The Annals of Applied Probability
2025, Vol. 35, No. 4, 2967–2982
https://doi.org/10.1214/25-AAP2202

UNCONDITIONAL LARGE DEVIATION PRINCIPLES FOR DIRICHLET POSTERIOR AND BAYESIAN BOOTSTRAP

BY SHUI FENG[a]

Department of Mathematics and Statistics, McMaster University, [a]*shuifeng@mcmaster.ca*

The unconditional or annealed large deviation principles are established for the Dirichlet posterior and the Bayesian bootstrap. The rate functions are identified explicitly, which provide new measurements of divergence between probabilities. As applications, we study the asymptotic efficiencies of the Dirichlet posterior mean and the Bayesian bootstrap mean.

1. Introduction. Let S be a compact Polish space with the Borel σ-field $\mathcal{S}$. Denote by $C(S)$ and $B(S)$ the spaces of continuous functions and bounded measurable functions on S, respectively. The space of probability measures on $(S, \mathcal{S})$ is denoted by $M_1(S)$ which is equipped with the weak topology. The space $M_1(M_1(S))$ is defined similarly. To avoid triviality, we assume that S contains at least two elements. For any f in $C(S)$ and μ in $M_1(S)$, the integral of f with respect to μ is denoted by $\langle \mu, f \rangle$.

For any $\theta > 0$, let $U_1, U_2, \ldots$ be i.i.d. with Beta$(1, \theta)$ distribution. Independently, let $\xi_1, \xi_2, \ldots$ be i.i.d. with common distribution ν_0 in $M_1(S)$. The Dirichlet process [8] on $(S, \mathcal{S})$ with mean distribution ν_0 and concentration parameter θ, denoted by Π_{θ,ν_0}, is the law of the random measure

$$\Xi_{\theta,\nu_0} = \sum_{i=1}^{\infty} V_i \delta_{\xi_i}, \tag{1.1}$$

where $V_1 = U_1$, $V_i = (1 - U_1) \cdots (1 - U_{i-1}) U_i$, $i \geq 2$ and δ_ξ denotes the Dirac measure at ξ. Given observations $X_1, \ldots, X_n$ from the Dirichlet process, set

$$\nu_n = \frac{\theta}{\theta + n} \nu_0 + \frac{n}{\theta + n} \sum_{i=1}^{n} \delta_{X_i}.$$

Then a version of the corresponding Dirichlet posterior distribution is given by $\Pi_{\theta+n,\nu_n}$, which is the law of the random measure

$$\Xi^{(n)}_{\theta,\nu_0} = U^{(n)} \Xi_{\theta,\nu_0} + \big(1 - U^{(n)}\big) \sum_{i=1}^{n} W_{n,i} \delta_{X_i} \tag{1.2}$$

where $U^{(n)}$ is Beta(θ, n) distributed, $(W_{n,1}, \ldots, W_{n,n})$ has Dirichlet$(1, \ldots, 1)$ distribution, and all random variables appearing on the right-hand side are independent given $X_1, \ldots, X_n$. The Bayesian bootstrap corresponds to $\theta = 0$.

For any ν_1 in $M_1(S)$, let ν_1^∞ denote the infinite product measure of ν_1. Given that $X_1, X_2, \ldots$ are i.i.d. with common distribution ν_1, the Dirichlet posterior is strongly consistent in the sense that, with ν_1^∞ probability one, $\Pi_{\theta+n,\nu_n}$ converges to δ_{ν_1} in $M_1(M_1(S))$ as n tends to infinity. In other words, for almost all observations under ν_1^∞, the random measure $\Xi^{(n)}_{\theta,\nu_0}$ converges in probability to ν_1 as n tends to infinity.

Received August 2023.
MSC2020 subject classifications. Primary 60G57; secondary 62F15.
Key words and phrases. Bayesian bootstrap, Dirichlet process, Dirichlet posterior, large deviation, relative entropy.

The study of large sample asymptotic behaviours for the Dirichlet posterior and the Bayesian bootstrap is part of frequentist Bayesian inference. Consistency and normal fluctuations (Bernstein–von Mises theorem) have been the focus of active research over the years [13, 16, 17, 21, 24]. In [11] a conditional large deviation principle was established for the Dirichlet posterior. More specifically, for any measurable set A with interior A° and closure $\bar{A}$, one has for almost all observations under ν_1^∞

$$-\inf_{\mu\in A^\circ} I(\mu) \le \varliminf_{n\to\infty} \frac{1}{n}\ln \Pi_{\theta+n,\nu_n}(A^\circ) \le \varlimsup_{n\to\infty} \frac{1}{n}\ln \Pi_{\theta+n,\nu_n}(\bar{A}) \le -\inf_{\mu\in\bar{A}} I(\mu),$$

where $I(\mu)$ is a good rate function, and, for ν_1 with support S, is equal to the Kullback–Leibler divergence or relative entropy

$$H(\nu_1|\mu) = \sup\bigl\{\langle \nu_1, f\rangle - \ln\langle \mu, e^f\rangle : f \in C(S)\bigr\}. \tag{1.3}$$

A comparison with the frequentist inference is revealing. Recall that for i.i.d. sequence $X_1, X_2, \ldots$ with common distribution ν_1, the empirical distribution

$$\mathcal{L}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$$

converges almost surely to ν_1 as n tends to infinity. In addition, the functional central limit theorem holds for $\mathcal{L}_n$. Thus $\Xi^{(n)}_{\theta,\nu_0}$ and $\mathcal{L}_n$ have similar large sample behaviours under ν_1^∞ in terms of law of large numbers and normal fluctuations. But the more refined large deviation results reveal the fundamental difference between the two. By Sanov's theorem [23], $\mathcal{L}_n$ satisfies a large deviation principle with the good rate function $H(\mu|\nu_1)$ while the conditional large deviation rate function for $\Xi^{(n)}_{\theta,\nu_0}$ has the reversed form $H(\nu_1|\mu)$.

The large deviation principle for $\mathcal{L}_n$ is a frequentist result while the conditional large deviation result for $\Xi^{(n)}_{\theta,\nu_0}$ is purely Bayesian. The objective of this paper is to understand the large deviation behaviour of the combined frequentist and Bayesian structure of the Dirichlet posterior. More specifically, we are interested in the large deviation principle for the law of $\Xi^{(n)}_{\theta,\nu_0}$ under $\Pi_{\theta,\nu_0}\times\nu_1^\infty$ denoted by $\Pi^{(n)}_{\theta,\nu_0,\nu_1}$. Using the terminology of statistical mechanics, the conditional large deviation principle in [11] can be viewed as quenched large deviations while the main result in this paper is the annealed large deviations. As application, we will study the asymptotic efficiency of the Dirichlet posterior mean in terms of large deviation rate functions. In particular we observe the following:

- The Dirichlet posterior mean is asymptotically less efficient than the Dirichlet mean and the sample mean.
- If ν_1 is uniform over $[0,\tau]$ for some $0<\tau\le 1$, then the Dirichlet mean and the sample mean have the same asymptotic efficiency while the Dirichlet posterior mean is strictly less efficient than both.

The development of the paper is as follows. The main result of annealed large deviation principle will be presented in Section 2. The rate function is identified explicitly. It consists of a frequentist part and a Bayesian part. The main result is applied in the study of asymptotic efficiencies of three random means in Section 3. The proof of the main result is contained in Section 4. All terms and definitions regarding large deviations are found in [6].

