Moderate deviations for Bayes posteriors

Peter Eichelsbacher* and Ayalvadi Ganesh †
University of Bielefeld and BRIMS, Bristol

Abstract

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

Key words: moderate deviations principle, Bayes posteriors, Fisher-information, exponentially good approximations

1 Introduction

Bayesian methods have become increasingly popular in statistics in recent years and there has been renewed interest in non-parametric Bayesian inference. In turn, this has stimulated much recent work on the consistency of these inference procedures. Freedman [12] and Diaconis and Freedman [8] 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 [27] 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 *et al.* [4] show

*Fakultät für Mathematik, Universität Bielefeld, Postfach 100131, D–33501 Bielefeld, Germany

†BRIMS, Hewlett-Packard Labs, Bristol, Filton Road, Stoke Gifford, Bristol BS12 QZ, UK. Research supported by ARC Grant 880 from the Anglo-German foundation.

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 $\int(\sqrt{f}-\sqrt{g})^2 d\mu$. As a distance between probability distribution, it does not depend on the choice of the reference measure.) Rates of convergence have been investigated by Schwartz [27], Johnson [20, 21], Ghosal *et al.* [18] and Shen and Wasserman [28]. Recent research on the consistency of Bayes methods is reviewed by Ghosal *et al.* [17] and Wasserman [29].

There has been comparatively little work on more refined asymptotics for Bayes posteriors. Results on asymptotic normality of the posterior are few and mixed. Le Cam [22] and Ibragimov and Has'minskii [19] 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. Non-parametric problems are studied by Cox [6] and Diaconis and Freedman [9]. Large deviation asymptotics have been studied by Fu and Kass [13] and Ganesh and O'Connell [14, 15]; also see Lynch and Sethuraman [24], 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 begin by deriving the result on a finite sample space and extend it to general state spaces using the concept of exponentially good approximations introduced by Eichelsbacher and Schmock [11].

Let $(\mathcal{X},\tau)$ be a Hausdorff topological space with Borel σ-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 semi-continuous 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 λ_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° 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 Ω is a finite set and denote by $M_1(\Omega)$ (respectively, $M_b(\Omega)$) the space of probability measures (respectively, finite signed measures) on Ω. Consider a sequence of independent random variables $(X_k)_{k\in\mathbb{N}}$ taking values in Ω, with common law μ. Denote by

$L_n : \Omega \to M_1(\Omega)$ 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 δ_{X_k} denotes unit mass at X_k. Sanov's theorem [26] 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

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

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 [25] 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 by de Acosta in [1]. The moderate deviation behaviour for $(\frac{n}{b_n}(L_n - \mu))_{n \in \mathbb{N}}$ was first considered in [5], the "full" moderate deviation principle in [2].

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 π. 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 Ω) such that ν_n has distribution $\pi^n(L_n)$ for each $n \geq 1$. Ganesh and O'Connell [14] showed that, on the set $\{L_n \to \mu\}$, for any μ 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

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

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 μ 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 general measurable 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 μ is that the posterior concentrates in weak neighbourhoods of the true distribution, μ; but Freedman [12] exhibited a prior on the natural numbers for which the posterior failed to concentrate in a weak neighbourhood of the true distribution.

2 The finite case

Let Ω be a finite set, and let $M_1(\Omega)$ denote the space of probability measures on Ω. Suppose $X_1, X_2, \ldots$ are i.i.d. Ω–valued random variables 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 $\operatorname{supp} \pi$. 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 π:

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

if the denominator is non-zero (which happens π-almost surely); π^n is undefined at μ_n if the denominator is zero. When defined, $\pi^n(\mu_n)$ denotes the posterior distribution, conditional on the observations $X_1, \ldots, X_n$ having empirical distribution $\mu_n \in M_1^n(\Omega)$ (this is immediate from Bayes' formula).

Definition: Let λ 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^Ω with probability distributions on Ω. We say that $\mu \in \operatorname{supp} \pi$ is a *regular point* of the support of π if there is a neighbourhood of μ (in total variation distance) in S^Ω on which π is absolutely continuous with respect to λ and $d\pi/d\lambda$, the density of π with respect to Lebesgue measure, is bounded away from zero and infinity on this neighbourhood.

