Discrepancies, dissimilarities, divergences, and distances

Frank Nielsen
Sony Computer Science Laboratories Inc.
Tokyo, Japan

13th August 2021, updated August 16, 2021

This is a working document which will be frequently updated with materials concerning the discrepancy between two distributions.

This document is also available in the PDF `Distance.pdf`

There are many other acronyms used in the literature for referencing a dissimilarity; For example, the following 5 D's: Discrepancies, deviations, dissimilarities, divergences, and distances.

Contents

1 Statistical distances between densities with computationally intractable normalizers

Consider a density $p(x) = \frac{\tilde{p}(x)}{Z_p}$ where $\tilde{p}(x)$ is an unnormalized *computable* density and $Z_p = \int p(x)\mathrm{d}\mu(x)$ the *computationally intractable* normalizer (also called in statistical physics the partition function or free energy). A statistical distance $D[p_1 : p_2]$ between two densities $p_1(x) = \frac{\tilde{p}_1(x)}{Z_{p_1}}$

and $p_2(x) = \frac{\tilde{p}_2(x)}{Z_{p_2}}$ with computationally intractable normalizers Z_{p_1} and Z_{p_2} is said *projective* (or two-sided *homogeneous*) if and only if

$$\forall \lambda_1 > 0, \lambda_2 > 0, \quad D[p_1 : p_2] = D[\lambda_1 p_1 : \lambda_2 p_2].$$

In particular, letting $\lambda_1 = Z_{p_1}$ and $\lambda_2 = Z_{p_2}$, we have

$$D[p_1 : p_2] = D[\tilde{p}_1 : \tilde{p}_2].$$

Notice that the rhs. does not rely on the computationally intractable normalizers. These projective distances are useful in statistical inference based on minimum distance estimators [2] (see next Section).

Here are a few statistical projective distances:

- **γ-divergences** ($\gamma > 0$) [10, 6]:

  $$D_\gamma[p:q] := \log\left(\int_{\mathbb{R}} q^{\alpha+1}\right) - \left(1+\frac{1}{\alpha}\right)\log\left(\int_{\mathbb{R}} q^\alpha p\right) + \frac{1}{\alpha}\log\left(\int_{\mathbb{R}} p^{\alpha+1}\right), \quad \gamma \geq 0$$

  When $\gamma \to 0$, we have [6] $D_\gamma[p:q] = D_{\mathrm{KL}}[p:q]$, the Kullback-Leibler divergence (KLD). For example, we can estimate the KLD between two densities of an exponential-polynomial family by Monte Carlo stochastic integration of the γ-divergence for a small value of γ [27].

  The γ-divergences (projective, Bregman-type=Cross-entropy-entropy) and the density power divergence [1] (non-projective, Bregman-type divergence):

  $$D_\alpha^{\mathrm{dpd}}[p:q] := \int_{\mathbb{R}} q^{\alpha+1} - \left(1+\frac{1}{\alpha}\right)\int_{\mathbb{R}} q^\alpha p + \frac{1}{\alpha}\int_{\mathbb{R}} p^{\alpha+1}, \quad \alpha \geq 0,$$

  can be encapsulated into the family of Φ-power divergences [37] (functional density power divergence class):

  $$D_{\phi,\alpha}[p:q] := \phi\left(\int_{\mathbb{R}} q^{\alpha+1}\right) - \left(1+\frac{1}{\alpha}\right)\phi\left(\int_{\mathbb{R}} q^\alpha p\right) + \frac{1}{\alpha}\phi\left(\int_{\mathbb{R}} p^{\alpha+1}\right), \quad \alpha \geq 0,$$

  where $\phi(e^x)$ convex and strictly increasing, ϕ continuous and twice continously differentiable with finite second order derivatives. We have $D_{\phi,0}[p:q] = \phi'(1)\int_{\mathbb{R}} p(x)\log\frac{p(x)}{q(x)}\mathrm{d}\mu(x) = \phi'(1)D_{\mathrm{KL}}[p:q]$.

