An inverse of Sanov's theorem

Ayalvadi Ganesh and Neil O'Connell

BRIMS, Hewlett-Packard Labs, Bristol

Abstract

Let X_k be a sequence of iid random variables taking values in a finite set, and consider the problem of estimating the law of X_1 in a Bayesian framework. We prove that the sequence of posterior distributions satisfies a large deviation principle, and give an explicit expression for the rate function. As an application, we obtain an asymptotic formula for the predictive probability of ruin in the classical gambler's ruin problem.

1 Introduction and preliminaries

Let $\mathcal{X}$ be a Hausdorff topological space with Borel σ-algebra $\mathcal{B}$, and let μ_n be a sequence of probability measures on $(\mathcal{X}, \mathcal{B})$. A *rate function* is a non-negative lower semicontinuous function on $\mathcal{X}$. We say that the sequence μ_n satisfies the *large deviation principle* (LDP) with rate function I, if for all $B \in \mathcal{B}$,

$$-\inf_{x \in B^\circ} I(x) \leq \liminf_n \frac{1}{n} \log \mu_n(B) \leq \limsup_n \frac{1}{n} \log \mu_n(B) \leq -\inf_{x \in \bar{B}} I(x).$$

Here B° and $\bar{B}$ denote the interior and closure of B, respectively.

Let Ω be a finite set and denote by $\mathcal{M}_1(\Omega)$ the space of probability measures on Ω. Consider a sequence of independent random variables X_k taking values in Ω, with common law $\mu \in \mathcal{M}_1(\Omega)$. Denote by L_n the empirical measure corresponding to the first n observations:

$$L_n = \frac{1}{n} \sum_{k=1}^{n} \delta_{X_k},$$

where δ_{X_k} denotes unit mass at X_k. We denote the law of L_n by $\mathcal{L}(L_n)$. For $\nu \in \mathcal{M}_1(\Omega)$ define its *relative entropy* (relative to μ) by

$$H(\nu|\mu) = \begin{cases} \int_\Omega \frac{d\nu}{d\mu} \log \frac{d\nu}{d\mu} d\mu & \nu \ll \mu \\ \infty & \text{otherwise.} \end{cases}$$

The statement of *Sanov's theorem* is that the sequence $\mathcal{L}(L_n)$ satisfies the LDP in $\mathcal{M}_1(\Omega)$ with rate function $H(\cdot|\mu)$.

In this paper we present an inverse of this result, which arises naturally in a Bayesian setting. The underlying distribution (of the X_k's) is unknown, and has a prior distribution $\pi \in \mathcal{M}_1(\mathcal{M}_1(\Omega))$. The posterior distribution, given the first n observations, is a function of the empirical measure L_n and will be denoted by $\pi^n(L_n)$. We prove that, on the set $\{L_n \to \mu\}$, for any fixed μ in the support of the prior, the sequence $\pi^n(L_n)$ satisfies the LDP in $\mathcal{M}_1(\Omega)$ with rate function given by $H(\mu|\cdot)$ on the support of the prior (otherwise it is infinite). Note that the roles played by the arguments of the relative entropy function are interchanged. This result can be understood intuitively as follows: in Sanov's theorem we ask how likely the empirical distribution is to be close to ν, given that the true distribution is μ, whereas in the Bayesian context we ask how likely it is that the true distribution is close to ν, given that the empirical distribution is close to μ. We regard these questions as natural inverses of each other.

As an application, we obtain an asymptotic formula for the *predictive* probability of ruin in the classical gambler's ruin problem.

Admittedly, the assumption that Ω is finite is very strong. However, it is the only case where the result can be stated *without making additional assumptions about the prior.* To see that this is a delicate issue, note that, since $H(\mu|\mu) = 0$, the LDP implies consistency of the posterior distribution: it was shown by Freedman [5] that Bayes estimates can be inconsistent even on countable Ω and even when the 'true' distribution is in the support of the prior; moreover, sufficient conditions for consistency which exist in the literature are quite disparate and in general far from being necessary. The problem of extending our result would seem to be an interesting (and formidable!) challenge for future research. We have recently extended the result to the case where Ω is a compact metric space, under the condition that the prior is of Dirichlet form [8].

