Monte Carlo Information-Geometric Structures

Frank Nielsen and Gaëtan Hadjeres

Abstract Exponential families and mixture families are parametric probability models that can be geometrically studied as smooth statistical manifolds with respect to any statistical divergence like the Kullback–Leibler (KL) divergence or the Hellinger divergence. When equipping a statistical manifold with the KL divergence, the induced manifold structure is dually flat, and the KL divergence between distributions amounts to an equivalent Bregman divergence on their corresponding parameters. In practice, the corresponding Bregman generators of mixture/exponential families require to perform definite integral calculus that can either be too time-consuming (for exponentially large discrete support case) or even do not admit closed-form formula (for continuous support case). In these cases, the dually flat construction remains theoretical and cannot be used by information-geometric algorithms. To bypass this problem, we consider performing stochastic Monte Carlo (MC) estimation of those integral-based mixture/exponential family Bregman generators. We show that, under natural assumptions, these MC generators are almost surely Bregman generators. We define a series of dually flat information geometries, termed Monte Carlo Information Geometries, that increasingly-finely approximate the untractable geometry. The advantage of this MCIG is that it allows a practical use of the Bregman algorithmic toolbox on a wide range of probability distribution families. We demonstrate our approach with a clustering task on a mixture family manifold. We then show how to generate MCIG for arbitrary separable statistical divergence between distributions belonging to a same parametric family of distributions.

F. Nielsen (✉)
Sony Computer Science Laboratories, Tokyo, Japan
e-mail: Frank.Nielsen@acm.org

G. Hadjeres
Sony Computer Science Laboratory, Paris, France
e-mail: Gaetan.Hadjeres@sony.com

F. Nielsen (ed.), *Geometric Structures of Information*, Signals and Communication Technology, https://doi.org/10.1007/978-3-030-02520-5_5

Keywords Computational information geometry · Statistical manifold
Dually flat information geometry · Bregman generator
Stochastic Monte Carlo integration · Mixture family · Exponential family
Clustering

1 Introduction

We concisely describe the construction and properties of dually flat spaces [1, 8] in Sect. 1.1, define the statistical manifolds of exponential families and mixture families in Sect. 1.2, and discuss about the computational tractability of Bregman algorithms in dually flat spaces in Sect. 1.3.

1.1 Dually Flat Space: Bregman Geometry

A smooth (potentially asymmetric) distance $D(\cdot, \cdot)$ is called a *divergence* in information geometry [1, 8], and induces a differential-geometric dualistic structure [1, 2, 8, 17]. In particular, a strictly convex and twice continuously differentiable D-dimensional real-valued function F, termed a *Bregman generator*, induces a dually connection-flat structure via a corresponding Bregman Divergence (BD) [4] $B_F(\cdot, \cdot)$ given by:

$$B_F(\theta_1 : \theta_2) := F(\theta_1) - F(\theta_2) - \langle \theta_1 - \theta_2, \nabla F(\theta_2) \rangle, \tag{1}$$

where $\langle y, x \rangle := y^\top x$ denotes the inner product, and $\nabla F(\theta) := (\partial_i F(\theta))_i$ denotes the gradient vector of partial first-order derivatives with respect to vector parameter θ. We use the standard notational convention of information geometry [1, 8]: $\partial_i :=: \frac{\partial}{\partial \theta^i}$ to indicate[1] a contravariant vector [18] $\theta = (\theta^i)_i$.

The Legendre–Fenchel transformation [30] :

$$F^*(\eta) = \sup_\theta \{\langle \theta, \eta \rangle - F(\theta)\}, \tag{2}$$

is at the heart of the duality of flat structures by defining two global affine coordinate systems: The *primal affine θ-coordinate system* and the *dual affine η-coordinate system*, so that any point P of the manifold $\mathcal{M}$ can either be accessed by its *primal* $\theta(P)$ coordinates or equivalently by its *dual* $\eta(P)$ coordinates. We can switch between these two dual coordinates as follows:

[1] The :=: symbol means it is a notational convention equality, like $\sum_{i=1}^k x_i :=: x_1 + \ldots x_k$. It differs from $a := b$ which denotes the symbol of a quantity equality by definition.

$$\eta = \eta(\theta) = \nabla F(\theta) = (\partial_i F(\theta))_i, \tag{3}$$
$$\theta = \theta(\eta) = \nabla F^*(\eta) = (\partial^i F^*(\eta))_i, \tag{4}$$

with reciprocal gradients $\nabla F^* := (\nabla F)^{-1}$. We used the notational convention $\partial^i := \frac{\partial}{\partial \eta_i}$ which indicates the covariant vector [18] $\eta = (\eta_i)_i$.

The metric tensor g of the dually flat structure $(\mathcal{M}, F)$ can either be expressed using the θ- or η-coordinates using the Hessians of the potential functions [53]:

$$G(\theta) = \nabla^2 F(\theta), \tag{5}$$
$$G^*(\eta) = \nabla^2 F^*(\eta). \tag{6}$$

It defines a smooth bilinear form $\langle v, v' \rangle_g$ on $\mathcal{M}$ so that for two vectors v, w of a tangent plane T_P:

$$\langle v, v' \rangle_g = \theta(v)^\top G(\theta)\theta(w), \tag{7}$$
$$= \eta(v)^\top G^*(\eta)\eta(w), \tag{8}$$

where $\theta(v) = (v^i)_i$ and $\eta(v) = (v_i)_i$ denote the contravariant coefficients and covariant coefficients of a vector v, respectively. That is, any vector $v \in T_P$ can be written either as $v = \sum_i v^i e_i$ or as $\sum_i v_i e^{*i}$, where $\{e_i\}_i$ and $\{e^{*i}\}_i$ are the dual basis [18] of the vector space structure of T_P.

Matrices $G(\theta)$ and $G^*(\eta)$ are symmetric positive definite (SPD, denoted by $G(\theta) \succ 0$ and $G^*(\eta) \succ 0$), and they satisfy the Crouzeix identity [13]:

$$G(\theta)G^*(\eta) = I, \tag{9}$$

where I stands for the $D \times D$ identity matrix. This indicates that at each tangent plane T_P, the dual coordinate systems are biorthogonal [57] (with $\{e_i\}_i$ and $\{e^{*i}\}_i$ forming a dual basis [18] of the vector space structure of T_P):

$$\langle e_i, e^{*j} \rangle = \delta_i^j, \tag{10}$$

with δ_i^j the Krönecker symbol: $\delta_i^j = 1$ if and only if (iff) $i = j$, and 0 otherwise. We have:

$$\frac{\partial \eta_i}{\partial \theta^j} = g_{ij}(\theta) = \langle e_i, e_j \rangle, \tag{11}$$
$$\frac{\partial \theta^i}{\partial \eta_j} = g^{ij}(\eta) = \langle e^{*i}, e^{*j} \rangle. \tag{12}$$

The convex conjugate functions $F(\theta)$ and $F^*(\eta)$ are called *dual potential functions*, and define the global metric [53].

