# Moderate deviations for Bayes posteriors

Peter Eichelsbacher
Ruhr-Universität Bochum

Ayalvadi Ganesh
Microsoft Research, UK

**Abstract**

Let $(X_k)_{k\in\mathbb{N}}$ be a sequence of i.i.d. random variables taking values in a set $\Omega$, and consider the problem of estimating the law of $X_1$ in a Bayesian framework. We prove, under mild conditions on the prior, that the sequence of posterior distributions satisfies a moderate deviation principle.

**Keywords**: Bayesian statistics, asymptotics, large deviations, moderate deviations, Polya tree, Dirichlet process.

## 1 Introduction

Bayesian methods have become increasingly popular in statistics in recent years and there has been renewed interest in nonparametric Bayesian inference. In turn, this has stimulated much recent work on the consistency of these inference procedures. Freedman (1963) and Diaconis and Freedman (1986) showed that even if the prior puts positive mass in every weak neighbourhood of the true distribution, it does not follow that the posterior mass of each weak neighbourhood tends to 1 (in fact, it can tend to zero!). Under the stronger condition that the prior puts positive mass in each Kullback-Leibler neighbourhood of the true distribution, Schwartz (1965) showed that asymptotically the posterior does concentrate on weak neighbourhoods of this distribution. If, in addition, the relevant space of probability distributions satisfies a 'metric entropy' condition, then Barron, Schervish, and Wasserman (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 $\mu$ is defined by $\int(\sqrt{f}-\sqrt{g})^2 d\mu$. As a distance between probability distribution, it does not depend on the choice of the reference measure.) Recent research on the consistency of Bayes methods is reviewed by Ghosal, Ghosh, and Ramamoorthi (1998) and Wasserman (1999).

There has been comparatively little work on more refined asymptotics for Bayes posteriors. For smooth parametric families Johnson (1967), (1970) obtains asymptotic expansions of the posterior distributions while Fu and Kass (1988) show that the posterior concentrates at an exponential rate. Rates of convergence in the nonparametric case have been investigated by Ghosal, Ghosh, and van der Vaart (2000) and Shen and Wasserman (1998). Results on asymptotic normality of the posterior are few and mixed. Le Cam (1973) and Ibragimov and Has'minskii (1981) prove central limit theorems for parametric models under certain smoothness conditions; the covariance of the limiting normal distribution is the inverse of the Fisher information. Nonparametric problems are studied by Cox (1993) and Freedman (1999). Large deviation asymptotics have been studied in Ganesh and O'Connell (1999) and Ganesh and O'Connell (2000); also see Lynch and Sethuraman (1987), who establish a large deviation principle for a sequence of Dirichlet distributions.
In this paper, we obtain a moderate deviation principle for a sequence of Bayes posteriors. We give the general statement of large and moderate deviation principles in the next section and compare our results for Bayes posteriors with Sanov's theorem for empirical measures. We derive the moderate deviation principle on a finite sample space in Section 3, and extend it to more general spaces in Section 4. The extension uses the concept of exponentially good approximations for completely regular topological spaces introduced by Eichelsbacher and Schmock (1998). We present some examples in Section 5 and conclude in Section 6.

## 2 Background

Let $(\mathcal{X}, \tau)$ be a Hausdorff topological space with Borel $\sigma$-algebra $\mathcal{B}$, and let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of probability measures on $(\mathcal{X}, \mathcal{B})$. A *rate function* is a non-negative lower semicontinuous function on $\mathcal{X}$. Let $(\lambda_n)_{n\in\mathbb{N}}$ be a positive sequence decreasing to zero. We say that the sequence $(\mu_n)_{n\in\mathbb{N}}$ satisfies a *large deviation principle* (LDP) on $\mathcal{X}$ with rate function $I$ and speed $\lambda_n$, if for all $B \in \mathcal{B}$,

$$-\inf_{x\in B^\circ} I(x) \le \liminf_{n\to\infty} \lambda_n \log \mu_n(B) \le \limsup_{n\to\infty} \lambda_n \log \mu_n(B) \le -\inf_{x\in \bar{B}} I(x).$$

Here $B^\circ$ and $\bar{B}$ denote the interior and closure of $B$, respectively. A rate function $I$ is *good*, when the level sets $\{x : I(x) \le L\}$, $L \ge 0$, are compact in $(\mathcal{X}, \tau)$.

First we consider the case that $\Omega$ is a finite set and denote by $M_1(\Omega)$ (respectively, $M_b(\Omega)$) the space of probability measures (respectively, finite signed measures) on $\Omega$. Consider a sequence of independent random variables $(X_k)_{k\in\mathbb{N}}$ taking values in $\Omega$, with common law $\mu$. For simplicity we assume that they are defined on the product space $\Omega^{\mathbb{N}}$. Denote by $L_n : \Omega^{\mathbb{N}} \to M_1(\Omega)$ the empirical measure corresponding to the first $n$ observations:

$$L_n(\omega) = \frac{1}{n}\sum_{k=1}^{n} \delta_{X_k}(\omega), \quad n \geq 1,$$

where $\delta_{X_k}$ denotes unit mass at $X_k$. Sanov's theorem (Sanov 1961) tells us that $(\mathbb{P}(L_n \in \cdot))_{n\in\mathbb{N}}$ satisfies the LDP on $M_1(\Omega)$, with speed $1/n$ and with rate function given by the *relative entropy* $H(\cdot|\mu)$. Here,

$$H(\nu|\mu) = \begin{cases} \sum_{x\in\Omega} \nu(x) \log \frac{\nu(x)}{\mu(x)}, & \text{if } \nu \ll \mu. \\ +\infty, & \text{otherwise.} \end{cases}$$

If $(b_n)_{n\in\mathbb{N}}$ is a positive sequence such that

$$\frac{b_n}{n} \to 0 \text{ and } \frac{b_n^2}{n} \to \infty \quad \text{as } n \to \infty, \tag{1}$$

then

$$\left(\mathbb{P}(\frac{n}{b_n}(L_n - \mu) \in \cdot)\right)_{n\in\mathbb{N}}$$

satisfies the LDP on $M_b(\Omega)$ with speed $n/b_n^2$ and rate function

$$I(\nu) = \begin{cases} \frac{1}{2}\sum_{x\in\Omega} \frac{\nu(x)^2}{\mu(x)}, & \text{if } \sum_{x\in\Omega} \nu(x) = 0, \\ +\infty, & \text{otherwise,} \end{cases} \tag{2}$$

where $\nu \in M_b(\Omega)$. $I$ is sometimes called the *Fisher-information*. This is the so-called *moderate deviation principle* (MDP). For a sequence of $\mathbb{R}^d$-valued i. i. d. random variables $X_i$ with a finite moment generating function Mogulskii (1976) investigated the moderate deviation behaviour for $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. A moderate deviation principle for $(S_n)_{n\in\mathbb{N}}$ was proved in de Acosta (1992). The moderate deviation behaviour for $(\frac{n}{b_n}(L_n - \mu))_{n\in\mathbb{N}}$ was first considered in Borovkov and Mogulskii (1978), the "full" moderate deviation principle in de Acosta (1994b).

Suppose next that the $X_k$'s are i.i.d., but with an unknown distribution which is to be inferred by a Bayesian procedure. Let the unknown probability distribution of $X_1$ be assigned a prior $\pi \in M_1(M_1(\Omega))$, with support

denoted by supp $\pi$. The posterior distribution, given the first $n$ observations, is a function of the empirical measure $L_n$ and will be denoted by $\pi^n(L_n)$. Given a sequence $(X_n)_{n\in\mathbb{N}}$ (equivalently, $(L_n)_{n\in\mathbb{N}}$), let $(\nu_n)_{n\in\mathbb{N}}$ be a sequence of $M_1(\Omega)$-valued random variables (i.e., random probability measures on $\Omega$) such that $\nu_n$ has distribution $\pi^n(L_n)$ for each $n \geq 1$. (Ganesh and O'Connell 1999) showed that, on the set $\{L_n \to \mu\}$, for any $\mu$ in the support of the prior, the sequence $(\mathbb{P}(\nu_n \in \cdot))_{n\in\mathbb{N}}$ $(= (\pi^n(L_n)(\cdot))_{n\in\mathbb{N}})$ satisfies the LDP on $M_1(\Omega)$ with speed $1/n$ and rate function $J(\cdot)$ given by

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

Let $(b_n)_{n\in\mathbb{N}}$ be a sequence satisfying (1). We prove that, on the set $\frac{n}{b_n}(L_n - \mu) \to 0$, for any "regular" point $\mu$ in the support of the prior (regularity is defined below), the sequence $(\mathbb{P}(\frac{n}{b_n}(\nu_n - \mu) \in \cdot))_{n\in\mathbb{N}}$ satisfies the LDP on $M_b(\Omega)$ with speed $n/b_n^2$ and the same rate function $I$ as in (2) above. In other words, the moderate deviations behaviour of the empirical process is identical to that of the corresponding Bayes posteriors. This is in marked contrast to their large deviations behaviour, where the rate function is $H(\cdot|\mu)$ for the empirical process and $H(\mu|\cdot)$ for the Bayesian posterior distributions. We extend the result to sequences taking values in a more general space $(S, \mathcal{S})$ under additional conditions on the prior. Note that we cannot expect the MDP to hold on general spaces for arbitrary priors because a corollary of the MDP and the fact that the rate function has a unique zero at $\mu$ is that the posterior concentrates in weak neighbourhoods of the true distribution, $\mu$; but Freedman (1963) exhibited a prior on the natural numbers for which the posterior failed to concentrate in a weak neighbourhood of the true distribution.

