is $O(1/\sqrt{n})$, the second term is $O_p(1/\sqrt{n})$, and the third term is of the form $\mu_0 N(0,1)+O_p(1\sqrt{n})$, the sum of which is of the form (1.4).

This example motivates the conjecture $a_n = O(n^{3/2})$ for (1.4). Moreover, examination of the proofs of Theorems 2.1 and A.1 reveals that the impediment to establishing this rate in general arises from the local term (2.7). It can be seen from the proofs that the error terms (2.8) and (2.9) go to zero at a rate $O_P(e^{-\xi n})$ for Theorem 2.1 and $O(e^{-\xi n})$ for Theorem A.1, for some $\xi > 0$.

The left-hand side of identity (1.2) can be interpreted as a Bayes redundancy or as a Shannon mutual information. The right-hand side is essentially determined by asymptotic normality of the posterior. This demonstrates a close connection between source coding in particular and the behavior of the posterior. The importance of asymptotic normality in coding and stochastic complexity has been recognized by [24]–[26], but the sense used there was the asymptotic normality of the MLE. Identity (1.2) is a Bayesian form, interpretable as in ([3, Sec. III]).

It is well known that the Bayes code in a parametric context is given by the integer part of $\log(1/m(x^n))$, to within one-bit accuracy. Reference [25] has shown that

$$\log\frac{1}{m(x^n)} = \log\frac{1}{p(x^n|\hat{\theta})}+\frac{d}{2}\log n+O(1) \tag{3.1}$$

uniformly in x^n, see also [1]. More recently, the stochastic complexity of a string x^n relative to a class of stochastic processes has been identified as

$$L(x^n) = \log\frac{1}{f(x^n|\hat{\theta})}+\frac{d}{2}\log\frac{n}{2\pi}+\log\int\sqrt{|I(\theta)|}\,d\theta+o(1) \tag{3.2}$$

see [26]. The main assumption in [26] is that the MLE $\hat{\theta}$ satisfy a Central Limit Theorem. In this context, Theorems

2.1 and A.1 can be restated to give the approximation to the Bayes codelengths

$$\begin{aligned}\log\frac{1}{m(x^n)} =& E\left(\log\frac{1}{p(x^n|\Theta)}\middle| x^n\right)+\frac{d}{2}\log\frac{n}{2\pi}\\ &+E\left(\log\frac{1}{w(\Theta)}\middle| x^n\right)+\frac{1}{2}\log|I(\theta_0)|\\ &-\frac{n}{2}\int(\theta-\hat{\theta})I(\theta_0)(\theta-\hat{\theta})w(\theta|x^n)\,d\theta+o(1)\end{aligned} \tag{3.3}$$

where the $o(1)$ is in P_{θ_0}-probability, or in $L^1(P_{\theta_0})$. The analogous statement for convergence assessed in the mixture follows from examining (1.2) and introduces an integration over the logarithm of the Fisher information. Apart from the mode of convergence and replacing the MLE with a posterior mean, the main difference between (3.2) and (3.3) is in the constants. It is seen that (1.4) suggests a smaller error may be obtainable.

Another implication of the theorems here is for sample-size selection for experimental design. Consider a fixed but arbitrary density for θ, say $q(\theta)$, and suppose that q is sufficiently concentrated around some central value that credibility sets of θ from it would provide satisfactory inferences. Now, Theorem A.1 implies that for any θ_0 we can find

$$N(\theta_0,q)=\arg\min_n E_{\theta_0}D(w(\cdot|X^n)\|q(\cdot)). \tag{3.4}$$

For given θ_0 and q, $N(\theta_0,q)$ is the sample size for which the posterior will be as close as possible to q when averaged with respect to P_{θ_0}.

It is seen that this $N(\theta_0,q)$ exists by writing

$$\begin{aligned}E_{\theta_0}D(w(\cdot|X^n)\|q) =& E_{\theta_0}D(w(\cdot|X^n)\|\hat{N})\\ &+E_{\theta_0}\int w(\theta|X^n)\log\frac{\varphi(\theta)}{q(\theta)}d\theta.\end{aligned}$$

By Theorem A.1, the first term tends to zero. Using the definition of φ gives that the second term is

$$\begin{aligned}\frac{d}{2}\log\frac{n}{2\pi}+\frac{1}{2}\log|I(\theta_0)|+E_{\theta_o}\int w(\theta|X^n)\log 1/q(\theta)&\\ -E_{\theta_o}\frac{n}{2}(\theta-\hat{\theta})I(\theta_o)(\theta-\hat{\theta}).&\end{aligned}$$

Posterior concentration at θ_o gives $\log 1/q(\theta_o)$ for the third term as n increases, and the last term tends to $-d/2$ by techniques similar to those used in the proof of Theorem A.1. Thus we get

$$\frac{d}{2}\log\frac{n}{2\pi e}+\frac{1}{2}\log|I(\theta_0)|+\log 1/q(\theta_0)+o(1)$$

i.e., it is increasing in n for n large.

Permitting θ_0 to vary over a plausible range of parameter values and q to vary over the collection of densities for θ which give credible sets of specified length means that $\max N(\theta_0,q)$ is a sample size sufficiently large to give, on average, inferences of the strength represented by the class of q's. This method is similar to that used in [10] in a pointwise sense.

Using Jeffreys noninformative prior in place of w gives a reference sample size; use of any other prior w gives the sample size required so that the data added to w will give a posterior sufficiently tight on average. Optimizing by using m_n in place of $p^n_{\theta_0}$ in (3.4) gives a degenerate result. Indeed,

$$E_mD(w(\cdot|X^n)\|q)=I(\Theta;X^n)+D(w\|q)$$

which is strictly increasing in n. An asymptotic expression for $E_{\theta_0}D(w(\cdot|X^n)\|q)$ accurate to order $o(1/n)$ can be differentiated to give an optimal sample size for given q and θ_0; we hope to pursue this elsewhere.

IV. Proof of the Theorem

To prove Theorem 2.1 we recall the decomposition at the end of Section II. Using it, (2.5) is

$$\int_{B(\theta_o,\varepsilon)} w(\theta|X^n)\log\frac{w(\theta|X^n)}{\varphi(\theta)} \tag{4.1}$$

$$+\int_{B^c(\theta_o,\varepsilon)} w(\theta|X^n)\log w(\theta|X^n)\,d\theta \tag{4.2}$$

$$-\int_{B^c(\theta_o,\varepsilon)} w(\theta|X^n)\log\varphi(\theta)\,d\theta. \tag{4.3}$$

It is enough to bound (4.1)–(4.3) from above, by quantities which go to zero.

We begin with (4.2), the first of the two error terms. It is enough to show that for $\varepsilon>0$ we have

$$P_{\theta_0}\left(\int_{B^c(\theta_0,\delta)} w(\theta|X^n)\log w(\theta|X^n)\,d\theta>\varepsilon\right)\longrightarrow 0 \tag{4.4}$$

as n increases.