Table 1 Overview of the dually differential-geometric structure $(\mathcal{M}, F)$ induced by a Bregman generator F. Notice that if F and ∇F^* are available in closed-form then so are ∇F and F^*

Manifold $(\mathcal{M}, F)$	Primal structure	Dual structure
Affine coordinate system	$\theta(\cdot)$	$\eta(\cdot)$
Conversion $\theta \leftrightarrow \eta$	$\theta(\eta) = \nabla F^*(\eta)$	$\eta(\theta) = \nabla F(\theta)$
Potential function	$F(\theta) = \langle \theta, \nabla F(\theta) \rangle - F^*(\nabla F(\theta))$	$F^*(\eta) = \langle \eta, \nabla F^*(\eta) \rangle - F(\nabla F^*(\eta))$
Metric tensor g	$G(\theta) = \nabla^2 F(\theta)$	$G^*(\eta) = \nabla^2 F^*(\eta)$
	$g_{ij} = \partial_i \partial_j F(\theta)$	$g^{ij} = \partial^i \partial^j F^*(\eta)$
Geodesic $(\lambda \in [0, 1])$	$\gamma(P, Q) := \{(PQ)_\lambda = (1-\lambda)\theta(P) + \lambda\theta(Q)\}_\lambda$	$\gamma^*(P, Q) := \{(PQ)^*_\lambda = (1-\lambda)\eta(P) + \lambda\eta(Q)\}_\lambda$

Table 1 summarizes the differential-geometric structures of dually flat spaces. Since Bregman divergences are *canonical divergences*[2] of dually flat spaces [1], the geometry of dually flat spaces is also referred to the *Bregman geometry* [15] in the literature.

Definition 1 (*Bregman generator*) A Bregman generator is a strictly convex and twice continuously differentiable real-valued function $F : \mathbb{R}^D \to \mathbb{R}$.

Let us cite the following well-known properties of Bregman generators [4]:

Property 1 *(Bregman generators are equivalent up to modulo affine terms) The Bregman generator $F_2(\theta) = F_1(\theta) + \langle a, \theta \rangle + b$ (with $a \in \mathbb{R}^D$ and $b \in \mathbb{R}$) yields the same Bregman divergence as the Bregman divergence induced by F_1, $B_{F_2}(\theta_1 : \theta_2) = B_{F_1}(\theta_1 : \theta_2)$, and therefore the same dually flat space $(\mathcal{M}, F_2) \cong (\mathcal{M}, F_1)$.*

Property 2 *(Linearity rule of Bregman generators) Let F_1, F_2 be two Bregman generators and $\lambda_1, \lambda_2 > 0$. Then $B_{\lambda_1 F_1 + \lambda_2 F_2}(\theta : \theta') = \lambda_1 B_{F_1}(\theta : \theta') + \lambda_2 B_{F_2}(\theta : \theta')$.*

In practice, the algorithmic toolbox in dually flat spaces (e.g., clustering [4], minimum enclosing balls [39], hypothesis testing [31] and Chernoff information [32], Voronoi diagrams [6, 34], proximity data-structures [45, 46], etc.) can be used whenever the dual Legendre convex conjugates F and F^* are both available in closed-form (see Type 1 of Table 4). In that case, both the primal $\gamma(P, Q) := \{(PQ)_\lambda\}_\lambda$ and dual $\gamma^*(P, Q) := \{(PQ)^*_\lambda\}_\lambda$ geodesics are available in closed form. These dual geodesics can either be expressed using the θ or η-coordinate systems as follows:

$$(PQ)_\lambda = \begin{cases} \theta((PQ)_\lambda) = \theta(P) + \lambda(\theta(Q) - \theta(P)), \\ \eta((PQ)_\lambda) = \nabla F(\theta((PQ)_\lambda)) = \nabla F(\nabla F^*(\eta(P)) + \lambda(\nabla F^*(\eta(Q)) - \nabla F^*(\eta(P)))), \end{cases} \tag{13}$$

[2] That is, we can associate to any dually flat manifold a divergence that amounts to a Bregman divergence [1].

Table 2 Some fundamental Bregman clustering algorithms [4, 22, 41] (of the Bregman algorithmic toolbox) that illustrate which closed-form are required to be run in practice

Algorithm	$F(\theta)$	$\eta(\theta) = \nabla F(\theta)$	$\theta(\eta) = \nabla F^*(\eta)$	$F^*(\eta)$
Right-sided Bregman clustering	✓	✓	×	×
Left-sided Bregman clustering	×	×	✓	✓
Symmetrized Bregman centroid	✓	✓	✓	✓
Mixed Bregman clustering	✓	✓	✓	✓
Maximum Likelihood Estimator for EFs	×	×	✓	×
Bregman soft clustering (≡ EM)	×	✓	✓	✓

$$(PQ)^*_\lambda = \begin{cases} \eta((PQ)^*_\lambda) = \eta(P) + \lambda(\eta(Q) - \eta(P)), \\ \theta((PQ)^*_\lambda) = \nabla F^*(\eta((PQ)^*_\lambda)) = \nabla F^*(\nabla F(\theta(P)) + \lambda(\nabla F(\theta(Q)) - \nabla F(\theta(P)))) \end{cases} \tag{14}$$

That is, the primal geodesic corresponds to a straight line in the primal coordinate system while the dual geodesic is a straight line in the dual coordinate system. However, in many interesting cases, the convex generator F or its dual F^* (or both) are not available in closed form or are computationally intractable, and the above Bregman toolbox cannot be used. Table 2 summarizes the closed-form formulas required to execute some fundamental clustering algorithms [4, 22, 41] in a Bregman geometry.

Let us notice that so far the points $P \in \mathcal{M}$ in the dually flat manifold have no particular meaning, and that the dually flat space structure is generic, not necessarily related to a statistical flat manifold. We shall now quickly review the dualistic structure of statistical manifolds [24].

