ON THE GEOMETRY OF MIXTURES OF PRESCRIBED DISTRIBUTIONS

Frank Nielsen

École Polytechnique, France
Sony Computer Science Laboratories, Japan
Frank.Nielsen@acm.org

Richard Nock

Data 61, Australia
The Australian National & Sydney Universities
Richard.Nock@data61.csiro.au

ABSTRACT

We consider the space of w-mixtures that are finite statistical mixtures sharing the same prescribed component distributions, like Gaussian mixture models sharing the same components. The information geometry induced by the Kullback-Leibler (KL) divergence yields a dually flat space where the KL divergence between two w-mixtures amounts to a Bregman divergence for the negative Shannon entropy generator, called the Shannon information. Furthermore, we prove that the skew Jensen-Shannon statistical divergence between w-mixtures amount to skew Jensen divergences on their parameters and state several divergence inequalities between w-mixtures and their closures.

1. INTRODUCTION AND BACKGROUND

Let $M_+^1(\Omega)$ denote the space of probability measures defined on a σ-algebra Ω of an observation space $\mathcal{X}$. Consider a *base measure* $\mu \in M_+^1(\Omega)$ (usually the Lebesgue or counting measure), and let $P_0, \ldots, P_{k-1}$ be k *prescribed* probability distributions, all dominated by μ ($P_i \ll \mu$), with $p_i = \frac{\mathrm{d}P_i}{\mathrm{d}\mu}$ the Radon-Nikodym derivative of P_i with respect to μ. The density $m(x;w) \in M_+^1(\Omega)$ of a *w-mixture* is defined by $m(x;w) := \sum_{i=0}^{k-1} w_i p_i(x)$, with $w := (w_0, \ldots, w_{k-1}) \in \Delta_{k-1}^\circ$, where Δ_{k-1}° is the $(k-1)$-dimensional open probability simplex sitting in $\mathbb{R}^k$. Thus w-mixtures are strictly convex weighted combinations of *fixed* component distributions: They form *special subfamilies* of finite statistical mixtures [1] that are closed by convex combinations.

Given multiple datasets $\mathcal{O}_1, \ldots, \mathcal{O}_n$, a set of w-mixtures $m_1 = m(x;w_1), \ldots, m_n = m(x, w_n)$ (called comixs [2]) can be learned *simultaneously* by generalizing the Expectation-Maximization (EM) or the Classification EM (CEM) algorithms. In particular, one can learn w-Gausian Mixture Models [2] (w-GMMs) where the prescribed mixture components are fixed Gaussian distributions.

The class of statistical f-divergences [3, 4, 5] between two distributions $p, q \ll \mu$ defined on support $\mathcal{X}$ is defined by

$$I_f(p:q) := \int_{\mathcal{X}} p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}\mu(x) \geq f(1), \qquad (1)$$

with f a convex function satisfying $f(1) = 0$. We have [6] $I_f(p:q) \leq \lim_{\epsilon \to 0} f(\epsilon) + \epsilon f(\frac{1}{\epsilon})$. For discrete distributions with probability mass functions $p := (p_0, \ldots, p_{d-1})$ and $q := (q_0, \ldots, q_{d-1})$, it comes that $I_f(p:q) = \sum_{i=0}^{d-1} p_i f(\frac{q_i}{p_i})$. The f-divergences include the KL divergence ($f(u) = -\log u$), the χ^2-divergence, the Hellinger divergence, the α-divergences, the total variation $\mathrm{TV}(p,q) := \frac{1}{2}\int_{\mathcal{X}} |p(x) - q(x)| \mathrm{d}\mu(x)$ (with $f(u) = \frac{1}{2}|1-u|$, the only f-divergence metric [7] satisfying the triangle inequality), etc. The dual divergence $I_f^*(p:q) := I_f(q:p)$ is obtained by taking the dual generator $f^\diamond(u) := uf\left(\frac{1}{u}\right)$: $I_{f^\diamond}(p:q) = I_f(q:p) = I_f^*(p:q)$. Thus f-divergences can always be symmetrized [8] by taking the generator $s(u) = f(u) + f^\diamond(u)$. Examples of symmetric f-divergences are the Jeffreys divergence [9] $J(p;q) := \mathrm{KL}(p:q) + \mathrm{KL}(q:p)$ and the Jensen-Shannon divergence [10] $\mathrm{JS}(p:q) := K(p:q) + K(q:p)$ with $K(p:q) := \mathrm{KL}(p:\frac{p+q}{2}) = \int p(x) \log \frac{2p(x)}{p(x)+q(x)} \mathrm{d}\mu(x)$. Depending on the generator f, the f-divergence may be either (1) unbounded when the integral diverges: $I_f(p:q) := +\infty$ (e.g., KL between a standard Cauchy distribution and a standard normal distributions), or (2) always bounded (e.g., Jensen-Shannon divergence bounded by $\log 2$).

