A large deviation principle for Dirichlet posteriors

A. J. Ganesh[1] *and Neil O'Connell*
Microsoft Research, 1 Guildhall Street, Cambridge CB2 3NH, U.K.
BRIMS, Hewlett-Packard Labs, Filton Road, Bristol BS12 6QZ, U.K.
E-mail: *ajg@microsoft.com, noc@hplb.hpl.hp.com*

Abstract

Let X_k be a sequence of independent and identically distributed random variables taking values in a compact metric space Ω, and consider the problem of estimating the law of X_1 in a Bayesian framework. A conjugate family of priors for non-parametric Bayesian inference is the Dirichlet process priors popularized by Ferguson. We prove that if the prior distribution is Dirichlet, then 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.

Keywords: Bayesian nonparametrics, large deviations, asymptotics, Dirichlet process.

1 Introduction

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 $B \in \mathcal{B}$,

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

Let Ω be a complete, separable metric space (Polish space) 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

[1]Research carried out in part while the author was at BRIMS, Hewlett-Packard Labs.

$\mu \in \mathcal{M}_1(\Omega)$. 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 Kullback-Leibler distance or *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 $\mathcal{M}_1(\Omega)$ equipped with the τ-topology (see Dembo and Zeitouni 1993, Theorem 6.2.10), with rate function $H(\cdot|\mu)$. As a corollary, the LDP also holds in the weak topology on $\mathcal{M}_1(\Omega)$, which is weaker than the τ-topology.

In an earlier paper (Ganesh and O'Connell 1999), we proved an inverse of this result, which arises naturally in a Bayesian setting, for finite sets Ω. 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 is denoted $\pi^n(L_n)$. We showed 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 compared to Sanov's theorem. We pointed out that the extension of the result to more general Ω would require 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 in the topology generated by Kullback-Leibler neighbourhoods; in particular, it implies weak consistency. But it was shown by Freedman (1963) that Bayes estimates can be inconsistent even for countable Ω; even if the true distribution is in the weak support of the prior, it does not follow that the posterior mass of each weak neighbourhood tends to 1 (in fact, it can tend to zero!).

There has recently been renewed interest in the consistency of non-parametric Bayes methods, prompted by their increasing popularity in applied work. A notable early result in this field is due to Schwartz (1965), who showed that if the prior assigns positive probability to every *Kullback-Leibler* neighbourhood of the true distribution, then the posterior is *weakly* consistent. If, in addition, the relevant space of probability distributions satisfies a 'metric

entropy' condition, then Barron *et al.* (1999) show that the posterior concentrates on neighbourhoods defined by the Hellinger metric; these are finer than weak neighbourhoods. (The Hellinger distance between two densities f and g with respect to a reference measure μ is defined by $\int(\sqrt{f}-\sqrt{g})^2 d\mu$.) Recent research on the consistency of Bayes methods is reviewed by Ghosal *et al.* (1999) and Wasserman (1998). Rates of convergence of the posterior have been investigated by Ghosal *et al.* (1998) and Shen and Wasserman (1998), but there is relatively little work on more refined asymptotics.

In this paper, we prove an LDP for the special (but nevertheless, useful) case of Dirichlet process priors on a compact metric space. The problem of extending our results to an arbitrary Polish space remains open.

An LDP with a similar flavour for a sequence of Dirichlet processes has been derived by Lynch and Sethuraman (1987); we compare our result with theirs following the statement of Theorem 1. The techniques we use in this paper are very different from those of Lynch and Sethuraman, who obtain their results as a consequence of an LDP they derive for processes with stationary, independent increments. We believe that our methods are of independent interest, and also that they can be generalized to a wider class of prior distributions.

The LDP for Dirichlet posteriors derived here has applications to queue and risk management that are discussed in Ganesh et al. (1998). Some questions of interest in this context are posed in terms of the ruin probability in the classical gambler's ruin problem. In Section 3, we use the LDP for the posterior distributions to obtain an asymptotic formula for the predictive probability of ruin.

2 The LDP

Let Ω be a compact metric space with Borel σ–algebra $\mathcal{F}$. Let $\mathcal{M}_1(\Omega)$ denote the space of probability measures on $(\Omega, \mathcal{F})$, and $\mathcal{B}(\mathcal{M}_1(\Omega))$ the Borel σ–algebra induced by the weak topology on $\mathcal{M}_1(\Omega)$. In this case, it is not possible to establish an LDP for Bayes posteriors corresponding to arbitrary prior distributions, for reasons discussed above. Therefore, we shall work with a specific family of priors, namely Dirichlet process priors; see Ferguson (1973) for a detailed discussion of their properties.

The n-dimensional *Dirichlet distribution* with parameter $a = (a_1, \ldots, a_n)$,

denoted $D(a)$, is defined to be the joint distribution of $(Z_1/Z, \ldots, Z_n/Z)$, where Z_i, $i = 1, \ldots, n$ are mutually independent, Z_i has the gamma distribution with shape parameter a_i and scale parameter 1, and $Z = Z_1 + \ldots + Z_n$.

