A Large Deviations Principle for Dirichlet Process Posteriors

A.J. Ganesh, Neil O'Connell
Basic Research Institute in Mathematical Sciences
HP Laboratories Bristol

HPL-BRIMS-98-05
February, 1998

large deviations,
Bayesian statistics,
Dirichlet process

Let X_k be a sequence of iid 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 are 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.

A LARGE DEVIATIONS PRINCIPLE FOR DIRICHLET PROCESS POSTERIORS

A. J. 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 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 are 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.

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 $\mu \in \mathcal{M}_1(\Omega)$. 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 $\mathcal{M}_1(\Omega)$ equipped with the τ-topology (see Dembo and Zeitouni [1, 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 [8], 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: it was shown by Freedman [6] 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. In this paper, we prove an LDP for the special (but nevertheless, useful) case of Dirichlet process priors on a compact metric space, Ω, equipped with the weak topology. Our techniques should easily generalize to a number of other popular choices of prior. If we assume only that the sequence of empirical measures L_n converges weakly to μ, then we can show that the LDP does not necessarily hold in the τ-topology on $\mathcal{M}_1(\Omega)$. The problem of extending our results to an arbitrary Polish space, Ω, remains open.

The LDP for Dirichlet posteriors derived here has applications to queue and risk management that are discussed in Ganesh et al. [7]. Some questions of interest in this context are posed in terms of the ruin probability in the classical gambler's ruin problem. The LDP for the posterior distributions can be used to obtain an asymptotic formula for the predictive probability of ruin, see [7, 8] for details.

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, which are introduced below. The discussion follows Ferguson [4].

Let $n > 0$ and $a = (a_1, \dots, a_n)$ be given. Suppose Z_i, $i = 1, \dots, n$ are independent real-valued random variables, with $Z_i \sim \mathcal{G}(a_i, 1)$, where $\mathcal{G}(a_i, 1)$ denotes the gamma distribution with shape parameter a_i and scale parameter 1, and $\sim$ denotes equality in distribution (if $a_i = 0$, we take $Z_i \equiv 0$). Let $Z = Z_1 + \ldots + Z_n$. The n-dimensional *Dirichlet distribution* with parameter $a = (a_1, \dots, a_n)$, denoted $D(a)$, is defined to be the joint distribution of $(Y_1, \dots, Y_n) = (Z_1/Z, \dots, Z_n/Z)$. This is a probability distribution on the n-simplex,

$$S^n = \{\mathbf{x} \in \mathbb{R}^n : x_i \geq 0, i = 1, \dots, n, \sum_{i=1}^{n} x_i = 1\},$$

and if all the a_i are strictly positive, it can be expressed by the density

$$f(x_1, \dots, x_{n-1}) = \frac{\Gamma(\sum_{i=1}^n a_i)}{\prod_{i=1}^n \Gamma(a_i)} \prod_{i=1}^{n-1} x_i^{a_i - 1} (1 - \sum_{i=1}^{n-1} x_i)^{a_n - 1}. \qquad (1)$$

Here $\Gamma(\cdot)$ denotes the gamma function: $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$, $z > 0$. Some interesting and useful properties of the Dirichlet distribution are stated below.

1. If $(Y_1, \dots, Y_k) \sim D(a_1, \dots, a_k)$ and $r_1, \dots, r_l$ are integers such that $0 < r_1 < \ldots < r_l = k$, then

$$\left(\sum_{1}^{r_1} Y_i, \sum_{r_1+1}^{r_2} Y_i, \dots, \sum_{r_{l-1}+1}^{r_l} Y_i \right) \sim D \left(\sum_{1}^{r_1} a_i, \sum_{r_1+1}^{r_2} a_i, \dots, \sum_{r_{l-1}+1}^{r_l} a_i \right).$$

This follows directly from the definition of the Dirichlet distribution and the additive property of the gamma distribution: if $Z_1 \sim \mathcal{G}(\alpha_1, 1)$ and $Z_2 \sim \mathcal{G}(\alpha_2, 1)$, and Z_1 and Z_2 are independent, then $Z_1 + Z_2 \sim \mathcal{G}(\alpha_1 + \alpha_2, 1)$.