The f-divergences between statistical mixtures [11, 12] is not available in closed form although it can be easily upper bounded by using the joint convexity property of f-divergences: $I_f(m:m') \leq \sum_{i,j} w_i w'_j I_f(p_i : p'_j)$ for two mixture models $m(x) = \sum_i w_i p_i(x)$ and $m'(x) = \sum_j w'_j p'_j(x)$. In practice, to bypass this intractability, one *estimates* the f-divergence using Monte Carlo (MC) stochastic integration [13] (Chapter 17): Let s iid. samples $x_1, \ldots, x_s \sim p(x)$, and define the estimator $\hat{I}_f^s(p:q) := \frac{1}{s}\sum_{i=1}^s f\left(\frac{q(x_i)}{p(x_i)}\right)$. It follows from the Law of Large Numbers (LLN) that $\lim_{s\to\infty} \hat{I}_f^s(p:q) = I_f(p:q)$ provided that the variance $\mathrm{Var}_p\left[f\left(\frac{q(x)}{p(x)}\right)\right]$ is bounded. The MC estimator is consistent (but the MC approximation does not hold when $I_f(p:q) = \infty$). Furthermore, using the Central Limit Theorem (CLT), the MC estimator is shown to be *normally* distributed: $\hat{I}_f^s(p:q) \sim \mathcal{N}\left(I_f(p:q), \frac{1}{s}\mathrm{Var}_p\left[f\left(\frac{q(x)}{p(x)}\right)\right]\right)$.

1.1. Contributions

In §2, we describe the dually flat geometry of the space of w-mixtures induced by the Kullback-Leibler (KL) divergence. It proves that the KL divergence between any two w-mixtures is equivalent to a Bregman divergence induced by the negative Shannon entropy generator. As a byproduct, this allows us to prove that the KL-averaging integration of w-mixtures used in distributed estimation [14] can be performed optimally without information loss. In §2.2, we show that the skew Jensen-Shannon divergences between w-mixtures amount to an equivalent skew Jensen α-divergences on their parameters. Finally, we consider several divergence inequalities between w-mixtures and their closures in §3.2.

2. GEOMETRY OF W-MIXTURES

We slightly depart from the constructions sketched in the textbooks [15, 9], in order to ease sanity checks.

When the k prescribed component distributions $p_0(x), \ldots, p_{k-1}(x)$ are *linearly independent*, the space $\mathcal{M} = \{m(x;w)\ ,\ w \in \Delta_{k-1}^{\circ}\}$ of w-mixtures forms a *mixture family* in information geometry [9, 16] with:

$$m(x;w) = m(x;\eta) = \sum_{i=1}^{k-1} \eta_i p_i(x) + \left(1 - \sum_{i=1}^{k-1} \eta_i\right) p_0(x), \tag{2}$$