- **Cauchy-Schwarz divergence** [9] (CSD, projective)

  $$D_{\mathrm{CS}}[p:q] = -\log\left(\frac{\int p(x)q(x)\mathrm{d}\mu(x)}{\sqrt{\int p(x)^2\mathrm{d}\mu(x)\int q(x)^2\mathrm{d}\mu(x)}}\right) = D_{\mathrm{CS}}[\lambda_1 p : \lambda_2 q], \forall \lambda_1 > 0, \lambda_2 > 0,$$

  and **Hölder divergences** [35] (HD, projective, which generalizes the CSD):

  $$D_{\alpha,\gamma}^{\text{Hölder}}[p:q] = -\log\left(\frac{\int_{\mathcal{X}} p(x)^{\gamma/\alpha} q(x)^{\gamma/\beta}\mathrm{d}x}{\left(\int_{\mathcal{X}} p(x)^\gamma \mathrm{d}x\right)^{1/\alpha}\left(\int_{\mathcal{X}} q(x)^\gamma \mathrm{d}x\right)^{1/\beta}}\right), \quad \frac{1}{\alpha}+\frac{1}{\beta} = 1.$$

We have

$$\forall \lambda_1 > 0, \lambda_2 > 0, D^{\text{Hölder}}_{\alpha,\gamma}[\lambda_1 p : \lambda_2 q] = D^{\text{Hölder}}_{\alpha,\gamma}[p : q],$$

and

$$D^{\text{Hölder}}_{2,2}[p : q] = D_{\text{CS}}[p : q].$$

Hölder divergences between two densities p_{θ_p} and p_{θ_q} of an exponential family with cumulant function $F(\theta)$ is available in closed-form [35]:

$$D^{\text{Hölder}}_{\alpha,\gamma}[p : q] = \frac{1}{\alpha} F(\gamma\theta_p) + \frac{1}{\beta} F(\gamma\theta_q) - F\left(\frac{\gamma}{\alpha}\theta_p + \frac{\gamma}{\beta}\theta_q\right)$$

The CSD is available in closed-form between mixtures of an exponential family with a conic natural parameter [18]: This includes the case of Gaussian mixture models [11].

- **Hilbert distance** [34] (projective): Consider two probability mass functions $p = (p_1, \ldots, p_d)$ and $q = (q_1, \ldots, q_d)$ of the d-dimensional probability simplex. Then the Hilbert distance is

$$D^{\text{Hilbert}}[p : q] = \log\left(\frac{\max_{i\in\{1,\ldots,d\}} \frac{p_i}{q_i}}{\min_{j\in\{1,\ldots,d\}} \frac{p_j}{q_j}}\right).$$

We have

$$\forall \lambda_1 > 0, \lambda_2 > 0, D^{\text{Hilbert}}[\lambda_1 p : \lambda_2 q] = D^{\text{Hilbert}}[p : q].$$

The Hilbert projective simplex distance can be extended to the cone of positive-definite matrices [34] (and its subspace of correlation matrices called the elliptope) as follows:

$$D^{\text{Hilbert}}[P : Q] = \log\left(\frac{\lambda_{\max}(PQ^{-1})}{\lambda_{\min}(PQ^{-1})}\right),$$

where $\lambda_{\max}(X)$ and $\lambda_{\min}(X)$ denote the largest and smallest eigenvalue of matrix X, respectively.

2 Statistical distances between empirical distributions and densities with computationally intractable normalizers

When estimating the parameter $\hat{\theta}$ for a parametric family of distributions $\{p_\theta\}$ from i.i.d. observations $\mathcal{S} = \{x_1, \ldots, x_n\}$, we can define a minimum distance estimator (MDE):

$$\hat{\theta} = \arg\min_\theta D[p_{\mathcal{S}} : p_\theta],$$

where $p_{\mathcal{S}} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ is the empirical distribution (normalized). Thus we need only a right-sided projective divergence to estimate models with computationally intractable normalizers. For example, the Maximum Likelihood Estimator (MLE) is a MDE wrt. the KLD:

$$\hat{\theta}_{\text{MLE}} = \arg\min_\theta D_{\text{KL}}[p_{\mathcal{S}} : p_\theta].$$

It is thus interesting to study the impact of the choice of the distance D to the properties of the corresponding estimator (e.g., γ-divergences yields provably robust estimators [6]).

- **Hyvärinen divergence** [7] (also called **Fisher divergence**):

$$D^{\text{Hyvärinen}}[p : p_\theta] := \frac{1}{2}\int \|\nabla_x \log p(x) - \nabla_x \log p_\theta(x)\|^2\ p(x)\mathrm{d}x.$$

  The Hyvarinen divergence has been extended for order-α Hyvarinen divergences [22] (for $\alpha > 0$):

$$D_\alpha^{\text{Hyvärinen}}[p : q] := \frac{1}{2}\int p(x)^\alpha \left(\nabla_x \log p(x) - \nabla_x \log q(x)\right)^2 \mathrm{d}x, \quad \alpha > 0.$$

3 The Jensen-Shannon divergence and some generalizations

3.1 Origins of the Jensen-Shannon divergence

Let $(\mathcal{X}, \mathcal{F}, \mu)$ be a measure space, and $(w_1, P_1), \ldots, (w_n, P_n)$ be n weighted probability measures dominated by a measure μ (with $w_i > 0$ and $\sum w_i = 1$). Denote by $\mathcal{P} := \{(w_1, p_1), \ldots, (w_n, p_n)\}$ the set of their weighted Radon-Nikodym densities $p_i = \frac{\mathrm{d}P_i}{\mathrm{d}\mu}$ with respect to μ.