## 3 The finite case

Let $\Omega$ be a finite set, and let $M_1(\Omega)$ denote the space of probability measures on $\Omega$. Suppose $X_1, X_2, \ldots$ are i.i.d. $\Omega$–valued random variables with unknown distribution. Let $\pi \in M_1(M_1(\Omega))$ denote the prior distribution of the law of $X_1$, with support denoted by supp $\pi$. For each $n$, set

$$M_1^n(\Omega) = \left\{ \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} : x \in \Omega^n \right\}.$$

Define a mapping $\pi^n : M_1^n(\Omega) \to M_1(M_1(\Omega))$ by its Radon-Nikodym derivative on the support of $\pi$:

$$\frac{d\pi^n(\mu_n)}{d\pi}(\nu) = \frac{\prod_{x\in\Omega} \nu(x)^{n\mu_n(x)}}{\int_{M_1(\Omega)} \prod_{x\in\Omega} \lambda(x)^{n\mu_n(x)} \pi(d\lambda)}, \tag{3}$$

if the denominator is non-zero (which happens $\pi$-almost surely); $\pi^n$ is undefined at $\mu_n$ 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 $\mu_n \in M_1^n(\Omega)$ (this is immediate from Bayes' formula).
**Definition**: Let $\lambda$ denote Lebesgue measure on the $|\Omega|$-simplex,

$$S^\Omega = \{\mathbf{x} \in \mathbb{R}^\Omega : x_i \geq 0 \ \forall \ i \in \Omega, \ \sum_{i\in\Omega} x_i = 1\}. \tag{4}$$

We identify the elements of $S^\Omega$ with probability distributions on $\Omega$. We say that $\mu \in \text{supp } \pi$ is a *regular point* of the support of $\pi$ if there is a neighbourhood of $\mu$ (in total variation distance) in $S^\Omega$ on which $\pi$ is absolutely continuous with respect to $\lambda$ and $d\pi/d\lambda$, the density of $\pi$ with respect to Lebesgue measure, is bounded away from zero and infinity on this neighbourhood.
Let $\nu$ be a signed measure and $f$ a real-valued function on $\Omega$. We define

$$\|\nu\|_1 = \sum_{x\in\Omega} |\nu(x)|, \quad \|f\|_\infty = \max_{x\in\Omega} f(x).$$

If $\mu, \nu$ are probability measures on $\Omega$, then the total variation distance between $\mu$ and $\nu$ is $d_{TV}(\mu,\nu) = \frac{1}{2}\|\mu - \nu\|_1$.

**Theorem 1** *Let $(b_n)_{n\in\mathbb{N}}$ be a positive sequence satisfying (1) and let $\mu$ be a regular point of the support of $\pi$ with $\mu(z) > 0$ for all $z \in \Omega$. Suppose $x \in \Omega^{\mathbb{N}}$ is such that the sequence $\mu_n = \sum_{i=1}^n \delta_{x_i}/n$ satisfies $\|\mu_n - \mu\|_1 = o(b_n/n)$. Let $(\nu_n)_{n\in\mathbb{N}}$ be $M_1(\Omega)$-valued random variables such that $\mathbb{P}(\nu_n \in \cdot) = \pi^n(\mu_n)(\cdot)$ for each $n \geq 1$. Then the sequence $\left(\mathbb{P}(\frac{n}{b_n}(\nu_n - \mu) \in \cdot)\right)_{n\in\mathbb{N}}$ satisfies the LDP on $M_b(\Omega)$ with speed $n/b_n^2$ and with convex good rate function $I$ given by*

$$I(\nu) = \begin{cases} \frac{1}{2}\sum_{x\in\Omega} \frac{\nu(x)^2}{\mu(x)}, & \text{if } \sum_{x\in\Omega} \nu(x) = 0, \\ +\infty, & \text{otherwise.} \end{cases}$$

**Proof**: Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence satisfying the conditions of the theorem, and let $(\pi^n(\mu_n))_{n\in\mathbb{N}}$ denote the corresponding sequence of Bayesian posterior distributions. Let $\nu_n$ be a random element of $M_1(\Omega)$, with distribution

$\pi^n(\mu_n)$. We want to establish a moderate deviations principle for the sequence of signed measures, $\eta_n = (n/b_n)(\nu_n - \mu)$. We shall do this by evaluating the "moment-generating functions", $E_{\pi^n}[\exp(b_n^2/n)\langle f, \eta_n\rangle]$ for arbitrary $f : \Omega \to \mathbb{R}$ and using the Gärtner-Ellis theorem (Dembo and Zeitouni (1998, Chapter 4.5)). Here, $\langle f, \eta\rangle$ denotes $\sum_{x\in\Omega} f(x)\eta(x)$. Observe from (3) and the definition of $\eta_n$ that

$$\begin{aligned} E_{\pi^n}\left[\exp\frac{b_n^2}{n}\langle f, \eta_n\rangle\right] &= E_{\pi^n}\left[\exp b_n\langle f, \nu_n - \mu\rangle\right] \\ &= \frac{\int_{M_1(\Omega)} \exp\sum_{x\in\Omega}\left[n\mu_n(x)\log\nu(x) + b_n f(x)(\nu(x) - \mu(x))\right]\pi(d\nu)}{\int_{M_1(\Omega)} \exp\sum_{x\in\Omega}\left[n\mu_n(x)\log\nu(x)\right]\pi(d\nu)} \\ &= \frac{\int_{M_1(\Omega)} \exp\left[-nH(\mu_n|\nu) + b_n\langle f, \nu - \mu\rangle\right]\pi(d\nu)}{\int_{M_1(\Omega)} \exp[-nH(\mu_n|\nu)]\pi(d\nu)}. \end{aligned} \tag{5}$$

We first find the value of $\nu$ that maximizes the exponent in the numerator above. Consider the constrained optimization problem:

$$\begin{aligned} \min \quad & nH(\mu_n|\nu) - b_n\langle f, \nu - \mu\rangle \\ \text{subject to} \quad & \sum_{x\in\Omega}\nu(x) - 1 = 0, \quad \nu(x) \geq 0 \ \forall\, x \in \Omega. \end{aligned}$$

The Kuhn-Tucker conditions (see, for example, Luenberger (1965), Section 10.6) for the optimality of a feasible solution $\lambda_n$ are that there exist $\alpha_n \in \mathbb{R}$ and $\beta_n(x) \geq 0$ for each $x \in \Omega$, such that

$$\begin{aligned} -\frac{n\mu_n(x)}{\lambda_n(x)} - b_n f(x) + \alpha_n - \beta_n(x) &= 0, \\ \text{and} \quad \lambda_n(x)\beta_n(x) &= 0, \end{aligned}$$

for all $x \in \Omega$. Now, $\mu(x) > 0$ for all $x \in \Omega$ and $\|\mu_n - \mu\|_1 \to 0$ by assumption, so, for sufficiently large $n$, $\mu_n(x) > 0$ for all $x \in \Omega$. Hence, it follows from the above that $\lambda_n(x) > 0$ and $\beta_n(x) = 0$ for all $x \in \Omega$. Thus, we have $\alpha_n - b_n f(x) > 0$ for all $x \in \Omega$ and the solution of the Kuhn-Tucker conditions is given by

$$\lambda_n(x) = \frac{n\mu_n(x)}{\alpha_n - b_n f(x)}, \tag{6}$$

where the constant $\alpha_n$ is chosen so that $\sum_{x\in\Omega}\lambda_n(x) = 1$. Now $f$ is fixed, $b_n/n \to 0$ by assumption and $\sum_{x\in\Omega}\mu_n(x) = 1$, so it follows that $\alpha_n = n + O(b_n)$. Note that $nH(\mu_n|\nu) - b_n\langle f, \nu - \mu\rangle$ is a convex function of $\nu$ (see,

for example, Dupuis and Ellis (1997), Lemma 1.4.3) and the constraints are linear in $\nu$, so the Kuhn-Tucker conditions are *sufficient* for the optimality of $\lambda_n$.
Now, using the fact $\alpha_n = O(n)$ and $b_n/n \to 0$, we can rewrite the constraint, $\sum_{x\in\Omega} \lambda_n(x) = 1$, as

$$\frac{n}{\alpha_n} + \frac{nb_n}{\alpha_n^2}\langle f, \mu_n\rangle + \frac{nb_n^2}{\alpha_n^3}\langle f^2, \mu_n\rangle + O\left(\frac{b_n^3}{n^3}\right) = 1. \tag{7}$$