2. If the prior distribution of $(Y_1, \ldots, Y_k)$ is $D(a_1, \ldots, a_k)$, and if

$$\mathbf{P}(X = j | Y_1, \ldots, Y_k) = Y_j, \quad j = 1, \ldots, k,$$

   then the posterior distribution of $(Y_1, \ldots, Y_k)$ given $X = j$ is the Dirichlet distribution $D(a_1^j, \ldots, a_k^j)$, where

$$a_i^j = \begin{cases} a_i, & \text{if } i \neq j \\ a_j + 1, & \text{if } i = j. \end{cases}$$

Recall that if $(Y_1, \ldots, Y_k)$ have a Dirichlet distribution, then each Y_i is non-negative and $\sum_{i=1}^k Y_i = 1$. So the Y_i can be interpreted as the parameters of a multinomial distribution on $\{1, \ldots, k\}$. Hence, Property 2 above says that the Dirichlet distribution is a conjugate prior for the parameters of a multinomial distribution: if the prior is a Dirichlet distribution, then so is the posterior.

Denote by $\mathcal{M}_+(\Omega)$ (respectively $\mathcal{M}_1(\Omega)$) the space of finite non-negative (respectively probability) measures on an arbitrary measure space $(\Omega, \mathcal{F})$. The "*Dirichlet process*" with parameter $\alpha \in \mathcal{M}_+(\Omega)$, which we denote by $\mathcal{D}(\alpha)$, is a probability distribution on $\mathcal{M}_1(\Omega)$, and is characterized as follows. A random probability measure, μ, on Ω has law $\mathcal{D}(\alpha)$ if, and only if, for each 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 (see [4] for details). A natural consistency criterion suggested by the above definition is the following: if $(A'_1, \ldots, A'_l)$ and $(A_1, \ldots, A_k)$ are measurable partitions of Ω, and if $(A'_1, \ldots, A'_l)$ is a refinement of $(A_1, \ldots, A_k)$ with $A_1 = \cup_1^{r_1} A'_i$, $A_2 = \cup_{r_1+1}^{r_2} A'_i$, $\cdots$, $A_k = \cup_{r_{k-1}+1}^{l} A'_i$, then

$$\left(\sum_1^{r_1} \mu(A'_i), \sum_{r_1+1}^{r_2} \mu(A'_i), \ldots, \sum_{r_{k-1}+1}^{l} \mu(A'_i) \right) \sim (\mu(A_1), \mu(A_2), \ldots, \mu(a_k)) .$$

It is clear from Property 1 of the Dirichlet distribution stated above that this consistency criterion is satisfied. It turns out that this condition is sufficient for the validity of the Kolmogorov consistency conditions for the finite-dimensional distributions defined above, and hence for the existence of the Dirichlet process: see Ferguson [4] for a proof and a more detailed discussion.

Suppose $\mathcal{P}$ is a prior distribution on the space $\mathcal{M}_1(\Omega)$ and assume $\mathcal{P}$ is a Dirichlet process with parameter α, denoted $\mathcal{D}(\alpha)$. Then, conditional on observing $\omega_1,\ldots,\omega_n$, it can be shown (see [4, 5]) that the posterior distribution is also a Dirichlet process, but with parameter $\alpha + \sum_{i=1}^n \delta_{\omega_i}$, where δ_x denotes the Dirac measure at x. 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*

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

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 iid 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 $I(\cdot) = H(\mu|\cdot)$.

Remark: Note that 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, we can confine ourselves to Ω_1. Since Ω_1 is closed (by definition of the support of a measure), it is compact and the requirements in the statement of the theorem are met.