2. Main result. Recall that S be a compact Polish space. By Urysohn's embedding lemma, the space S is homeomorphic to a compact subspace of $\mathbb{R}^\infty$. For ease of presentation, we choose $S=[0,1]$ in the sequel, which contains all the ingredient for the proof of

general cases. The space of probability measures on S, $M_1(S)$, is equipped with the weak topology, which is generated by the metric

$$\rho(\upsilon, \mu) = \sum_{i=1}^{\infty} \frac{|\langle \upsilon - \mu, f_i \rangle| \wedge 1}{2^i}, \quad \upsilon, \mu \in M_1(S),$$

where $\{f_i \in C(S) : i = 1, 2, \ldots\}$ be a countable dense subset of $C(S)$. The main result of this paper is the following theorem.

THEOREM 2.1. *Fix ν_1 in $M_1(S)$ with topological support* $\operatorname{supp}(\nu_1)$ *containing at least two points in S. Then the family $\{\Pi^{(n)}_{\theta,\nu_0,\nu_1} : n \geq 1\}$ satisfies a large deviation principle on $M_1(S)$ with speed n and good rate function*

$$J_{\nu_1}(\mu) = \begin{cases} \inf\limits_{\upsilon \in M_1(S)} \{H(\upsilon|\mu) + H(\upsilon|\nu_1)\} & \operatorname{supp}(\mu) \subset \operatorname{supp}(\nu_1) \\ \infty & \textit{else}. \end{cases} \tag{2.1}$$

Equivalently, due to the compactness of $M_1(S)$, we have for any μ in $M_1(S)$

$$\begin{aligned} \lim_{\delta \to 0} \underline{\lim}_{n \to \infty} \frac{1}{n} \ln \Pi^{(n)}_{\theta,\nu_0,\nu_1}(\rho(\upsilon, \mu) < \delta) &= \lim_{\delta \to 0} \overline{\lim}_{n \to \infty} \frac{1}{n} \ln \Pi^{(n)}_{\theta,\nu_0,\nu_1}(\rho(\upsilon, \mu) \leq \delta) \\ &= -J_{\nu_1}(\mu) \end{aligned} \tag{2.2}$$

and, for any $c \geq 0$, the set $\{\upsilon \in M_1(S) : J_{\nu_1}(\upsilon) \leq c\}$ is compact.

REMARKS.

(a) The large deviation result does not depend on θ and ν_0. Thus the Dirichlet posterior and the Bayesian bootstrap have the same large deviation behaviour.

(b) There are two sources of randomness in the Bayesian bootstrap $\sum_{i=1}^{n} W_{n,i}\delta_{X_i}$: the prior reflected in the random weights and the i.i.d. observations (frequentist component). These correspond to the two parts in $J_{\nu_1}(\mu)$. Noting that

$$H(\mu|\nu_1) = H(\mu|\mu) + H(\mu|\nu_1), \qquad H(\nu_1|\mu) = H(\nu_1|\mu) + H(\nu_1|\nu_1),$$

it follows that for $\operatorname{supp}(\mu) \subset \operatorname{supp}(\nu_1)$, $J_{\nu_1}(\mu)$ is less than both $H(\mu|\nu_1)$ and $H(\nu_1|\mu)$. This reflects the impact of the combined randomness.

(c) If the Dirichlet$(1, \ldots, 1)$ weights in the Bayesian bootstrap are replaced with Dirichlet$(a_n, \ldots, a_n)$ weights for a_n tending to infinity, then the unconditional large deviation principle would be similar to the large deviation principle for the empirical distribution $\mathcal{L}_n$. On the other hand, if the i.i.d. observations are replaced with triangular arrays $X_1^{(n)}, \ldots, X_n^{(n)}$ and the empirical distribution

$$\frac{1}{n} \sum_{k=1}^{n} \delta_{X_k^{(n)}}$$

satisfies a large deviation principle with speed n^γ for some $\gamma > 1$, then the unconditional large deviation principle will be similar to the conditional large deviation principle in [11]. The latter has been established in [10] and [9], where the corresponding triangular array are the eigenvalues of random matrices.

(d) For any μ, ν in $M_1(S)$, the function $J_\nu(\mu)$ defined as in (2.1) provides a new measurement of divergence between μ and ν, which we call the J-divergence. It is nonnegative, symmetric, convex, and equals to zero only when $\mu = \nu$. Since the triangle inequality does

not hold, it is not a metric. But the finiteness of $J_\nu(\mu)$ does not require the absolute continuity between μ and ν. This can be seen from the following example:

$$\upsilon(dx) = dx, \qquad \mu(dx) = \frac{1}{2}\big[\upsilon(dx) + \delta_{\{0\}}(dx)\big],$$

$$\nu(dx) = \frac{1}{2}\big[\upsilon(dx) + \delta_{\{1\}}(dx)\big], \quad x \in S.$$

It follows from direct calculation that

$$H(\upsilon|\mu) = \ln 2 = H(\upsilon|\nu), \qquad J_\nu(\mu) \le 2\ln 2.$$

Clearly $H(\mu|\nu) = H(\nu|\mu) = \infty$. Thus $J_\nu(\mu)$ can be strictly less than the minimum of $H(\mu|\nu)$ and $H(\nu|\mu)$. This helps in quantifying the relative information between probabilities that have no absolute continuity relation.

(e) Let d_h and d_{tv} denote the Hellinger distance and the total variation distance on $M_1(S)$ respectively. Then, for any μ, ν in $M_1(S)$ we have by Pinsker's inequality,

$$d_h^4(\mu, \nu) \le 4d_{\mathrm{tv}}^2(\mu, \nu) \le 16 J_\nu(\mu).$$

3. Application. Let ν_1 in $M_1(S)$ have mean value κ and variance σ^2. Assume the observations $X_1, X_2, \ldots$ are i.i.d.with common distribution ν_1. Recall that for any f in $C(S)$ and μ in $M_1(S)$, the integral of f with respect to μ is denoted by $\langle \mu, f\rangle$. The mean of μ corresponds to $f(x) = x$. In this section, we will apply the main result to the analysis of the asymptotic behaviours of following three random means:

$$\text{sample mean} = \langle \mathcal{L}_n, x\rangle,$$

$$\text{Dirichlet mean} = \langle \Xi_{n,\nu_0}, x\rangle,$$

$$\text{Dirichlet posterior mean} = \big\langle \Xi^{(n)}_{\theta,\nu_0}, x\big\rangle.$$

The study of random means has been an active research area over the years. Two comprehensive surveys on the subject can be found in [15] and [19].

By direct calculation, we obtain that

$$\mathbb{E}\big[\langle \mathcal{L}_n, x\rangle\big] = \mathbb{E}\big[\langle \Xi_{n,\nu_0}, x\rangle\big] = \kappa,$$

$$\mathbb{E}\big[\big\langle \Xi^{(n)}_{\theta,\nu_0}, x\big\rangle\big] = \frac{n}{n+\theta}\kappa + \frac{\theta}{n+\theta}\langle \nu_0, x\rangle,$$

and

$$\mathrm{Var}\big[\langle \mathcal{L}_n, x\rangle\big] = \frac{\sigma^2}{n},$$

$$\mathrm{Var}\big[\langle \Xi_{n,\nu_0}, x\rangle\big] = \frac{\sigma^2}{n+1},$$

$$\mathrm{Var}\big[\big\langle \Xi^{(n)}_{\theta,\nu_0}, x\big\rangle\big] = \frac{2n\sigma^2}{(n+\theta)(n+\theta+1)} + \frac{\theta}{(n+\theta)^2(n+\theta+1)}\big[\kappa^2 + (n+\theta)\big\langle \nu_0, x^2\big\rangle - \theta\langle \nu_0, x\rangle^2 - 2n\kappa\langle \nu_0, x\rangle\big].$$

Thus all three means are either unbiased or asymptotically unbiased estimators of κ, and have the same magnitude of fluctuations. It is natural to ask which of these provide a better or more efficient estimation?