Denote by $\mathcal{M}_+(\Omega)$ the space of finite non-negative measures on $(\Omega, \mathcal{F})$. The "*Dirichlet process*" with parameter $\alpha \in \mathcal{M}_+(\Omega)$, denoted $\mathcal{D}(\alpha)$, is a probability distribution on $\mathcal{M}_1(\Omega)$. A random probability measure, μ, on Ω is said to have law $\mathcal{D}(\alpha)$ if, for every finite measurable partition $(A_1, \ldots, A_n)$ of Ω, the vector $(\mu(A_1), \ldots, \mu(A_n))$ has the n-dimensional Dirichlet distribution $D(\alpha(A_1), \ldots, \alpha(A_n))$. The distribution of $(\mu(B_1), \ldots, \mu(B_n))$ for arbitrary measurable $B_1, \ldots, B_n$ follows in an obvious way from the distributions for partitions.

Let π be a Dirichlet process prior, $\mathcal{D}(\alpha)$, on the space $\mathcal{M}_1(\Omega)$. Then, conditional on observing ω_1,...,ω_n, the posterior distribution is also a Dirichlet process, but with parameter $\alpha + \sum_{i=1}^n \delta_{\omega_i}$, where δ_x denotes Dirac measure at x (see Ferguson 1973, 1974). In other words, the Dirichlet processes $\mathcal{D}(\alpha)$, $\alpha \in \mathcal{M}_+(\Omega)$ are a conjugate family of priors. This property greatly facilitates computation of posterior distributions and is very useful in analytical work. We now prove a large deviation principle for the sequence of distributions, $\{\mathcal{D}(\alpha + \sum_{i=1}^n \delta_{\omega_i}), n = 1, 2, \ldots\}$.

Theorem 1 *Let α be a finite non-negative measure on $(\Omega, \mathcal{B}(\Omega))$, with support Ω. Let μ be a probability measure on $(\Omega, \mathcal{B}(\Omega))$, and let $\{x_n\}$ be an Ω-valued sequence such that*

$$\frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \to \mu \quad \textit{weakly},$$

where δ_{x_i} denotes Dirac measure at x_i. Then the sequence of probability measures, $\mathcal{D}(\alpha + \sum_{i=1}^n \delta_{x_i})$, satisfies an LDP in $\mathcal{M}_1(\Omega)$ equipped with its weak topology, with rate function $I(\cdot)$ given by

$$I(\nu) = H(\mu|\nu),$$

where $H(\mu|\nu)$ denotes the relative entropy of μ with respect to ν.

Corollary: If $X_i, i \in \mathbb{N}$ are independent and identically distributed with common law μ, then the sequence of empirical distributions, $(1/n) \sum_{i=1}^n \delta_{X_i}$, converges weakly to μ with probability one. Hence, the sequence of random

probability measures $\mathcal{D}(\alpha + \sum_{i=1}^{n} \delta_{X_i})$ almost surely satisfies an LDP (on $\mathcal{M}_1(\Omega)$ equipped with its weak topology) with rate function $I(\cdot) = H(\mu|\cdot)$.

Remarks: (1) There is no loss of generality in the assumption that the support of the prior, α, is Ω. Indeed, if the prior were supported on some smaller set Ω_1, then since the posterior assigns no mass outside Ω_1, we can confine ourselves to the closed set Ω_1, which is also a compact metric space. (2) Lynch and Sethuraman (1987) prove an LDP for the sequence of Dirichlet distributions, $\mathcal{D}(n\mu)$, on $[0, 1]$. Their result would be equivalent to our theorem, for $\Omega = [0, 1]$, if $\mathcal{D}(n\mu)$ and $\mathcal{D}(\alpha + n\mu_n)$ are exponentially equivalent whenever μ_n converges weakly to μ; however, establishing exponential equivalence does not appear to be trivial.

We now sketch the main ideas behind the proof before proceeding with a formal derivation. Let μ_n be a random element of $\mathcal{M}_1(\Omega)$ with distribution $\mathcal{D}(\alpha + \sum_{i=1}^{n} \delta_{x_i})$ as above. For bounded measurable functions $f : \Omega \to \mathbb{R}$, we define

$$\Lambda_n(f) = \log E \left[\exp \int_\Omega f d\mu_n \right]. \tag{1}$$

We show in Lemma 1 below that, for finite measurable partitions $(A_1, \ldots, A_k)$ of Ω, the vector $(\mu_n(A_1), \ldots, \mu_n(A_k))$ satisfies the LDP in $\mathbb{R}^k$. We then use *Varadhan's integral lemma* (Dembo and Zeitouni 1993, Theorem 4.3.1) to infer the existence of the limit

$$\Lambda(f) = \lim_{n\to\infty} \frac{1}{n} \Lambda(nf),$$

for simple functions $f = \sum_{i=1}^{k} c_i 1_{A_i}$; here 1_{A_i} denotes the indicator of A_i. This is extended in Lemma 3 to all bounded continuous functions on Ω, using the continuity of $\Lambda(\cdot)$. By Theorem 4.5.3 in Dembo and Zeitouni (1993), the existence of the limiting logarithmic moment generating function, Λ, implies the large deviation upper bound for the sequence $\{\mu_n\}$, for all compact subsets of $M_1(\Omega)$. Since Ω was assumed to be compact, $M_1(\Omega)$ is compact in the weak topology and so the upper bound holds for all closed sets. The rate function for this upper bound is the convex conjugate of Λ, which we identify to be $H(\mu|\cdot)$.

