KYBERNETIKA — VOLUME 57 (2021), NUMBER 6, PAGES 879–907

ON THE JENSEN–SHANNON DIVERGENCE AND THE VARIATION DISTANCE FOR CATEGORICAL PROBABILITY DISTRIBUTIONS

Jukka Corander, Ulpu Remes, Timo Koski

We establish a decomposition of the Jensen–Shannon divergence into a linear combination of a scaled Jeffreys' divergence and a reversed Jensen–Shannon divergence. Upper and lower bounds for the Jensen–Shannon divergence are then found in terms of the squared (total) variation distance. The derivations rely upon the Pinsker inequality and the reverse Pinsker inequality. We use these bounds to prove the asymptotic equivalence of the maximum likelihood estimate and minimum Jensen–Shannon divergence estimate as well as the asymptotic consistency of the minimum Jensen–Shannon divergence estimate. These are key properties for likelihood-free simulator-based inference.

Keywords: blended divergences, Chan–Darwiche metric, likelihood-free inference, implicit maximum likelihood, reverse Pinsker inequality, simulator-based inference

Classification: 62B10, 62H05, 94A17

1. INTRODUCTION

1.1. Background: Simulator-based modeling and inference

The Jensen–Shannon Divergence (JSD) is a numerical quantity expressing the degree of disagreement between two probability distributions. JSD is a special case of the family of ϕ-divergences between two probability distributions, see [13] and [49] for general presentations. There is an abundance of applications of various ϕ-divergences in, e. g., statistical inference, signal processing and machine learning, see [2] for a survey. The study of JSD in the present paper has its background in the development of likelihood-free inference in [20] for the context of simulator-based modeling. The paper [35] is a survey of inference by approximate Bayesian computing in simulator-based modeling. The more recent survey [11] covers also a wider range of methods for inference in simulator-based modeling like, e. g., probabilistic programming.

Many recent mechanistic models, e. g., in genetics, medicine and molecular biology, describe a data generating process in nature by complex, high-accuracy simulator models and the existing computing power gives the ability to generate synthetic data from them. However, simulator models are often not amenable to a tractable analytical treatment.

DOI: 10.14736/kyb-2021-6-0879

For example, the likelihood functions of the model parameters cannot be written down explicitly, since, e. g., a marginalization over some usually large space of latent variables is required. In the terminology of Diggle and Gratton in [15, p.193] such a likelihood function is called implicit. In the same vein, a prescribed model is a model with the likelihood specified by a data distribution in closed form [15, lo.cit.].

Simulator-based models are functions $\mathbb{M}_c$ that map the model parameters and some random variables to synthetic data. The functions $\mathbb{M}_c$ are generally implemented as computer programs, where the parameter values are provided as input and where the random variables are drawn sequentially by making calls to a random number generator. The parameters govern the interesting regularities of a data source in nature, whereas the random variables represent the stochastic variation inherent to the simulated process. The mapping $\mathbb{M}_c$ may be as complex as is asked for, and this generality of simulator-based models allows one to implement a scientifically plausible implicit generative model, which needs not be ruled out on the grounds of mathematical intractability. Of course, to paraphrase [15, p.210], there is no merit in fitting an arbitrary implicit model in favour of a simpler prescribed model, unless the former has a scientific justification.

The notations to be used next serve the purpose of a quick intuitive introduction and will be made more precise in the paper. Let $\mathbf{X}$ be the observed data set and $\mathbf{X}_\theta$ be an output of $\mathbb{M}_c$ at θ, both with values in a finite alphabet $\mathcal{A}$. A key part of simulator-based inference is frequently the selection of summary statistics. In [20] $\mathbf{X}_\theta$ and $\mathbf{X}$ are summarized by their respective empirical categorical distributions $\widehat{P}_{\mathbf{X}_\theta}$ and $\widehat{P}_{\mathbf{X}}$ on $\mathcal{A}$. One option for an approximation of the implicit likelihood is to evaluate the JSD between $\widehat{P}_{\mathbf{X}_\theta}$ and $\widehat{P}_{\mathbf{X}}$ to be minimized as a function of θ. This minimization can be done by BOLFI (Bayesian optimization for likelihood-free inference), see [4] and [20, Section 6.], a method implemented in ELFI, which is a statistical software package suitable for such a mininization. ELFI has been developed by the team of authors in [36]. Concisely stated, the JSD between $\widehat{P}_{\mathbf{X}_\theta}$ and $\widehat{P}_{\mathbf{X}}$ in view of [20] is an estimate of the implicit likelihood function. We prove here that the minimum of JSD-estimate of θ is asymptotically equivalent to the maximum likelihood estimate (MLE) of θ based on $\mathbf{X}$. Here we take advantage of the general properties of ϕ-divergences, and in particular those of the Kullback-Leibler divergence, the Jeffreys' divergence and the (total) variation distance. The work in [33] minimizes an expected Euclidean distance between $\mathbf{X}$ and $\mathbf{X}_\theta$, as data in a general Euclidean space, to find the implicit MLE. It is shown in [33], under fairly restrictive conditions, that a certain non-trivial scaling of the estimate is the MLE.

1.2. Total variation distance and JSD

JSD was up to our understanding first introduced in 1969 by Sibson in [43], under the name *information radius* (of first order), the biological impetus can be found in [26] and is clarified briefly in the observation 7.3 below. Independently, Topsøe dealt with JSD in [46] and called it *capacitory information*. Sibson proved, amongst other, that certain JSD-neighborhoods form a basis for the variation distance topology, we shall state this fact more precisely in Section 2.3. We bound JSD upwards and downwards by the squared variation distance.

The variation distance is effectively the only ϕ-divergence that is a metric, see [27],

see also [50, Corollary p.898] for a transparent proof, in the sense that, if a ϕ-divergence is a metric, then it is proportional to the variation distance. The square root of the symmetric JSD is a metric, see [16], a special case of a general result on metricity of positive powers of ϕ-divergences [50, Thm 1, p.891]. Total variation distance appears like a center node in a graph when a number of inequalities to other measures of difference between probability measures used in statistics are depicted pictorially, see [18, Figure1, p.421]. In the comprehensive and deep study [41] the inequalities between total variation and ϕ-divergences are summarized in [41, pp.5973–5974]. The most recent techniques for deriving inequalities for ϕ-divergencies appear in [40]. A survey of the inequalities for ϕ-divergences of categorical distributions is found in [1]. The survey [18, p.429] provides a summary of inequalities for total variation distance w.r.t. the well known metrics of probability theory.

JSD is operationally relevant in machine learning, see, e. g., [38], [42] and [44]. Due to the applications we have in mind, we are restricting ourselves to probabilities on finite discrete sets (alphabets), to be called categorical distributions. Chan and Darwiche have introduced a metric between two categorical distributions in [6] and [7]. There is no quick way of comparing a ϕ-divergence to the Chan–Darwiche metric. When we consider the special case of the multivariate Bernoulli distributions on the binary hypercube, we can express the bounds on JSD in terms of the Chan–Darwiche metric.

1.3. Organization of the paper

The notations and definitions, including Chan and Darwiche metric, ϕ-divergences, JSD, and reversed JSD, are presented in Section 2. Here JSD and reversed JSD are introduced by the notion of blending in [29]. The Pinsker inequality and reverse Pinsker inequality, see [41] are used for the first upper and lower bounds for JSD in terms of the variation distance in Section 3. The bounds based on a decomposition of the JSD as a linear combination of a scaled Jeffreys' divergence and a reversed JSD are in Section 4. The decomposition is derived by means of the compensation identity [46] and is bounded by the reverse Pinsker inequality. In Section 5, the proof of consistency of the minimization of JSD, when there is a true distribution inside the model, is based on taking the square root of the inequalities and then using the triange inequality for the variation distance. Thereafter we can apply the complete convergence results in [3] and [14]. The same argument shows also that the minimum JSD-estimate is asymptotically equivalent to the MLE. We can thus actually compute the MLE by the minimum JSD, if the likelihood is implicit. There are simulation studies in Section 6 for comparison of MLE and the minimum JSD-estimate. Instances of the derived bounds with explicit expressions for the multivariate Bernoulli distributions are given in Section 7, which partly are based on the special representation of the multivariate Bernoulli distribution found in [22].

2. CATEGORICAL DISTRIBUTIONS AND THE JENSEN–SHANNON DIVERGENCE

2.1. Categorical distributions and a metric

Let $\mathcal{A} = \{a_1, \ldots, a_k\}$ be a finite alphabet, $k \geq 2$. $\mathcal{A}$ and k are known. $\mathbb{P}$ is the set of all probability distributions on $\mathcal{A}$. Every $P \in \mathbb{P}$ is called a categorical distribution, and

can be represented as,

$$P(x) = \prod_{i=1}^{k} p_i^{[x=a_i]}, \quad x \in \mathcal{A}, \tag{1}$$

where $[x = a_i] = 1$, if $x = a_i$ and $[x = a_i] = 0$ otherwise (the Iverson bracket), $(0^0 = 1, 0^1 = 0)$, and $p_i \geq 0$, $\sum_{i=1}^{k} p_i = 1$. If we have a random variable X assuming values in $\mathcal{A}$ such that the probability of $X = a_i$ is p_i, we write $X \sim P$. The support of $P \in \mathbb{P}$ is $\text{supp}(P) = \{a_i \in \mathcal{A} | \quad p_i = P(a_i) > 0\}$. Models $\mathbb{M}$ are subsets of $\mathbb{P}$, where $P \in \mathbb{M}$ is depending on a finite number of real valued parameters with the generic symbol $\theta \in \Theta$. For k functions $p_i(\theta)$ of θ, such that $p_i(\theta) \geq 0$, $\sum_{i=1}^{k} p_i(\theta) = 1$ for all $\theta \in \Theta$ we have that $P_\theta(x) := \prod_{i=1}^{k} p_i(\theta)^{[x=a_i]}$, $x \in \mathcal{A}$, is a parametric distribution $P_\theta \in \mathbb{P}$ and the corresponding model is $\mathbb{M} = \{P_\theta \mid \theta \in \Theta\}$. In simulation-based inference the simulator itself induces $\mathbb{M}$. The notion of implicit likelihood due to Diggle and Gratton [15] means in the case of $P_\theta \in \mathbb{P}$ that all of $p_i(\theta)$ are implicit or intractable, i. e., cannot be expressed in closed form. We cite two cases.

Example 2.1. In [9] Wright−Fisher theory about the multilocus negative frequency-dependent selection is used as the model of the evolution of genotype frequencies. The categories are vaccine-type statuses combined with sequence clusters of binary strings representing the presence and absence of a number of antibiotic-resistance phenotypes present in a certain bacterial population. The category probabilities are functions of the carrying capacity of an environment, migration rate, and the vaccine selection pressure and its magnitude. The likelihood function for these parameters based on observed frequencies $\widehat{P}$ of each category is intractable, but one can simulate $\mathbb{M}$, in this case the generative Wright−Fisher model, to get the synthetic relative frequencies under any parameter setting and then find the (symmetric) $\widehat{\theta}_{n,\text{JS}}$. Further data-analysis with minimum symmetric JSD in this framework is presented in [23].

Example 2.2. Suppose we have a theory domain entertaining k functions $g_i(\theta)$, $i = 1, \ldots, k$ of interest, tractable or not. Then a categorical distribution is determined by the soft-max assignments $p_i(\theta) = e^{g_i(\theta)}/C(\theta)$, $i = 1, \ldots, k$. In these situations, that occur, e. g., in neural networks and discriminative classification, frequently no closed form exists for the normalization $C(\theta)$. But synthetic samples of P_θ are still readily generated without access to a tractable $C(\theta)$ by, e. g., the well known Gumbel-Max trick, see [51, Lemma 6, p.123].

Following Chan and Darwiche, see [6] and [7], we introduce for $P \in \mathbb{P}$ and $Q \in \mathbb{P}$ the quantity

$$D_{\text{CD}}(P, Q) := \max_{x \in \mathcal{A}} \ln \frac{P(x)}{Q(x)} - \min_{x \in \mathcal{A}} \ln \frac{P(x)}{Q(x)}. \tag{2}$$

If $\text{supp}(P) \cap \text{supp}(Q) = \emptyset$, then $D_{\text{CD}}(P, Q) = +\infty$. We take $0/0 = 1$. It is shown in [6, Thm 2.1] that $D_{\text{CD}}(P, Q)$ is a metric on $\mathbb{P}^2$. The following simple example will play a part in Section 7 via Appendix A.

Example 2.3. Here $\mathcal{A} = \{0, 1\}$ and $\theta \in (0, 1)$. Let

$$p_\theta(x) := (1 - \theta)^{[x=0]} \cdot \theta^{[x=1]}, \quad x \in \{0, 1\}. \tag{3}$$

Then $\mathbb{M}_{\mathrm{Be}} = \{p_\theta | 0 < \theta < 1\}$ is the family of Bernoulli distributions. We take $q(x) = p_{1/2}(x) = 1/2$ for $x \in \{0, 1\}$. Then (2) yields

$$D_{\mathrm{CD}}(p_\theta, q) = \begin{cases} \ln \frac{\theta}{1-\theta} & \text{if } 1/2 \leq \theta < 1; \\ \ln \frac{1-\theta}{\theta} & \text{if } 0 < \theta < 1/2. \end{cases} \tag{4}$$

Hence it is seen that $D_{\mathrm{CD}}(p_\theta, q) > 0$ if $\theta \neq 1/2$ and $D_{\mathrm{CD}}(p_{1/2}, q) = 0$. The expression (4) is the link connecting D_{CD} to JSD between two multivariate Bernoulli distributions.

The ϕ-divergences to be discussed next are not metrics in general. $D_{\mathrm{CD}}(P, Q)$ is constructed to measure the distance between P and Q, when P is an up-date of Q on a set of events, e. g., Jeffrey's update and Pearl's virtual up-date. It is shown in [6] and [7] that, e. g., Kullback-Leibler divergence can fail to make a meaningful comparison of up-dates.

2.2. ϕ-divergences

We define first a divergence function, see, e. g., [29, Def.1, p.44] or [40, Def. 3, p.32].

