Ann Inst Stat Math (2016) 68:439–468
DOI 10.1007/s10463-014-0501-x

Escort distributions minimizing the Kullback–Leibler divergence for a large deviations principle and tests of entropy level

Valérie Girardin · Philippe Regnault

Received: 22 July 2013 / Revised: 19 December 2014 / Published online: 20 January 2015

Abstract Kullback–Leibler divergence is minimized among finite distributions with finite state spaces under various constraints of Shannon entropy. Minimization is closely linked to escort distributions whose main properties related to entropy are proven. This allows a large deviations principle to be stated for the sequence of plug-in empirical estimators of Shannon entropy of any finite distributions. Since no closed-form expression of the rate function can be obtained, an explicit approximating function is constructed. This approximation is accurate enough to provide good results in all applications. Tests of entropy level, using both the large deviations principle and the minimization results, are constructed and shown to have a good behavior in terms of errors.

Keywords Escort distributions · Estimation · Information geometry · Kullback–Leibler divergence · Large deviations principle · Shannon entropy · Tests

1 Introduction

The concept of entropy has been introduced in the field of probability by Shannon (1948) through defining

V. Girardin (✉)
Laboratoire de Mathématiques N. Oresme, UMR6139, Campus II,
Université de Caen Basse Normandie, BP 5186, 14032 Caen, France
e-mail: valerie.girardin@unicaen.fr

P. Regnault
Laboratoire de Mathématiques de Reims, EA4535, UFR Sciences Exactes et Naturelles,
Moulin de la Housse, BP 1039, 51687 Reims, France
e-mail: philippe.regnault@univ-reims.fr

$$\mathbb{S}(P) = -\sum_{i \in E} P(i) \log P(i),$$

for any P belonging to the set $\mathcal{D}$ of all probability distributions on a finite set E, with the convention $0 \log 0 = 0$. We will set $E = \{0, \ldots, N\} = [\![0, N]\!]$, only for the sake of simplicity.

Kullback and Leibler (1951) introduced the Kullback–Leibler divergence (KL-divergence) of a distribution Q relative to another P as

$$\mathbb{K}(Q|P) = \sum_{i \in E} Q(i) \log \frac{Q(i)}{P(i)}$$

with the conventions $0 \log(0/a) = 0$ for $0 \leq a \leq 1$ and $a \log(a/0) = +\infty$, for $0 < a \leq 1$. KL-divergence appears through $\mathbb{K}(P|U) = \mathbb{S}(U) - \mathbb{S}(P)$ as a measure of variation of information from U to P, where U is the uniform distribution on E. See Cover and Thomas (1991) for a detailed study of Shannon entropy, KL-divergence, and their fields of application.

The need to minimize the KL-divergence under constraints arises in numerous applications. Linear constraints on Q are classical; see Csiszár (1975) and the references therein. Further, in thermodynamics, the equilibrium distribution of a system maximizes the entropy, or in other words minimizes the KL-divergence with respect to the uniform distribution, subject to a given average energy level. In parametric statistics, the maximum likelihood estimator of the parameter is the minimizer of the KL-divergence of the empirical distribution with respect to the parametric distribution. We are here interested in highly non-linear entropic constraints. Precisely, we will minimize $\mathbb{K}(Q|P)$, first subject to $\mathbb{S}(Q) = s_0$, then to $\mathbb{S}(P) = s_1$, and finally determine the minimum under both constraints, that is

$$\inf_{P \in \mathcal{S}_{s_1}} \inf_{Q \in \mathcal{S}_{s_0}} \mathbb{K}(Q|P) = \mathbb{K}(\mathcal{S}_{s_0}|\mathcal{S}_{s_1}), \tag{1}$$

where $\mathcal{S}_s = \mathbb{S}^{-1}(\{s\}) = \{P \in \mathcal{D} : \mathbb{S}(P) = s\}$. The minimization process will use the Lagrange multipliers method and also rely on concepts of information geometry. Indeed, $\mathbb{K}(\mathcal{S}_{s_0}|\mathcal{S}_{s_1})$ appears as the KL-divergence between entropic spheres $\mathcal{S}_{s_0}$ and $\mathcal{S}_{s_1}$ in information geometry. So doing, an explicit expression will be obtained for $\mathbb{K}(\mathcal{S}_{s_0}|\mathcal{S}_{s_1})$; originally requiring minimization with respect to $(|E| - 1)^2$ variables, it will be shown to amount to the numerical determination of only two parameters.

The concept of information geometry is associated to the Riemannian manifold structure induced by the KL-divergence on $\mathcal{D}$. Its metric is Fisher information. The families of escort distributions defined by

$$\mathcal{E}_P^k(i) = \frac{P(i)^k}{\sum_{j \in E} P(j)^k}, \quad i \in E, \tag{2}$$

where $P \in \mathcal{D}$ and $k \in \mathbb{R}^* = \mathbb{R} \backslash \{0\}$, play a prominent part in the field: they constitute the geodesics of information geometry with respect to an affine connection naturally induced by the KL-divergence; see Amari and Nagaoka (2000), Sgarro (1978), Csiszár (1975), Regnault (2011). They also provide a tool for zooming at different parts of P, or for adapting P to constraints through their ability to scan its structure. Initially introduced in Beck and Schlogl (1993), their general properties linked to Tsallis entropy are proven in Tsallis (2009). Motivated by the applications to follow, first we will establish many of their entropic properties.

The KL-divergence is naturally involved in large deviations principles; see Dembo and Zeitouni (1998), Ellis (1985), Csiszár and Shieds (2004). We will establish, study and apply a large deviations principle (LDP) for the sequence of plug-in estimators $\widehat{S}_n = \mathbb{S}(\widehat{P}_n)$ of $\mathbb{S}(P)$ based on the empirical distribution $\widehat{P}_n$ of an n-sample of a finite distribution P. The rate function $I_{\mathbb{S}}$ involved in the LDP will be shown to depend on the number m of modes of P; precisely, $s \mapsto I_{\mathbb{S}}(s, P)$ is equal to $\mathbb{K}(\mathcal{E}_P^k|P)$ for $k > 0$ such that $\mathbb{S}(\mathcal{E}_P^k) = s$ when $s \geq m$, and otherwise to $-s - \log p$, where p is the modes' weight of P. Since no closed-form expression is available for k as a function of the entropy level s, as an alternative, we will build an approximation I_M converging uniformly to $I_{\mathbb{S}}$ and valid for all values of s. The original rate function can be replaced by its approximation in all applications involving the LDP, without significative loss of accuracy.

Finally, we will develop one of the numerous applications of LDP, to tests of entropy level. In data compression and coding theory, bounds on entropy level are well-known basic tools; see Cover and Thomas (1991). In goodness-of-fit theory, a test of entropy level of an independent and identically distributed (i.i.d.) sample constitutes a first approach to decide if it comes from such or such distribution; see Girardin and Lequesne (2013) and the references therein. Accordingly, we will consider the following tests of entropy level:

$$H_0 : \text{“}\mathbb{S}(P) = s_0\text{”} \quad \text{against} \quad H_1 : \text{“}\mathbb{S}(P) = s_1\text{”}, \tag{3}$$

$$H_0 : \text{“}\mathbb{S}(P) = s_0\text{”} \quad \text{against} \quad H_1 : \text{“}\mathbb{S}(P) \neq s_0\text{”}, \tag{4}$$

$$H_0 : \text{“}\mathbb{S}(P) = s_0\text{”} \quad \text{against} \quad H_1 : \text{“}\mathbb{S}(P) < s_0\text{”}. \tag{5}$$

The statistics for all above tests strongly rely on both the LDP and the KL-divergence between entropic spheres. First, a rejection region will be obtained for both tests (3) and (4) by assuming that $\mathbb{K}(\mathcal{S}_{\widehat{S}_n}|\mathcal{S}_{s_0})$ is greater than a threshold depending on the number n of observations. The error of the first kind will be shown to decrease linearly with n while the error of the second kind decreases exponentially fast with $\mathbb{K}(\mathcal{S}_{s_0}|\mathcal{S}_{s_1})$. The procedure is similar for Test (5). Examples will be given for Test (3) with different entropy level alternatives. The test of uniformity, equivalent to Test (5) with s_0 equal to the entropy of the uniform distribution, will also be illustrated.

The paper is organized as follows. In Sect. 2, the KL-divergence is minimized, after the main properties of escort distributions linked to entropy have been stated. In Sect. 3.1, the LDP based on the minimization of KL-divergence by escort distributions is proven to hold, after the asymptotic properties of the estimators of entropy have been recalled; the explicit approximation of the rate function is built in Sect. 3.2. Finally, in Sect. 4, tests of entropy levels are constructed and illustrated, using both the LDP and the double minimization of the KL-divergence.

2 Escort distributions and minimization of the KL-divergence

This section mainly aims at obtaining and making explicit in Sect. 2.3 the double minimum (1). First, we will establish new properties of the escort distributions in Sect. 2.1, both interesting in themselves and necessary to prove the results to follow. Then, in Sect. 2.2, we will determine the minimum under each of the two constraints, separately. Escort distributions of P will appear in minimizing $\mathbb{K}(Q|P)$ under constraint $\mathbb{S}(Q) = s$. The minimum of $\mathbb{K}(Q|P)$ under constraint $\mathbb{S}(P) = s$ will be shown to be obtained for an implicit function of Q. Both this function and escort distributions will play a crucial role in the double minimization.

The set $\mathcal{D}$ is identified to the simplex $\{(p_1, \dots, p_N) \in \mathbb{R}_+^N : \sum_{i=1}^N p_i \leq 1\}$ of $\mathbb{R}^N$. Any topological or differential property is thus related to the classical normed vector space structure on $\mathbb{R}^N$. In particular, the interior of $\mathcal{D}$ is the set $\mathcal{D}^\circ = \{P \in \mathcal{D} : \forall i \in E, P(i) > 0\}$ of all distributions supported by E. Since $\sum_{i=0}^N P(i) = 1$, the entropy $\mathbb{S}(P)$ of any $P \in \mathcal{D}^\circ$ is an explicit smooth function of $P' = (P(1), \dots, P(N))$, in mathematical words

$$\mathbb{S}(P) = -\sum_{i=1}^N P(i) \log P(i) - \left[1 - \sum_{i=1}^N P(i)\right] \log \left[1 - \sum_{i=1}^N P(i)\right] = S(P'). \quad (6)$$

Still, in order to simplify notation, we will keep on denoting $S(P')$ by $\mathbb{S}(P)$ when no confusion can ensue; for instance, differentiating $\mathbb{S}$ with respect to P will mean differentiating S with respect to P', with partial derivatives $\frac{\partial}{\partial P(i)}\mathbb{S}(P) = \frac{\partial}{\partial P(i)} S(P')$ for $i \in [\![1, N]\!]$.

The following lemma constitutes a basic tool which will be of use many times below.

Lemma 1 *Let $P \in \mathcal{D}$ have m modes. Then $\mathbb{S}(P) \in [\log m, \log \nu]$, where ν is the cardinal of the support of P.*

Proof Let p denote the weight of the modes of P. Let $P_m \in \mathcal{D}$ have the same modes and only one other non-zero weight, necessarily $1 - mp$. One of the basic properties of Shannon entropy says that $\mathbb{S}(P) \geq \mathbb{S}(P_m)$; see, e.g., Girardin and Limnios (2014).

Differentiating $\mathbb{S}(P_m) = -mp \log p - (1 - mp) \log(1 - mp)$ with respect to p shows that the function is decreasing on $[1/(m + 1), 1/m[$. Its infimum is $\log m$, obtained when p tends to $1/m$.

Clearly, the maximum value of $\mathbb{S}(P)$ is $\log \nu$, the entropy of the uniform distribution U on $[\![0, \nu - 1]\!]$. □

2.1 Properties of the escort distributions and their entropy

The ultimate goal of this part is to show that the entropy of an escort distribution is a bijective function of its parameter. For any non-uniform fixed $P \in \mathcal{D}^\circ$, let $\mathbf{s}_P$ be defined on $\mathbb{R}^*$ by $\mathbf{s}_P(k) = \mathbb{S}(\mathcal{E}_P^k)$. The properties stated in Proposition 1 will specifically allow us to prove in Proposition 2 that $\mathbf{s}_P$ restricted to either $\mathbb{R}_+^*$ or $\mathbb{R}_-^*$ is an

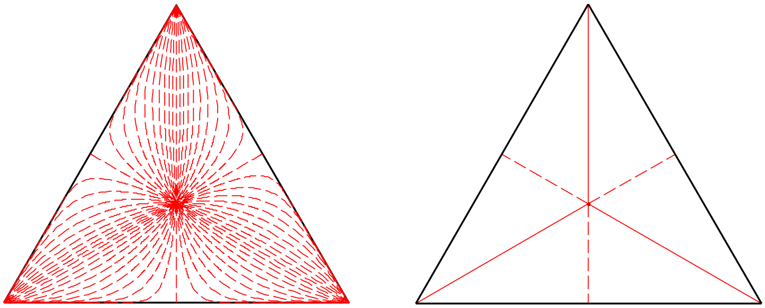


Fig. 1 *Left* escort families of $\mathcal{D}$ for $E = \{0, 1, 2\}$. *Right* distributions with $N = 2$ modes (*dashed lines*) and with 1 mode and $N = 2$ equal weights (*plain lines*); the *center* is U and the submits are Dirac distributions

invertible function. Further, the inequalities established in Proposition 3 are analogue in information geometry to the classical Pythagorean inequalities in Euclidean geometry.