Clearly, $p(x^n|\theta)\le e^{-n\hat{H}(\hat{\theta})}$ where $\hat{\theta}$ is the MLE and

$$\hat{H}(\theta)=(1/n)\log(1/p(x^n|\theta)).$$

Let $w_B(\theta)$ denote the restriction of w to $B(\theta_o,\varepsilon)$. Lower bounding the mixture of densities gives

$$\begin{aligned}&\int w(\theta)p(x^n|\theta)\,d\theta\\ &\quad\ge\int_{B(\theta_0,\varepsilon)} w(\theta)p(x^n|n)\,d\theta\\ &\quad=\int_{B(\theta_0,\varepsilon)}\frac{w(\theta)}{W(B(\theta_0,\varepsilon))}e^{-n\hat{H}(\theta)}\,d\theta W(B(\theta_0,\varepsilon))\\ &\quad\ge e^{-n\int_{B(\theta_0,\varepsilon)}w_B(\theta)\hat{H}(\theta)\,d\theta}W(B(\theta_0,\varepsilon))\end{aligned}$$

by use of Jensen's inequality for the exponential function. Now

$$\log w(\theta|x^n)\le\log\frac{w(\theta)e^{-n\hat{H}(\hat{\theta})}}{W(B(\theta_0,\varepsilon))e^{-n\int_{B(\theta_0,\varepsilon)}w_B(\theta)\hat{H}(\theta)\,d\theta}}. \tag{4.5}$$

Since w is bounded, $\log w(\theta)\le M$ for some $M>0$. So, intersect the event in (4.4) with the event $|\hat{\theta}-\theta_0|\le\varepsilon$ and its complement. This gives two terms whose sum bounds (4.4). One of the terms, the intersection with the complement, has probability going to zero by consistency of the MLE. The other term is bounded by

$$\begin{aligned}P_{\theta_0}\Bigg(&|\hat{\theta}-\theta_o|\le\varepsilon \text{ and } W(B^c(\theta_0,\varepsilon)|X^n)\\ &\cdot (M-\log W(B(\theta_0,\varepsilon)))+nW(B^c(\theta_0,\varepsilon)|X^n)\\ &\cdot\left(\int_{B(\theta_0,\varepsilon)} w_{B(\theta_0,\varepsilon)}(\theta)\hat{H}(\theta)\,d\theta-\hat{H}(\hat{\theta})\right)>\varepsilon\Bigg). \qquad (4.6)\end{aligned}$$

The first term in the event in (4.6) goes to zero in P_{θ_0} probability by Bayes consistency. Proposition 2.1 gives that $nW(B^c(\theta_0,\varepsilon)|X^n)$ goes to zero in P_{θ_0} probability. By Slutsky's theorem, to finish the proof that (4.5) goes to zero, it is enough to show that

$$\left(\int_{B(\theta_0,\varepsilon)} w_{B(\theta_0,\varepsilon)}(\theta)\hat{H}(\theta)\,d\theta-\hat{H}(\hat{\theta})\right)$$

is bounded in P_{θ_0} probability as n increases.

On $|\hat{\theta}-\theta_0|<\varepsilon$ we have

$$|\hat{H}(\hat{\theta})|\le\frac{1}{n}\left|\sum_{i=1}^{n}\log p(x_i|\hat{\theta})\right|\le\frac{1}{n}\sum_{i=1}^{n}\sup_{|\theta'-\theta_0|<\varepsilon}|\log p(x|\theta')|$$

and for $\theta\in B(\theta_0,\varepsilon)$ we have

$$|\hat{H}(\theta)|\le\frac{1}{n}\sum_{i=1}^{n}\sup_{|\theta'-\theta_0|<\varepsilon}|\log p(x|\theta')|. \qquad (4.7)$$

Expression (4.7) has a finite mean by assumption. Thus

$$\left(\int_{B(\theta_0,\varepsilon)} w_{B(\theta_0,\varepsilon)}(\theta)\hat{H}(\theta)\,d\theta-\hat{H}(\hat{\theta})\right)$$

is bounded above and below by random variables which converge in probability to constants. In particular,

$$\left(\int_{B(\theta_0,\varepsilon)} w_{B(\theta_0,\varepsilon)}(\theta)\hat{H}(\theta)\,d\theta-\hat{H}(\hat{\theta})\right)$$

is bounded in probability. This means that (4.6) goes to zero and hence so does (4.4).

Next we examine (4.3), the other error term. It is

$$\begin{aligned}-W(B^c(\theta_o,\varepsilon)|X^n)\,&\frac{d}{2}\log\left(\frac{n|I(\theta_o)|}{2\pi}\right)\\ &+\int_{B(\theta_o,\varepsilon)^c} w(\theta|X^n)\left(\frac{n}{2}\right)(\theta-\hat{\theta})I(\theta_o)\,(\theta-\hat{\theta})\,d\theta. \qquad (4.8)\end{aligned}$$

The first term in (4.8) goes to zero in P_{θ_o}-probability by noting that for $\varepsilon'>0$ there is an $r>0$ so that $P_{\theta_o}(W(B^c(\theta_o,\varepsilon)|X^n)\log n>\varepsilon')$ is bounded above by

$$\begin{aligned}P_{\theta_o}\Bigg(&(\log n)\int_{B^c(\theta_o,\varepsilon)} w(\theta)p(X^n|\theta)\,d\theta\\ &>\varepsilon'\int_{B(\theta_o,\varepsilon)} w(\theta)p(X^n|\theta)\,d\theta\Bigg)\\ \le P_{\theta_o}\Bigg(&e^{nr}\int_{B^c(\theta_o,\varepsilon)} w(\theta)p(X^n|\theta)\,d\theta\\ &>\int_{B(\theta_o,\varepsilon)} w(\theta)p(X^n|\theta)\,d\theta\Bigg). \qquad (4.9)\end{aligned}$$

[11, Proposition 6.3] shows that the right-hand side of (4.9) goes to zero at rate $O(1/n)$.

For the second term in (4.8), set $\nu=\sqrt{n}(\theta-\hat{\theta})$ and add and subtract $\phi(\nu)$ in the integrand, where $\phi(\nu)$ is the normal density for ν with mean zero and variance $I^{-1}(\theta)$. Jensen's inequality for absolute value now gives that (4.8) is bounded from above by

$$\begin{aligned}\int \nu^t\nu&\left|\frac{w(\hat{\theta}+\nu/\sqrt{n}|X^n)}{n^{d/2}}-\phi(\nu)\right|d\nu\\ &+\int_{|\hat{\theta}-\theta_o-\nu/\sqrt{n}|>\varepsilon}\nu^t\nu\phi(\nu)\,d\nu. \qquad (4.10)\end{aligned}$$

The first term goes to zero in P_{θ_o}-probability: This follows from correcting the proof of [6, Theorem 2.2], or from [9, Proposition 3.4.1]. (We comment that the small error in the work of [6] admits a simple correction: Theorem 2.2 is corrected by replacing expression (2.14) in assumption A.29, p. 261 by the three Wald assumptions recorded here, namely, (2.6a,b) and $p(x|\theta)\to 0$ as $||x||\to\infty$. These three assumptions are needed in [6, Lemma 2.6] for the Wald covering argument. Their proof [6, Lemma 2.6, pp. 263–264], is corrected by covering $B(\theta_o,\epsilon)^c$ by finitely many sets of the form $B(\theta_i,\epsilon_i)$ for $i=1,\cdots,k$ with $\epsilon_i>0$ on which (2.6b) is satisfied and a single set $B(\theta_o,\rho)^c$ where ρ is large enough to satisfy (2.6a).)

The second term in (4.10) is a normal tail probability that also goes to zero. Indeed, in more familiar notation it is

$$\int_{|\theta-\theta_o|>\varepsilon} n(\theta-\hat{\theta})^t(\theta-\hat{\theta})\frac{|I(\theta_o)|}{(2\pi)^{d/2}}e^{-(n/2)(\theta-\hat{\theta})I(\theta_o)(\theta-\hat{\theta})}\,d\theta. \qquad (4.11)$$

Now, by consistency of the MLE, if $|\theta_o-\hat{\theta}|<\varepsilon/2$ and we have that $|\theta-\theta_o|>\varepsilon$, then the triangle inequaliy gives $|\theta-\hat{\theta}|>\varepsilon/2$. Using

$$(\theta-\hat{\theta})I(\theta_o)(\theta-\hat{\theta})\le(1/2)(\theta-\hat{\theta})I(\theta_o)(\theta-\hat{\theta})+\varepsilon/2$$

in (4.11) gives a factor $e^{-n\varepsilon/4}$ which goes to zero.