To obtain the above, we have used the fact that $\mu_n$ is a probability distribution, so that $\sum_{x\in\Omega} \mu_n(x) = 1$. Now, (7) implies the asymptotic expansion,

$$\alpha_n = n\left[1 + a_1\frac{b_n}{n} + a_2\frac{b_n^2}{n^2} + O\left(\frac{b_n^3}{n^3}\right)\right],$$

where $a_1$ and $a_2$ are unknown polynomials in $\langle f, \mu_n\rangle$ and $\langle f^2, \mu_n\rangle$, which are $O(1)$ quantities. Substituting this in (7) and simplifying, we obtain the solution,

$$a_1 = \langle f, \mu_n\rangle, \quad a_2 = \mathrm{Var}_n f := \langle f^2, \mu_n\rangle - \langle f, \mu_n\rangle^2,$$

and so,

$$\alpha_n = n + b_n\langle f, \mu_n\rangle + \frac{b_n^2}{n}\mathrm{Var}_n f + O\left(\frac{b_n^3}{n^2}\right), \tag{8}$$

as can be verified by substituting back in (7).
In order to evaluate $E_{\pi^n}[\exp\frac{b_n^2}{n}\langle f, \eta_n\rangle]$, we rewrite (5) as

$$\begin{aligned}
&E_{\pi^n}\left[\exp\frac{b_n^2}{n}\langle f, \eta_n\rangle\right] = \exp\left[-nH(\mu_n|\lambda_n) + b_n\langle f, \lambda_n - \mu_n\rangle\right]\frac{Z_1}{Z}, \quad \text{where}\\
&Z_1 = \int_{M_1(\Omega)} \exp\left[-nH(\mu_n|\nu) + nH(\mu_n|\lambda_n) + b_n\langle f, \nu - \lambda_n\rangle\right]\pi(d\nu),\\
&Z = \int_{M_1(\Omega)} \exp[-nH(\mu_n|\nu)]\pi(d\nu).
\end{aligned} \tag{9}$$

We have from (6) and (8) that

$$\begin{aligned}
&-nH(\mu_n|\lambda_n) = -n\sum_{x\in\Omega}\mu_n(x)\log\left[1 + \frac{\mu_n(x) - \lambda_n(x)}{\lambda_n(x)}\right]\\
= \; &-\sum_{x\in\Omega}(\alpha_n - b_n f(x))(\mu_n(x) - \lambda_n(x))
\end{aligned}$$

$$
\begin{aligned}
&+\frac{n}{2}\sum_{x\in\Omega}\mu_n(x)\left(\frac{\alpha_n}{n}-\frac{b_n}{n}f(x)-1\right)^2+O\left(\frac{b_n^3}{n^2}\right)\\
=\;& b_n\langle f,\mu_n-\lambda_n\rangle+\sum_{x\in\Omega}\frac{n\mu_n(x)}{2}\frac{b_n^2}{n^2}(\langle f,\mu_n\rangle-f(x))^2+O\left(\frac{b_n^3}{n^2}\right).
\end{aligned}
$$

Here, we have used the Taylor expansion, $\log(1+x)=x-(x^2/2)+O(x^3)$, and the fact that $\|\lambda_n-\mu_n\|_1=O(b_n/n)$ to obtain the second equality above, and the fact that $\sum\lambda_n(x)=\sum\mu_n(x)=1$ to obtain the last equality. On simplifying the above, we get

$$
-nH(\mu_n|\lambda_n)+b_n\langle f,\lambda_n-\mu_n\rangle=\frac{b_n^2}{2n}\mathrm{Var}_n f+O\left(\frac{b_n^3}{n^2}\right).
$$

Define $\mathrm{Var}f=\langle f^2,\mu\rangle-\langle f,\mu\rangle^2$. By the assumption that $\|\mu_n-\mu\|_1=o(b_n/n)$, we have

$$
\mathrm{Var}f=\mathrm{Var}_n f+o(b_n/n),
$$

and it follows that

$$
-nH(\mu_n|\lambda_n)+b_n\langle f,\lambda_n-\mu_n\rangle=\frac{b_n^2}{2n}\mathrm{Var}f+O\left(\frac{b_n^3}{n^2}\right). \tag{10}
$$

We now proceed to derive upper and lower bounds on $Z_1$, $Z$ defined in (9). Let $\alpha,\delta>0$ be arbitrary, and define

$$
\begin{aligned}
\mathcal{A} &= \{\nu\in M_1(\Omega):\ |\nu(x)-\lambda_n(x)|\le\alpha n^{-1/2}\ \text{for all } x\in\Omega\},\\
\mathcal{A}_\delta &= \{\nu\in M_1(\Omega):\ |\nu(x)-\lambda_n(x)|\le\delta\ \text{for all } x\in\Omega\}.
\end{aligned}
$$

For any $\alpha,\delta>0$,

$$
\mathcal{A}\subseteq\mathcal{A}_\delta\subseteq\{\nu\in M_1(\Omega):\ \|\mu-\nu\|_1\le 2|\Omega|\delta\}
$$

for all $n$ sufficiently large, because $b_n/n\to 0$, $\|\mu-\mu_n\|_1=o(b_n/n)$ by assumption, and it is clear from (6) and (8) that $\|\mu_n-\lambda_n\|_1=O(b_n/n)$. Hence, it follows from the assumption that $\mu$ is a regular point of the support of $\pi$ and the definition of regularity that, if $\delta$ is sufficiently small, then

$$
\exists\ k>0,\ K<\infty:\ k\le\frac{d\pi}{d\lambda}\le K\ \text{on}\ \mathcal{A}\ \text{and}\ \mathcal{A}_\delta, \tag{11}
$$

where $\lambda$ denotes Lebesgue measure on the $|\Omega|$-simplex, $S^\Omega$.

For $\nu \in \mathcal{A}$, we have

$$\begin{aligned} g(\nu) &:= b_n\langle f, \nu - \lambda_n\rangle - nH(\mu_n|\nu) + nH(\mu_n|\lambda_n) \\ &= b_n\langle f, \nu - \lambda_n\rangle + n\sum_{x\in\Omega}\mu_n(x)\log\left[1 + \frac{\nu(x)-\lambda_n(x)}{\lambda_n(x)}\right] \\ &= b_n\langle f, \nu - \lambda_n\rangle + \sum_{x\in\Omega}(\alpha_n - b_n f(x))(\nu(x)-\lambda_n(x)) \\ &\qquad -\frac{n}{2}\sum_{x\in\Omega}\mu_n(x)\left(\frac{\nu(x)-\lambda_n(x)}{\lambda_n(x)}\right)^2 + O(n^{-1/2}) \\ &\geq -c, \end{aligned} \tag{12}$$

for some constant $c > 0$ and all $n$ sufficiently large. We have used the fact that $\sum\nu(x) = \sum\lambda_n(x) = 1$ to obtain the inequality above. Thus, we have from (9), (11) and the definition of $g$, that

$$\begin{aligned} Z_1 &= \int_{M_1(\Omega)} e^{g(\nu)}\pi(d\nu) \;\geq\; \int_{\mathcal{A}} e^{g(\nu)}\pi(d\nu) \\ &\geq \int_{\mathcal{A}} ke^{-c}d\lambda \;=\; ke^{-c}\text{ vol }(\mathcal{A}) \\ &\geq cn^{-\frac{|\Omega|-1}{2}}, \end{aligned} \tag{13}$$

for a generic constant $c$ that does not depend on $n$. Here, vol $(\mathcal{A})$ denotes the volume of $\mathcal{A}$ as a subset of $\mathbb{R}^{|\Omega|-1}$.
Observe that $g$ is a concave function of $\nu$, $g(\lambda_n) = 0$ and $\lambda_n$ was defined so that $g$ attains its maximum over $M_1(\Omega)$ at $\lambda_n$. Thus,

$$g(\nu) \leq 0 \quad \text{for all } \nu \in \mathcal{A}. \tag{14}$$

Next, if $\nu \in M_1(\Omega)\backslash\mathcal{A}$, let

$$\epsilon = \frac{\alpha n^{-1/2}}{\max_{x\in\Omega}|\nu(x)-\lambda_n(x)|}, \quad \nu_\epsilon = \epsilon\nu + (1-\epsilon)\lambda_n. \tag{15}$$

