Ann. Inst. Statist. Math.
Vol. 40, No. 4, 683–691 (1988)

THE EXPONENTIAL RATES OF CONVERGENCE OF POSTERIOR DISTRIBUTIONS*

JAMES C. FU[1] AND ROBERT E. KASS[2]

[1]*Department of Statistics, University of Manitoba, Winnipeg, Manitoba, Canada*
[2]*Department of Statistics, Carnegie-Mellon University, Pittsburgh, PA 15213, U.S.A.*

(Received September 10, 1986; revised February 3, 1987)

Abstract. After the observations were observed, the posterior distribution under mild conditions becomes more concentrated in the neighbourhood of the mode of the posterior distribution as sample size n increase. In this paper, the exponential rate of convergence of posterior distribution around the mode is established by using the generalized Laplace method. An example is also given.

Key words and phrases: Bayesian, posterior probability, exponential rate, f-uniform convergence, Q-uniform convergence.

1. Introduction

After the observations were observed, from the Bayesian point of view, the statistical inference about the unknown parameter θ is expressed by the posterior distribution only. The posterior distribution, under mild conditions becomes more concentrate as sample size increase. There is considerable literature on the asymptotic behaviour of posterior distribution in a neighbourhood of the mode (or maximum likelihood estimator), for example, LeCam (1953), Freedman (1963), Lindley (1965), Johnson (1967), Walker (1969), Brenner *et al.* (1983) and Chen (1983). Mathematically, if a and b are constant, the posterior probability

$$(1.1)\qquad P_{\theta|s}(\hat{\theta}_n + a\sigma_n < \theta < \hat{\theta}_n + b\sigma_n) = \int_{\hat{\theta}_n + a\sigma_n}^{\hat{\theta}_n + b\sigma_n} f_n(\theta|s)d\theta\,,$$

converges to $\Phi(b) - \Phi(a)$ when $n \to \infty$, where $f_n(\theta|s)$ is posterior density function, $\hat{\theta}_n$ is the mode of posterior density (or m.l.e.), $s = (x_1, x_2, \dots)$ stands for observations, $\sigma_n^2 = (-l_n^{(2)}(s|\hat{\theta}_n))^{-1} = O(n^{-1})$, $l_n^{(2)}$ is the second

*This work was supported in part by the Natural Sciences and Engineering Research Council of Canada under Grant NSERC A-9216.

derivative of logarithm of likelihood function, and Φ stands for cumulative distribution function of standard normal distribution.

Schwartz (1965) and Johnson (1967, 1970) study, under certain regularity conditions, asymptotic expansions in powers of $n^{-1/2}$ associated with posterior distributions having the standard normal as the leading term.

Fu and Kass (1983) obtain the exponential rate for the ε-tail posterior probability around the mode of posterior distribution tending to zero as $n \to \infty$; i.e.,

$$\lim_{n\to\infty} \frac{1}{n} \log P_{\theta|s}(|\theta - \hat{\theta}| > \varepsilon) = -\beta(\hat{\theta}, \varepsilon) , \tag{1.2}$$

almost surely $[P_{\theta_0}]$, where $\hat{\theta}$ is mode, $\beta(\hat{\theta}, \varepsilon) = \min\{K(\theta_0, \hat{\theta} + \varepsilon), K(\theta_0, \hat{\theta} - \varepsilon)\}$, and $K(\theta', \theta)$ is the Kullback-Leibler information

$$K(\theta', \theta) = \int_{-\infty}^{\infty} \left(\log \frac{f(x|\theta')}{f(x|\theta)} \right) f(x|\theta') dx . \tag{1.3}$$