Finally, we show the local term (4.1) goes to zero in probability. First note that for $\varepsilon>0\exists\rho_\varepsilon>0$ so that

$$B(\theta_o,\varepsilon)\subseteq B(\hat{\theta},\rho_\varepsilon)=\{(\theta-\hat{\theta})I^*(\hat{\theta})(\theta-\hat{\theta})<\rho_\varepsilon\} \qquad (4.12)$$

with P_{θ_o}-probability at least $1-O(1/n)$. This follows from using [11, eq. (4.16)] for the Fisher information and setting up an application of Chebyshev's inequality on

$$P_{\theta_o}\left(\sup_{|\theta-\theta_o|>\delta}\log\frac{p(X^n|\theta)}{p(X^n|\theta_o)}>-n\varepsilon\right)$$

in a [29] style bounding argument.

Now (4.1) is

$$\begin{aligned}\int_{B(\theta_o,\varepsilon)} &w(\theta|X^n)\log\frac{w(\theta|X^n)}{\varphi(\theta)}d\theta\\ &\le\int_{B(\hat{\theta},\rho_\varepsilon)} w(\theta|X^n)\,(\log w(\theta|X^n))^+\,d\theta\\ &\quad-\int_{B(\theta_o,\varepsilon)}\Bigg(\frac{d}{2}\log\frac{n}{2\pi}+\frac{1}{2}\log|I(\theta_o)|\\ &\quad-\frac{n}{2}\,(\theta-\hat{\theta})I(\theta_o)\,(\theta-\hat{\theta})\Bigg)w(\theta|X^n)\,d\theta \qquad (4.13)\end{aligned}$$

with P_{θ_o} probability $1-O\,(1/n)$ where a superscript $+$ denotes the positive part of a random variable. The second term is

$$-\left(\frac{d}{2}\log\frac{n}{2\pi}+\frac{1}{2}\log|I(\theta_o)|-\frac{d}{2}+\varepsilon\right)$$
$$+\int_{B^c(\theta_o,\varepsilon)}\left(\frac{d}{2}\log\frac{n}{2\pi}+\frac{1}{2}\log|I(\theta_o)|\right.$$
$$\left.-\frac{n}{2}\,(\theta-\hat{\theta})I(\theta_o)\,(\theta-\hat{\theta})\right)w(\theta|X^n)\,d\theta \tag{4.14}$$

the second term of which $\xrightarrow{P_{\theta_o}} 0$ by the reasoning following expression (4.9). So, it is enough to show that the first term in (4.13) added to the first term in (4.14) goes to zero in P_{θ_o}-probability.

Since we want to bound $w(\theta|x^n)=w(\theta)p(x^n|\theta)/m(x^n)$ from above, we start by following [28] and bounding $m(x^n)$ from below by Laplace's method. Restricting the domain of integration in the definition of $m(x^n)$ to $B(\hat{\theta},\rho_\varepsilon)$ we deal with the integrand locally. On $B(\hat{\theta},\varepsilon)$ we assume that $\tau>0$ has been chosen so that

$$(1-\tau)\,(\theta-\hat{\theta})I^*(\hat{\theta})\,(\theta-\hat{\theta})\le(\theta-\hat{\theta})I^*(\theta^{**})\,(\theta-\hat{\theta})$$
$$\le(1+\tau)\,(\theta-\hat{\theta})I^*(\hat{\theta})\,(\theta-\hat{\theta}) \tag{4.15}$$

with P_{θ_o}-probability at least $1-O\,(1/n)$ where θ^{**} is on the line joining θ_o and $\hat{\theta}$ and the empirical Fisher information is

$$I^*(\theta)=\left(-\frac{1}{n}\sum\frac{\partial}{\partial\theta_i\partial\theta_j}\log p_\theta(x_k)\right)\Bigg|_{i,j=1\cdots d}$$

where the sum is over $k=1,\cdots,n$. Therefore, we may use

$$p(x^n|\theta)=p(x^n|\hat{\theta})e^{-n/2(\theta-\hat{\theta})I^*(\theta^{**})\,(\theta-\hat{\theta})} \tag{4.16}$$

for $\theta\in B(\hat{\theta},\varepsilon)$ by Taylor expanding. Assume that an $\varepsilon''>0$ has been chosen so that with P_{θ_o}-probability at least $1-O\,(1/n)$ we have that $w(\theta)\ge w(\hat{\theta})\,(1-\varepsilon'')$. Now, it is straightforward to show that there is an $\varepsilon'>0$ so that

$$m(x^n)\ge p(x^n|\hat{\theta})w(\hat{\theta})\,(1-\varepsilon'')\left(\frac{2\pi}{n(1+\tau)}\right)^{d/2}$$
$$\cdot\,(\det I^*(\hat{\theta}))^{-1/2}(1-e^{-n\varepsilon'}). \tag{4.17}$$

Using this to bound the posterior density gives

$$w(\theta|x^n)\le\left(\frac{w(\theta)}{w(\hat{\theta})}\,\frac{p(x^n|\theta)}{p(x^n|\hat{\theta})\,(1-\varepsilon'')}\,\left(\frac{n(1+\tau)}{2\pi}\right)^{d/2}\right.$$
$$\left.\cdot\,\frac{(\det I^*(\hat{\theta}))^{1/2}}{(1-\varepsilon^{-n\varepsilon'})}\right)$$
$$\le\left(\frac{w(\theta)}{w(\hat{\theta})}e^{-(n/2)(\hat{\theta}-\theta)I^*(\hat{\theta})\,(\hat{\theta}-\theta)}\right.$$
$$\left.\cdot\,\left(\frac{n(1+\tau)}{2\pi}\right)^{d/2}\frac{\sqrt{\det I^*(\hat{\theta})}}{(1-e^{-n\varepsilon'})}\right). \tag{4.18}$$

Taking logarithms on both sides of (4.18) gives an upper bound on the sum of term one in (4.13) plus term one in (4.14)

$$\int_{B(\hat{\theta},\rho_\varepsilon)}w(\theta|x^n)\,(\log w(\theta|x^n))^+d\theta$$
$$-\left(\frac{d}{2}\log\frac{n}{2\pi}+\frac{1}{2}\log|I(\theta_o)|-\frac{d}{2}+\varepsilon\right)$$
$$\le\int_{B(\hat{\theta},\rho_\varepsilon)}w(\theta|x^n)$$
$$\cdot\left(\log\left(\frac{w(\theta)}{w(\hat{\theta})}\right)-\frac{n}{2}\,(\hat{\theta}-\theta)I^*(\hat{\theta})\,(\hat{\theta}-\theta)\right.$$
$$+\frac{d}{2}\log\left(\frac{n(1+\tau)}{2\pi}\right)+\frac{1}{2}\log|I^*(\hat{\theta})|$$
$$\left.-\log(1-e^{-n\varepsilon'})\right)^+d\theta-\frac{d}{2}\log\frac{n}{2\pi}$$
$$-\frac{1}{2}\log|I(\theta_o)|+\frac{d}{2}. \tag{4.19}$$

The quantity with the superscript plus is seen to be positive because the $\log n$ term dominates.