For the sake of simplicity, we will state all results for $k \in \mathbb{R}_+^*$ – from which symmetric properties can be deduced for $k \in \mathbb{R}_-^*$. Also, all results will be stated for $P \in \mathcal{D}^\circ$, essentially for differentiability purpose. Actually, they also apply to any distribution P in the border $\partial\mathcal{D} = \{P \in \mathcal{D} : \exists i \in E, P(i) = 0\}$ of $\mathcal{D}$, by restricting the metric to the set of probability distributions with the same support as P.

Additional notation will be of use in the following. Let $\mathcal{E}$ be the escort transformation defined on $\mathcal{D} \times \mathbb{R}^*$ by $\mathcal{E}(P, k) = \mathcal{E}_P^k$, where $\mathcal{E}_P^k(i)$ is given in (2). Clearly, $\mathcal{E}$ is continuous on $\mathcal{D} \times \mathbb{R}^*$ and continuously differentiable on $\mathcal{D}^\circ \times \mathbb{R}^*$. For any $P \in \mathcal{D}^\circ$, let the partial derivatives of $\mathcal{E}$ with respect to $k \in \mathbb{R}^*$ be denoted by $\partial_k \mathcal{E}_P^k$ and with respect to $P(i)$, for $i \in [\![1, N]\!]$, by $\frac{\partial}{\partial P(i)}\mathcal{E}(P, k)$.

Illustration 1 For $E = \{0, 1, 2\}$, the set $\mathcal{D}$ is identified to an equilateral triangle in Fig. 1. The uniform distribution U is the centre of the triangle and the Dirac distributions are its summits. The families of escort distributions are represented by dashed curves on the left. The escort distributions with $N = 2$ equal weights are segments represented on the right; they will play a special role below in the determination of the KL-divergence between spheres.

Note that Fig. 1 and all figures below have been generated with a computer algebra system, so represent the true shapes of escort families and entropy level sets.

First, the escort transformation constitutes a scanning tool on the original distribution, and does not modify the ordering of weights for $k > 0$. In particular, the set of modes

$$\mathcal{M} = \left\{ i \in E : P(i) = \max_{j \in E} P(j) = p \right\},$$

with cardinal m, is an invariant of the transformation. For $m = N + 1$, it yields that for any $k > 0$, $\mathcal{E}(U, k) = U$.

Proposition 1 *Let $P \in \mathcal{D}^\circ$ have m modes.*

1. *For $0 < k < 1$, the escort transformation makes the distribution more uniform, whereas for $k > 1$, it concentrates weight on the modes. Asymptotically*:

a. *The escort distribution $\mathcal{E}_P^k$ converges to the uniform distribution as k tends to 0^+;*
b. *$\mathcal{E}_P^k$ converges to $\sum_{i\in\mathcal{M}}\delta_i/m$ as k tends to infinity, where δ_i is the Dirac measure at i.*

2. *Both the escort transformation and its derivative with respect to $k \in \mathbb{R}_+^*$ preserve the weight ordering. Precisely:*
   a. *If $P(i) > P(j)$, then $\mathcal{E}_P^k(i) > \mathcal{E}_P^k(j)$ and $\partial_k\mathcal{E}_P^k(i) > \partial_k\mathcal{E}_P^k(j)$, while if $P(i) = P(j)$, then $\mathcal{E}_P^k(i) = \mathcal{E}_P^k(j)$ and $\partial_k\mathcal{E}_P^k(i) = \partial_k\mathcal{E}_P^k(j)$.*
   b. *For $P \neq U$, if $i^* \in \mathcal{M}$ and $i_* \in E$ are such that $P(i_*) = \min_{i\in E} P(i)$, then $\partial_k\mathcal{E}_P^k(i^*) > 0$ and $\partial_k\mathcal{E}_P^k(i_*) < 0$.*

Proof 1. *a*. Since $P(j) \in]0, 1[$ for all $j \in E$, clearly $P(j)^k$ converges to 1 as k tends to 0. Thus,

$$\Lambda(k) = \sum_{j\in E} P(j)^k \tag{7}$$

converges to $N + 1$ while $\mathcal{E}_P^k(i) = P(i)^k/\Lambda(k)$ converges to $1/(N + 1)$.

1. *b*. We compute

$$\mathcal{E}_P^k(i) = \frac{P(i)^k}{\Lambda(k)} = \frac{P(i)^k/p^k}{\sum_{j\in E} P(j)/p^k} = \frac{[P(i)/p]^k}{m + \sum_{j\notin\mathcal{M}}[P(j)/p]^k}.$$

If $i \notin \mathcal{M}$, then $[P(i)/p]^k$ converges to 0, whereas $P(i)/p = 1$ for $i \in \mathcal{M}$. Hence $\mathcal{E}_P^k(i)$ tends to $1/m$ if $i \in \mathcal{M}$ and to 0 otherwise.

2. *a*. If $P(i) > P(j) > 0$, then $P(i)^k > P(j)^k > 0$ and hence $\mathcal{E}_P^k(i) > \mathcal{E}_P^k(j)$. Moreover,

$$\begin{aligned}\partial_k\mathcal{E}_P^k(i) &= \frac{P(i)^k}{[\Lambda(k)]^2}\sum_{l\in E} P(l)^k \log\left[\frac{P(i)}{P(l)}\right]\\ &> \frac{P(j)^k}{[\Lambda(k)]^2}\sum_{l\in E} P(l)^k \log\left[\frac{P(j)}{P(l)}\right] = \partial_k\mathcal{E}_P^k(j).\end{aligned}$$

The same arguments apply to prove the second part of the assertion.

2. *b*. If $\partial_k\mathcal{E}_P^k(i^*)$ was negative, then, according to Point 2. a., $\partial_k\mathcal{E}_P^k(i)$ would also be negative for all $i \in E$. But, since $\sum_{i\in E}\mathcal{E}_P^k(i) = 1$, we know that $\sum_{i\in E}\partial_k\mathcal{E}_P^k(i) = 0$. Therefore, $\partial_k\mathcal{E}_P^k(i)$ would be null for all $i \in E$ and all $k > 0$, so that $\mathcal{E}_P^k = P$, which is impossible since P is neither a uniform nor a Dirac distribution.

The proof for the minimum is similar. □

Now we can prove that $\mathbf{s}_P$ is bijective from $\mathbb{R}_+^*$ to $]\log(m), \log(N + 1)[$. Note that it can similarly be proven to be bijective from $\mathbb{R}_-^*$ to $]\log(m), \log(N + 1)[$.

Proposition 2 *Let $P \in \mathcal{D}^\circ$ have m modes, with $m \neq N + 1$. The function $\mathbf{s}_P$ defined on $\mathbb{R}_+^*$ by $\mathbf{s}_P(k) = \mathbb{S}(\mathcal{E}_P^k)$ is a bijection from $\mathbb{R}_+^*$ to $]\log(m), \log(N + 1)[$. Precisely, $\mathbf{s}_P$ is twice continuously differentiable on $\mathbb{R}_+^*$ with a negative derivative.*

Proof First, $\mathbf{s}_P$ is clearly infinitely continuously differentiable on $\mathbb{R}_+^*$. Its derivative, obtained by the chain rule, is

$$\mathbf{s}'_P(k) = -\sum_{i=1}^{N} \partial_k \mathcal{E}_P^k(i) \log \left[\frac{\mathcal{E}_P^k(i)}{\mathcal{E}_P^k(0)} \right], \quad k \in \mathbb{R}_+^*. \tag{8}$$

Let us order the elements of E according to their weights and hence suppose that $P(0) \geq P(1) \geq \cdots \geq P(N)$. Thanks to Point 1. a. of Proposition 1, $\mathcal{E}_P^k(0) \geq \cdots \geq \mathcal{E}_P^k(N)$ and $\partial_k \mathcal{E}_P^k(0) \geq \cdots \geq \partial_k \mathcal{E}_P^k(N)$. Thanks to Point 2., $\partial_k \mathcal{E}_P^k(0) > 0 > \partial_k \mathcal{E}_P^k(N)$, so that for all $k > 0$, some $j \in E$ exists such that $\partial_k \mathcal{E}_P^k(j) \geq 0 > \partial_k \mathcal{E}_P^k(j+1)$. Since $x \mapsto -\log x$ is a decreasing function, we get

$$\max_{i \in [\![1, j]\!]} -\log \left[\frac{\mathcal{E}_P^k(i)}{\mathcal{E}_P^k(0)} \right] = -\log \left[\frac{\mathcal{E}_P^k(j)}{\mathcal{E}_P^k(0)} \right],$$
$$\min_{i \in [\![j+1, \ldots, N]\!]} -\log \left[\frac{\mathcal{E}_P^k(i)}{\mathcal{E}_P^k(0)} \right] = -\log \left[\frac{\mathcal{E}_P^k(j+1)}{\mathcal{E}_P^k(0)} \right].$$

Cutting the sum in (8) into two sums thus yields

$$\mathbf{s}'_P(k) \leq -\log \left[\frac{\mathcal{E}_P^k(j)}{\mathcal{E}_P^k(0)} \right] \sum_{i=1}^{j} \partial_k \mathcal{E}_P^k(i) - \log \left[\frac{\mathcal{E}_P^k(j+1)}{\mathcal{E}_P^k(0)} \right] \sum_{i=j+1}^{N} \partial_k \mathcal{E}_P^k(i),$$

where the sum from 1 to j is empty if $j = 0$. Since $\sum_{i \in E} \mathcal{E}_P^k(i) = 1$, we have $\sum_{i=j+1}^{N} \partial_k \mathcal{E}_P^k(i) = -\sum_{i=1}^{j} \partial_k \mathcal{E}_P^k(i) - \partial_k \mathcal{E}_P^k(0)$, and hence

$$\begin{aligned} \mathbf{s}'_P(k) &\leq \log \left[\frac{\mathcal{E}_P^k(j)}{\mathcal{E}_P^k(j+1)} \right] \sum_{i=j+1}^{N} \partial_k \mathcal{E}_P^k(i) + \log \left[\frac{\mathcal{E}_P^k(j)}{\mathcal{E}_P^k(0)} \right] \partial_k \mathcal{E}_P^k(0) \\ &\leq \log \left[\frac{\mathcal{E}_P^k(j)}{\mathcal{E}_P^k(j+1)} \right] \sum_{i=j+1}^{N} \partial_k \mathcal{E}_P^k(i). \end{aligned}$$

Since $\mathcal{E}_P^k(j) \geq \mathcal{E}_P^k(j+1)$, the derivative $\mathbf{s}'_P(k)$ is negative. Moreover, the three above inequalities are strict ones as soon as some $i \in E$ exists such that $P(i) \neq P(i+1)$, that is for any non-uniform P.

Now, let P have m modes. Thanks to Point 2. of Proposition 1, $\mathcal{E}_P^k$ has m modes too, and hence, thanks to Lemma 1, the minimum value of $\mathbb{S}(\mathcal{E}_P^k)$ is $\log m$ again. □

Finally, the sign of $\mathbb{K}(\mathcal{E}_P^k|U) + \mathbb{K}(U|P) - \mathbb{K}(\mathcal{E}_P^k|P)$, for $k \in \mathbb{R}$, is given in Proposition 3, depending on the respective positions of P, $\mathcal{E}_P^k$ and U.

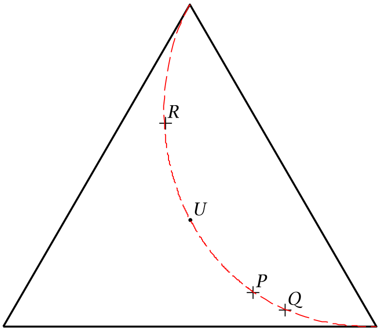


Fig. 2 Positions of a distribution P and two of its escorts, $Q = \mathcal{E}_P^{k_+}$ with $k_+ > 0$ and $R = \mathcal{E}_P^{k_-}$ with $k_- < 0$, with respect to U

Proposition 3 *Let* $P \in \mathcal{D}$ *and let* $\mathcal{E}_P^k$ *for* $k \in \mathbb{R}$ *be its escort distributions. Then*

$$\mathbb{K}(\mathcal{E}_P^k|U) + \mathbb{K}(U|P) \leq \mathbb{K}(\mathcal{E}_P^k|P), \quad k < 0, \tag{9}$$

$$\mathbb{K}(\mathcal{E}_P^k|U) + \mathbb{K}(U|P) \geq \mathbb{K}(\mathcal{E}_P^k|P), \quad k > 0, \tag{10}$$

where U *denotes the uniform distribution on* $E = [\![0, N]\!]$.

