Large Deviation Principles for Posterior Distributions of the Normal Parameters

Takuhisa Shikimi

Abstract

Suppose that $X_1, X_2, \ldots$ are conditionally i.i.d. random variables with distribution P_θ given $\vartheta = \theta$, where ϑ is an unknown parameter. If P_θ is a normal distribution with mean θ and known variance σ^2, and if the prior of ϑ is chosen from the conjugate family $N(\mu, v^2)$ or proportional to the Lebesgue measure, then it follows that the posterior distributions given $X_1, \ldots, X_n$ obey a large deviation principle with a rate function. If P_θ is a normal distribution with known mean and unknown precision θ, and if as a prior we choose the gamma distribution with parameters α and β or the improper distribution $(1/\theta)d\theta$, the Jeffereys' prior, then the posterior distributions of ϑ given $X_1, \ldots, X_n$ are shown to satisfy a large deviation principle. The Gärtner-Ellis theorem plays the key role to prove these large deviation principles for the posterior distributions.

Keywords: large deviations; posterior distributions; the Gärtner-Ellis theorem.

1 Introduction

Let ϑ be an unknown parameter with prior π and $X_1, X_2, \ldots$ are conditionally i.i.d. random variables with conditional distribution P_θ given $\vartheta = \theta$. In this paper, we will show the large deviation principles for the posterior dis-

tribution of ϑ given $X_1, \ldots, X_n$ when P_θ is either $N(\theta, \sigma^2)$ with known σ^2 or $N(\mu, 1/\theta)$ with known μ. There is comparatively little literature on the exponential rate of convergence of posterior distributions. Fu and Kass (1988) studies the rate of convergence of posterior distributions in the neighborhood of the mode. In the nonparametric Bayesian framework, Shen and Wasserman (2001) studies the rate at which the posterior distribution concentrates around the true parameter, and Ganesh and O'Connell (1999) proves the large deviation principle for posterior distributions given i.i.d. random variables taking values in a finite set. Moderate deviation asymptotic results in a Bayesian setting are studied by Eichelsbacher and Ganesh (2002).

We begin with constructing the Bayesian framework appropriate for formulating and proving our results. Let $(\Theta, \mathscr{U})$ and $(\mathscr{X}, \mathscr{A})$ be Polish spaces (complete separable metric spaces) endowed with the Borel σ-algebras $\mathscr{U}$ and $\mathscr{A}$. A stochastic kernel from $(\Theta, \mathscr{U})$ to $(\mathscr{X}, \mathscr{A})$ is a family $(P_\theta : \theta \in \Theta)$ of probability measures on $(\mathscr{X}, \mathscr{A})$ indexed by $\theta \in \Theta$, namely a mapping $\theta \in \Theta \mapsto P_\theta \in \mathscr{P}_\mathscr{X}$, where $\mathscr{P}_\mathscr{X}$ is the set of probability measures on $(\mathscr{X}, \mathscr{A})$, such that for each $A \in \mathscr{A}$, $\theta \in \Theta \mapsto P_\theta(A) \in [0,1]$ is measurable. The space $\mathscr{P}_\mathscr{X}$ equipped with the weak topology is metrized by the Lévy-Prohorov metric, and with respect to the metric $\mathscr{P}_\mathscr{X}$ is Polish. Further, a family $(P_\theta : \theta \in \Theta)$ of probability measures is a stochastic kernel from $(\Theta, \mathscr{U})$ to $(\mathscr{X}, \mathscr{A})$ if and only if $\theta \in \Theta \mapsto P_\theta \in \mathscr{P}_\mathscr{X}$ is measurable with respect to $\mathscr{U}$ and the Borel σ-algebra generated by the weak topology of $\mathscr{P}_\mathscr{X}$. In the literature of statistics, $(\Theta, \mathscr{U})$ is called a parameter space and a stochastic kernel from $(\Theta, \mathscr{U})$ to $(\mathscr{X}, \mathscr{A})$ is called a statistical model on $(\mathscr{X}, \mathscr{A})$. Let $(\mathscr{X}^{(n)}, \mathscr{A}^{(n)}) = (\mathscr{X}, \mathscr{A}) \times \cdots \times (\mathscr{X}, \mathscr{A})$ be the n-dimensional product measurable space and let $P_\theta^{(n)} = P_\theta \times \cdots \times P_\theta$ be the n-dimensional product measure of P_θ with itself. It is easy to show