Definition 2.4. A divergence function $\phi(x)$ is a continuous convex function $(0, +\infty) \overset{\phi}{\mapsto} \mathbb{R} \cup +\infty$, $\phi(1) = 0$, and strictly convex at $x = 1$. If $p = 0$, we take $q\phi(p/q) = q\phi(0)$, where $\phi(0) = \lim_{x \downarrow 0} \phi(x)$. If $q = 0$, we take $q\phi(p/q)$ as $p \lim_{x \to +\infty} \phi(x)/x$.

For two categorical distributions $Q \leftrightarrow Q(x) = \prod_{i=1}^{k} q_i^{[x=a_i]} \in \mathbb{P}$ and $P \leftrightarrow P(x) = \prod_{i=1}^{k} p_i^{[x=a_i]} \in \mathbb{P}$ and for a divergence function ϕ, we define the ϕ-divergence [49] between P and Q as

$$D_\phi(P, Q) := \sum_{x \in \mathcal{A}} Q(x) \phi\left(\frac{P(x)}{Q(x)}\right) = \sum_{i=1}^{k} q_i \phi\left(\frac{p_i}{q_i}\right), \tag{5}$$

where we applied (1). We have the adjoined function, also known as the conjugated function,

$$\phi^*(x) = x\phi\left(\frac{1}{x}\right), \quad x \in (0, +\infty). \tag{6}$$

The adjoined function has the property $D_{\phi^*}(P, Q) = D_\phi(Q, P)$. It holds that

$$\phi(1) \leq D_\phi(P, Q) \leq \phi(0) + \phi^*(0). \tag{7}$$

This gives $D_\phi(P, Q) = 0 \Leftrightarrow P = Q$. The range property in (7) is proven by Liese and Vajda in [32, Thm 5].

Several distinct ϕ-divergences appear in the sequel. For the first instance, we select in (5) $\phi(x) = x \ln x$, the resulting divergence is known as the Kullback-Leibler divergence (KL) and is denoted by $D_{\mathrm{KL}}(P, Q)$. In general $D_{\mathrm{KL}}(P, Q) \neq D_{\mathrm{KL}}(Q, P)$, if $P \neq Q$. If there is a pair $p_i > 0$ and $q_i = 0$, then $D_{\mathrm{KL}}(P, Q) = +\infty$ by the properties of divergence functions stated in definition 2.4. The expression

$$D_{\mathrm{Je}}(P, Q) := D_{\mathrm{KL}}(P, Q) + D_{\mathrm{KL}}(Q, P) \tag{8}$$

defines Jeffreys' symmetric ϕ-divergence. It corresponds to the divergence function $\phi_{\mathrm{Je}}(x) := (x-1)\ln x$, or by (6), $\phi_{\mathrm{Je}}(x) = \phi(x) + \phi^*(x)$, for $\phi(x) = x\ln x$.

When $\phi(x) = |x-1|$, the ϕ-divergence in (5) is denoted by $V(P,Q)$, and called the (total) variation distance. In [41, Thm 4, p.5980] Sason and Verdu prove that

$$D_\phi(P,Q) \leq \frac{1}{2}\left[\phi(0) + \phi^*(0)\right] V(P,Q). \tag{9}$$

We note the elementary equality $\min\{x,y\} = \frac{1}{2}\left(x + y - |x-y|\right)$, valid for any real numbers x and y. It follows that

$$\frac{1}{2}V(P,Q) = 1 - \sum_{i=1}^{k} \min\{p_i, q_i\}, \tag{10}$$

which plays a role in the sequel. This equality is a special case of the equality in [31, p.803].

2.3. Jensen–Shannon and reversed Jensen–Shannon divergence

Jensen–Shannon divergence and the Reversed Jensen–Shannon divergence can be introduced in a unified fashion by the notion of blended divergence in [29, Def. 2., p.48]. We recapitulate this definition. Let ϕ_1 and ϕ_2 be two divergence functions and set for $x > 0$

$$\phi_B(x) := (1-\pi+\pi x)\phi_1\left(\frac{x}{1-\pi+\pi x}\right) + \phi_2\left(1-\pi+\pi x\right). \tag{11}$$

Then ϕ_B can be shown to be a divergence function, in the sense of definition 2.4, and is called the blended divergence function (blending of ϕ_1 and ϕ_2). For $0 < \pi < 1$ the mixture $M \in \mathbb{P}$ is defined by $M = \pi P + (1-\pi)Q$ with the category probabilities $m_i := \pi p_i + (1-\pi)q_i$. When we blend the divergence functions $\pi\phi_1$ and $(1-\pi)\phi_2$, respectively, we get the blended divergence

$$D_{\phi_B}(P,Q) = \pi D_{\phi_1}(P;M) + (1-\pi)D_{\phi_2}(M,Q), \tag{12}$$

as follows by [29, Corollary 1, Corollary 2., p.47] noting the rule $aD_\phi(P,Q) = D_{a\phi}(P,Q)$. Let us take $\phi_1 = \phi$ and $\phi_2 = \phi^*$. Then the property $D_{\phi^*}(P,Q) = D_\phi(Q,P)$ entails that

$$D_{\phi_B}(P,Q) = \pi D_\phi(P;M) + (1-\pi)D_\phi(Q,M),$$

c.f. [29, Specification 1, p.48]. When $\phi(x) = x\ln x$, we set $D_{\mathrm{JS}}(P,Q) := D_{\phi_B}(P,Q)$, and find

$$D_{\mathrm{JS}}(P,Q) = \pi D_{\mathrm{KL}}(P,M) + (1-\pi)D_{\mathrm{KL}}(Q,M), \quad 0 < \pi < 1. \tag{13}$$

This is the Jensen–Shannon divergence (JSD) with the divergence function

$$\phi(x) = \pi\cdot x\ln x - (\pi x + (1-\pi))\ln(\pi x + (1-\pi)), \tag{14}$$

by (11), see also [52, Example 1, p.1038] for another derivation. Blending has produced a smoothing of KL, as $D_{\mathrm{KL}}(P,M) < +\infty$ and $D_{\mathrm{KL}}(Q,M) < +\infty$ for all Q and P,

since M dominates both P and Q. Actually $D_{\mathrm{JS}}(P,Q)$ is uniformly bounded, the upper bound will be pointed out below. Vajda and Österreicher did show in [52] that $D_{\mathrm{JS}}(P,Q)$ is a measure of statistical information in terms of Bayesian risk of discrimination with a logarithmic risk function.

By (5) and (14) it follows, as pointed out by a reviewer, after a small piece of algebra that

$$D_{\mathrm{JS}}(P,Q) = \pi D_{\mathrm{KL}}(P,Q) - D_{\mathrm{KL}}(M,Q). \tag{15}$$

This expression will be invoked in the last paragraph of Section 5.

The decomposition of JSD in the sequel produces the following quantity

$$D_{\mathrm{RJS}}(P,Q) := \pi D_{\mathrm{KL}}(M,P) + (1-\pi) D_{\mathrm{KL}}(M,Q) \tag{16}$$

called the reversed JSD, c.f. [50, p.891]. The reversed JSD is another blended divergence. Let $\phi(x) = x \ln x$ and specify (12) for $x > 0$ with $\phi_1(x) = -\pi\phi^*(x) = -\pi \ln x$ and $\phi_2(x) = (1-\pi)\phi(x)$. This entails (16) by (12). The corresponding blended divergence function is derived by (11) as

$$\phi_{\mathrm{RJS}}(x) = (\pi x + (1-\pi)) \left[\ln(\pi x + (1-\pi)) - \pi \ln x\right]. \tag{17}$$

When dealing with $D_{\mathrm{RJS}}(P,Q)$ we are going to assume $\mathrm{supp}(P) = \mathrm{supp}(Q) (= \mathcal{A}$ without loss of much generality). By (7) we get

$$0 \le D_{\mathrm{JS}}(P,Q) \le B(\pi) \le \ln(2), \tag{18}$$

where

$$B(\pi) := -\pi \ln(\pi) - (1-\pi)\ln(1-\pi) \tag{19}$$

is the binary entropy function. From (7) and (9) we obtain

$$D_{\mathrm{JS}}(P,Q) \le \frac{B(\pi)}{2} V(P,Q) \le \frac{\ln 2}{2} V(P,Q). \tag{20}$$

For the case $\pi = 1/2$ we introduce the notation $D_{\mathrm{JS},1/2}(P,Q)$. Hence $D_{\mathrm{JS},1/2}$ is a smoothed and symmetrized blend of KL. $D_{\mathrm{JS},1/2}$ is often used in applying the JSD in data analysis by simulator modeling. The paper [47] provides alternative explicit expressions (e. g., in terms of infinite series) and bounds for $D_{\mathrm{JS},1/2}$. By [30, pp.108-111], $D_{\mathrm{JS},1/2}$ is an instance of extended Matusita divergences and by [30, pp.106-108], $D_{\mathrm{JS},1/2}$ is an instance of extended absolute power divergences. In our effort we have not so far taken advantage of selecting ϕ-divergences with most desirable properties using smooth transitions between the cited divergences and $D_{\mathrm{JS},1/2}$.

2.4. JSD and total variation distance topology

For any $P \in \mathbb{P}$ we define an open JSD - neighborhood around P as

$$N(P,\epsilon) := \{Q \in \mathbb{P} | D_{\mathrm{JS}}(P,Q) < \epsilon\}. \tag{21}$$

The following is from [43, Thm 2.7., p.153].

Proposition 2.5. For fixed π, varying $P \in \mathbb{P}$ and ϵ, $N(P,\epsilon)$ form a basis for the variation distance topology of $\mathbb{P}$.

We shall now make a statistical interpretation. The empirical distribution

$$\widehat{P}_n(x) := \prod_{i=1}^{k} \widehat{p}_i^{[x=a_i]} \in \mathbb{P}, \quad x \in \mathcal{A}, \tag{22}$$

is determined by the relative frequencies $\widehat{p}_i = \frac{n_i}{n}$, $i = 1, \ldots, k$, found in $\mathbf{X} = (X_1, \ldots, X_n)$, where X_l are outcomes of independent, identically distributed (i.i.d.) random variables with values in $\mathcal{A}$, where n_i = the number of samples X_l in $\mathbf{X}$ such that $X_l = a_i$. We define the set of 'true' distributions, this notion is from [8, p. 2254], by

$$\mathbf{B}(\widehat{P}_n) := \{P \in \mathbb{P} | V(P, \widehat{P}_n) < \epsilon\}, \quad \epsilon \in (0, 2). \tag{23}$$

Let $\mathcal{B}$ denote the variation distance topology, thus $\mathbf{B}(\widehat{P}_n) \in \mathcal{B}$. Suppose $P_\theta \in \mathbf{B}(\widehat{P}_n)$. A first property of a topological basis is that there exists (for a fixed π) some P_{θ_o} and $\epsilon_o > 0$ such that $P_\theta \in N(P_{\theta_o}, \epsilon_o) \subseteq \mathbf{B}(\widehat{P}_n)$. In words, for every nominal distribution $P_\theta \in \mathbf{B}(\widehat{P}_n)$, there exists a JSD neighborhood of true distributions P that covers P_θ.

The algorithms for simulator-modeling include often a rejection step, c.f., [35, p. e70]. In the total variation setting above, a summarization P_θ of synthetic output of $\mathbb{M}_c$ at θ is accepted, if it lies in $\mathbf{B}(\widehat{P}_n)$ for a chosen $\epsilon > 0$, otherwise rejected. Hence JSD is inherent at the rejection step with relative frequencies and V as rejection distance, which supports intuitively the idea that JSD is an estimate of the implicit likelihood.

3. BOUNDS FOR THE JENSEN–SHANNON DIVERGENCE

3.1. Applications of Pinsker and reverse Pinsker inequalities

We recall the Pinsker inequality, see [10, Lemma 11.6.1]

$$D_{\mathrm{KL}}(P, Q) \geq \frac{1}{2} V(P, Q)^2. \tag{24}$$

An account of the steps of discovery and further refinements on (24) are available in [17]. The result in the next Lemma is known as the reverse Pinsker inequality, found in [41, Eq. (335), p.5991].

Lemma 3.1. $P \in \mathbb{P}$ and $Q \in \mathbb{P}$. Assume $\mathrm{supp(Q)} = \mathcal{A}$. Set $Q_{\min} := \min_{x \in \mathcal{A}} Q(x)$. Then

$$D_{\mathrm{KL}}(P, Q) \leq \frac{1}{2Q_{\min}} V(P, Q)^2. \tag{25}$$

An elementary proof of a less sharp (the factor 1/2 is missing in the right hand side) version of (25) is developed by [12, Lemma 6.3, p.1012], see also [18, Thm 5, p.429]. For the statements to follow we compute the second derivative at $x = 1$ from (14) as

$$\phi''(1) = \pi(1 - \pi). \tag{26}$$

We write $f(x) \asymp g(x)$, when $cg(x) \leq f(x) \leq Cg(x)$ for all x in some domain, where $c < C$. The inequalities in (27) read then as $D_{\mathrm{JS}}(P, Q) \asymp \phi''(1) V(P, Q)^2$.

Proposition 3.2. $P \in \mathbb{P}$ and $Q \in \mathbb{P}$. Assume supp(P) = supp(Q) = $\mathcal{A}$. Set $P_{\min} = \min_{x\in\mathcal{A}} P(x)$ and $Q_{\min} = \min_{x\in\mathcal{A}} Q(x)$. Then

$$\frac{\phi^{''}(1)}{2} V(P,Q)^2 \le D_{\mathrm{JS}}(P,Q) \le \frac{\phi^{''}(1)}{2\min\{\pi P_{\min}, (1-\pi)Q_{\min}\}} V(P,Q)^2, \tag{27}$$

where $\phi^{''}(1)$ is given by (14).