We use the LDP for $(\mu_n(A_1), \ldots, \mu_n(A_k))$ and the *contraction principle* (Dembo and Zeitouni 1993, Theorem 4.2.1) to get an LDP for $\sum_{i=1}^{k} c_i \mu_n(A_i)$, for arbitrary constants c_i. Thus, we obtain a large deviation lower bound

for sets of the form

$$U(\phi_k, x, \delta) = \{\nu \in M_1(\Omega) : \int_\Omega \phi_k d\nu \ \in \ (x-\delta, x+\delta)\},$$

with rate function $H_k(\mu|\cdot)$ given in Lemma 2. Here $x \in \mathbb{R}$ and $\delta > 0$ are arbitrary, and ϕ_k is any simple function $\phi_k = \sum_{i=1}^k c_i 1_{A_i}$, so that $\int \phi_k d\nu = \sum c_i \nu(A_i)$. We extend the lower bound to sets of the form $U(\phi, x, \delta)$ where ϕ is any bounded continuous function on Ω, by using increasingly fine partitions of Ω to approximate ϕ by simple functions. The rate function for this lower bound is the limit of $H_k(\mu|\cdot)$ as the partitions indexed by k get finer, which is shown in Lemma 2 to be $H(\mu|\cdot)$. Since the sets $U(\phi, x, \delta)$ constitute a base for the weak topology on $M_1(\Omega)$, this establishes the large deviation lower bound for all open sets.

The proof of Theorem 1 uses the following lemmas, whose proofs are in the appendix.

Lemma 1 *Let $(A_1, \ldots, A_k)$ be a measurable partition of Ω and suppose that the interior of A_i is non-empty for each $i = 1, \ldots, k$. Let f be bounded and measurable with respect to $\sigma(A_1, \ldots, A_k)$, the σ-algebra generated by the sets $A_1, \ldots, A_k$. Then,*

$$\Lambda(f) := \lim_{n\to\infty} \frac{1}{n} \Lambda_n(nf) \tag{2}$$

exists and is finite, and is given by

$$\Lambda(f) = \sup_{\nu \in \mathcal{M}_1(\Omega)} \left[\int_\Omega f d\nu - H(\mu|\nu) \right]. \tag{3}$$

Lemma 2 *Let $A^k = (A_1^k, \ldots, A_{n_k}^k), k \in \mathbb{N}$, be a sequence of partitions of Ω such that the corresponding σ-algebras, $\sigma(A^k)$, increase to $\mathcal{B}(\Omega)$, the Borel σ-algebra on Ω. Then, for all $\nu \in \mathcal{M}_1(\Omega)$, we have*

$$H(\mu|\nu) = \sup_k H_k(\mu|\nu) = \lim_{k\to\infty} H_k(\mu|\nu),$$

where $H_k(\mu|\nu) := \sum_{i=1}^{n_k} \mu(A_i^k) \log[\mu(A_i^k)/\nu(A_i^k)]$.

This result is well known, see Georgii (1988) for example.

Lemma 3 *For all bounded, continuous functions $f : \Omega \to \mathbb{R}$, the limit in (2) exists and is finite. The map $\Lambda : \mathcal{C}_b(\Omega) \to \mathbb{R}$ is convex and continuous, and we have*

$$\Lambda(f) = \sup_{\nu \in \mathcal{M}_1(\Omega)} \left[\int f d\nu - H(\mu|\nu) \right].$$

Here, $\mathcal{C}_b(\Omega)$ denotes the space of bounded continuous functions from Ω to $\mathbb{R}$, equipped with the supremum norm, $\|f\|_\infty = \sup_{x \in \Omega} |f(x)|$.

Proof of Theorem 1 : We have from Lemma 3 that Λ is the convex conjugate of $H(\mu|\cdot)$. But $H(\mu|\cdot)$ is convex, and lower semicontinuous in the weak topology (see Dupuis and Ellis 1997, Lemma 1.4.3), and recall that $\mathcal{M}_1(\Omega)$ is Polish since Ω is a Polish space). Hence, $H(\mu|\cdot)$ and $\Lambda(\cdot)$ are convex duals of each other. The large deviations upper bound for compact subsets of $\mathcal{M}_1(\Omega)$ now follows from Dembo and Zeitouni (1993), Theorem 4.5.3. But Ω was assumed to be compact, hence $\mathcal{M}_1(\Omega)$ is compact in the weak topology, so the upper bound holds for all closed sets in $\mathcal{M}_1(\Omega)$. We now turn to the proof of the large deviations lower bound.

The weak topology on $\mathcal{M}_1(\Omega)$ is generated by the sets

$$U_{\phi,x,\delta} = \left\{ \nu \in \mathcal{M}_1(\Omega) : \left| \int_\Omega \phi d\nu - x \right| < \delta \right\}, \quad \phi \in \mathcal{C}_b(\Omega), x \in \mathbb{R}, \delta > 0.$$

Given such a set and $\epsilon > 0$, we can find a sequence of measurable partitions $A^k = (A_1^k, \ldots, A_{n_k}^k)$ of Ω, and a sequence of simple functions ϕ_k measurable with respect to $\sigma(A^k)$, with the following properties: the σ-algebras $\sigma(A^k)$ increase to $\mathcal{B}(\Omega)$, the Borel σ-algebra on Ω; for all k and all $i \in \{1, \ldots, n_k\}$, A_i^k is a μ-continuity set with non-empty interior; for some $K > 0$ and all $k > K$, $\|\phi_k - \phi\|_\infty < \epsilon$. We shall assume that $\epsilon < \delta/3$. We now have

$$P(\mu_n \in U_{\phi,x,\delta}) \geq P\left(\left| \int_\Omega \phi_k d\mu_n - x \right| < \delta - \epsilon \right) \quad \forall\, k > K. \tag{4}$$