Clearly, $\nu_\epsilon \in \partial\mathcal{A}$, the boundary of $\mathcal{A}$, and we have by the concavity of $g$ that

$$g(\nu_\epsilon) \geq \epsilon g(\nu) + (1-\epsilon)g(\lambda_n) = \epsilon g(\nu), \quad \text{i.e.,} \quad g(\nu) \leq \frac{g(\nu_\epsilon)}{\epsilon}. \tag{16}$$

Moreover, arguing along the lines in (12) yields the reverse inequality

$$\sup_{\nu\in\partial\mathcal{A}} g(\nu) \;\leq\; -c \;<\; 0, \tag{17}$$

for some constant $c$ and all $n$ sufficiently large. It follows from (15), (16) and (17) that for all $\nu \in M_1(\Omega)\backslash\mathcal{A}$ ($\supseteq M_1(\Omega)\backslash\mathcal{A}_\delta$), we have

$$g(\nu) \leq \frac{-cn^{1/2}}{\alpha} \max_{x\in\Omega} |\nu(x) - \lambda_n(x)| \leq \frac{-cn^{1/2}}{\alpha|\Omega|} \sum_{x\in\Omega} |\nu(x) - \lambda_n(x)|. \tag{18}$$

Thus, using (14) and (18), we get the upper bound,

$$\begin{aligned}
Z_1 &= \int_{M_1(\Omega)} e^{g(\nu)}\pi(d\nu) \\
&= \int_{\mathcal{A}} e^{g(\nu)}\pi(d\nu) + \int_{\mathcal{A}_\delta\backslash\mathcal{A}} e^{g(\nu)}\pi(d\nu) + \int_{M_1(\Omega)\backslash\mathcal{A}_\delta} e^{g(\nu)}\pi(d\nu) \\
&\leq \int_{\mathcal{A}} \pi(d\nu) + \int_{\mathcal{A}_\delta\backslash\mathcal{A}} \exp\Big[-cn^{1/2} \sum_{x\in\Omega} |\nu(x) - \lambda_n(x)|\Big]\pi(d\nu) \\
&\qquad + \int_{M_1(\Omega)\backslash\mathcal{A}_\delta} \exp\Big[-cn^{1/2} \sum_{x\in\Omega} |\nu(x) - \lambda_n(x)|\Big]\pi(d\nu) \\
&\leq K \text{ vol } (\mathcal{A}) + \prod_{j=1}^{|\Omega|-1} \int_{-\delta}^{\delta} Ke^{-c\sqrt{n}|x_j|}dx_j + \int_{M_1(\Omega)\backslash\mathcal{A}_\delta} e^{-c\sqrt{n}\delta}\pi(d\nu) \\
&\leq c_1 n^{-\frac{|\Omega|-1}{2}} + c_2 n^{-\frac{|\Omega|-1}{2}} + e^{-c_3\sqrt{n}},
\end{aligned}$$

for some non-zero constants $c_1$, $c_2$, $c_3$. Thus, for some constant $c$ and large enough $n$, we have

$$Z_1 \leq cn^{-\frac{|\Omega|-1}{2}}. \tag{19}$$

The derivation of upper and lower bounds on $Z$ proceeds along virtually the same lines, except that the sets $\mathcal{A}$, $\mathcal{A}_\delta$ are centered at $\mu_n$ rather than $\lambda_n$, and the Taylor expansion of $H(\mu_n|\nu)$ is also done around $\mu_n$ (for deriving the upper bound on $Z$, we can use the well-known inequality, $H(\mu_n|\nu) \geq \frac{1}{2}\|\mu_n - \nu\|_1^2$). We omit the details and simply state the result: there are non-zero constants $c_1$ and $c_2$ such that

$$c_1 n^{-\frac{|\Omega|-1}{2}} \leq Z \leq c_2 n^{-\frac{|\Omega|-1}{2}} \tag{20}$$

for all $n$ sufficiently large. Now, we have from (9), (10), (13), (19) and (20) that, for arbitrary $f : \Omega \to \mathbb{R}$,

$$\lim_{n\to\infty} \frac{n}{b_n^2} \log E_{\pi^n}\left[\exp \frac{b_n^2}{n}\langle f, \eta_n\rangle\right] = \frac{1}{2}\text{Var}f, \tag{21}$$

where $\eta_n = (n/b_n)(\nu_n - \mu)$. Hence, it follows from the Gärtner-Ellis theorem (Dembo and Zeitouni (1998, Theorem 2.3.6)) that the sequence $\mathbb{P}(\frac{n}{b_n}(\nu_n - \mu) \in \cdot)_{n\in\mathbb{N}}$ satisfies the LDP on $M_b(\Omega)$ with speed $n/b_n^2$ and with convex good rate function $I$ which is the convex dual of $\mathrm{Var}f/2$. In other words,

$$I(\nu) = \sup_f \langle f, \nu\rangle - \frac{1}{2}\mathrm{Var}f. \tag{22}$$

Suppose $\nu \in M_b(\Omega)$ is such that $\sum_{x\in\Omega}\nu(x) \neq 0$. Take $f$ to be a constant, so that $\mathrm{Var}f = 0$. By choosing the constant appropriately, we can make the right hand side of (22) arbitrarily large; so $I(\nu)$ is infinite. Otherwise, if $\sum_{x\in\Omega}\nu(x) = 0$, then we find that the supremum in (22) is achieved at $f$ given by $f(x) = \nu(x)/\mu(x)$ for all $x \in \Omega$. Substituting this in (22), we find that

$$I(\nu) = \sum_{x\in\Omega}\frac{\nu(x)^2}{2\mu(x)}.$$

This completes the proof the theorem.

## 4 The general case

Now let $M_1(S)$ and $M(S)$, respectively, denote the set of probability measures and of finite signed measures on a Polish space $S$ with Borel $\sigma$-algebra $\mathcal{S}$. The $\tau$-topology on $M_1(S)$ ($M(S)$ respectively) is defined to be the coarsest topology which makes the maps $M_1(S) \ni \mu \mapsto \mu(A)$ continuous for every $A \in \mathcal{S}$. Let $\mathcal{B}(M_1(S))$ ($\mathcal{B}(M(S))$ respectively) be the $\sigma$-algebra generated by the maps $M_1(S) \ni \mu \mapsto \mu(A)$ with $A \in \mathcal{S}$. Let $\{X_k\}_{k\in\mathbb{N}}$ denote the projection maps on the product space $(\Omega, \mathcal{A}) \equiv (S^{\mathbb{N}}, \mathcal{S}^{\otimes\mathbb{N}})$ and define the empirical measure $L_n(\omega) = \frac{1}{n}\sum_{k=1}^n \delta_{X_k(\omega)} \in M_1(S)$ for every $\omega \in \Omega$ and $n \in \mathbb{N}$. Note that $L_n : \Omega \to M_1(S)$ is $\mathcal{A}$-$\mathcal{B}(M_1(S))$-measurable. Given $\mu \in M_1(S)$, define

$$I(\nu) = \begin{cases} \frac{1}{2}\int_S (\frac{d\nu}{d\mu})^2\, d\mu, & \text{if } \nu \ll \mu \text{ and } \nu(S) = 0. \\ +\infty, & \text{otherwise.} \end{cases}$$