P r o o f. By (13) and (24)

$$D_{\mathrm{JS}}(P,Q) \ge \frac{\pi}{2} V(P,M)^2 + \frac{(1-\pi)}{2} V(Q,M)^2. \tag{28}$$

By definition of the total variation distance

$$V(P,M) = \sum_{i=1}^{k} |p_i - (\pi p_i - (1-\pi)q_i)| = \sum_{i=1}^{k} |(1-\pi)p_i - (1-\pi)q_i| = (1-\pi)V(P,Q), \tag{29}$$

and

$$V(Q,M) = \sum_{i=1}^{k} |q_i - (\pi p_i + (1-\pi)q_i)| = \sum_{i=1}^{k} |\pi q_i - \pi p_i| = \pi V(Q,P) = \pi V(P,Q). \tag{30}$$

When we insert (29) and (30) in the right hand side of (28) we get

$$D_{\mathrm{JS}}(P,Q) \ge \frac{1}{2}\left[\pi(1-\pi)^2 + (1-\pi)\pi^2\right] V(P,Q)^2. \tag{31}$$

Here $\pi(1-\pi)^2 + (1-\pi)\pi^2 = \phi^{''}(1)$ by (26). Hence we have the lower bound in (27). To prove the upper bound of (31) we set $M_{\min} := \min_{x\in\mathcal{A}} [\pi P(x) + (1-\pi)Q(x)]$. By Lemma 3.1, (25), and since

$$M_{\min} \ge \min\{\pi P_{\min}, (1-\pi)Q_{\min}\} > 0,$$

we obtain

$$D_{\mathrm{KL}}(P,M) \le \frac{1}{2M_{\min}} V(P,M)^2 \le \frac{1}{2\min\{\pi P_{\min}, (1-\pi)Q_{\min}\}} V(P,M)^2 \tag{32}$$

and similarly for $D_{\mathrm{KL}}(Q,M)$ with Q replacing P. When we insert the rightmost expressions in (29) and (30) in the right hand sides of the bounds on $D_{\mathrm{KL}}(P,M)$ and $D_{\mathrm{KL}}(Q,M)$ above, respectively, we obtain by the same computation as above the right hand side inequality in (27). □

The upper bound (27) is sharper than (20), if $V(P,Q) \le \min\{P_{\min}, Q_{\min}\} \cdot \ln 2/\phi^{''}(1)$. As $\phi^{''}(1) = \pi(1-\pi)$ is maximized by $\pi = 1/2$, then $V(P,Q) \le \min\{P_{\min}, Q_{\min}\} \cdot 4\ln 2$ is the lowest range. When (10) is applied in (27) we obtain

$$\sqrt{2}\left(\frac{1}{2} - \mathbf{P}_e(P,Q)\right) \le \sqrt{D_{\mathrm{JS},1/2}(P,Q)} \le 2c\left(\frac{1}{2} - \mathbf{P}_e(P,Q)\right), \tag{33}$$

where $c = \sqrt{1/\min\{P_{\min}, Q_{\min}\}}$ and

$$\mathbf{P}_e(P,Q) := \sum_{i=1}^{k} \min\left\{\frac{1}{2}p_i, \frac{1}{2}q_i\right\}. \tag{34}$$

The observation in (33) has a significant meaning. Suppose that the prior probabilities of P and Q are $pr(P) = pr(Q) = 1/2$. The Bayesian rule of discrimination of $X = x \in \mathcal{A}$ tells to decide x as being sampled from P, if $\frac{1}{2}P(x) > \frac{1}{2}Q(x)$ and conversely, ties resolved arbitrarily. This rule minimizes the probability of error. The optimal probability of error equals $\mathbf{P}_e(P,Q)$ in (34), one proof is found in [28, ch.9.4].

Suppose now that $Q_o \in \mathbf{B}(P) = \{Q \in \mathbb{P} \mid V(P,Q) < \epsilon\}$ for some $\epsilon > 0$. This implies by (10) that $\frac{1}{2}\left(1 - \frac{\epsilon}{2}\right) < \mathbf{P}_e(P,Q_o) \leq \frac{1}{2}$. In view of discussion in Section 2.4 there exist P_o and $\epsilon_o > 0$ such that $Q_o \in \{Q \in \mathbb{P} | D_{\mathrm{JS},1/2}(P_o,Q) < \epsilon_o\} \subseteq \mathbf{B}(P)$. Hence, when there is P such that $D_{\mathrm{JS},1/2}(P,Q_o)$ is small, then $\mathbf{P}_e(P,Q_o)$ must be close to $1/2$. Thus, the Bayes' discrimination rule for such P and Q_o performs equally well as discrimination by tossing an unbiased coin without looking at x. This is what a good simulator-model with outputs distributed by Q_o should achieve against a given P of observed data.

3.2. Lower bounds for JSD by refinements of Pinsker's inequality

There is also the lower bound due to Vajda [48], see also [17, Corollary 4, p.1494],

$$D_{\mathrm{KL}}(P,Q) \geq \ln\left(\frac{2+V(P,Q)}{2-V(P,Q)}\right) - \frac{2V(P,Q)}{2+V(P,Q)}, \quad V(P,Q) \in [0,2). \tag{35}$$

Proposition 3.3. $P \in \mathbb{P}$ and $Q \in \mathbb{P}$. Then

$$D_{\mathrm{JS},1/2}(P,Q) \geq \ln\left(\frac{1+\left(\frac{1}{2}-\mathbf{P}_e(P,Q)\right)}{1-\left(\frac{1}{2}-\mathbf{P}_e(P,Q)\right)}\right) - \frac{2\left(\frac{1}{2}-\mathbf{P}_e(P,Q)\right)}{1+\left(\frac{1}{2}-\mathbf{P}_e(P,Q)\right)}. \tag{36}$$

Proof. When (29) and (30) and definition of $\mathbf{P}_e(P,Q)$ are applied in (35) with $V \in [0,2)$, we obtain (36). □

In [40, Remark 12., Eqn. (156)] one finds, as pointed out by a reviewer, another refinement of the Pinsker bound.

$$D_{\mathrm{KL}}(P,Q) \geq -\ln\left(1 - \frac{1}{4}V(P,Q)^2\right), \quad \text{for } V(P,Q) \in [0,2), \tag{37}$$

arising from a bound on E_γ-divergences for $\gamma = 1$. By (37) we obtain the following.

Proposition 3.4. $P \in \mathbb{P}$ and $Q \in \mathbb{P}$. Then for $V(P,Q) \in [0,2)$

$$D_{\mathrm{JS},1/2}(P,Q) \geq -\ln\left(1 - \frac{1}{16}V(P,Q)^2\right). \tag{38}$$

P r o o f . We apply first (37) in the two terms in $D_{\mathrm{JS}}(P,Q)$ and then (29) and (30) in the two KL- divergences to obtain a lower bound, where we set $\pi = \frac{1}{2}$. □

We set for economy of writing $V = V(P,Q)$ for a moment. Vajda's lower bound improves Pinsker's bound in the sense that if $V \uparrow 2$, then the lower bound in (35) turns to $+\infty$, as does $D_{\mathrm{KL}}(P,Q)$, since $V = 2$ is equivalent to $\mathrm{supp}(P) \cap \mathrm{supp}(Q) = \emptyset$. For $V \uparrow 2$, the optimal error $\mathbf{P}_e(P,Q) \to 0$ and the lower bound (36) converges to $\ln(3) - 2/3 = 0.432 < \ln(2) = 0.693$. Thus, too much smoothing was injected above. Hence, the inequality in (36) is mainly of interest for small values of V.

The right hand side of (38) is well-defined even for $V = 2$ and equals there $\ln\left(\frac{4}{3}\right) = 0.288$. For $\pi = \frac{1}{2}$ the lower bound (27) is always sharper than the bound (38), i. e., $\frac{1}{8}V^2 > -\ln\left(1 - \frac{1}{16}V^2\right)$ for all $V \in (0,2]$. In [19, Example II.4, p.2389] one finds

$$D_{\mathrm{JS},1/2}(P,Q) \geq \frac{1}{2}\left[\left(1+\frac{V}{2}\right)\ln\left(1+\frac{V}{2}\right) + \left(1-\frac{V}{2}\right)\ln\left(1-\frac{V}{2}\right)\right].$$

This lower bound by Guntuboyina is sharper than the lower bound $\frac{1}{8}V^2$ in (27). Let us denote the expression in the right hand side of Guntuboyina lower bound by $f(V)$. One gets $f(2) = \ln 2$, which is perfect. However, $f(1) = 0.1308$, $f(0.1) = 0.013$, where the corresponding values for $\frac{1}{8}V^2$ are 0.1250 and 0.013. By a plot one sees that the bounds $f(V)$ and $\frac{1}{8}V^2$ have indistinguishable graphs for $0 < V < 1$ for all practical purposes.

4. DECOMPOSITION OF JSD BY MEANS OF A SUM OF JEFFREYS' DIVERGENCE AND REVERSED JSD

In all statements in this Section we are assuming that $\mathrm{supp}(P) = \mathrm{supp}(Q) = \mathcal{A}$ for $P \in \mathbb{P}$ and $Q \in \mathbb{P}$.

4.1. A decomposition by means of an escort distribution

Consider $P \in \mathbb{P}$ and $Q \in \mathbb{P}$, $\mathrm{supp}(P) = \mathrm{supp}(Q)$, and set for $i = 1, \ldots, k$

$$g_i := \frac{p_i^\pi q_i^{1-\pi}}{c(\pi)}, \quad \pi \in [0,1]. \tag{39}$$

Here $c(\pi) = \sum_{i=1}^k p_i^\pi q_i^{1-\pi}$. Then $G(x) = \prod_{i=1}^k g_i^{[x=a_i]}$ and $G \in \mathbb{P}$ is often called the escort distribution. For π_o such that $c(\pi_o) \leq c(\pi)$ for all $\pi \in (0,1)$, G is the barycenter of $\{P,Q\}$, and represents $\{P,Q\}$ by a minimax property, see [25, p.48].

Proposition 4.1. With Jeffreys divergence (8) and $D_{\mathrm{RJS}}(P,Q)$ in (16) it holds for any $P \in \mathbb{P}$ and $Q \in \mathbb{P}$ that

$$D_{\mathrm{JS}}(P,Q) = \phi''(1) D_{\mathrm{Je}}(P,Q) - D_{\mathrm{RJS}}(P,Q). \tag{40}$$

P r o o f . The compensation identity of [46, Lemma 7], see also [47, p.1603], shows that, since $\mathrm{supp}(G) = \mathcal{A}$,

$$D_{\mathrm{JS}}(P,Q) = \pi D_{\mathrm{KL}}(P,G) + (1-\pi) D_{\mathrm{KL}}(Q,G) - D_{\mathrm{KL}}(M,G). \tag{41}$$

Here $D_{\mathrm{KL}}(P,G) = \sum_{i=1}^k p_i \ln\left(\frac{p_i}{p_i^\pi q_i^{1-\pi}}\right) + \ln c(\pi)$. Furthermore, $\sum_{i=1}^k p_i \ln\left(\frac{p_i}{p_i^\pi q_i^{1-\pi}}\right) = \sum_{i=1}^k p_i \ln\left(\frac{p_i^\pi p_i^{1-\pi}}{p_i^\pi q_i^{1-\pi}}\right) = (1-\pi)D_{\mathrm{KL}}(P,Q)$, and thus $\pi D_{\mathrm{KL}}(P,G) = (1-\pi)\pi D_{\mathrm{KL}}(P,Q) + \pi \ln c(\pi)$. In the same way we obtain $(1-\pi)D_{\mathrm{KL}}(Q,G) = \pi(1-\pi)D_{\mathrm{KL}}(Q,P) + (1-\pi)\ln c(\pi)$.

Next, $D_{\mathrm{KL}}(M,G) = \sum_{i=1}^n m_i \ln \frac{m_i}{p_i^\pi q_i^{1-\pi}} + \ln c(\pi)$. We continue with the same split of exponent as above, $\sum_{i=1}^n m_i \ln \frac{m_i}{p_i^\pi q_i^{1-\pi}} = \sum_{i=1}^n m_i \ln \frac{m_i^\pi m_i^{1-\pi}}{p_i^\pi q_i^{1-\pi}} = \pi \sum_{i=1}^n m_i \ln \frac{m_i}{p_i} + (1-\pi)\sum_{i=1}^n m_i \ln \frac{m_i}{q_i} = \pi D_{\mathrm{KL}}(M,P) + (1-\pi)D_{\mathrm{KL}}(M,Q)$. When the preceding expressions are inserted in (41), remembering (26), the assertion in (40) is proven. □

The result in the Proposition decomposes $D_{\mathrm{JS}}(P,Q)$ into a sum of a symmetric term and an asymmetric term, the reversed JSD. Since JSD is nonnegative, $\pi \cdot (1-\pi)D_{\mathrm{Je}}(P,Q) \geq D_{\mathrm{RJS}}(P,Q$, hence the reversed JSD is antithetic to $D_{\mathrm{JS}}(P,Q)$ in the sense that if $D_{\mathrm{JS}}(P,Q)$ is small, or close to its minimum zero, then $D_{\mathrm{RJS}}(P,Q)$ has to be large, and vice versa. We can write $T(P,Q) := D_{\mathrm{KL}}(M,G) - \ln c(\pi)$, and in light of the terminology in [45, p.12] we can call $T(P,Q)$ the weighted arithmetic-geometric divergence. Next, $H(P) := -\sum_{x\in\mathcal{A}} P(x)\ln P(x)$, is the Shannon entropy of $P \in \mathbb{P}$ in natural logarithm. Explicit expansions of the KL-distances appearing in the right hand side of (40), re-organizations and a number of terms taking out each other give the following formula

$$D_{\mathrm{JS}}(P,Q) = H(\pi P + (1-\pi)Q) - \pi H(P) - (1-\pi)H(Q). \tag{42}$$

The right hand side of (42) is nothing else than the Shannon-Jensen divergence $D_{\mathrm{JS}}(P,Q)$ as written down and named in [34]. Of course, the right hand side is well-defined also when $\mathrm{supp}(P) = \mathrm{supp}(Q) = \mathcal{A}$ does not hold.