A *statistical divergence* $D[p : q]$ is a measure of dissimilarity between two densities p and q (i.e., a 2-point distance) such that $D[p : q] \geq 0$ with equality if and only if $p(x) = q(x)$ μ-almost everywhere. A *statistical diversity index* $D(\mathcal{P})$ is a measure of variation of the weighted densities in $\mathcal{P}$ related to a measure of centrality, i.e., a n-point distance which generalizes the notion of 2-point distance when $\mathcal{P}_2(p, q) := \{(\frac{1}{2}, p_1), (\frac{1}{2}, p_2)\}$:

$$D[p : q] := D(\mathcal{P}_2(p, q)).$$

The fundamental measure of dissimilarity in information theory is the *I-divergence* (also called the *Kullback-Leibler divergence*, KLD, see Equation (2.5) page 5 of [12]):

$$D_{\mathrm{KL}}[p : q] := \int_{\mathcal{X}} p(x) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}\mu(x).$$

The KLD is asymmetric (hence the delimiter notation ":" instead of ',') but can be symmetrized by defining the Jeffreys *J-divergence* (Jeffreys divergence, denoted by I_2 in Equation (1) in 1946's paper [8]):

$$D_J[p, q] := D_{\mathrm{KL}}[p : q] + D_{\mathrm{KL}}[q : p] = \int_{\mathcal{X}} (p(x) - q(x)) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}\mu(x).$$

Although symmetric, any positive power of Jeffreys divergence fails to satisfy the triangle inequality: That is, D_J^α is never a metric distance for any $\alpha > 0$, and furthermore D_J^α cannot be upper bounded.

In 1991, Lin proposed the asymmetric *K-divergence* (Equation (3.2) in [14]):

$$D_K[p : q] := D_{\mathrm{KL}}\left[p : \frac{p+q}{2}\right],$$

and defined the *L-divergence* by analogy to Jeffreys's symmetrization of the KLD (Equation (3.4) in [14]):

$$D_L[p, q] = D_K[p : q] + D_K[q : p].$$

By noticing that

$$D_L[p,q] = 2h\left[\frac{p+q}{2}\right] - (h[p]+h[q]),$$

where h denotes Shannon entropy (Equation (3.14) in [14]), Lin coined the (skewed) *Jensen-Shannon divergence* between two weighted densities $(1-\alpha,p)$ and (α,q) for $\alpha\in(0,1)$ as follows (Equation (4.1) in [14]):

$$D_{\mathrm{JS},\alpha}[p,q] = h[(1-\alpha)p+\alpha q] - (1-\alpha)h[p] - \alpha h[q]. \tag{1}$$

Finally, Lin defined the *generalized Jensen-Shannon divergence* (Equation (5.1) in [14]) for a finite weighted set of densities:

$$D_{\mathrm{JS}}[\mathcal{P}] = h\left[\sum_i w_i p_i\right] - \sum_i w_i h[p_i].$$

This generalized Jensen-Shannon divergence is nowadays called the *Jensen-Shannon diversity index*.

To contrast with the Jeffreys' divergence, the Jensen-Shannon divergence (JSD) $D_{\mathrm{JS}} := D_{\mathrm{JS},\frac{1}{2}}$ is upper bounded by $\log 2$ (does not require the densities to have the same support), and $\sqrt{D_{\mathrm{JS}}}$ is a metric distance [4, 5]. Lin cited precursor work [42, 15] yielding definition of the Jensen-Shannon divergence: The Jensen-Shannon divergence of Eq. equation1 is the so-called "increments of entropy" defined in (19) and (20) of [42].