that $(P_\theta^{(n)}: \theta \in \Theta)$ is a stochastic kernel from $(\Theta, \mathscr{U})$ to $(\mathscr{X}^{(n)}, \mathscr{A}^{(n)})$. We assume that each P_θ is absolutely continuous with respect to a σ-finite measure λ on $(\mathscr{X}, \mathscr{A})$, i.e., $P_\theta \ll \lambda$ for all $\theta \in \Theta$. If $f(x \mid \theta)$ is a density of P_θ with respect to λ

$$P_\theta^{(n)}(A) = \int_A \prod_{i=1}^n f(x_i \mid \theta) d(\lambda \times \cdots \times \lambda), \quad A \in \mathscr{A}^{(n)}.$$

We further assume that $f(x \mid \theta)$ is measurable as a function of $(\theta, x) \in \Theta \times \mathscr{X}$.

Let π be a prior distribution on $(\Theta, \mathscr{U})$. Define probability measures $\mathbb{P}^{(n)}$ on $(\Theta \times \mathscr{X}^{(n)}, \mathscr{U} \times \mathscr{A}^{(n)})$ by

$$\mathbb{P}^{(n)}(C) = \int_C \prod_{i=1}^n f(x_i \mid \theta) d(\pi \times \lambda \times \cdots \times \lambda), \quad C \in \mathscr{U} \times \mathscr{A}^{(n)}, \quad n \geqslant 1.$$

Since the sequence $\mathbb{P}^{(1)}, \mathbb{P}^{(2)}, \ldots$ is consistent, that is,

$$\mathbb{P}^{(n+1)}(C \times \mathscr{X}) = \mathbb{P}^{(n)}(C)$$

for all $n \geqslant 1$ and $C \in \mathscr{U} \times \mathscr{A}^{(n)}$, it follows by Kolmogorov's consistency theorem that there is a unique probability measure $\mathbb{P} = \mathbb{P}^\pi$ on the infinite product space $(\Omega, \mathscr{F}) = (\Theta \times \mathscr{X}^{(\infty)}, \mathscr{U} \times \mathscr{A}^{(\infty)})$ such that

$$\mathbb{P}(C \times \mathscr{X} \times \mathscr{X} \times \cdots) = \mathbb{P}^{(n)}(C), \quad n \geqslant 1, \quad C \in \mathscr{U} \times \mathscr{A}^{(n)}.$$

Now let us introduce the coordinate functions. For $\omega = (\theta, (x_i)) \in \Omega$ define

$$\vartheta(\omega) = \theta \in \Theta,$$
$$X_i(\omega) = x_i \in \mathscr{X}, \quad i = 1, 2, \ldots.$$

We think of ϑ as the unknown parameter and $X_1, X_2, \ldots$ as data. If $U \in \mathscr{U}$ and $A \in \mathscr{A}^{(n)}$, then

$$\mathbb{P}(\vartheta \in U, (X_1, \ldots, X_n) \in A) = \mathbb{P}(U \times A \times \mathscr{X} \times \mathscr{X} \times \cdots)$$

$$= \mathbb{P}^{(n)}(U \times A) = \int_{U \times A} \prod_{i=1}^{n} f(x_i|\theta) d(\pi \times \lambda \times \cdots \times \lambda)$$

$$= \int_U \int_A \prod_{i=1}^{n} f(x_i|\theta) d(\lambda \times \cdots \times \lambda) d\pi$$

$$= \int_U P_\theta^{(n)}(A) \pi(d\theta).$$

In particular, the marginal distributions of ϑ, $(X_1, \ldots, X_n)$ and X_i are given as follows:

$$\mathbb{P}(\vartheta \in U) = \pi(U), \quad U \in \mathscr{U},$$

$$\mathbb{P}((X_1, \ldots, X_n) \in A) = \int_\Theta P_\theta^{(n)}(A) \pi(d\theta), \quad A \in \mathscr{A}^{(n)},$$

$$\mathbb{P}(X_i \in A) = \int_\Theta P_\theta(A) \pi(d\theta), \quad A \in \mathscr{A}, \quad i = 1, 2, \ldots.$$