Proof We compute

$$\mathbb{K}(\mathcal{E}_P^k|P) - \mathbb{K}(U|P) - \mathbb{K}(\mathcal{E}_P^k|U) = -\sum_{i=0}^{N} \mathcal{E}_P^k(i) \log P(i) + \sum_{i=0}^{N} \frac{1}{N+1} \log P(i).$$

Let us set $f(k) = \sum_{i \in E} \mathcal{E}_P^k(i) \log P(i) = \Lambda'(k)/\Lambda(k)$, where $\Lambda(k)$ is defined in (7). In dynamical systems theory, this normalizing function Λ is called the Dirichlet series of fundamental measures of depth 1; see Vallée (2001). Cauchy–Schwarz inequality says that Λ is log-convex. In particular, $\log \Lambda(k+h) + \log \Lambda(k-h) - 2 \log \Lambda(k) \geq 0$ for all $h \geq 0$. Making h tend to 0, we get that $[\log \Lambda(k+h) + \log \Lambda(k-h) - 2 \log \Lambda(k)]/h^2$ converges to $(\log \Lambda)''(k) = f'(k)$, which shows that f' is non-negative. Therefore, f is an increasing function on both $\mathbb{R}_-^*$ and $\mathbb{R}_+^*$ so that $\lim_{k \to 0} f(k) = \sum_{i \in E} [\log P(i)]/(N+1)$ is its supremum value for $k \in \mathbb{R}_-$ and its infimum value for $k \in \mathbb{R}_+$. Both (9) and (10) are proven. □

Illustration 2 Figure 2 shows the positions of a three-state distribution P and two of its escort distributions, $Q = \mathcal{E}_P^{k_+}$ with $k_+ > 0$ and $R = \mathcal{E}_P^{k_-}$ with $k_- < 0$, with respect to U. The analogy between the inequalities of Proposition 3 in information geometry and the classical Pythagorean inequalities in Euclidean geometry is thus illustrated, through the interpretation of the families of escort distributions as segments; see Regnault (2011) for details.

2.2 Minimum of $\mathbb{K}(Q|P)$ with either P or Q fixed

On the one hand, minimizing $\mathbb{K}(Q|P)$ with fixed P, based on the properties proven in Sect. 2.1, will highlight the role of escort distributions in information geometry. On the other hand, minimizing $\mathbb{K}(Q|P)$ with fixed Q will be performed only as much as used in the double minimization to come in Sect. 2.3.

Even if the KL-divergence is not a distance by lack of symmetry and triangular inequality, projection can nevertheless be considered; see Csiszár (1975), Amari and Nagaoka (2000). The escort distributions of any $P \in \mathcal{D}$ will be proven to be its projections on the spheres centred at the distribution U uniform on E; in mathematical words,

$$\mathbb{K}(\mathcal{E}_P^k|P) = \inf\left\{\mathbb{K}(Q|P): \ Q \in \mathcal{D} \text{ such that } \mathbb{K}(Q|U) = \log(N+1) - s\right\},$$

for $k > 0$ such that $\mathbb{S}(\mathcal{E}_P^k) = s$, provided that $s > \log m$, where m is the number of modes of P. This was first stated in Sgarro (1978), then in theory of tests by Cover and Thomas (1991) Section 12.7, page 309, in large deviations theory by Dembo and Zeitouni (1998) Exercice 3.4.14 , and also in relation to Tsallis entropy by Bercher (2009). Unfortunately, the Lagrange multipliers method used by all above authors fails to yield the projection of distributions when $s \leq \log m$. Indeed, this method provides local extrema only over an open set of values, here $\mathcal{D}^\circ$ for $s > \log m$, while for $m > 1$ and $s \leq \log m$, the infimum is achieved on $\partial\mathcal{D}$. The following result takes all cases into account.

Theorem 1 *Let $P \in \mathcal{D}$ have m modes with weight p. Let ν be the cardinal of its support. Let $s \in \mathbb{R}_+$ and $\mathcal{S}_s = \{Q \in \mathcal{D} : \mathbb{S}(Q) = s\}$. Then*

$$\inf_{Q\in\mathcal{S}_s} \mathbb{K}(Q|P) =$$

$$\begin{array}{lll} -s - \log p & \textit{if} \ \ 0 \leq s \leq \log m, & (11)\\ \mathbb{K}(\mathcal{E}_P^k|P) & \textit{if} \ \ \log m < s \leq \log \nu, & (12)\\ +\infty & \textit{if} \ \ \log \nu < s, & \end{array}$$

where $\mathcal{E}_P^k$ is the escort distribution of P such that $\mathbb{S}(\mathcal{E}_P^k) = s$ with $k > 0$.

Proof In order to simplify notation, we will give the proof for $P \in \mathcal{D}^\circ$, that is supported by E; it applies to any $P \in \partial\mathcal{D}$, by replacing E by the support E_P of P and $N+1$ by the cardinal ν of E_P. Indeed, the support of any $Q \in \mathcal{S}_s$ achieving the infimum of $\mathbb{K}(Q|P)$ is included in E_P since otherwise $\mathbb{K}(Q|P)$ is infinite, so that all the arguments developed below easily transpose to the set of distributions supported by E_P.

We can assume without loss of generality that the m modes of $P \in \mathcal{D}^\circ$ are at $0, \ldots, m-1$. Let Q be any distribution in $\mathcal{S}_s$. Since $\mathbb{K}(Q|P) = -s - \sum_{i\in E} Q(i) \log P(i)$, where s is known, the quantity to be minimized is the average of the quantities $-\log P(i)$ weighted by Q, that is

$$-\sum_{i=0}^{N} Q(i) \log P(i) = -\log p \sum_{i=0}^{m-1} Q(i) - \sum_{i=m}^{N} Q(i) \log P(i).$$

The infimum is achieved at the distribution Q favoring the smallest of them, that is $-\log p$. Hence, any Q such that $Q(m) = \cdots = Q(N) = 0$ is a solution, provided that $\mathbb{S}(Q) = s$.

If $s \leq \log m$, such distributions do exist, so that

$$\inf_{Q \in \mathcal{S}_s} -\sum_{i \in E} Q(i) \log P(i) = -\log p,$$

and hence the searched minimum is $-s - \log p$.

If $s > \log m$, the infimum cannot be achieved in this way. Let us use the Lagrange multipliers method. The constraints are $\sum_{i \in E} Q(i) = 1$ and $\mathbb{S}(Q) = -\sum_{i \in E} Q(i) \log Q(i) = s$. Differentiating the Lagrangian

$$-\sum_{i \in E} Q(i) \log P(i) - \eta \sum_{i \in E} Q(i) \log Q(i) - \mu \sum_{i \in E} Q(i)$$

with respect to $Q(i)$ yields $-\log P(i) - \eta[\log Q(i) + 1] - \mu = 0$, from which it follows that the infimum takes the form $Q^{\min}(i) = C P(i)^k$, for $i \in E$. Since $\sum_{i \in E} Q^{\min}(i) = 1$, we get $C = 1/\sum_{j \in E} P(j)^k$, and hence $Q^{\min}$ is an escort distribution of P.

Finally, let $k' < 0 < k$ be the two real numbers such that $\mathbb{S}(\mathcal{E}_P^{k'}) = \mathbb{S}(\mathcal{E}_P^{k}) = s$ (see Proposition 2). We have $\mathbb{K}(\mathcal{E}_P^{k'}|U) = \mathbb{S}(U) - s = \mathbb{K}(\mathcal{E}_P^{k}|U)$, and hence we get from Proposition 3 that

$$\mathbb{K}\left(\mathcal{E}_P^{k}|P\right) \leq \mathbb{S}(U) - s + \mathbb{K}(U|P) \leq \mathbb{K}\left(\mathcal{E}_P^{k'}|P\right),$$

which proves that the minimum is obtained for $k > 0$. □

As shown by Illustration 3, Theorem 1 is strongly related to a vector space structure on $\mathcal{D}^\circ$, introduced by Sgarro (1978) and detailed in Regnault (2011). Indeed, $\mathcal{D}^\circ$ equipped with the operations

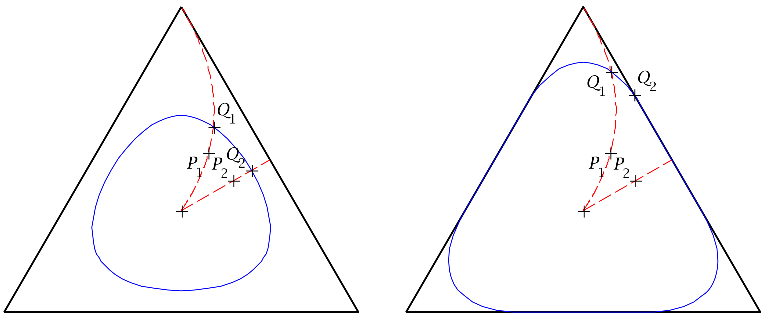
$$P \oplus Q(i) = \frac{P(i)Q(i)}{\sum_{j \in E} P(j)Q(j)} \quad \text{and} \quad k \odot P(i) = \mathcal{E}_P^k(i), \quad i \in E, \tag{13}$$

is an N-dimensional vector space, on which the KL-divergence behaves similarly to the square of the distance induced by a norm.

Illustration 3 In Fig. 3, plain lines represent entropy level sets $\mathcal{S}_s$, for $s > \log 2$ on the left and $s < \log 2$ on the right. Dashed lines are the segments with respect to the vector space structure (13)—or geodesics in information geometry, equal to the sets of all escort distributions of given distributions.

Two distributions, P_1 with one mode, and P_2 with two modes, are shown together with their respective projections on $\mathcal{S}_s$. For P_1, the infimum $\mathbb{K}(\mathcal{S}_s|P_1)$ is achieved in $\mathcal{D}^\circ$, whatever be the entropy level s. For P_2, the infimum $\mathbb{K}(\mathcal{S}_s|P_2)$ is also achieved in $\mathcal{D}^\circ$ when $s > \log 2$, but in $\partial\mathcal{D}$ when $s \leq \log 2$ since in this case P_2 has no escort distribution in $\mathcal{S}_s$.

Let us now state a necessary condition for a distribution $\widetilde{P} \in \mathcal{D}$ to achieve the infimum of $\mathbb{K}(Q|P)$ with Q fixed, subject to an entropic level constraint. In information



$s \in]\log 2, \log 3[$ $\qquad$ $s \in]0, \log 2[$

Fig. 3 Respective positions of P_1 with one mode, and P_2 with two modes, and their projections on $\mathcal{S}_s$ (*plain line*): $Q_1 = \mathcal{E}^k_{P_1} \in \mathcal{D}^\circ$ for P_1, and $Q_2 = \mathcal{E}^k_{P_2} \in \mathcal{D}^\circ$ when $s > \log 2$ (*left*), and $Q_2 \in \partial\mathcal{D}$ when $s \leq \log 2$ (*right*) for P_2

geometry, $\widetilde{P}$ appears while projecting Q on a Shannon entropic sphere centred at U, through

$$\mathbb{K}(Q|\widetilde{P}) = \inf \left\{ \mathbb{K}(Q|P) : P \in \mathcal{D} \text{ such that } \mathbb{K}(P|U) = \log(N+1) - s \right\}.$$

Proposition 4 *For any $Q \in \mathcal{D}$ and any $s \in]0, \log(N+1)[$, distributions $\widetilde{P} \in \mathcal{S}_s = \{P \in \mathcal{D} : \mathbb{S}(P) = s\}$ exist such that*

$$\inf_{P \in \mathcal{S}_s} \mathbb{K}(Q|P) = \mathbb{K}(Q|\widetilde{P}).$$

Moreover, they satisfy

$$-\eta \widetilde{P}(i) \log \widetilde{P}(i) + (1 - \eta s)\widetilde{P}(i) = Q(i), \quad i \in E, \tag{14}$$

where η is such that $\widetilde{P} \in \mathcal{S}_s$.

Proof Again, we will give the proof for $Q \in \mathcal{D}^\circ$; it extends to any $Q \in \partial\mathcal{D}$ by replacing E by the support of Q.

First, since distributions P do exist for which $\mathbb{K}(Q|P)$ is finite, the infimum is finite. This minimum is obtained at some $\widetilde{P} \in \mathcal{D}^\circ$, because $\mathbb{K}(Q|P)$ is infinite for all $P \in \partial\mathcal{D}$.

Then (14) is a direct application of the Lagrange multipliers method. Indeed, since $\mathbb{K}(Q|P) = \sum_{i \in E} Q(i) \log Q(i) - \sum_{i \in E} Q(i) \log P(i)$, differentiating the Lagrangian

$$-\sum_{i \in E} Q(i) \log P(i) - \eta \sum_{i \in E} P(i) \log P(i) - \mu \sum_{i \in E} P(i)$$

with respect to $P(i)$ yields $-\frac{Q(i)}{P(i)} - \eta(\log P(i) + 1) - \mu = 0$, for $i \in E$, or

$$Q(i) = -\eta P(i) \log P(i) - (\eta + \mu) P(i).$$

Furthermore,

$$\sum_{i \in E} Q(i) = -\eta \sum_{i \in E} P(i) \log P(i) - (\eta + \mu) \sum_{i \in E} P(i),$$

from which it follows that $-(\eta + \mu) = 1 - \eta s$, and finally (14). □

Note that Amari and Nagaoka (2000) Corollary 3.11 provides the projection of Q on Burg entropic spheres, that is

$$\inf \left\{ \mathbb{K}(Q|P) : P \in \mathcal{D} \text{ such that } -\sum_{i \in E} \log P(i) = c \right\} = \mathbb{K}(Q|(1-t)U + tQ),$$

for $t > 0$ such that $-\sum_{i \in E} \log((1-t)U + tQ) = c$.