To see that (4.19) goes to zero in P_θ probability, note that ρ_ε in (4.12) can go to zero slowly. Now, for $\log\,(w(\theta)/w(\hat{\theta}))$ it is enough to note that the posterior concentrates at θ_o and $\hat{\theta}$ goes to θ_o. For the second term in the integrand, note that $I^*(\hat{\theta})$ converges to $I(\theta_o)$ in probability by the expected supremum conditions and the consistency of the MLE. Thus by the reasoning after (4.9) that term converges to $-d/2$ canceling a later term. The two $\log n$ terms cancel because $(\log n)W(B^c(\hat{\theta},\rho_\varepsilon)|X^n)$ is less than or equal to $(\log n)W(B^c(\theta_o,\varepsilon|X^n)$ which goes to zero by the reasoning after (4.8). The remaining terms are routine. □

Appendix

Here we state and prove a result analogous to Theorem 2.1. We give conditions under which the relative entropy between the posterior and a limiting normal converges to zero in expectation. We write $\ell(\theta)$ to mean the log-likelihood.

Theorem A.1: Assume the conclusion of Proposition 2.1 and the regularity conditions at the beginning of Section II. Assume also that $p(x|\theta)$ satisfies the Lipschitz condition

$$|\log p(x|\theta)-\log p(x|\theta')|\le g(x)|\theta-\theta'|^\alpha \tag{A.1a}$$

for some $\alpha>0$, and that for some function $g(x)$ and $\alpha'>0$ we have that

$$n^{\alpha'}\log E_{\theta_o}e^{g(X_1)/n^{\alpha/2}}\longrightarrow 0 \tag{A.1b}$$

as $n\to\infty$, and $E_{\theta_o}g^2(X_1)$ finite, where θ, θ' are arbitrary. We require the "stronger" [29] hypotheses: For each θ, there is a $\delta>0$ and an $\eta>0$ so that

$$E_{\theta_0}\sup_{|\theta-\theta'|<\delta}\left|\log\frac{p(X|\theta')}{p(X|\theta_0)}\right|^{1+\eta}<\infty \tag{A.2a}$$

and there is a ρ large enough that

$$E_{\theta_0}\sup_{|\theta-\theta_0|>\rho}\left|\log\frac{p(X|\theta)}{p(X|\theta_0)}\right|^{1+\eta}<\infty. \tag{A.2b}$$

We also assume that for each $\varepsilon > 0$

$$E_{\theta_o} \int_{B^c(\theta_o,\varepsilon)} n|\theta - \theta^*|^2 w(\theta|X^n)\, d\theta \tag{A.3}$$

goes to zero as n gets large, where $\theta^* = \theta_0 + I(\theta_0)^{-1}\ell'(\theta_0)$ and that the random variables $(\partial/\partial\theta_j)\log p(X_1|\theta_o)$ and

$$Y = \sup_{|\theta' - \theta_o| < \varepsilon} \left| \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(X_1|\theta') - I_{i,j}(\theta_o) \right| \tag{A.4}$$

for $i, j = 1, \cdots, d$ have moment-generating functions under P_{θ_o} that are finite on some neighborhood of zero.

Finally, suppose $|\theta|^2$ has finite prior mean, with $|\theta|^2 w(\theta)$ bounded and let $\tilde{\theta}$ be the posterior mean, assumed to satisfy $E_{\theta_o}(\theta_o - \tilde{\theta})^2$ goes to zero. Now, we have

$$E_{\theta_0} D(w(\cdot|X^n)||N(\theta^*, (nI(\theta_0))^{-1})) \to 0. \tag{A.5}$$

Remarks: Theorem A.1 takes more work than Theorem 2.1. First, we show that the expected posterior and expected squared posterior are subexponential in the sense that any factor of the form $e^{-\gamma n}$ with $\gamma > 0$ makes them go to zero. (For Theorem 2.1, the analogous convergence in probability results were used.) This uses the Lipschitz condition (A.1a,b). Controlling the expectation of (2.8) requires this, Proposition 2.1, and the stronger Wald conditions (A.2a,b). The expectation of (2.9) uses (A.3) which is similar to assuming that the expected posterior variance converges to the Fisher information. Indeed, [6, Theorem 2.2] uses the location $\hat{\theta}$ rather than θ^* and gives convergence in distribution. The content of the assumption is that Bickel and Yahav's result holds in L^1 also. That this might be true is not surprising, but it does not seem easy to give general, readily verifiable conditions. The expectation of the local term uses the moment-generating function hypotheses (A.4) to ensure that the Laplace's method argument from Theorem 2.1 continues to apply. The hypothesis that the expected squared error of the posterior mean goes to zero is not very strong; in many examples it is $O(1/n)$.

Proof: We begin by showing that $E_{\theta_o} w(\theta_o|X^n)$ and $E_{\theta_o} w^2(\theta_o|X^n)$ are subexponential under the Lipschitz-like condition (A.1a,b).

The empirical entropy is $\hat{H}(\theta) = -(1/n)\log p(x^n|\theta)$ and we have that

$$p(x^n|\theta_o) = e^{-\hat{H}(\theta_o)}.$$

This gives a lower bound on the mixture. Write $w_B(\theta)$ to mean $w(\theta)/W(B(\theta_o, \varepsilon))$ We have

$$\begin{aligned} m(x^n) &\ge W(B(\theta_o,\varepsilon)) \int_{B(\theta_o,\varepsilon)} w_B(\theta) e^{-n\hat{H}(\theta)}\, d\theta \\ &\ge \exp\left[-n \int_{B(\theta_o,\varepsilon)} w_B(\theta)\hat{H}(\theta)\, d\theta \right] \end{aligned}$$

which gives a lower bound on the posterior density

$$\begin{aligned} &w(\theta_o|x^n) \\ &\quad\le w_B(\theta_o) \exp\left[-n \int_{B(\theta_o,\varepsilon)} w_B(\theta)\,(\hat{H}(\theta_o) - \hat{H}(\theta))\, d\theta \right]. \end{aligned} \tag{A.6}$$

If we choose $\varepsilon = 1/\sqrt{n}$ then a Taylor expansion argument gives that $W(B(\theta_o, \varepsilon)) \ge K(1/n)^{d/2}$ for some $K > 0$. The difference in entropy in the last exponent gives a term of the same form as in (A.1a). Condition (A.1b) ensures the upper bound goes to zero. Indeed, (A.6) gives

$$E_{\theta_o} w(\theta_o|X^n) \le K n^{d/2} E_{\theta_o} \exp(1/n^{\alpha/2}) \sum_{i=1}^{n} g(X_i)$$

since $|\theta - \theta_o| \le 1/\sqrt{n}$. The n summands in the exponent on the right are independent and identically distributed so the right-hand side is upper-bounded by an expression of the form

$$K n^{d/2} \exp n \log E_{\theta_o} e^{g(X_1)/n^{\alpha/2}}.$$

Since the quantity in (A.1b) is bounded, this in turn is bounded by

$$K n^{d/2} \exp n^{1-\alpha'} \le K \exp n^{1-\alpha'/2}$$

since $n^{d/2} \le e^{n^{\alpha'/2}}$. This shows that $E_{\theta_o} w(\theta_o|X^n)$ is subexponenital in the sense that multiplying it by $e^{-\gamma n}$ gives a limit of zero as n increases, for any $\gamma > 0$. As similar argument holds for the square $E_{\theta_o} w^2(\theta_o|X^n)$, or any power, because powers of the posterior are converted to differences of entropies in the exponent.