The Jensen-Shannon diversity index was also obtained very differently by Sibson in 1969 when he defined the *information radius* [40] of order α using Rényi α-means and Rényi α-entropies [38]. In particular, the information radius IR_1 of order 1 of a weighted set $\mathcal{P}$ of densities is a diversity index obtained by solving the following variational optimization problem:

$$\mathrm{IR}_1[\mathcal{P}] := \min_c \sum_{i=1}^n w_i D_{\mathrm{KL}}[p_i : c]. \tag{2}$$

Sibson solved a more general optimization problem, and obtained the following expression (term K_1 in Corollary 2.3 [40]):

$$\mathrm{IR}_1[\mathcal{P}] = h\left[\sum_i w_i p_i\right] - \sum_i w_i h[p_i] := D_{\mathrm{JS}}[\mathcal{P}].$$

Thus Eq. equation2 is a variational definition of the Jensen-Shannon divergence.

3.2 Some extensions of the Jensen-Shannon divergence

- **Skewing the JSD.**

  The K-divergence of Lin can be skewed with a scalar parameter $\alpha\in(0,1)$ to give

  $$D_{K,\alpha}[p:q] := D_{\mathrm{KL}}\left[p : (1-\alpha)p+\alpha q\right]. \tag{3}$$

  Skewing parameter α was first studied in [13] (2001, see Table 2 of [13]). We proposed to unify the Jeffreys divergence with the Jensen-Shannon divergence as follows (Equation 19 in [17]):

  $$D^J_{K,\alpha}[p:q] := \frac{D_{K,\alpha}[p:q] + D_{K,\alpha}[q:p]}{2}. \tag{4}$$

When $\alpha = \frac{1}{2}$, we have $D^J_{K,\frac{1}{2}} = D_{\mathrm{JS}}$, and when $\alpha = 1$, we get $D^J_{K,1} = \frac{1}{2} D_J$.

Notice that

$$D^{\alpha,\beta}_{\mathrm{JS}}[p;q] := (1-\beta) D_{\mathrm{KL}}[p : (1-\alpha)p + \alpha q] + \beta D_{\mathrm{KL}}[q : (1-\alpha)p + \alpha q]$$

amounts to calculate

$$h^{\times}[(1-\beta)p + \beta q : (1-\alpha)p + \alpha q] - ((1-\beta)h[p] + \beta h[q])$$

where

$$h^{\times}[p,q] := \int -p(x)\log q(x) \mathrm{d}\mu(x)$$

denotes the *cross-entropy*. By choosing $\alpha = \beta$, we have $h^{\times}[(1-\beta)p + \beta q : (1-\alpha)p + \alpha q] = h[(1-\alpha)p + \alpha q]$, and thus recover the skewed Jensen-Shannon divergence of Eq. equation1.

In [21] (2020), we considered a positive *skewing vector* $\alpha \in [0,1]^k$ and a unit positive weight w belonging to the standard simplex Δ_k, and defined the following *vector-skewed Jensen-Shannon divergence*:

$$\begin{aligned} D^{\alpha,w}_{\mathrm{JS}}[p:q] &:= \sum_{i=1}^{k} D_{\mathrm{KL}}[(1-\alpha_i)p + \alpha_i q : (1-\bar{\alpha})p + \bar{\alpha} q], && (5) \\ &= h[(1-\bar{\alpha})p + \bar{\alpha} q] - \sum_{i=1}^{k} h[(1-\alpha_i)p + \alpha_i q], && (6) \end{aligned}$$

where $\bar{\alpha} = \sum_{i=1}^{k} w_i \alpha_i$. The divergence $D^{\alpha,w}_{\mathrm{JS}}$ generalizes the (scalar) skew Jensen-Shannon divergence when $k = 1$, and is a Ali-Silvey-Csiszár f-divergence upper bounded by $\log \frac{1}{\bar{\alpha}(1-\bar{\alpha})}$ [21].

- **A priori mid-density**. The JSD can be interpreted as the total divergence of the densities to the *mid-density* $\bar{p} = \sum_{i=1}^{n} w_i p_i$, a statistical mixture:

$$D_{\mathrm{JS}}[\mathcal{P}] = \sum_{i=1}^{n} w_i D_{\mathrm{KL}}[p_i : \bar{p}] = h[\bar{p}] - \sum_{i=1}^{n} w_i h[p_i].$$