4.2. Bounds for JSD using the decomposition

Proposition 4.2. Let $P \in \mathbb{P}$ and $Q \in \mathbb{P}$. $P_{\min} = \min_{x\in\mathcal{A}} P(x)$ and $Q_{\min} = \min_{x\in\mathcal{A}} Q(x)$. Then it holds that

$$D_{\mathrm{JS}}(P,Q) \leq \frac{\phi''(1)}{2}\left(\frac{P_{\min}+Q_{\min}}{Q_{\min}P_{\min}} - 1\right)V(P,Q)^2. \tag{43}$$

Proof. Due to (8) and the reverse Pinsker inequality (25), it follows that

$$D_{\mathrm{Je}}(P,Q) \leq \frac{1}{2Q_{\min}}V(P,Q)^2 + \frac{1}{2P_{\min}}V(Q,P)^2 = \frac{P_{\min}+Q_{\min}}{2Q_{\min}P_{\min}}V(P,Q)^2. \tag{44}$$

Next we bound the reversed JSD by the Pinsker inequality (24). This entails $D_{\mathrm{RJS}}(P,Q;\pi) \geq \frac{1}{2}\left(\pi V(M,P)^2 + (1-\pi)\right)V(M,Q)^2)$. But $V(M,P)$ and $V(M,Q)$ are calculated in (29) and (30). Hence we have $D_{\mathrm{RJS}}(P,Q;\pi) \geq \frac{1}{2}\phi''(1)V(P,Q)^2$. Then the assertion follows in view of the decomposition (40) in proposition 4.1. □

We note that $2Q_{\min}P_{\min} = Q_{\min}P_{\min} + Q_{\min}P_{\min} < P_{\min} + Q_{\min}$. Thus the second factor multiplying $V(P,Q)^2$ is greater than 1, as it should be. By [19, Example II, p.2390], for $\pi = \frac{1}{2}$ there is the bound $D_{\mathrm{RJS}}(P,Q) \geq -\frac{1}{2}\ln\left(1 - V(P,Q)^2\right)$, which is valid for probabilities on general sample spaces. By elementary calculus, $-\frac{1}{2}\ln\left(1 - V(P,Q)^2\right) > \frac{1}{8}V(P,Q)^2$, the lower bound above for $\pi = \frac{1}{2}$.

Corollary 4.3.

$$D_{\mathrm{JS}}(P,Q) \leq 2\phi''(1)\left[\frac{e^{D_{\mathrm{CD}}(P,Q)}}{e^{D_{\mathrm{CD}}(P,Q)}-1} - \ln\frac{D_{\mathrm{CD}}(P,Q)}{e^{D_{\mathrm{CD}}(P,Q)}-1} - 1\right]. \tag{45}$$

The corollary follows by the bound [6, Thm 3.4,p.157] for $D_{\mathrm{KL}}(P,Q)$ in terms of $D_{\mathrm{CD}}(P,Q)$ from (2)

$$D_{\mathrm{KL}}(P,Q) \leq \frac{e^{D_{\mathrm{CD}}(P,Q)}}{e^{D_{\mathrm{CD}}(P,Q)}-1} - \ln\frac{D_{\mathrm{CD}}(P,Q)}{e^{D_{\mathrm{CD}}(P,Q)}-1} - 1, \tag{46}$$

and by (8), (40) and (46), since $D_{\mathrm{RJS}}(P,Q) \geq 0$.

4.3. Two examples

Example 4.4. Fragile sites on chromosomes are points at which the chromosome is likely to break. Identification of fragile sites is thought to contribute to the detection of genetic abnormalities. The authors of [5] propose a statistical genetics model for fragile sites named as the fragile site multinomial. Let k be the total number of sites on a chromosome, k_1 is the number of non-fragile sites, i. e., the k categories in $\mathcal{A}$ are split up into two subsets with k_1 ($< k$) and $k - k_1$ elements. In [5] the probabilities p_i satisfy $p_1 = \ldots = p_{k_1}$ and $p_i < \frac{1}{k}, i = 1, \ldots, k_1$, and $p_i > \frac{1}{k}, i = k_1 + 1, \ldots, k$, $\sum_{i=1}^{k} p_i = 1$. The work in [24] finds a method to estimate k_1 and tests it on different listed structures for two such subsets.

We insert θ, a pressure of breaking, into one of the structures of [24, Table 2, p.438] by setting $p_i = \frac{\theta}{k_1}$ for $i = 1, 2, \ldots k_1$, and $p_i = \frac{(1-\theta)}{(k-k_1)}, i = k_1 + 1, \ldots, k$. If $0 < \theta < \frac{k_1}{k}$, we have $\frac{\theta}{k_1} < \frac{1}{k}$, $\frac{(1-\theta)}{(k-k_1)} > \frac{1}{k}$ and $\sum_{i=1}^{k} p_i = 1$. The model $\mathbb{M}_{\mathrm{FSM}}$ is

$$\mathbb{M}_{\mathrm{FSM}} = \left\{ P_\theta | P_\theta(x) = \prod_{i=1}^{k_1}\left(\frac{\theta}{k_1}\right)^{[x=a_i]} \cdot \prod_{i=k_1+1}^{k}\left(\frac{1-\theta}{k-k_1}\right)^{[x=a_i]}, 0 < \theta < \frac{k_1}{k} \right\}. \tag{47}$$

Hence $\mathbb{M}_{\mathrm{FSM}}$ is an example of a prescribed model. For $\theta \in [0,1]$, $B(\theta)$ is given in (19). The Shannon entropy of $P_\theta \in \mathbb{M}_{\mathrm{FSM}}$ in (47) equals

$$H\big(P_\theta; \mathrm{FSM}\big) = B(\theta) + \ln(k - k_1) + \theta \cdot \ln\left(\frac{k_1}{k-k_1}\right). \tag{48}$$

Let $P_{\theta_1} \in \mathbb{M}_{\mathrm{FSM}}$ and $P_{\theta_2} \in \mathbb{M}_{\mathrm{FSM}}$. When (48) is applied in (42), we obtain

$$D_{\mathrm{JS}}\big(P_{\theta_1}, P_{\theta_2}; \mathrm{FSM}\big) = B\left(\pi\theta_1 + (1-\pi)\theta_2\right) - \pi B(\theta_1) - (1-\pi)B(\theta_2). \tag{49}$$

If $p_{\theta_i} \in \mathbb{M}_{\mathrm{Be}}$, $i = 1, 2$ in (3), we find $D_{\mathrm{JS}}\big(P_{\theta_1}, P_{\theta_2}; \mathrm{FSM}\big) = D_{\mathrm{JS}}(p_{\theta_1}, p_{\theta_2})$. The requirements in (47) make $\mathbb{M}_{\mathrm{FSM}}$, in terms of the statistical information measured by JSD,

nothing else than a model on a binary category set. If $P_{\theta_1} \in \mathbb{M}_{\mathrm{FSM}}$ and $P_{\theta_2} \in \mathbb{M}_{\mathrm{FSM}}$, then $V(P_{\theta_1}, P_{\theta_2}) = 2|\theta_1 - \theta_2| = V(p_{\theta_1}, p_{\theta_2})$. In the bounds of (27) we have

$$c_1|\theta_1 - \theta_2|^2 \leq D_{\mathrm{JS}}\big(P_{\theta_1}, P_{\theta_2}; \mathrm{FSM}\big) \leq c_{u,1}|\theta_1 - \theta_2|^2. \tag{50}$$

where $c_1 = 2\phi^{''}(1)$, $c_{u,1} = 4.5\phi^{''}(1)/\min\{\pi\theta_1, (1-\pi)\theta_2\}$. By the upper bound of (43), $c_1|\theta_1 - \theta_2|^2 \leq D_{\mathrm{JS}}(P_{\theta_1}, P_{\theta_2}; \mathrm{FSM}) \leq c_{u,2}|\theta_1 - \theta_2|^2$, where $c_{u,2} = \frac{3\phi^{''}(1)(\theta_1+\theta_2)}{2(\theta_1\theta_2-1)}$. When the optimal error of discrimination (34) is computed, we get

$$\frac{1}{2} - \mathbf{P}_e\left(P_{\theta_1}, P_{\theta_2}; \mathrm{FSM}\right) = \max\{\theta_1, \theta_2\} - \min\{\theta_1, \theta_2\}.$$

We choose $\pi = 1/2$ and set $c(\theta_1, \theta_2) = \sqrt{1/\min\{\theta_1, \theta_2\}}$ and $d(\theta_1, \theta_2) := \max\{\theta_1, \theta_2\} - \min\{\theta_1, \theta_2\}$. Hence we get by (33) the inequalities

$$\sqrt{2}d(\theta_1, \theta_2) \leq D_{\mathrm{JS},1/2}(P_{\theta_1}, P_{\theta_2}; \mathrm{FSM})^{1/2} \leq 2c(\theta_1, \theta_2)\, d(\theta_1, \theta_2),$$

or $D_{\mathrm{JS},1/2}(P_{\theta_1}, P_{\theta_2}; \mathrm{FSM})^{1/2} \asymp d(\theta_1, \theta_2)$.

Example 4.5. In [37, p.360] there is the following model $\mathbb{M}_{\mathrm{F}}$ with four categories

$$\left\{ P_\theta | P_\theta(x) = \left(\frac{2+\theta}{4}\right)^{[x=a_1]} \left(\frac{1-\theta}{4}\right)^{[x=a_2]} \left(\frac{1-\theta}{4}\right)^{[x=a_3]} \left(\frac{1-\theta}{4}\right)^{[x=a_4]}, 0 < \theta < 1 \right\}.$$

The reference in [37, loc.cit.] is to R.A. Fisher's research in plant genetics. It is possible to write down various explicit expressions for $D_{\mathrm{JS}}(P_{\theta_1}, P_{\theta_2}; \mathrm{F})$, e.g, by using (42). But the resulting sums containing a number of logarithmic terms seem unwieldy. However, for $P_{\theta_1} \in \mathbb{M}_{\mathrm{F}}$ and $P_{\theta_2} \in \mathbb{M}_{\mathrm{F}}$, we have $V(P_{\theta_1}, P_{\theta_2}) = |\theta_1 - \theta_2|$. Hence we obtain by (27) and $c(\theta_1, \theta_2) = (1/\min\{\theta_1, \theta_2\})^{1/2}$ that $\frac{\sqrt{2}}{4}|\theta_1 - \theta_2| \leq D_{\mathrm{JS},1/2}(P_{\theta_1}, P_{\theta_2}; \mathrm{F})^{1/2} \leq 2c(\theta_1, \theta_2)|\theta_1 - \theta_2|$, or $D_{\mathrm{JS},1/2}(P_{\theta_1}, P_{\theta_2}; \mathrm{F})^{1/2} \asymp |\theta_1 - \theta_2|$, which is both manageable and informative.

5. ON ASYMPTOTIC EQUIVALENCE BETWEEN MAXIMUM LIKELIHOOD ESTIMATE AND MINIMUM JSD-DIVERGENCE ESTIMATE

In this section, $\mathbb{M} = \{P_\theta \mid \theta \in \Theta\}$ is a parametric model in $\mathbb{P}$ with full support for every $\theta \in \Theta$, $\mathbf{X} = (X_1, \ldots, X_n) \sim P_{\theta_o} \in \mathbb{M}$ are i.i.d. random variables. As is readily checked, see e. g., [37, Eqn. (3), p.350], the maximum likelihood estimate is

$$\widehat{\theta}_n \in \arg\min_{\theta \in \Theta} D_{\mathrm{KL}}(\widehat{P}_n, P_\theta), \tag{51}$$

where $\widehat{P}_n$ is given by (22) for $\mathbf{X}$. In the Lemma below the assumption of full support $\mathrm{supp}(\widehat{P}_n) = \mathcal{A}$ for every $n > N$ seems natural, as we are assuming that every a_i is found in the sample for a large N and will appear infinitely often as $n \to +\infty$.

Lemma 5.1. $P_{\theta_o} \in \mathbb{M}$, $\mathrm{supp}(P_\theta) = \mathcal{A}$ for every $\theta \in \Theta$, and $\mathrm{supp}(\widehat{P}_n) = \mathcal{A}$ for every $n \geq N$ for some $N > 1$ with P_{θ_o}-probability one. Then it holds that

$$D_{\mathrm{JS}}\left(P_{\theta_o}, P_{\widehat{\theta}_n}\right) \to 0, \quad P_{\theta_o}\text{- almost surely, as } n \to +\infty. \tag{52}$$

P r o o f . By (43) and triangle inequality it holds that

$$\sqrt{D_{\mathrm{JS}}\left(P_{\theta_o}, P_{\widehat{\theta}_n}\right)} \leq c(n) V\left(P_{\theta_o}, P_{\widehat{\theta}_n}\right) \leq c(n)\left[V\left(P_{\theta_o}, \widehat{P}_n\right) + V\left(\widehat{P}_n, P_{\widehat{\theta}_n}\right)\right], \tag{53}$$

where we have simplified the writing by letting $c(n)$ denote the square root of the factor multiplying $V\left(P_{\theta_o}, P^{(n)}_{\widehat{\theta}}\right)$ in (43).

It follows by [14, Theorem 1, p.896], see in particular [14, Lemma 3, p.898] and [3, Eqn. (17), p.1258], that there is complete convergence $\widehat{P}_n \to P_{\theta_o}$ implying by Borel-Cantelli that

$$V(\widehat{P}_n, P_{\theta_o}) \to 0, \quad P_{\theta_o}\text{-almost surely, as } n \to +\infty. \tag{54}$$

Therefore, since $P_{\theta_o} \in \mathbb{M}$, (51) and the reverse Pinsker's inequality (25) entail, when $P_{\theta_o,\min} := \min_{x\in\mathcal{A}} P_{\theta_o}(x) > 0$,

$$D_{\mathrm{KL}}\left(\widehat{P}_n, P_{\widehat{\theta}_n}\right) \leq D_{\mathrm{KL}}\left(\widehat{P}_n, P_{\theta_o}\right) \leq \frac{1}{2P_{\theta_o,\min}} V\left(\widehat{P}_n, P_{\theta_o}\right)^2 \to 0 \quad P_{\theta_o} \text{ a.s.} \tag{55}$$

Then (24) gives

$$V\left(\widehat{P}_n, P_{\widehat{\theta}_n}\right) \to 0, P_{\theta_o} \text{ a.s..} \tag{56}$$

By (43), and since $\widehat{P}_{n,min} \to P_{\theta_o,\min}$ as $n \to +\infty$ P_{θ_o}-a.s., as is easily verified, $c(n)$ converges P_{θ_o}-almost surely to a finite positive limit by the assumptions about full supports and by the continuous mapping theorem for a.s. convergence. The assertion now follows by continuity of the square root. □

We define the minimum JSD -divergence estimate $\widehat{\theta}_{\mathrm{JS}}$ given $\mathbf{X}$ as

$$\widehat{\theta}_{n,\mathrm{JS}} \in \arg\min_{\theta\in\Theta} D_{\mathrm{JS}}(\widehat{P}_n, P_\theta). \tag{57}$$

This minimization can be done in practice, e. g., by the techniques applied in Section 6 below. The result in the next Proposition shows that the computable $\widehat{\theta}_{n,\mathrm{JS}}$ is asymptotically equivalent to the maximum likelihood estimate. This holds, as soon as the generative model is written in an programming language with expressiveness to include the true distribution, even if the likelihood is implicit or intractable.