We use the same decomposition as for Theorem 2.1. That is, we write the left-hand side of (A.5) as

$$\begin{aligned} &E_{\theta_0} D(w(\cdot|X^n)||N(\theta^*, (nI(\theta_0))^{-1})) \\ &\quad= E_{\theta_0} \int_{B(\theta_0,\varepsilon)} w(\theta|X^n) \log \frac{w(\theta|X^n)}{\varphi(\theta)} d\theta && \text{(A.7)} \\ &\qquad+ E_{\theta_0} \int_{B^c(\theta_0,\varepsilon)} w(\theta|X^n) \log w(\theta|X^n)\, d\theta && \text{(A.8)} \\ &\qquad- E_{\theta_0} \int_{B^c(\theta_0,\varepsilon)} w(\theta|X^n) \log \varphi(\theta)\, d\theta. && \text{(A.9)} \end{aligned}$$

We begin, as before, with the first error term (A.8). Write it as

$$\begin{aligned} &E_{\theta_0} \int_{B^c(\theta_0,\varepsilon)} w(\theta|X^n) \log \frac{w(\theta)}{w(\theta_0)} d\theta && \text{(A.10)} \\ &\quad+ E_{\theta_0} \int_{B^c(\theta_0,\varepsilon)} w(\theta|X^n) \log w(\theta_0|X^n)\, d\theta && \text{(A.11)} \\ &\quad+ E_{\theta_0} \int_{B^c(\theta_0,\varepsilon)} w(\theta|X^n) \log \frac{p(x^n|\theta)}{p(x^n|\theta_0)} d\theta. && \text{(A.12)} \end{aligned}$$

Since $w(\theta)$ is bounded, (A.10) is bounded by

$$K E_{\theta_0} W(B^c(\theta_0, \varepsilon)|X^n)$$

for some $K > 0$, which is $O(e^{-n\alpha})$ for some $\alpha > 0$ by (2.4). Similarly, using $\log x \le x$ gives that (A.11) is bounded from above by

$$\begin{aligned} &E_{\theta_0} W(B^c(\theta_0, \varepsilon)|X^n) w(\theta_0|X^n)) \\ &\qquad\le \sqrt{E_{\theta_0} W(B^c(\theta_0, \varepsilon)|X^n)^2) E_{\theta_0} w(\theta_0|X^n)^2}. \end{aligned}$$

The last bound is a product; one factor is $O(e^{-\alpha n})$ as just noted and the other is subexponential as just shown. Together, they go to zero. The third term (A.12) is

$$\begin{aligned} &nE_{\theta_0}\int_{B^c(\theta_0,\varepsilon)} w(\theta|X^n)\log\frac{p(X_1|\theta)}{p(X_1|\theta_0)}d\theta \\ &\le nE_{\theta_0}\sup_{\theta\in B^c(\theta_0,\varepsilon)}\log\frac{p(X_1|\theta)}{p(X_1|\theta_0)}W(B(\theta_0,\varepsilon)^c|X^n) \\ &\le n(E_{\theta_0}W(B^c(\theta_0,\varepsilon)|X^n)^{((1+\eta)/\eta)})^{\eta/(1+\eta)} \\ &\cdot\left(E_{\theta_0}\sup_{\theta\in B^c(\theta_0,\varepsilon)}\left(\log\frac{p(X_1|\theta)}{p(X_1|\theta_0)}\right)^{1+\eta}\right)^{1/(1+\eta)}. \end{aligned} \tag{A.13}$$

The first factor is $O\left(ne^{-n\alpha}\right)$ by (2.4). To see that the second factor is finite under (A.2a,b), let Δ_j be a finite open cover of the closure of $B(\theta_o,r)-B(\theta_o,\varepsilon)$, where r is chosen so large that (A.2b) is satisfied, and each Δ_j is of the form $B(\theta,\delta)$ so that (A.2a) is satisfied. Now

$$\begin{aligned} &E_{\theta_0}\sup_{\theta\in B^c(\theta_0,\varepsilon)}\left(\log\frac{p(X_1|\theta)}{p(X_1|\theta_0)}\right)^{1+\eta} \\ &\le\sum_{i=1}^{R}E_{\theta_0}\sup_{\theta\in\Delta_j}\left(\log\frac{p(X_1|\theta)}{p(X_1|\theta_0)}\right)^{1+\eta} \\ &+E_{\theta_0}\sup_{|\theta|>r}\left(\log\frac{p(X_1|\theta)}{p(X_1|\theta_0)}\right)^{1+\eta} \end{aligned} \tag{A.14}$$

where

$$\bigcup_{j=1}^{R}\Delta_j\cup\{|\theta|>r\}\supset B(\theta_0,\varepsilon)^c$$

but omits an open set around θ_0. So, the product in (A.13) goes to zero.

Next we deal with the other error term, (A.9). Note that

$$\begin{aligned} \log\varphi(\theta)=&\frac{d}{2}\log\frac{n}{2\pi}+\frac{1}{2}\log|I(\theta_0)| \\ &-\frac{n}{2}\,(\theta-\theta^*)I(\theta_0)\,(\theta-\theta^*). \end{aligned}$$

Using this expression, (A.9) is of the form

$$\begin{aligned} &K(\log n)E_{\theta_0}w(B^c(\theta_0,\varepsilon)|X^n) \\ &+KE_{\theta_0}\int_{B^c(\theta_0,\varepsilon)}nw(\theta|X^n)(\theta-\theta^*)I(\theta_0)(\theta-\theta^*)\,d\theta. \end{aligned} \tag{A.15}$$

The first term is $O\left((\log n)e^{-n\alpha}\right)$, by (2.4). Apart from a positive constant to bound the quadratic form defined by $I(\theta_o)$, the second term is controlled by (A.3).

At last we turn to the local term (A.7). Recall the sets $B_n(\theta_0,\alpha,\varepsilon)$ and $C_n(\delta,\theta_0)$ in the sample space, defined for $\delta,\ \varepsilon>0$ by

$$\begin{aligned} B_n(\theta_0,\alpha,\varepsilon)=\{x^n|&(1-\varepsilon)\,(\theta-\theta_0)I(\theta_0)\,(\theta-\theta_0) \\ &\le(\theta-\theta_0)I^*(\theta')\,(\theta-\theta_0) \\ &\le(1+\varepsilon)\,(\theta-\theta_0)I(\theta_0)\,(\theta-\theta_0) \\ &\text{for }\theta,\theta'\in B(\theta_0,\alpha)\} \end{aligned}$$

and let

$$C_n(\theta_0,\delta)=\{x^n|\ell_n'(\theta_0)I(\theta_0)^{-1}\ell_n'(\theta_0)\le\delta^2\}$$

where $\ell'(\theta)=(1/n)\nabla\log p_\theta(x^n)$. Under the regularity assumptions at the beginning of Section II, the technique of proof of Theorem 2.5 in ([11, p. 465]) gives

$$P_{\theta_0}(B_n^c),P_{\theta_0}(C_n^c)=o(1/n).$$

Now we decompose (A.7) into

$$E_{\theta_0}\int_{B(\theta_0,\varepsilon)}w(\theta|X^n)\log w(\theta|X^n)\,d\theta\,\chi_{B_n^c\cup C_n^c} \tag{A.16}$$

$$+E_{\theta_0}\int_{B(\theta_0,\varepsilon)}w(\theta|X^n)\log w(\theta|X^n)\,d\theta\,\chi_{B_n\cap C_n} \tag{A.17}$$

$$\begin{aligned} &-E_{\theta_0}\int_{B(\theta_0,\varepsilon)}w(\theta|X^n) \\ &\cdot\log\frac{|nI(\theta_0)|^{1/2}e^{-(n/2)\,(\theta-\theta^*)I(\theta_0)\,(\theta-\theta^*)}}{(2\pi)^{d/2}}d\theta. \end{aligned} \tag{A.18}$$

Under our hypotheses expression (A.16) tends to zero. Write (A.16) as the sum

$$E_{\theta_0}\int_{B(\theta_0,\varepsilon)}w(\theta|X^n)\log\frac{w(\theta)}{w(\theta_o)}d\theta\,\chi_{B_n^c\cup C_n^c} \tag{A.19}$$

$$+E_{\theta_0}\int_{B(\theta_0,\varepsilon)}w(\theta|X^n)\log w(\theta_o|X^n)\,d\theta\,\chi_{B_n^c\cup C_n^c} \tag{A.20}$$

$$+E_{\theta_0}\chi_{B_n^c\cup C_n^c}\int_{B(\theta_0,\varepsilon)}w(\theta|X^n)\log\frac{p(x^n|\theta)}{p(x^n|\theta_o)}d\theta. \tag{A.21}$$

Now, (A.19) goes to zero as n increases becasue w is bounded and $P_{\theta_o}(B_n^c\cup C_n^c)$ is $o(1/n)$ by [11, Proposition 6.3].