There is a considerable literature on the asymptotic behaviour of Bayes posteriors; see, for example, Le Cam [11, 12], Freedman [5], Lindley [13], Schwartz [14], Johnson [9, 10], Brenner *et al.* [1], Chen [2], Diaconis and Freedman [4] and Fu and Kass [6]. Freedman, and Diaconis and Freedman study conditions under which the Bayes estimate (the mean of the posterior distribution) is consistent. Le Cam, Freedman, Brenner *et al.* and Chen establish asymptotic normality of the posterior distribution under different conditions on the parameter space and the prior. Schwartz and Johnson study asymptotic expansions of the posterior distribution in powers of $n^{-1/2}$, having a normal distribution as the leading term. Fu and Kass establish an LDP for the posterior distribution of the parameter in a one-dimensional parametric family, under certain regularity conditions.

2 The LDP

Let Ω be a finite set, and let $\mathcal{M}_1(\Omega)$ denote the space of probability measures on Ω. Suppose $X_1,\ X_2, \ldots$ are i.i.d. Ω–valued random variables.

Let $\pi \in \mathcal{M}_1(\mathcal{M}_1(\Omega))$ denote the prior distribution on the space $\mathcal{M}_1(\Omega)$, with support denoted by supp π. For each n, set

$$\mathcal{M}_1^n(\Omega) = \left\{ \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \,:\, x \in \Omega^n \right\}.$$

Define a mapping $\pi^n : \mathcal{M}_1^n(\Omega) \to \mathcal{M}_1(\mathcal{M}_1(\Omega))$ by its Radon-Nikodym derivative on the support of π:

$$\frac{d\pi^n(\mu_n)}{d\pi}(\nu) = \frac{\prod_{x\in\Omega} \nu(x)^{n\mu_n(x)}}{\int_{\mathcal{M}_1(\Omega)} \prod_{x\in\Omega} \lambda(x)^{n\mu_n(x)} \pi(d\lambda)}, \tag{1}$$

if the denominator is non-zero; π^n is undefined at μ_n if the denominator is zero. When defined, $\pi^n(\mu_n)$ denotes the posterior distribution, conditional on the observations $X_1, \ldots, X_n$ having empirical distribution $\mu_n \in \mathcal{M}_1^n(\Omega)$. (This follows immediately from Bayes formula; there are no measurability concerns here since Ω is finite.)

Theorem 1 *Suppose $x \in \Omega^{\mathbb{N}}$ is such that the sequence $\mu_n = \sum_{i=1}^n \delta_{x_i}/n$ converges weakly to $\mu \in$ supp π and that $\mu(x) = 0$ implies that $\mu_n(x) = 0$ for all n. Then the sequence of laws $\pi^n(\mu_n)$ satisfies a large deviation principle (LDP) with good rate function*

$$I(\nu) = \begin{cases} H(\mu|\nu), & \textit{if } \nu \in \textit{supp } \pi \\ \infty, & \textit{else.} \end{cases}$$

The rate function $I(\cdot)$ is convex if supp π is convex.

Proof : Observe that

$$\begin{aligned}\frac{1}{n}\log\prod_{x\in\Omega}\lambda(x)^{n\mu_n(x)} &= \sum_{x\in\Omega}\mu_n(x)\log\mu_n(x)-\sum_{x\in\Omega}\mu_n(x)\log\frac{\mu_n(x)}{\lambda(x)}\\ &= -H(\mu_n)-H(\mu_n|\lambda)\ \le\ -H(\mu_n).\end{aligned}$$

Here, $H(\mu_n)=\sum_{x\in\Omega}\mu_n(x)\log\mu_n(x)$ denotes the entropy of μ_n. The last inequality follows from the fact that relative entropy is non-negative. It follows, since π is a probability measure on $\mathcal{M}_1(\Omega)$, that

$$\int_{\mathcal{M}_1(\Omega)}\prod_{x\in\Omega}\lambda(x)^{n\mu_n(x)}\pi(d\lambda)\le\exp(-nH(\mu_n)).$$