2.3 Double minimization of $\mathbb{K}(Q|P)$

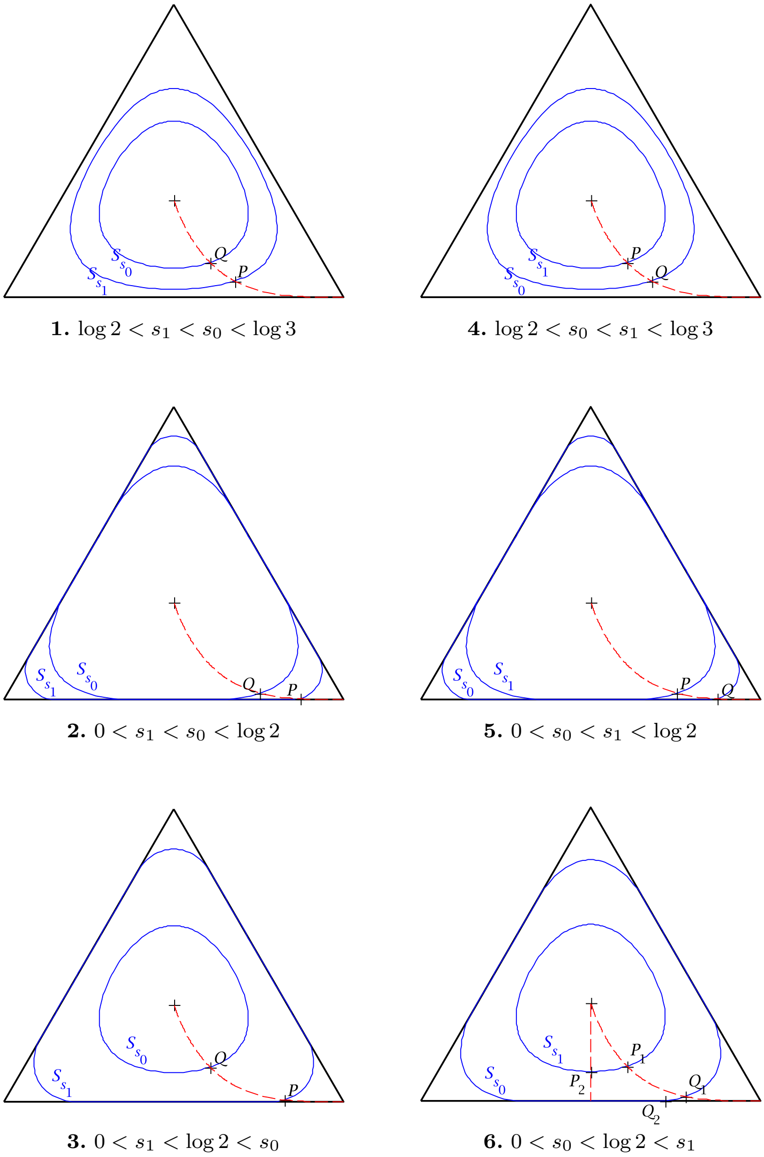
The aim of this section is to determine the KL-divergence $\mathbb{K}(\mathcal{S}_{s_0}|\mathcal{S}_{s_1})$ defined in (1). To the best of our knowledge, this quantity has never been fully explicited, even if its existence has been proven, for example in Udrişte (1994). The most natural way would be, using either Theorem 1 or Proposition 4, to minimize, through the Lagrange multipliers method, either $\mathbb{K}(Q|\widetilde{P})$ for $Q \in \mathcal{S}_{s_0}$, or both $\mathbb{K}(\mathcal{E}_P^k|P)$ and $-s_0 - \log p$ for $P \in \mathcal{S}_{s_1}$. Unfortunately, since no closed-form expressions are available for k and η, the related equations lead to nowhere.

For a better understanding of the geometric arguments that will be involved, let us look closely at the shapes and relative positions of the entropic spheres $\mathcal{S}_{s_0}$ and $\mathcal{S}_{s_1}$ for $(s_0, s_1) \in]0, \log(N+1)[^2$. As usual, Fig. 4 provides illustration for the three-state case.

Let us set

$$m_l = \max\{m \leq N : s_l > \log m\}, \quad \text{for} \quad l = 0, 1. \tag{15}$$

First, the nature of $\mathcal{S}_{s_l} \cap \partial\mathcal{D}$ depends on m_l. If $m_l = N$, then $\mathcal{S}_{s_l} \subset \mathcal{D}^\circ$, whereas if $m_l \leq N - 1$, the intersection of $\mathcal{S}_{s_l}$ with each face of the simplex $\mathcal{D}$ is an entropic sphere of this face. Lemma 1 states that all $P \in \mathcal{S}_{s_l}$ have at most m_l modes. Further, since the minimum of $\mathbb{K}(Q|P)$ subject to $\mathbb{S}(Q) = s_0$ depends on the number of modes of P, the double minimum highly depends on whether the entropic constraints allow the existence of P with several modes or not. It also depends on the respective positions of the spheres. If either $s_0 > s_1$ (Cases 1 to 3 of Fig. 4) or $s_0 < s_1$ and $m_0 = m_1$ (Cases 4 and 5), then for all $P \in \mathcal{S}_{s_1}$, some k exists such that $\mathcal{E}_P^k \in \mathcal{S}_{s_0}$. Otherwise $m_0 < m_1$ (Case 6), and $P \in \mathcal{S}_{s_1}$ with more than m_0 modes exists, with no escort distributions in $\mathcal{S}_{s_0}$; this will appear as a particular case in the proof of Theorem 2 below.



1. $\log 2 < s_1 < s_0 < \log 3$

4. $\log 2 < s_0 < s_1 < \log 3$

2. $0 < s_1 < s_0 < \log 2$

5. $0 < s_0 < s_1 < \log 2$

3. $0 < s_1 < \log 2 < s_0$

6. $0 < s_0 < \log 2 < s_1$

Fig. 4 Respective positions and forms of entropic spheres $\mathcal{S}_{s_0}$ and $\mathcal{S}_{s_1}$ for $s_0 > s_1$ (*left*), and $s_0 < s_1$ (*right*), with $|E| = 3$. *Cases 1–5* the distance between any P and $\mathcal{S}_{s_0}$ is obtained at one escort of P. *Case 6* ($s_0 < s_1$ and $m_0 < m_1$, with m_0 and m_1 defined by (15)): the distance between P_2 with 2 modes and $\mathcal{S}_{s_0}$ is obtained at $Q_2 = P_2^{\min}$ on the border of $\mathcal{S}_{s_0}$

The distributions $P_m^\nu \in \mathcal{D}$, with m modes and $\nu - m$ non-zero equal weights, will play a prominent role in that proof, especially for $\nu = N + 1$ and $m = 1$. In this aim, let us define for $(\nu, p) \in [m, N + 1] \times [1/(N + 1), 1/m]$,

$$\varphi_m(\nu, p) = -mp \log p - (1 - mp) \log \frac{1 - mp}{\nu - m}, \tag{16}$$

with $\varphi_m(m, 1/m) = \log m$. Note that $\varphi_m(\nu, p) = \mathbb{S}(P_m^\nu)$ for $\nu \in [\![m + 1, N]\!]$.

Basic arguments on entropy show that for all ν, the partial function $\varphi_m(\nu, \cdot)$ is decreasing and bijective from $[1/\nu, 1/m]$ onto $[\log m, \log \nu]$. Let $\psi_m(\nu, \cdot)$ be its reciprocal; for any $s \in [\log m, \log \nu]$, the quantities $\psi_m(\nu, s)$ are the modes' weights of the distributions P_m^ν such that $\mathbb{S}(P_m^\nu) = s$. Their determination will be of fundamental importance in the expression of $\mathbb{K}(\mathcal{S}_{s_0}|\mathcal{S}_{s_1})$. Since no closed-form expression can be obtained for the function ψ_m, this will finally involve a numerical procedure. The following technical results will be also of use.

Lemma 2 *For $m \in [\![1, N]\!]$, $\nu \in [m + 1, N]$ and $s \in [\log m, \log \nu]$, let ψ_m be the reciprocal of $\varphi_m(\nu, \cdot)$ defined by (16). Then $\psi_m(\nu, s) < \psi_m(N + 1, s)$ and $-\log \psi_m(\nu, s) < s$.*

Proof Note that $\nu = m = 1/p$ if and only if $e^s = m \in [\![1, N]\!]$; otherwise $\nu > m$.

First, due to the implicit function theorem, the differential of $\psi_m(\nu, s)$ with respect to ν is

$$\frac{\partial}{\partial \nu}\psi_m(\nu, s) = -\left[\frac{\partial}{\partial p}\varphi_m(\nu, p)\right]^{-1} \frac{\partial}{\partial \nu}\varphi_m(\nu, p) = -\left[m \log \frac{(1 - mp)}{(\nu - m)p}\right]^{-1} \frac{1 - mp}{\nu - m},$$

and hence $\frac{\partial}{\partial \nu}\psi_m(\nu, s) > 0$, so that $\psi_m(\nu, s)$ is increasing in ν. In particular, $\psi_m(\nu, s) \leq \psi_m(N + 1, s)$, with equality if and only if $\nu = N + 1$.

For proving the second inequality, let us study the function f_m^ν defined on $[1/\nu, 1/m]$ by $f_m^\nu(p) = -\log p - \varphi_m(\nu, p)$. For $p \in]1/\nu, 1/m[$, we compute

$$(f_m^\nu)'(p) = -\frac{1}{p} + m \log \frac{\nu - m}{\frac{1}{p} - m} \quad \text{and} \quad (f_m^\nu)''(p) = \frac{1}{p^2} + \frac{m}{p(1 - mp)} > 0,$$

with $(f_m^\nu)'(1/\nu) = -\nu$ and $(f_m^\nu)'(p)$ tending to infinity when p tends to $1/m$. Since $f_m^\nu(1/\nu) = f_m^\nu(1/m) = 0$, the function f_m^ν is negative.

In particular, $f_m^\nu(\psi_m(\nu, s)) = -\log \psi_m(\nu, s) - s < 0$. □

We can now state and prove the main result of the section.

Theorem 2 *Let $s_0 \neq s_1$ belong to $]0, \log(N + 1)[$. If $e^{s_1} \notin \mathbb{N}$, then the KL-divergence between two entropic spheres $\mathcal{S}_{s_0}$ and $\mathcal{S}_{s_1}$ defined in (1) is*

$$\mathbb{K}(\mathcal{S}_{s_0}|\mathcal{S}_{s_1}) = q^* \log \left(\frac{q^*}{p^*}\right) + (1 - q^*) \log \left(\frac{1 - q^*}{1 - p^*}\right) = \mathbb{K}(Q^*|P^*), \tag{17}$$

where the distributions $P^* \in \mathcal{D}^\circ$ *and* $Q^* \in \mathcal{D}^\circ$*, both with one mode (with respective weights* p^* *and* q^**) and N other equal weights, satisfy*

$$\mathbb{S}(P^*) = -p^* \log p^* - (1-p^*) \log \frac{1-p^*}{N} = s_1, \tag{18}$$

$$\mathbb{S}(Q^*) = -q^* \log q^* - (1-q^*) \log \frac{1-q^*}{N} = s_0. \tag{19}$$

Moreover, $\mathbb{K}(\mathcal{S}_{s_0}|\mathcal{S}_{s_1}) \leq s_1 - s_0$*, with equality if* $e^{s_1} \in \mathbb{N}$.

Note that Q^* is the escort distribution of $P^* \in \mathcal{S}_{s_1}$ belonging to $\mathcal{S}_{s_0}$.

Proof The infimum (1) is obviously attained on $\mathcal{D}^2$ and hence is non-negative and finite. Moreover, it is obtained at $(P, Q) \in \mathcal{D}^2$ such that the support of Q is included in the support of P, since otherwise $\mathbb{K}(Q|P)$ is infinite.

Let us set $\mathcal{D}^-_{s_0} = \{P \in \mathcal{D} : m \leq m_0\}$ where m_0 is defined in (15), and $\mathcal{D}^+_{s_0} = \mathcal{D} \backslash \mathcal{D}^-_{s_0}$, where m is the number of modes of P, with weight p. Theorem 1 says that if $P \in \mathcal{D}^+_{s_0}$ then (11) holds while if $P \in \mathcal{D}^-_{s_0}$ then (12) holds. Clearly,

$$\begin{aligned}\mathbb{K}(\mathcal{S}_{s_0}|\mathcal{S}_{s_1}) &= \min\left[\inf_{P \in \mathcal{S}_{s_1} \cap \mathcal{D}^-_{s_0}} \inf_{Q \in \mathcal{S}_{s_0}} \mathbb{K}(Q|P), \inf_{P \in \mathcal{S}_{s_1} \cap \mathcal{D}^+_{s_0}} \inf_{Q \in \mathcal{S}_{s_0}} \mathbb{K}(Q|P)\right] \\ &= \min\left[\mathbf{K}^-, \mathbf{K}^+\right].\end{aligned}$$

If $s_0 > \log m_1$ (Cases 1 to 5 of Fig. 4), then $\mathcal{S}_{s_1} \cap \mathcal{D}^+_{s_0}$ is empty so that $\mathbb{K}(\mathcal{S}_{s_0}|\mathcal{S}_{s_1}) = \mathbf{K}^-$, which we will determine later on.