Let $\phi_k = \sum_{i=1}^{n_k} c_i^k 1_{A_i^k}$. Then,

$$\int_\Omega \phi_k d\mu_n = \sum_{i=1}^{n_k} c_i^k \mu_n(A_i^k).$$

It is shown in the proof of Lemma 1 (see equation (13)) that the sequence $(\mu_n(A_1^k), \ldots, \mu_n(A_{n_k}^k))_{n \geq 0}$ satisfies an LDP with rate function I_k given by

$$I_k(y_1, \ldots, y_{n_k}) = \begin{cases} \sum_{j=1}^{n_k} \mu(A_j) \log \dfrac{\mu(A_j)}{y_j}, & \text{if } y \in \mathbb{R}_+^{n_k} \text{ and } \sum_{i=1}^{n_k} y_i = 1, \\ +\infty, & \text{otherwise.} \end{cases}$$

It follows from the Contraction Principle (Dembo and Zeitouni 1993, Theorem 4.2.1) that $\sum_{i=1}^{n_k} c_i^k \mu_n(A_i^k)$ satisfies an LDP with rate function J_k given by

$$\begin{aligned} J_k(x) &= \inf\left\{ I_k(y) : \sum_{i=1}^{n_k} c_i^k y_i = x \right\} \\ &= \inf\left\{ H_k(\mu|\nu) : \nu \in \mathcal{M}_1(\Omega), \int_\Omega \phi_k d\nu = x \right\}. \end{aligned}$$

In particular, we obtain the large deviations lower bound,

$$\begin{aligned} & \liminf_{n\to\infty} \frac{1}{n} \log P\left(\left| \int_\Omega \phi_k d\mu_n - x \right| < \delta - \epsilon \right) \qquad (5) \\ \geq\ & -\inf\{ J_k(y) : |y - x| < \delta - \epsilon \} \\ =\ & -\inf\left\{ H_k(\mu|\nu) : \nu \in \mathcal{M}_1(\Omega), \left| \int_\Omega \phi_k d\nu - x \right| < \delta - \epsilon \right\}. \qquad (6) \end{aligned}$$

Now, $\|\phi - \phi_k\|_\infty < \epsilon$ for all $k > K$, so we have for all $\nu \in \mathcal{M}_1(\Omega)$ that,

$$\left| \int_\Omega \phi d\nu - x \right| < \delta - 2\epsilon \quad \Rightarrow \quad \left| \int_\Omega \phi_k d\nu - x \right| < \delta - \epsilon \quad \forall\ k > K.$$

It now follows from (4) and (6) that, for all $k > K$,

$$\liminf_{n\to\infty} \frac{1}{n} \log P(\mu_n \in U_{\phi,x,\delta}) \geq -\inf\left\{ H_k(\mu|\nu) : \left| \int_\Omega \phi d\nu - x \right| < \delta - 2\epsilon \right\}.$$

Hence, we have from Lemma 2 that

$$\liminf_{n\to\infty} \frac{1}{n} \log P(\mu_n \in U_{\phi,x,\delta}) \geq -\inf\left\{ H(\mu|\nu) : \left| \int_\Omega \phi d\nu - x \right| < \delta - 2\epsilon \right\}.$$

Since $\epsilon > 0$ was arbitrary, we can let ϵ decrease to zero, to get

$$\liminf_{n\to\infty} \frac{1}{n} \log P(\mu_n \in U_{\phi,x,\delta}) \geq -\inf\left\{ H(\mu|\nu) : \left| \int_\Omega \phi d\nu - x \right| < \delta \right\},$$

which is the desired large deviations lower bound for the set $U_{\phi,x,\delta}$, with rate function $H(\mu|\cdot)$. We have thus established the large deviations lower bound for a base of the weak topology on $\mathcal{M}_1(\Omega)$, and hence for all open sets in this topology. Combined with the upper bound above, this completes the proof of the theorem. ∎

We have established an LDP for the sequence of Dirichlet posterior distributions in the weak topology on $\mathcal{M}_1(\Omega)$, with rate function $I(\nu) = H(\mu|\nu)$.

The rate function differs from that in Sanov's theorem in that its argument, ν, enters as the second rather than the first argument in the relative entropy function. (Sanov's theorem says that the empirical distribution of a sequence of independent and identically distributed Ω-valued random variables with common law μ satisfies an LDP with rate function $J(\nu) = H(\nu|\mu)$). Intuitively, this is because, in Sanov's theorem we are asking how likely we are to observe ν, given that the true distribution is μ, whereas in this paper we are asking how likely it is that the true distribution is ν, given that we observe μ.