Thus,

$$\limsup_{n\to\infty}\frac{1}{n}\log\int_{\mathcal{M}_1(\Omega)}\prod_{x\in\Omega}\lambda(x)^{n\mu_n(x)}\pi(d\lambda)\le\limsup_{n\to\infty}-H(\mu_n)=-H(\mu);\quad(2)$$

here we are using the fact that $H(\cdot)$ is continuous.

Next, since μ_n converges to $\mu\in$ supp π, we have that for all $\epsilon>0$, $\pi(B(\mu,\epsilon))>0$ and $\mu_n\in B(\mu,\epsilon)$ for all n sufficiently large. (Here, $B(\mu,\epsilon)$ denotes the set of probability distributions on Ω that are within ϵ of μ in total variation distance—note that this generates the weak topology since Ω is finite.) Therefore,

$$\frac{1}{n}\log\int_{\mathcal{M}_1(\Omega)}\prod_{x\in\Omega}\lambda(x)^{n\mu_n(x)}\pi(d\lambda)\ \ge\ \frac{1}{n}\log\int_{B(\mu,\epsilon)}\prod_{x\in\Omega}\lambda(x)^{n\mu_n(x)}\pi(d\lambda)\ \ge$$
$$\frac{1}{n}\log\pi\Big(B(\mu,\epsilon)\Big)+\sum_{x\in\Omega}\mu_n(x)\log\mu_n(x)-\sup_{\lambda\in B(\mu,\epsilon)}\sum_{x\in\Omega:\mu(x)>0}\mu_n(x)\log\frac{\mu_n(x)}{\lambda(x)}.$$

To obtain the last inequality, we have used the assumption that, if $\mu(x)=0$, then $\mu_n(x)=0$ for all n. We also use the convention that $0\log 0=0$. Since $\pi(B(\mu,\epsilon))>0$, it follows from the above, again using the continuity of $H(\cdot)$, that

$$\begin{aligned} & \liminf_{n\to\infty} \frac{1}{n} \log \int_{\mathcal{M}_1(\Omega)} \prod_{x\in\Omega} \lambda(x)^{n\mu_n(x)} \pi(d\lambda) \\ \geq & \quad -H(\mu) - \sup_{\lambda,\rho\in B(\mu,\epsilon)} \sum_{x\in\Omega:\mu(x)>0} \rho(x) \log \frac{\rho(x)}{\lambda(x)}. \end{aligned} \tag{3}$$

Letting $\epsilon \to 0$, we get from (2) and (3) that

$$\lim_{n\to\infty} \frac{1}{n} \log \int_{\mathcal{M}_1(\Omega)} \prod_{x\in\Omega} \lambda(x)^{n\mu_n(x)} \pi(d\lambda) = -H(\mu). \tag{4}$$

LD upper bound: Let F be an arbitrary closed subset of $\mathcal{M}_1(\Omega)$. If $\pi(F) = 0$, then $\pi^n(\mu_n)(F) = 0$ for all n and the LD upper bound,

$$\limsup_{n\to\infty} \frac{1}{n} \log \pi^n(\mu_n)(F) \leq -\inf_{\nu\in F} I(\nu),$$

holds trivially. Otherwise, if $\pi(F) > 0$, observe from (1) and (4) that

$$\begin{aligned} & \limsup_{n\to\infty} \frac{1}{n} \log \pi^n(\mu_n)(F) - H(\mu) \;=\; \limsup_{n\to\infty} \frac{1}{n} \log \int_F \lambda(x)^{n\mu_n(x)} \pi(d\lambda) \\ \leq & \quad \limsup_{n\to\infty} \frac{1}{n} \log \pi(F) + \limsup_{n\to\infty} \sup_{\lambda\in F\cap \text{supp } \pi} \sum_{x\in\Omega} \mu_n(x) \log \lambda(x) \\ = & \quad \limsup_{n\to\infty} \sup_{\lambda\in F\cap \text{supp } \pi} [-H(\mu_n) - H(\mu_n|\lambda)] \\ \leq & \quad \sup_{\rho\in B(\mu,\delta)} \sup_{\lambda\in F\cap \text{supp } \pi} [-H(\rho) - H(\rho|\lambda)] \quad \forall\, \delta > 0, \end{aligned}$$