By Lemma 2.1 in de Acosta (1994a), the level sets of $I$ are $\tau$-compact, therefore $I$ is a good rate function. In this general setting de Acosta proved the MDP for $\{n/b_n(L_n - \mu)\}_{n\in\mathbb{N}}$ with rate $I$. He applied the *projective limit approach* (see, for example, Theorem 4.6.1 in Dembo and Zeitouni (1998)). Here, we are interested in an MDP for a sequence of Bayes posteriors corresponding to the empirical distributions, $L_n$. For *large deviations* Ganesh

and O'Connell (2000) pointed out that an extension of a result for finite state spaces $S$ to a general space $S$ would require additional assumptions about the prior distribution. They proved a large deviation principle for compact spaces $S$ and for Dirichlet priors.
On the moderate scale, defined by (1), we obtain a result in a topology which is weaker than the $\tau$-topology on $M(S)$.
We assume that the prior $\pi$ is *exchangeable with respect to a sequence of nested partitions* and we impose a regularity assumption on the support of the prior. We will discuss examples of priors satisfying the exchangeability assumption thereafter.
By a partition $P$ of $S$ we mean a finite collection $\{A_1, \ldots, A_n\} \subset \mathcal{S}$ of disjoint sets whose union is $S$. Let $\mathcal{P}$ be the set of all partitions of $S$. For $P \in \mathcal{P}$ let $\sigma(P)$ denote the $\sigma$-algebra generated by $P$.
In what follows, we shall restrict ourselves to a specific sequence of *nested* measurable partitions, $(P_m)_{m\in\mathbb{N}}$, (sometimes called a *tree* of measurable partitions) of $S$; $P_1, P_2, \ldots$ is a sequence of measurable partitions such that $P_{m+1}$ is a refinement of $P_m$ (means $P_m \subset \sigma(P_{m+1})$) for each $m = 1, 2, \ldots$, and such that $\cup_{m\geq 1} P_m$ generates the $\sigma$-algebra $\mathcal{S}$. We note that $(P_m)_{m\in\mathbb{N}}$ is a directed set with respect to $\preceq$, where $P \preceq P'$ for $P, P' \in \mathcal{P}$ means $P \subset \sigma(P')$, that is, $P'$ is a refinement of $P$. (We say $\mathcal{P}$ is a directed set if it is partially ordered and right filtering, i.e., given $P, P' \in \mathcal{P}$, there is a $P'' \in \mathcal{P}$ such that $P \preceq P''$ and $P' \preceq P''$. Since $(P_m)_{m\in\mathbb{N}}$ is totally ordered, it is clearly directed).
The restriction of a measure $\mu \in M_1(S)$ to the $\sigma$-algebra $\sigma(P_m)$ is denoted by $\mu_{P_m}$, i.e., $\mu_{P_m} = E[\mu|\sigma(P_m)]$. For a prior $\pi \in M_1(M_1(S))$ we denote by $\pi_{P_m}$ the corresponding element in $M_1(M_1(S, \sigma(P_m)))$, thus the restriction of $\pi$ to the Borel $\sigma$-algebra $\mathcal{B}(M_1(S, \sigma(P_m)))$. More precisely, we define $\pi_{P_m}$ by setting

$$\pi_{P_m}(B) = \pi(\{\nu \in M_1(S) : \nu_{P_m} \in B\}),$$

for all $B \in \mathcal{B}(M_1(S, \sigma(P_m)))$. For $\mu_n \in M_1(S)$ and $\pi^n(\mu_n) \in M_1(M_1(S))$ the elements $\mu_{n,P_m}$ and $(\pi^n(\mu_n))_{P_m}$ are defined analogously. Here $\pi^n(\mu_n)$ is (a version of) the posterior distribution conditional on the observations $X_1, \ldots, X_n$ having empirical distribution $\mu_n \in M_1(S)$.
We denote by $\pi^n_{P_m}(\mu_{n,P_m})$ the posterior distribution on $M_1(S, \sigma(P_m))$ corresponding to the prior $\pi_{P_m}$ and the empirical distribution restricted to $\sigma(P_m)$, $\mu_{n,P_m}$.

Now we can define
**Definition**: A prior measure $\pi \in M_1(M_1(S))$ is *exchangeable with respect to a sequence of nested partitions* if, for every partition $P_m$ of the sequence

$(P_m)_{m\in\mathbb{N}}$, we have the identity

$$(\pi^n(\mu_n))_{P_m} = \pi^n_{P_m}(\mu_{n,P_m}).$$

The interpretation of the definition is the following: let $P_m = \{P_1^m, P_2^m, \ldots, P_{k_m}^m\}$ and $a_1^m(n), a_2^m(n), \ldots, a_{k_m}^m(n)$ be the number of observations from $X_1, X_2, \ldots, X_n$ in the cells $P_l^m$, $l \in \{1, \ldots, k_m\}$. Then $\pi$ is called exchangeable with respect to $(P_m)_m$, if for all $n, m \in \mathbb{N}$, the posterior distribution of $(\nu(P_1^m), \nu(P_2^m), \ldots, \nu(P_{k_m}^m))$ given $X_1, X_2, \ldots, X_n$ is the same as that given $a_1^m(n), a_2^m(n), \ldots, a_{k_m}^m(n)$. Recall that the posterior distribution may not be unique. We only require the above to hold for some version of the posterior; the conclusions of Theorem 2 below will then apply to that version. We will discuss examples of exchangeable priors at the end of this section.
A probability measure $\mu \in M_1(S)$ is called a *regular point* of the support of $\pi$ with respect to the sequence of partitions $(P_m)_{m\in\mathbb{N}}$ if $\mu_{P_m}$ is a regular point of the support of $\pi_{P_m}$ (as defined preceding the statement of Theorem 1) for every $m \in \mathbb{N}$.
The *projective limit topology* on $M_1(S)$ ($M(S)$ respectively) is defined to be the coarsest topology which makes the maps $M_1(S) \ni \mu \mapsto \mu(A)$ continuous for every $A \in P_m$ and $m \in \mathbb{N}$. In other words $\mu_n \to \mu$ in the projective limit topology if $\mu_{n,P_m} \to \mu_{P_m}$ for all $m \in \mathbb{N}$.

**Theorem 2** *Let $(b_n)_{n\in\mathbb{N}}$ be a positive sequence satisfying (1). Let $\pi$ be an exchangeable prior and $\mu$ a regular point of the support of $\pi$ with respect to a sequence of nested partitions, $(P_m)_{m\in\mathbb{N}}$, such that $\mu(A) > 0$ for all $A \in \cup_{m\in\mathbb{N}} P_m$. Suppose $x \in S^{\mathbb{N}}$ is such that the sequence $\mu_n = \sum_{i=1}^n \delta_{x_i}/n$ satisfies $\|\mu_{n,P_m} - \mu_{P_m}\|_1 = o(b_n/n)$ for all $m \in \mathbb{N}$. Let $(\nu_n)_{n\in\mathbb{N}}$ be a sequence of $M_1(S)$-valued random variables such that $\mathbb{P}(\nu_n \in \cdot) = \pi^n(\mu_n)(\cdot)$ for all $n \geq 1$. Then the sequence $(\frac{n}{b_n}(\nu_n - \mu))_{n\in\mathbb{N}}$ satisfies the LDP on $M(S)$ with respect to the projective limit topology, with speed $n/b_n^2$ and with convex good rate function $I$ given by*

$$I(\nu) = \begin{cases} \frac{1}{2}\int_S (\frac{d\nu}{d\mu})^2 \, d\mu, & \text{if } \nu \ll \mu \text{ and } \nu(S) = 0. \\ +\infty, & \text{otherwise.} \end{cases}$$

In proving the Theorem, we will use the concept of exponential approximations in completely regular topological spaces introduced in Eichelsbacher and Schmock (1998). A completely regular topological space $(Y, \mathcal{T})$ is a space where the topology $\mathcal{T}$ is generated by a family $\mathcal{D}$ of pseudo-metrics which is separating, that is, for each pair of points $x \neq y$ in $Y$ there exists a

pseudo-metric $d \in \mathcal{D}$ such that $d(x,y) \neq 0$. Let $\mathcal{D}'$ be the smallest family of pseudo-metrics on $Y$ which contains $\mathcal{D}$ and is closed with respect to finite maxima.
We observe that the projective limit topology turns $M_1(S)$ as well as $M(S)$ into a completely regular space with the separating family $\{d_A\}_{A \in \cup_{m \geq 1} P_m}$ of pseudo-metrics, where $d_A(\mu,\nu) := |\mu(A) - \nu(A)|$ for all $\mu, \nu \in M_1(S)(M(S))$. Define the balls $B(y,d,\delta) := \{x \in Y | d(x,y) < \delta\}$ with $y \in Y, d \in \mathcal{D}$ and $\delta > 0$.
A collection $\{\mu_{n,i}\}_{n \in \mathbb{N}, i \in \mathcal{I}} \subset M_1(Y)$ with $(\mathcal{I}, \preceq)$ a nonempty directed set is called a $\mathcal{D}$-exponentially good approximation of a sequence $\{\tilde{\mu}_n\}_{n \in \mathbb{N}} \subset M_1(Y)$, if for every $d \in \mathcal{D}'$, $n \in \mathbb{N}$ and $i \in \mathcal{I}$ there exists a probability measure $\nu_{d,n,i}$ on a $\sigma$-algebra $\mathcal{Y}_{d,n,i}$ containing $\mathcal{Y}^{\otimes 2}$ such that the two marginals are $\mu_{n,i}$ and $\tilde{\mu}_n$, respectively, and