Let us assume that $s_0 \leq \log m_1$ (and hence $m_0 < m_1$, Case 6). Theorem 1 says that $\inf_{Q \in \mathcal{S}_{s_0}} \mathbb{K}(Q|P) = -s - \log p$, for all $P \in \mathcal{S}_{s_1} \cap \mathcal{D}^+_{s_0}$, so that $\mathbf{K}^+ = \min_{P \in \mathcal{S}_{s_1} \cap \mathcal{D}^+_{s_0}} (-s - \log p)$. Writing $\mathcal{S}_{s_1} \cap \mathcal{D}^+_{s_0} = \bigcup_{\nu=m_1+1}^{N+1} \bigcup_{m=m_0+1}^{\nu} \mathcal{F}^\nu_m$, where $\mathcal{F}^\nu_m$ is the subset of $\mathcal{S}_{s_1} \cap \mathcal{D}^+_{s_0}$ of distributions with support of cardinal ν and m modes, we get

$$\mathbf{K}^+ = \min_{\nu \in \{m_1+1,\ldots,N+1\}} \min_{m \in \{m_0+1,\ldots,\nu\}} \min_{P \in \mathcal{F}^\nu_m} (-s_0 - \log p).$$

For all $\nu \in \{m_1+1, \ldots, N+1\}$, $\min_{m \in \{m_0+1,\ldots,\nu\}} \min_{P \in \mathcal{F}^\nu_m} (-s_0 - \log p)$ is obtained when p is maximum. Since $p < 1/m$, this maximum is obtained when m is minimum, that is for $m = m_0 + 1$. Applying the Lagrange multipliers method to the minimization of $-\log p$ subject to

$$\sum_{i \notin \mathcal{M}} P(i) + (m_0+1)p = 1 \text{ and } -\sum_{i \notin \mathcal{M}} P(i) \log P(i) - (m_0+1) p \log p = s_1,$$

where $\mathcal{M}$ is the set of modes of P, readily yields that all $P(i)$ for $i \notin \mathcal{M}$ are equal. Finally, Lemma 2 induces that

$$\mathbf{K}^+ = \min_{\nu \in \{m_1+1,\ldots,N+1\}} -s_0 - \psi_{m_0+1}(\nu, s_1) = -\log \psi_{m_0+1}(N+1, s_1) - s_0. \tag{20}$$

Now, let $\mathbf{K}^-$ be obtained at some $(P^-, Q^-) \in \mathcal{D}_{s_0}^- \times \mathcal{D}$. We will determine P^- and Q^- by taking advantage of the symmetric relation

$$\inf_{Q\in\mathcal{S}_{s_0}} \inf_{P\in\mathcal{S}_{s_1}\cap\mathcal{D}_{s_0}^-} \mathbb{K}(Q|P) = \inf_{P\in\mathcal{S}_{s_1}\cap\mathcal{D}_{s_0}^-} \inf_{Q\in\mathcal{S}_{s_0}} \mathbb{K}(Q|P).$$

On the one hand, Theorem 1 says that some $k > 0$ exists such that $Q^- = \mathcal{E}_{P^-}^k$, with the same support and number of modes as P^-. On the other hand, Proposition 4 says that some η exists such that $Q^- = -\eta P^- \log P^- + (1 - \eta s_1)P^-$. Therefore,

$$\frac{P^-(i)^k}{\Lambda(k)} = -\eta P^-(i) \log P^-(i) + (1 - \eta s_1)P^-(i), \quad i \in E.$$

Let us solve the equation $x^{k-1} = \Lambda(-\eta \log x - \eta s_1 + 1)$ in $x \in]0, 1[$, for any possible given values of $\eta \in \mathbb{R}^*$, $k \in \mathbb{R}_+^*$ and $\Lambda \in \mathbb{R}_+^*$. Setting $z = \log x + s_1 < s_1$, this is equivalent to solving in $z < s_1$ the equation $e^{-(k-1)s_1}e^{(k-1)z} = -\Lambda\eta z + \Lambda$. The intersection between the graph of $z \to e^{-(k-1)s_1}e^{(k-1)z}$ and the line $z \to \Lambda - \Lambda\eta z$ clearly contains at most two points $z < s_1$.

It contains one point if and only if $P^- \in \mathcal{S}_{s_1}$ is uniform on its support with cardinal $m_1 + 1$, and $e^{s_1} = m_1 + 1 \in \mathbb{N}$. Then $\mathbb{K}(Q^-|P^-) = \mathbb{S}(P^-) - \mathbb{S}(Q^-) = s_1 - s_0 \geq 0$, and hence $\mathbb{K}(\mathcal{S}_{s_0}|\mathcal{S}_{s_1}) = \mathbb{K}^- = \mathbb{K}^+ = s_1 - s_0$.

Otherwise, it contains two points and the searched distribution P^- takes exactly two values, say $P^-(i) = p_m^\nu \in]0, 1[$ for m indices i and $P^-(i) = (1-mp_m^\nu)/(\nu-m) < p_m^\nu$ for the $\nu - m$ others, where ν is the cardinal of the support of P^-. Hence, $P^- = P_m^\nu$ for some ν and m. Lemma 2 shows that p_m^ν is uniquely determined through $p_m^\nu = \psi_m(\nu, s_1)$. Lemma 1 implies that $\nu \geq \max(m_0, m_1) + 1$ and also that $m \leq \min(m_0, m_1)$. Therefore,

$$\begin{aligned}\mathbf{K}^- &= \min_\nu \min_m \mathbb{K}(Q_m^\nu|P_m^\nu),\\ &= \min_\nu \min_m \left[q_m^\nu \log\left(\frac{q_m^\nu}{p_m^\nu}\right) + (1 - q_m^\nu) \log\left(\frac{1 - q_m^\nu}{1 - p_m^\nu}\right)\right] \qquad (21)\end{aligned}$$

where $\nu \in \{\max(m_0, m_1) + 1, \ldots, N + 1\}$ and $m \in \{1, \ldots, \min(m_0, m_1)\}$, with Q_m^ν denoting the escort distribution of P_m^ν with entropy s_0, and $q_m^\nu = \psi_m(\nu, s_0)$ its modes' weight. Since ψ_m is an implicit function, the minimization problem (21) has to be solved numerically, which induces no computational difficulty. We have implemented a numerical procedure (available upon request) showing that, for any fixed values of N and $s_0 \neq s_1$, we have $\mathbb{K}(Q_m^\nu|P_m^\nu) > \mathbb{K}(Q_1^\nu|P_1^\nu) > \mathbb{K}(Q_1^{N+1}|P_1^{N+1})$ for all $m \in [\![2, \min(m_0, m_1)]\!]$ and $\nu \in [\![\max(m_0, m_1) + 1, N]\!]$. For illustration, Table 1 compares $\mathbb{K}(Q_m^\nu|P_m^\nu)$ for all $m \in [\![1, \min(m_0, m_1)]\!]$ and Table 2 compares $\mathbb{K}(Q_1^\nu|P_1^\nu)$ for all possible values of ν for $N = 4$.

Therefore, by setting $P^* = P_1^{N+1}$, $Q^* = Q_1^{N+1}$, with $q^* = q_{m_0}^1 = \psi_{m_0}(N+1, s_0)$ and $p^* = p_{m_1}^1 = \psi_{m_1}(N + 1, s_1)$, we get that $\mathbf{K}^- = \mathbb{K}(Q^*|P^*)$ in (17).

Again, since ψ_m is an implicit function, no closed-form expression can be obtained for p^*, q^* and $\psi_{m_0+1}(N + 1, s_1)$. Nevertheless, numerical comparison for any fixed

Table 1 Comparison of $\mathbb{K}(Q_m^\nu|P_m^\nu)$ for different entropic levels s_0 and s_1 and $\nu = |E| = 5$

m_0	s_0	m_1	s_1	m	$\mathbb{K}(Q_m^\nu|P_m^\nu)$
2	0.82830	2	0.79451	1	**0.000472**
2	0.82830	2	0.79451	2	0.001571
2	0.82830	3	1.17053	1	**0.052755**
2	0.82830	3	1.17053	2	0.082430
2	0.82830	4	1.44208	1	**0.220048**
2	0.82830	4	1.44208	2	0.273061
3	1.19451	2	0.79451	1	**0.080862**
3	1.19451	2	0.79451	2	0.207856
3	1.19451	3	1.17053	1	**0.000342**
3	1.19451	3	1.17053	2	0.000499
3	1.19451	3	1.17053	3	0.001333
3	1.19451	4	1.44208	1	**0.052137**
3	1.19451	4	1.44208	2	0.063211
3	1.19451	4	1.44208	3	0.082316
4	1.46068	2	0.79451	1	**0.297856**
4	1.46068	2	0.79451	2	0.687013
4	1.46068	3	1.17053	1	**0.073658**
4	1.46068	3	1.17053	2	0.104428
4	1.46068	3	1.17053	3	0.220123
4	1.46068	4	1.44208	1	**0.000499**
4	1.46068	4	1.44208	2	0.000599
4	1.46068	4	1.44208	3	0.000746
4	1.46068	4	1.44208	4	0.001536

values N and $s_0 \neq s_1$ easily shows that $\mathbb{K}(Q^*|P^*) \leq -\log \psi_{m_0+1}(N+1, s_1) - s_0$; for illustration, Table 3 presents the results for $3 \leq |E| \leq 7$. In other words, by using (20), we get $\mathbb{K}(\mathcal{S}_{s_0}|\mathcal{S}_{s_1}) = \mathbf{K}^- \leq \mathbf{K}^+$.

Finally, if $e^{s_1} \notin \mathbb{N}$, then $\mathbb{K}(Q^*|P^*) = s_1 - s_0 - (q^* - p^*)\log[Np^*/(1-p^*)]$. Since $q^* > p^*$ by Proposition 1 and $p^* > (1-p^*)/N$ by definition, we get that $\mathbb{K}(Q^*|P^*) \leq s_1 - s_0$. □

3 LDP for the sequence of plug-in estimators of entropy

Let $(X_1, \ldots, X_n)$ be an i.i.d. n-sample of $P \in \mathcal{D}$. Since only states with positive probability can be observed and estimated, we will suppose in the following that $P \in \mathcal{D}^\circ$. A natural estimator $\widehat{S}_n$ of $\mathbb{S}(P)$ is obtained by plug-in from the empirical estimator

$$\widehat{P}_n(i) = \frac{1}{n}\sum_{m=1}^{n} \mathbb{1}_{\{X_m=i\}}, \quad i \in E, \tag{22}$$

through

$$\widehat{S}_n = \mathbb{S}(\widehat{P}_n) = S(\widehat{P}'_n), \tag{23}$$

Table 2 Comparison of $\mathbb{K}(Q_1^\nu | P_1^\nu)$ for different entropic levels s_0 and s_1 and $\nu \leq |E| = 5$

m_0	s_0	m_1	s_1	ν	$\mathbb{K}(Q_1^\nu \| P_1^\nu)$
1	0.231049	1	0.173287	2	0.004300
1	0.231049	1	0.173287	3	0.003032
1	0.231049	1	0.173287	4	0.002595
1	0.231049	1	0.173287	5	**0.002357**
1	0.231049	2	0.794513	3	0.183667
1	0.231049	2	0.794513	4	0.134231
1	0.231049	2	0.794513	5	**0.113772**
1	0.231049	3	1.17053	4	0.409901
1	0.231049	3	1.17053	5	**0.309968**
2	0.828302	1	0.173287	3	0.440681
2	0.828302	1	0.173287	4	0.318929
2	0.828302	1	0.173287	5	**0.269408**
2	0.828302	2	0.794513	3	0.001131
2	0.828302	2	0.794513	4	0.000620
2	0.828302	2	0.794513	5	**0.000472**
2	0.828302	3	1.17053	4	0.081904
2	0.828302	3	1.17053	5	**0.052755**
3	1.19451	1	0.173287	4	0.919693
3	1.19451	1	0.173287	5	**0.700784**
3	1.19451	2	0.794513	4	0.125149
3	1.19451	2	0.794513	5	**0.080862**
3	1.19451	3	1.17053	4	0.000689
3	1.19451	3	1.17053	5	**0.000342**

where $\widehat{P}_n' = (\widehat{P}_n(1), \dots, \widehat{P}_n(N))$ and S is explicitly defined by (6).

The aim of this section is to state an LDP for the sequence $(\widehat{S}_n)$. It will be based on the minimum of KL-divergence computed in Theorem 1. First, we will recall the asymptotic properties of $(\widehat{S}_n)$. Then, since no closed-form expression of the rate function is available, we will construct an explicit approximating function.

3.1 The large deviations principle

The plug-in empirical estimator $\widehat{S}_n$ of $\mathbb{S}(P)$ has been first considered in the 1950s. Basharin (1959) proves that it is biased, but strongly consistent and asymptotically normal. As a particular case of a complicated series scheme of observations, Zubkov (1973) shows asymptotic normality holds only if P is not uniform on E, that is if entropy is not maximum; see also Harris (1977) and the references therein. Therefore, we here present the proof of the asymptotic properties of $\widehat{S}_n$ only for the uniform distribution.