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

$$\Lambda_n(f) = \log E\left[\exp \int_\Omega f d\mu_n\right], \tag{2}$$

We shall need the following lemmas, whose proofs are in the appendix.

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

$$\Lambda(f) := \lim_{n\to\infty} \frac{1}{n} \Lambda_n(nf) \tag{3}$$

exists and is finite, and is given by

$$\Lambda(f) = \sup_{\nu \in \mathcal{M}_1(\Omega)} \int_\Omega f d\nu - H_k(\mu|\nu), \tag{4}$$

where

$$H_k(\mu|\nu) := \sum_{i=1}^{k} \mu(A_i) \log \frac{\mu(A_i)}{\nu(A_i)}. \tag{5}$$

If f is $\sigma(A_1, \dots, A_k)$-measurable and the partition $(B_1, \dots, B_r)$ is a refinement of $(A_1, \dots, A_k)$, then f is also $\sigma(B_1, \dots, B_r)$-measurable, so it may appear at first glance that the right hand side of (4) is not well-defined. We show in Lemma 4 that (4) can, in fact, be rewritten as $\Lambda(f) = \int f d\nu - H(\mu|\nu)$, so that there is no ambiguity in the definition of $\Lambda(f)$ for simple functions.

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

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

Lemma 3 *Let $A^k = (A_1^k, \dots, A_{n_k}^k), k \in \mathbb{N}$, 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*

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

Lemma 4 *For all $f \in \mathcal{C}_b(\Omega)$, we have*

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

Proof of Theorem 1 : We have from Lemma 4 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 [2, Lemma 1.4.3], and recall that $\mathcal{M}_1(\Omega)$ is a Polish space since Ω is Polish). 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 [1, 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 has 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

$$P(\mu_n \in U_{\phi,x,\delta}) \geq P\left(\left| \int_\Omega \phi_k d\mu_n - x \right| < \delta - \epsilon \right) \quad \forall\, k > K. \qquad (6)$$

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 (15)) 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 (see [1, 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,

$$\begin{aligned}
&\liminf_{n\to\infty}\frac{1}{n}\log P\left(\left|\int_\Omega \phi_k d\mu_n - x\right| < \delta - \epsilon\right) &(7)\\
\geq\ & -\inf\{J_k(y) : |y-x| < \delta - \epsilon\}\\
=\ & -\inf\left\{H_k(\mu|\nu) : \nu \in \mathcal{M}_1(\Omega), \left|\int_\Omega \phi_k d\nu - x\right| < \delta - \epsilon\right\}. &(8)
\end{aligned}$$

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

$$\left|\int_\Omega \phi d\nu - x\right| < \delta - 2\epsilon \quad \Leftrightarrow \quad \left|\int_\Omega \phi_k d\nu - x\right| < \delta - \epsilon.$$

It now follows from (6) and (8) 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 3 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 Bayesian 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 iid 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.

3 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, the extension of our approach to other commonly used classes of priors does not appear to be difficult.

References

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

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

[3] R. Durrett, *Probability: Theory and Examples*, Wadsworth & Brooks/Cole, 1991.

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

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

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

[7] Ayalvadi Ganesh, Peter Green, Neil O'Connell and Susan Pitts, Bayesian network management. To appear in *Queueing Systems*, 1998.

[8] A. J. Ganesh and Neil O'Connell, An inverse of Sanov's theorem, HP Technical Report HPL-BRIMS-97-25.