Proposition 5.2. $P_{\theta_o} \in \mathbb{M}$, $\mathrm{supp}(P_\theta) = \mathcal{A}$ for every $\theta \in \Theta$, and $\mathrm{supp}(\widehat{P}_n) = \mathcal{A}$ for every $n \geq N$ for some $N > 1$ with P_{θ_o}-probability one. Then it holds that

$$D_{\mathrm{JS}}\left(P_{\widehat{\theta}_{n,\mathrm{JS}}}, P_{\widehat{\theta}_n}\right) \to 0, \quad P_{\theta_o}\text{- almost surely, as } n \to +\infty, \tag{58}$$

and

$$V\left(P_{\theta_o}, P_{\widehat{\theta}_{n,\mathrm{JS}}}\right) \to 0, \quad P_{\theta_o}\text{- almost surely, as } n \to +\infty. \tag{59}$$

P r o o f . As in the proof of preceding lemma we have

$$\sqrt{D_{\mathrm{JS}}\left(P_{\widehat{\theta}_{n,\mathrm{JS}}}, P_{\widehat{\theta}_n}\right)} \leq c(n)\left(V\left(P_{\widehat{\theta}_{n,\mathrm{JS}}}, P_{\theta_o}\right) + V\left(P_{\theta_o}, P_{\widehat{\theta}_n}\right)\right). \tag{60}$$

The triangle inequality gives furthermore

$$V\left(P_{\widehat{\theta}_{n,\mathrm{JS}}}, P_{\theta_o}\right) \leq V\left(P_{\theta_o}, \widehat{P}_n\right) + V\left(\widehat{P}_n, P_{\widehat{\theta}_{n,\mathrm{JS}}}\right). \tag{61}$$

By the left hand side inequality in (43) of proposition 3.2 and (57), since $P_{\theta_o} \in \mathbb{M}$,

$$V\left(\widehat{P}_n, P_{\widehat{\theta}_{n,\mathrm{JS}}}\right) \leq c_1\sqrt{D_{\mathrm{JS}}\left(\widehat{P}_n, P_{\widehat{\theta}_{n,\mathrm{JS}}}\right)} \leq c_1\sqrt{D_{\mathrm{JS}}\left(\widehat{P}_n, P_{\theta_o}\right)}. \tag{62}$$

By the right hand side inequality in (43) in proposition 3.2

$$\sqrt{D_{\mathrm{JS}}\left(\widehat{P}_n, P_{\theta_o}\right)} \leq c(n)V\left(\widehat{P}_n, P_{\theta_o}\right) \to 0, \quad \text{a.s. } P_{\theta_o}, \tag{63}$$

since by the proof of lemma 5.1, $c(n)$ converges P_{θ_o} a.s. to a finite limit as $n \to +\infty$. Thus by (62) and (61) $V\left(P_{\widehat{\theta}_{n,\mathrm{JS}}}, P_{\theta_o}\right) \to 0$, P_{θ_o} a.s., as $n \to +\infty$. It holds by a result in the proof of Lemma 5.1 that $V\left(P_{\theta_o}, P_{\widehat{\theta}_n}\right) \to 0$, P_{θ_o} a.s., as $n \to +\infty$. When these facts are used in the right hand side of (60), the proof is completed. □

When $\mathbf{X}$ is an i..i.d. n-sample, then with multiplications of (1) and n_i is the number of occurrences of a_i in $\mathbf{X}$, the loglikelihood function of θ is for $P_\theta \in \mathbb{P}$

$$l_{\mathbf{X}}(\theta) = \sum_{i=1}^{k} n_i \ln p_i(\theta). \tag{64}$$

By (15), $D_{\mathrm{JS}}(\widehat{P}_n, P_\theta) = \pi D_{\mathrm{KL}}(\widehat{P}_n, P_\theta) - D_{\mathrm{KL}}(\widehat{M}_n, P_\theta)$, where $\widehat{M}_n = \pi\widehat{P}_n + (1-\pi)P_\theta$. By definition of KL, $D_{\mathrm{KL}}(\widehat{P}_n, P_\theta) = -\frac{\pi}{n}l_{\mathbf{X}}(\theta) - H\left(\widehat{P}_n\right)$. This entails, since $H\left(\widehat{P}_n\right)$ does not depend on θ, that

$$\arg\min_{\theta\in\Theta} D_{\mathrm{JS}}(\widehat{P}_n, P_\theta) = \arg\min_{\theta\in\Theta}\left[-\frac{\pi}{n}l_{\mathbf{X}}(\theta) - D_{\mathrm{KL}}(\widehat{M}_n, P_\theta)\right]. \tag{65}$$

Some auxiliary piece of notation is helpful for simplification. $L(\theta) := -\frac{\pi}{n}l_{\mathbf{X}}(\theta)$ and $D(\theta) := D_{\mathrm{KL}}(\widehat{M}_n, P_\theta)$. Both $L(\theta)$ and $D(\theta)$ are non-negative. Set $A = \{\theta \in \Theta | L(\theta) \geq D(\theta)\}$, A^c is the complement set. $\mathbf{I}_A(\theta)$ denotes the indicator function. Then

$$L(\theta) - D(\theta) = \mathbf{I}_A(\theta)\cdot|L(\theta) - D(\theta)| - \mathbf{I}_{A^c}(\theta)\cdot|L(\theta) - D(\theta)|.$$

We are thus in the right hand side of (65) searching for $\theta \in \Theta$ such that $|L(\theta) - D(\theta)|$ is minimized, i. e., θ such that $|L(\theta) - D(\theta)|$ should be as close to zero, as possible.

By reverse Pinsker (25) $D_{\mathrm{KL}}(\widehat{M}_n, P_\theta) \le V\left(M_n, P_\theta\right)^2 / 2 \min_i p_i(\theta)$. By Pinsker (24) $D_{\mathrm{KL}}(\widehat{M}_n, P_\theta) \ge V\left(\widehat{M}_n, P_\theta\right)^2 / 2$. Here $V\left(\widehat{M}_n, P_\theta\right) = \pi \sum_{i=1}^k |\widehat{p}_i - p_i(\theta)|$. By standard probability, $V\left(\widehat{M}_n, P_\theta\right)$ is an outcome of $\pi \sum_{i=1}^d \left|\frac{\xi_i}{n} - p_i(\theta)\right|$, where each ξ_i is binomially distributed $\mathrm{Bin}(n, p_i(\theta))$. In view of [3, p.1258], $\sum_{i=1}^d \left|\frac{\xi_i}{n} - p_i(\theta)\right| \to 0$, P_θ-a.s., as $n \to +\infty$. Hence we would expect, at least when θ is sufficiently close to the true parameter value, that with a small distortion proportional to $\sum_{i=1}^k |\widehat{p}_i - p_i(\theta)|$,

$$-\frac{\pi}{n} l_{\mathbf{X}}(\theta) - D_{\mathrm{KL}}(\widehat{M}_n, P_\theta) \approx -\frac{\pi}{n} l_{\mathbf{X}}(\theta) - O\left(\sum_{i=1}^k |\widehat{p}_i - p_i(\theta)|\right). \tag{66}$$

One so-called non-parametric kernel method for simulator based likelihood-free inference on θ, c.f. [20, p.10], is exemplified by

$$\overline{D}_{\mathrm{JS}}^{(m)}(\theta) := \frac{1}{m} \sum_{l=1}^m D_{\mathrm{JS}}(\widehat{P}_n, \widehat{Q}^{(l)}), \tag{67}$$

where $\widehat{Q}^{(l)}$ are empirical distributions corresponding to m independent synthetic i.i.d. n-samples $\sim P_\theta$, $l = 1, \ldots, m$. This is called a kernel-GAN in [44], where the asymptotics for $m \to +\infty$ is studied. Then from (65) and (66)

$$\arg\min_{\theta\in\Theta} \overline{D}_{\mathrm{JS}}^{(m)}(\theta) \approx \arg\min_{\theta\in\Theta} \left[-\frac{\pi}{n} \sum_{i=1}^k n_i \frac{1}{m} \sum_{l=1}^m \ln \widehat{q}_i^{(l)} - O\left(\sum_{i=1}^k \frac{1}{m} \sum_{l=1}^m \left|\widehat{q}_i^{(l)} - p_i(\theta)\right|\right)\right].$$

Here for large m and n, we expect $\frac{1}{m} \sum_{l=1}^m \ln \widehat{q}_i^{(l)} \approx \ln p_i(\theta)$ and the distortion term $O \approx 0$. Hence, a minimized $\overline{D}_{\mathrm{JS}}^{(m)}$ seems to deliver an estimate of MLE for an implicit likelihood function. A rigorous analysis of $\overline{D}_{\mathrm{JS}}^{(m)}$ will be presented elsewhere. We illustrate this by means of a simulation recycling the example 4.4.

Example 5.3. The model in (47) is, of course, not implicit and the MLE based on an n-sample is explicit, i. e., $\widehat{\theta}_n = \frac{1}{n} \sum_{i=1}^{k_1} n_i$, when k_1 is given in advance. Let us take $k = 4$ and $k_1 = 3$ in example 4.4. In this experiment we simulate $\widehat{P}_n$, too. One hundred i.i.d. random samples from $P_{0.25}$ yielded $\widehat{P}_{100}$ with $\widehat{p}_1 = 0.08$, $\widehat{p}_2 = 0.09$, $\widehat{p}_3 = 0.06$, $\widehat{p}_4 = 0.77$, so that $\widehat{\theta}_{100} = (8 + 9 + 6)/100 = 0.23$. Thereafter we compute $\overline{D}_{\mathrm{JS},1/2}^{(1000)}(\theta)$ in (67) for $\theta \in (0, \frac{3}{4})$ with the step 0.01 on a grid, on which $D_{\mathrm{JS},1/2}(\widehat{P}_{100}, P_\theta; \mathrm{FSM})$ is also evaluated by (49). $\overline{D}_{\mathrm{JS},1/2}^{(1000)}(\theta)$ and $D_{\mathrm{JS},1/2}(\widehat{P}_{100}, P_\theta; \mathrm{FSM})$ can then be plotted. We plot also a third function of θ to be defined next.

The following statements are valid for all $\mathbb{M}$. By standard probability, the frequencies $\left(\xi_j^{(l)}\right)_{j=1}^k$ in any $\widehat{Q}^{(l)}$ have a multinomial distribution w.r.t. P_θ. Hence it holds for any $\widehat{Q}^{(l)}$ that

$$p_\theta := P_\theta\left(\widehat{Q}^{(l)} = \widehat{P}_n\right) = P_\theta\left(\xi_1^{(l)} = n_1, \ldots, \xi_k^{(l)} = n_k\right) = \frac{n!}{\prod_{i=1}^k n_i!} \prod_{i=1}^k p_i(\theta)^{n_i}.$$

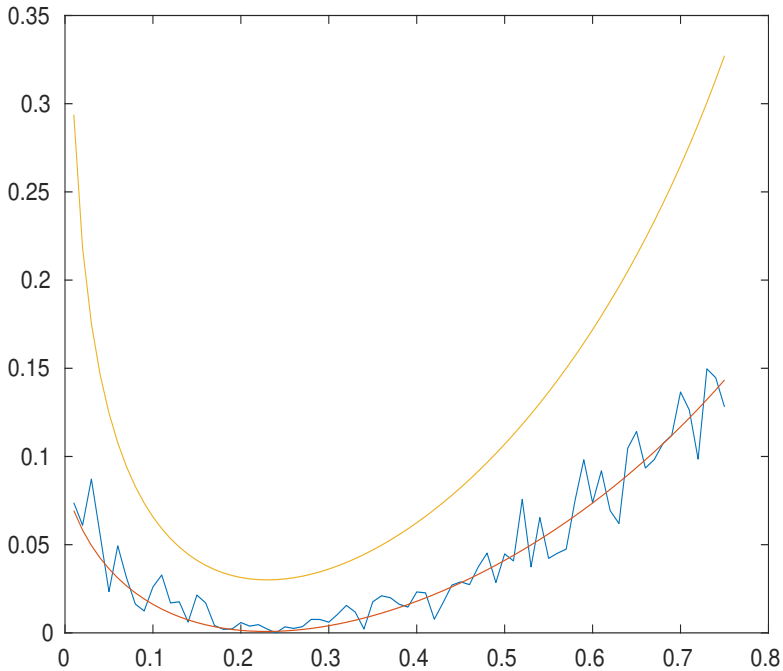


Fig. 1. $\overline{D}^{(1000)}_{\mathrm{JS},1/2}(\theta)$ (blue) , $D_{\mathrm{JS},1/2}(\widehat{P}_{100}, P_\theta; \mathrm{FSM})$ (red), and $-\frac{1}{200}\ln(p_\theta)$ (yellow).

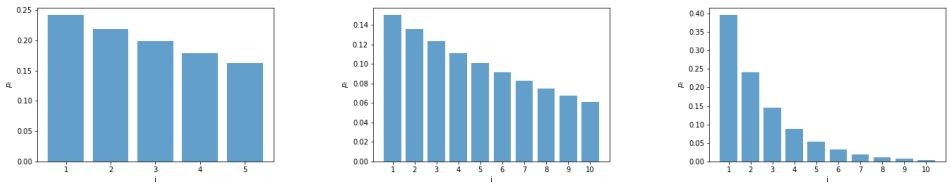
This connects to the method of types, [10, Chapter 11.1]. Then it can be shown for sufficiently large n and any $m > 1$, that $\overline{D}^{(m)}_{\mathrm{JS}}(\theta) \leq -\frac{\pi}{n}\ln(p_\theta)$. This is valid for any D_{JS}. The proof, which would require additional machinery, is omitted here and only a graphical evidence is given. We plot $\overline{D}^{(1000)}_{\mathrm{JS},1/2}(\theta)$, $D_{\mathrm{JS},1/2}(\widehat{P}_{100}, P_\theta; \mathrm{FSM})$ and $-\frac{1}{200}\ln(p_\theta)$ in Figure 1, where $\overline{D}^{(1000)}_{\mathrm{JS},1/2}(\theta)$ is, by graphical inspection, seen to be small close to $\theta_o = 0.25$.