Unfortunately, the JSD between two Gaussian densities is not known in closed form because of the definite integral of a log-sum term (i.e., K-divergence between a density and a mixture density $\bar{p}$). For the special case of the Cauchy family, a closed-form formula [29] for the JSD between two Cauchy densities was obtained. Thus we may choose a *geometric mixture distribution* [19] instead of the ordinary arithmetic mixture $\bar{p}$. More generally, we can choose any weighted mean M_α (say, the geometric mean, or the harmonic mean, or any other power mean) and define a generalization of the K-divergence of Equation equation3:

$$D^{M_\alpha}_K[p:q] := D_K[p : (pq)_{M_\alpha}], \tag{7}$$

where

$$(pq)_{M_\alpha}(x) := \frac{M_\alpha(p(x), q(x))}{Z_{M_\alpha}(p:q)}$$

is a statistical M-mixture with $Z_{M_\alpha}(p,q)$ denoting the normalizing coefficient:

$$Z_{M_\alpha}(p:q) = \int M_\alpha(p(x), q(x))\mathrm{d}\mu(x)$$

so that $\int (pq)_{M_\alpha}(x)\mathrm{d}\mu(x) = 1$. These M-mixtures are well-defined provided the convergence of the definite integrals.

Then we define a generalization of the JSD [19] termed (M_α, N_β)*-Jensen-Shannon divergence* as follows:

$$D_{\mathrm{JS}}^{M_\alpha,N_\beta}[p:q] := N_\beta\left(D_K[p:(pq)_{M_\alpha}], D_K[q:(pq)_{M_\alpha}]\right), \tag{8}$$

where N_β is yet another weighted mean to average the two M_α-K-divergences. We have $D_{\mathrm{JS}} = D_{\mathrm{JS}}^{A,A}$ where $A(a,b) = \frac{a+b}{2}$ is the arithmetic mean. The geometric JSD yields a closed-form formula between two multivariate Gaussians, and has been used in deep learning [3]. More generally, we may consider the Jensen-Shannon symmetrization of an arbitrary distance D as

$$D^{\mathrm{JS}}_{M_\alpha,N_\beta}[p:q] := N_\beta\left(D[p:(pq)_{M_\alpha}], D[q:(pq)_{M_\alpha}]\right). \tag{9}$$

- **A posteriori mid-density**. We consider a generalization of Sibson's information radius [40]. Let $S_w(a_1,\ldots,a_n)$ denote a generic weighted mean of n positive scalars $a_1,\ldots,a_n$, with weight vector $w \in \Delta_n$. Then we define the S*-variational Jensen-Shannon diversity index* [24] as

  $$D_{\mathrm{vJS}}^{S_w}(\mathcal{P}) := \min_c S_w\left(D_{\mathrm{KL}}[p_1:c], D_{\mathrm{KL}}[p_n:c]\right). \tag{10}$$

  When $S_w = A_w$ (with $A_w(a_1,\ldots,a_n) = \sum_{i=1}^n w_i a_i$ the arithmetic weighted mean), we recover the ordinary Jensen-Shannon diversity index. More generally, we define the S*-Jensen-Shannon index of an arbitrary distance* D as

  $$D^{\mathrm{vJS}}_{S_w}(\mathcal{P}) := \min_c S_w\left(D[p_1:c],\ldots,D[p_n:c]\right). \tag{11}$$

  When $n=2$, this yields a Jensen-Shannon-symmetrization of distance D.

  The variational optimization defining the JSD can also be constrained to a (parametric) family of densities $\mathcal{D}$, thus defining the $(S,\mathcal{D})$*-relative Jensen-Shannon diversity index*:

  $$D_{\mathrm{vJS}}^{S_w,\mathcal{D}}(\mathcal{P}) := \min_{c\in\mathcal{D}} S_w\left(D_{\mathrm{KL}}[p_1:c],\ldots,D_{\mathrm{KL}}[p_n:c]\right). \tag{12}$$

  The relative Jensen-Shannon divergences are useful for clustering applications: Let p_{θ_1} and p_{θ_2} be two densities of an exponential family $\mathcal{E}$ with cumulant function $F(\theta)$. Then the $\mathcal{E}$-relative Jensen-Shannon divergence is the Bregman information of $\mathcal{P}_2(p,q)$ for the conjugate function $F^*(\eta) = -h[p_\theta]$ (with $\eta = \nabla F(\theta)$). The $\mathcal{E}$-relative JSD amounts to a *Jensen divergence* for F^*:

$$
\begin{aligned}
D_{\mathrm{vJS}}[p_{\theta_1}, p_{\theta_2}] &= \min_\theta \frac{1}{2}\left\{D_{\mathrm{KL}}[p_{\theta_1} : p_\theta] + D_{\mathrm{KL}}[p_{\theta_2} : p_\theta]\right\}, & (13)\\
&= \min_\theta \frac{1}{2}\left\{B_F[\theta : \theta_1] + B_F[\theta : \theta_2]\right\}, & (14)\\
&= \min_\eta \frac{1}{2}\left\{B_{F^*}[\eta_1 : \eta] + B_{F^*}[\eta_2 : \eta]\right\}, & (15)\\
&= \frac{F^*(\eta_1)+F^*(\eta_2)}{2} - F^*(\eta^*), & (16)\\
&=: J_{F^*}(\eta_1, \eta_2), & (17)
\end{aligned}
$$

since $\eta^* := \frac{\eta_1+\eta_2}{2}$ (a right-sided *Bregman centroid* [26]).