1.2 Geometry of Statistical Manifolds

Let $I_1(x; y)$ denote a *scalar divergence*. A *statistical divergence* between two probability distributions P and Q, with Radon-Nikodym derivatives $p(x)$ and $q(x)$ with respect to (w.r.t.) a base measure μ defined on the support $\mathcal{X}$, is defined as:

$$I(P : Q) = \int_{x \in \mathcal{X}} I_1 \left(p(x) : q(x)\right) \mathrm{d}\mu(x). \tag{15}$$

A statistical divergence is a measure of dissimilarity/discrimination that satisfies $I(P : Q) \geq 0$ with equality iff. $P = Q$ (a.e., reflexivity property) . For example, the Kullback–Leibler divergence is a statistical divergence:

$$\mathrm{KL}(P : Q) := \int_{x \in \mathcal{X}} \mathrm{kl}(p(x) : q(x)) \mathrm{d}\mu(x), \tag{16}$$

with corresponding scalar divergence:

$$\mathrm{kl}(x : y) := x \log \frac{x}{y}. \tag{17}$$

The KL divergence between P and Q is also called the *relative entropy* [11] because it is the difference of the *cross-entropy* $h^\times(P : Q)$ between P and Q with the Shannon entropy $h(P)$ of P:

$$\mathrm{KL}(P : Q) = h^\times(P : Q) - h(P), \tag{18}$$

$$h^\times(P : Q) := \int_{x \in \mathcal{X}} p(x) \log \frac{1}{q(x)} \mathrm{d}\mu(x), \tag{19}$$

$$h(P) := \int_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} \mathrm{d}\mu(x) = h^\times(P : P). \tag{20}$$

Thus we distinguish a statistical divergence from a parameter divergence by stating that a statistical divergence is a separable divergence that is the definite integral on the support of a scalar divergence.

In information geometry [1, 8], we equip a probability manifold $\mathcal{M} = \{p(x; \theta) : \theta \in \Theta\}$ with a *metric tensor* g (for measuring angles between vectors and lengths of vectors in tangent planes) and a *pair of dual torsion-free connections* ∇ and ∇^* (for defining parallel transports and geodesics) that are defined by their Christoffel symbols Γ_{ijk} and Γ^*_{ijk}. These geometric structures $(\mathcal{M}, D) := (\mathcal{M}, g_D, \nabla_D, \nabla^*_D)$ can be induced by *any smooth* C^∞ divergence $D(\cdot : \cdot)$ [1, 2, 8, 17] as follows:

$$g_{ij}(x) = \left. \frac{\partial^2}{\partial x_i \partial x_j} D(x : y) \right|_{y=x}, \tag{21}$$

$$\Gamma_{ijk}(x) = \left. -\frac{\partial^3}{\partial x_i \partial x_j \partial y_k} D(x : y) \right|_{y=x}. \tag{22}$$

The *dual divergence* $D^*(p : q) := D(q : p)$ highlights the *reference duality* [57], and the dual connection ∇^* is induced by the dual divergence $D^*(\cdot : \cdot)$ (∇^* is defined by $\Gamma^*_{ijk}(x) = \left. -\frac{\partial^3}{\partial x_i \partial x_j \partial y_k} D^*(x : y) \right|_{y=x}$). Observe that the metric tensor is self-dual: $g^* = g$.

Let us give some examples of parametric probability families and their statistical manifolds induced by the Kullback–Leibler divergence.

1.2.1 Exponential Family Manifold (EFM)

We start by a definition:

Definition 2 (*Exponential family*) Let μ be a prescribed base measure and $t(x)$ a sufficient statistic vector. We can build a corresponding exponential family:

$$\mathcal{E}_{t,\mu} := \{p(x;\theta) \propto \exp(\langle t(x), \theta\rangle)\}_\theta, \tag{23}$$

where $p(x;\theta) := \frac{\mathrm{d}P(\theta)}{\mathrm{d}\mu}(x)$.

The densities are normalized by the cumulant function F:

$$F(\theta) := \log\left(\int_{x\in\mathcal{X}} \exp(\langle t(x), \theta\rangle)\mathrm{d}\mu(x)\right), \tag{24}$$

so that:

$$p(x;\theta) = \exp(\langle t(x), \theta\rangle - F(\theta)). \tag{25}$$

The function F is a Bregman generator on the natural parameter space:

$$\Theta := \left\{\theta : \int_{x\in\mathcal{X}} \exp(\langle t(x), \theta\rangle)\mathrm{d}\mu(x) < \infty\right\}. \tag{26}$$

If we add an extra carrier term $k(x)$ and consider the measure $\nu(x) := \frac{\mu(x)}{\exp(k(x))}$, we get the generic form of an exponential family [36]:

$$\mathcal{E}_{t,k,\nu} := \{p(x;\theta) \propto \exp(\langle t(x), \theta\rangle + k(x)) : \theta \in \Theta\}. \tag{27}$$

We call the function F the *Exponential Family Bregman Generator*, or EFBG for short in the remainder.

It turns out that $(\mathcal{E}_{t,\mu}, \mathrm{KL}, \nabla_{\mathrm{KL}}, \nabla^*_{\mathrm{KL}}) \cong (\mathcal{M}, F)$ (meaning the information-geometric structure of the statistical manifold is isomorphic to the information-geometry of a dually flat manifold) so that:

$$\mathrm{KL}(p(x;\theta_1) : p(x;\theta_2) = B_F(\theta_2 : \theta_1), \tag{28}$$

$$= B_{F^*}(\eta_1 : \eta_2), \tag{29}$$

with $\eta = E_{p(x;\theta)}[t(x)]$ the dual parameter called the expectation parameter or moment parameter.

1.2.2 Mixture Family Manifold (MFM)

Another important type of families of probability distributions are the mixture families:

Definition 3 (*Mixture family*) Given a set of k prescribed statistical distributions $p_0(x), \ldots, p_{k-1}(x)$, all sharing the same support $\mathcal{X}$ (say, $\mathbb{R}$), a *mixture family* $\mathcal{M}$ of order $D = k - 1$ consists of all *strictly convex combinations* of these component distributions [43, 44]:

$$\mathcal{M} := \left\{ m(x;\eta) = \sum_{i=1}^{k-1} \eta_i p_i(x) + \left(1 - \sum_{i=1}^{k-1} \eta_i\right) p_0(x) : \eta_i > 0, \sum_{i=1}^{k-1} \eta_i < 1 \right\}. \tag{30}$$

It shall be understood from the context that $\mathcal{M}$ is a shorthand for $\mathcal{M}_{p_0(x),\ldots,p_D}$.

It turns out that $(\mathcal{M}, \mathrm{KL}, \nabla_{\mathrm{KL}}, \nabla^*_{\mathrm{KL}}) \cong (\mathcal{M}, G)$ so that:

$$\mathrm{KL}(m(x;\eta) : m(x;\eta')) = B_G(\eta : \eta'), \tag{31}$$

for the Bregman generator being the Shannon negative entropy (also called Shannon information):

$$G(\eta) = -h(m(x;\eta)) = \int_{x\in\mathcal{X}} m(x;\eta) \log m(x;\eta) \mathrm{d}\mu(x). \tag{32}$$

We call function G the *Mixture Family Bregman Generator*, or MFBG for short in the remainder.