$$\limsup_{i \in \mathcal{I}} \limsup_{n \to \infty} \frac{1}{n} \log \nu_{d,n,i}(\{(y,\tilde{y}) \in Y^2 | d(y,\tilde{y}) > \delta\}) = -\infty \tag{23}$$

for every $\delta > 0$. If the measures $\nu_{d,n,i}$ and the $\sigma$-algebras $\mathcal{Y}_{d,n,i}$ do not depend on $d \in \mathcal{D}$, then condition (23) for all $d \in \mathcal{D}$ implies condition (23) for all $d \in \mathcal{D}'$.
We define as usual $\limsup_{i \in \mathcal{I}} a_i = \inf_{j \in \mathcal{I}} \sup_{i \in \mathcal{I}, j \preceq i} a_i$ and $\liminf_{i \in \mathcal{I}} a_i = \sup_{j \in \mathcal{I}} \inf_{i \in \mathcal{I}, j \preceq i} a_i$. We will apply the following Theorem (Theorem 1.6 in Eichelsbacher and Schmock (1998)):

**Theorem 3** *If $\{\mu_{n,i}\}_{n \in \mathbb{N}, i \in \mathcal{I}} \subset M_1(Y)$ is a $\mathcal{D}$-exponentially good approximation of $\{\tilde{\mu}_n\}_{n \in \mathbb{N}}$ and if, for every $i \in \mathcal{I}$, the family $\{\mu_{n,i}\}_{n \in \mathbb{N}}$ satisfies a LDP with a not necessarily good rate function $J_i$, then the following statements hold:*

1. *$\{\tilde{\mu}_n\}_{n \in \mathbb{N}}$ satisfies a weak LDP with rate function*

$$J(y) \equiv \sup_{d \in \mathcal{D}', \delta > 0} \liminf_{i \in \mathcal{I}} \inf_{z \in B(y,d,\delta)} J_i(z), \quad y \in Y. \tag{24}$$

2. *If $J$ is a good rate function and if for every measurable closed subset $C$ of $Y$*

$$\inf_{y \in C} J(y) \leq \limsup_{i \in \mathcal{I}} \inf_{y \in C} J_i(y), \tag{25}$$

*then the LDP holds for $\{\tilde{\mu}_n\}_{n \in \mathbb{N}}$ with the good rate function $J$.*

**Proof of Theorem 2:** For $m \in \mathbb{N}$, let $\nu_n^{P_m}$ be a random element of $M_1(S, \sigma(P_m))$ with distribution $\pi^n_{P_m}(\mu_{n,P_m})$. It follows from Theorem 1 for finite state spaces that, for each fixed $m$, the sequence

$$\eta_n^{P_m} := \frac{n}{b_n}(\nu_n^{P_m} - \mu_{P_m})$$

satisfies an LDP on $M(S, \sigma(P_m))$ with speed $n/b_n^2$ and with the good rate function

$$I_{P_m}(\nu) = \begin{cases} \frac{1}{2}\sum_{A\in P_m} \frac{\nu(A)^2}{\mu(A)}, & \text{if } \sum_{A\in P_m}\nu(A) = 0, \\ +\infty, & \text{otherwise.} \end{cases} \tag{26}$$

This is well defined since $\mu(A) > 0$ for all $A \in P_m$ by assumption. We define

$$M(S, \sigma(P_m)) \ni \nu \mapsto \Psi_{P_m}(\nu) = \sum_{A\in P_m} \nu(A)\frac{\mu(\,\cdot\,\cap A)}{\mu(A)} \in M(S).$$

Note that $\Psi_{P_m}(\nu)(S) = 0$ for all $\nu \in M(S, \sigma(P_m))$ having $I_{P_m}(\nu) < \infty$. The maps $\Psi_{P_m}$ are measurable and continuous with respect to the projective limit topology. By the contraction principle, we obtain, for each $P_m$, that the sequence of random elements $\tilde\eta_n^{P_m} \equiv \Psi_{P_m}(\eta_n^{P_m})$ satisfies an LDP in the projective limit topology on $M(S)$ with the good rate function $J_{P_m} : M(S) \to [0, \infty]$, given by

$$J_{P_m}(\nu) = \begin{cases} I_{P_m}(\tilde\nu), & \text{if } \nu = \Psi_{P_m}(\tilde\nu) \text{ with } \tilde\nu \in M(S, \sigma(P_m)), \\ \infty, & \text{otherwise.} \end{cases} \tag{27}$$

We will show for the tree $(P_m)_{m\in\mathbb{N}}$ that $\{\mathbb{P}\circ(\tilde\eta_n^{P_m})^{-1}\}_{n\in\mathbb{N},m\in\mathbb{N}}$ is a $\{d_A,\ A \in \cup_{m\in\mathbb{N}} P_m\}$-exponentially good approximation of $\{\mathbb{P}\circ\eta_n^{-1}\}_{n\in\mathbb{N}}$ in $M_1(M(S))$. Here $\eta_n := (n/b_n)(\nu_n - \mu)$, where $\nu_n$ is a random element of $M_1(S)$ with distribution $\pi^n(\mu_n)$. It follows from the assumed exchangeability of the prior $\pi$ that, for each $m \in \mathbb{N}$ and $A \in \sigma(P_m)$, $\eta_n(A)$ has the same distribution as $\tilde\eta_n^{P_m}(A) = \eta_n^{P_m}(A)$. Thus we may construct $\eta_n$ and $(\tilde\eta_n^{P_m})_{m\in\mathbb{N}}$ on the same probability space simply by taking

$$\tilde\eta_n^{P_m} = \eta_{n,P_m},$$

where $\eta_{n,P_m}$ denotes the restriction of $\eta_n$ to $\sigma(P_m)$. It is immediate from this construction that $\mathbb{P}(d_A(\eta_n, \tilde\eta_n^{P_m}) > 0) = 0$ for all $n \in \mathbb{N}$ and $A \in \sigma(P_k)$, $k \le m$. Thus (23) is satisfied.

To apply Theorem 3, it remains to show that (24) and (25) are satisfied when we make the following identifications: $Y = M(S)$, $\mathcal{T}$ is the projective

limit topology, which is generated by the family of pseudometrics $\mathcal{D} = \{d_A : A \in \cup_{m\geq 1} P_m\}$, $J = I$, for $I$ defined in the statement of Theorem 2, $\mathcal{I} = (P_m)_{m\in\mathbb{N}}$, and $J_i = J_{P_m}$.
We first note that the definitions of $I_{P_m}(\nu)$ and $\Psi_{P_m}(\nu)$ extend from $\nu \in M(S, \sigma(P_m))$ to $\nu \in M(S)$ for every finite partition, $P_m$, of the nested sequence. Moreover, we may replace the collection of balls, $\{B(y, d, \delta)\}_{d\in\mathcal{D}', \delta>0}$ in (24) by any filterbase of neighbourhoods of $y$ converging to $y$ without changing the rate function. In particular, the sets

$$B_m(\nu, \delta) := \{\, \tilde{\nu} \in M(S) | \max_{A\in P_m} d_A(\tilde{\nu}, \nu) < \delta \,\}, \quad m \in \mathbb{N},\ \delta > 0,$$

form a neighbourhood filterbase of $\nu \in M(S)$. We can now rewrite (24) as

$$I(\nu) = \sup_{k\in\mathbb{N},\delta>0} \liminf_{m\in\mathbb{N}} \inf_{\tilde{\nu}\in B_k(\nu,\delta)} J_{P_m}(\tilde{\nu}). \tag{28}$$