6. SIMULATION EXPERIMENTS

In this Section we run a simulation experiment to study the properties of maximum likelihood and minimum JSD estimates. We apply software documented in [4]. In our simulations we use the following categorical distribution on $\mathcal{A} = \{i \in \mathbb{Z}_+ \mid i = 1, \ldots, k\}$

$$p_i(\theta) = C(\theta)^{-1} e^{-\theta(i-1)}, \quad i \in \mathcal{A}, \tag{68}$$

where, if the model parameter $\theta > 0$, the normalization constant $C(\theta) = e^\theta A(\theta)$ with $A(\theta) = (e^{-\theta} - e^{-\theta(k+1)})/\left(1 - e^{-\theta}\right)$. If $\theta < 0$, then $A(\theta) = (e^{-\theta(k+1)} - e^{-\theta})/\left(e^{-\theta} - 1\right)$.



(a) (b) (c)

Fig. 2. P_θ calculated based on (a) $k = 5$, $\theta = 0.1$, (b) $k = 10$, $\theta = 0.1$, and (c) $k = 10$, $\theta = 0.5$.

Here the likelihood is neither implicit nor analytically intractable. The loglikelihood function in (64) becomes $l_{\mathbf{X}}(\theta) = -\theta \sum_{i=1}^{k} i n_i - n \ln A(\theta)$, and the MLE $\widehat{\theta}_n$ satisfies the equation $A'\left(\widehat{\theta}_n\right)/A\left(\widehat{\theta}_n\right) = -\sum_{i=1}^{k} i\widehat{p}_i$, where $\widehat{p}_i = n_i/n$.

We test parameter estimation with (a) $k = 5$, $\theta = 0.1$, (b) $k = 10$, $\theta = 0.1$, and (c) $k = 10$, $\theta = 0.5$ (Figure 2). We use the categorical distributions (a)–(c) to simulate 1000 observation sets with $n = \{50, 100, 500, 1000\}$ data points. Each observation set is then used to calculate the maximum likelihood estimate $\widehat{\theta}_n$ and the minimum JSD estimate $\widehat{\theta}_{n,\mathrm{JS}}$, see (57). Moreover we calculate minimum JSD estimates based on JSD with $\pi = \{0.4, 0.5, 0.6\}$. Figure 3 shows the estimated parameter values.

To evaluate and compare the parameter estimates, we calculate the root mean squared error (RMSE) between the estimated and true parameter values. This captures both the estimator bias and variance. The errors are presented in Table 1. We observe that when the sample size is small, the minimum JSD estimates are associated with larger errors than the maximum likelihood estimates, but the difference disappears as sample size increases. This is as expected, since the estimates are asymptotically equivalent. The estimation error also approaches zero as sample size increases, and depends on both k and θ. Namely we observe that the increase in k between setups (a) and (b) decreases the estimation error and the increase in θ between setups (b) and (c) increases the error. The increase in error is due to an overestimation bias seen as a right-hand tail in the histograms in Figure 3 (c). We observe a bias here because the model used in this experiment is such that the parameter θ tends to be overestimated when the expected counts np_i are small in some categories i. This effect will appear for $\mathcal{A}$ with a large cardinality, as there will be several cases of rarely observed categories.

7. THIRD EXAMPLE: MULTIVARIATE BERNOULLI DISTRIBUTION

7.1. Bounds on JSD of two multivariate bernoulli distributions

Here we use the boundings and the decomposition above to establish that the JSD for d -variate Bernouli distributions behaves like $D_{\mathrm{JS}}(P_{\theta_1}, P_{\theta_2}) \asymp \phi''(1) \sum_{j=1}^{d} (\theta_{j,1} - \theta_{j,2})^2$, where $\theta_{j,l}$ are the respective marginal probabilities of success of the two d-variate Bernoulli distributions.

Let $k = 2^d$ and $\mathcal{A} = \{\mathbf{a}^{(1)}, \ldots \mathbf{a}^{(k)}\} = \{0, 1\}^d$ is the binary hypercube, where $\mathbf{a}^{(i)} =$

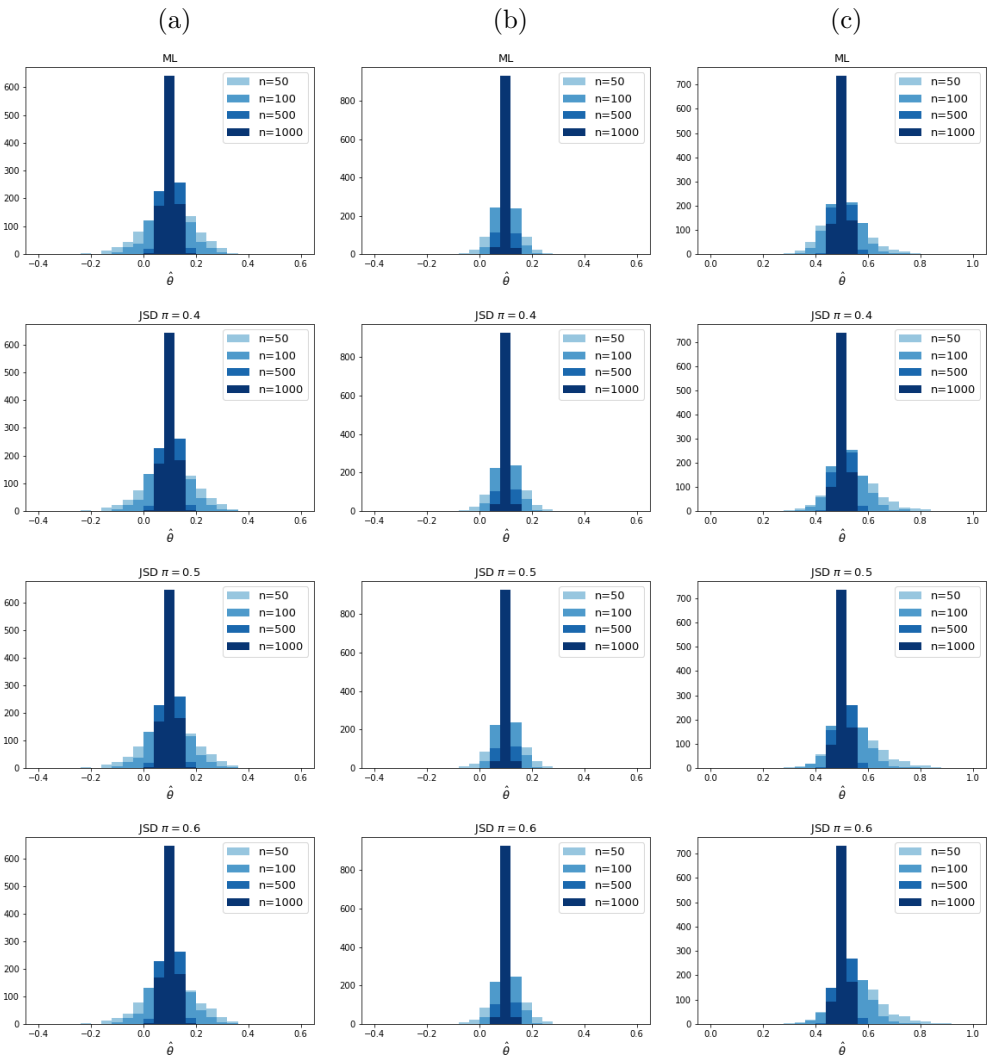


Fig. 3. Distribution over 1000 ML and minimum JSD estimates calculated based on n observations when (a) $k = 5$, $\theta = 0.1$ (b) $k = 10$, $\theta = 0.1$ (c) $k = 10$, $\theta = 0.5$.

	n	50	100	500	1000
	ML	0.097	0.070	0.031	0.022
(a)	min JSD $\pi = 0.4$	0.099	0.070	0.031	0.022
	min JSD $\pi = 0.5$	0.100	0.071	0.031	0.022
	min JSD $\pi = 0.6$	0.100	0.071	0.031	0.022
	n	50	100	500	1000
	ML	0.051	0.035	0.016	0.011
(b)	min JSD $\pi = 0.4$	0.054	0.036	0.016	0.011
	min JSD $\pi = 0.5$	0.056	0.037	0.016	0.011
	min JSD $\pi = 0.6$	0.057	0.037	0.016	0.011
	n	50	100	500	1000
	ML	0.082	0.058	0.026	0.018
(c)	min JSD $\pi = 0.4$	0.099	0.067	0.026	0.018
	min JSD $\pi = 0.5$	0.106	0.070	0.027	0.018
	min JSD $\pi = 0.6$	0.114	0.074	0.027	0.018

Tab. 1. RMSE evaluated based on 1000 MLE or minimum JSD estimates calculated based on n observations when (a) $k = 5$, $\theta = 0.1$ (b) $k = 10$, $\theta = 0.1$ (c) $k = 10$, $\theta = 0.5$.

$\left(a_1^{(i)}, \ldots, a_d^{(i)}\right)$, $a_j^{(i)} \in \{0, 1\}$. We take $\theta_j \in (0, 1)$, $j = 1, \ldots, d$. Let $\theta = (\theta_1, \ldots, \theta_d)$. We set

$$p_i(\theta) := \prod_{j=1}^{d} p_{\theta_j}(a_j^{(i)}), \tag{69}$$

where $p_{\theta_j}(a_j^{(i)})$ is given as in (3), i. e., it is the probability mass function of a Bernoulli random variable with θ_j as the probability of success. Then with

$$P_\theta(x) := \prod_{i=1}^{k} p_i(\theta)^{[x=\mathbf{a}^{(i)}]}, \quad x \in \mathcal{A}, \tag{70}$$

the model $\mathbb{M}_{\mathrm{MBe}} = \left\{P_\theta \mid \theta \in \Theta = (0, 1)^d\right\}$ is the family of multivariate Bernoulli distributions on the binary hypercube $\{0, 1\}^d$. This means modeling the bits $a_j^{(i)}$ as independent Bernoulli r.v.'s. It holds that $\operatorname{supp}(P_\theta) = \mathcal{A}$ for all $\theta \in \Theta$.

Let $P_{\theta_1} \in \mathbb{M}_{\mathrm{MBe}}$ and $P_{\theta_2} \in \mathbb{M}_{\mathrm{MBe}}$, where $\theta_s = (\theta_{1,s}, \ldots, \theta_{d,s})$ for $s = 1, 2$. If we set $M = \pi P_{\theta_1} + (1-\pi) P_{\theta_2}$, then for any $x \in \{0, 1\}^d$, we have $M(x) = \pi \prod_{i=1}^{k} p_i(\theta_1)^{[x=\mathbf{a}^{(i)}]} + (1 - \pi) \prod_{i=1}^{k} p_i\,(\theta_2)^{[x=\mathbf{a}^{(i)}]}$. This mixture of distributions is not in $\mathbb{M}_{\mathrm{MBe}}$. In fact, mixtures of multivariate Bernoulli distributions are not identifiable, [21]. Of course, the variation distance is also hard to evaluate analytically for multivariate Bernoulli distributions. However, another round of reverse Pinsker inequalities will render the bounds in Proposition 4.2 useful.

Proposition 7.1. $P_{\theta_1} \in \mathbb{M}_{\mathrm{MBe}}$ and $P_{\theta_2} \in \mathbb{M}_{\mathrm{MBe}}$. Then

$$D_{\mathrm{JS}}\left(P_{\theta_1}, P_{\theta_2}\right) \geq 4\phi^{''}(1)\left[\prod_{j=1}^{d} \min(1-\theta_{j,2}, \theta_{j,2})\right] \cdot \sum_{j=1}^{d} \left(\theta_{i,1}-\theta_{j,2}\right)^2. \tag{71}$$

P r o o f. For the proof of (71), we plan to use the left hand inequality in (27). The following steps are done.

Step 1. Let, as in Example 2.3, $p_{\theta_{j,1}} \in \mathbb{M}_{\mathrm{Be}}$ and $p_{\theta_{j,2}} \in \mathbb{M}_{\mathrm{Be}}$ be two Bernoulli distributions with probabilities of success $\theta_{j,1}$ and $\theta_{j,2}$, respectively. Then [39, Lemma 2 (a), pp.29–30] tells that

$$D_{\mathrm{KL}}\left(p_{\theta_{j,1}}, p_{\theta_{j,2}}\right) \geq 2\left(\theta_{j,1}-\theta_{j,2}\right)^2. \tag{72}$$

Step 2. In view of the definition of D_{KL} we get by (70) and (69)

$$D_{\mathrm{KL}}(P_{\theta_1}, P_{\theta_2}) = \sum_{i=1}^{k}\prod_{j=1}^{d} p_{\theta_{j,1}}(a_j^{(i)}) \ln \frac{\prod_{j=1}^{d} p_{\theta_{j,1}}(a_j^{(i)})}{\prod_{j=1}^{d} p_{\theta_{j,2}}(a_j^{(i)})}.$$

Hence, from a standard property of KL for product distributions, we get that

$$D_{\mathrm{KL}}(P_{\theta_1}, P_{\theta_2}) = \sum_{j=1}^{d} D_{\mathrm{KL}}\left(p_{\theta_{j,1}}, p_{\theta_{j,2}}\right) \geq 2\sum_{j=1}^{d}\left(\theta_{j,1}-\theta_{j,2}\right)^2. \tag{73}$$

Step 3. By reverse Pinsker (25), (27) and (73)

$$D_{\mathrm{JS}}\left(P_{\theta_1}, P_{\theta_2}\right) \geq 4\pi(1-\pi)\sum_{j=1}^{d}\left(\theta_{j,1}-\theta_{j,2}\right)^2 \min_{x\in\{0,1\}^d} P_{\theta_2}(x). \tag{74}$$

Step 4. It remains to establish that $\min_{x\in\{0,1\}^d} P_{\theta_2}(x) = \prod_{j=1}^{d} \min\{1-\theta_{j,2}, \theta_{j,2}\}$. This requires some additional auxiliary quantities and is done in Appendix A.1.