We believe that our result holds for a wider class of priors, of the form described below. Let $\mathcal{P}$ be the set of all finite measurable partitions of Ω. For $P \in \mathcal{P}$ let $\sigma(P)$ denote the σ-algebra generated by P. The restriction of a measure $\nu \in \mathcal{M}_1(\Omega)$ to the σ-algebra $\sigma(P)$ is denoted ν_P. In other words, $\nu_P = E[\nu|\sigma(P)]$. For a prior $\pi \in \mathcal{M}_1(\mathcal{M}_1(\Omega))$ we denote by π_P the corresponding element in $\mathcal{M}_1(\mathcal{M}_1(\Omega, \sigma(P)))$, thus the restriction of π to the Borel σ-algebra $\mathcal{B}(\mathcal{M}_1(\Omega, \sigma(P)))$. We fix a subset $\mathcal{P}'$ of $\mathcal{P}$ and say that a prior measure $\pi \in \mathcal{M}_1(\mathcal{M}_1(\Omega))$ is *exchangeable with respect to finite projections in* $\mathcal{P}'$ if for every $P \in \mathcal{P}'$ we have

$$[\pi^n(\mu_n)]_P = \pi_P^n(\mu_{n,P}).$$

Here $\pi^n(\mu_n)$ denotes the posterior distribution on $M_1(\Omega, \mathcal{B}(\Omega))$ corresponding to the prior P and the empirical distribution μ_n, $[\pi^n(\mu_n)]_P$ its restriction to $\sigma(P)$ and $\pi_P^n(\mu_{n,P})$ the posterior distribution on $M_1(\Omega, \sigma(P))$ corresponding to the prior π_P and the empirical distribution restricted to $\sigma(P)$.

The essential property of the Dirichlet process that we have used in the proof of Theorem 1 is its exchangeability with respect to $\mathcal{P}'$, where $\mathcal{P}'$ is the collection of finite partitions consisting of sets with non-empty interiors. This collection is large enough to generate the Borel σ-algebra on Ω. We believe that our methods can be generalized to priors which are exchangeable with respect to finite projections in $\mathcal{P}'$, for some $\mathcal{P}'$ which generates the Borel σ-algebra on Ω, although there do seem to be some technical difficulties which we hope to address in future work. An example of a class of priors which are exchangeable with respect to finite projections are the Polya tree distributions studied by Mauldin *et al.* (1992) and Lavine (1992), which generalize the Dirichlet process.

3 Application to the gambler's ruin problem

Suppose now that Ω is a compact subset of $\mathbb{R}$. As before, $\{X_k\}$ is a sequence of independent, identically distributed 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

$$\mathbf{P}_n(Q, \mu_n) = \int R(Q, \lambda)\pi^n(d\lambda),$$

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 corresponding to some prior, π, and the empirical distribution, μ_n. A standard refinement of Wald's approximation yields

$$Ce^{-\delta(\mu)Q} \leq R(Q, \mu) \leq e^{-\delta(\mu)Q},$$

for some $C > 0$, where

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

Thus,

$$C \int_{\mathcal{M}_1(\Omega)} e^{-\delta(\lambda)Q}\ \pi^n(d\lambda) \leq \mathbf{P}_n(Q, \mu_n) \leq \int_{\mathcal{M}_1(\Omega)} e^{-\delta(\lambda)Q}\ \pi^n(d\lambda).$$

Now, if π is the Dirichlet process, $\mathcal{D}(\alpha)$, parametrized by an arbitrary finite positive measure α whose support is all of Ω, then the sequence π^n obeys an LDP by Theorem 1, and we can apply Varadhan's lemma (see, for example, Dembo and Zeitouni (1993) Theorem 4.3.1) to obtain the asymptotic formula, for $q > 0$,

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

on the set $\mu_n \to \mu$. Here, we are using the easy (Ω is compact) 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.* (1998).

4 Conclusion

In this paper, we establish a large deviations principle for the sequence of Bayesian posteriors induced by a Dirichlet prior on a compact metric space, Ω. Can the result be extended to an arbitrary Polish space? Our approach yields the large deviation lower bound for arbitrary open subsets of this space, and the upper bound for compact subsets. In other words, we can prove a weak LDP on a Polish space. This could be strengthened to a full LDP if the sequence of Dirichlet posteriors were exponentially tight. However, exponential tightness of this sequence would imply the goodness of the rate function $H(\mu|\cdot)$, which we know not to be true in general. For example, take $\Omega = \mathbb{R}$, $\mu = \delta_0$, the unit mass at 0, and $\nu_n = (1/2)\delta_0 + (1/2)\delta_n$. Then $H(\mu|\nu_n) = \log 2$ for all n, but the sequence ν_n is not tight. This implies that $H(\mu|\cdot)$ doesn't have compact level sets, i.e., it is not a good rate function. Hence, our method cannot be easily extended to arbitrary Polish spaces. Finally, while we have worked with Dirichlet process priors, we believe that our approach can be extended to priors with the appropriate exchangeability properties, as discussed at the end of Section 2. However, there do appear to be some technical difficulties with this approach, which we hope to address in future work.

Acknowledgements: We would like to thank the referees for their comments, which have helped to improve the presentation of the paper.

References

[1] A. Barron, M. J. Schervish and L. Wasserman (1999). The consistency of posterior distributions in nonparametric problems, *Ann. Statist.*, 27(2) 536–561.

[2] A. Dembo and O. Zeitouni (1993). *Large Deviations Techniques and Applications*, Jones and Bartlett, Boston & London.