4 Statistical distances between mixtures

Pearson [36] first considered a unimodal Gaussian mixture of two components for modeling distributions crabs in 1894. Statistical mixtures [16] like the Gaussian mixture models (GMMs) are often met in information sciences, and therefore it is important to assess their dissimilarities. Let $m(x) = \sum_{i=1}^k w_i p_i(x)$ and $m'(x) = \sum_{i=1}^{k'} w_i' p_i'(x)$ be two finite statistical mixtures. The KLD between two GMMs m and m' is not analytic [41] because of the log-sum terms:

$$
D_{\mathrm{KL}}[m : m'] = \int m(x) \log \frac{m(x)}{m'(x)} \mathrm{d}x.
$$

However, the KLD between two GMMs with the same prescribed components $p_i(x) = p_i'(x) = p_{\mu_i,\Sigma_i}(x)$ (i.e., $k = k'$, and only the normalized positive weights may differ) is provably a Bregman divergence [28] for the differential negentropy $F(\theta)$:

$$
D_{\mathrm{KL}}[m(\theta) : m(\theta')] = B_F(\theta, \theta'),
$$

where $m(\theta) = \sum_{i=1}^{k-1} w_i p_i(x) + (1 - \sum_{i=1}^{k-1} w_i) p_k(x)$ and $F(\theta) = \int m(\theta) \log m(\theta) \mathrm{d}x$. The family $\{m_\theta \quad \theta \in \Delta_{k-1}^\circ\}$ is called a mixture family in information geometry, where Δ_{k-1}° denotes the $(k-1)$-dimensional open standard simplex. However, $F(\theta)$ is usually not available in closed-form because of the log-sum integral. In some special cases like the mixture of two prescribed Cauchy distributions, we get a closed-form formula for the KLD, JSD, etc. [29, 25]. Thus when dealing with mixtures (like GMMs), we either need efficient approximating (§subsection4.1), bounding (§subsection4.2) KLD techniques, or new distances (§subsection4.3) that yields closed-form formula between mixture densities.

4.1 Approximating and/or fast statistical distances between mixtures

- The Jeffreys divergence (JD) $D_J[m, m'] = D_{\mathrm{KL}}[m : m'] + D_{\mathrm{KL}}[m' : m]$ between two (Gaussian) MMs is not available in closed-form, and can be estimated using Monte Carlo integration as

$$
\hat{D}_J^{\mathcal{S}_s}[m, m'] := \frac{1}{s} \sum_{i=1}^s 2 \frac{(m(x_i) - m'(x_i))}{m(x_i) + m'(x_i)} \log\left(\frac{m(x_i)}{m'(x_i)}\right),
$$

where $\mathcal{S}_s = \{x_1, \ldots, x_s\}$ are s IID samples from the mid mixture $m_{12}(x) := \frac{1}{2}(m(x) + m'(x))$ (with $\lim_{s\to\infty} \hat{D}_J^{\mathcal{S}_s}[m, m'] = D_J[m, m']$). In [23], the mixtures m and m' are converted into densities of an exponential-polynomial family. The JD between densities p_θ and $p_{\theta'}$ of an exponential family with cumulant function F is available in closed-form:

$$D_J[p_\theta, p_{\theta'}] = (\theta' - \theta) \cdot (\eta' - \eta),$$

with $\eta = \nabla F(\theta)$ and $\theta = \nabla F^*(\eta)$, where F^* denotes the convex conjugate. Any smooth density r (includes a mixture $r = m$) is converted into close densities $p_{\theta_r^{\mathrm{MLE}}}$ and $p^{\eta_r^{\mathrm{SME}}}$ of a exponential-polynomial family using extensions of the Maximum Likelihood Estimator (MLE) and Score Matching Estimator (SME). Then JD between mixtures is approximated as follows

$$D_J[m, m'] \simeq ({\theta'}^{\mathrm{SME}} - \theta^{\mathrm{SME}}) \cdot ({\eta'}^{\mathrm{MLE}} - \eta^{\mathrm{MLE}}).$$

- Given a finite set of mixtures $\{m_i(x)\}$ sharing the same components (e.g., points on a mixture family manifold), we precompute the KLD between pairwise components to obtain fast approximation of the KLD $D_{\mathrm{KL}}[m_i : m_j]$ between any two mixtures m_i and m_j, see [39].