For (A.21), write the log of the product as the sum of the log's. Now, by symmetry, (A.21) is n times the first of these summands. Taking the absolute value inside the integral gives

$$nE_{\theta_0}\chi_{B_n^c\cup C_n^c}\int_{B(\theta_0,\varepsilon)}w(\theta|X^n)\left|\log\frac{p(X_1|\theta)}{p(X_1|\theta_o)}\right|\,d\theta$$

as an upper bound. The Lipschitz condition (A.1a) now gives

$$\begin{aligned} &nE_{\theta_0}\chi_{B_n^c\cup C_n^c}\int_{B(\theta_0,\varepsilon)}w(\theta|X^n)g(X_1)|\theta-\theta_o|^\alpha\,d\theta \\ &\le KnE_{\theta_o}\chi_{B_n^c\cup C_n^c}g(X_1) \end{aligned} \tag{A.22}$$

for some $K>0$. Using the Cauchy–Schwartz inequality on the right-hand side to separate the indicator function from g and using the fact that $P_{\theta_o}(B_n^c\cup C_n^c)$ is $o(1/n)$ gives that (A.21) goes to zero.

Turning to (A.20), we see that it is bounded from above by

$$E_{\theta_o}\chi_{B_n^c\cup C_n^c}w(\theta_o|X^n)\le\sqrt{E_{\theta_o}\chi_{B_n^c\cup C_n^c}E_{\theta_o}w^2(\theta_o|X^n)} \tag{A.23}$$

since $W(B(\theta_0,\varepsilon)|x^n)\le 1$. Now, by the subexponentiality of $E_{\theta_o}w^2(\theta_o|X^n)$, it is enough to note that $P_{\theta_o}(B_n^c\cup C_n^c)$ is $O\left(e^{-n\gamma}\right)$ for some $\gamma>0$. For this it is enough to recall that by the standard Cramer–Chernoff large deviation theory, the moment-generating function hypotheses on the score and second derivative of the log likelihood ensure that $P_{\theta_o}(B^c)$ and

$P_{\theta_o}(C^c)$ are $O(e^{-n\gamma})$. This gives that (A.23), and therefore, (A.16), are $o(1)$.

To finish the proof of Theorem A.1 we show (A.17) and (A.18) cancel. Taylor-expand $p(x^n|\theta)$ at θ_0 to get

$$p(x^n|\theta) = e^{n\ell'(\theta_0)\,(\theta-\theta_0)-(n/2)\,(\theta-\theta_0)I^*(\theta^{**})\,(\theta-\theta_0)}p(x^n|\theta_0) \tag{A.24}$$

where $\theta \in B(\theta_0, \varepsilon)$ and θ^{**} is on the straight line joining θ_0 and θ. By the restriction to B_n we have

$$p(x^n|\theta) \le p(x^n|\theta_0)e^{n\ell'(\theta_0)\,(\theta-\theta_0)-(n/2)\,(1-\varepsilon)\,(\theta-\theta_0)I(\theta_0)\,(\theta-\theta_0)} \tag{A.25a}$$

and

$$p(x^n|\theta) \ge p(x^n|\theta_0)e^{n\ell'(\theta_0)\,(\theta-\theta_0)-(n/2)\,(1+\varepsilon)\,(\theta-\theta_0)I(\theta_0)\,(\theta-\theta_0)}. \tag{A.25b}$$

Letting

$$\theta_u = \theta_0 + \frac{1}{1-\varepsilon}I(\theta_0)^{-1}\ell'(\theta_0)$$

and

$$\theta_\ell = \theta_0 + \frac{1}{1+\varepsilon}I(\theta_0)^{-1}\ell'(\theta_0)$$

and completing the square in the exponents of (A.25a,b) gives (A.26a) and (A.26b) at the bottom of this page. These provide local bounds at θ_0 for the posterior (A.27), at the bottom of this page. Using (A.27) we get an upper bound on (A.17)

$$\begin{aligned} E_{\theta_0} \int_{B(\theta_0,\varepsilon)} & w(\theta|X^n) \log w(\theta|X^n)\, d\theta\, \chi_{B_n \cap C_n} \\ &\le E_{\theta_0}\chi_{B_n\cap C_n} \int_{B(\theta_0,\varepsilon)} w(\theta|X^n) \log w(\theta)\, d\theta \qquad \text{(A.28a)} \\ &\quad - E_{\theta_0}\chi_{B_n\cap C_n} \int_{B(\theta_0,\varepsilon)} w(\theta|X^n) \left(\frac{n}{2}\right) (\theta - \theta_u) \\ &\quad \cdot (1-\varepsilon) I(\theta_0)\, (\theta-\theta_u)\, d\theta \qquad \text{(A.28b)} \\ &\quad + E_{\theta_0}\chi_{B_n\cap C_n} \int_{B(\theta_0,\varepsilon)} w(\theta|X^n) \left(\frac{n}{2}\right) \ell'(\theta_0) \\ &\quad \cdot \left(\frac{1}{1-\varepsilon} - \frac{1}{1+\varepsilon}\right) I(\theta_0)^{-1}\ell'(\theta_0)\, d\theta \qquad \text{(A.28c)} \\ &\quad - E_{\theta_0}\chi_{B_n\cap C_n} \int_{B(\theta_0,\varepsilon)} w(\theta|X^n) \log \int_{B(\theta_0,\varepsilon)} \\ &\quad \cdot w(\theta) e^{-(n/2)\,(\theta-\theta_\ell)\,(1+\varepsilon) I(\theta_0)\,(\theta-\theta_\ell)}\, d\theta. \qquad \text{(A.28d)} \end{aligned}$$

Also, we express (A.18) as

$$- E_{\theta_0} W(B(\theta_0,\varepsilon)|X^n) \cdot \left(\log\left(\frac{n}{2\pi}\right)^{d/2} + \frac{1}{2}\log|I(\theta_0)|\right) \tag{A.29a}$$

$$+ E_{\theta_0} \int_{B(\theta,\varepsilon)} (\frac{n}{2})\, (\theta-\theta^*) I(\theta_0) (\theta-\theta^*) w(\theta|X^n)\, d\theta. \tag{A.29b}$$