Following [1], we compare these means through the study of their asymptotic efficiencies characterized by the large deviation rate functions. The asymptotically more efficient estimators will have bigger rate functions. This criterion has been used in the comparison of weighted bootstraps [2]. Rate functions are usually proportional to the Fisher information, which is consistent with the well known Cramér–Rao theory.

Noting that the map

$$M_1(S) \to \mathbb{R}, \quad \mu \to \langle \mu, x \rangle$$

is continuous, the following result follows by a direct application of Theorem 2.1, the large deviations for the Dirichlet process [5, 18], the Sanov's theorem, and the contraction principle.

THEOREM 3.1. *The laws of the families* $\{\langle \mathcal{L}_n, x\rangle : n \geq 1\}$, $\{\langle \Xi_{n,\nu_0}, x\rangle : n \geq 1\}$, *and* $\{\langle \Xi^{(n)}_{\theta,\nu_0}, x\rangle : n \geq 1\}$ *satisfy large deviation principles on S with the same speed n and respective good rate functions*

$$I_1(u;\nu_1) = \inf_{\mu \in M_1(S)} \{H(\mu|\nu_1) : \langle \mu, x\rangle = u\},$$

$$I_2(u;\nu_1) = \inf_{\mu \in M_1(S)} \{H(\nu_1|\mu) : \langle \mu, x\rangle = u\},$$

$$I_3(u;\nu_1) = \inf_{\mu \in M_1(S)} \{J_{\nu_1}(\mu) : \langle \mu, x\rangle = u\}.$$

By Theorem 2.1, we have $I_3(u;\nu_1) \leq \min\{I_1(u;\nu_1), I_2(u;\nu_1)\}$ for all u. Hence the Dirichlet posterior mean is less efficient than both the sample mean and the Dirichlet mean. The inequality can be strict as the following example indicates:

$$\nu_1(dx) = \frac{1}{2}[\delta_0 + \delta_1],$$

$$\mu(dx) = \frac{3}{4}\delta_0 + \frac{1}{4}\delta_1,$$

$$I_1\left(\frac{1}{4};\nu_1\right) = H(\mu|\nu_1) = \frac{3}{4}\ln 3 - \ln 2,$$

$$I_2\left(\frac{1}{4};\nu_1\right) = H(\nu_1|\mu) = \ln 2 - \frac{\ln 3}{2},$$

and

$$I_3\left(\frac{1}{4};\nu_1\right) = \ln\frac{4}{2+\sqrt{3}}.$$

Clearly we have

$$I_3\left(\frac{1}{4};\nu_1\right) < I_1\left(\frac{1}{4};\nu_1\right) < I_2\left(\frac{1}{4};\nu_1\right).$$

The rate functions $I_1(u;\nu_1)$ and $I_2(u;\nu_1)$ are the forward information projection and the reverse information projection, respectively. There is no simple order between them in general. It is known that the forward information projection over convex set has an explicit solution [3], and the reverse information projection has a solution over any log-convex domain [4]. Since the set $\{\mu \in M_1(S) : \langle \mu, x\rangle = u\}$ is not log-convex, it is not clear whether one can identify the minimizer, and thus $I_2(u)$ explicitly.

Our next result shows that if $\nu_1(dx)$ is uniform over $[0,\tau]$ for some $0 < \tau \leq 1$, then we have $I_3(u,\nu_1) < I_1(u;\nu_1) = I_2(u;\nu_1)$ for $u \neq \tau/2$. In other words, the sample mean and the Dirichlet mean have the same efficiency while the Dirichlet posterior mean is strictly less efficient than both.

THEOREM 3.2. *Let ν_1 be the uniform distribution over $[0,\tau]$. Then, for any $0 \le u \le \tau$,*

$$\begin{aligned} I_1(u,\nu_1) &= I_2(u;\nu_1) \\ &= \int_0^\tau \ln(\alpha+\beta x)\nu_1(dx) \\ &= F^{-1}(u/\tau)(u-\tau) + \ln\bigl(1+F^{-1}(u/\tau)u\bigr), \end{aligned} \tag{3.1}$$

where

$$\alpha+\tau\beta = \alpha e^{\tau\beta}, \qquad \alpha+\beta u = 1$$

and

$$F(z) = \begin{cases} \dfrac{1}{2} & z=0, \\ \dfrac{e^z}{e^z-1} - \dfrac{1}{z} & \textit{else}. \end{cases}$$

In addition, we have for $u \ne \tau/2$

$$I_3(u,\nu_1) < I_1(u,\nu_1) = I_2(u;\nu_1). \tag{3.2}$$

PROOF. By the minimum discrimination information theorem (pages 36–39 in [14]), the infimum of $I_1(u;\nu_1)$ is achieved at a probability measure μ_0 satisfying

$$\mu_0(dx) = ce^{rx}\nu_1(dx).$$

The constraints

$$\langle \nu_1, ce^{rx}\rangle = 1, \qquad \langle \nu_1, cxe^{rx}\rangle = u$$

imply that

$$c = \frac{r\tau}{e^{r\tau}-1}, \qquad ce^{r\tau} - 1 = ur,$$

and

$$u/\tau = \frac{e^{r\tau}}{e^{r\tau}-1} - \frac{1}{r\tau} = F(r\tau).$$

By direct calculation,

$$\begin{aligned} I_1(u;\nu_1) &= H(\mu_0|\nu_1) \\ &= \frac{1}{\tau}\int_0^\tau ce^{rx}(\ln c + rx)\,dx \\ &= \ln c + ru \\ &= \tau r(u-1) + \ln(1+\tau ru) \\ &= F^{-1}(u/\tau)(u-\tau) + \ln\bigl(1+F^{-1}(u/\tau)u\bigr). \end{aligned}$$

On the other hand, for any μ in $M_1(S)$, let $\mu = \mu_1 + \mu_2$ be the Lebesgue decomposition of μ with respect to ν_1, with $\mu_1 \ll \nu_1$ and $\mu_2 \perp \nu_1$. Set $f(x) = \frac{d\mu_1}{d\nu_1}$, the Radon–Nikodym derivative of μ_1 with respect to ν_1. Then for any u in S we have

$$\begin{aligned} I_2(u;\nu_1) &= \inf\bigl\{\langle \nu_1, -\ln f(x)\rangle : \langle \mu_1+\mu_2, x\rangle = u, \mu_1 \equiv \nu_1\bigr\} \\ &= \inf\bigl\{\langle \nu_1, -\ln f(x)\rangle : f>0, \nu_1\text{-a.s.}, \langle \nu_1, f(x)\rangle \le 1, \langle \nu_1, xf(x)\rangle \le u\bigr\} \\ &= \inf_{a\in(0,1], b\in(0,a\wedge u]} \inf_{f\in\Gamma_{a,b}} \bigl\{\langle \nu_1, -\ln f(x)\rangle\bigr\}, \end{aligned} \tag{3.3}$$

where

$$\Gamma_{a,b} = \{f \in B(S) : f > 0, \nu_1\text{-a.s.}, \langle \nu_1, f(x)\rangle = a, \langle \nu_1, xf(x)\rangle = b\}.$$