For a mixture family, we prefer to use the notation η instead of θ for indexing the distribution parameters as it is customary in textbooks of information geometry [1, 8]. One reason comes from the fact that the KL divergence between two mixtures amounts to a BD on their respective parameters (Eq. 31) while the KL divergence between exponential family distributions is equivalent to a BD on the swapped order of their respective parameters (Eq. 28), see [3, 19]. Thus in order to get the same order of arguments for the KL between two exponential family distributions, we need to use the dual Bregman divergence on the dual η parameter, see Eq. 29.

1.2.3 Cauchy Family Manifold (CFM)

This example is only given to emphasize that probability families may neither be exponential nor mixture families [28].

A Cauchy distribution has probability density defined on the support $\mathcal{X} = \mathbb{R}$ by:

$$p(x;\mu,\sigma) = \frac{1}{\pi\sigma\left(1 + \left(\frac{x-\mu}{\sigma}\right)^2\right)}. \tag{33}$$

The space of all Cauchy distributions:

$$\mathcal{C} = \{p(x;\mu,\sigma) : \mu \in \mathbb{R}, \sigma > 0\}. \tag{34}$$

is a location-scale family [23]. It is not an exponential family nor a mixture family.

Table 3 compares the dually flat structures of mixture families with exponential families. In information geometry, $(\mathcal{E}_{t,k,\mu}, \mathrm{KL}, \nabla_{\mathrm{KL}}, \nabla^*_{\mathrm{KL}}) = (\mathcal{E}_{t,k,\mu}, g, \nabla^e, \nabla^m)$ and $(\mathcal{M}, \mathrm{KL}, \nabla_{\mathrm{KL}}, \nabla^*_{\mathrm{KL}}) = (\mathcal{M}, g, \nabla^m, \nabla^e)$ where g is the *Fisher information metric*

Table 3 Characteristics of the dually flat geometries of Exponential Families (EFs) and Mixture Families (MFs)

	Exponential Family	Mixture Family
Density	$p(x;\theta) = \exp(\langle\theta, x\rangle - F(\theta))$	$m(x;\eta) = \sum_{i=1}^{k-1} \eta_i f_i(x) + c(x)$
		$f_i(x) = p_i(x) - p_0(x)$
Family/Manifold	$\mathcal{M} = \{p(x;\theta) \ : \ \theta \in \Theta^\circ\}$	$\mathcal{M} = \{m(x;\eta) \ : \ \eta \in H^\circ\}$
Convex function ($\equiv ax + b$)	F: cumulant	F^*: negative entropy
Dual coordinates	moment $\eta = E[t(x)]$	$\theta^i = h^\times(p_0 : m) - h^\times(p_i : m)$
Fisher Information $g = (g_{ij})_{ij}$	$g_{ij}(\theta) = \partial_i \partial_j F(\theta)$ $g = \mathrm{Var}[t(X)]$	$g_{ij}(\eta) = \int_{\mathcal{X}} \frac{f_i(x) f_j(x)}{m(x;\eta)} \mathrm{d}\mu(x)$
Christoffel symbol	$\Gamma_{ij,k} = \frac{1}{2}\partial_i \partial_j \partial_k F(\theta)$	$g_{ij}(\eta) = -\partial_i \partial_j h(\eta)$ $\Gamma_{ij,k} = -\frac{1}{2} \int_{\mathcal{X}} \frac{f_i(x) f_j(x) f_k(x)}{m^2(x;\eta)} \mathrm{d}\mu(x)$
Entropy	$-F^*(\eta)$	$-F^*(\eta)$
Kullback–Leibler divergence	$B_F(\theta_2 : \theta_1)$	$B_{F^*}(\eta_1 : \eta_2)$
	$= B_{F^*}(\eta_1 : \eta_2)$	$= B_F(\theta_2 : \theta_1)$

tensor and ∇^e and ∇^m are the exponential and mixture connections, respectively. These connections are dual to each others, see [8].

1.3 Computational Tractability of Dually Flat Statistical Manifolds

The previous section explained the dually flat structures (i.e., Bregman geometry) of the exponential family manifold and of the mixture family manifold. However these geometries may be purely theoretical as the Bregman generator F may not be available in closed form so that the Bregman toolbox cannot be used in practice. This work tackles this problem faced in exponential and mixture family manifolds by proposing the novel framework of *Monte Carlo Information Geometry* (MCIG). MCIG approximates the untractable Bregman geometry by considering the Monte Carlo stochastic integration of the definite integral-based ideal Bregman generator.

But first, let us quickly review the five types of tractability of Bregman geometry in the context of statistical manifolds by giving an illustrating family example for each type:

Type 1. F and ∇F^* are both available in closed-form, and so are ∇F and F^*. For example, this is the case of the *the Gaussian exponential family*. The normal distribution [36] has sufficient statistic vector $t(x) = (x, x^2)$ so that its log-normalizer is

$$F(\theta) = \log\left(\int_{-\infty}^{+\infty} \exp(\theta_1 x + \theta_2 x^2)\mathrm{d}x\right). \tag{35}$$

Since $\int_{-\infty}^{\infty} \exp(\theta_1 x + \theta_2 x^2) = \sqrt{\frac{\pi}{-\theta_2}} \exp(-\frac{\theta_1^2}{4\theta_2})$ for $\theta_2 < 0$, we find:

$$F(\theta) = \log\left(\int \exp(\theta_1 x + \theta_2 x^2)\mathrm{d}x\right) = -\frac{\theta_1^2}{4\theta_2} + \frac{1}{2}\log\frac{\pi}{-\theta_2}. \tag{36}$$

This is in accordance with the direct canonical decomposition [36] of the density $p(x;\theta) = \exp(\langle t(x), \theta\rangle - F(\theta))$ of the normal density $p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$.

Remark 1 When $F(\theta)$ can be expressed using the canonical decomposition of exponential families, this means that the definite integral $\log(\int \exp(\langle t(x), \theta\rangle + k(x))\mathrm{d}x)$ is available in closed form, and vice-versa.