with $\eta_i = w_i$ for $i \in [k-1] := \{1, \ldots, k-1\}$ and $w_0 = 1 - \sum_{i=1}^{k-1} \eta_i = 1 - \sum_{i=1}^{k-1} w_i$. Let $D = k-1$ denote the *order* of the mixture family, that is its number of degrees of freedom. We have $m(x;w) = m(x;\eta)$, where vector w is k-dimensional while vector η is $(k-1)$-dimensional. Let $f_i(x) = p_i(x) - p_0(x)$ for $i \in [D]$, and $c(x) = p_0(x)$. Then $\mathcal{M}$ can be written in the *canonical form* of a *mixture family* in information geometry [9]: $\mathcal{M} = \left\{ m(x;\eta) = \sum_{i=1}^{k-1} \eta_i f_i(x) + c(x), \quad \eta \in \Delta_D^{\circ} \right\}$, where the $f_i(x)$'s and $c(x)$ are linearly independent. By convention, we define $\eta_0 = 1 - \sum_{i=1}^{D} \eta_i$, the weight of p_0. Beware that η_0 is *not* a vector component of $\eta = (\eta_1, \ldots, \eta_D) \in \Delta_D^{\circ}$, the $D = (k-1)$-dimensional *open* probability simplex sitting in $\mathbb{R}^d$.

We consider $\mathcal{M}$ as a smooth manifold of η-mixtures. The Shannon differential entropy [17] of a mixture $m(x)$:

$$h(m) := -\int_{\mathcal{X}} m(x) \log m(x) \mathrm{d}\mu(x) \tag{3}$$

is usually not available in closed-form [11, 12] because of the log-sum term. Both lower and upper bounds on the entropy of mixtures are reported in [11, 18]. For η-mixtures, the parametric function $E(\eta) = -h(m(x;\eta))$, is strictly convex and differentiable. Thus we can form a *dually flat manifold* [15, 9] where the Kullback-Leibler divergence between two mixtures $m(x;\eta_1)$ and $m(x;\eta_2)$ amounts to calculate a *Bregman divergence* [19] $B_{F^*}(\eta_1 : \eta_2)$ for the *negative Shannon information generator* shifted by one [9]:

$$\begin{aligned} F^*(\eta) &= \int (m(x;\eta) \log m(x;\eta) - m(x;\eta)) \mathrm{d}\mu(x), && (4) \\ &= \int m(x;\eta) \log m(x;\eta) \mathrm{d}\mu(x) - 1. && (5) \end{aligned}$$

Since the Shannon entropy is strictly concave, the negative Shannon entropy called *Shannon information* [20] is strictly convex (and a dually flat manifold can be built from *any* C^3 convex function [5]). Let $m_1(x) = m(x;\eta_1)$ and $m_2(x) = m(x;\eta_2)$ for short. We have

$$\begin{aligned} \mathrm{KL}(m_1 : m_2) &= \int m(x;\eta_1) \log \frac{m(x;\eta_1)}{m(x;\eta_2)} \mathrm{d}\mu(x), \\ &= F^*(\eta_1) - F^*(\eta_2) - \langle \eta_1 - \eta_2, \nabla F^*(\eta_2) \rangle, \\ &= B_{F^*}(\eta_1 : \eta_2), \end{aligned}$$

where $\langle x, y \rangle = x^\top y$ denotes the scalar product of $\mathbb{R}^D$. Although the Shannon information of a w-mixture is a convex function of η, it is not available in closed-form [21, 22]. The η parameter is traditionally called the *"expectation" parameter* in information geometry (although this stems from a property of the exponential family manifolds [9]). The dual parameters $\theta = (\theta^1, \ldots, \theta^D)$, called the *natural parameters*, are defined by

$$\theta^i(\eta) = (\nabla_\eta F^*(\eta))_i = \int (p_i(x) - p_0(x)) \log m(x;\eta) \mathrm{d}\mu(x), \tag{6}$$

since $(\nabla_\eta m(x;\eta))_i = p_i(x) - p_0(x)$ and swapping $\nabla \int = \int \nabla$ (under regularity condition of Leibniz integral rule). The extra constant $-1 = -\int m(x;\eta)\mathrm{d}\mu(x)$ term in Eq. 4 is added to get a nice expression of the θ^i's in Eq. 6. The dual Legendre convex conjugate [23] $F(\theta)$ of $F^*(\eta)$ defined by the Legendre-Fenchel transform $F(\theta) = \sup_\theta \{\langle \theta, \eta \rangle - F^*(\eta)\}$ is

$$\begin{aligned} F(\theta) &= -\int (p_0(x) \log m(x;\eta) - m(x;\eta)) \mathrm{d}\mu(x), && (7) \\ &= -\int p_0(x) \log m(x;\eta) \mathrm{d}\mu(x) + 1 && (8) \end{aligned}$$