For any $0 < a \le 1$, and $0 < b \le u \wedge (\tau a)$, let $\lambda_1(a,b)$ and $\lambda_2(a,b)$ be such that

$$\lambda_1(a,b) + \tau\lambda_2(a,b) = \lambda_1(a,b)e^{a\tau\lambda_2(a,b)}, \qquad a\lambda_1(a,b) + b\lambda_2(a,b) = 1. \tag{3.4}$$

Since $b \le \tau \wedge a$, it follows that

$$\lambda_1(a,b) > 0, \qquad \lambda_1(a,b) + \lambda_2(a,b) > 0.$$

Thus the nonnegative function

$$g_{a,b}(x) = \frac{1}{\lambda_1(a,b) + \lambda_2(a,b)x}$$

is well defined. It follows from direct calculation that

$$\langle \nu_1, g_{a,b}(x)\rangle = a, \qquad \langle \nu_1, xg_{a,b}(x)\rangle = b.$$

Thus $g_{a,b}$ is in $\Gamma_{a,b}$. For any f in $\Gamma_{a,b}$, we obtain

$$\int_0^\tau \frac{f(x)}{g_{a,b}(x)}\nu_1(dx) = \int_0^\tau \big(\lambda_1(a,b) + \lambda_2(a,b)x\big)f(x)\nu_1(dx) = 1$$

and

$$\int_0^\tau \ln\frac{f(x)}{g_{a,b}(x)}\nu_1(dx) \le \ln\bigg(\int_0^\tau \frac{f(x)}{g_{a,b}(x)}\nu_1(dx)\bigg) = 0.$$

Hence

$$\langle \nu_1, -\ln f(x)\rangle \ge \langle \nu_1, -\ln g_{a,b}(x)\rangle,$$

and the infimum of $\langle \nu_1, -\ln f(x)\rangle$ is achieved at $g_{a,b}$. It is not difficult to see that the map from (a,b) to (λ_1, λ_2) is one-to-one, and

$$\frac{\partial a}{\partial \lambda_1} < 0, \qquad \frac{\partial a}{\partial \lambda_2} < 0, \qquad \frac{\partial b}{\partial \lambda_1} < 0, \qquad \frac{\partial b}{\partial \lambda_2} < 0.$$

It follows that $\langle \nu_1, -\ln g_{a,b}(x)\rangle$ is decreasing in both a and b, and the infimum is achieved for $a = 1$, $b = u$. Solving the equations in (3.4), we obtain

$$u/\tau = F\big(-\tau\lambda_2(1,u)\big).$$

Thus by (3.3) we obtain

$$\begin{aligned} I_2(u;\nu_1) &= -\int_0^\tau \ln g_{1,u}(x)\nu_1(dx) \\ &= \int_0^\tau \ln(\alpha + \beta x)\nu_1(dx) \end{aligned}$$

which implies (3.1) by taking $\alpha = \lambda_1(1,u)$, $\beta = \lambda_2(1,u)$.

Finally we turn to the proof of (3.2). For any $u \ne \tau/2$, let

$$\mu_1(dx) = g_{1,u}(x)\nu_1(dx)$$

and

$$\mu_\lambda = \lambda\mu_0 + (1-\lambda)\mu_1, \quad 0 \le \lambda \le 1.$$

It is clear that μ_λ is in $M_1(S)$ and $\langle \mu_\lambda, x\rangle = u$ for all λ. By direct calculation we obtain

$$\begin{aligned} I_3(u,\nu_1) &\le J_{\nu_1}(\mu_\lambda) \\ &\le H\big(\lambda\mu_0 + (1-\lambda)\nu_1 | \mu_\lambda\big) + H\big(\big(\lambda\mu_0 + (1-\lambda)\nu_1 | \nu_1\big). \end{aligned}$$

It follows from the pair convexity that

$$H\big(\big(\lambda\mu_0 + (1-\lambda)\nu_1 | \nu_1\big) \le \lambda I_1(u,\nu_1)$$

and

$$H\big(\lambda\mu_0 + (1-\lambda)\nu_1 | \mu_\lambda\big) \le (1-\lambda) I_2(u,\nu_1).$$

Since the first inequality is strict for some λ in (0,1), it follows that

$$\begin{aligned} &H\big(\lambda\mu_0 + (1-\lambda)\nu_1 | \mu_\lambda\big) + H\big(\big(\lambda\mu_0 + (1-\lambda)\nu_1 | \nu_1\big) \\ &< \lambda I_1(u,\nu_1) + (1-\lambda) I_2(u,\nu_1) \\ &= I_1(u,\nu_1). \end{aligned}$$

Putting all these together we obtain (3.2). □

REMARKS.

(a) Let ν_2 denote the law of $\langle \Xi_{1,\nu_0}, x\rangle$ and $\eta_1, \eta_2, \ldots$ be i.i.d. with common distribution ν_2. Then by the Gamma–Dirichlet algebra (Theorem 1.1 in [7]) we have

$$\langle \Xi_{n,\nu_0}, x\rangle \stackrel{d}{=} \sum_{i=1}^{n} W_{n,i}\eta_i,$$

which implies that

$$I_2(u;\nu_1) = I_3(u;\nu_2).$$

(b) The distribution of the Dirichlet mean has been derived explicitly in many cases in [22]. For diffuse probability measure ν_1, the law of the Dirichlet mean $\langle \Xi_{n,\nu_0}, x\rangle$ for $n \ge 2$ is absolutely continuous with respect to the Lebesgue measure with the Radon–Nikodym derivative

$$q_n(x) = \frac{n-1}{\pi}\int_0^x (x-y)^{n-2} e^{-n\int_0^1 \ln|y-z|\nu_1(dz)} \sin(n\pi\nu_1\big([0,y]\big)\, dy, \quad 0 \le x \le 1.$$

But it is not clear how this can be used to obtain the explicit form of $I_2(u,\nu_1)$.

4. Proof of the main result. For ease of presentation, we assume that the topological support of ν_1 is S. The general case follows directly by defining the rate function to be infinity for μ with $\text{supp}(\mu) \not\subset \text{supp}(\nu_1)$.

For any $m \ge 1$ and any partition $A_1, \ldots, A_m$ of S, define the map

$$\pi : M_1(S) \to \triangle_m, \qquad \upsilon \to \big(\upsilon(A_1), \ldots, \upsilon(A_m)\big),$$

where

$$\triangle_m = \left\{(x_1, \ldots, x_m) \in [0,1] \times \cdots \times [0,1] : \sum_{k=1}^{m} x_k = 1\right\}.$$

Let $a_i = \theta\nu_0(A_i)$, $p_i = \nu_1(A_i)$, $i = 1, \ldots, m$, and

$$n_i = \#\{1 \le k \le n : X_k \in A_i\}, \quad i = 1, \ldots, m.$$

Then $\pi(\Xi^{(n)}_{\theta,\nu_0})$ has Dirichlet$(a_1+n_1,\dots,a_m+n_m)$ distribution, and $\pi(\mathcal{L}_n)$ has multinomial distribution with parameters $n,\ p_1,\dots,p_m$. If $a_i+n_i=0$ ($p_i=0$) for some i, then the corresponding coordinate in the Dirichlet (multinomial) distribution will be zero and the distribution can be viewed as a Dirichlet (multinomial) distribution on a lower dimensional space. Thus, without loss of generality, we assume $a_i+n_i>0$, $p_i>0$ for $i=1,\dots,m$.