Let ν be a signed measure and f a real-valued function on Ω. We define

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

If μ, ν are probability measures on Ω, then the total variation distance between μ and ν 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 μ be a regular point of the support of π 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 ν_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 [7]. Here, $\langle f, \eta \rangle$ denotes $\sum_{x \in \Omega} f(x)\eta(x)$. Observe from (3) and the definition of η_n that

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

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

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

The Kuhn-Tucker conditions (see [23, Section 10.6]) for the optimality of a feasible solution λ_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

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

where the constant α_n is chosen so that $\sum_{x\in\Omega}\lambda_n(x) = 1$. Now f is fixed, $b_n/nto0$ 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 ν (see, for example, [10, Lemma 1.4.3]) and the constraints are linear in ν, so the Kuhn-Tucker conditions are *sufficient* for the optimality of λ_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 μ_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 constants. 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,

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

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

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

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 $\alpha, \delta > 0$,

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

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 μ is a regular point of the support of π and the definition of regularity that, if δ 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 λ denotes Lebesgue measure on the $|\Omega|$-simplex, S^Ω.

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

$$\begin{aligned}
g(\nu) &:= b_n\langle f, \nu - \lambda_n\rangle - nH(\mu_n|\nu) + nH(\mu_n|\lambda_n) \\
&= b_n\langle f, \nu - \lambda_n\rangle + n\sum_{x\in\Omega} \mu_n \log\left[1 + \frac{\nu(x) - \lambda_n(x)}{\lambda_n(x)}\right]
\end{aligned}$$

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

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

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

for a generic constant c that does not depend on n. Here, $\mathrm{vol}\ (\mathcal{A})$ denotes the volume of $\mathcal{A}$ as a subset of $\mathbb{R}^{|\Omega|-1}$.

Observe that g is a concave function of ν, $g(\lambda_n) = 0$ and λ_n was defined so that g attains its maximum over $M_1(\Omega)$ at λ_n. Thus,

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

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

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

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

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

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

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

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

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

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

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

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 μ_n rather than λ_n, and the Taylor expansion of $H(\mu_n|\nu)$ is also done around μ_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 [7, 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

$$f(x) = \frac{\nu(x)}{\mu(x)} + c \quad \forall\, x \in \Omega,$$

where c is an arbitrary constant. Substituting this in (22), we find that the value of $I(\nu)$ does not depend on c, and is given by

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

This completes the proof the theorem.

3 The general case

Now let $M_1(S)$ and $M(S)$, respectively, denote the set of probability measures (finite signed measures) on a measurable space $(S, \mathcal{S})$. The τ-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 σ-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 [3, Lemma 2.1], the level sets of I are τ-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 [7]).

For *large deviations* Ganesh and O'Connell pointed out in [15] 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. Nevertheless on the moderate scale, defined by (1), we obtain a result for *arbitrary measurable state spaces* S, if the prior is *exchangeable with respect to finite projections* and a regularity assumption on the support of the prior π is fulfilled. The Dirichlet prior is exchangeable for example.

By a partition P of S we mean a finite collection $\{A_1, \dots, A_n\} \subset \mathcal{S}$ of disjoint sets whose union is S. Let $\mathcal{P}$ be the set of all these partitions of S. For $P \in \mathcal{P}$ let $\sigma(P)$ denote the σ-algebra generated by P. Fix $\mu \in M_1(S)$ and define $\mathcal{P}_\mu = \{\, P \in \mathcal{P} \mid \mu(A) > 0 \text{ for all } A \in P \,\}$. Note that $(\mathcal{P}_\mu, \preceq)$ is a directed set, provided that $P \preceq P'$ for $P, P' \in \mathcal{P}_\mu$ means $P \subset \sigma(P')$, that is, P' is a refinement of P. Given $P \in \mathcal{P}$, the empirical measure $L_n(\omega)$ can be considered as a probability measure on the measurable space $(S, \sigma(P))$. Note that $M_1(S, \sigma(P))$ can be identified with $\{\, (x_A)_{A \in P} \in [0,1]^P \mid \sum_{A \in P} x_A = 1 \,\}$ and that the τ-topology and the σ-algebra induced on $M_1(S, \sigma(P))$ by the maps $M_1(S, \sigma(P)) \ni \nu \mapsto \nu(A)$ with $A \in P$ coincides with the natural topology and Borel σ-algebra of this subset of $[0,1]^P$. The same is true for $M(S, \sigma(S))$.