Function $F(\theta)$ is convex with respect to θ, and the gradients of the convex conjugates are reciprocal, allowing one to convert *theoretically* from one coordinate system into the dual one: $\eta = \nabla F(\theta)$ and $\theta = \nabla F^*(\eta)$. However, since neither F or F^* are available in closed forms (except for the multinomial family that are w-mixtures with prescribed Dirac component distributions), those conversions are computationally intractable. It follows that the KL divergence between two η-mixture distributions of $\mathcal{M}$ can be equivalently written as

$$
\begin{aligned}
\mathrm{KL}(m_1 : m_2) &= \int m(x;\eta_1)\log\frac{m(x;\eta_1)}{m(x;\eta_2)}\mathrm{d}\mu(x), \\
&= B_{F^*}(\eta_1 : \eta_2) = B_F(\theta_2 : \theta_1), \qquad (9) \\
&= D_{F^*,F}(\eta_1 : \theta_2) = D_{F,F^*}(\theta_2 : \eta_1),
\end{aligned}
$$

where $D_{F^*,F}(\eta_1 : \theta_2) = F^*(\eta_1) + F(\theta_2) - \langle \eta_1, \theta_2\rangle$ denotes the *canonical divergence* [9] in dually flat spaces written using the mixed θ/η-coordinate systems.

Theorem 1 (KL of w-mixtures as a Bregman divergence)**.** *The Kullback-Leibler divergence between two η-mixtures (or w-mixtures) is equivalent to a Bregman divergence defined for the convex Shannon information generator on the η-parameters.*

The information geometry of $(\mathcal{M}, \mathrm{KL})$ is said *dually flat* [9] because the dual Christoffel symbol coefficients Γ_{ijk} and Γ^*_{ijk} have all their coefficients equal to zero [24]. Thus geodesics (autoparallel curves) are visualized as straight Euclidean lines in either the η- or the θ-affine coordinate systems.

Corollary 1 (KL of w-GMMS as a Bregman divergence)**.** *The KL between Gaussian Mixture Models sharing the same components (w-GMM [2]) is equivalent (theoretically) to a Bregman divergence.*

2.1. Application: Optimal KL-averaging integration

Let us consider a *computer cluster* [26] of m machines $M_1, \ldots, M_m$ with the independently and identically sampled data-set $\mathcal{O}$ partitioned into m pieces: $\mathcal{O}_1, \ldots, \mathcal{O}_m$ with $|\mathcal{O}_i| = n_i$. Dataset $\mathcal{O}_i$ is stored *locally* in the memory of machine M_i. Liu and Ihler [14] proposed (1) to estimate the m models locally (say, via Maximum Likelihood Estimators, MLEs, $\hat{\eta}_i$'s on the local samples $\mathcal{O}_i$), and then (2) to merge/aggregate those local model estimates on a *central node* by performing *KL-averaging integration*. When the models all belong to the same exponential family (e.g., Gaussian models), they showed that the *KL-averaging* model integration yields no information loss: For exponential families with log-density $t(x)^\top\theta - F(\theta)$ (with θ the natural parameters, sufficient statistics $t(x)$ and $F(\theta)$ the log-normalizer), the KL-averaging integration [14] yields $\hat{\theta}^{\mathrm{KL}} = \nabla F^{-1}\left(\frac{1}{m}\sum_{i=1}^m \nabla F(\hat{\theta}_i)\right)$ without information loss (with MLE $\hat{\eta}_i = \nabla F(\hat{\theta}_i) = \frac{1}{n_i}\sum_{x\in\mathcal{O}_i} t(x)$). Notice that it requires to manipulate explicitly both the log-normalizer $F(\theta)$ and its inverse gradient function ∇F^{-1}, see [14]. Interestingly, they also report experiments on GMMs [14] that are not exponential families with information loss.