4.2 Bounding statistical distances between mixtures

- **Log-Sum-Exp bounds**: In [31, 32], we lower and upper bound the cross-entropy between mixtures using the fact that the log-sum term $\log m(x)$ and be interpreted as a LSE function. We then compute lower envelopes and upper envelopes of density functions using technique of computational geometry to report deterministic lower and upper bounds on the KLD and α-divergences. These bounds are said combinatorial because we decompose the support into elementary intervals. Bounds between the Total Variation Distance (TVD) between univariate mixtures are reported in [33].

4.3 Newly designed statistical distances yielding closed-form formula for mixtures

- **Statistical Minkowski distances** [20]: Consider the Lebesgue space

$$L_\alpha(\mu) := \left\{ f \in \mathbb{F} \ : \ \int_{\mathcal{X}} |f(x)|^\alpha \mathrm{d}\mu(x) < \infty \right\}$$

for $\alpha \geq 1$, where $\mathbb{F}$ denotes the set of all real-valued measurable functions defined on the support $\mathcal{X}$. Minkowski's inequality writes as $\|p + q\|_\alpha \leq \|p\|_\alpha + \|q\|_\alpha$ for $\alpha \in [1, \infty)$. The statistical Minkowski difference distance between $p, q \in L_\alpha(\mu)$ is defined as

$$D_\alpha^{\mathrm{Minkowski}}[p, q] := \|p\|_\alpha + \|q\|_\alpha - \|p + q\|_\alpha \geq 0. \quad (18)$$

The statistical Minkowski log-ratio distance is defined by:

$$L_\alpha^{\mathrm{Minkowski}}[p, q] := -\log \frac{\|p + q\|_\alpha}{\|p\|_\alpha + \|q\|_\alpha} \geq 0. \quad (19)$$

These statistical Minkowski distances are symmetric, and $L_\alpha[p, q]$ is scale-invariant. For even integers $\alpha \geq 2$, $D_\alpha^{\mathrm{Minkowski}}[m : m']$ is available in closed-form.

- We show that the Cauchy-Schwarz divergence (CSD), the quadratic Jensen-Rényi divergence [**?**] (JRD), and the total square Distance (TSD) between two GMMs, and more generally two mixtures of exponential families, can be obtained in closed-form in [18].

Initially created 13th August 2021 (last updated August 16, 2021).

References

[1] Ayanendranath Basu, Ian R Harris, Nils L Hjort, and MC Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998.

[2] Ayanendranath Basu, Hiroyuki Shioya, and Chanseok Park. *Statistical inference: the minimum distance approach*. Chapman and Hall/CRC, 2019.

[3] Jacob Deasy, Nikola Simidjievski, and Pietro Liò. Constraining Variational Inference with Geometric Jensen-Shannon Divergence. In *Advances in Neural Information Processing Systems*, 2020.

[4] Dominik Maria Endres and Johannes E Schindelin. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860, 2003.

[5] Bent Fuglede and Flemming Topsoe. Jensen-Shannon divergence and Hilbert space embedding. In *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, page 31. IEEE, 2004.

[6] Hironori Fujisawa and Shinto Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9):2053–2081, 2008.

[7] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

[8] Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

[9] Robert Jenssen, Jose C Principe, Deniz Erdogmus, and Torbjørn Eltoft. The Cauchy–Schwarz divergence and Parzen windowing: Connections to graph theory and Mercer kernels. *Journal of the Franklin Institute*, 343(6):614–629, 2006.

[10] MC Jones, Nils Lid Hjort, Ian R Harris, and Ayanendranath Basu. A comparison of related density-based minimum divergence estimators. *Biometrika*, 88(3):865–873, 2001.

[11] Kittipat Kampa, Erion Hasanbelliu, and Jose C Principe. Closed-form Cauchy-Schwarz PDF divergence for mixture of Gaussians. In *The 2011 International Joint Conference on Neural Networks*, pages 2578–2585. IEEE, 2011.

[12] Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

[13] Lillian Lee. On the effectiveness of the skew divergence for statistical language analysis. In *Artificial Intelligence and Statistics (AISTATS)*, page 65?72, 2001.

[14] Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[15] Jianhua Lin and SKM Wong. Approximation of discrete probability distributions based on a new divergence measure. *Congressus Numerantium (Winnipeg)*, 61:75–80, 1988.