The restriction of a measure $\mu \in M_1(S)$ to the σ-algebra $\sigma(P)$ with $P \in \mathcal{P}_\mu$ is denoted by μ_P. For a prior $\pi \in M_1(M_1(S))$ we denote by π_P the corresponding element in $M_1(M_1(S, \sigma(P)))$, thus the restriction of π to the Borel σ-algebra $\mathcal{B}(M_1(S, \sigma(P)))$. For $\mu_n \in M_1^n(S)$ and $\pi^n(\mu_n) \in M_1(M_1(S))$ the elements $\mu_{n,P}$ and $\pi_P^n(\mu_n)$ are defined analogously. Now we can define

Definition: A prior measure $\pi \in M_1(M_1(S))$ is *exchangeable with respect to finite projections* if for every $P \in \mathcal{P}_\mu$ we have

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

Let ν be a signed measure on S. We define

$$\|\nu\|_1 := \sup\{\int g \, d\nu : \ g \in C_b(S, \mathrm{I\!R}) \text{ with } \|g\|_\infty \le 1\}$$

and $\|g\|_\infty := \sup_{x \in \mathrm{I\!R}} |g(x)|$, where $C_b(S, \mathrm{I\!R})$ denotes the set of bounded and continuous real valued functions. The total variation distance between $\mu, \nu \in M_1(S)$ is $d_{TV}(\mu, \nu) = \frac{1}{2}\|\mu - \nu\|_1$. A probability measure $\mu \in M_1(S)$ is now called a regular point of the support of π, whenever μ_P is a regular point of the support of π_P for every $P \in \mathcal{P}$.

Theorem 2 *Let $(b_n)_{n\in\mathbb{N}}$ be a positive sequence satisfying (1) and let μ be a regular point of the support of π, where π is assumed to be an exchangeable prior with respect to finite projections. Suppose $x \in S^{\mathbb{N}}$ is such that the sequence $\mu_n = \sum_{i=1}^n \delta_{x_i}/n$ converges in the τ-topology to μ such that $\|\mu_{n,P} - \mu_P\|_1 = o(b_n/n)$ for every $P \in \mathcal{P}$. 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 speed n/b_n^2 and with convex good rate function I given by*

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

Proving the Theorem we will apply the concept of exponential approximations in completely regular topological spaces introduced in [11]. 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 $d(x,y) \neq 0$. Note that the τ-topology turns $M_1(S)$ as well as $M(S)$ into a completely regular space with the separating family $\mathcal{D} := \{d_A\}_{A\in\mathcal{S}}$ of pseudo-metrics, where $d_A(\mu,\nu) := |\mu(A) - \nu(A)|$ for all $\mu, \nu \in M_1(S)(M(S))$. Note that for a completely regular space the collections of balls $B(y,d,\delta) := \{x \in Y | d(x,y) < \delta\}$ with $y \in Y, d \in \mathcal{D}$ and $\delta > 0$ is a sub-basis of the topology $\mathcal{T}$. Let $\mathcal{D}'$ be the smallest family of pseudo-metrics on Y which contains $\mathcal{D}$ and is closed with respect to maxima. Then the collection of all balls $B(y,d,\delta)$ with $y \in T, d \in \mathcal{D}'$ and $\delta > 0$ is a basis of $\mathcal{T}$.

A collection $\{\mu_{n,i}\}_{n\in\mathbb{N},i\in I} \subset M_1(Y)$ with $(I, \preceq)$ a nonempty directed set is called $\mathcal{D}$-exponentially 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 $i \in I$ there exists a probability measure $\nu_{d,n,i}$ on a σ-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

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

for every $\delta > 0$. If the measures $\nu_{d,\varepsilon,i}$ and the σ-algebras $\mathcal{Y}_{d,\varepsilon,i}$ 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 $\limsup_{i\in I} a_i = \inf_{j\in I} \sup_{i\in I, i\preceq j} a_i$ and $\liminf_{i\in I} a_i = \sup_{j\in I} \inf_{i\in I, i\preceq j} a_i$. We will apply the following Theorem (Theorem 1.6 in [11]):

Theorem 3 *If $\{\mu_{n,i}\}_{n\in\mathbb{N},i\in I} \subset M_1(Y)$ is a $\mathcal{D}$-exponentially good approximation of $\{\tilde{\mu}_n\}_{n\in\mathbb{N}}$ and if, for every $i \in I$, 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*