However, for η-mixtures (mixture families), the KL-averaging integration [14, 27] is defined by the following optimization problem:

$$
\begin{aligned}
\hat{\eta}^{\mathrm{KL}} &= \arg\min_{\eta} \sum_{i=1}^m \mathrm{KL}(m(x;\hat{\eta}_i) : m(x;\eta)), \qquad (10) \\
&= \arg\min_{\eta} \sum_{i=1}^m B_{F^*}(\hat{\eta}_i : \eta). \qquad (11)
\end{aligned}
$$

Since the *right-sided Bregman centroid* [28] is always the center of mass *whatever* the chosen Bregman generator[1], we end up with the *optimal integration* (best parameter) for η-mixtures: $\hat{\eta}^{\mathrm{KL}} = \frac{1}{m}\sum_{i=1}^m \hat{\eta}_i$ (or equivalently, $\hat{w}^{\mathrm{KL}} = \frac{1}{m}\sum_{i=1}^m \hat{w}_i$).

Theorem 2 (Optimal KL-averaging integration)**.** *The KL-averaging integration of w-mixtures can be performed optimally without information loss.*

Note that the local model estimators of mixtures may not be consistent nor efficient. In fact, the global Maximum Likelihood (ML) optimization requires to tackle an untractable log-sum maximization for mixtures, and the exact MLE solution for these mixtures maybe transcendental [29]. (In a separate report, we study how w-mixtures can be inferred efficiently.)

2.2. Skew Jensen-Shannon divergences of w-mixtures

Let the skew α-Jensen-Shannon divergence be defined by

$$
\mathrm{JS}_\alpha(p : q) := (1-\alpha)\mathrm{KL}(p : m_\alpha) + \alpha\mathrm{KL}(q : m_\alpha),
$$

for $\alpha \in [0,1]$, and $m_\alpha = (1-\alpha)p + \alpha q$. Define the α-Jensen divergences [30, 31] by $J_{F^*,\alpha}(\eta_1 : \eta_2) := (1-\alpha)F^*(\eta_1) + \alpha F^*(\eta_2) - F^*((1-\alpha)\eta_1 + \alpha\eta_2)$, for the Shannon information $F^*(\eta) = -h(m(x;\eta))$. We have in the limit cases [30, 31] for $m_1(x) = m(x;\eta_1)$ and $m_2(x) = m(x;\eta_2)$:

$$
\lim_{\alpha\to 1^-} \frac{J_{F^*,\alpha}(\eta_1 : \eta_2)}{\alpha(1-\alpha)} = B_{F^*}(\eta_1 : \eta_2) = \mathrm{KL}(m_1 : m_2)
$$
$$
\lim_{\alpha\to 0^+} \frac{J_{F^*,\alpha}(\eta_1 : \eta_2)}{\alpha(1-\alpha)} = B_{F^*}(\eta_2 : \eta_1) = \mathrm{KL}(m_2 : m_1)
$$

Since the combination of w-mixtures is a w-mixture, $m_\alpha(x) := (1-\alpha)m(x;\eta_1) + \alpha m(x;\eta_2) = m(x;\eta_\alpha = (1-\alpha)\eta_1 + \alpha\eta_2)$, plugging Shannon entropy h, we get $J_{F^*,\alpha}(\eta_1 : \eta_2) = h(m_\alpha) - (1-\alpha)h(m_1) - \alpha h(m_2)$. Therefore we rewrite

$$
J_{F^*,\alpha}(\eta_1 : \eta_2) = \int \left((1-\alpha)m_1(x)\log\frac{m_1(x)}{m_\alpha(x)} + \alpha m_2(x)\log\frac{m_2(x)}{m_\alpha(x)} \right)\mathrm{d}\mu(x)
$$

and get $J_{F^*,\alpha}(\eta_1 : \eta_2) = (1-\alpha)\mathrm{KL}(m_1 : m_\alpha) + \alpha\mathrm{KL}(m_2 : m_\alpha) = \mathrm{JS}_\alpha(m_1 : m_2)$. In particular, when $\alpha = \frac{1}{2}$,

[1]Here, it is specially interesting since F^* (the negative entropy) is not available in closed form, and we bypass its use.

$J_{F^*,\frac{1}{2}}(\eta_1 : \eta_2) = \frac{1}{2}\mathrm{JS}(m_1 : m_2)$ is the Jensen-Shannon divergence [10], and when $\alpha \to 1$, $\frac{1}{1-\alpha} J_{F^*,\alpha}(\eta_1 : \eta_2) = \mathrm{KL}(m_1 : m_2)$.