LEMMA 4.1. *For any μ, υ in $M_1(S)$, set*

$$\pi(\mu)=(q_1,\dots,q_m)=\mathbf{q},\qquad \pi(\upsilon)=(o_1,\dots,o_m)=\mathbf{o}.$$

Define

$$H\big(\pi(\upsilon)|\pi(\mu)\big)=\sum_{i=1}^m o_i\ln\frac{o_i}{q_i},\qquad H\big(\pi(\upsilon)|\pi(\nu_1)\big)=\sum_{i=1}^m o_i\ln\frac{o_i}{p_i},$$

where $0\ln 0=0\ln\frac{0}{0}=0$. *Then we have*

$$\begin{aligned}&\lim_{\delta\to0}\varliminf_{n\to\infty}\frac{1}{n}\ln P\big\{\big|\big(\pi\big(\Xi^{(n)}_{\theta,\nu_0}\big),\pi(\mathcal{L}_n)\big)-\big(\pi(\mu),\pi(\upsilon)\big)\big|<\delta\big\}\\&\quad=\lim_{\delta\to0}\varlimsup_{n\to\infty}\frac{1}{n}\ln P\big\{\big|\big(\pi\big(\Xi^{(n)}_{\theta,\nu_0}\big),\pi(\mathcal{L}_n)\big)-\big(\pi(\mu),\pi(\upsilon)\big)\big|\le\delta\big\}\\&\quad=-\big[H\big(\pi(\upsilon)|\pi(\mu)\big)+H\big(\pi(\upsilon)|\pi(\nu_1)\big)\big],\end{aligned}\tag{4.1}$$

where

$$\big|\big(\pi\big(\Xi^{(n)}_{\theta,\nu_0}\big),\pi(\mathcal{L}_n)\big)-\big(\pi(\mu),\pi(\upsilon)\big)\big|=\sum_{i=1}^m\big[\big|\Xi^{(n)}_{\theta,\nu_0}(A_i)-q_i\big|+\big|\mathcal{L}_n(A_i)-o_i\big|\big].$$

PROOF. Fix μ, ν in $M_1(S)$. For any $\delta>0$, let

$$F(n_1,\dots,n_m;\delta)=\frac{\Gamma(n+\theta)}{\Gamma(n_1+a_1)\cdots\Gamma(n_m+a_m)}\int\cdots\int_{\mathbf{D}_{\delta,m}}\prod_{i=1}^m x_i^{a_i+n_i-1}\,dx_1\dots dx_{m-1},$$

where

$$\mathbf{D}_{\delta,m}=\left\{(x_1,\dots,x_l)\in\triangle_m:\sum_{i=1}^m|x_i-q_i|<\delta\right\}.$$

Then we have

$$\begin{aligned}&P\big\{\big|\big(\pi\big(\Xi^{(n)}_{\theta,\nu_0}\big),\pi(\mathcal{L}_n)\big)-\big(\pi(\mu),\pi(\upsilon)\big)\big|<\delta\big\}\\&\quad=\sum_{\sum_{k=1}^m|\frac{n_i}{n}-o_i|<\delta}A(n_1,\dots,n_m;\delta),\end{aligned}\tag{4.2}$$

where

$$A(n_1,\dots,n_m;\delta)=\binom{n}{n_1\cdots n_m}F(n_1,\dots,n_m;\delta)\prod_{i=1}^m p_i^{n_i}.$$

Similarly we have

$$\begin{aligned}&P\big\{\big|\big(\pi\big(\Xi^{(n)}_{\theta,\nu_0}\big),\pi(\mathcal{L}_n)\big)-\big(\pi(\mu),\pi(\upsilon)\big)\big|\le\delta\big\}\\&\quad=\sum_{\sum_{k=1}^m|\frac{n_i}{n}-o_i|\le\delta}\bar{A}(n_1,\dots,n_m;\delta),\end{aligned}\tag{4.3}$$

where

$$\bar{A}(n_1,\dots,n_m;\delta)=\binom{n}{n_1\cdots n_m}\bar{F}(n_1,\dots,n_m;\delta)\prod_{i=1}^m p_i^{n_i},$$

and $\bar{F}(n_1,\dots,n_m;\delta)$ is defined by replacing $\mathbf{D}_{\delta,m}$ with

$$\bar{\mathbf{D}}_{\delta,m}=\left\{(x_1,\dots,x_m)\in\triangle_m:\sum_{i=1}^m|x_i-q_i|\le\delta\right\}.$$

We begin with the case $\theta=0$. The assumption $a_i+n_i>0$, $1\le i\le m$ implies that $n_i\ge 1$ for all i.

By Stirling's formula for the gamma function, there exist positive constants $c_1<c_2$ such that for all $n\ge 1$ and $n_i\ge 1$, $\sum_{i=1}^m n_i=n$

$$c_1\prod_{i=1}^m\left(\frac{n_i}{n}\right)^{-n_i}\sqrt{\frac{n_1\dots n_m}{n}}\le\frac{\Gamma(n)}{\Gamma(n_1)\cdots\Gamma(n_m)}\le c_2\prod_{i=1}^m\left(\frac{n_i}{n}\right)^{-n_i}\sqrt{\frac{n_1\dots n_m}{n}}.$$

Noting that for any integer $n\ge 1$

$$\sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n}\le n!\le en^{n+\frac{1}{2}}e^{-n},$$

it follows that there exist constants $c_3<c_4$ such that

$$c_3\prod_{i=1}^m\left(\frac{n_i}{n}\right)^{-n_i}\sqrt{\frac{n}{n_1\dots n_m}}\le\binom{n}{n_1\dots n_m}\le c_4\prod_{i=1}^m\left(\frac{n_i}{n}\right)^{-n_i}\sqrt{\frac{n}{n_1\dots n_m}}.$$

Putting these together it follows that

$$c_5\prod_{i=1}^m\left(\frac{n_i}{n}\right)^{-2n_i}\le\binom{n}{n_1\dots n_m}\frac{\Gamma(n)}{\Gamma(n_1)\cdots\Gamma(n_m)}\le c_6\prod_{i=1}^m\left(\frac{n_i}{n}\right)^{-2n_i}$$

for some positive constants $c_5<c_6$.

For any $\mathbf{x}=(x_1,\dots,x_m)$, $\mathbf{y}=(y_1,\dots,y_m)$ in $\triangle_m$, define the function

$$\Psi(\mathbf{x},\mathbf{y})=H(\mathbf{y}|\mathbf{x})+H\big(\mathbf{y}|\pi(\nu_1)\big).$$

Then we have for $\mathbf{y}=(\frac{n_1}{n},\dots,\frac{n_m}{n})$

$$\begin{aligned}&c_5\int\cdots\int_{\mathbf{D}_{\delta,m}}\exp\left\{-n\left(\Psi(\mathbf{x},\mathbf{y})+n^{-1}\ln\prod_{i=1}^m x_i\right)\right\}dx_1,\dots dx_{m-1}\\ &\le A(n_1,\dots,n_m;\delta)\le\bar{A}(n_1,\dots,n_m;\delta)\\ &\le c_6\int\cdots\int_{\bar{\mathbf{D}}_{\delta,m}}\exp\left\{-n(\Psi(\mathbf{x},\mathbf{y})+n^{-1}\ln\prod_{i=1}^m x_i\right\}dx_1,\dots dx_{m-1}.\end{aligned}\tag{4.4}$$

We now divide the upper estimations into three cases.