Theorem 3 *Let $\widehat{P}_n$ denote the empirical estimator defined in* (22) *of $P \in \mathcal{D}^\circ$. The plug-in estimator $\widehat{S}_n = \mathbb{S}(\widehat{P}_n)$ is a strongly consistent estimator of the entropy $\mathbb{S}(P)$. Moreover*:

Table 3 $\mathbb{K}(Q_1^{N+1}|P_1^{N+1}) = \mathbf{K}^- \leq \mathbf{K}^+ = -\log \psi_{m_0+1}(N+1, s_1) - s_0$ for different values of $|E|$, s_0 and s_1

$\lvert E\rvert$	m_1	s_1	m_0	s_0	$\mathbf{K}^-$	$\mathbf{K}^+$
3	2	0.82830	1	0.17328	0.244192	0.558585
4	2	0.82830	1	0.17328	0.178434	0.550442
4	3	1.19451	1	0.17328	0.481118	0.953279
4	3	1.19451	2	0.79451	0.112046	0.332052
5	2	0.82830	1	0.17328	0.151577	0.547245
5	3	1.19451	1	0.17328	0.362760	0.947297
5	3	1.19451	2	0.79451	0.071617	0.326071
5	4	1.46068	1	0.17328	0.678375	1.234900
5	4	1.46068	2	0.79451	0.260531	0.613669
5	4	1.46068	3	1.17053	0.072280	0.237650
6	2	0.82830	1	0.17328	0.136195	0.545395
6	3	1.19451	1	0.17328	0.309936	0.944968
6	3	1.19451	2	0.79451	0.056527	0.323741
6	4	1.46068	1	0.17328	0.523632	1.230170
6	4	1.46068	2	0.79451	0.178933	0.608940
6	4	1.46068	3	1.17053	0.043294	0.232921
6	5	1.67021	1	0.17328	0.844839	1.454140
6	5	1.67021	2	0.79451	0.400629	0.832916
6	5	1.67021	3	1.17053	0.178487	0.456896
6	5	1.67021	4	1.44208	0.053203	0.185349
7	2	0.82830	1	0.17328	0.125955	0.544140
7	3	1.19451	1	0.17328	0.278447	0.943624
7	3	1.19451	2	0.79451	0.048375	0.322397
7	4	1.46068	1	0.17328	0.450619	1.228330
7	4	1.46068	2	0.79451	0.144658	0.607107
7	4	1.46068	3	1.17053	0.032876	0.231088
7	5	1.67021	1	0.17328	0.663920	1.450230
7	5	1.67021	2	0.79451	0.287511	0.829005
7	5	1.67021	3	1.17053	0.116686	0.452985
7	5	1.67021	4	1.44208	0.030603	0.181438
7	6	1.84314	1	0.17328	0.988230	1.633750
7	6	1.84314	2	0.79451	0.527289	1.012520
7	6	1.84314	3	1.17053	0.285612	0.636501
7	6	1.84314	4	1.44208	0.135490	0.364953
7	6	1.84314	5	1.65502	0.042044	0.152015

If P is not uniform, then $\sqrt{n}[\widehat{\mathbb{S}}_n - \mathbb{S}(P)]$ converges in distribution to a centered normal distribution with variance

$$\Sigma_{\mathbb{S}}^2 = \sum_{i=1}^{N} \left[\log \frac{P(i)}{1 - \sum_{j=1}^{N} P(j)} \right]^2 P(i)[1 - P(i)].$$

If $P = U$ is uniform, then $2n[\widehat{S}_n - \mathbb{S}(U)]$ converges to $\sum_{i=1}^{N} \beta_i Y_i$, where the Y_i are i.i.d. $\chi^2(1)$-distributed random variables and $\beta_i \in \mathbb{R}$ for $i \in [\![1, N]\!]$.

Proof When $P = U$, the asymptotic distribution of $2n[\widehat{S}_n - \mathbb{S}(U)]$ derives from the second-order Taylor expansion of $\mathbb{S}$ at U. Indeed,

$$\widehat{S}_n - \mathbb{S}(U) = D_{\mathbb{S}}(U)(\widehat{P}_n - U) + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{\partial^2}{\partial P(i)\partial P(j)}\mathbb{S}(P)\bigg|_{P=U} \right.$$
$$\left. \times \left[\widehat{P}_n(i) - \frac{1}{N+1}\right]\left[\widehat{P}_n(j) - \frac{1}{N+1}\right]\right) + o_{\mathbb{P}}(\|\widehat{P}_n - U\|^2).$$

Since entropy is maximum at U, the differential $D_{\mathbb{S}}(U)$ is null, and we get

$$2n[\widehat{S}_n - \mathbb{S}(U)] = \sum_{i=1}^{N}\sum_{j=1}^{N}\left[\frac{\partial^2}{\partial P(i)\partial P(j)}\mathbb{S}(U)\right]\mathbf{P}_{n,i}\mathbf{P}_{n,j} + o_{\mathbb{P}}(\|\mathbf{P}_n\|^2),$$

where $\mathbf{P}_n = (\mathbf{P}_{n,i})_{i\in[\![1,N]\!]}$, with $\mathbf{P}_{n,i} = \sqrt{n}[\widehat{P}_n(i) - 1/(N+1)]$.

The covariance matrix of $\mathbf{P}_n$ induces a scalar product over $\mathbb{R}^N$, while $D^2_{\mathbb{S}}(U)$ induces a quadratic form. Hence the simultaneous diagonalization theorem of quadratic forms applies: a basis $\mathbf{B}$ of $\mathbb{R}^N$ exists which is orthonormal for the covariance and orthogonal for the Hessian quadratic form. Let $(\alpha_{i,j})_{(i,j)\in[\![1,N]\!]^2}$ denote the change of basis matrix from the canonical matrix to $\mathbf{B}$ and let $(\beta_i)_{i\in[\![1,N]\!]}$ be the diagonal coefficients of $D^2_{\mathbb{S}}(U)$ in $\mathbf{B}$. Then $2n[\widehat{S}_n - \mathbb{S}(U)] = \sum_{i=1}^{N} \beta_i Z_i^2 + o_{\mathbb{P}}(\|\mathbf{P}_n\|^2)$, where $Z_i = \sum_{j=1}^{N} \alpha_{i,j}\mathbf{P}_{n,j}$ are uncorrelated random variables with variance 1 for all $i \in [\![1, N]\!]$. Since $\mathbf{P}_n$ is asymptotically a Gaussian vector, each sum $\sum_{j=1}^{N} \alpha_{i,j}\mathbf{P}_{n_j}$ is asymptotically normal and hence $Y_i = Z_i^2$ is asymptotically $\chi^2(1)$-distributed.

Finally, thanks to Prohorov's theorem (see, e.g., Van der Vaart 1998), since $\mathbf{P}_n$ is asymptotically Gaussian with zero mean and diagonal variance, $\|\mathbf{P}_n\|^2$ converges in distribution to a $\chi^2(N)$-distribution with $o_{\mathbb{P}}(\|\mathbf{P}_n\|^2) = o_{\mathbb{P}}(1)$. The conclusion follows from Slutsky's theorem. □

The sequence $(\widehat{S}_n)$ satisfies an LDP with good rate function depending on the number of modes of P: it is either the KL-divergence with respect to P of one of its escort distributions or minus the entropy level minus the logarithm of the modes' weight.

Theorem 4 *Let $(X_n)_{n\in\mathbb{N}^*}$ be a sequence of i.i.d. random variables taking values in a finite set E, with distribution $P \in \mathcal{D}^\circ$.*

The sequence of estimators $(\widehat{S}_n)_{n\in\mathbb{N}^*}$ defined in (23) of the entropy $\mathbb{S}(P)$ satisfies the large deviations principle

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\widehat{S}_n \in A) = -\inf_{s\in A} I_{\mathbb{S}}(s, P), \qquad (24)$$

for all Borel sets $A \subseteq [0, \log(N+1)]$ *with non-empty interior, with good rate function* $I_{\mathbb{S}}$ *defined by*

$$I_{\mathbb{S}}(s, P) = \begin{cases} -s - \log p & \text{if } 0 \leq s \leq \log m, \\ \mathbb{K}(\mathcal{E}_P^k | P) & \text{if } \log(m) < s \leq \log(N+1), \ \textit{with } k > 0 \textit{ such} \\ & \textit{that } \mathbb{S}(\mathcal{E}_P^k) = s, \\ +\infty & \text{otherwise,} \end{cases} \tag{25}$$

where m is the number of modes of P, *with weight* p.

If $P = U$ *is uniform, then* $I_{\mathbb{S}}(s, U) = \log(N+1) - s$.

Proof Since $\mathbb{S}$ is a continuous function from $\mathcal{D}$ to $[0, \log(N+1)]$, a straightforward application of both Sanov's theorem and the contraction principle (see, e.g., Dembo and Zeitouni 1998) yields

$$\mathbb{P}(\widehat{S}_n \in A) \leq \binom{n+N}{N} \exp\left[-n \inf_{s \in \overline{A}} I_{\mathbb{S}}(s, P)\right] \tag{26}$$

where $I_{\mathbb{S}}(s, P) = \mathbb{K}(\mathcal{S}_s | P)$, so

$$\limsup_{n \to \infty} \frac{1}{n} \log \mathbb{P}\left(\widehat{S}_n \in A\right) \leq - \inf_{s \in \overline{A}} I_{\mathbb{S}}(s, P), \tag{27}$$

and also

$$- \inf_{s \in A^\circ} I_{\mathbb{S}}(s, P) \leq \liminf_{n \to \infty} \frac{1}{n} \log \mathbb{P}\left(\widehat{S}_n \in A\right), \tag{28}$$

and hence Theorem 1 gives (25).

Proposition 2 says that $\mathbf{s}_P$ is a bijection from $\mathbb{R}_+^*$ to $]\gamma, \Gamma[=]\log(m), \log(N+1)[$ with twice continuously differentiable inverse $\mathbf{s}_P^{-1}$. Since for all $s \in]\gamma, \Gamma[$, we have $I_{\mathbb{S}}(s, P) = \mathbb{K}\left(\mathcal{E}(P, \mathbf{s}_P^{-1}(s)) | P\right)$, we get that $I_{\mathbb{S}}$ too is twice continuously differentiable with respect to s on $]\gamma, \Gamma[$.

In particular, $I_{\mathbb{S}}$ is continuous on $]\gamma, \Gamma[$. It is continuous on $[0, \gamma[$ too, as a linear mapping, and finally at $\log(m)$ as a straightforward consequence of Proposition 1 and continuity of the function $Q \to \mathbb{K}(Q|P)$. The continuity of $I_{\mathbb{S}}$ induces that $I_{\mathbb{S}}$ is a good rate function and also that $\inf_{s \in A^\circ} I_{\mathbb{S}}(s, P) = \inf_{s \in \overline{A}} I_{\mathbb{S}}(s, P)$ so that the lower and upper bounds in (27) and (28) are equal, and hence (24) follows. □

Note that Chazottes and Gabrielli (2005) Proposition 3.3 gives the good rate function governing the LDP of the plug-in estimators of Shannon entropy of the so-called g-measures. This may be applied to distributions on a finite set E. Nevertheless, even if this rate function is shown to be linear for small values, neither the threshold nor the role of the escort distributions are explained.

3.2 Approximation of the rate function

To obtain an explicit expression of the rate function $I_{\mathbb{S}}$, it would be necessary to obtain first an explicit expression of the inverse of $\mathbf{s}_P$. As an alternative, we will build an approximation of this function. For simplification, we will denote $I_{\mathbb{S}}(s, P)$ by $I_{\mathbb{S}}(s)$ in this section.

Since $\mathbb{S}(\mathcal{E}_P^k)$ is a decreasing function of k from Γ to γ, the equation $\mathbb{S}(\mathcal{E}_P^k) = s$ has a unique solution for any $s \in]\gamma, \Gamma[$. In other words, the infimum $I_{\mathbb{S}}(s)$ is achieved at a unique escort distribution. Let us set $y_l = \gamma + l(\Gamma - \gamma)/M$, thus defining a uniformly distributed finite set $\{y_1, \dots, y_{M-1}\}$ in $]\gamma, \Gamma[$. Since $\mathbf{s}_P$ is monotonic and smooth, numerous numerical methods are available which provide approximated solutions of $\mathbf{s}_P(k) = y_l$, for $l \in [\![1, M-1]\!]$. We are interested here in the effect of this approximation on the LDP, and not in discussing the best possible method of approximation; the point is that approximated solutions $\widetilde{k}_1, \dots, \widetilde{k}_{M-1}$ as close as necessary to the true solutions $k_1, \dots, k_{M-1}$ can be obtained. If $\xi > 0$ is the required accuracy given in (33) below, we will choose $\widetilde{k}_l$ such that $|\widetilde{k}_l - k_l| < \xi$ for all l. An approximation $I_M(y_l)$ of $I_{\mathbb{S}}(y_l)$ is obtained by setting $I_M(y_l) = \mathbb{K}(\mathcal{E}_P^{\widetilde{k}_l}|P)$ for $l \in [\![1, M-1]\!]$. The approximation of $I_{\mathbb{S}}$ will be the interpolating piece-wise affine and continuous function I_M built from the above points.

Proposition 5 *Set $\alpha \in [0, 1]$ and $y_l = \gamma + l(\Gamma - \gamma)/M$, for $l \in [\![1, M-1]\!]$. Let I_M be defined*:

for $s = \alpha y_1 + (1 - \alpha)\gamma$ by

$$I_M(s) = \alpha I_M(y_1) - (1-\alpha)\left(\log\left[\max_{i\in E}\{P(i)\}\right] + \gamma\right);$$

for $s = \alpha\Gamma + (1-\alpha)y_{M-1}$ by

$$I_M(s) = -\alpha\left[\Gamma + \frac{1}{N+1}\sum_{i\in E}\log P(i)\right] + (1-\alpha)I_M(y_{M-1});$$

and for $s = \alpha y_l + (1 - \alpha y_{l-1})$ by

$$I_M(s) = \alpha I_M(y_l) + (1-\alpha)I_M(y_{l-1}).$$

Then, for all $s \in [y_1, y_{M-1}]$,

$$|I_M(s) - I_{\mathbb{S}}(s)| \le \frac{2}{M}, \tag{29}$$

with $M \ge (\Gamma - \gamma)^2 \max\{1, \sup_{s\in[y_1, y_{M-1}]} I_{\mathbb{S}}''(s)/8\}$, where $I_{\mathbb{S}}''(s)$ is the second order derivative of $I_{\mathbb{S}}$.