Theorem 3. *The α-Jensen-Shannon statistical divergences between η-mixtures amount to α-Jensen divergences between their corresponding η-mixture parameters:* $\mathrm{JS}_\alpha(m(x;\eta_1) : m(x;\eta_2)) = J_{F^*,\alpha}(\eta_1 : \eta_2)$.

3. ON CLOSURES AND DIVERGENCES

3.1. Divergence inequalities for w-mixtures

Theorem 4 (Upper bound on f-divergences of w-mixtures)**.** *The f-divergence $I_f(m(x;w) : m(x;w'))$ between any two w-mixtures is upper bounded by $I_f(w : w') = \sum_{i=0}^{k-1} w_i f(\frac{w_i'}{w_i})$.*

Proof. We use a generalization of the log-sum inequality for any convex function f (see [32], p. 448): For two finite positive number sequences $A = \{a_i\}_{i=0}^{k-1}$ and $B = \{b_i\}_{i=0}^{k-1}$, we have $\sum_i a_i f\left(\frac{b_i}{a_i}\right) \geq af\left(\frac{b}{a}\right)$. It follows that $m(x;w) f\left(\frac{m(x;w')}{m(x;w)}\right) \leq \sum_{i=0}^{k-1} w_i p_i(x) f\left(\frac{w_i' p_i(x)}{w_i p_i(x)}\right) = \sum_{i=0}^{k-1} w_i f\left(\frac{w_i'}{w_i}\right) p_i(x)$ Carrying out integration on the support $\mathcal{X}$, we get $I_f(m(x;w) : m(x;w')) \leq I_f(w : w')$ since $\int_\mathcal{X} p_i(x)\mathrm{d}\mu(x) = 1$. Recall that the KL divergence is a f-divergence obtained for the generator $f(u) = -\log u$. ☐

3.2. Closures of w-mixtures

The manifold $\mathcal{M}$ of w-mixtures is parameterized by the *open* probability simplex Δ_{k-1}°. When topologically closing the manifold $\mathcal{M}$, we consider $\bar{\Delta}_{k-1}$. Take a l-face of the $(d-1)$-dimensional simplex Δ_{k-1}°. When $l > 0$, the sub-simplex $\sigma \in \bar{\Delta}_{k-1}$ is a l-dimensional simplex, and σ° parameterizes a w-mixture family of order $l > 0$. In the extreme case, we consider order-1 w-mixture induced by a simplex edge $\sigma_1 \in \Delta_{k-1}^\circ$ with extremity component distributions p and q. Define $m^\epsilon(p,q) = (1-\epsilon)p + \epsilon q = p + \epsilon(q-p) = m^{1-\epsilon}(q:p)$ for $\epsilon \in [0,1]$. In the limit cases, the w-mixtures m^ϵ yields (with $w \in \Delta_1^\circ$): $\lim_{\epsilon\to 0} m^\epsilon(p,q) = \lim_{\epsilon\to 1} m^\epsilon(q,p) = p$ and $\lim_{\epsilon\to 1} m^\epsilon(p,q) = \lim_{\epsilon\to 0} m^\epsilon(q,p) = q$. Let $I_f^\epsilon(p:q){:=}I_f(m^\epsilon(p,q) : m^\epsilon(q,p))$. How "far" is $I_f^\epsilon(p:q)$ from its closure $I_f(p:q)$?

On one hand, we have the following theorem:

Theorem 5 (Total variation continuity)**.** *We have the following identity* $\mathrm{TV}^\epsilon(p,q) = |1-2\epsilon|\mathrm{TV}(p,q)$ *(since* $m^\epsilon(p,q) - m^\epsilon(q,p) = (1-2\epsilon)(p-q)$*) that yields* $\lim_{\epsilon\to 0}\mathrm{TV}^\epsilon(p,q) = \lim_{\epsilon\to 1}\mathrm{TV}^\epsilon(p,q) = \mathrm{TV}(p,q)$.