Case 1. $q_i>0$ for all i. For any $\tau>0$ one can choose n large and δ small such that

$$\left|\Psi(\mathbf{x},\mathbf{y})+n^{-1}\ln\prod_{i=1}^m x_i-\Psi\big(\pi(\mu),\pi(\upsilon)\big)-n^{-1}\ln\prod_{i=1}^m q_i\right|\le\tau,$$

which implies that

$$\begin{aligned}&\bar{A}(n_1,\dots,n_m;\delta)\\ &\le c_6[2\delta]^m\exp\left\{-n\left(\Psi\big(\pi(\mu),\pi(\upsilon)\big)+n^{-1}\ln\prod_{i=1}^m q_i-\tau\right)\right\}.\end{aligned}\tag{4.5}$$

Since the total number of terms in (4.3) is at most $(n+1)^m$ and τ is arbitrary, we obtain that

$$\begin{aligned}
&\lim_{\delta\to 0}\overline{\lim_{n\to\infty}}\frac{1}{n}\ln P\big\{\big|\big(\pi\big(\Xi^{(n)}_{\theta,\nu_0}\big),\pi(\mathcal{L}_n)\big)-\big(\pi(\mu),\pi(\upsilon)\big)\big|\le\delta\big\}\\
&\quad\le-\sum_{i=1}^{l}o_i\bigg(\ln\frac{o_i}{q_i}+\ln\frac{o_i}{p_i}\bigg)\qquad(4.6)\\
&\quad=-\big[H\big(\pi(\upsilon)|\pi(\mu)\big)+H\big(\pi(\upsilon)|\pi(\nu_1)\big)\big].
\end{aligned}$$

Next we turn to the situation where $q_i=0$ for some i. Without loss of generality, we assume that there exists an $1<l<m$ such that $q_i>0$ for $1\le i\le l$ and $q_i=0$ for $i>l$.

Case 2. $o_i>0$ for some $i>l$. In this case the term $\frac{n_i-1}{n}\ln\frac{n_i/n}{x_i}$ in

$$\Psi(\mathbf{x},\mathbf{y})+n^{-1}\ln\prod_{i=1}^{m}x_i$$

converges to infinity as δ tends to zero. Thus

$$\lim_{\delta\to 0}\overline{\lim_{n\to\infty}}\frac{1}{n}\ln P\big\{\big|\big(\pi\big(\Xi^{(n)}_{\theta,\nu_0}\big),\pi(\mathcal{L}_n)\big)-\big(\pi(\mu),\pi(\upsilon)\big)\big|\le\delta\big\}\le-\infty.$$

The result (4.1) follows from the fact that

$$H\big(\pi(\upsilon)|\pi(\mu)\big)+H\big(\pi(\upsilon)|\pi(\nu_0)\big)=\infty.$$

Case 3. $o_i=0$ for $i>l$. It follows from direct calculation that

$$\begin{aligned}
&\exp\bigg\{-n\bigg(\Psi(\mathbf{x},\mathbf{y})+n^{-1}\ln\prod_{i=1}^{m}x_i\bigg)\bigg\}\\
&\quad=\bigg(\prod_{i=l+1}^{m}x_i^{n_i-1}\bigg)\exp\bigg\{-n\bigg(\sum_{i=1}^{l}y_i\bigg(\ln\frac{y_i}{x_i}+\ln\frac{y_i}{p_i}\bigg)+n^{-1}\ln\prod_{i=1}^{l}x_i\bigg)\bigg\}\\
&\qquad\times\exp\bigg\{-n\bigg(\sum_{i=l+1}^{m}y_i\bigg(\ln y_i+\ln\frac{y_i}{p_i}\bigg)\bigg)\bigg\}\\
&\quad\le\exp\bigg\{-n\bigg(\sum_{i=1}^{l}y_i\bigg(\ln\frac{y_i}{x_i}+\ln\frac{y_i}{p_i}\bigg)+n^{-1}\ln\prod_{i=1}^{l}x_i\bigg)\bigg\}\\
&\qquad\times\exp\bigg\{-n\bigg(\sum_{i=l+1}^{m}y_i\bigg(\ln y_i+\ln\frac{y_i}{p_i}\bigg)\bigg)\bigg\}
\end{aligned}$$

on the domain $\mathbf{D}_{\delta,m}$. The exponential term

$$\exp\bigg\{-n\bigg(\sum_{i=1}^{l}y_i\bigg(\ln\frac{y_i}{x_i}+\ln\frac{y_i}{p_i}\bigg)+n^{-1}\ln\prod_{i=1}^{l}x_i\bigg)\bigg\}$$

can be estimated by an argument similar to that used in deriving (4.5).

For the second exponential term we have

$$\begin{aligned}
&\lim_{\delta\to 0}\overline{\lim_{n\to\infty}}\frac{1}{n}\ln\exp\left\{-n\left(\sum_{i=l+1}^{m} y_i\left(\ln y_i+\ln\frac{y_i}{p_i}\right)\right)\right\}\\
&\quad\le -\lim_{\delta\to 0}\inf_{\sum_{i=l+1}^{m} y_i\le\delta}\left\{\sum_{i=l+1}^{m} y_i\left(\ln y_i+\ln\frac{y_i}{p_i}\right)\right\}\\
&\quad= 0.
\end{aligned}$$

Thus (4.6) also holds in this case. It remains to check the lower bound in **Cases 1** and **3**.

In **Case 1**, the function $\Psi(\mathbf{x},\mathbf{y})+n^{-1}\ln\prod_{i=1}^{m} x_i$ is continuous at $(\mathbf{q},\mathbf{o})$. Thus for any $\tau>0$ one can choose δ small so that for

$$|(\mathbf{x},\mathbf{y})-(\mathbf{q},\mathbf{o})|<\delta$$

and

$$\left|\Psi(\mathbf{x},\mathbf{y})-\Psi(\mathbf{q},\mathbf{o})+n^{-1}\left(\ln\prod_{i=1}^{m} x_i-\ln\prod_{i=1}^{m} q_i\right)\right|<\tau.$$

By (4.4) we have

$$\begin{aligned}
A(n_1,\dots,n_m;\delta)&\ge c_5\int\cdots\int_{\mathbf{D}_{\delta,m}}\exp\left\{-n\left[\Psi(\mathbf{x},\mathbf{y})+n^{-1}\ln\prod_{i=1}^{m} x_i\right]\right\}dx_1,\dots dx_m\\
&\ge c_5 e^{-n\tau}\exp\left\{-n\left[\Psi(\mathbf{q},\mathbf{o})+n^{-1}\prod_{i=1}^{m} q_i\right]\right\}\int\cdots\int_{\mathbf{D}_{\delta,m}} dx_1\cdots dx_m
\end{aligned}$$

which combined with (4.2) implies

$$\begin{aligned}
&\lim_{\delta\to 0}\underline{\lim_{n\to\infty}}\frac{1}{n}\ln P\left\{\left|\left(\pi\left(\Xi_{\theta,\nu_0}^{(n)}\right),\pi(\mathcal{L}_n)\right)-\left(\pi(\mu),\pi(\upsilon)\right)\right|<\delta\right\}\\
&\quad\ge -\left[H\left(\pi(\upsilon)|\pi(\mu)\right)+H\left(\pi(\upsilon)|\pi(\nu_1)\right)\right].
\end{aligned}\tag{4.7}$$

In **Case 3**, let $T=\{1\le i\le m: q_i=0\}(o_i=0 \text{ for } i\in T)$ and define

$$\Psi_1(\mathbf{x},\mathbf{y})=\sum_{i\notin T} y_i\ln\frac{y_i}{x_i},\qquad \Psi_2(\mathbf{x},\mathbf{y})=\sum_{i\in T} y_i\ln\frac{y_i}{x_i}.$$

The function $\Psi_1(\mathbf{x},\mathbf{y})+n^{-1}\ln\prod_{i\notin T} x_i$ is clearly continuous at $(\mathbf{q},\mathbf{o})$. On the other hand, set

$$\mathbf{C}_{\delta,m}=\left\{\mathbf{x}=(x,\dots,x_m)\in\triangle_m:\frac{1}{2}\left(q_i+\frac{\delta}{m}\right)<x_i<q_i+\frac{\delta}{m},\text{ for all } i\right\}.$$

Then the following holds on $\mathbf{C}_{\delta,m}$:

$$\Psi_2(\mathbf{x},\mathbf{y})+n^{-1}\ln\prod_{i\in T} x_i\le -m\delta\ln\frac{\delta}{2m}+\frac{|T|}{n}\ln\frac{\delta}{m},$$

where $|T|$ is the cardinality of T.