Type 2. F is available in closed form (and so is ∇F) but ∇F^* is not available in closed form (and therefore F^* is not available too). This is for example the *Beta exponential family*. A Beta distribution $\mathrm{Be}(\alpha,\beta)$ has density on support $x \in (0,1)$:

$$p(x;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \tag{37}$$

where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, and $(\alpha > 0, \beta > 0)$ are the shape parameters. The Beta family of distributions is an exponential family with $\theta = (\alpha,\beta)$, $t(x) = (\log(x), \log(1-x))$, $k(x) = -\log(x) - \log(1-x)$ and $F(\theta) = \log B(\theta_1,\theta_2) = \log\Gamma(\theta_1) + \log\Gamma(\theta_2) - \log\Gamma(\theta_1+\theta_2)$. Note that we could also have chosen $\theta = (\alpha-1, \beta-1)$ and $k(x) = 0$. Thus $\nabla F(\theta) = (\psi(\theta_1) - \psi(\theta_1+\theta_2), \psi(\theta_2) - \psi(\theta_1+\theta_2))$ where $\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ is the digamma function. Inverting the gradient $\nabla F(\theta) = \eta$ to get $\eta = \nabla F^*(\theta)$ is not available in closed-form.[3]

Type 3. This type of families has discrete support $\mathcal{X}$ and thus requires an exponential time to compute the log-normalizer. For example, consider the Ising models [5, 9, 21]: Let $G = (V, E)$ be an undirected graph of $|V|$ nodes and $|E|$ edges. Each node $v \in V$ is associated with a binary random variable $x_v \in \{0,1\}$. The probability of an Ising model is defined as follows:

$$p(x;\theta) = \exp\left(\sum_{v\in V} \theta_v x_v + \sum_{(v,w)\in E} \theta_{vw} x_v x_w - F(\theta)\right). \tag{38}$$

[3] To see this, consider the digamma difference property: $f_\Delta(\theta) = \psi(\theta) - \psi(\theta+\Delta) = -\sum_{i=0}^{\Delta-1} \frac{1}{x+i}$ for $\Delta \in \mathbb{N}$. We cannot invert $f_\Delta(\theta)$ since it involves solving the root of a high-degree polynomial.

The vector $t(x) = (\ldots, x_v, \ldots, x_{vw}, \ldots)$ of sufficient statistics is D-dimensional with $D = |V| + |E|$. The log-normalizer is:

$$F(\theta) = \log \left(\sum_{(x_v)_v \in \{0,1\}^{|V|}} \left(\exp \sum_{v \in V} \theta_v x_v + \sum_{(v,w) \in E} \theta_{vw} x_v x_w \right) \right). \tag{39}$$

It requires to sum up $2^{|V|}$ terms.

Type 4. This type of families has a Bregman generator which is not available in closed-form. For example, this is the case of the *Polynomial Exponential Family* [10, 42] (PEF) which are helpful to model a multimodal distribution (instead of using a statistical mixture). Consider the following vector of sufficient statistics $t(x) = (x, x^2, \ldots, x^D)$ for defining an exponential family:

$$\mathcal{E}_{t(x),\mu} = \left\{ p(x;\theta) = \exp\left(\sum_{i=1}^{D} \theta_i x^i - F(\theta) \right) : \theta \in \Theta \right\}. \tag{40}$$

(Beware that here, $x^i = \mathrm{Pow}(x, i) := \underbrace{x \times \cdots \times x}_{i \text{ times}}$ denotes the ith power of x (monomial of degree i), and not a contravariant coefficient of a vector x.)

In general, the definite integral of the cumulant function (the Exponential Family Bregman Generator, EFBG) of Eq. 24 does not admit a closed form, but is analytic. For example, choosing $t(x) = x^8$, we have:

$$F(\theta) = \log \int_{-\infty}^{\infty} \exp(\theta x^8) \mathrm{d}x = \log 2 + \log \Gamma(9/8) - \frac{1}{8} \log(-\theta), \tag{41}$$

for $\theta < 0$. But $\int_{-\infty}^{\infty} \exp(-x^8 - x^4 - x^2) \mathrm{d}x \simeq 1.295$ is not available in closed form.

Type 5. This last category is even more challenging from a computational point of view because of log-sum terms. For example, the *mixture family*. As already stated, the negative Shannon entropy (i.e., the Mixture Family Bregman Generator, MFBG) is not available in closed form for statistical mixture models [43]. It is in fact even worse, as the Shannon entropy of mixtures is not analytic [56].

This paper considers approximating the computationally untractable generators of statistical exponential/mixture families (type 4 and type 5) using stochastic Monte Carlo approximations.

In [12], Critchley et al. take a different approach of the computational tractability by discretizing the support $\mathcal{X}$ into a finite number of bins, and considering the corresponding discrete distribution. However, this approach does not scale well with the dimension of the support. Our Monte Carlo Information Geometry scales to arbitrary high dimensions because it relies on the fact that the Monte Carlo stochastic estimator is independent of the dimension [52].

1.4 Paper Organization

In Sect. 2, we consider the MCIG structure of mixture families: Namely, Sect. 2.1 considers first the uni-order families to illustrate the basic principle. It is followed by the general case in Sect. 2.2. Similarly, Sect. 3 handles the exponential family case by first explaining the uni-order case in Sect. 3.1 before tackling the general case in Sect. 3.2. Sect. 4 presents an application of the computationally-friendly MCIG structures for clustering distributions in dually flat statistical mixture manifolds. In Sect. 5, we show how to construct non-flat MCIG structures of a parametric family of distributions given by a statistical separable divergence. Finally, we conclude and discuss several perspectives in Sect. 6.

2 Monte Carlo Information Geometry of Mixture Families

Recall the definition of a statistical mixture model (Definition 3): Given a set of k prescribed statistical distributions $p_0(x), \ldots, p_{k-1}(x)$, all sharing the same support $\mathcal{X}$, a *mixture family* $\mathcal{M}$ of order $D = k - 1$ consists in all *strictly convex combinations* of the $p_i(x)$'s [43]:

$$\mathcal{M} := \left\{ m(x;\eta) = \sum_{i=1}^{k-1} \eta_i p_i(x) + \left(1 - \sum_{i=1}^{k-1} \eta_i\right) p_0(x) : \eta_i > 0, \sum_{i=1}^{k-1} \eta_i < 1 \right\}. \tag{42}$$

The differential-geometric structure of $\mathcal{M}$ is well studied in information geometry [1, 8] (although much less than for the exponential families), where it is known that:

$$\mathrm{KL}(m(x;\eta) : m(x;\eta')) = B_G(\eta : \eta'), \tag{43}$$

for the Bregman generator being the Shannon negative entropy (MFBG):

$$G(\eta) = -h(m(x;\eta)) = \int_{x\in\mathcal{X}} m(x;\eta) \log m(x;\eta) \mathrm{d}\mu(x). \tag{44}$$

The negative entropy $G(\eta) = \int_{x\in\mathcal{X}} m(x;\eta) \log m(x;\eta) \mathrm{d}\mu(x)$ is a smooth and strictly convex function which induces a dually flat structure with Legendre convex conjugate:

$$F(\theta) = G^*(\theta) = -\int_{x\in\mathcal{X}} p_0(x) \log m(x;\eta) \mathrm{d}\mu(x) = h^\times(p_0(x) : m(x;\eta)), \tag{45}$$

interpretable as the cross-entropy of $p_0(x)$ with the mixture $m(x;\eta)$ [43].