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

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

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

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

Proof of Theorem 2: We go along the arguments of Example 2.4 in [11]: Given $P \in \mathcal{P}$ the empirical measure $\mu_n \in M_1^n(S)$ can be considered as a probability measure on the measurable space $(S, \sigma(P))$. It follows from Theorem 1 for finite state spaces that for every partition $P \in \mathcal{P}$ the sequence

$$\eta_n \equiv \frac{n}{b_n}\Big(\pi^n(\mu_{n,P}) - \mu_P\Big) = \frac{n}{b_n}\Big(\pi_P^n(\mu_n) - \mu_P\Big)$$

satisfies a LDP. The corresponding good rate function is given by

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

For every $P \in \mathcal{P}_\mu$ define

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

Notice that $\Psi_P(\nu)(S) = 0$ for $\nu \in \{\tilde{\nu} \in M(S,\sigma(S)) : I_P(\tilde{\nu}) < \infty\}$ with $P \in \mathcal{P}_\mu$. These maps are measurable and continuous with respect to the τ-topologies. By the contraction principle the measures $\{\mathbb{P} \circ \eta_{n,P}^{-1}\}_{n\in\mathbb{N}} \subset M_1(M(S))$ with $\eta_{n,P} \equiv \Psi_P(\eta_n)$ satisfy a LDP in the τ-topology on $M(S)$ with the good rate function $J_P : M(S) \to [0,\infty]$, given by

$$J_P(\nu) \equiv \begin{cases} I_P(\tilde{\nu}), & \text{if } \nu = \Psi_P(\tilde{\nu}) \text{ with } \tilde{\nu} \in M(S,\sigma(P)), \\ \infty, & \text{otherwise.} \end{cases}$$

We will show that $\{\mathbb{P} \circ \eta_{n,P}^{-1}\}_{n\in\mathbb{N},P\in\mathcal{P}_\mu}$ is a $\{d_A\}_{A\in\mathcal{S}}$-exponentially good approximation of $\{\mathbb{P} \circ \eta_n^{-1}\}_{n\in\mathbb{N}}$ in $M_1(M(S))$. If $A \in \mathcal{S}$ satisfies $\mu(A) \in \{0,1\}$, then $\mathbb{P}(d_A(\eta_n, \eta_{n,P}) > 0) = 0$ for all $n \in \mathbb{N}$ and $P \in \mathcal{P}_\mu$; if $\mu(A) \in (0,1)$, then this is true at least for all $P \in \mathcal{P}_\mu$ which are refinements of $\{A, A^{\mathrm{c}}\}$.

To apply Theorem 3, it remains to show that (24) and (25) are satisfied. First note that the definitions of $I_P(\nu)$ and $\Psi_P(\nu)$ extend from $\nu \in M(S, \sigma(P))$ to $\nu \in M(S)$ for every $P \in \mathcal{P}$ or $P \in \mathcal{P}_\mu$, respectively. Note that in (24) we may replace the collection $\{B(y, d, \delta)\}_{d\in\mathcal{D}',\delta>0}$ of balls by any filterbase of neighbourhoods of y converging to y without changing $J(\cdot)$. Here all sets $\{\,\tilde\nu \in M(S) |\max_{A\in P} d_A(\tilde\nu, \nu) < \delta\,\}$ with $P \in \mathcal{P}$ and $\delta > 0$ is a neighbourhood filterbase of $\nu \in M(S)$.

First we prove (25): consider any $P \in \mathcal{P}_\mu$ and $\nu \in M(S)$ satisfying $J_P(\nu) < \infty$. Then $\nu = \Psi_P(\nu)$ and $S \ni s \mapsto \sum_{A\in P}(\nu(A)/\mu(A))1_A(s)$ is a density of ν with respect to μ, hence $J(\nu) = J_P(\nu)$ and (25) is proved.

To prove "$\geq$" in (24) it suffices to consider $\nu \in M(S)$ satisfying $J(\nu) < \infty$. Note that $\nu \ll \mu$ in this case. Given $P \in \mathcal{P}$, we can join all μ-null sets of P to one $A \in P$ with $\mu(A) > 0$ to obtain a $P' \in \mathcal{P}_\mu$ with $P' \subset \sigma(P)$. Define $\tilde\nu = \Psi_{P'}(\nu)$. Since $\nu \ll \mu$, it follows that $d_A(\tilde\nu, \nu) = 0$ for all $A \in P$. By Lemma 1 every refinement $P'' \in \mathcal{P}_\mu$ of P' satisfies $J_{P''}(\tilde\nu) = I_{P''}(\tilde\nu) \leq I(\nu)$. This shows that the right-hand side of (24) is bounded by $I(\nu)$.