They proved this result under a very strong condition that

$$\lambda_n(s, \theta_0, \theta) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(x_i|\theta_0)}{f(x_i|\theta)} \to K(\theta_0, \theta) , \tag{1.4}$$

as $n \to \infty$ a.s. $[P_{\theta_0}]$ f-uniformly with respect to prior distribution $\mu(\theta)$ (see Parzen (1953)) i.e.,

$$\operatorname*{ess.sup}_{\theta} |\lambda_n(s, \theta_0, \theta) - K(\theta_0, \theta)| \to 0 , \tag{1.5}$$

as $n \to \infty$ a.s. $[P_{\theta_0}]$, where the essential supremum is taken with respect to the prior measure $\mu(\theta)$. This condition is usually satisfied only if the proper prior measure $\mu(\theta)$ has a compact support (or the parameter space Θ is a compact subset of R). Hence it excluded many important cases, for example when $x_1, \ldots, x_n$ are i.i.d. observations from a normal distribution $N(\theta, 1)$ and the prior distribution $\mu(\theta)$ is also a normal distribution on the real line. The main purpose of this article is to extend the result with broader scope, in particular, when the condition convergence f-uniformly doesn't meet.

2. Main results

To extend the result of Fu and Kass (1983) we need to extend the concept of f-uniform convergence for broad scopes.

DEFINITION. The sequence $\lambda_n(s,\theta_0,\theta)$ converges to $K(\theta_0,\theta)$ as $n\to\infty$ a.s. $[P_{\theta_0}]$ and $Q(\theta)$-uniformly with respect to prior measure $\mu(\theta)$ (a proper prior distribution) if there exists a function $Q(\theta)$ ($Q(\theta)$ may depend on θ_0) such that

$$\text{ess.sup}_{\theta} \,|Q^{-1}(\theta)[\lambda_n(s,\theta_0,\theta)-K(\theta_0,\theta)]| \to 0\,, \tag{2.1}$$

as $n\to\infty$ a.s. $[P_{\theta_0}]$, where the essential supremum is taken with respect to the prior measure $\mu(\theta)$.

For given positive constant δ, we write

$$H(\theta_0,\theta,\pm\delta)=K(\theta_0,\theta)\pm\delta|Q(\theta)|\,. \tag{2.2}$$

To prove our results, we assume the following conditions hold:

CONDITION 1. For every θ and θ' ($\theta\neq\theta'$), the probability measure given by the density $f(x|\theta)$ is absolutely continuous with respect to the probability measure given by $f(x|\theta')$.

CONDITION 2. For θ_0 and $\theta\in\Theta$, the Kullback-Leibler information $K(\theta_0,\theta)$ is finite and m times differentiable

$$K^{(i)}(\theta_0,\theta)=\frac{d^i}{d\theta^i}K(\theta_0,\theta)<\infty,\qquad i=1,\dots,m\ (m\geq 2)\,.$$

CONDITION 3. The prior distribution has density function $g(\theta)$. The $g(\theta)$ is m times differentiable ($g^{(i)}(\theta)<\infty$, $i=1,2,\dots,m$) and bounded.

CONDITION 4. The sequence $\lambda_n(s,\theta_0,\theta)$ converges to $K(\theta_0,\theta)$ as $n\to\infty$ a.s. $[P_{\theta_0}]$ $Q(\theta)$-uniformly with respect to prior measure $\mu(\theta)$ and $Q(\theta)$ is m times differentiable ($Q^{(m)}(\theta)<\infty$), $Q(\theta_0)\equiv 0$, and $|Q(\theta)|\uparrow$ as $|\theta-\theta_0|\uparrow$.

CONDITION 5. For every $\delta>0$, $H(\theta_0,\theta,\pm\delta)$ are convex and m ($m\geq 2$) times differentiable functions in the intervals $(-\infty,\theta_0)$ and (θ_0,∞).

CONDITION 6. $\hat\theta_n(s)$ converges to θ_0 a.s. $[P_{\theta_0}]$ as $n\to\infty$.

THEOREM 2.1. *Let* $s=(x_1,\dots,x_n)$ *be* n *i.i.d. observations from the population* $f(x|\theta_0)$ *and* $\mu(\theta)$ *be prior distribution of* θ. *If the Conditions* 1–6 *are satisfied and* $\hat\theta$ *is the mode of the posterior density function, then*

$$\alpha_n^+=P_{\theta|s}(\theta>\hat\theta+\varepsilon)=e^{-n[K(\theta_0,\hat\theta+\varepsilon)+o(1)]}\,, \tag{2.3}$$

and

$$\alpha_n^- = P_{\theta|s}(\theta < \hat{\theta} - \varepsilon) = e^{-n[K(\theta_0, \hat{\theta} - \varepsilon) + o(1)]}, \tag{2.4}$$

as $n \to \infty$ *almost surely* $[P_{\theta_0}]$.