For each ω, $P_{\vartheta(\omega)}^{(n)}$ is a probability measure on $(\mathscr{X}^{(n)}, \mathscr{A}^{(n)})$. *If* $A \in \mathscr{A}^{(n)}$, $P_{\vartheta(\omega)}^{(n)}(A)$ *is* $\sigma(\vartheta)$-measurable as a function of ω and

$$\int_{\vartheta \in U} P_{\vartheta(\omega)}^{(n)}(A) d\mathbb{P} = \int_U P_\theta^{(n)}(A) \pi(d\theta) = \mathbb{P}(\vartheta \in U, (X_1, \ldots, X_n) \in A)),$$

which shows that $P_\vartheta^{(n)}(A)$ is a conditional probability of $(X_1, \ldots, X_n) \in A$ given ϑ. Thus $P_{\vartheta(\omega)}^{(n)}(A)$ is a regular conditional distribution of $(X_1, \ldots, X_n)$ given ϑ. In the same way, $P_{\vartheta(\omega)}(A)$, $A \in \mathscr{A}$ is shown to be a regular conditional distribution of X_i given ϑ. In particular, $\mathbb{P}(X_i \in A \mid \vartheta) = P_\vartheta(A)$ a.s., $A \in \mathscr{A}$. Since

$$\mathbb{P}((X_1 \in A_1, \ldots, X_n \in A_n \mid \vartheta) = P_\vartheta^{(n)}(A_1 \times \cdots \times A_n)$$
$$= P_\vartheta(A_1) \cdots P_\vartheta(A_n)$$

$$= \mathbb{P}(X_1 \in A_1 \mid \vartheta) \cdots \mathbb{P}(X_1 \in A_n \mid \vartheta) \text{ a.s.},$$

it follows that $X_1, X_2, \ldots$ are conditionally i.i.d. given ϑ.

Since ϑ is a random variable with values in a Polish space $(\Theta, \mathscr{U})$, there is a regular conditional distribution of ϑ given $X_1, \ldots, X_n$, which we will call the posterior distribution of ϑ given $X_1, \ldots, X_n$ and denote it by $\hat{\mathbb{P}}(\vartheta \in d\theta \mid X_1, \ldots, X_n)(\omega)$ or $\pi_n^{\omega}(d\theta)$.

2 The large deviation principle

Let (Q_n) be a sequence of probability measures on a Polish space $(S, \mathscr{B}(S))$, where $\mathscr{B}(S)$ is the Borel σ-algebra. A function $I : S \to [0, \infty]$ is called a rate function on S if for each $M < \infty$ the level set $\{x \in S : I(x) \leqslant M\}$ is a compact subset of S. It follows necessarily that I is a lower semicontinuous function. In most applications, I is a convex function. The large deviation principle focuses on the asymptotic behavior of the sequence (Q_n) in terms of a rate function I. The precise formulation is as follows.

Definition 1. Let (Q_n) be a sequence of probability measures on S and I a rate function on S. We say that (Q_n) satisfies the large deviation principle with rate function I if the following two conditions hold.

(i) For each closed subset $F \subset S$

$$\limsup_{n \to \infty} \frac{1}{n} \log Q_n(F) \leqslant -\inf_{x \in F} I(x).$$

(ii) For each open subset $G \subset S$

$$\liminf_{n \to \infty} \frac{1}{n} \log Q_n(G) \geqslant -\inf_{x \in G} I(x).$$

It is well known that if (Q_n) satisfies the large deviation principle with a

rate function, then the rate function is uniquely determined. We refer the reader to Dembo and Zeitouni (1998) or Deuschel and Strook (2000) for basic results and a wide variety of large deviation techniques.

A convex function $\psi:\mathbb{R}\to(-\infty,\infty]$ is said to be essentially smooth if

(i) E_ψ° is not empty, where E_ψ° is the interior of $E_\psi \triangleq \{t\in\mathbb{R} : \psi(t)<\infty\}$;

(ii) ψ is differentiable on E_ψ°;

(iii) ψ is steep, namely

$$\lim_{n\to\infty} |\psi'(t_n)| = \infty$$

whenever (t_n) is a sequence in E_ψ° converging to a boundary point of E_ψ°.