Since $\mathbf{C}_{\delta,m}$ is a subset of $\mathbf{D}_{\delta,m}$, it follows from (4.4) and an argument similar to **Case 1** that (4.7) holds in **Case 3**. Putting together (4.6) and (4.7) we obtain the result for $\theta=0$.

For $\theta > 0$, we can write the probability density function of $\pi(\Xi^{(n)}_{\theta,\upsilon_0})$ in the form

$$f_\theta(x_1,\dots,x_m) = \frac{\Gamma(n+\theta)}{\Gamma(n_1+a_1)\cdots\Gamma(n_m+a_+{}_m)}\prod_{i=1}^m x_i^{(n_i+a_i)-1}$$
$$= g_\theta(x_1,\dots,x_m)\,f_0(x_1,\dots,x_m),$$

where $f_0(x_1,\dots,x_m$ is the corresponding density function for $\theta = 0$ and

$$g_\theta(x_1,\dots,x_m) = \frac{\Gamma(n+\theta)}{\Gamma(n)}\frac{\Gamma(n_1)\cdots\Gamma(n_m)}{\Gamma(n_1+a_1)\cdots\Gamma(n_m+a_m)}\prod_{i=1}^m x_i^{a_i}.$$

The lemma follows from the observation that the factor

$$\prod_{i=1}^m x_i^{a_i} = \exp\left\{n\left[\frac{1}{n}\sum_{i=1}^m a_i \ln x_i\right]\right\}$$

does not change the estimates above, and

$$\lim_{n\to\infty}\frac{1}{n}\ln\frac{\Gamma(n+\theta)}{\Gamma(n)}\frac{\Gamma(n_1)\cdots\Gamma(n_m)}{\Gamma(n_1+a_1)\cdots\Gamma(n_m+a_m)} = 0. \qquad \square$$

For any μ in $M_1(S)$, set

$$S_\mu = \{t \in S : \mu(\{t\}) = 0\}.$$

For any $0 < t_1 < \cdots < t_m < 1$ in S_μ, define

$$\pi_{t_1,\dots,t_m}(\mu) = \big(\mu([0,t_1)),\dots,\mu([t_k,t_{k+1})),\dots,\mu([t_m,1])\big).$$

Then the following holds.

LEMMA 4.2. *For any υ, μ in $M_1(S)$, we have*

$$H(\upsilon|\mu) + H(\upsilon|\nu_1) = \sup_{0<t_1<\cdots<t_m<1\in\, S_\upsilon\cap S_\mu}\big\{H\big(\pi_{t_1,\dots,t_m}(\upsilon)|\pi_{t_1,\dots,t_m}(\mu)\big) + H(\pi_{t_1,\dots,t_m}(\upsilon)|\pi_{t_1,\dots,t_m}(\nu_1)\big\}. \tag{4.8}$$

PROOF. It is known [5, 12] that for any υ, μ in $M_1(S)$

$$H(\upsilon|\mu) = \sup_{0<t_1<\cdots<t_m<1\,\in\, S_\mu}\big\{H(\pi_{t_1,\dots,t_m}(\upsilon)|\pi_{t_1,\dots,t_m}(\mu)\big\}.$$

Since the supremum of sums is less than or equal to the sum of supremums, it follows that

$$H(\upsilon|\mu) + H(\upsilon|\nu_1) \ge \sup_{0<t_1<\cdots<t_m<1 \text{ in } S_\upsilon\cap S_\mu}\big\{H\big(\pi_{t_1,\dots,t_m}(\upsilon)|\pi_{t_1,\dots,t_m}(\mu)\big) + H(\pi_{t_1,\dots,t_m}(\upsilon)|\pi_{t_1,\dots,t_m}(\nu_1)\big\}.$$

To prove the other direction, we first recall the variational form (1.3) of the relative entropy

$$H(\upsilon|\mu) = \sup_{g\in C(S)}\big\{\langle\upsilon,g\rangle - \ln\langle\mu,e^g\rangle\big\}.$$

For any $\tau > 0$, there are g, h in $C(S)$ such that

$$H(\upsilon|\mu) \le \langle\upsilon,g\rangle - \ln\langle\mu,e^g\rangle + \tau$$

and

$$H(\upsilon|\nu_1) \le \langle\upsilon,h\rangle - \ln\langle\mu,e^h\rangle + \tau.$$

Since $S_\upsilon \cap S_\mu$ is dense in S, there exist $0 < t_{n1} < \cdots < t_{nn} < 1$ in $S_\upsilon \cap S_\mu$ such that

$$\lim_{n\to\infty} \max_{1\le i\le n+1} \{\max\{|g(x) - g(y)| : x, y \in [t_{n(i-1)}, t_{ni}]\} = 0,$$

$$\lim_{n\to\infty} \max_{1\le i\le n+1} \{\max\{|h(x) - h(y)| : x, y \in [t_{n(i-1)}, t_{ni}]\}\} = 0,$$

where $t_{n0} = 0$, $t_{n(n+1)} = 1$. Set

$$\alpha_{ni} = g(t_{ni}), \qquad \beta_{ni} = h(t_{ni}), \quad i = 1, \dots, n+1$$

and

$$A_{n(n+1)} = [t_{nn}, 1], \qquad A_{ni} = [t_{n(i-1)}, t_{ni}), \quad i = 1, \dots, n.$$

Then there exists $c_n(g, h)$ such that

$$\lim_{n\to\infty} c_n(g, h) = 0$$

$$\begin{aligned} H(\upsilon|\mu) &\le \sum_{i=1}^{n+1} \alpha_{ni}\upsilon(A_{ni}) - \ln \sum_{i=1}^{n+1} e^{\alpha_{ni}}\mu(A_{ni}) + \tau + c_n(g, h) \\ &\le H\big(\pi_{t_{n1},\dots,t_{nn}}(\upsilon)|\pi_{t_{n1},\dots,t_{nn}}(\mu)\big) + \tau + c_n(g, h) \end{aligned}$$

and

$$\begin{aligned} H(\upsilon|\nu_1) &\le \sum_{i=1}^{n+1} \beta_{ni}\upsilon(A_{ni}) - \ln \sum_{i=1}^{n+1} e^{\beta_{ni}}\nu_1(A_{ni}) + \tau + c_n(g, h) \\ &\le H\big(\pi_{t_{n1},\dots,t_{nn}}(\upsilon)|\pi_{t_{n1},\dots,t_{nn}}(\nu_1)\big) + \tau + c_n(g, h). \end{aligned}$$