[3] P. Dupuis and R. S. Ellis (1997). *A Weak Convergence Approach to the Theory of Large Deviations*, John Wiley & Sons.

[4] T. S. Ferguson (1973). A Bayesian analysis of some non-parametric problems, *Ann. Statist.* 1, 209–230.

[5] T. S. Ferguson (1974). Prior distributions on spaces of probability measures, *Ann. Statist.* 2, 615–629.

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

[7] Ayalvadi Ganesh, Peter Green, Neil O'Connell and Susan Pitts (1998). Bayesian network management, *Queueing Systems* 28, 267–282.

[8] A. J. Ganesh and Neil O'Connell (1999). An inverse of Sanov's theorem, *Stat. and Prob. Letters* 42, 201–206.

[9] Hans-Otto Georgii (1988). *Gibbs Measures and Phase Transitions*, De Gruyter.

[10] S. Ghosal, J. K. Ghosh and R. V. Ramamoorthi (1999). Consistency issues in Bayesian nonparametrics, in *Asymptotics, Nonparametrics and Time Series*, Subir Ghosh ed., Marcel Dekker Inc., New York.

[11] S. Ghosal, J. K. Ghosh and A. W. van der Vaart (1998) *Convergence rates of posterior distributions*, Preprint.

[12] M. Lavine (1992). Some aspects of Polya tree distributions for statistical modelling, *Ann. Statist.* 20, 1222–1235.

[13] James Lynch and Jayaram Sethuraman (1987). Large deviations for processes with independent increments, *Ann. Probab.* 15, 610–627.

[14] R. D. Mauldin, W. D. Sudderth and S. C. Williams (1992). Polya trees and random distributions, *Ann. Statist.* 20, 1203–1221.

[15] R. T. Rockafellar (1974) *Conjugate Duality and Optimization*, Society for Industrial and Applied Mathematics.

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

[17] X. Shen and L. Wasserman (1998). *Rates of convergence of posterior distributions*, Technical report, Dept. of Statistics, Carnegie-Mellon University.

[18] L. Wasserman (1998). *Asymptotic properties of nonparametric Bayesian procedures*, Technical report, Dept. of Statistics, Carnegie-Mellon University.

A Proofs

Proof of Lemma 1: Let $(A_1, \ldots, A_k)$ be a partition of Ω such that the interior of A_i is non-empty and that A_i is a μ-continuity set for every $i = 1 \ldots, k$. Let f be bounded and measurable with respect to the σ-algebra generated by the partition. Then we can write

$$f = \sum_{i=1}^{k} c_i 1_{A_i}, \tag{7}$$

for some constants c_i, where 1_{A_i} denotes the indicator of A_i. Then, by (1),

$$\Lambda_n(f) = \log E\Big[\exp \sum_{i=1}^{k} c_i \mu_n(A_i)\Big]. \tag{8}$$

By the assumption that each A_i has non-empty interior and that the support of α is Ω, we have

$$\alpha_n(A_j) := \alpha(A_j) + \sum_{i=1}^{n} \delta_{x_i}(A_j) > 0 \quad \forall\, n \in \mathbb{N} \text{ and } j = 1, \ldots, k; \tag{9}$$

We have from the definition of the Dirichlet distribution that

$$(\mu_n(A_1), \ldots, \mu_n(A_k)) \sim \left(\frac{Z_n^1}{\sum_{i=1}^k Z_n^i}, \ldots, \frac{Z_n^k}{\sum_{i=1}^k Z_n^i} \right),$$

where the Z_n^i are independent gamma random variables, with

$$Z_n^j \sim \mathcal{G}(\alpha_n(A_j), 1),$$

and α_n is defined in (9). Here, $\mathcal{G}(\alpha, 1)$ denotes the gamma distribution with shape parameter α and scale parameter 1. It is straightforward to evaluate the cumulant generating functions of the Z_n^j. We have

$$\lambda_n^j(\theta) := \log E[\exp(\theta Z_n^j)] = \begin{cases} -\alpha_n(A_j) \log(1-\theta), & \text{if } \theta < 1, \\ +\infty, & \text{otherwise.} \end{cases}$$

Since $\sum_{i=1}^n \delta_{x_i}(A_j)/n \to \mu(A_j)$ by assumption, we get

$$\lambda_j(\theta) := \lim_{n\to\infty} \frac{1}{n} \lambda_n^j(\theta) = \begin{cases} -\mu(A_j) \log(1-\theta), & \text{if } \theta < 1, \\ +\infty, & \text{otherwise.} \end{cases}$$

Hence, by the Gärtner-Ellis theorem (see Dembo and Zeitouni (1993) Theorem 2.3.6), the sequence of random variables Z_n^j/n satisfies an LDP in $\mathbb{R}$ with rate function λ_j^* which is the convex dual of λ_j, i.e.,

$$\lambda_j^*(x) = \sup_{\theta \in \mathbb{R}} [\theta x - \lambda_j(\theta)] = \begin{cases} x - \mu(A_j) + \mu(A_j) \log \dfrac{\mu(A_j)}{x}, & \text{if } x > 0, \\ +\infty, & \text{else.} \end{cases} \tag{10}$$

If $\mu(A_j) = 0$, then the assumption of steepness of λ_j is not satisfied, so the Gärtner-Ellis theorem doesn't apply. However, it is not hard to verify directly in this case that Z_n^j/n does indeed satisfy an LDP with the above rate function.