[16] Geoffrey J McLachlan and Kaye E Basford. *Mixture models: Inference and applications to clustering*, volume 38. M. Dekker New York, 1988.

[17] Frank Nielsen. A family of statistical symmetric divergences based on Jensen's inequality. *arXiv preprint arXiv:1009.4004*, 2010.

[18] Frank Nielsen. Closed-form information-theoretic divergences for statistical mixtures. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR)*, pages 1723–1726. IEEE, 2012.

[19] Frank Nielsen. On the Jensen?Shannon Symmetrization of Distances Relying on Abstract Means. *Entropy*, 21(5), 2019.

[20] Frank Nielsen. The statistical Minkowski distances: Closed-form formula for Gaussian mixture models. In *International Conference on Geometric Science of Information*, pages 359–367. Springer, 2019.

[21] Frank Nielsen. On a Generalization of the Jensen?Shannon Divergence and the Jensen?Shannon Centroid. *Entropy*, 22(2), 2020.

[22] Frank Nielsen. Fast approximations of the Jeffreys divergence between univariate Gaussian mixture models via exponential polynomial densities. *arXiv preprint arXiv:2107.05901*, 2021.

[23] Frank Nielsen. Fast approximations of the jeffreys divergence between univariate gaussian mixture models via exponential polynomial densities. *arXiv preprint arXiv:2107.05901*, 2021.

[24] Frank Nielsen. On a Variational Definition for the Jensen-Shannon Symmetrization of Distances Based on the Information Radius. *Entropy*, 23(4), 2021.

[25] Frank Nielsen. The dually flat information geometry of the mixture family of two prescribed Cauchy components. *arXiv preprint arXiv:2104.13801*, 2021.

[26] Frank Nielsen and Richard Nock. Sided and symmetrized Bregman centroids. *IEEE transactions on Information Theory*, 55(6):2882–2904, 2009.

[27] Frank Nielsen and Richard Nock. Patch matching with polynomial exponential families and projective divergences. In *International Conference on Similarity Search and Applications*, pages 109–116. Springer, 2016.

[28] Frank Nielsen and Richard Nock. On the geometry of mixtures of prescribed distributions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2861–2865. IEEE, 2018.

[29] Frank Nielsen and Kazuki Okamura. On f-divergences between cauchy distributions. *arXiv:2101.12459*, 2021.

[30] Frank Nielsen and Kazuki Okamura. On f-divergences between Cauchy distributions. *arXiv preprint arXiv:2101.12459*, 2021.

[31] Frank Nielsen and Ke Sun. Guaranteed bounds on information-theoretic measures of univariate mixtures using piecewise log-sum-exp inequalities. *Entropy*, 18(12):442, 2016.

[32] Frank Nielsen and Ke Sun. Combinatorial bounds on the α-divergence of univariate mixture models. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 4476–4480. IEEE, 2017.

[33] Frank Nielsen and Ke Sun. Guaranteed Deterministic Bounds on the total variation distance between univariate mixtures. In *28th IEEE International Workshop on Machine Learning for Signal Processing, MLSP 2018, Aalborg, Denmark, September 17-20, 2018*, pages 1–6. IEEE, 2018.

[34] Frank Nielsen and Ke Sun. Clustering in Hilbert's projective geometry: The case studies of the probability simplex and the elliptope of correlation matrices. In *Geometric Structures of Information*, pages 297–331. Springer, 2019.

[35] Frank Nielsen, Ke Sun, and Stéphane Marchand-Maillet. On Hölder projective divergences. *Entropy*, 19(3):122, 2017.

[36] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

[37] Souvik Ray, Subrata Pal, Sumit Kumar Kar, and Ayanendranath Basu. Characterizing the functional density power divergence class. *arXiv preprint arXiv:2105.06094*, 2021.

[38] Alfréd Rényi et al. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

[39] Olivier Schwander, Stéphane Marchand-Maillet, and Frank Nielsen. Comix: Joint estimation and lightspeed comparison of mixture models. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, pages 2449–2453. IEEE, 2016.

[40] Robin Sibson. Information radius. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 14(2):149–160, 1969.

[41] Sumio Watanabe, Keisuke Yamazaki, and Miki Aoyagi. Kullback information of normal mixture is not an analytic function. *IEICE technical report. Neurocomputing*, 104(225):41–46, 2004.

[42] Andrew KC Wong and Manlai You. Entropy and distance of random graphs with application to structural pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (5):599–609, 1985.