Proof For getting an upper bound for $|I_M(s) - I_{\mathbb{S}}(s)|$ for all $s \in [y_1, y_{M-1}]$, let us consider the linear interpolation $\widetilde{I}_{\mathbb{S}}$ of $I_{\mathbb{S}}$ built from $I_{\mathbb{S}}(y_l)$ for $l \in [\![1, M-1]\!]$. We have

$$|I_M(s) - I_{\mathbb{S}}(s)| \leq |I_M(s) - \widetilde{I}_{\mathbb{S}}(s)| + |\widetilde{I}_{\mathbb{S}}(s) - I_{\mathbb{S}}(s)|.$$

We will separately bound $|\widetilde{I}_{\mathbb{S}}(s) - I_{\mathbb{S}}(s)|$, the loss inherent to linear interpolation, and $|I_M(s) - \widetilde{I}_{\mathbb{S}}(s)|$, resulting from solving $\mathbb{S}(\mathcal{E}_P^k) = y_l$ numerically.

First, since $I_{\mathbb{S}}$ is twice continuously differentiable with respect to s,

$$|\widetilde{I}_{\mathbb{S}}(s) - I_{\mathbb{S}}(s)| \leq \frac{(\Gamma - \gamma)^2}{8M^2} C, \quad s \in [y_1, y_{M-1}], \tag{30}$$

where $C = \sup_{s \in [y_1, y_{M-1}]} |I''_{\mathbb{S}}(s)|$.

Second, $s = \alpha y_{l+1} + (1-\alpha) y_l$ for some $l \in [\![1, M-2]\!]$, and hence,

$$\begin{aligned} |I_M(s) - \widetilde{I}_{\mathbb{S}}(s)| &= |\alpha(I_M(y_{l+1}) - I_{\mathbb{S}}(y_{l+1})) + (1-\alpha)(I_M(y_l) - I_{\mathbb{S}}(y_l))| \\ &\leq \alpha |I_M(y_{l+1}) - I_{\mathbb{S}}(y_{l+1})| + (1-\alpha)|I_M(y_l) - I_{\mathbb{S}}(y_l)|. \end{aligned} \tag{31}$$

We compute $|I_M(y_l) - I_{\mathbb{S}}(y_l)| = |\mathbb{K}(\mathcal{E}_P^{\widetilde{k}_l}|P) - \mathbb{K}(\mathcal{E}_P^k|P)|$ for $l \in [\![1, M-1]\!]$, with k such that $\mathbb{S}(\mathcal{E}_P^k) = s$. Since $\mathbb{K}(\mathcal{E}(P, .)|P)$ is twice continuously differentiable with respect to k on $\mathbb{R}_+^*$, it is a Lipschitz function and

$$D = \sup_{(k,k') \in \mathbf{s}_P^{-1}([y_1, y_{m-1}])^2} \frac{1}{|k - k'|} |\mathbb{K}(\mathcal{E}_P^k|P) - \mathbb{K}(\mathcal{E}_P^{k'}|P)|$$

is finite, so that $|I_M(y_l) - I_{\mathbb{S}}(y_l)| \leq D\xi$ for $l \in [\![1, M-1]\!]$, where ξ is the accuracy to be fixed of the numerical method used for solving the equations $\mathbb{S}(\mathcal{E}_P^{k_l}) = y_l$. This in turn gives in (31)

$$|I_M(s) - \widetilde{I}_{\mathbb{S}}(s)| \leq \alpha D\xi + (1-\alpha) D\xi = D\xi. \tag{32}$$

Inequalities (30) and (32) together lead to

$$|I_M(s) - I_{\mathbb{S}}(s)| \leq \frac{(\Gamma - \gamma)^2}{8M^2} C + D\xi, \quad s \in [y_1, y_{M-1}].$$

For $M \geq (\Gamma - \gamma)^2 C/8$, the left term of the above sum is upper bounded by $1/M$. Finally, choosing

$$\xi \leq 1/DM, \tag{33}$$

we get the searched inequality (29). □

We have illustrated this approximation through simulation. For showing how I_M depends on P, we have chosen five distributions on a space E with four elements. The first three ones have one mode: $P_1 = (0.5, 0.1, 0.05, 0.35)$ is taken at random,

$P_2 = (0.2, 0.3, 0.22, 0.28)$ is closed to uniform and $P_3 = (0.95, 0.04, 0.007, 0.003)$ is closed to a Dirac. The fourth distribution, $P_4 = (0.4, 0.4, 0.15, 0.05)$, has two modes and the fifth one, $P_5 = (0.3, 0.3, 0.3, 0.1)$, has three modes. The function I_M is constructed by computing the solutions of the equations $\mathbf{s}_P(k) = y_l$ thanks to the dichotomy method.

Proposition 5 states that $I_{\mathbb{S}}$ lies between $I_M + 2/M$ and $I_M - 2/M$ for M large enough. Figure 5 shows on the left the curves $I_M + 1/M$ and $I_M - 1/M$ for P_1 and $M = 50$ (top) and $M = 500$ (bottom); actually, the approximation is so good for $M = 500$ that differentiating between the two curves is impossible. In the middle I_M is shown for $M = 500$ for P_2 (top) and P_3 (bottom) and finally on the right for P_4 with two modes (top) and P_5 with three modes (bottom). As expected, the shapes of the different curves are significantly different.

4 Application to entropy level testing

For showing one possible application, let us construct tests of entropy levels based on both the LDP and the double minimization of KL-divergence. Other statistical applications of LDP may be found in Birgé et al. (1979).

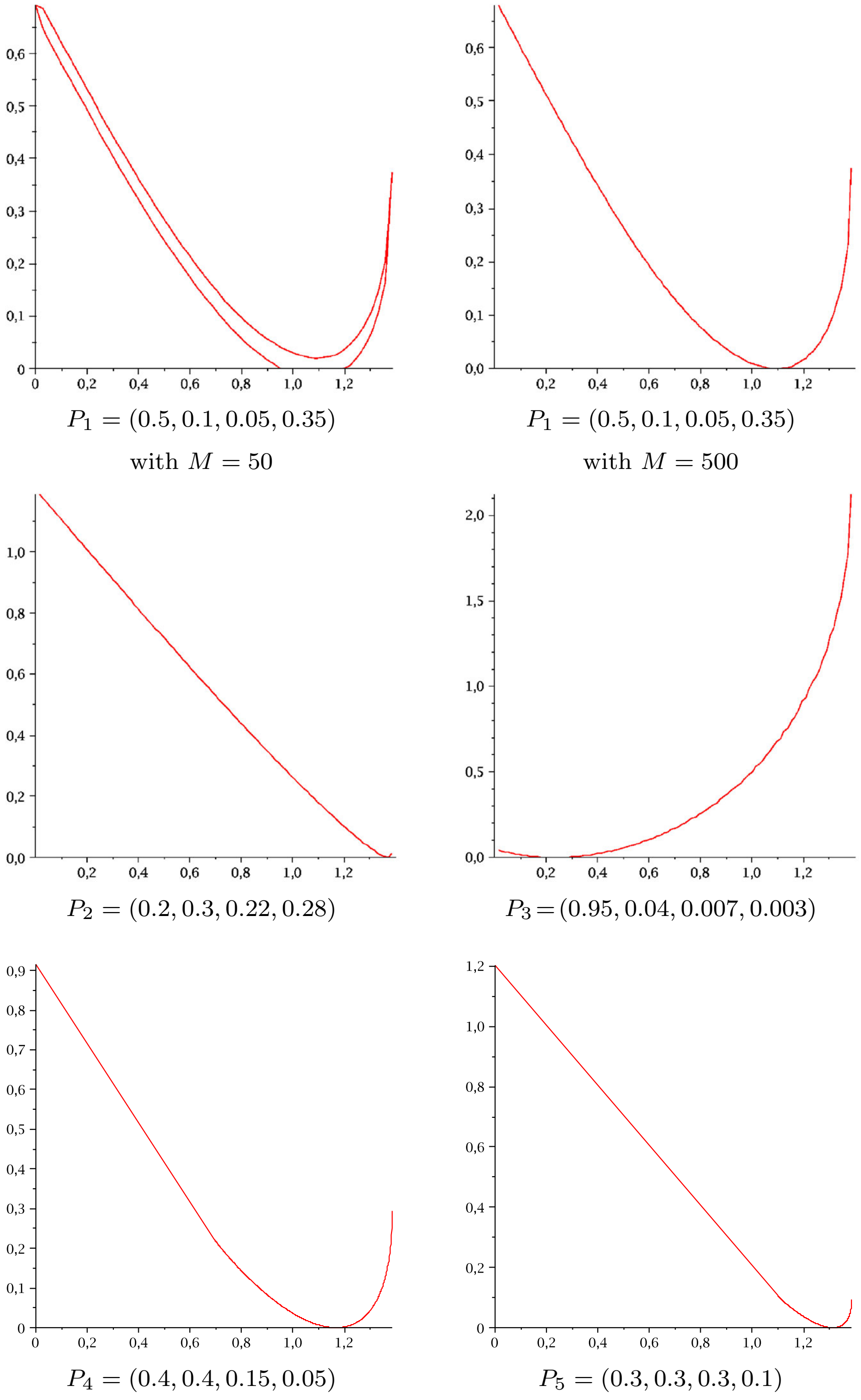
In goodness-of-fit testing, statistics based on difference of entropy of distributions are usual for discriminating between distributions. Now applied to all classical distributions, they have been developed from Vasicek (1976) for testing normality, in relation with the maximum entropy principle; see Lequesne (2015) for details and fields of application.

In data compression theory, the entropy of a distribution P is well known to be the lower bound of the average length per symbol of any encoded i.i.d. sequence of symbols drawn according to P; see Cover and Thomas (1991). Several codes, such as Huffman code, achieve asymptotically that lower bound, provided that P is known. If P is unknown, universal codes still compress the sequence up to a limit rate s, provided that s is more than the entropy of P. It may then be necessary to decide, according to observation of an i.i.d. sample, whether $\mathbb{S}(P) < s$ or not. Such a decision can be taken from testing the entropy level of P. The entropy level tests are also classical in testing random numbers for randomness; see Rukhin et al. (2010)[1].

In biology, the potential application of Shannon entropy as a measure of diversity has early been acknowledged, for example in Pielou (1967). Indeed, an obvious analogy exists between a biological collection consisting of various numbers of different species of organisms, and a coded message consisting of various numbers of different kinds of symbols. Identifying the members of a collection to the right species is formally identical to identifying the symbols in a message, one by one. The total diversity of a collection of n individuals belonging to $N + 1$ species with $\mathbf{N}_n(i)$ individuals in the ith species is given by Brillouin's formula

$$B = \log \frac{n!}{\mathbf{N}_n(0)! \dots \mathbf{N}_n(N)!},$$

[1] see also http://www.random.org/analysis/.



$P_1 = (0.5, 0.1, 0.05, 0.35)$ with $M = 50$

$P_1 = (0.5, 0.1, 0.05, 0.35)$ with $M = 500$

$P_2 = (0.2, 0.3, 0.22, 0.28)$

$P_3 = (0.95, 0.04, 0.007, 0.003)$

$P_4 = (0.4, 0.4, 0.15, 0.05)$

$P_5 = (0.3, 0.3, 0.3, 0.1)$

Fig. 5 The original and approximating rate functions for different distributions

which can be approximated through Stirling's formula by the plug-in estimator $\widehat{S}_n$ of Shannon entropy of the population under study. A test on the entropy level of the collection then constitutes a first approach to decide whether it comes from a known population or not.

For tests of entropy level (3) to (5), we consider rejection regions C_n^1 and C_n^2 obtained by assuming that the divergence $\mathbb{K}(\mathcal{S}_{\widehat{S}_n}|\mathcal{S}_{s_0})$ between the entropic spheres with entropy levels $\widehat{S}_n$ and the null hypothesis s_0 is greater than a threshold depending on the number of observations. The error of the first kind, namely $\alpha_n^j = \sup_{P\in\mathcal{S}_{s_0}} P^n\left(C_n^j\right)$, will be shown to decrease with $1/n$. For the test (3), the error of the second kind $\beta_n^1 = \sup_{P\in\mathcal{S}_{s_1}} P^n(E^n\backslash C_n^1)$ will be shown to decrease exponentially fast with the KL-divergence $\mathbb{K}(\mathcal{S}_{s_0}|\mathcal{S}_{s_1})$. The proofs will derive from a slight modification of the classical proof of Sanov's theorem; see Csiszár and Shieds (2004). The tests are thus proven to be consistent.