where the last inequality is because μ_n converges to μ. Letting $\delta \to 0$, we get from the continuity of $H(\cdot)$ and $H(\cdot|\cdot)$ that

$$\limsup_{n\to\infty} \frac{1}{n} \log \pi^n(\mu_n)(F) \leq -\inf_{\lambda\in F\cap \text{supp } \pi} H(\mu|\lambda) = -\inf_{\lambda\in F} I(\lambda), \tag{5}$$

where the last equality follows from the definition of I in the statement of the theorem. This completes the proof of the large deviations upper bound.

LD lower bound: Fix $\nu \in$ supp π, and let $B(\nu, \epsilon)$ denote the set of probability distributions on Ω that are within ϵ of ν in total variation. Then,

$\pi(B(\nu,\epsilon)) > 0$ for any $\epsilon > 0$. Using (1) and (4) we thus have, for all $\delta \in (0,\epsilon)$,

$$\begin{aligned} & \liminf_{n\to\infty} \frac{1}{n} \log \pi^n(\mu_n)\Big(B(\nu,\epsilon)\Big) - H(\mu) \\ \geq\ & \liminf_{n\to\infty} \frac{1}{n} \log \int_{B(\nu,\delta)} \lambda(x)^{n\mu_n(x)} \pi(d\lambda) \\ \geq\ & \liminf_{n\to\infty} \frac{1}{n} \log \pi\Big(B(\nu,\delta)\Big) + \liminf_{n\to\infty} \inf_{\lambda\in B(\nu,\delta)} \sum_{x\in\Omega} \mu_n(x) \log \lambda(x) \\ \geq\ & \inf_{\rho\in B(\mu,\delta)} \inf_{\lambda\in B(\nu,\delta)} \left[-H(\rho) - H(\rho|\lambda)\right]. \end{aligned}$$

Letting $\delta \to 0$, we have by the continuity of $H(\cdot)$ and $H(\cdot|\cdot)$ that

$$\liminf_{n\to\infty} \frac{1}{n} \log \pi^n(\mu_n)\Big(B(\nu,\epsilon)\Big) \geq -H(\mu|\nu) = -I(\nu), \tag{6}$$

where the last equality is because we took ν to be in supp π.

Let G be an arbitrary open subset of $\mathcal{M}_1(\Omega)$. If $G \cap \text{supp } \pi$ is empty, then, by the definition of I in the statement of the theorem, $\inf_{\nu\in G} I(\nu) = \infty$. Therefore, the large deviations lower bound,

$$\liminf_{n\to\infty} \frac{1}{n} \log \pi^n(\mu_n)(G) \geq -\inf_{\nu\in G} I(\nu),$$

holds trivially. On the other hand, if $G \cap \text{supp } \pi$ isn't empty, then we can pick $\nu \in G$ such that $I(\nu)$ is arbitrarily close to $\inf_{\lambda\in G} I(\lambda)$, and $\epsilon > 0$ such that $B(\nu,\epsilon) \subseteq G$. So $\pi^n(\mu_n)(G) \geq \pi^n(\mu_n)\Big(B(\nu,\epsilon)\Big)$ for all n, and it follows from (6) that

$$\liminf_{n\to\infty} \frac{1}{n} \log \pi^n(\mu_n)(G) \geq -\inf_{\lambda\in G} I(\lambda).$$

This completes the proof of the large deviations lower bound, and of the theorem.