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

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 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

$$f = \sum_{i=1} c_i 1_{A_i}, \tag{9}$$

for some constants c_i, where 1_{A_i} denotes the indicator of A_i. Then, by (2),

$$\Lambda_n(f) = \log E\Big[\exp \sum_{i=1}^{k} c_i \mu_n(A_i)\Big]. \tag{10}$$

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

$$\alpha_n(A_j) := \alpha(A_j) + \sum_{i=1}^{n} \delta_{x_i}(A_j) > 0 \quad \forall\, n \in \mathbb{N} \text{ and } j = 1, \ldots, k; \tag{11}$$

It follows 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 (11). 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 [1]), 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.,

$$\lambda_j^*(x) = \sup_{\theta\in\mathbb{R}}[\theta x - \lambda_j(\theta)] = \begin{cases} x - \mu(A_j) + \mu(A_j)\log\frac{\mu(A_j)}{x}, & \text{if } x > 0, \\ +\infty, & \text{else.} \end{cases} \tag{12}$$

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 $\{Z_n^j, j = 1,\dots,k\}$ are independent, $\{Z_n^j/n, j = 1,\dots,k\}$ jointly satisfy an LDP in $\mathbb{R}^k$ with rate function $\lambda^*(x) = \sum_{j=1}^{k}\lambda_j^*(x_j)$, where $x = (x_1,\dots,x_k)$ and λ_j^* is given by (12).

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

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

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

$$I(y_1,\dots,y_k) = \inf\left\{\sum_{j=1}^{k}\lambda_j^*(z_j) : y_j = \frac{z_j}{\sum_{i=1}^{k} z_i}, j = 1,\dots,k\right\}. \tag{13}$$

If $y_j < 0$ for some j, then any z included in the infimum in (13) must have $z_i < 0$ for some i and so, by (12), $I(y) = \infty$. 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 $j = 1,\dots,k$, then we can write $z = \beta y$ for some $\beta > 0$. Now (13) gives

$$I(y_1,\dots,y_k) = \inf_{\beta>0}\sum_{j=1}^{k}\lambda_j^*(\beta y_j). \tag{14}$$

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 (14) 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 (12) 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

$$I(y) = \begin{cases} \sum_{j=1}^{k} \mu(A_j) \log \frac{\mu(A_j)}{y_j}, & \text{if } y \in \mathbb{R}_+^k \text{ and } \sum_{i=1}^{k} y_i = 1, \\ +\infty, & \text{otherwise.} \end{cases} \tag{15}$$

Observe from (9) 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 [1, Theorem 4.3.1] and the LDP for $(\mu_n(A_1), \ldots, \mu_n(A_k))$ that

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

Using (15), we can rewrite the above as

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

where

$$H_k(\mu|\nu) = \sum_{i=1}^{k} \mu(A_i) \log \frac{\mu(A_i)}{\nu(A_i)}.$$

This completes the proof of the lemma.

Proof of Lemma 2: 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 have non-empty interiors. Now, by (2) and the fact that each μ_n is a probability distribution,

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

so that, by (3),

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

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 $|\Lambda(f) - \Lambda(g)| \leq \|f - g\|_\infty$. The convexity of Λ follows from Hölder's inequality since, for each $n \in \mathbb{N}$, and for all $f, g \in \mathcal{C}_b(\Omega)$ and $\alpha \in [0, 1]$, we have

$$\begin{aligned}&\Lambda_n(n[\alpha f + (1-\alpha)g]) = \log E\left[\Big(\exp\int_\Omega nfd\mu_n\Big)^\alpha\Big(\exp\int_\Omega ngd\mu_n\Big)^{1-\alpha}\right] \\ &\leq \log\left\{E\Big[\exp\int_\Omega nfd\mu_n\Big]^\alpha E\Big[\exp\int_\Omega ngd\mu_n\Big]^{1-\alpha}\right\} \\ &= \alpha\Lambda_n(nf) + (1-\alpha)\Lambda_n(ng).\end{aligned}$$

Proof of Lemma 3: Suppose first that μ is absolutely continuous with respect to ν, and denote by $d\mu/d\nu$ the Radon–Nikodym derivative of μ with respect to ν. We now have, for all $k \in \mathbb{N}$, that

$$\begin{aligned}H(\mu|\nu) &= \int_\Omega \frac{d\mu}{d\nu}\log\frac{d\mu}{d\nu}d\nu = \sum_{i=1}^{n_k}\nu(A_i^k)\int_{A_i^k}\frac{d\mu}{d\nu}\log\frac{d\mu}{d\nu}\frac{d\nu}{\nu(A_i^k)} \\ &\geq \sum_{i=1}^{n_k}\mu(A_i^k)\log\frac{\mu(A_i^k)}{\nu(A_i^k)} = H_k(\mu|\nu). \qquad (16)\end{aligned}$$

