HEWLETT PACKARD

An Inverse of Sanov's Theorem

Ayalvadi Ganesh, Neil O'Connell
Basic Research Institute in the Mathematical Sciences
HP Laboratories Bristol
HPL-BRIMS-97-25
November, 1997

large deviations,
non-parametric
Bayes

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 distribution 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.

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 $B \in \mathcal{B}$,

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

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:

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

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.

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

It is clear to us that the result should hold for more general Ω, but not 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 [2] 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.

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 suppπ. 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:

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

here, $\pi^n(\mu_n)$ denotes the posterior distribution, conditional on the observations $X_1, \ldots, X_n$ having empirical distribution $\mu_n \in \mathcal{M}_1^n(\Omega)$.

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 $\mu \in$ suppπ 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*

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

The rate function $I(\cdot)$ is convex if suppπ 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 $\mu\in\text{supp}\pi$, 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,

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

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

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

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

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

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

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

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 $\nu \in \text{supp}\pi$, 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 $\delta \in (0, \epsilon)$,

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

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 $\text{supp}\pi$.

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

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

for some $C > 0$, where

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

Thus,

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

and we can apply Varadhan's lemma (see, for example, [1, 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 [3].

References

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

[2] D. Freedman, On the asymptotic behavior of Bayes estimates in the discrete case, *Ann. Statist.* 34 (1963) 1386–1403.

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