□

We set next

$$b\left(\theta_1, \theta_2\right) = \frac{1}{\min\left\{\left[\prod_{j=1}^{d} \min\{1-\theta_{j,1}, \theta_{j,1}\}\right], \left[\prod_{j=1}^{d} \min\{1-\theta_{j,1}, \theta_{j,1}\}\right]\right\}}.$$

Proposition 7.2. $P_{\theta_1} \in \mathbb{M}_{\mathrm{MBe}}$ and $P_{\theta_2} \in \mathbb{M}_{\mathrm{MBe}}$. Then

$$D_{\mathrm{JS}}\left(P_{\theta_1}, P_{\theta_2}\right) \leq \phi^{''}(1) \cdot b\left(\theta_1, \theta_2\right)\sum_{j=1}^{d}\left(\theta_{j,1}-\theta_{j,2}\right)^2. \tag{75}$$

P r o o f. By construction $\operatorname{supp}(P_{\theta_1}) = \operatorname{supp}(P_{\theta_2}) = \{0,1\}^d$. Thus we can use (40), where $0 \le D_{\mathrm{RJS}}(P_{\theta_1}, P_{\theta_2}) < +\infty$, to get $D_{\mathrm{JS}}(P_{\theta_1}, P_{\theta_2}) \le \phi^{''}(1) D_{\mathrm{Je}}(P_{\theta_1}, P_{\theta_2})$. Let us apply the reverse Pinsker (25) of Lemma 3.1 in each individual $D_{\mathrm{KL}}\left(p_{\theta_{j,1}}, p_{\theta_{j,2}}\right)$ in the sum in the right hand side of (73). This gives $D_{\mathrm{KL}}(P_{\theta_1}, P_{\theta_2}) \le 4\sum_{j=1}^d c_1(j)\left(\theta_{j,1} - \theta_{j,2}\right)^2$, where $c_1(j) = 1/(2\min\{1-\theta_{j,2}, \theta_{j,2}\})$. We repeat this for $D_{\mathrm{KL}}(P_{\theta_2}, P_{\theta_1})$. By the definition of Jeffreys' divergence (8) we obtain $D_{\mathrm{Je}}(P_{\theta_1}, P_{\theta_2}) \le 4\sum_{j=1}^d c_1(j)\left(\theta_{j,1} - \theta_{j,2}\right)^2 + 4\sum_{j=1}^d c_2(j)\left(\theta_{j,1} - \theta_{j,2}\right)^2$, where $c_2(j) = 1/(2\min\{1-\theta_{j,1}, \theta_{j,1}\})$. With

$$c(\theta_1, \theta_2) := \min_{1 \le j \le d} \{\min\{1-\theta_{j,1}, \theta_{j,1}\}, \min\{1-\theta_{j,2}, \theta_{j,2}\}\},$$

we get $D_{\mathrm{Je}}(P_{\theta_1}, P_{\theta_2}) \le \frac{4}{c(\theta_1,\theta_2)} \sum_{j=1}^d \left(\theta_{j,1} - \theta_{j,2}\right)^2$. Since the positive numbers $1-\theta_{j,l}$ and $\theta_{j,l}$ are strictly smaller than 1 for all j and $l = 1, 2$, it holds that

$$c(\theta_1, \theta_2) > \min\left\{\left[\prod_{j=1}^d \min(1-\theta_{j,1}, \theta_{j,1})\right], \left[\prod_{j=1}^d \min(1-\theta_{j,1}, \theta_{j,1})\right]\right\}.$$

Hence we have (75) as asserted. □

An instance of the bound (43) in Proposition 4.2 for $P_{\theta_1} \in \mathbb{M}_{\mathrm{MBe}}$ and $P_{\theta_2} \in \mathbb{M}_{\mathrm{MBe}}$ is

$$D_{\mathrm{JS}}(P_{\theta_1}, P_{\theta_2}) \le \phi^{''}(1) a(\theta_1, \theta_2) V(P_{\theta_1}, P_{\theta_2})^2, \tag{76}$$

where

$$a(\theta, \theta^{(o)}) = \left(\frac{\prod_{j=1}^d \min\{1-\theta_{j,1}, \theta_{j,1}\} + \prod_{j=1}^d \min\{1-\theta_{j,2}, \theta_{j,2}\}}{\prod_{j=1}^d \min\{1-\theta_{j,2}, \theta_{j,2}\} \prod_{j=1}^d \min\{1-\theta_{j,1}, \theta_{j,1}\}} - 1\right).$$

To verify this, we need to compute $\min_{x \in \{0,1\}^d} P_{\theta_1}(x)$ and $\min_{x \in \{0,1\}^d} P_{\theta_2}(x)$. This is found in Appendix A.1.

Observation 7.3. Let us consider an n sample of data denoted by $\mathbf{X} = (X_1, \ldots, X_n)$, where $X_l \in \{0,1\}^d$, $X_l = \left(x_1^{(l)}, \ldots, x_d^{(l)}\right)$, and $x_j^{(l)} \in \{0,1\}$, $j = 1, \ldots, d$. The MLE of θ in $\mathbb{M}_{\mathrm{MBe}}$ is given by $\widehat{\theta}_n = (\widehat{\theta}_{1,n}, \ldots, \widehat{\theta}_{d,n})$, where

$$\widehat{\theta}_{j,n} := \frac{\text{number of } l \text{ s.t. } x_j^{(l)} = 1}{n}.$$

We assume that $\widehat{\theta}_{j,n} > 0$ for every j. Then we set $p_{\widehat{\theta}_{j,n}}\left(a_j^{(i)}\right) = \widehat{\theta}_{j,n}^{[\left(a_j^{(i)}\right)=1]} \cdot (1 - \widehat{\theta}_{j,n})^{[\left(a_j^{(i)}\right)=0]}, a_j^{(i)} \in \{0,1\}$, and $p_i(\widehat{\theta}_n) := \prod_{j=1}^d p_{\widehat{\theta}_{j,n}}\left(a_j^{(i)}\right)$. When $P_{\widehat{\theta},n}(x)$ is defined following (22), we note that $P_{\widehat{\theta},n} \in \mathbb{M}_{\mathrm{MBe}}$.

To get to the point of this observation, we need to complicate the notation by setting $\widehat{\theta}_n(\mathbf{X}) = \widehat{\theta}_n$. For another sample $\mathbf{Y} = (Y_1, \ldots, Y_m)$, $Y_l \in \{0,1\}^d$, we find as above $P_{\widehat{\theta}_m(\mathbf{Y})} \in \mathbb{M}_{\mathrm{MBe}}$. Thus we can analyze $D_{\mathrm{JS}}\left(P_{\widehat{\theta}_n(\mathbf{X})}, P_{\widehat{\theta}_m(\mathbf{Y})}\right)$ by the inequalities above.

Suppose $\mathbf{X}$ and $\mathbf{Y}$ are data from two different data generating sources, or two operative taxonomic units in the terminology of [26] or two different simulator based models. A third sample $\mathbf{Z} = (Z_1, \ldots, Z_r)$, $Z_l \in \{0,1\}^d$ of data is to be analyzed w.r.t. similarity to operative taxonomic units represented by $\mathbf{X}$ and $\mathbf{Y}$, respectively. This comparison or discrimination can be done by finding the smaller of $D_{\mathrm{JS}}\left(P_{\widehat{\theta}_n(\mathbf{X})}, P_{\widehat{\theta}_r(\mathbf{Z})}\right)$ and $D_{\mathrm{JS}}\left(P_{\widehat{\theta}_m(\mathbf{Y})}, P_{\widehat{\theta}_r(\mathbf{Z})}\right)$. By the bounds above, this is graphically performed by expressions of the form $D_{\mathrm{JS}}\left(P_{\widehat{\theta}_n(\mathbf{X})}, P_{\widehat{\theta}_r(\mathbf{Z})}\right) \asymp \phi''(1) \sum_{j=1}^{d} \left(\widehat{\theta}_{j,r}(\mathbf{Z}) - \widehat{\theta}_{j,n}(\mathbf{X})\right)^2$. This concept of discrimination between operative taxonomic units was the rationale for JSD in [26]. Jardine and Sibson argue in [26, pp.13−16] that $D_{\mathrm{JS}}(P,Q)$ is a mathematical model for the notion of D-similarity in biological taxonomy. Actually, in their numerical studies in Appendix 1 of [26] these authors restrict themselves to $D_{\mathrm{JS},1/2}$ of two categorical distributions on a binary $\mathcal{A}$. We may thus also perceive JSD as Jardine-Sibson Divergence.

8. DISCUSSION

In our work we have shown how the JSD can be conveniently decomposed such that useful upper and lower bounds can be derived in explicit terms. The bounds are used to prove consistency and asymptotic equivalence results for the JSD based estimator. These statements provide a foundation for practical applications, where ML estimator would not be available. This observation and the simulation experiments provided and discussed in Section 6 suggest that there is a rich field of additional theoretical and numerical questions related the JSD to be considered from an inferential perspective.

In the simulation example it was observed that the parameter θ tends to be overestimated when all categories are not represented in the observation set, which occurs when the expected counts np_i are small in some categories i. The overestimation bias can be observed as a right-hand tail in the histograms in Figure 3 (c). Hence, even when the assumption of a full support is fulfilled, there may be practical predicaments.

In the simulation example the mapping between model parameters and category probabilities is known so that we are able to calculate P_θ based on θ. This allowed comparison between the maximum likelihood and minimum JSD estimates. However, we emphasise that our main interest and motivation behind the present work are complex simulator models, where the exact dependencies between the model parameters and category probabilities are unknown and maximum likelihood estimation has to be done by a simulator based inference.

9. ACKNOWLEDGMENTS

The authors thank the two reviewers for very careful reading of the submitted paper, for pointing out errors and for several suggestions that greatly improved both presentation and content. J.C. and U.R. are supported by ERC grant 742158 and T.K. is supported by FCAI (=Finnish Center for Artificial Intelligence).

A. APPENDIX: AUXILIARIES ON MULTIVARIATE BERNOULLI DISTRIBUTONS

A.1. The minimum probability

In this Section we find expressions for $\min_{x\in\{0,1\}^d} P_\theta(x)$. Let first $\mathbf{a}^* = (a_1^*, \dots, a_d^*) \in \{0,1\}^d$, where

$$a_j^* := \begin{cases} 1 & 1/2 < \theta_j < 1; \\ 0 & 0 < \theta_j < 1/2, \end{cases} \tag{77}$$

with $\theta_j = 1/2$ being resolved arbitrarily. Let $\overline{\mathbf{a}^*} \in \{0,1\}^d$ be the binary complement of $\mathbf{a}^*$, i. e., it satisfies $\overline{a_j^*} = 1$, if $a_j^* = 0$ and $\overline{a_j^*} = 0$ if $a_j^* = 1$.

Proposition A.1. For every $x \in \{0,1\}^d$ and every $P_\theta \in \mathbb{M}_{\mathrm{MBe}}$ it holds that

$$P_\theta(\overline{\mathbf{a}^*}) \le P_\theta(x) \le P_\theta(\mathbf{a}^*). \tag{78}$$

P r o o f. Let

$$\alpha_j(\theta) := \begin{cases} \frac{\theta_j}{1-\theta_j} & \text{if } 1/2 \le \theta_j < 1; \\ \frac{1-\theta_j}{\theta_j} & \text{if } 0 < \theta_j < 1/2. \end{cases} \tag{79}$$

Then every $P_\theta \in \mathbb{M}_{\mathrm{MBe}}$ can be written as

$$P_\theta(x) = P_\theta(\mathbf{a}^*) \cdot \prod_{j=1}^{d} \alpha_j(\theta)^{-|x_j - a_j^*|}, \tag{80}$$

see [22, pp.222–223]. We cite (4) to get

$$D_{\mathrm{CD}}(p_{\theta_j}, q) = \ln \alpha_j(\theta), \tag{81}$$

and we can up-date the expression in (80) as

$$P_\theta(x) = P_\theta(\mathbf{a}^*) \cdot e^{-\sum_{j=1}^{d} D_{\mathrm{CD}}(p_{\theta_j},q)\cdot|x_j - a_j^*|}. \tag{82}$$

Since $D_{\mathrm{CD}}(p_{\theta_j}, q) \ge 0$, $\sum_{j=1}^{d} D_{\mathrm{CD}}(p_{\theta_j}, q) \cdot |x_j - a_j^*| \ge 0$, we have established the right hand inequality in (78). Next, by (82)

$$P_\theta(\overline{\mathbf{a}^*}) = P_\theta(\mathbf{a}^*) \cdot e^{-\sum_{j=1}^{d} D_{\mathrm{CD}}(p_{\theta_j},q)\cdot|\overline{a_j^*} - a_j^*|} = P_\theta(\mathbf{a}^*) \cdot e^{-\sum_{j=1}^{d} D_{\mathrm{CD}}(p_{\theta_j},q)}, \tag{83}$$

since $|\overline{a_j^*} - a_j^*| = 1$ for all j by definition of the complement. Clearly, $e^{-\sum_{j=1}^{d} D_{\mathrm{CD}}(p_{\theta_j},q)} \le e^{-\sum_{j=1}^{d} D_{\mathrm{CD}}(p_{\theta_j},q)\cdot|x_j - a_j^*|}$ for every $x \in \{0,1\}^d$, and hence we have shown that $P_\theta(\overline{\mathbf{a}^*}) \le P_\theta(x)$ for every $x \in \{0,1\}^d$. □

Next we prove the formula completing the proof of Proposition 7.1 and the inequality (76).

Proposition A.2.

$$\min_{x\in\{0,1\}^d} P_\theta(x) = \prod_{j=1}^{d} \min\{1-\theta_j, \theta_j\}. \tag{84}$$

Proof. From (78), $P_\theta(\overline{\mathbf{a}^*}) = \min_{x\in\{0,1\}^d} P_\theta(x)$. We have in view of (83) and (81) that

$$P_\theta(\overline{\mathbf{a}^*}) = P_\theta(\mathbf{a}^*)\cdot e^{-\sum_{j=1}^d D_{\mathrm{CD}}(p_{\theta_j},q)} = P_\theta(\mathbf{a}^*)\cdot\prod_{j=1}^d \alpha_j(\theta)^{-1}. \tag{85}$$

By (69), (70) and (77) we have $P_\theta(\mathbf{a}^*) = \prod_{j=1}^d \theta_j^{[a_j^*=1]}(1-\theta_j)^{[a_j^*=0]} = \prod_{j=1}^d \max\{1-\theta_j,\theta_j\}$, which entails $P_\theta(\overline{\mathbf{a}^*}) = \prod_{j=1}^d \frac{\max\{1-\theta_j,\theta_j\}}{\alpha_j(\theta)}$. We take a generic factor in this product. If $1/2 < \theta_j < 1$, then $1-\theta_j < 1-1/2 = 1/2$ and $\max\{1-\theta_j,\theta_j\} = \theta_j$. The case $1/2 < \theta_j < 1$ in (79) gives $\alpha_j(\theta) = \frac{\theta_j}{1-\theta_j}$. Hence

$$\frac{\max\{1-\theta_j,\theta_j\}}{\alpha_j(\theta)} = \frac{\theta_j}{\frac{\theta_j}{1-\theta_j}} = 1-\theta_j = \min\{1-\theta_j,\theta_j\}.$$

The case $0 < \theta_j < 1/2$ is handled in the same manner. Finally, in case $\theta_j = 1/2$, $\alpha_j(\theta) = 1$ and $\max\{1-\theta_j,\theta_j\} = \min\{1-\theta_j,\theta_j\}$. Thus the proof is completed. □

A.2. A bound in terms of the Chan–Darwich metric

We define the uniform distribution $Q \in \mathbb{M}_{\mathrm{MBe}}$ by $Q(x) = \frac{1}{2^d}$ for all $x \in \{0,1\}^d$. We recall Chan–Darwich metric in (2).