The inequality above follows from the positivity of the measures μ and ν, the convexity of $x \log x$ on $[0, \infty)$, and Jensen's inequality. Since the above holds for all partitions A^k of Ω, we get

$$H(\mu|\nu) \geq \sup_k \; H_k(\mu|\nu). \tag{17}$$

Let $\{A^k, k \in \mathbb{N}\}$, be a sequence of partitions of Ω such that $\sigma(A^k) \uparrow \mathcal{B}(\Omega)$. Here, $\sigma(A^k)$ denotes the σ-algebra generated by $(A_1^k, \ldots, A_{n_k}^k)$ and $\mathcal{B}(\Omega)$ is the Borel σ-algebra on Ω. Define the sequence of functions $g_k : \Omega \to \mathbb{R}$ by

$$g_k(x) = \frac{\mu(A_i^k)}{\nu(A_i^k)}, \quad x \in A_i^k,\ i = 1, \ldots, n_k.$$

Then, by Theorem 3.4, Chapter 4 of Durrett [3],

$$g_k \to \frac{d\mu}{d\nu} \quad \nu - \text{a.s. as } k \to \infty,$$

and so, by continuity of $x \to x \log x$ for $x \geq 0$, $g_k \log g_k \to (d\mu/d\nu)\log(d\mu/d\nu)$, ν almost surely. Hence, it follows from Fatou's lemma that

$$\begin{aligned} H(\mu|\nu) &= \int_\Omega \frac{d\mu}{d\nu} \log \frac{d\mu}{d\nu} d\nu \;\leq\; \liminf_{k\to\infty} \int_\Omega g_k \log g_k d\nu \\ &= \liminf_{k\to\infty} \sum_{i=1}^{n_k} \mu(A_i^k) \log \frac{\mu(A_i^k)}{\nu(A_i^k)} \;=\; \liminf_{k\to\infty} \; H_k(\mu|\nu). \end{aligned} \tag{18}$$

We get from (17) and (18) that

$$H(\mu|\nu) = \lim_{k\to\infty} H_k(\mu|\nu) = \sup_k \; H_k(\mu|\nu). \tag{19}$$

Suppose next that μ is not absolutely continuous with respect to ν. Then $H(\mu|\nu) = \infty$, so (17) holds trivially. Moreover, there exists a Borel measurable set B such that $\mu(B) > 0$ and $\nu(B) = 0$. Since $\sigma(A^k) \uparrow \mathcal{B}(\Omega)$, we can write $B = \bigcap_{k\in\mathbb{N}} B_k$, where $B_k \in \sigma(A^k)$, the σ-algebra generated by the partition A^k. Re-ordering the sets in A^k if necessary, we can write

$$B_k = \bigcup_{i=1}^{m_k} A_i^k, \quad k \in \mathbb{N}.$$

We have from Jensen's inequality and the convexity of $x \log x$ on $[0, \infty)$, that

$$\begin{aligned} \sum_{i=1}^{m_k} \mu(A_i^k) \log \frac{\mu(A_i^k)}{\nu(A_i^k)} &= \nu(B_k) \sum_{i=1}^{m_k} \left[\frac{\mu(A_i^k)}{\nu(A_i^k)} \log \frac{\mu(A_i^k)}{\nu(A_i^k)} \right] \frac{\nu(A_i^k)}{\nu(B_k)} \\ &\geq \mu(B_k) \log \frac{\mu(B_k)}{\nu(B_k)}. \end{aligned}$$