On the other hand, $\mathrm{KL}^\epsilon(p:q){:=}\mathrm{KL}(m^\epsilon(p,q) : m^\epsilon(q,p))$ has been shown to amount to a (univariate) Bregman divergence. That is, $\mathrm{KL}^\epsilon(p:q) = B_{F^*}(\epsilon : 1-\epsilon)$ for 1D generator $F^*(\eta) = \int_\mathcal{X}(p(x) + \eta(q(x)-p(x)))\log(p(x) + \eta(q(x)-p(x)))\mathrm{d}\mu(x)$. By using the fact that the Bregman divergence is the tail of a first-order Taylor expansion [9], we get using Lagrange exact remainder: $\mathrm{KL}^\epsilon(p:q) = \frac{1}{2}(1-2\epsilon)^2(F^*)''(\eta)$, for $\eta \in [\epsilon, 1-\epsilon]$ (assuming $\epsilon \leq \frac{1}{2}$). However, the KL between p and q may potentially be infinite so that in general $\forall\epsilon \neq 0, \mathrm{KL}^\epsilon(p:q) \neq \mathrm{KL}(p:q)$ (Bregman divergences are always finite). Using the *joint convexity* of the KL divergence, we can show that

$$\mathrm{KL}^\epsilon(p:q) \leq \mathrm{KL}(p:q) + \epsilon^2 J(p;q), \tag{12}$$

where J denotes the Jeffreys divergence.

Let us relate the f-divergence between the 1D η-mixture and its extremities (closure) as follows:

Theorem 6 (f-divergence inequalities)**.** *We have*

$$I_f^\epsilon(p:q) \leq (1-\epsilon)I_f(p:q) + \epsilon I_f(q:p), \tag{13}$$

$$I_f^\epsilon(p:q) \leq (1-\epsilon)f\left(\frac{\epsilon}{1-\epsilon}\right) + \epsilon f\left(\frac{1-\epsilon}{\epsilon}\right). \tag{14}$$

When I_f is symmetric ($f = f^\diamond$), $I_f^\epsilon(p:q) \leq I_f(p:q)$. That is, mixing distributions decrease symmetrized f-divergences.

Proof. Apply the convex-sum inequality on $A{:=}\{(1-\epsilon)p(x), \epsilon q(x)\}$ and $B{:=}\{(1-\epsilon)q(x), \epsilon p(x)\}$, so that $a = m^\epsilon(p,q)$ and $b = m^\epsilon(q,p)$. First, let $a_0{:=}(1-\epsilon)p(x)$, $b_0{:=}(1-\epsilon)q(x)$, and $a_1{:=}\epsilon q(x)$ and $b_1{:=}\epsilon p(x)$. We get Ineq. 13. Second, let $a_0{:=}(1-\epsilon)p(x)$, $b_0{:=}\epsilon p(x)$, and $a_1{:=}\epsilon q(x)$ and $b_1{:=}(1-\epsilon)q(x)$. We get Ineq. 14. Note that when $\epsilon \to 0$, the second right-hand-side inequality yields $f(0) + 0f(\infty)$, similar to $I_f \leq f(0) + \frac{f(\infty)}{\infty}$ of [6]. ☐

Supplementary material at
`https://FrankNielsen.github.io/wmixture/`

4. REFERENCES

[1] G. McLachlan and D. Peel, *Finite Mixture Models*, Wiley series in probability and statistics: Applied probability and statistics. Wiley, 2004.

[2] Olivier Schwander, Stéphane Marchand-Maillet, and Frank Nielsen, "Comix: Joint estimation and lightspeed comparison of mixture models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 2449–2453.

[3] Imre Csiszár, "Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten," *Magyar. Tud. Akad. Mat. Kutató Int. Közl*, vol. 8, pp. 85–108, 1963.

[4] Syed Mumtaz Ali and Samuel D Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 131–142, 1966.

[5] Frank Nielsen, "What is... an information projection," *Notices of the AMS*, vol. 65, no. 3, pp. 321–324, 2018.

[6] F. Liese and I. Vajda, *Convex Statistical Distances*, Teubner, Leipzig, 1987.

[7] Mohammadali Khosravifard, Dariush Fooladivanda, and T Aaron Gulliver, "Confliction of the convexity and metric properties in f-divergences," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 90, no. 9, pp. 1848–1853, 2007.