PROOF. For $\varepsilon > 0$, the ε-tail probability of posterior

$$\alpha_n^+ = P_{\theta|s}(\theta > \hat{\theta} + \varepsilon), \tag{2.5}$$

can be written as

$$\alpha_n^+ = C_n(s, \theta_0) \int_{\hat{\theta}+\varepsilon}^{\infty} g(\theta) e^{-n\lambda_n(s, \theta_0, \theta)} d\theta, \tag{2.6}$$

where

$$C_n(s, \theta_0) = \left(\int_{-\infty}^{\infty} g(\theta) e^{-n\lambda_n(s, \theta_0, \theta)} d\theta \right)^{-1},$$

and $\lambda_n(s, \theta_0, \theta)$ is defined by (1.4). We write

$$\begin{aligned} C_n^{-1}(s, \theta_0) &= \int_{-\infty}^{\infty} g(\theta) e^{-n\lambda_n(s, \theta_0, \theta)} d\theta \\ &= \int_{-\infty}^{\infty} g(\theta) e^{-n[K(\theta_0, \theta) - Q(\theta) R_n(s, \theta_0, \theta)]} d\theta, \end{aligned} \tag{2.7}$$

where

$$R_n(s, \theta_0, \theta) = Q^{-1}(\theta)[\lambda_n(s, \theta_0, \theta) - K(\theta_0, \theta)]. \tag{2.8}$$

Since, by Condition 4, $R_n(s, \theta_0, \theta)$ converges to zero as $n \to \infty$ a.s. $[P_{\theta_0}]$ uniformly in θ hence for any *small* $\delta > 0$ we have, for n sufficiently large,

$$\int_{-\infty}^{\infty} g(\theta) e^{-nH(\theta_0, \theta, \delta)} d\theta \le C_n^{-1}(s, \theta) \le \int_{-\infty}^{\infty} g(\theta) e^{-nH(\theta_0, \theta, -\delta)} d\theta, \tag{2.9}$$

(a.s. $[P_{\theta_0}]$). Note that $\lim_{\delta \to 0} H(\theta_0, \theta, \pm\delta) = K(\theta_0, \theta)$. For *arbitrary* $r > 0$, it follows from Conditions 4 and 5 that there exists δ such that

$$\int_{-\infty}^{\infty} g(\theta) e^{-nH(\theta_0, \theta, \delta)} d\theta = O(e^{nr}), \tag{2.10}$$

and

$$\int_{-\infty}^{\infty} g(\theta)e^{-nH(\theta_0,\theta,-\delta)}d\theta = O(e^{nr}) . \tag{2.11}$$

Since, by Condition 6, $\hat{\theta}_n(s)$ converges to θ_0 as $n \to \infty$ $[P_{\theta_0}]$ hence without loss of generality we assume $\theta_0 < \hat{\theta}_n(s) + \varepsilon < b$ a.s. $[P_{\theta_0}]$ for all $b > \theta_0 + \varepsilon$. Denote

$$\begin{aligned} I_n &= \int_{\hat{\theta}_n(s)+\varepsilon}^{\infty} g(\theta)e^{-n\lambda_n(s,\theta_0,\theta)}d\theta \\ &= \left[\int_{\hat{\theta}_n(s)+\varepsilon}^{b} + \int_{b}^{\infty}\right] g(\theta)e^{-n\lambda_n(s,\theta_0,\theta)}d\theta \\ &= I_{n,1} + I_{n,2} . \end{aligned} \tag{2.12}$$

By Conditions 3 to 5, there exist δ and $\zeta > 0$ such that

$$I_{n,2} = O(e^{-(n-\zeta)H(\theta_0,b,\delta)}) , \tag{2.13}$$

a.s. $[P_{\theta_0}]$. Now we need only to consider the integral

$$I_{n,1} = \int_{\hat{\theta}+\varepsilon}^{b} g(\theta)e^{-\lambda_n(s,\theta_0,\theta)}d\theta . \tag{2.14}$$