Note that $w(\theta) \ge w(\theta_0) - \varepsilon'$ on $B(\theta_0,\varepsilon)$ for use in (A.28d). Adding the terms of (A.28) and (A.29) and rearranging gives the upper bound on the sum of (A.17) and (A.18)

$$\begin{aligned} & E_{\theta_0}\chi_{B_n\cap C_n} \int_{B(\theta_0,\varepsilon)} w(\theta|X^n) \log\left(\frac{w(\theta)}{w(\theta_0)-\varepsilon'}\right) d\theta \qquad \text{(A.30a)} \\ &\quad + \left[\frac{n}{2} E_{\theta_0} \int_{B(\theta_0,\varepsilon)} (\theta-\theta^*) I(\theta_0)\, (\theta-\theta^*)\, w(\theta|X^n)\, d\theta \right. \qquad \text{(A.30b)} \\ &\quad - \frac{n}{2} E_{\theta_0} \int_{B(\theta_0,\varepsilon)} \chi_{B_n\cap C_n} w(\theta|X^n)\, (\theta-\theta_u) \\ &\quad \left. \cdot (1-\varepsilon) I(\theta_0)\, (\theta-\theta_u)\, d\theta \right] \\ &\quad + E_{\theta_0}\chi_{B_n\cap C_n} \int_{B(\theta_0,\varepsilon)} w(\theta|X^n) \left(\frac{n}{2}\right) \ell'(\theta_0) \\ &\quad \cdot I(\theta_0)^{-1}\ell'(\theta_0)\, d\theta \frac{2\varepsilon}{(1+\varepsilon)\,(1-\varepsilon)} \qquad \text{(A.30c)} \\ &\quad - E_{\theta_0} W(B(\theta_0,\varepsilon)|X^n)\, (\log\left(\frac{n}{2\pi}\right)^{d/2} + \frac{1}{2}\log|I(\theta_0)|) \\ &\quad - E_{\theta_0}\chi_{B_n\cap C_n} \int_{B(\theta_0,\varepsilon)} w(\theta|X^n) \log \\ &\quad \cdot \left(\int_{B(\theta_0,\varepsilon)} e^{-(n/2)\,(\theta'-\theta_\ell)\,(1+\varepsilon) I(\theta_0)\,(\theta'-\theta_\ell)}\, d\theta'\right) d\theta. \qquad \text{(A.30d)} \end{aligned}$$

$$p(x^n|\theta) \le p(x^n|\theta_0)e^{-(n/2)\,(\theta-\theta_u)\,(1-\varepsilon)I(\theta_0)\,(\theta-\theta_u)+(n/2)\ell'(\theta_0)\,(1-\varepsilon)^{-1}I(\theta_0)^{-1}\ell'(\theta_0)} \tag{A.26a}$$

and

$$p(x^n|\theta) \ge p(x^n|\theta_0)e^{-(n/2)\,(\theta-\theta_\ell)\,(1+\varepsilon)I(\theta_0)\,(\theta-\theta_\ell)+(n/2)\ell'(\theta_0)\,(1+\varepsilon)^{-1}I(\theta_0)^{-1}\ell'(\theta_0)}. \tag{A.26b}$$

$$\begin{aligned} w(\theta|x^n) &\le \frac{w(\theta)p(x^n|\theta)}{\displaystyle\int_{B(\theta_0,\varepsilon)} w(\theta)p(x^n|\theta)\, d\theta} \\ &\le \frac{w(\theta)e^{-(n/2)\,(\theta-\theta_u)\,(1-\varepsilon)I(\theta_0)\,(\theta-\theta_u)+(n/2)\ell'(\theta_0)\,(1-\varepsilon)^{-1}I(\theta_0)^{-1}\ell'(\theta_0)}}{\displaystyle\int_{B(\theta_0,\varepsilon)} w(\theta)e^{-(n/2)\,(\theta-\theta_\ell)\,(1+\varepsilon)I(\theta_0)\,(\theta-\theta_\ell)+(n/2)\ell'(\theta_0)\,(1+\varepsilon)^{-1}I(\theta_0)^{-1}\ell'(\theta_0)}\, d\theta}. \end{aligned} \tag{A.27}$$

Continuity of w at θ_0 implies that (A.30a) $\to 0$ as $n \to \infty$, provided ε' decreases as ε decreases. Equally easy is (A.30c) which is bounded from above by

$$\varepsilon n E_{\theta_0} \ell'(\theta_0) I(\theta_0)^{-1} \ell'(\theta_0) = \varepsilon I_{d\times d}.$$

(Apply *trace*, take it inside the expectation, rearrange the factors under trace, and recognize the expectation is the identity $I_{d\times d}$.) Expression (A.30d) is

$$\begin{aligned}&- E_{\theta_0} W(B(\theta_0,\varepsilon)|X^n) \log \left(\frac{n}{2\pi}\right)^{d/2} + \frac{1}{2}\log|I(\theta_0)| \\ &\quad - E_{\theta_0} \chi_{B_n\cap C_n} W(B(\theta_0,\varepsilon)|X^n) \\ &\quad \cdot \log \int_{B(\theta_0,\varepsilon)} e^{-(n/2)\,(1+\varepsilon)\,(\theta-\theta_\ell) I(\theta_0)\,(\theta-\theta_\ell)}\, d\theta.\end{aligned} \tag{A.31}$$

Because $x^n \in B_n \cap C_n$ and $\theta \in B(\theta_0,\varepsilon)$ we have that $|\theta-\theta_\ell|$ is bounded away from zero. Consequently, $\exists \eta > 0$ so that for $\varepsilon > 0$ (in $B(\theta_0,\varepsilon)$ and in $(1+\varepsilon)$) small enough and n large enough when the integration is performed we get an upper bound on (A.31) of the form

$$\begin{aligned}&- E_{\theta_0} W(B(\theta_0,\varepsilon)|X^n)\,(\log \left(\frac{n}{2\pi}\right)^{d/2} + \frac{1}{2}\log|I(\theta_0)|) \\ &\quad + \eta - E_{\theta_0} \chi_{B_n\cap C_n} W(B(\theta_0,\varepsilon)|X^n) \\ &\quad \cdot \log \int_{R^d} e^{-(n/2)\,(\theta-\theta_\ell)\,(1+\varepsilon) I(\theta_0)\,(\theta-\theta_\ell)}\, d\theta\end{aligned} \tag{A.32}$$

where η goes to zero slowly as n increases. After adding and subtracting $E_{\theta_0} W(B(\theta_0,\varepsilon)^c|X^n) = o(e^{-n\alpha})$ in both places in (A.32) where the posterior probability appears, evaluating the integral in the second term of (A.32) gives the upper bound

$$\begin{aligned}&- \log \left(\frac{n}{2\pi}\right)^{d/2} + \frac{1}{2}\log|I(\theta_0)| + O(e^{-n\alpha}) \\ &+\eta + E_{\theta_0}\chi_{B_n\cap C_n} \left(\log\left(\frac{n}{2\pi}\right)^{d/2} - \frac{1}{2}\log|I(\theta_0)|\right).\end{aligned} \tag{A.33}$$

Since $P_{\theta_0}((B_n \cap C_n)^c) = o(1/n)$ (A.33) goes to zero.

Finally, we deal with (A.30b). It is

$$\begin{aligned}&\frac{n}{2} E_{\theta_0} \int [(\theta-\theta^*) I(\theta_0)\,(\theta-\theta^*) \\ &\quad - (\theta-\theta_u) I(\theta_o)\,(\theta-\theta_u)] w(\theta|X^n)\, d\theta && \text{(A.34a)} \\ &\quad - \frac{n}{2} E_{\theta_0} \int_{B^c(\theta_0,\varepsilon)} (\theta-\theta^*) I(\theta_0)(\theta-\theta^*) w(\theta|X^n)\, d\theta \\ & && \text{(A.34b)} \\ &\quad + \frac{n}{2} E_{\theta_0} \chi_{(B_n\cap C_n)^c} \int_{B(\theta_0,\varepsilon)} \\ &\qquad \cdot (\theta-\theta_u) I(\theta_o)\,(\theta-\theta_u) w(\theta|X^n)\, d\theta && \text{(A.34c)} \\ &\quad + \frac{n}{2} E_{\theta_0} \int_{B^c(\theta_0,\varepsilon)} (\theta-\theta_u) I(\theta_o)(\theta-\theta_u) w(\theta|X^n)\, d\theta. \\ & && \text{(A.34d)}\end{aligned}$$

Term (A.34c) is easy. The integrand is bounded in θ. The posterior probability of the domain of integration is also bounded and the probability of $(B_n \cap C_n)^c$ is $o(1/n)$. Terms (A.34b,d) go to zero by the same argument as was used for (A.15).