3 Application to the gambler's ruin problem

Suppose now that Ω is a finite subset of $\mathbb{R}$. As before, X_k is a sequence of iid random variables with common law $\mu \in \mathcal{M}_1(\Omega)$, and we are interested in level-crossing probabilities for the random walk $S_n = X_1 + \cdots + X_n$. For $Q > 0$, denote by $R(Q,\mu)$ the probability that the walk ever exceeds the level Q. If a gambler has initial capital Q, and loses amount X_k on the k^{th} bet, then $R(Q,\mu)$ is the probability of ultimate ruin. If the underlying distribution μ is unknown, the gambler may wish to assess this probability based on experience: this leads to a *predictive* probability of ruin, given by the formula

$$P_n(Q,\mu_n) = \int R(Q,\lambda)\pi^n(d\lambda),$$

where, as before, μ_n is the empirical distribution of the first n observations and $\pi^n \equiv \pi^n(\mu_n)$ is the posterior distribution as defined in equation (1). A standard refinement of Wald's approximation yields

$$C\exp(-\delta(\mu)Q) \leq R(Q,\mu) \leq \exp(-\delta(\mu)Q),$$

for some $C > 0$, where

$$\delta(\mu) = \sup\{\theta \geq 0 : \int e^{\theta x}\mu(dx) \leq 1\}.$$

Thus,

$$C\int \exp(-\delta(\lambda)Q)\pi^n(d\lambda) \leq P_n(Q,\mu_n) \leq \int \exp(-\delta(\lambda)Q)\pi^n(d\lambda),$$

and we can apply Varadhan's lemma (see, for example, [3, Theorem 4.3.1]) to obtain the asymptotic formula, for $q > 0$,

$$\lim_{n\to\infty} \frac{1}{n}\log P_n(qn,\mu_n) = -\inf\{H(\mu|\nu) + \delta(\nu)q : \nu \in \text{supp } \pi\},$$

on the set $\mu_n \to \mu$. We are also assuming, as in Theorem 1, that $\mu_n(x) = 0$, for all n, whenever $\mu(x) = 0$, and using the easy (Ω is finite) fact that $\delta : \mathcal{M}_1(\Omega) \to \mathbb{R}_+$ is continuous. This formula can be simplified in special cases. Its implications for risk and network management are discussed in Ganesh *et al.* [7].

References

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

[2] C. F. Chen, On asymptotic normality of limiting density functions with Bayesian implications, *J. Roy. Statist. Soc. Ser. B*, 47(3): 540–546, 1985.

[3] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, Jones and Bartlett, 1993.

[4] Persi Diaconis and David Freedman, On the consistency of Bayes estimates, *Ann. Statist.*, 14(1): 1–26, 1986.

[5] David A. Freedman, On the asymptotic behavior of Bayes estimates in the discrete case, *Ann. Math. Statist.* 34: 1386–1403, 1963.

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

[7] Ayalvadi Ganesh, Peter Green, Neil O'Connell and Susan Pitts, Bayesian network management. To appear in *Queueing Systems.*

[8] A. J. Ganesh and Neil O'Connell, A large deviations principle for Dirichlet process posteriors, Hewlett-Packard Technical Report, HPL-BRIMS-98-05, 1998.

[9] R. A. Johnson, An asymptotic expansion for posterior distributions, *Ann. Math. Statist.*, 38: 1899–1907, 1967.

[10] R. A. Johnson, Asymptotic expansions associated with posterior distributions, *Ann. Math. Statist.*, 41: 851–864, 1970.

[11] L. Le Cam, On some asymptotic properties of maximum likelihood estimates and related Bayes estimates, *Univ. Calif. Publ. Statist.*, 1: 227–330, 1953.

[12] L. Le Cam, Locally asymptotically normal families of distributions, *Univ. Calif. Publ. Statist.*, 3: 37–98, 1957.

[13] D. V. Lindley, *Introduction to Probability and Statistics from a Bayesian Viewpoint*, Cambridge University Press, 1965.

[14] L. Schwartz, On Bayes procedures, *Z. Wahrsch. Verw. Gebiete*, 4: 10–26, 1965.