To prove "$\geq$" in (28) observe that $\Psi_{P_k}(\nu) \in B_k(\nu, \delta)$ for all $\delta > 0$ since $d_A(\nu, \Psi_{P_k}(\nu)) = 0$ for all $A \in P_k$. Thus, it is enough to show that

$$I(\nu) \geq \sup_k \liminf_m J_{P_m}(\Psi_{P_k}(\nu)). \tag{29}$$

But $d\Psi_{P_k}(\nu)/d\mu$ is constant on each $A \in P_k$ by definition of $\Psi_{P_k}$, so in fact $J_{P_m}(\Psi_{P_k}(\nu)) = J_{P_k}(\Psi_{P_k}(\nu))$ for all $m \geq k$. The inequality (29) is now immediate from (27) and Lemma 1 below.
To prove "$\leq$" in (28) for a $\nu \in M(S)$, we consider two cases:
If $\nu \not\ll \mu$, then there exists a $B \in \mathcal{S}$ satisfying $\mu(B) = 0$ and $|\nu(B)| = b > 0$. Hence, given $\delta > 0$, we can find $k$ large enough and $A \in \sigma(P_k)$ such that $|\nu(A)| > b - \delta$ and $\mu(A) < \delta$. Now, for any $\lambda \in B_k(\nu, \delta)$, we have $|\lambda(A)| > b - 2\delta$, and so

$$I_{P_k}(\lambda) \geq \frac{(b-2\delta)^2}{2\delta} \quad \forall\, \lambda \in B_k(\nu, \delta).$$

By Lemma 1, we get for all $m > k$ that

$$I_{P_m}(\lambda) \geq \frac{(b-2\delta)^2}{2\delta} \quad \forall\, \lambda \in B_k(\nu, \delta).$$

Since $\delta > 0$ can be taken arbitrarily small, the right hand side of (28) is infinity (note that either $\lambda = \Psi_{P_m}(\lambda)$ and $J_{P_m}(\lambda) = I_{P_m}(\lambda)$ or $\lambda \neq \Psi_{P_m}(\tilde{\lambda})$ for any $\tilde{\lambda} \in M(S, \sigma(P_m))$ and $J_{P_m}(\lambda) = \infty$).
If $\nu \ll \mu$, then, given $r < I(\nu)$, there exists $P_k$ such that $r < I_{P_k}(\nu)$, see Lemma 1 below. Since $\mu(A) > 0$ for all $A \in P_k$ by the assumption of Theorem 2, $x \mapsto x^2/(\mu(A))$ is continuous for every $A \in P_k$, and there exists

$\delta > 0$ such that $r < I_{P_k}(\tilde{\nu})$ for all $\tilde{\nu} \in M(S)$ with $\max_{A \in P_k} d_A(\nu, \tilde{\nu}) < \delta$. We apply Lemma 1 again to obtain $I_{P_k}(\tilde{\nu}) \leq I_{P_m}(\tilde{\nu})$ for all refinements $P_m$ of $P_k$. Hence, $r < I_{P_m}(\tilde{\nu})$ for all $m > k$ and all $\tilde{\nu} \in M(S)$ with $\max_{A \in P_k} d_A(\nu, \tilde{\nu}) < \delta$, which implies "$\leq$" in (24).
Now it only remains to verify (25). Consider any $P_m$ and $\nu \in M(S)$ satisfying $J_{P_m}(\nu) < \infty$. Then $\nu = \Psi_{P_m}(\nu)$ and $S \ni s \mapsto \sum_{A \in P_m} (\nu(A)/\mu(A)) 1_A(s)$ is a density of $\nu$ with respect to $\mu$. Hence $I(\nu) = J_{P_m}(\nu)$ and (25) holds. This completes the proof of Theorem 2.

The proof of the following Lemma is an adaptation of the proofs of Proposition 15.5 and Corollary 15.7 in Georgii (1988).

**Lemma 1** *Let $\nu \in M(S)$ and $(P_m)_{m \in \mathbb{N}}$ a nested sequence of partitions of $S$ such that $\sigma(\cup_{m \geq 1} P_m) = \mathcal{S}$. Then $I_{P_m}(\nu)$ is an increasing function of $\sigma(P_m)$, and*

$$I(\nu) = \lim_{m \to \infty} I_{P_m}(\nu) = \sup_{m \in \mathbb{N}} I_{P_m}(\nu).$$

**Proof:** Consider $P \preceq P'$, thus $\sigma(P) \subseteq \sigma(P')$. Without loss of generality we can assume that $\nu \ll \mu$ on $\sigma(P')$. Let $f_{\sigma(P')} \geq 0$ denote the $\sigma(P')$-measurable density of $\nu$ with respect to $\mu$. Then $f_{\sigma(P)} = E_\mu(f_{\sigma(P')} | \sigma(P))$ is the density of $\nu$ with respect to $\mu$ on $\sigma(P)$. Using Jensen's inequality for conditional expectations, we obtain

$$I_P(\nu) = \frac{1}{2} \int_S E_\mu(f_{\sigma(P')} | \sigma(P))^2 \, d\mu \leq \frac{1}{2} \int_S E_\mu(f^2_{\sigma(P')} | \sigma(P)) \, d\mu = I_{P'}(\nu).$$

This proves the first claim of the lemma. By the same argument, we obtain

$$I(\nu) \geq \sup_{m \in \mathbb{N}} I_{P_m}(\nu) = \lim_{m \to \infty} I_{P_m}(\nu) \equiv b. \tag{30}$$

Thus it remains to prove $I(\nu) \leq b$ for all $\nu$ such that $b < \infty$. In this case, the restriction of $\nu$ to each $\sigma(P_m)$ is absolutely continuous with respect to the corresponding restriction of $\mu$, and so the densities $f_m = d\nu_{P_m}/d\mu_{P_m}$ exist. It is easy to see that $f_m$ is a martingale relative to $\mu$ and the filtration $(\sigma(P_m))_{m \in \mathbb{N}}$. Moreover,

$$E_\mu\Big(f_m \, 1_{\{f_m \geq k\}}\Big) \leq \frac{1}{k} E_\mu\Big(f_m^2 \, 1_{\{f_m \geq k\}}\Big) \leq \frac{b+1}{k}$$

for every $k > 0$, and so we obtain that $(f_m)_{m \in \mathbb{N}}$ is uniformly $\mu$-integrable. It is also Cauchy; the contrary is impossible because $(f_m)_{m \in \mathbb{N}}$ is a uniformly integrable martingale. Thus $f_m$ converges in $L_1(\mu)$-norm to the $\mathcal{S}$-measurable function, $f = d\mu/d\nu$.

Now for $N \in \mathbb{N}$ consider the functions $g = N \wedge \frac{1}{2}(f^2)$ and $g_m = N \wedge \frac{1}{2}(f_m^2)$. Since $f_m$ converges to $f$, $g_m$ converges to $g$. But since $0 \leq g_m, g \leq N$, this implies that $E_\mu(|g_m - g|) \to 0$ and therefore

$$E_\mu(g) = \lim_{m\to\infty} E_\mu(g_m) \leq \lim_{m\to\infty} I_{P_m}(\nu) = b.$$

Letting $N$ go to $\infty$ yields $I(\nu) \leq b$ which, together with the reverse inequality in (30), yields the claim of the lemma.

## 5 Examples

**Pólya trees:** We now exhibit a class of priors, namely *Pólya tree priors*, which satisfy the exchangeability assumption. A detailed discussion of the properties of Pólya trees together with some applications can be found in Mauldina, Sudderth, and Williams (1992) and Lavine (1992); we summarize the facts that our relevant to us below.

Let $S$ be a Polish space equipped with Borel $\sigma$-algebra $\mathcal{S}$ and consider a sequence of nested partitions, $P_0 = S$, $P_1 = (B_0, B_1)$, $P_2 = (B_{00}, B_{01}, B_{10}, B_{11})$ and so on, such that $\cup_{m\geq 0} P_m$ generates $\mathcal{S}$. Here, for each $m$ and $\epsilon = \epsilon_1 \cdots \epsilon_m \in \{0,1\}^m$, $(B_{\epsilon 0}, B_{\epsilon 1})$ is a measurable partition of $B_\epsilon$ (and $B_\emptyset = S$). Degenerate splits are allowed, so we could have $B_{\epsilon 0} = B_\epsilon$ and $B_{\epsilon 1} = \emptyset$. We denote by $\mathcal{P}$ the set of nested partitions $\{P_0, P_1, P_2, \ldots\}$.