Theorem 5 *Let $(X_1,\ldots,X_n)$ be an n-sample of a random variable X with distribution P supported by $E = [\![0, N]\!]$.*

For both tests (3) and (4), let the critical region be

$$C_n^1 = \left\{\mathbb{K}(\mathcal{S}_{\widehat{S}_n}|\mathcal{S}_{s_0}) \geq \delta_n\right\}, \tag{34}$$

where $\widehat{S}_n$ is defined in (23), with

$$\delta_n = \frac{1}{n}\log\left[n\binom{n+N}{N}\right]. \tag{35}$$

The error of the first kind α_n^1 satisfies

$$\alpha_n^1 \leq 1/n. \tag{36}$$

For the test (3), the error of the second kind β_n^1 linked to C_n^1 satisfies

$$\limsup_{n\to\infty}\frac{1}{n}\log\beta_n^1 \leq -\mathbb{K}(\mathcal{S}_{s_0}|\mathcal{S}_{s_1}). \tag{37}$$

For the test (5), let the critical region be

$$C_n^2 = \begin{cases}\left\{\mathbb{K}(\mathcal{S}_{\widehat{S}_n}|\mathcal{S}_{s_0}) \geq \delta_n\right\} & \textit{if} \quad \widehat{S}_n < s_0,\\ \emptyset & \textit{if} \quad \widehat{S}_n \geq s_0.\end{cases}$$

The error of the first kind satisfies $\alpha_n^2 \leq 1/n$, with δ_n given by (35).

Proof Let $P \in \mathcal{S}_{s_0}$. Obviously, $P^n(C_n^1) \leq \mathbb{P}(\widehat{S}_n \in A_n)$, where $A_n = \{s \in [0, \log(N+1)] : I_{\mathbb{S}}(s, P) \geq \delta_n\}$. Thanks to (26), by definition of δ_n,

$$P^n(C_n^1) \leq \binom{n+N}{N}\exp\left(-n\delta_n\right) = \frac{1}{n}.$$

Since this relation holds for all $P \in \mathcal{S}_{s_0}$, (36) is shown.

It remains to prove (37). Since the empirical distribution $\widehat{P}_n$ built from any sequence of observations $x_1^n = (x_1, \ldots, x_n)$ clearly belongs to the set

$$\mathcal{Q} = \{Q \in \mathcal{D} : Q(i) = k_i/n, i \in E, k_i \in \mathbb{N}\},$$

the only possible values for the estimators $\widehat{S}_n = \mathbb{S}(\widehat{P}_n)$ are $\mathbb{S}(Q)$, with $Q \in \mathcal{Q}$. Let $R \in \mathcal{S}_{s_1}$, with $s_1 \neq s_0$. Then

$$R^n\left(E^n \setminus C_n^1\right) = R^n\left(\bigcup_{Q \in \mathcal{Q}(\delta_n)} T_Q\right) = \sum_{Q \in \mathcal{Q}(\delta_n)} R^n(T_Q),$$

where $\mathcal{Q}(\delta_n) = \left\{Q \in \mathcal{Q} : \mathbb{K}(\mathcal{S}_{\mathbb{S}(Q)}|\mathcal{S}_{s_0}) < \delta_n\right\}$ and $T_Q = \{x_1^n = (x_1, \ldots, x_n) \in E^n : \widehat{P}_n = Q\}$.

For any $x_1^n \in T_Q$, the number of x_k such that $x_k = i$ is $nQ(i)$, so that

$$\frac{R^n(x_1^n)}{Q_n(x_1^n)} = \prod_{i \in E}\left[\frac{R(i)}{Q(i)}\right]^{nQ(i)} = \mathrm{e}^{-n\mathbb{K}(Q|R)},$$

and hence for all $Q \in \mathcal{Q}(\delta_n)$,

$$R^n(T_Q) \leq Q^n(T_Q)\mathrm{e}^{-n\mathbb{K}(Q|R)} \leq \exp\left[-n \inf_{R \in \mathcal{S}_{s_1}} \mathbb{K}(Q|R)\right] \leq \mathrm{e}^{-n\eta_n},$$

where $\eta_n = \inf_{Q \in \mathcal{Q}(\delta_n)} \inf_{R \in \mathcal{S}_{s_1}} \mathbb{K}(Q|R)$.

Using some simple combinatorics, we get

$$R^n\left(E^n \setminus C_n^1\right) \leq |\mathcal{Q}(\delta_n)|\mathrm{e}^{-n\eta_n} \leq \binom{n+N}{N}\mathrm{e}^{-n\eta_n}.$$

For all $Q \in \mathcal{Q}(\delta_n)$, when n tends to infinity, $\mathbb{S}(Q)$ converges to s_0, and hence η_n converges to $\inf_{Q \in \mathcal{S}_{s_0}} \inf_{R \in \mathcal{S}_{s_1}} \mathbb{K}(Q|R) = \mathbb{K}(\mathcal{S}_{s_0}|\mathcal{S}_{s_1})$. Therefore,

$$\limsup_{n \to \infty} \frac{1}{n} \log R^n\left(E^n \setminus C_n^1\right) \leq -\mathbb{K}(\mathcal{S}_{s_0}|\mathcal{S}_{s_1}),$$

and the searched inequality is proven.

The proof for the test (5) follows the same lines. □

Finally, let us give illustrative examples of entropy level tests. For the test (4), the rejection region is simply given in (34). For the test (3) of entropy level $s_0 = \log 2/2 \simeq 0.347$ for $N = 3$, with significance level 0.05, we have chosen to take as alternatives the entropy levels of distributions P_1 to P_5 already considered in Sect. 3.2 and another

Table 4 Critical sample sizes for Test (4) for different alternatives, with signification level 0.05, from 1,000 simulated samples

Sampled distribution	Entropy of P_j	Divergence	Critical sample size n
$P_1 = (0.5, 0.1, 0.05, 0.35)$	1.094	0.399	38
$P_2 = (0.2, 0.3, 0.22, 0.28)$	1.373	0.829	12
$P_3 = (0.95, 0.04, 0.007, 0.003)$	0.23	0.00866	3192
$P_4 = (0.4, 0.4, 0.15, 0.05)$	1.167	0.489	29
$P_5 = (0.3, 0.3, 0.3, 0.1)$	1.314	0.713	18
$P_6 = (0.5, 0.167, 0.167, 0.167)$	0.94	0.284	64

distribution $P_6 = (0.7, 0.1, 0.1, 0.1)$ with one mode and all other equal weights. Their entropy levels are given in Table 4.

Samples of increasing size $n \geq 10$ are drawn through simulation according to each distribution P_1 to P_6, yielding the related plug-in estimator $\widehat{S}_n$. The divergence $\mathbb{K}(\mathcal{S}_{\widehat{S}_n}|\mathcal{S}_{s_0})$ is then computed through (17), by first numerically solving (18) in p^* with $s_1 = s_0$ and (19) in q^* with $s_0 = \widehat{S}_n$, as stated in Theorem 2. Applying inequality (26) gives the following upper bound $B(n)$ for the p values,

$$B(n) = \binom{n+N}{N} e^{-n\mathbb{K}\left(\mathcal{S}_{\widehat{S}_n}|\mathcal{S}_{s_0}\right)}, \tag{38}$$

that becomes of use only if less than 1.

Table 4 gives numerical values of $\mathbb{K}(\mathcal{S}_{\mathbb{S}(P_i)}|\mathcal{S}_{s_0})$ for $i \in [\![1, 6]\!]$ – denoted by Divergence, and the (critical) sample sizes obtained by the Monte Carlo method, through simulation of 1,000 samples of increasing size for P_1 to P_6. These critical sample sizes are minimum required for the upper-bound $B(n)$ to be less than 0.05, that is for the test to rightly reject the null hypothesis at the chosen significance level. Note that the sample size obtained for P_3 differs greatly from the others. Indeed, $\mathbb{S}(P_3) = 0.23$ is very close to s_0, so that $\mathbb{K}(\mathcal{S}_{\mathbb{S}(P_3)}|\mathcal{S}_{s_0})$ is very close to 0 and the exponential factor in $B(n)$ hardly overcomes the polynomial factor. This upper bound is to be refined in a further study, for smaller critical sample sizes to be obtained for moderate deviations from the null hypothesis. However, tests based on large deviations such as developed above are known to be more powerful than tests based on asymptotic normality mainly for small significance levels; see Dembo and Zeitouni (1998).

We also present simulation results for a test of uniformity deduced from the test (5). Indeed, since $\mathcal{S}_{\log(N+1)} = \{U\}$ and $\mathbb{S}(P) < \mathbb{S}(U)$ for all $P \neq U$, this test for $s_0 = \log(N+1)$ is equivalent to the goodness-of-fit test

$$H_0 : \text{“}P = U\text{”} \quad \text{against} \quad H_1 : \text{“}P \neq U\text{”}. \tag{39}$$

Table 5 presents the critical sample sizes obtained for this test from 200 simulated samples drawn according to probability distributions with one mode and all other equal weights (yielding the KL-divergence between spheres), whose supports have

Table 5 Critical sample size for the uniformity test (39) with signification level 0.05

State space size $N+1$	$s_0 = \log(N+1)$	Sample's entropy	Critical sample size n
4	1.39	0.795	15
10	2.3	1.83	39
50	3.91	3.4	165

cardinal either 4, 10 or 50. Again, the p values are upper-bounded by $B(n)$ in (38), where $\mathcal{S}_{s_0} = \mathcal{S}_{\log(N+1)}$, and $\mathbb{K}(\mathcal{S}_{\widehat{S}_n} | \mathcal{S}_{\log(N+1)}) = \log(N+1) - \widehat{S}_n$, according to Theorem 1.

Acknowledgments The present paper has greatly benefited from fruitful discussions with Claudio Macci on the topic of large deviations principles for estimators of entropy. We also gratefully acknowledge some pertinent comments from anonymous reviewers.

References

Amari, S., Nagaoka, H. (2000). *Methods of information geometry*. Oxford: Oxford University Press.

Basharin, G. P. (1959). On a statistical estimation for the entropy of a sequence of independent random variables. *Theory of probability and its applications*, *4*, 333–336.

Beck, C., Schlogl, F. (1993). *Thermodynamics of chaotic systems*. Cambridge: Cambridge University Press.

Bercher, J.-F. (2009). Tsallis distribution as a standard maximum entropy solution with Tail constraint. *Physics Letters A*, *372*, 5657–5659.

Birgé, L., Bretagnolle, J., Dacunha-Castell, D., Duflo, M., Deshayes, J., Maigret, N., Picard, D., Ruget, G. (1979). Grandes déviations et applications statistiques—Séminaire Orsay 1977–1978. *Astérisque 68*. Paris:S.M.F.

Chazottes, J.-R., Gabrielli, D. (2005). Large deviations for empirical entropies of g-measures. *Nonlinearity*, *18*, 2545–2563.

Cover, L., Thomas, J. (1991). *Elements of information theory*. New York: Wiley Series in Telecommunications.

Csiszár, I. (1975). *I*-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, *3*, 141–158.

Csiszár, I., Shieds, P. C. (2004). Information theory and statistics: a tutorial. *Foundations and Trends in Communications and Information Theory*, *1*(Issue 4).

Dembo, A., Zeitouni, O. (1998). *Large deviations techniques and applications* (2nd ed.). New York: Springer.

Ellis, R. (1985). *Entropy, large deviations, and statistical mechanics*. New York: Springer.

Girardin, V., Lequesne, J. (2013). Entropy based goodness-of-fit tests. In: *Proceedings 2d Marrakesh international conference on probability and statistics*.

Girardin, V., Limnios, N. (2014). *Probabilités avec une introduction à la statistique* (3rd ed.). Paris: Vuibert.

Harris, B. (1977). The statistical estimation of entropy in the non-parametric case. *Colloquia Mathematica Societatis Jnos Bolyai* (vol. 16, pp. 323–355). Amsterdam: North-Holland.

Kullback, S., Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *29*, 79–86.

Lequesne, J. (2015). Tests basés sur la théorie de l'information. Application en démographie et en biologie PhD Thesis Université de Caen Basse Normandie, France.

Pielou, E.C. (1967). The use of information theory in the study of the diversity of biological populations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (vol. 4, pp. 163–177).

Regnault, P. (2011). Différents problèmes liés à l'estimation de l'entropie de Shannon d'une loi, d'un processus de Markov. PhD thesis, Université de Caen Basse Normandie, France.

Rukhin, A., Sto, J., Nechvatal, J., Levenson, M., Vangel, M., Banks, D., Heckert, A., Dray, J., Vo, S. (2010). Statistical test suite for random and pseudorandom number generators for cryptographic applications. NIST special publications, pp. 800–822.

Sgarro, A. (1978). An informational divergence geometry for stochastic matrices. *Calcolo*, *15*, 41–49.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423, 623–656.

Tsallis, C. (2009). *Introduction to Nonextensive Statistical Mechanics*. New York: Springer.

Udrişte, C. (1994). *Convex functions and optimization methods on Riemannian manifolds*. Dordrecht: Kluwer Academic Publishers.

Vallée, B. (2001). Dynamical sources in information theory: fundamental intervals and words prefixes. *Algorithmica*, *29*, 262–306.

Van der Vaart, A. (1998). *Asymptotic statistics*. Cambridge: Cambridge Series in Statistical and Probabilistic Mathematics.

Vasicek, O. (1976). A Test for Normality Based on Sample Entropy. *Journal of the Royal Statistical Society*, *38*, 54–59.

Zubkov, A. M. (1973). Limit distribution for a statistical estimator of the entropy. *Theory of Probability and its Applications*, *18*, 611–618.