[8] Frank Nielsen, "A family of statistical symmetric divergences based on Jensen's inequality," *arxiv:1009.4004*, 2010.

[9] Shun-ichi Amari, *Information Geometry and Its Applications*, vol. 194, Springer, 2016.

[10] Jianhua Lin, "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information theory*, vol. 37, no. 1, pp. 145–151, 1991.

[11] Frank Nielsen and Ke Sun, "Guaranteed bounds on the Kullback-Leibler divergence of univariate mixtures," *IEEE Signal Processing Letters*, vol. 23, no. 11, pp. 1543–154, 2016.

[12] Frank Nielsen and Ke Sun, "Combinatorial bounds on the α-divergence of univariate mixture models," *IEEE ICASSP*, pp. 4476-4480, 2017.

[13] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning*, MIT Press, 2016, `http://www.deeplearningbook.org`.

[14] Qiang Liu and Alexander T Ihler, "Distributed estimation, information loss and exponential families," in *Advances in Neural Information Processing Systems*, 2014, pp. 1098–1106.

[15] Ovidiu Calin and Constantin Udriste, *Geometric modeling in probability and statistics*, Springer, 2014.

[16] Frank Nielsen and Richard Nock, "On w-mixtures: Finite convex combinations of prescribed component distributions," *CoRR*, vol. abs/1708.00568, 2017.

[17] Thomas M Cover and Joy A Thomas, *Elements of information theory*, John Wiley & Sons, 2012.

[18] Artemy Kolchinsky and Brendan D. Tracey, "Estimating mixture entropy with pairwise distances," *Entropy*, vol. 19, no. 7, 2017.

[19] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh, "Clustering with Bregman divergences," *Journal of machine learning research*, vol. 6, no. Oct, pp. 1705–1749, 2005.

[20] Frank Nielsen, "The dual geometry of Shannon information and its applications ," `https://www.youtube.com/watch?v=aGxZoKSk6CQ`, 2016.

[21] Sumio Watanabe, Keisuke Yamazaki, and Miki Aoyagi, "Kullback information of normal mixture is not an analytic function," *Technical report of IEICE (in Japanese)*, pp. 41–46, 2004.

[22] Kamyar Moshksar and Amir K. Khandani, "Arbitrarily tight bounds on differential entropy of gaussian mixtures," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3340–3354, June 2016.

[23] Frank Nielsen, "Cramér-Rao lower bound and information geometry," In *Connected at Infinity II*, pp. 18-37, Hindustan Book Agency, 2013.

[24] Shun-ichi Amari and Andrzej Cichocki, "Information geometry of divergence functions," *Bulletin of the Polish Academy of Sciences: Technical Sciences*, vol. 58, no. 1, pp. 183–195, 2010.

[25] Richard Nock and Frank Nielsen, "Fitting the smallest enclosing Bregman ball," in *ECML*. Springer, 2005, pp. 649–656.

[26] Frank Nielsen, "Introduction to HPC with MPI for Data Science," Springer, 2016.

[27] Shun-ichi Amari, "Integration of stochastic models by minimizing α-divergence," *Neural computation*, vol. 19, no. 10, pp. 2780–2796, 2007.

[28] Frank Nielsen and Richard Nock, "Sided and symmetrized Bregman centroids," *IEEE transactions on Information Theory*, vol. 55, no. 6, pp. 2882–2904, 2009.

[29] Carlos Améndola, Mathias Drton, and Bernd Sturmfels, "Maximum likelihood estimates for gaussian mixtures are transcendental," in *International Conference on Mathematical Aspects of Computer and Information Sciences*. Springer, 2015, pp. 579–590.

[30] Jun Zhang, "Divergence function, duality, and convex analysis," *Neural Computation*, vol. 16, no. 1, pp. 159–195, 2004.

[31] Frank Nielsen and Sylvain Boltz, "The Burbea-Rao and Bhattacharyya centroids," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5455–5466, 2011.

[32] Imre Csiszár and Paul C Shields, "Information theory and statistics: A tutorial," *Foundations and Trends® in Communications and Information Theory*, vol. 1, no. 4, pp. 417–528, 2004.