Notice that the component distributions may be heterogeneous like $p_0(x)$ being a fixed Cauchy distribution, $p_1(x)$ being a fixed Gaussian distribution, $p_2(x)$ a Laplace

distribution, etc. Except for the case of the finite categorical distributions (that are both interpretable as either a mixture family and an exponential family, see [1]), $G(\eta)$ provably does not admit a closed form [56] (i.e., meaning that the definite integral of Eq. 32 does not admit a simple formula using common standard functions). Thus the dually-flat geometry $(\mathcal{M}, G)$ is a theoretical construction which cannot be explicitly used by Bregman algorithms.

One way to tackle the lack of closed form in Eq. 32, is to approximate the definite integrals whenever they are used by using Monte Carlo stochastic integration. However, this is computationally very expensive, and, even worse, it cannot guarantee that the overall computation is consistent.

Let us briefly explain the meaning of *consistency*: We can estimate the KL between two distributions p and q by drawing m variates $x_1, \ldots, x_m \sim p(x)$, and use the following MC KL estimator:

$$\widehat{\mathrm{KL}}_m(p : q) := \frac{1}{m} \sum_{i=1}^{m} \log \frac{p(x_i)}{q(x_i)}. \tag{46}$$

Now, suppose we have $\mathrm{KL}(p : q) \leq \mathrm{KL}(q : r)$, then their MC estimates may not satisfy $\widehat{\mathrm{KL}}_m(p : q) < \widehat{\mathrm{KL}}_m(q : r)$ (since each time we evaluate a $\widehat{\mathrm{KL}}_m$ we draw different samples). Thus when running a KL/Bregman algorithm, the more MC stochastic approximations of integrals are performed in the algorithm, the less likely is the output consistent. For example, consider computing the Bregman Voronoi diagram [34] of a set of n mixtures belonging to a mixture family manifold (say, with $D = 2$) using the algorithm explained in [34]: Since we use for each BD calculation or predicate evaluation relying on F or F^* stochastic Monte Carlo integral approximations, this MC algorithm may likely not deliver a proper combinatorial structure of the Voronoi diagram: The Voronoi structure is likely to be inconsistent.

Let us now show how Monte Carlo Information Geometry (MCIG) approximates this computationally untractable $(\mathcal{M}, G)$ geometric structure by defining a consistent and computationally-friendly dually-flat information geometry $(\mathcal{M}, \tilde{G}_{\mathcal{S}})$ for a finite number m of identically and independently distributed (iid) random samples $\mathcal{S}$.

2.1 MCIG of Order-1 Mixture Family

In order to highlight the principle of MCIGs, let us first consider a mixture family of order $D = 1$. That is, we consider a set of mixtures of $k = 2$ components with density:

$$m(x; \eta) = \eta p_1(x) + (1 - \eta) p_0(x) = p_0(x) + \eta(p_1(x) - p_0(x)), \tag{47}$$

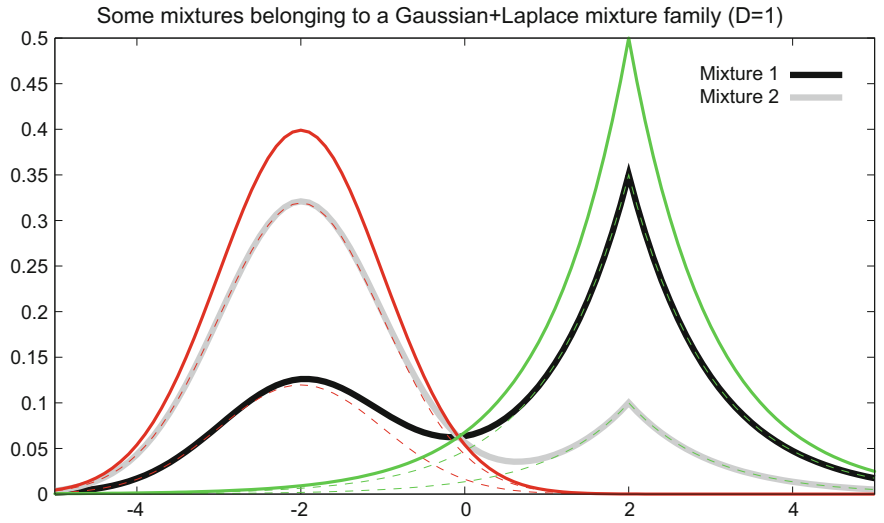

Fig. 1 Example of a mixture family of order $D = 1$ ($k = 2$): $p_0(x) \sim$ Gaussian$(-2, 1)$ (red) and $p_1(x) \sim$ Laplace$(2, 1)$ (green). The two mixtures are $m_1(x) = m(x; \eta_1)$ (black) with $\eta_1 = 0.7$ and $m_2(x) = m(x; \eta_2)$ (grey) with $\eta_2 = 0.2$. Weighted component distributions are displayed in dashed

with parameter η ranging in $(0, 1)$. The two prescribed component densities $p_0(x)$ and $p_1(x)$ (with respect to a base measure μ, say the Lebesgue measure) are defined on a common support $\mathcal{X}$. Densities $p_0(x)$ and $p_1(x)$ are assumed to be linearly independent [8].

Figure 1 displays an example of uni-order mixture family with heterogeneous components: $p_0(x)$ is chosen as a Gaussian distribution while $p_1(x)$ is taken as a Laplace distribution. A mixture $m(x; \eta)$ of $\mathcal{M}$ is visualized as a point P (here, one-dimensional) with $\eta(P) = \eta$.

Let $\mathcal{S} = \{x_1, \ldots, x_m\}$ denote a iid sample from a fixed *proposal distribution* $q(x)$ (with $q(x) > 0$ for $x \in \mathcal{X}$, and $q(x)$ independent of η). We approximate the Bregman generator $G(\eta)$ using Monte Carlo stochastic integration with importance sampling as follows:

$$G(\eta) \simeq \tilde{G}_{\mathcal{S}}(\eta) := \frac{1}{m} \sum_{i=1}^{m} \frac{1}{q(x_i)} m(x_i; \eta) \log m(x_i; \eta). \tag{48}$$

Let us prove that the Monte Carlo function $\tilde{G}_{\mathcal{S}}(\eta)$ is a proper Bregman generator. That is, that $\tilde{G}_{\mathcal{S}}(\eta)$ is strictly convex and twice continuously differentiable (Definition 1).