To prove "$\leq$" in (24) for a $\nu \in M(S)$, we consider two cases:

If $\nu \not\ll \mu$, then there exists an $A \in \mathcal{S}$ satisfying $\mu(A) = 0$ and $\nu(A) > 0$. Hence, $B(\nu, d_A, \nu(A)) \cap \Psi_P(M(S, \sigma(P))) = \emptyset$ for every $P \in \mathcal{P}_\mu$ because $\Psi_P(\tilde\nu)(A) = 0$ for all $\tilde\nu \in M(S, \sigma(P))$. Therefore, the right-hand side of (3.1) equals infinity.

If $\nu \ll \mu$, then, given $r < I(\nu)$, there exists $P \in \mathcal{P}$ such that $r < I_P(\nu)$, see Lemma 2. By joining all μ-null sets of P to one $A \in P$ with $\mu(A) > 0$, if necessary, we may assume that $P \in \mathcal{P}_\mu$. Since $[0,1] \ni x \mapsto x^2/(\mu(A))$ is continuous for every $A \in P$, there exists $\delta > 0$ such that $r < I_P(\tilde\nu)$ for all $\tilde\nu \in M(S)$ with $\max_{A\in P} d_A(\nu, \tilde\nu) < \delta$. Again we apply Lemma 1 to obtain $I_P(\tilde\nu) \leq I_{P'}(\tilde\nu)$ for all refinements $P' \in \mathcal{P}_\mu$ of P. Hence, $r < I_{P'}(\tilde\nu)$ for all refinements $P' \in \mathcal{P}_\mu$ of P and all $\tilde\nu \in M(S)$ with $\max_{A\in P} d_A(\nu, \tilde\nu) < \delta$, which implies "$\leq$" in (24). Thus Theorem 2 is proved.

The proofs of the following Lemmata are an adaption of the proofs of Proposition 15.5 and Corollary 15.7 in [16].

Lemma 1 *Let $\nu \in M(S)$ and $\sigma(P)$ the σ-algebra generated by $P \in \mathcal{P}$. Then $I_P(\nu)$ is an increasing function of $\sigma(P)$.*

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

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

The Lemma is proved.

Define for a σ-subalgebra $\mathcal{A}$ of $\mathcal{S}$ the Fisher information

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

where $f_{\mathcal{A}}$ is any Radon-Nikodym density of $\nu|_{\mathcal{A}}$ relative to $\mu|_{\mathcal{A}}$.

Lemma 2 *Let $\mathcal{A}$ be a σ-subalgebra of $\mathcal{S}$. Then*

$$I_{\mathcal{A}}(\nu) = \sup_{P \in \mathcal{P}_{\mathcal{A}}} I_P(\nu),$$

where $\mathcal{P}_{\mathcal{A}}$ denotes the collection of all finite $\mathcal{A}$-measurable partitions of S.

Proof: First of all, we prove the following: let $(\mathcal{A}_a)_{a\in D}$ be an increasing net of σ-subalgebras of $\mathcal{S}$ and $\mathcal{A} = \sigma\left(\bigcup_{a\in D} \mathcal{A}_a\right)$ the smalest σ-algebra that contains all $\mathcal{A}_a$'s. Then

$$I_{\mathcal{A}}(\nu) = \lim_{a\in D} I_{\mathcal{A}_a}(\nu) = \sup_{a\in D} I_{\mathcal{A}_a}(\nu). \tag{27}$$

Lemma 1 shows that

$$I_{\mathcal{A}}(\nu) \geq \lim_{a\in D} I_{\mathcal{A}_a}(\nu) = \sup_{a\in D} I_{\mathcal{A}_a}(\nu) \equiv b.$$

Thus it remains to prove $I_{\mathcal{A}}(\nu) \leq b$ with $b < \infty$. Since ν is μ-continuous on each $\mathcal{A}_a$, there is a net $(f_a)_{a\in D}$ of $\mathcal{A}_a$-measurable densities $\frac{d\nu}{d\mu}|_{\mathcal{A}_a} = f_a$.