Henceforth, we shall assume without loss of generality that the partitions A^k have been chosen such that A^{k+1} is a refinement of A^k for every k; then the sets B_k can be chosen such that $B_{k+1} \subseteq B_k$ for all k. Then $B_k \downarrow B$, so $\nu(B_k) \downarrow \nu(B) = 0$, whereas $\mu(B_k) \geq \mu(B) > 0$ for all $k \in \mathbb{R}$. It follows from the above that

$$\lim_{k\to\infty} \sum_{i=1}^{m_k} \mu(A_i^k) \log \frac{\mu(A_i^k)}{\nu(A_i^k)} \geq \lim_{k\to\infty} \mu(B_k) \log \frac{\mu(B_k)}{\nu(B_k)} = \infty.$$

But

$$H_k(\mu|\nu) = \sum_{i=1}^{m_k} \mu(A_i^k) \log \frac{\mu(A_i^k)}{\nu(A_i^k)} + \sum_{i=m_k+1}^{n_k} \mu(A_i^k) \log \frac{\mu(A_i^k)}{\nu(A_i^k)}.$$

The first sum above goes to infinity as $k \to \infty$, whereas it is easy to verify, using the inequality $\log x \geq 1 - (1/x)$ for $x \geq 0$ and the fact that ν is a probability distribution, that the second sum is bounded below by -1. Therefore,

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

i.e, (19) holds even for μ not absolutely continuous with respect to ν. This completes the proof of the lemma.

Proof of Lemma 4: We begin by establishing the claim of the lemma for simple functions of the form (9), where the A_i have non-empty interiors and partition Ω. Let $\nu \in \mathcal{M}_1(\Omega)$ be such that $\mu \ll \nu$. Define $\lambda \in \mathcal{M}_1(\Omega)$ by setting $\lambda \equiv \nu$ on A_i if $\mu(A_i) = 0$; if $\mu(A_i) > 0$, define 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

$$\begin{aligned} H(\mu|\lambda) &= \int_\Omega d\mu \log \frac{d\mu}{d\lambda} = \sum_{i:\mu(A_i)>0} \int_{A_i} d\mu \log \frac{d\mu}{d\lambda} \\ &= \sum_{i:\mu(A_i)>0} \mu(A_i) \log \frac{\mu(A_i)}{\nu(A_i)} = H_k(\mu|\nu). \end{aligned} \tag{20}$$

Also,

$$\int_\Omega f d\nu = \sum_{i=1}^{k} c_i \nu(A_i) = \sum_{i=1}^{k} c_i \lambda(A_i) = \int_\Omega f d\lambda. \tag{21}$$

Since $\nu \in \mathcal{M}_1(\Omega)$ was arbitrary, we obtain from (20) and (21) that

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

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 3. Hence, equality holds above and, by (4), this establishes the claim of the lemma for simple functions of the form we consider.

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

$$H^*(f) = \sup_{\nu \in \mathcal{M}_1(\Omega)} \int_\Omega f d\nu - H(\mu|\nu), \tag{22}$$

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 $H(\mu|\mu) = 0$. It follows that $H^*(f)$ is finite for all $f \in L^\infty(\Omega)$; in fact, $|H^*(f)| \leq \|f\|_\infty$. Thus, H^* is a convex function with domain $L^\infty(\Omega)$, which is bounded on the open neighbourhood, $\{f : \|f\|_\infty < 1\}$. Hence, by [9, Theorem 8], H^* is continuous on the interior of its domain, which is all of $L^\infty(\Omega)$.

We have established the lemma for simple functions of the form in (9). Hence, 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 has non-empty interior. Since such functions are dense in $C_b(\Omega)$, Λ is continuous on $C_b(\Omega)$ by Lemma 2 and H^* was shown to be continuous on $L^\infty(\Omega) \supseteq C_b(\Omega)$, it follows that $\Lambda = H^*$ on all of $C_b(\Omega)$. The claim of the lemma is now immediate from the definition of H^*.