Write for short $m_x(\eta) := m(x;\eta)$ so that $G(\eta) = \int_{x\in\mathcal{X}} m_x(\eta)\log m_x(\eta)\mathrm{d}\mu(x)$ is approximated by $\frac{1}{m}\sum_{i=1}^{m}\frac{1}{q(x_i)}m_{x_i}(\eta)\log m_{x_i}(\eta)$. Since $\frac{1}{m}\frac{1}{q(x_i)} > 0$, it suffices to prove that the function $g_x(\eta) = m_x(\eta)\log m_x(\eta)$ is strictly convex wrt parameter η. Then we shall conclude that $\tilde{G}_{\mathcal{S}}(\eta)$ is strictly convex because it is a finite positively weighted sum of strictly convex functions.

Let us write the first and second derivatives of $g_x(\eta)$ as follows:

$$g_x(\eta)' = m_x(\eta)'(\log m_x(\eta) + 1), \tag{49}$$

$$g_x(\eta)'' = m_x(\eta)''(\log m_x(\eta) + 1) + \frac{(m_x(\eta)')^2}{m_x(\eta)}. \tag{50}$$

Since $m_x'(\eta) = p_1(x) - p_0(x)$ and $m_x''(\eta) = 0$, we get:

$$g_x(\eta)'' = \frac{(p_1(x) - p_0(x))^2}{m_x(\eta)}. \tag{51}$$

Thus it follows that:

$$\tilde{G}_{\mathcal{S}}''(\eta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{q(x_i)}\frac{(p_1(x_i) - p_0(x_i))^2}{m(x_i;\eta)} \geq 0. \tag{52}$$

It is strictly convex provided that there exists at least one x_i such that $p_1(x_i) \neq p_0(x_i)$.

Let $\mathcal{D} \subset \mathcal{X}$ denote the degenerate set $\mathcal{D} = \{x \in \mathcal{X} : p_1(x) = p_0(x)\}$. For example, if $p_0(x)$ and $p_1(x)$ are two distinct univariate normal distributions, then $|\mathcal{D}| = 2$ (roots of a quadratic equation), and

$$\mu_q(\mathcal{D}) := \int_{x\in\mathcal{X}} 1_{[p_0(x)=p_1(x)]}q(x)\mathrm{d}\mu(x) = 0. \tag{53}$$

Assumption 1 *(AMF1D)* We assume that $p_0(x)$ and $p_1(x)$ are linearly independent (non-singular statistical model, see [8]), and that $\mu_q(\mathcal{D}) = 0$.

Lemma 1 (Monte Carlo Mixture Family Function is a Bregman generator) *The Monte Carlo Mixture Family Function (MCMFF) $\tilde{F}_{\mathcal{S}}(\theta)$ is a Bregman generator almost surely.*

Proof When there exists a sample $x \in \mathcal{S}$ with two distinct densities $p_0(x)$ and $p_1(x)$, we have $(p_1(x_i) - p_0(x_i))^2 > 0$ and therefore $\tilde{G}_{\mathcal{S}}''(\eta) > 0$. The probability to get a degenerate sample is almost zero.

To recap, the MCMFF of the MCIG of uni-order family has the following characteristics:

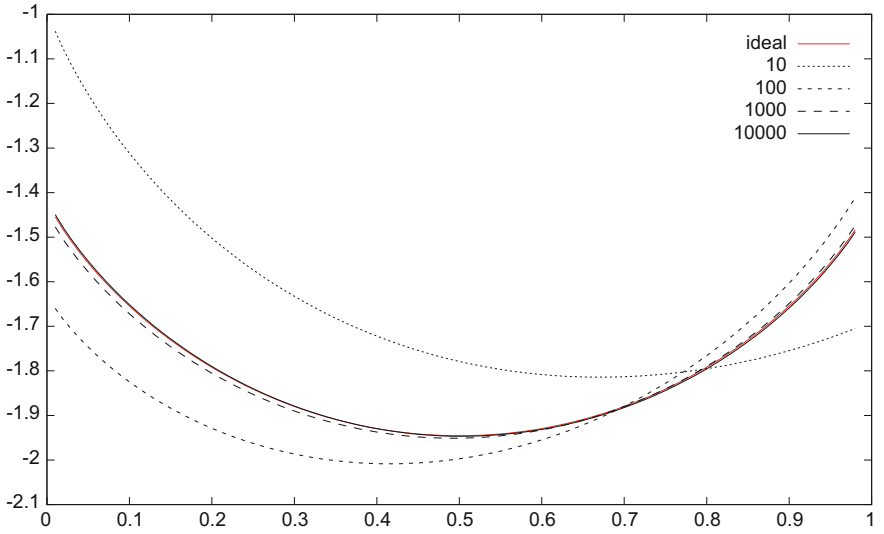

Fig. 2 A series $G_{\mathcal{S}}(\eta)$ of Bregman Monte Carlo Mixture Family generators (for $m = |\mathcal{S}| \in \{10, 100, 1000, 10000\}$) approximating the untractable ideal negentropy generator $G(\eta) = -h(m(x;\eta))$ (red) of a mixture family with prescribed Gaussian distributions $m(x;\eta) = (1-\eta)p(x;0,3) + \eta p(x;2,1)$ for the proposal distribution $q(x) = m(x;\frac{1}{2})$

Monte Carlo Mixture Family Generator 1D:

$$\tilde{G}_{\mathcal{S}}(\eta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{q(x_i)}m(x_i;\eta)\log m(x_i;\eta), \tag{54}$$

$$\tilde{G}'_{\mathcal{S}}(\eta) = \theta = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{q(x_i)}(p_1(x_i)-p_0(x_i))(1+\log m(x_i;\eta)), \tag{55}$$

$$\tilde{G}''_{\mathcal{S}}(\eta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{q(x_i)}\frac{(p_1(x_i)-p_0(x_i))^2}{m(x_i;\eta)}. \tag{56}$$

Note that $(G^*)'$ and G^* may be calculated numerically but not in closed-form. We may also MC approximate ∇G^* since $\theta = (h^\times(p_0 : m) - h^\times(p_i : m))_i$.

Thus we change from type 5 to type 2 the computational tractability of mixtures by adopting the MCIG approximation.

Figure 2 displays a series of Bregman mixture family MC generators for a mixture family for different values of $|\mathcal{S}| = m$.

As we increase the sample size of $\mathcal{S}$, the MCMFF Bregman generator tends to the ideal mixture family Bregman generator.