Putting all these together we obtain (4.8). □

PROOF OF THEOREM 2.1. For any υ, μ in $M_1(S)$, and any $0 < t_1 < \cdots < t_m < 1$ in $S_\upsilon \cap S_\mu$, let

$$A_{m+1} = [t_m, 1], \qquad A_i = [t_{i-1}, t_i), \quad i = 1, \dots, m.$$

It follows from Lemma 4.1 that

$$\begin{aligned} &\lim_{\delta\to 0} \varliminf_{n\to\infty} \frac{1}{n} \ln P\big\{\big|\big(\pi_{t_1,\dots,t_m}\big(\Xi^{(n)}_{\theta,\nu_0}\big), \pi_{t_1,\dots,t_m}(\mathcal{L}_n)\big) \\ &\qquad - \big(\pi_{t_1,\dots,t_m}(\mu), \pi_{t_1,\dots,t_m}(\upsilon)\big)\big| < \delta\big\} \\ &= \lim_{\delta\to 0} \varlimsup_{n\to\infty} \frac{1}{n} \ln P\big\{\big|\big(\pi_{t_1,\dots,t_m}\big(\Xi^{(n)}_{\theta,\nu_0}\big), \pi_{t_1,\dots,t_m}(\mathcal{L}_n)\big) \\ &\qquad - \big(\pi_{t_1,\dots,t_m}(\mu), \pi_{t_1,\dots,t_m}(\upsilon)\big)\big| \le \delta\big\} \\ &= -\big[H\big(\pi_{t_1,\dots,t_m}(\upsilon)|\pi_{t_1,\dots,t_m}(\mu)\big) + H\big(\pi_{t_1,\dots,t_m}(\upsilon)|\pi_{t_1,\dots,t_m}(\nu_1)\big)\big]. \end{aligned}$$

By approximation and Lemma 4.2, we obtain

$$\begin{aligned} &\lim_{\delta\to 0} \varliminf_{n\to\infty} \frac{1}{n} \ln P\big(\big\{\rho\big(\Xi^{(n)}_{\theta,\nu_0}, \mu\big) < \delta, \rho(\mathcal{L}_n, \nu) < \delta\big\}\big) \\ &\quad = \lim_{\delta\to 0} \varlimsup_{n\to\infty} \frac{1}{n} \ln P\big(\big\{\rho\big(\Xi^{(n)}_{\theta,\nu_0}, \mu\big) \le \delta, \rho(\mathcal{L}_n, \nu) \le \delta\big\}\big) \\ &\quad = -\big[H(\upsilon|\mu) + H(\upsilon|\nu_1)\big]. \end{aligned}$$

Since $M_1(S) \times M_1(S)$ is compact, it follows from Theorem (P) in [20] that the family of the laws of $(\Xi^{(n)}_{\theta,\nu_0}, \mathcal{L}_n)$ satisfies a large deviation principle with rate function $H(\upsilon|\mu) + H(\upsilon|\nu_1)$.

Noting that $\Xi^{(n)}_{\theta,\nu_0}$ is the continuous image of $(\Xi^{(n)}_{\theta,\nu_0}, \mathcal{L}_n)$ through projection, we obtain the main result by the contraction principle. □

Acknowledgements. The author would like to thank the referees for helpful comments and suggestions.

Funding. This research was supported by a discovery grant from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

[1] BAHADUR, R. R. (1960). On the asymptotic efficiency of tests and estimates. *Sankhyā* **22** 229–252. MR0293767

[2] BARBE, P. and BERTAIL, P. (1995). *The Weighted Bootstrap*. *Lecture Notes in Statistics* **98**. Springer, New York. MR2195545 https://doi.org/10.1007/978-1-4612-2532-4

[3] CSISZÁR, I. (1975). I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* **3** 146–158. MR0365798 https://doi.org/10.1214/aop/1176996454

[4] CSISZÁR, I. and MATÚŠ, F. (2003). Information projections revisited. *IEEE Trans. Inf. Theory* **49** 1474–1490. MR1984936 https://doi.org/10.1109/TIT.2003.810633

[5] DAWSON, D. A. and FENG, S. (2001). Large deviations for the Fleming–Viot process with neutral mutation and selection. II. *Stochastic Process. Appl.* **92** 131–162. MR1815182 https://doi.org/10.1016/S0304-4149(00)00070-3

[6] DEMBO, A. and ZEITOUNI, O. (1998). *Large Deviations Techniques and Applications*, 2nd ed. *Applications of Mathematics* (*New York*) **38**. Springer, New York. MR1619036 https://doi.org/10.1007/978-1-4612-5320-4

[7] FENG, S. (2010). *The Poisson–Dirichlet Distribution and Related Topics*: *Models and Asymptotic Behaviors*. *Probability and Its Applications* (*New York*). Springer, Heidelberg. MR2663265 https://doi.org/10.1007/978-3-642-11194-5

[8] FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949

[9] GAMBOA, F., NAGEL, J. and ROUAULT, A. (2016). Sum rules via large deviations. *J. Funct. Anal.* **270** 509–559. MR3425894 https://doi.org/10.1016/j.jfa.2015.08.009

[10] GAMBOA, F. and ROUAULT, A. (2010). Canonical moments and random spectral measures. *J. Theoret. Probab.* **23** 1015–1038. MR2735735 https://doi.org/10.1007/s10959-009-0239-1

[11] GANESH, A. J. and O'CONNELL, N. (2000). A large-deviation principle for Dirichlet posteriors. *Bernoulli* **6** 1021–1034. MR1809733 https://doi.org/10.2307/3318469

[12] GEORGII, H.-O. (1988). *Gibbs Measures and Phase Transitions*. *De Gruyter Studies in Mathematics* **9**. de Gruyter, Berlin. MR0956646 https://doi.org/10.1515/9783110850147

[13] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007 https://doi.org/10.1214/aos/1016218228

[14] KULLBACK, S. (1959). *Information Theory and Statistics*. Wiley, New York. MR0103557

[15] LIJOI, A. and PRÜNSTER, I. (2009). Distributional properties of means of random probability measures. *Stat. Surv.* **3** 47–95. MR2529667 https://doi.org/10.1214/09-SS041

[16] LO, A. Y. (1986). A remark on the limiting posterior distribution of the multiparameter Dirichlet process. *Sankhyā Ser. A* **48** 247–249. MR0905464

[17] LO, A. Y. (1987). A large sample study of the Bayesian bootstrap. *Ann. Statist.* **15** 360–375. MR0885742 https://doi.org/10.1214/aos/1176350271

[18] LYNCH, J. and SETHURAMAN, J. (1987). Large deviations for processes with independent increments. *Ann. Probab.* **15** 610–627. MR0885133

[19] PITMAN, J. (2018). Random weighted averages, partition structures and generalized arcsine laws. Available at arXiv:1804.07896v1.

[20] PUHALSKII, A. (1991). On functional principle of large deviations. In *New Trends in Probability and Statistics*, *Vol.* 1 (*Bakuriani*, 1990) (V. Sazonov and T. Shervashidze, eds.). 198–218. VSP, Utrecht. MR1200917

[21] RAY, K. and VAN DER VAART, A. (2021). On the Bernstein–von Mises theorem for the Dirichlet process. *Electron. J. Stat.* **15** 2224–2246. MR4255307 https://doi.org/10.1214/21-ejs1821

[22] REGAZZINI, E., GUGLIELMI, A. and DI NUNNO, G. (2002). Theory and numerical analysis for exact distributions of functionals of a Dirichlet process. *Ann. Statist.* **30** 1376–1411. MR1936323 https://doi.org/10.1214/aos/1035844980

[23] SANOV, I. N. (1961). On the probability of large deviations of random variables. In *Select. Transl. Math. Statist. and Probability*, *Vol.* 1 213–244. Am. Math. Soc., Providence, RI. MR0116378

[24] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*: *With Applications to Statistics*. *Springer Series in Statistics*. Springer, New York. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2