Here we state the Gärtner-Ellis theorem, an extension of Crámer's theorem to non-i.i.d. random variables. This is a key theorem in this paper for proving the large deviation principles for posterior distributions of the normal parameters.

Theorem 2. *Let (Q_n) be a sequence of probability measures on $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ such that the cumulant generating function*

$$\psi(t) = \lim_{n\to\infty} \frac{1}{n} \log \int_{-\infty}^{\infty} e^{ntx} Q_n(dx) \tag{1}$$

exists in $[-\infty,\infty]$ for each $t\in\mathbb{R}$. If E_ψ° contains the origin, then $\psi(t)>-\infty$, ψ is convex, and the Fenchel-Legendre transform of ψ (the conjugate function of ψ)

$$I(x) \triangleq \sup_{t\in\mathbb{R}} [tx-\psi(t)] \tag{2}$$

is a rate function on $\mathbb{R}$.

Theorem 3 (**Gärtner-Ellis**). *Let (Q_n) be a sequence of probability measures on $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ such that (1) exists for all $t\in\mathbb{R}$. If E_ψ° contains the origin, then the followings hold with the rate function I defined by (2).*

(i) *For each closed subset* $F \subset S$

$$\limsup_{n\to\infty} \frac{1}{n} \log Q_n(F) \leqslant -\inf_{x\in F} I(x).$$

(ii) For each open subset $G \subset S$

$$\liminf_{n\to\infty} \frac{1}{n} \log Q_n(G) \geqslant -\inf_{x\in G\cap H} I(x),$$

where H *is the set of exposed points of* I*:*

$$H=\{x\in\mathbb{R} : \exists t\in E_\psi \forall y\neq x\ ty-I(y)>tx-I(x)\}.$$

If, in addition, ψ *is essentially smooth and lower semicontinuous, then* (Q_n) *satisfies the large deviation principle with the rate function* I*.*

For the proofs of Theorem 2 and Theorem 3, see Dembo and Zeitouni (1998). Our purpose is to obtain the large deviation principles for posteriors of the normal parameters given data. Accordingly, we have to slightly modify the definitions of rate function and the large deviation principle in order to cover a sequence of conditional distributions. Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space, $(\mathscr{G}_n)$ an increasing sequence of sub σ-algebras, and Y random variable taking values in $(S, \mathscr{B}(S))$. Then for each $n \geqslant 1$ there exists a function $Q_n : (\omega, B) \in \Omega \times \mathscr{B}(S) \mapsto Q_n^\omega(B) \in [0,1]$ such that

(i) for each $\omega \in \Omega$ $Q_n^\omega(B)$ is a probability measure on $(S, \mathscr{B}(S))$;

(ii) for each $B \in \mathscr{B}(S)$ $Q_n^\omega(B)$ is a version of $\mathbb{P}(Y \in B \mid \mathscr{G}_n)(\omega)$.

The function Q_n is called a regular conditional distribution of Y given $\mathscr{G}_n$.

We define a measurable function $I : \Omega \times S \to [0, \infty]$ to be a rate function on S if $I(\omega, \cdot)$ is a rate function for each $\omega \in \Omega$. It is natural to define the large deviation principle for a sequence of conditional distributions (Q_n) as follows.

Definition 4. Let $I : \Omega \times S \to [0, \infty]$ be a rate function on S. We say that a

sequence (Q_n) satisfies the large deviation principle with rate function I if the following two conditions hold.