Again, for any $\delta > 0$, by Condition 4, we have the following inequality

$$\begin{aligned} \underline{I}_{n,1} &= \int_{\hat{\theta}+\varepsilon}^{b} g(\theta)e^{-nH(\theta_0,\theta,\delta)}d\theta \le I_{n,1} \\ &\le \int_{\hat{\theta}+\varepsilon}^{b} g(\theta)e^{-nH(\theta_0,\theta,-\delta)}d\theta = \bar{I}_{n,1} . \end{aligned} \tag{2.15}$$

Now we need to evaluate the rates of $\bar{I}_{n,1}$ and $\underline{I}_{n,1}$ tending to zero. Let $a = \hat{\theta} + \varepsilon$ $(0 < a < b)$. Note that, by Condition 4,

$$H^{(1)}(\theta_0,\theta,\delta) > 0, \qquad \text{for all} \quad a \le \theta \le b . \tag{2.16}$$

Therefore the minimum of $H(\theta_0,\theta,\delta)$ for $a \le \theta \le b$ occurs at $\theta = a$. It follows, for arbitrary $\delta > 0$,

$$\underline{I}_{n,1} = e^{-nH(\theta_0,a,\delta)} \int_a^b G(\theta;n)e^{-nH^{(1)}(\theta_0,a,\delta)(\theta-a)}d\theta , \tag{2.17}$$

where

$$G(\theta;n) = g(\theta)\exp\left\{-n\left[\frac{1}{2}H^{(2)}(\theta_0,a,\delta)(\theta-a)^2 + \cdots\right]\right\} . \tag{2.18}$$

By Conditions 2 to 4, the Taylor expansion of $G(\theta; n)$ at $\theta = a$ yields

$$G(\theta; n) = \sum_{i=0}^{m-1} \frac{1}{i!} G^{(i)}(a; n)(\theta - a)^i + \frac{1}{m!} G^{(m)}(\xi; n)(\theta - a)^m , \tag{2.19}$$

where $a \leq \xi \leq b$. Note that, by Condition 3, the numbers

$$M_i = \sup_{a<\theta<b} |g^{(i)}(\theta)|, \quad i = 0, 1, \ldots, m , \tag{2.20}$$

are bounded. Furthermore, by using the inequality $x^l \exp(-x^2) < l!$ for $x > 0$ and $l = 0, 1, 2, \ldots$ it is easy to see that

$$G^{(i)}(\theta; n) = O(n^{i/2}) , \tag{2.21}$$

for all $i = 0, 1, \ldots, m - 1$ and $a \leq \theta \leq b$. Inserting the equation (2.19) into equation (2.17) and integrating term by term, we have

$$\begin{aligned} \underline{I}_{n,1} = e^{-nH(\theta_0, a, \delta)} \Bigg\{ & \sum_{i=0}^{m-1} \frac{1}{i!} G^{(i)}(a; n) \\ & \cdot \int_a^b (\theta - a)^i e^{-nH^{(1)}(\theta_0, a, \delta)(\theta - a)} d\theta + \delta_{n,m} \Bigg\} , \end{aligned} \tag{2.22}$$

where the remainder term $\delta_{n,m}$ satisfies

$$\begin{aligned} |\delta_{n,m}| &= O\left(n^{m/2} \int_a^b (\theta - a)^m e^{-nH^{(1)}(\theta_0, a, \delta)(\theta - a)} d\theta \right) \\ &= O(n^{-(m+2)/2}), \quad \text{as} \quad n \to \infty . \end{aligned} \tag{2.23}$$

Note that each term in equation (2.22) is a gamma-type integration, hence for each $i = 0, 1, \ldots$ there exists a positive constant ε_i such that

$$\int_a^b (\theta - a)^i e^{-nH^{(1)}(\theta_0, a, \delta)(\theta - a)} d\theta = \frac{i!}{[nH^{(1)}(\theta_0, a, \delta)]^{i+1}} + O(e^{-n\varepsilon_i}) . \tag{2.24}$$