Term (A.34a) requires a bit more. Neglecting the factor of $n/2$, the difference of quadratic forms in the integrand can be rewritten as $-(tI(\theta_o)t + 2tI(\theta_o)\,(\theta-\theta^*)$, where $t = (\varepsilon/(1-\varepsilon))I^{-1}(\theta_o)\ell'(\theta_o)$. Using this in (A.34a), integrating over θ, and adding and subtracting $(\theta^*-\tilde{\theta})I(\theta_o)\,(\theta^*-\tilde{\theta})$, where $\tilde{\theta}$ is the posterior mean, gives that (A.34a) is

$$\begin{aligned}&nE_{\theta_o}(t-(\theta^*-\tilde{\theta}))I(\theta_o)\,(t-(\theta^*-\tilde{\theta})) \\ &\qquad\qquad -nE_{\theta_o}(\theta^*-\tilde{\theta})I(\theta_o)\,(\theta^*-\tilde{\theta}).\end{aligned} \tag{A.35}$$

Now, using the definitions

$$\theta^* = \theta_o + I^{-1}(\theta_o)\ell'(\theta_o) \quad \text{and} \quad t = (\varepsilon/(1-\varepsilon))I^{-1}(\theta_o)\ell'(\theta_o)$$

(A.35) reduces to

$$\begin{aligned}&\frac{2-\varepsilon}{1-\varepsilon} nE_{\theta_o}(\theta_o-\tilde{\theta})I^{-1}(\theta_o)\ell'(\theta_o) \\ &\qquad + \frac{1-(1-\varepsilon)^2}{(1-\varepsilon)^2} E_{\theta_o}[n\ell'(\theta_o)I^{-1}(\theta_o)\ell'(\theta_o)]\end{aligned} \tag{A.36}$$

after canceling some terms and collecting others. The expectation in the second term of (A.36) is finite, so the whole term goes to zero as ε goes to zero. It is seen that the first term in (A.36) goes to zero by taking absolute value in the expectation and applying the Cauchy–Schwartz inequality to get, apart from bounded factors, $E_{\theta_o}(\theta_o-\tilde{\theta})^2$ times $nE_{\theta_o}\ell'(\theta_o)\ell'$. The second factor is bounded and the first goes to zero.

Acknowledgment

The author would like to express his gratitude to A. R. Barron and two anonymous referees whose suggestions greatly improved this paper.

References

[1] A. R. Barron, "Logically smooth density estimation," Ph.D. dissertation, Dept. Elec. Eng., Stanford Univ., Stanford, CA, Aug. 1985.
[2] ——, "Entropy and the central limit theorem," *Ann. Probab.*, vol. 14, bo. 1, pp. 336–342, 1986.
[3] A. R. Barron and T. Cover, "Minimum complexity density estimation," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1034–1054, July 1991.
[4] A. R. Barron, L. Gyorfi, and E. C. van der Meulen, "Distribution estimates consistent in total variation and in two types of information divergence," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1437–1454, 1992.
[5] ——, "The asymptotic normality of information divergence error for a histogram based density estimate," in *Proc. 1994 IEEE–IMS Workshop on Information Theory and Statistics*, Oct. 1994, p. 39.
[6] P. Bickel and J. Yahav, "Some contributions to the asymptotic theory of Bayes' solutions," *Z. Wahrsch. verw. Gebiete*, vol. 11, pp. 257–276, 1969.
[7] F. Brouaye, "Asymptotic normality of some Hermitian forms with complex noisy data," *IEEE Trans. Inform. Theory*, vol. 40, pp. 236–239, Jan. 1994.
[8] N. Cesa-Binachi, A. Krogh, and M. K. Warmuth, "Bounds on the approximate steepest descent for likelihood maximization in exponential families," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1215–1218, July 1994.
[9] B. Clarke, "Asymptotic cumulative risk and Bayes risk under entropy loss, with applications," Ph.D. dissertation, Dept. Statistics, Univ. Illinois, Urbaba, IL, July 1989.
[10] ——, "Implications of reference priors for prior information and sample size," *J. Amer. Statist. Assoc.*, vol. 91, pp. 173–184, 1996. 1995.
[11] B. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory*, vol. 36, no. 3, May 1990.
[12] ——, "Jeffreys' prior is asymptotically least favorable under entropy risk," *J. Statist. Planning Infer.*, vol. 41, pp. 37–60, 1994.
[13] I. Csiszár, "I-Divergence geometry of probability distributions and minimization problems," *Ann. Probab.*, vol. 3, no. 1, pp. 146–158, 1975.

[14] ——, “Sanov property, generalized I-projection and a conditional limit theorem,” *Ann. Probab.*, vol. 12, pp. 768–793, 1984.

[15] ——, “Generalized cutoff rates and Renyi’s information measures,” *IEEE Trans. Inform. Theory*, vol. 41, pp. 26–34, Jan. 1995.

[16] Efroimovich, “Information contained in a sequence of observations,” *Probl. Inform. Transm.*, vol. 15, pp. 24–39, 1980.

[17] D. A. S. Fraser and P. McDunnough, “Further remarks on asymptotic normality of likelihood and conditional analyses,” *Canad. J. Statist.*, vol. 12, no. 3, pp. 183–190, 1984.

[18] J. Hartigan, *Bayes’ Theory*. New York: Springer-Verlag, 1983.

[19] I. Ibragimov and R. Has’minskii, “On the information in a sample about a parameter,” in *Proc. Int. Symp. Information Theory*. Budapest, Hungary: Akademiai Kiado, 1973, pp. 295–309.

[20] ——, *Statistical Estimation: Asymptotic Theory*. New York: Springer-Verlag, 1980.

[21] L. Le Cam, “Les proprietes asymptotiques des solutions de Bayes,” *Publ. Inst. Statist. Univ. Paris*, vol. 7, pp. 18–35, 1958.

[22] E. L. Lehmann, *Theory of Point Estimation*. New York: Wiley, 1983.

[23] M. E. Meyer and D. V. Gokhale, “Kullback–Leibler information measure for studying convergence rates of densities and distributions,” *IEEE Trans. Inform. Theory*, vol. 39, pp. 1401–1404, July 1993.

[24] J. Rissanen, “Universal coding, information, prediction, and estimation,” *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 629–636, July 1984.

[25] ——, “Stochastic complexity,” *J. Roy. Statist. Soc. Ser. B*, vol. 49, no. 3, pp. 223–239, 1987.

[26] ——, “Fisher information and stochastic complexity,” in *Proc. 1994 IEEE–IMS Workshop on Information Theory and Statistics*, Oct. 1994, p. 5.

[27] J. Rissanen, T. Speed, and B. Yu, “Density estimation by stochastic complexity,” *IEEE Trans. Inform. Theory*, vol. 38, pp. 315–323, 1992.

[28] A. M. Walker, “On the asymptotic behaviour of posterior distributions,” *J. Roy. Statist. Soc.*, vol. 31, pp. 80–88, 1967.

[29] A. Wald, “Note on the consistency of the maximum likelihood estimate,” *Ann. Math. Statist.*, vol. 20, no. 4, pp. 595–601, Dec. 1949.