Proposition A.3. For any $P_\theta \in \mathbb{M}_{\mathrm{MBe}}$ and the uniform distribution Q in $\mathbb{M}_{\mathrm{MBe}}$ we have

$$D_{\mathrm{CD}}(P_\theta, Q) = \sum_{j=1}^d D_{\mathrm{CD}}(p_{\theta_j}, q), \tag{86}$$

where $D_{\mathrm{CD}}(p_{\theta_j}, q)$ is given in (4).

Proof. The first step applies the definition (2) to write $D_{\mathrm{CD}}(P_\theta, Q)$ in terms of $\overline{\mathbf{a}^*}$ and $\mathbf{a}^*$. The second step combines step one with the preceding findings.

Step 1. We prove first that for every $x \in \{0,1\}^d$ by (78) $\ln\frac{P_\theta(x)}{Q(x)} \le d\ln 2 + \ln P_\theta(\mathbf{a}^*)$ and $\ln\frac{P_\theta(x)}{Q(x)} \ge d\ln 2 + \ln P_\theta(\overline{\mathbf{a}^*})$. Hence by the definition in (2) the equality

$$D_{\mathrm{CD}}(P_\theta, Q) = \ln\frac{P_\theta(\mathbf{a}^*)}{P_\theta(\overline{\mathbf{a}^*})} \tag{87}$$

follows as asserted.

Step 2. Next we observe that (85) gives also that $\frac{P_\theta(\mathbf{a}^*)}{P_\theta(\overline{\mathbf{a}^*})} = \prod_{j=1}^d \alpha_j(\theta)$. Hence (81) and (87) establish (86). We note that (86) agrees with (4) for $d = 1$, since then Q equals q of Example 2.3.

□

In view of (87), $P_\theta(\overline{\mathbf{a}^*}) = P_\theta(\mathbf{a}^*)\cdot e^{-D_{\mathrm{CD}}(P_\theta,Q)}$, which can be used in the bound (45) to get an upper bound for $D_{\mathrm{JS}}(P_{\theta_1}, P_{\theta_2})$. The upper bound in (45) is strictly increasing in D_{CD}. We apply the triangle inequality

$$e^{D_{\mathrm{CD}}(P_{\theta_1},P_{\theta_2})} \le e^{D_{\mathrm{CD}}(P_{\theta_1},Q)} e^{D_{\mathrm{CD}}(P_{\theta_2},Q)} = \frac{P_{\theta_1}(\mathbf{a}_1^*)}{P_{\theta_1}(\overline{\mathbf{a}_1^*})}\cdot\frac{P_{\theta_2}(\mathbf{a}_2^*)}{P_{\theta_2}(\overline{\mathbf{a}_2^*})}$$

and from (86) we obtain that $D_{\mathrm{CD}}(P_{\theta_1}, P_{\theta_2}) \le \sum_{j=1}^d D_{\mathrm{CD}}(p_{\theta_{j,1}}, q) + \sum_{j=1}^d D_{\mathrm{CD}}(p_{\theta_{j,2}}, q)$. These two inequalities give obviously an upper bound for $D_{\mathrm{JS}}(P_{\theta_1}, P_{\theta_2})$ by insertion in (45).

(Received December 8, 2020)

REFERENCES

[1] N. S. Barnet and S. Dragomir: A survey of recent inequalities for ϕ-divergences of discrete probability distributions. In: Advances in Inequalities from Probability Theory and Statistics (N. S. Barnett and S. S. Dragomir, eds.), Nova Science Publishing, New York 2008, pp. 1–85. DOI:10.1002/ev.254

[2] M. Basseville: Divergence measures for statistical data processing − An annotated bibliography. Signal Processing *93* (2013), 621–633. DOI:10.1016/j.sigpro.2012.09.003

[3] D. Berend and A. Kontorovich: A sharp estimate of the binomial mean absolute deviation with applications. Stat. Probab. Lett. *83* (2013), 1254–259.

[4] BOLFI Tutorial and Manual: https://elfi.readthedocs.io/en/latest/usage/BOLFI.html, 2017.

[5] U. Böhm, P. F. Dahm, B. F. McAllister, and I. F. Greenbaum: Identifying chromosomal fragile sites from individuals: a multinomial statistical model. Human Genetics *95* (1995), 249–256.

[6] H. Chan and A. Darwiche: A distance measure for bounding probabilistic belief change. Int. J. Approx. Reasoning *38* (2005), 149–174. DOI:10.1016/j.ijar.2004.07.001

[7] H. Chan and A. Darwiche: On the revision of probabilistic beliefs using uncertain evidence. Artif. Intell. *163* (2005), 67–90.

[8] C. D. Charalambous, I. Tzortzis, S. Loyka, and T. Charalambous: Extremum problems with total variation distance and their applications. IEEE Trans. Automat. Control *59* (2014), 2353–2368. DOI:10.1109/TAC.2014.2321951

[9] J. Corander, C. Fraser, M. U. Gutmann, B. Arnold, W. P. Hanage, S. D. Bentley, M. Lipsitch, and N. J. Croucher: Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. Nature Ecology Evolution *1* (2017), 1950-1960. DOI:10.1038/s41559-017-0337-x

[10] Th. M. Cover and J. A. Thomas: Elements of Information Theory. Second edition. John Wiley and Sons, New York 2012.

[11] K. Cranmer, J. Brehmer and G. Louppe: The frontier of simulation-based inference. Proc. Natl. Acad. Sci. USA *117* (2020), 30055–30062. DOI:10.1073/pnas.1912789117

[12] I. Csiszár and Z. Talata: Context tree estimation for not necessarily finite memory processes, via BIC and MDL. IEEE Trans. Inform. Theory *52* (2006), 1007–1016. DOI:10.1109/TIT.2005.864431

[13] I. Csiszár and P. C. Shields: Information Theory and Statistics: A tutorial. Now Publishers Inc, Delft 2004.

[14] L. Devroye: The equivalence of weak, strong and complete convergence in L_1 for kernel density estimates. Ann. Statist. *11* (1983), 896–904.

[15] P. J. Diggle and R. J. Gratton: Monte Carlo methods of inference for implicit statistical models. J. R. Stat. Soc. Ser. B. Stat. Methodol. *46*, (1984), 193–212.

[16] D. M.Endre and J. E. Schindelin: A new metric for probability distributions. IEEE Trans. Inform. Theory *49* (2003), 1858–1860. DOI:10.1109/TIT.2003.813506

[17] A. A. Fedotov, P. Harremoës, and F. Topsøe: Refinements of Pinsker's inequality. IEEE Trans. Inform. Theory *49* (2003), 1491–1498. DOI:10.1109/TIT.2003.811927

[18] A. L. Gibbs and F. E. Su: On choosing and bounding probability metrics. Int. Stat. Rev. *70* (2002), 419–435. DOI:10.1111/j.1751-5823.2002.tb00178.x

[19] A. Guntuboyina: Lower bounds for the minimax risk using f-divergences, and applications. IEEE Trans. Inform. Theory *57* (2011), 2386–2399. DOI:10.1109/TIT.2011.2110791

[20] M. U. Gutmann and J. Corander: Bayesian optimization for likelihood-free inference of simulator-based statistical models. J. Mach. Learn. Res. *17*, (2016), 4256–4302.

[21] M. Gyllenberg, T. Koski, E. Reilink, and M. Verlaan: Non-uniqueness in probabilistic numerical identification of bacteria. J. App. Prob. *31* (1994), 542–548. DOI:10.1017/S0021900200045034

[22] M. Gyllenberg and T. Koski: Numerical taxonomy and the principle of maximum entropy. J. Classification *13* (1996), 213–229. DOI:10.1007/BF01246099

[23] I. Holopainen: Evaluating Uncertainty with Jensen–Shannon Divergence. Master's Thesis, Faculty of Science, University of Helsinki 2021.

[24] C-D. Hou, J. Chiang, and J. J. Tai: Identifying chromosomal fragile sites from a hierarchical-clustering point of view. Biometrics *57* (2001), 435–440. DOI:10.1111/j.0006-341X.2001.00435.x

[25] M. Janžura and P. Boček: A method for knowledge integration. Kybernetika *34* (1998), 41–55.

[26] N. Jardine and R. Sibson: Mathematical Taxonomy. J. Wiley and Sons, London 1971.

[27] M. Khosravifard, D. Fooladivanda, and T. A. Gulliver: Exceptionality of the variational distance. In: 2006 IEEE Information Theory Workshop-ITW'06 Chengdu 2006, pp. 274–276.

[28] T. Koski: Probability Calculus for Data Science. Studentlitteratur, Lund 2020.

[29] V. Kůs: Blended ϕ-divergences with examples. Kybernetika *39* (2003), 43–54.

[30] V. Kůs, D. Morales, and I. Vajda: Extensions of the parametric families of divergences used in statistical inference. Kybernetika *44* (2008), 95–112. DOI:10.1111/j.1399-0004.1993.tb03860.x

[31] L. LeCam: On the assumptions used to prove asymptotic normality of maximum likelihood estimates. Ann. Math. Statist. *41* (1970), 802–828. DOI:10.1214/aoms/1177696960

[32] F. Liese and I. Vajda: On divergences and informations in statistics and information theory. IEEE Trans. Inform. Theory *52* (2006), 4394–4412. DOI:10.1109/TIT.2006.881731

[33] K. Li and J. Mitendra: Implicit maximum likelihood estimation. arXiv preprint arXiv:1809.09087, 2018).

[34] J. Lin: Divergence measures based on the Shannon entropy. IEEE Trans. Inform. Theory *37* (1991), 145–151. DOI:10.1109/18.61115

[35] J. Lintusaari, M. U Gutmann, R. Dutta, S. Kaski, and J. Corander: Fundamentals and recent developments in approximate Bayesian computation. Systematic Biology *66* (2017), e66–e82.

[36] J. Lintusaari, H. Vuollekoski, A. Kangasrääsiö, K. Skytén, M. Järvenpää, P. Marttinen, M. U. Gutmann, A. Vehtari, J. Corander, and S. Kaski: ELFI: Engine for likelihood-free inference. J. Mach. Learn. Res. *19* (2018), 1–7.

[37] D. Morales, L. Pardo, and I. Vajda: Asymptotic divergence of estimates of discrete distributions. J. Statist. Plann. Inference *48* (1995), 347–369. DOI:10.1016/0378-3758(95)00013-Y

[38] S. Nowozin, B. Cseke, and R. Tomioka: f-gan: Training generative neural samplers using variational divergence minimization. Advances Neural Inform. Process. Systems (2016), 271–279.

[39] M. Okamoto: Some inequalities relating to the partial sum of binomial probabilities. Ann. Inst.of Statist. Math. *10* (1959), 29–35. DOI:10.1007/BF02883985

[40] I. Sason: On f-divergences: Integral representations, local behavior, and inequalities. Entropy *20* (2018), 383–405. DOI:10.3390/e20050383

[41] I. Sason and S. Verdu: f-divergence inequalities. IEEE Trans. Inform. Theory *62* (2016), 5973–6006. DOI:10.1109/TIT.2016.2603151

[42] M. Shannon: Properties of f-divergences and f-GAN training. arXiv preprint arXiv:2009.00757, 2020.

[43] R. Sibson: Information radius. Z. Wahrsch. Verw. Geb. *14* (1969), 149–160. DOI:10.1007/BF00537520

[44] M. Sinn and A. Rawat: Non-parametric estimation of Jensen–Shannon divergence in generative adversarial network training. In: International Conference on Artificial Intelligence and Statistics 2018, pp. 642–651.

[45] I. J. Taneja: On mean divergence measures. In: Advances in Inequalities from Probability Theory and Statistics (N. S. Barnett and S. S. Dragomir, eds.), Nova Science Publishing, New York 2008, pp. 169–186.

[46] F. Topsøe: Information-theoretical optimization techniques. Kybernetika *15* (1979), 8–27.

[47] F. Topsøe: Some inequalities for information divergence and related measures of discrimination. IEEE Trans. Inform. Theory *46* (2000), 1602–1609. DOI:10.1109/18.850703

[48] I. Vajda: Note on discrimination information and variation (Corresp.). IEEE Trans. Inform. Theory *16* (1970), 771–773. DOI:10.1109/TIT.1970.1054557

[49] I. Vajda: Theory of Statistical Inference and Information. Kluwer Academic Publ., Delft 1989.

[50] I. Vajda: On metric divergences of probability measures. Kybernetika *45* (2009), 885–900. DOI:10.1145/1932682.1869533

[51] J. I. Yellott Jr.: The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution. J. Math. Psych. *15* (1977), 109–144. DOI:10.1016/0022-2496(77)90026-8

[52] F. Österreicher and I. Vajda: Statistical information and discrimination. IEEE Trans. Inform. Theory *39* (1993), 1036–1039. DOI:10.1109/18.256536

Jukka Corander, Department of Biostatistics, Institute of Basic Medical Sciences, Faculty of Medicine, University in Oslo. Norway.
e-mail: jukka.corander@medisin.uio.no

Ulpu Remes, Department of Biostatistics, Institute of Basic Medical Sciences, Faculty of Medicine, University in Oslo. Norway.
e-mail: umvremes@medisin.uio.no

Timo Koski, Department of Mathematics and Statistics, University of Helsinki and Helsinki Institute of Information Technology HIIT. Finland.
e-mail: timo.jt.koski@helsinki.fi