Since $\{Z_n^j, j = 1, \ldots, k\}$ are independent, $\{Z_n^j/n, j = 1, \ldots, k\}$ jointly satisfy an LDP in $\mathbb{R}^k$ with rate function $\lambda^*(x) = \sum_{j=1}^k \lambda_j^*(x_j)$, where $x = (x_1, \ldots, x_k)$ and λ_j^* is given by (10).

Define $Y_n^j = Z_n^j / \sum_{i=1}^k Z_n^i$. Since $\sum_{i=1}^k Z_n^i$ is strictly positive with probability 1, the maps

$$(Z_n^1, \ldots, Z_n^k) \to (Y_n^1, \ldots, Y_n^k)$$

are almost surely continuous for every n. It follows from the Contraction Principle (Dembo and Zeitouni (1993) Theorem 4.2.1) that $\{Y_n^j, j = 1, \ldots, k\}$ jointly satisfy an LDP with rate function I given by

$$I(y_1, \ldots, y_k) = \inf \left\{ \sum_{j=1}^k \lambda_j^*(z_j) : y_j = \frac{z_j}{\sum_{i=1}^k z_i},\ j = 1, \ldots, k \right\}. \tag{11}$$

If $y_j < 0$ for some j, then any z included in the infimum in (11) must have $z_i < 0$ for some i and so, by (10), $I(y) = \infty$. Next, if $y_j = 0$ for all j or if $\sum_{i=1}^n y_i \neq 1$, then there does not exist $z \in \mathbb{R}^k$ such that $y_j = z_j / \sum_{i=1}^k z_i$ for all j. Hence $I(y)$, being the infimum of an empty set, is again $+\infty$.

In the following, we shall confine attention to $y \in \mathbb{R}^k$ such that $y \geq 0$ and $\sum_{i=1}^k y_i = 1$. If $z \in \mathbb{R}^k$ is such that $y_j = z_j / \sum_{i=1}^k z_i$ for all $j = 1, \ldots, k$, then we can write $z = \beta y$ for some $\beta > 0$. Now (11) gives

$$I(y_1, \ldots, y_k) = \inf_{\beta > 0} \sum_{j=1}^k \lambda_j^*(\beta y_j). \tag{12}$$

Setting the derivative of the sum on the right with respect to β equal to zero yields

$$0 = \sum_{j=1}^k \left(y_j - \frac{\mu(A_j)}{\beta} \right) = 1 - \frac{1}{\beta}.$$

To obtain the last equality, we have used the fact that $\sum_{j=1}^k y_j = 1$ by assumption, while $\sum_{j=1}^k \mu(A_j) = 1$ as μ is a probability distribution and $A_1, \ldots, A_k$ partition Ω. Since each λ_j^* is convex, the above implies that the infimum in (12) is achieved at $\beta = 1$, and

$$\begin{aligned} I(y) &= \sum_{j=1}^k \lambda_j^*(y_j) = \sum_{j=1}^k y_j - \mu(A_j) + \mu(A_j) \log \frac{\mu(A_j)}{y_j} \\ &= \sum_{j=1}^k \mu(A_j) \log \frac{\mu(A_j)}{y_j}. \end{aligned}$$

The second equality above comes from (10) and the third follows from the fact that μ and y are both probability distributions, hence sum to 1. It follows from the preceding discussion that the sequence of random vectors $(\mu_n(A_1), \ldots, \mu_n(A_k))$ satisfy an LDP in $\mathbb{R}^k$ with rate function

$$I(y) = \begin{cases} \sum_{j=1}^k \mu(A_j) \log \frac{\mu(A_j)}{y_j}, & \text{if } y \in \mathbb{R}_+^k \text{ and } \sum_{i=1}^k y_i = 1, \\ +\infty, & \text{otherwise.} \end{cases} \tag{13}$$

Observe from (7) that $|\int f d\mu_n| \leq \max_{i=1}^k |c_i|$ as μ_n is a probability distribution. Hence, we have from Varadhan's lemma (Dembo and Zeitouni 1993, Theorem 4.3.1) and the LDP for $(\mu_n(A_1), \ldots, \mu_n(A_k))$ that

$$\Lambda(f) := \lim_{n \to \infty} \frac{1}{n} \Lambda_n(nf) = \sup_{y \in \mathbb{R}^k} \left[\sum_{i=1}^k c_i y_i - I(y) \right].$$

Using (13), we can rewrite the above as

$$\begin{aligned} \Lambda(f) &= \sup_{\nu \in \mathcal{M}_1(\Omega)} \left[\sum_{i=1}^k c_i \nu(A_i) - \sum_{i=1}^k \mu(A_i) \log \frac{\mu(A_i)}{\nu(A_i)} \right] \\ &= \sup_{\nu \in \mathcal{M}_1(\Omega)} \left[\int_\Omega f d\nu - H_k(\mu|\nu) \right], \end{aligned} \tag{14}$$

where $H_k(\mu|\nu)$ is defined in Lemma 2. We now show that we can replace $H_k(\mu|\nu)$ in the supremum above by $H(\mu|\nu)$.