Equations (2.15), (2.22), (2.23), (2.24) and $O(n^{-(m+2)/2}) = O(e^{-[(m+1)\log n]/2})$ yield

$$\begin{aligned} \underline{I}_{n,1} &= e^{-nH(\theta_0, a, \delta)} \left\{ \sum_{i=0}^{m-1} G^{(i)}(a; n)[nH^{(1)}(\theta_0, a, \delta)]^{-i-1} + O(n^{-(m+2)/2}) \right\} \\ &= e^{-n[H(\theta_0, a, -\delta) + o(1)]} . \end{aligned} \tag{2.25}$$

Similarly, for given $\varepsilon > 0$ and arbitrary small $\delta > 0$, we have

$$\bar{I}_{n,1} = e^{-n[H(\theta_0, a, -\delta) + o(1)]} . \tag{2.26}$$

Results (2.25) and (2.26) and inequality (2.15) yield

$$e^{-n[K(\theta_0, \hat{\theta}+\varepsilon) + \delta|Q(\hat{\theta}+\varepsilon)| + o(1)]} \leq I_{n,1} \leq e^{-n[K(\theta_0, \hat{\theta}+\varepsilon) - \delta|Q(\hat{\theta}+\varepsilon)| + o(1)]} , \tag{2.27}$$

for any small δ. Since $Q(\theta)$ is bounded in the interval $a \leq \theta \leq b$ and δ is arbitrary it follows from (2.27)

$$I_{n,1} = e^{-n[K(\theta_0, \hat{\theta}+\varepsilon) + o(1)]} . \tag{2.28}$$

Again since r is arbitrary it follows from (2.6), (2.10), (2.11), (2.12), (2.13) and (2.28) that

$$\begin{aligned} \alpha_n^+ &= P_{\theta|s}(\theta > \hat{\theta}_n + \varepsilon) \\ &= C_n^{-1}(s, \theta_0)[I_{n,1} + I_{n,2}] \\ &= e^{-n[K(\theta_0, \hat{\theta}_n + \varepsilon) + o(1)]} , \end{aligned}$$

a.s. $[P_{\theta_0}]$ as $n \to \infty$. Similarly, we have

$$\alpha_n^- = e^{-n[K(\theta_0, \hat{\theta}_n - \varepsilon) + o(1)]} ,$$

a.s. $[P_{\theta_0}]$ as $n \to \infty$. This completes our proof.

3. Example

Suppose $x_1, \ldots, x_n$ are i.i.d. observations from a normal population having mean θ, variance one and density function

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}(x-\theta)^2 \right\} . \tag{3.1}$$

Assume the prior distribution $\mu(\theta)$ for the location parameter θ has a density function

$$g(\theta) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\theta^2 \right\} . \tag{3.2}$$

For θ_0, $\theta \in \Theta$, the Kullback-Leibler information is

$$K(\theta_0, \theta) = (\theta - \theta_0)^2/2 . \tag{3.3}$$

For any θ_0 and arbitrarily small $r > 0$ the posterior distribution of parameter θ given data s can be written as

$$f_n(\theta|s) = C_n e^{-n(\theta-\theta_0)(\theta+\theta_0-2\bar{x})/2} g(\theta) , \tag{3.4}$$

where

$$C_n^{-1} = \int_{-\infty}^{\infty} e^{-n(\theta-\theta_0)(\theta+\theta_0-2\bar{x})/2} g(\theta) d\theta$$
$$= O(e^{nr}) ,$$

almost surely $[P_{\theta_0}]$ as $n \to \infty$. Let $\theta_0 = 0$. The log-likelihood ratio statistic $\lambda_n(s, 0, \theta)$ is

$$\lambda_n(s, 0, \theta) = \frac{\theta^2}{2} - \bar{x}\theta . \tag{3.5}$$

For every θ, $\lambda_n(s, 0, \theta)$ converges to $K(0, \theta)$ a.s. $[P_0]$. Note that

$$\operatorname*{ess.sup}_{\theta} |\lambda_n(s, 0, \theta) - K(0, \theta)| \to \infty , \tag{3.6}$$