(i) For each closed subset $F \subset S$

$$\limsup_{n\to\infty} \frac{1}{n} \log Q_n^{\omega}(F) \leqslant -\inf_{x\in F} I(\omega, x) \quad \text{a.s.}$$

(ii) For each open subset $G \subset S$

$$\liminf_{n\to\infty} \frac{1}{n} \log Q_n^{\omega}(G) \geqslant -\inf_{x\in G} I(\omega, x) \quad \text{a.s.}$$

3 The normal model with unknown mean and known variance

Suppose that $(\Theta, \mathscr{U}) = (\mathscr{X}, \mathscr{A}) = (\mathbb{R}, \mathscr{B}(\mathbb{R}))$, where $\mathscr{B}(\mathbb{R})$ is the Borel σ-algebra of $\mathbb{R}$. Let $P_\theta = \mathrm{N}(\theta, \sigma^2)$, the normal distribution with mean $\theta \in \Theta$ and known variance σ^2. If we choose the normal distribution $N(\mu, v^2)$ for the prior distribution of ϑ, then the posterior distribution of ϑ given $X_1, \ldots, X_n$ is the normal distribution $N(\mu_n, v_n^2)$, where

$$\mu_n = \frac{\sigma^2/n}{\sigma^2/n + v^2}\mu + \frac{v^2}{\sigma^2/n + v^2}\bar{X}_n, \quad \bar{X}_n = \frac{X_1 + \cdots + X_n}{n},$$

$$\frac{1}{v_n^2} = \frac{1}{v^2} + \frac{1}{\sigma^2/n}.$$

Since $X_1, X_2, \ldots$ are conditionally i.i.d. given ϑ and X_1 is integrable under $\mathbb{P}$, the law of large numbers for conditionally i.i.d. random variables entails the convergence $\bar{X}_n \to \mathbb{E}(X_1 \mid \vartheta) = \vartheta$ a.s., and hence we have

$$\mu_n \to \vartheta \quad a.s.$$

Since

$$\int_{-\infty}^{\infty} e^{nt\theta}\,\pi_n^{\omega}(d\theta)=\exp\left(\mu_n nt+\frac{v_n^2 n^2 t^2}{2}\right)$$

we have the cumulant generating function

$$\psi(t)=\lim_{n\to\infty}\frac{1}{n}\log\int_{-\infty}^{\infty} e^{t\theta}\,\pi_n^{\omega}(d\theta)=\vartheta t+\frac{\sigma^2 t^2}{2} \tag{3}$$

which is that of $N(\vartheta,\sigma^2)$. If we select the Lebesgue measure on $\mathbb{R}$ as the prior for ϑ, then the posterior distribution is given by $N(\bar{X}_n,\sigma^2/n)$, which is the limiting distribution of $N(\mu_n,v_n^2)$ as $v^2\to\infty$. It follows that the cumulant generating function of the posterior distribution is the same as (3). It is easy to see that $E_\psi^\circ=\mathbb{R}$, ψ is differentiable and steep for each ω, so that ψ is essentially smooth and continuous for each ω. It is not difficult to show that the Fenchel-Legendre transform of ψ for each ω is given by

$$I(\omega,x)=\frac{(x-\vartheta(\omega))^2}{2\sigma^2}. \tag{4}$$

Now we obtain the following theorem.

Theorem 5. *Suppose that $X_1,X_2,\ldots$ are conditionally i.i.d. with distribution $N(\theta,\sigma^2)$ given $\vartheta=\theta\in\Theta=\mathbb{R}$. If $N(\mu,v^2)$, $\mu\in\mathbb{R}$, $v^2>0$, is the prior for ϑ, then the posterior distributions satisfy the large deviation principle with rate function (4). If we choose the Legesgue measure as the prior for ϑ, the same result follows.*

4 The normal model with known mean and unknown precision

Let $X_1,X_2,\ldots$ be conditionally i.i.d. random variables sampled from $P_\theta=N(\mu,1/\theta)$, $\theta\in\Theta=(0,\infty)$. Assume that the prior distribution of the precision ϑ is a gamma distribution $G(\alpha,\beta)$, $\alpha>0$, $\beta>0$. Then the posterior dis-

tribution of ϑ given $X_1, \ldots, X_n$ is a gamma $G(\alpha_n, \beta_n)$, where

$$\alpha_n = \alpha + \frac{n}{2},$$

$$\beta_n = \beta + \frac{n}{2} S_n^2, \quad S_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_1 - \mu)^2.$$

Note that by the law of large numbers for conditionally i.i.d random variables

$$S_n^2 \to \mathbb{E}[(X_1 - \mu)^2 \mid \vartheta] = \frac{1}{\vartheta} \quad \text{a.s.}$$