The net is a martingal relative to μ, which converges in $L_1(\mu)$-norm to some $\mathcal{A}$-measurable function $f \geq 0$: Since

$$E_\mu\left(f_a\, 1_{\{f_a \geq k\}}\right) \leq \frac{1}{k} E_\mu\left(f_a^2\, 1_{\{f_a \geq k\}}\right) \leq \frac{2(b+1)}{k}$$

for every $k > 0$, we obtain that $(f_a)_{a \in D}$ is uniformly μ-integrable. Moreover it is Cauchy, since the contrary is impossible because $(f_a)_{a \in D}$ is a uniformly integrable martingal. Thus the $L_1(\mu)$-convergence to some f is checked.

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

$$E_\mu(g) = \lim_{a \in D} E_\mu(g_a) \leq \lim_{a \in D} I_{\mathcal{A}_a}(\nu) = b.$$

Hence (27) is checked.

For each $P \in \mathcal{P}_\mathcal{A}$ again let $\sigma(P)$ be the finite σ-algebra which is generated by P. With the partial order on $\mathcal{P}_\mathcal{A}$, $(\sigma(P))_{P \in \mathcal{P}_\mathcal{A}}$ is an increasing net and since $\mathcal{A} = \sigma(\sigma(P) : P \in \mathcal{P}_\mathcal{A})$, the result follows from (27).

References

[1] A. de Acosta, *Moderate deviations and associated Laplace approximations for sums of independent random vectors*, Trans. Amer. Math. Soc., vol 329, no 1, 357–375, 1992.

[2] A. de Acosta, *Projective systems in large deviation theory II: some applications*, Probability in Banach Spaces 9, J. Hoffman-Jørgensen and J. Kuelbs and M. B. Marcus, Birkhäuser, Boston, Progress in Probability, vol 35, 241–250, 1994.

[3] A. de Acosta, *On large deviations of empirical measures in the τ-topology*, J. Appl. Probab., vol 31A (special volume), 41–47, 1994.

[4] A. Barron, M. J. Schervish and L. Wasserman, *The consistency of posterior distributions in nonparametric problems*, Technical report, Dept. of Statistics, Carnegie-Mellon University, 1997.

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

[6] D. Cox, *An analysis of Bayesian inference for nonparametric regression*, Ann. Statist., vol 21, 903–923, 1993.

[7] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, Springer, New York, 1998.

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

[9] P. Diaconis and D. Freedman, *On the Bernstein-von Mises theorem with infinite dimensional parameters*, unpublished manuscript.

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

[11] P. Eichelsbacher and U. Schmock, *Exponential approximations in completely regular topological spaces and extensions of Sanov's theorem*, Stoch. Proc. and their Appl., vol 77, no 2, 233–251, 1998.

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

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

[14] A. J. Ganesh and Neil O'Connell, *An inverse of Sanov's theorem*, to appear in Stat. Probab. Letters, 1999.

[15] A. J. Ganesh and Neil O'Connell, *A large deviations principle for Dirichlet process posteriors*, BRIMS, HP Laboratories Bristol, Technical report 98-05, 1998.

[16] H.-O. Georgii, *Gibbs Measures and Phase Transitions*, Walter de Gruyter, Berlin, 1988.

[17] S. Ghosal, J. K. Ghosh and R. V. Ramamoorthi, *Consistency issues in Bayesian nonparametrics*, in Asymptotics, Nonparametrics and Time Series: A Tribute to Madan Lal Puri (Subir Ghosh ed.), Marcel Dekker, 1998.

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

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

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

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

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

[23] D.G. Luenberger, *Introduction to linear and nonlinear programming*, Addison-Wesley, 1965.

[24] J. Lynch and J. Sethuraman, *Large deviations for processes with independent increments*, Annals of Probab. 15: 610–627, 1987.

[25] A.A. Mogulskii, *Large deviations for trajectories of multidimensional random walks*, Theory Probab. Appl., vol 21, 300–315, 1976.

[26] I. N. Sanov, *On the probability of large deviations of random variables*, Selected Transl. Math. Statist. and Prob., I, 214–244, 1961.

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

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

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

Peter Eichelsbacher, Fakultät für Mathematik, Universität Bielefeld, Postfach 100131, D–33501 Bielefeld, Germany