A random probability distribution $\nu \in M_1(S)$ is said to have a Pólya tree distribution with parameter $(\mathcal{P}, \mathcal{A})$ if there exists a sequence of nested partitions $\mathcal{P}$ as above, non-negative real numbers $\mathcal{A} = (\alpha_0, \alpha_1, \alpha_{00}, \ldots)$ and random variables $\mathcal{C} = (C_0, C_{00}, C_{10}, \ldots)$ such that the following hold:

(i) the random variables in $\mathcal{C}$ are independent;

(ii) for every $\epsilon$, $C_{\epsilon 0}$ has the beta distribution, $\mathcal{B}(\alpha_{\epsilon 0}, \alpha_{\epsilon 1})$;

(iii) for every $m \in \mathbb{N}$ and $\epsilon = \epsilon_1 \cdots \epsilon_m$,

$$\nu(B_\epsilon) = \prod_{j=1}^{m} \left[ C_{\epsilon_1 \cdots \epsilon_j 0} \mathbf{1}_{\epsilon_j = 0} + (1 - C_{\epsilon_1 \cdots \epsilon_j 0}) \mathbf{1}_{\epsilon_j = 1} \right].$$

By identifying each $B_\epsilon$ with a node in a binary tree, we can interpret $\nu(B_\epsilon)$ as the (random) probability that a random walk visits the node when, from each node $B_\epsilon$, the walker chooses to move to $B_{\epsilon 0}$ with probability $C_{\epsilon 0}$ and to $B_{\epsilon 1}$ with probability $1 - C_{\epsilon 0}$.

The Pólya tree distribution is defined above for an infinite sequence of nested partitions but can be defined analogously for a finite sequence by terminating the process at some finite level $m$. If $\pi \in M_1(M_1(S))$ is a Pólya tree prior

with parameter $(\mathcal{P}, \mathcal{A})$, then its restriction $\pi_{P_m} \in M_1(M_1(S, \sigma(P_m)))$ to the partition $P_m$ is a Pólya tree prior with parameter $(\mathcal{P}^m, \mathcal{A}^m)$, where $\mathcal{P}^m = (P_0, \ldots, P_m)$ and $\mathcal{A}^m = (\alpha_\epsilon, \epsilon \in \{0,1\}^k, k = 1, \ldots, m)$. Moreover, if $\pi$ is a Pólya tree prior, then the posterior conditional on $X_1$ is also a Pólya tree distribution, but with parameters $(\mathcal{P}, \tilde{\mathcal{A}})$ where

$$\tilde{\alpha}_\epsilon = \alpha_\epsilon + \mathbf{1}_{X_1 \in B_\epsilon}, \quad \epsilon \in \{0,1\}^k, \ k = 1, 2, \ldots$$

The posterior corresponding to $\pi_{P_m}$ is described analogously. It is clear from this description that the Pólya tree prior, $\pi$, is exchangeable with respect to the nested partitions $\mathcal{P}$ involved in its definition.

**Dirichlet processes:** The *Dirichlet process* is obtained as a special case of the Pólya tree distribution by restricting $\mathcal{A}$ to be such that $\alpha_{\epsilon 0} + \alpha_{\epsilon 1} = \alpha_\epsilon$ for all $\epsilon \in \{0,1\}^k$, $k = 0, 1, \ldots$. In this case, we can find identify $\alpha_\epsilon$ with $\alpha(B_\epsilon)$, for some finite non-negative measure $\alpha$ on $(S, \mathcal{S})$. Thus, the Dirichlet process is also exchangeable. In fact, the Dirichlet process is exchangeable with respect to *any* sequence of nested partitions and it was shown in Corollary 2.1 of Doksum (1974) that it is essentially the only distribution with this property.
If the assumptions of Theorem 2 hold simultaneously for all partitions of $S$ and the prior is Dirichlet then the sequence of Dirichlet posteriors satisfies the MDP in the projective limit topology corresponding to any sequence of nested partitions. But the topology generated by all partitions is simply the $\tau$-topology on $M(S)$. Note however that it is not possible for $\mu_{n,P}$ to converge to $\mu_P$ for all partitions $P$ unless $S$ is countable. This is the price we need to pay to have the MDP in the $\tau$-topology.

## 6 Concluding remarks

We established a moderate deviation principle for Bayes posteriors on finite sample spaces. We extended the result to Polish spaces for priors which are exchangeable with respect to a nested sequence of partitions. It remains an open problem to identify distributions other than Pólya trees which satisfy the exchangeability requirement and also to explore the connections between exchangeability and seemingly related properties like neutrality and tailfreeness which have been studied extensively in the literature on Bayes' methods.

**Acknowledgements:** Research carried out in part while the first author was at Universität Bielefeld, Germany, and the second author at BRIMS,

Hewlett-Packard Labs, UK, and supported by ARC Grant 880 from the Anglo-German foundation. We would like to thank the referees for their careful reading of the paper and for their suggestions, which have improved its presentation.

## References

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

Borovkov, A. and A. Mogulskii (1978). Probabilities of large deviations in topological spaces I. *Siberian Math. J. 19*, 697–709.

Cox, D. (1993). An analysis of bayesian inference for nonparametric regression. *Ann. Statist. 21*, 903–923.

de Acosta, A. (1992). Moderate deviations and associated Laplace approximations for sums of independent random vectors. *Trans. Amer. Math. Soc. 329*(1), 357–375.

de Acosta, A. (1994a). On large deviations of empirical measures in the $\tau$-topology. *J. Appl. Probab. 31A*, 41–47. (special volume).

de Acosta, A. (1994b). Projective systems in large deviation theory II: some applications. In J. Hoffman-Jørgensen, J. Kuelbs, and M. Marcus (Eds.), *Probability in Banach Spaces, 9*, Volume 35 of *Progress in Probability*, pp. 241–250. Birkhäuser, Boston.

Dembo, A. and O. Zeitouni (1998). *Large Deviations Techniques and Applications*. New York: Springer.

Diaconis, P. and D. Freedman (1986). On the consistency of bayes estimates. *Ann. Statist. 14*(1), 1–26.

Doksum, K. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Ann. Probab. 2*(2), 183–201.

Dupuis, P. and R. S. Ellis (1997). *A Weak Convergence Approach to the Theory of Large Deviations*. New York: John Wiley & Sons, Inc.

Eichelsbacher, P. and U. Schmock (1998). Exponential approximations in completely regular topological spaces and extensions of Sanov's theorem. *Stochastic Process. Appl. 77*(2), 233–251.

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

Freedman, D. A. (1999). On the Bernstein-von Mises theorem with infinite dimensional parameters. *Ann. Statist. 27*(4), 1119–1140.

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

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

Ganesh, A. J. and N. O'Connell (2000). A large deviation principle for Dirichlet posteriors. *Bernoulli 6*(6), 1021–1034.

Georgii, H.-O. (1988). *Gibbs Measures and Phase Transitions*. Berlin: Walter de Gruyter.

Ghosal, S., J. K. Ghosh, and R. V. Ramamoorthi (1998). Consistency issues in bayesian nonparametrics. In S. Ghosh (Ed.), *Asymptotics, Nonparametrics and Time Series: A Tribute to Madan Lal Puri*. Marcel Dekker.

Ghosal, S., J. K. Ghosh, and A. W. van der Vaart (2000). Convergence rates of posterior distributions. *Ann. Statist. 28*(2), 501–531.

Ibragimov, I. A. and R. Z. Has'minskii (1981). *Statistical Estimation: Asymptotic Theory*. New York: Springer-Verlag.

Johnson, R. (1967). An asymptotic expansion for posterior distributions. *Ann. Math. Statist. 38*, 1899–1907.

Johnson, R. (1970). Asymptotic expansions associated with posterior distributions. *Ann. Math. Statist. 41*, 851–864.

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

Le Cam, L. M. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist. 22*, 38–53.

Luenberger, D. G. (1965). *Introduction to linear and nonlinear programming*. New-York: Addison-Wesley.

Lynch, M. and M. Sethuraman (1987). Large deviations for processes with independent increments. *Ann. Probab. 15*, 610–627.

Mauldina, R. D., W. D. Sudderth, and S. C. Williams (1992). Pólya trees and random distributions. *Ann. Statist. 20*, 1203–1221.

Mogulskii, A. (1976). Large deviations for trajectories of multidimensional random walks. *Theory Probab. Appl. 21*, 300–315.

Sanov, I. (1961). On the probability of large deviations of random variables. *Selected Transl. Math. Statist. and Prob. I*, 214–244.

Schwartz, L. (1965). On Bayes procedures. *Z. Wahrscheinlichkeitstheorie verw. Geb. 4*, 10–26.

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

Wasserman, L. (1999). Asymptotic properties of nonparametric Bayesian procedures. In *Practical nonparametric and semiparametric Bayesian statistics*, Volume 133 of *Lecture Notes in Statist.*, pp. 293–304. Springer.

Peter Eichelsbacher
Fakultät für Mathematik, Ruhr-Universität Bochum, NA 3/68
D–44780 Bochum, Germany.
E-mail: peter.eichelsbacher@ruhr-uni-bochum.de.

Ayalvadi Ganesh
Microsoft Research, St. George House, 1 Guildhall Street
Cambridge CB2 3NH, UK.
E-mail: ajg@microsoft.com.