By the moment generating function of $G(\alpha_n, \beta_n)$, we have

$$\int_0^\infty e^{nt\theta} \pi_n^\omega(d\theta) = \begin{cases} \left(\frac{\beta_n}{\beta_n - nt}\right)^{\alpha_n}, & t < \frac{\beta_n}{n}, \\ \infty, & t \geqslant \frac{\beta_n}{n}, \end{cases}$$

and hence we obtain the cumulant generating function

$$\psi(t) \triangleq \lim_{n\to\infty} \frac{1}{n} \log \int_0^\infty e^{nt\theta} \pi_n^\omega(d\theta)$$

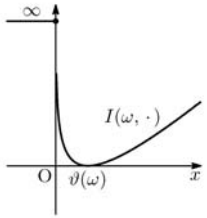
$$= \begin{cases} \log\left(\frac{1/(2\vartheta)}{1/(2\vartheta) - t}\right)^{1/2}, & t < \frac{1}{2\vartheta}, \\ \infty, & t \geqslant \frac{1}{2\vartheta}. \end{cases} \tag{5}$$

This is the cumulant generating function of $G(1/2, 1/(2\vartheta))$. If we utilize the improper prior

$$\pi(d\theta) \propto \frac{1}{\theta} d\theta \tag{6}$$

which is the Jefferys' prior for ϑ and obtained formally by letting $\alpha \to 0$ and $\beta \to 0$ in $G(\alpha, \beta)$, then the same cumulant generating function as (5) is obtained. The Fenchel-Legendre transform of ψ has the following form.

$$I(\omega,x)=\begin{cases}\dfrac{1}{2}\left[\dfrac{x}{\vartheta(\omega)}-\log\dfrac{x}{\vartheta(\omega)}-1\right], & x>0,\\ \infty, & x\leqslant 0.\end{cases} \tag{7}$$



***I*(ω, ·) defined by (7).**

By Theorem 2, $I(\omega,\cdot)$ is a convex rate function for each ω, and by Theorem 3 we have the following result.

Theorem 6. *Suppose that $X_1, X_2, \ldots$ are conditionally i.i.d. with distribution $N(\mu,1/\theta)$ given $\vartheta=\theta\in\Theta=(0,\infty)$. If a gamma distribution $G(\alpha,\beta)$, $\alpha>0$, $\beta>0$ is the prior for ϑ, then*

$$\limsup_{n\to\infty}\frac{1}{n}\log\hat{\mathbb{P}}(\vartheta\in G\mid X_1,\ldots,X_n)(\omega)\leqslant-\inf_{x\in G}I(\omega,x)\ a.s.$$

for all open sets $G\subset(0,\infty)$, and

$$\liminf_{n\to\infty}\frac{1}{n}\log\hat{\mathbb{P}}(\vartheta\in F\mid X_1,\ldots,X_n)(\omega)\geqslant-\inf_{x\in F\cap H_\omega}I(\omega,x)\ a.s.$$

for all closed sets $F\subset(0,\infty)$, where H_ω is the set of exposed points of $I(\omega,\cdot)$. If we select the Jeffreys's prior (6) for ϑ, the same result follows.

References

[1] Dembo, A. and Zeitouni, O. (1998). *Large Deviations Techniques and Applications*, 2nd ed., Springer-Verlag, New York.

[2] Deuschel, J. D. and Stroock, D. W. (2000). *Large Deviations*, AMS Chelsea Publishing, Amer. Math. Soc.

[3] Eichelsbacher, P. and Ganash, A.J. (2002). Moderate deviations for Bayes posteriors. *Scand. J. Statist.* 29, 153-167.

[4] Fu, J. C. and Kass, R. E. (1988). The exponential rate of convergence of posterior distributions, *Ann. Inst. Statist. Math.*, 40, 683-691.

[5] Ganesh, A. and O'Connell, N. (1999). An inverse of Sanov's theorem, *Statist. Probab. Lett.* 42, 201-206.

[6] Shen, X. and Wasserman, L. (1998). Rates of convergence of posterior distributions, *Ann. Probab.*, 29, 687-714.