Let $\nu \in \mathcal{M}_1(\Omega)$ be arbitrary. If $\mu(A_i) > 0$ and $\nu(A_i) = 0$ for some A_i, then $\mu \not\ll \nu$ and $H(\mu|\nu)$ and $H_k(\mu|\nu)$ are both infinite. Hence, such ν can be excluded from consideration of the supremum above, and we shall suppose

without loss of generality that $\mu(A_i) = 0$ whenever $\nu(A_i) = 0$. We now define $\lambda \in \mathcal{M}_1(\Omega)$ as follows. Set $\lambda \equiv \nu$ on A_i if $\mu(A_i) = 0$; if $\mu(A_i) > 0$, take λ to be absolutely continuous with respect to μ on A_i, with Radon-Nikodym derivative

$$\frac{d\lambda}{d\mu} \equiv \frac{\nu(A_i)}{\mu(A_i)} > 0.$$

Then μ is absolutely continuous with respect to λ and we have

$$\begin{aligned} H(\mu|\lambda) &= \int_\Omega d\mu \log \frac{d\mu}{d\lambda} = \sum_{i:\mu(A_i)>0} \int_{A_i} d\mu \log \frac{d\mu}{d\lambda} \\ &= \sum_{i:\mu(A_i)>0} \mu(A_i) \log \frac{\mu(A_i)}{\nu(A_i)} = H_k(\mu|\nu). \end{aligned} \tag{15}$$

Also,

$$\int_\Omega f d\nu = \sum_{i=1}^k c_i \nu(A_i) = \sum_{i=1}^k c_i \lambda(A_i) = \int_\Omega f d\lambda. \tag{16}$$

Since $\nu \in \mathcal{M}_1(\Omega)$ was arbitrary, we obtain from (15) and (16) that

$$\sup_{\nu \in \mathcal{M}_1(\Omega)} \left[\int_\Omega f d\nu - H_k(\mu|\nu) \right] \leq \sup_{\lambda \in \mathcal{M}_1(\Omega)} \left[\int_\Omega f d\lambda - H(\mu|\lambda) \right].$$

The reverse inequality holds as well because $H_k(\mu|\nu) \leq H(\mu|\nu)$ for all $\nu \in \mathcal{M}_1(\Omega)$ by Lemma 2 (or by the convexity of $x \mapsto x \log x$ on $[0, \infty)$). Hence, equality holds above and the claim of the lemma follows from (14). ∎

Proof of Lemma 3: Let $\epsilon > 0$ be given, and let $f : \Omega \to \mathbb{R}$ be bounded and continuous. We can find $k > 0$ and a simple function $g = \sum_{i=1}^k c_i 1_{A_i}$ such that $\|f - g\|_\infty < \epsilon$. Since f is continuous, we can in fact choose the A_i to be μ-continuity sets with non-empty interiors. Now, by (1) and the fact that each μ_n is a probability distribution,

$$\begin{aligned} \Lambda_n(nf) &= \log E\left[\exp \int_\Omega nf d\mu_n\right] \leq \log E\left[\exp\Big(\int_\Omega ng d\mu_n + n\epsilon\Big)\right] \\ &= n\epsilon + \Lambda_n(ng), \end{aligned}$$

so that, by (2),

$$\limsup_{n\to\infty} \frac{1}{n} \Lambda_n(nf) \leq \Lambda(g) + \epsilon.$$

Likewise, $\liminf_{n\to\infty} \Lambda_n(nf)/n \geq \Lambda(g) - \epsilon$. Since $\epsilon > 0$ is arbitrary, it follows that

$$\Lambda(f) := \lim_{n\to\infty} \frac{1}{n} \Lambda_n(nf)$$

exists and is finite for all bounded, continuous $f : \Omega \to \mathbb{R}$. The arguments above also show that $\Lambda : \mathcal{C}_b(\Omega) \to \mathbb{R}$ is continuous, with $|\Lambda(f) - \Lambda(g)| \leq \|f - g\|_\infty$.

For $f \in L^\infty(\Omega)$, define

$$H^*(f) = \sup_{\nu \in \mathcal{M}_1(\Omega)} \left[\int_\Omega f d\nu - H(\mu|\nu) \right], \tag{17}$$

i.e., H^* is the convex conjugate of $H(\mu|\cdot)$. Now $|\int f d\nu| \leq \|f\|_\infty$ for all $\nu \in \mathcal{M}_1(\Omega)$, while $H(\mu|\cdot)$ is non-negative, with $H(\mu|\mu) = 0$. Thus, $|H^*(f)| \leq \|f\|_\infty$. Since H^* is a convex function with domain $L^\infty(\Omega)$, which is bounded on the open neighbourhood, $\{f : \|f\|_\infty < 1\}$, we have by (Rockafellar 1974, Theorem 8) that H^* is continuous on the interior of its domain, which is all of $L^\infty(\Omega)$.

By Lemma 1, H^* and Λ agree on functions of the form $f = \sum_{i=1}^k c_i 1_{A_i}$, where the A_i partition Ω and each A_i is a μ-continuity set with non-empty interior. Since such functions are dense in $\mathcal{C}_b(\Omega)$, Λ was shown to be continuous on $\mathcal{C}_b(\Omega)$ and H^* to be continuous on $L^\infty(\Omega) \supseteq \mathcal{C}_b(\Omega)$, it follows that $\Lambda = H^*$ on all of $\mathcal{C}_b(\Omega)$ and, consequently, that Λ is convex. ■