a.s. $[P_0]$ as $n \to \infty$, where the essential supremum is taken with respect to the prior distribution $\mu(\theta)$. This means that $\lambda_n(s, 0, \theta)$ does not converge to $K(0, \theta)$ a.s. $[P_0]$ f-uniformly. Let $Q(\theta) = \theta$. Note that

$$|Q^{-1}(\theta)\lambda_n(s, 0, \theta) - Q^{-1}(\theta)K(0, \theta)| = |\bar{X}_n| , \tag{3.7}$$

which is *independent* of θ. Furthermore, $|\bar{X}_n| \to 0$ a.s. $[P_0]$ and the convergence is independent of θ. Hence the sequence $\lambda_n(s, 0, \theta)$ converges to $K(0, \theta)$ a.s. $[P_0]$ and *$Q(\theta)$-uniformly* with respect to prior measure $\mu(\theta)$ as $n \to \infty$. For any $\delta > 0$, it follows from (3.7) for sufficiently large n

$$|Q^{-1}(\theta)\lambda_n(s, 0, \theta) - Q^{-1}(\theta)K(0, \theta)| < \delta . \tag{3.8}$$

For every δ, the $H(0, \theta, \pm\delta) = K(0, \theta) \pm \delta|Q(\theta)|$ are convex functions and $\lim_{\delta \to 0} H(0, \theta, \pm\delta) = K(0, \theta)$. This can be seen from the following Fig. 1. All the conditions stated in Section 2 are satisfied. It follows from Theorem 2.1 we have

$$\alpha_n^+ = e^{-n[(\hat{\theta}_n(s)+\varepsilon)^2/2+o(1)]} , \tag{3.9}$$

and

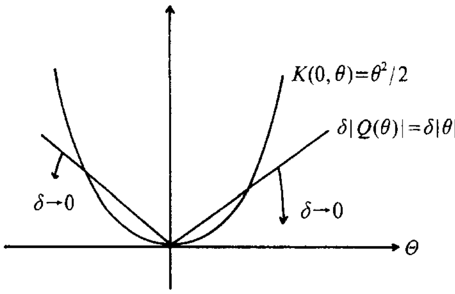

Fig. 1.

$$\alpha_n^- = e^{-n[(\hat{\theta}_n(s)-\varepsilon)^2/2+o(1)]} . \tag{3.10}$$

The results (3.9) and (3.10) can also be obtained by integrating directly. One could see that this example does not fit the framework of Fu and Kass (1983). This result provides more broad expression for a large deviation probability that is relevant to a Bayesian, more importantly, a characteristic of parametric families that affects the quality of the inferences a Bayesian can make.

REFERENCES

Brenner, D., Fraser, D. A. S. and McDunnough, P. (1983). On asymptotic normality of likelihood and conditional analysis, *Canad. J. Statist.*, **10**, 163–172.

Chen, C. F. (1983). On asymptotic normality of limiting density functions with Bayesian implications, Technical Report B-83-18, Institute of Statistics, Academia Sinica, China.

Freedman, D. A. (1963). On the asymptotic behaviour of Bayes estimates in discrete case, *Ann. Math. Statist.*, **34**, 1386–1403.

Fu, J. C. and Kass, R. E. (1983). The rate at which posterior distributions become concentrated, Technical Report, Dept. of Statist., Univ. of Manitoba, Canada.

Johnson, R. A. (1967). An asymptotic expansion for posterior distributions, *Ann. Math. Statist.*, **38**, 1899–1907.

Johnson, R. A. (1970). Asymptotic expansions associated with posterior distributions, *Ann. Math. Statist.*, **41**, 851–864.

LeCam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates, *Univ. California Publ. Statist.*, **1**, 227–330.

Lindley, D. V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint*, University Press, Cambridge.

Parzen, E. (1953). On uniform convergence of families or sequences of random variables, *Univ. California Publ. Statist.*, **1**, 23–53.

Schwartz, L. (1965). On Bayes procedures, *Z. Wahrsch. Verw. Gebiete*, **4**, 10–26.

Walker, A. M. (1969). On the asymptotic behaviour of posterior distributions, *J. Roy. Statist. Soc. Ser. B*, **31**, 80–88.