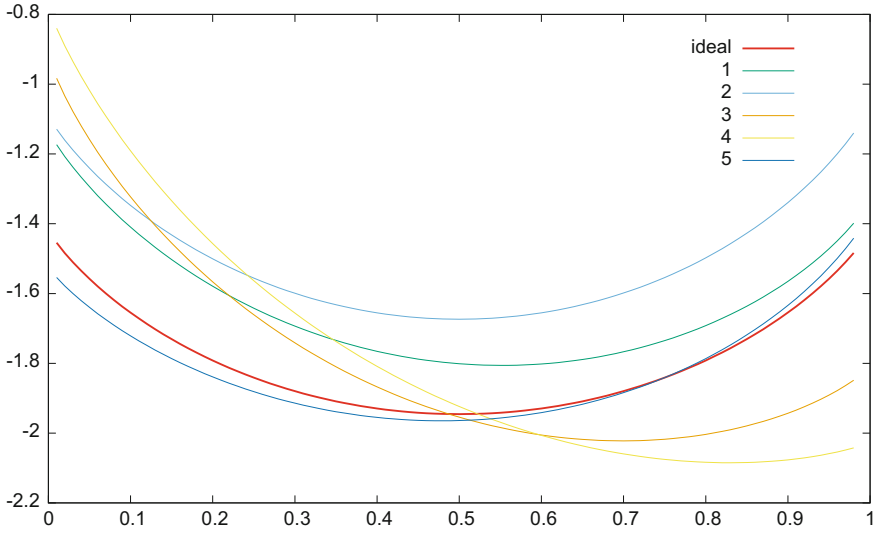

Fig. 3 The Monte Carlo Mixture Family Generator $\hat{G}_{10}$ (MCMFG) considered as a random variable: Here, we show five realizations (i.e., $\mathcal{S}_1, \ldots, \mathcal{S}_5$) of the randomized generator for $m = 5$. The ideal generator is plot in thick red

Theorem 1 (Consistency of MCIG) *Almost surely,* $\lim_{m\to\infty}(\mathcal{M}, \tilde{G}_{\mathcal{S}}) = (\mathcal{M}, G)$ *when* $\mu_q(\mathcal{D}) = 0$.

Proof It suffices to prove that $\lim_{m\to\infty} \tilde{G}_{\mathcal{S}}(\eta) = G(\eta)$. The general theory of Monte Carlo stochastic integration yields a consistent estimator provided that the following variance is bounded

$$\mathrm{Var}_q\left[\frac{m(x;\eta)\log m(x;\eta)}{q(x)}\right] < \infty. \tag{57}$$

For example, when $m(x;\eta)$ is a mixture of prescribed isotropic gaussians (say, from a KDE), and $q(x)$ is also an isotropic Gaussian, the variance is bounded. Note that q is the proposal density wrt the base measure μ.

In practice, the proposal distribution $q(x)$ can be chosen as the uniform mixture of the fixed component distributions:

$$q(x) = \frac{1}{m}\sum_{i=0}^{D} p_i(x). \tag{58}$$

Notice that the Monte Carlo Mixture Family Function is a random variable (r.v. for short) estimator itself by considering a vector of iid variables instead of a sample variate: $\hat{G}_m(\eta)$. Figure 3 displays five realizations of the random variable $\hat{G}_m(\eta)$ for $m = 10$.

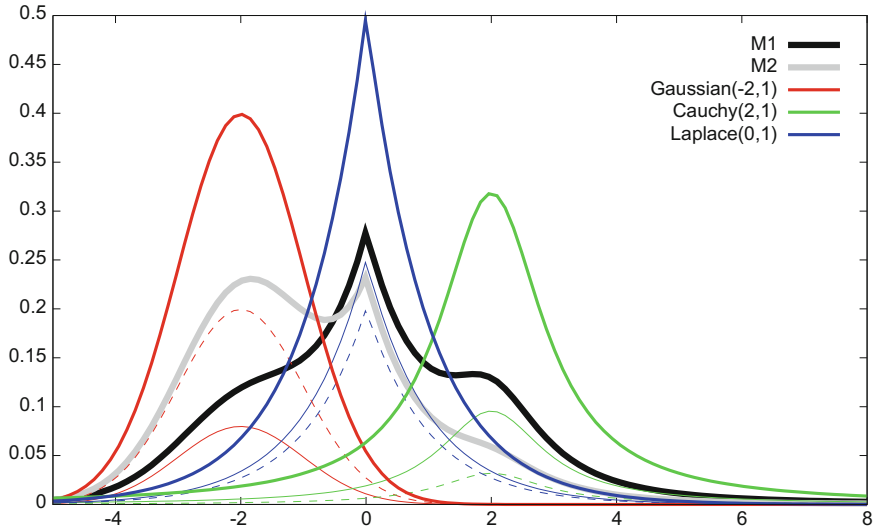

Fig. 4 Example of a mixture family of order $D = 2$ ($k = 3$): $p_0(x) \sim \text{Gaussian}(-2, 1)$ (red), $p_1(x) \sim \text{Laplace}(0, 1)$ (blue) and $p_2(x) \sim \text{Cauchy}(2, 1)$ (green). The two mixtures are $m_1(x) = m(x; \eta_1)$ (black) with $\eta_1 = (0.3, 0.5)$ and $m_2(x) = m(x; \eta)$ (gray) with $\eta = (0.1, 0.4)$

2.2 General D-Order Mixture Case

Here, we consider statistical mixtures with $k = D + 1 > 2$ prescribed distributions $p_0(x), \ldots, p_D(x)$. The component distributions are linearly independent so that they define a non-singular statistical model [8].

We further strengthen conditions on the prescribed distributions as follows:

Assumption 2 *(AMF)* We assume that the linearly independent prescribed distributions further satisfy:

$$\sup_{B \in \mathcal{B}} \left\{ \mu_q(B) : \exists \lambda \neq (0), \sum_{i \neq j} \lambda_i \left(p_i|_B - p_j|_B \right) = 0 \right\} = 0, \quad \forall j, \tag{59}$$

where the supremum is over all subsets B of the σ-algebra $\mathcal{B}$ of the probability space with support $\mathcal{X}$ and measure μ, with $p_i|_B$ denoting the restriction of p_i to subset B. In other words, we impose that the components $(p_i)_i$ still constitute an affinely independent family when restricted to any subset of positive measure.

For example, Figure 4 displays two mixture distributions belonging to a 2D mixture family with Gaussian, Laplace and Cauchy component distributions.

Recall that the mixture family Monte Carlo generator is:

$$\tilde{G}_{\mathcal{S}}(\eta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{q(x_i)}m(x_i;\eta)\log m(x_i;\eta). \tag{60}$$

In order to prove that G is strictly convex, we shall prove that $\nabla^2\tilde{G}_{\mathcal{S}}(\eta) \succ 0$ almost surely. It suffices to consider the basic Hessian matrix $\nabla^2 g_x = (\partial^i\partial^j g_x(\eta))_{ij}$ of $g_x(\eta) = m_x(\eta)\log m_x(\eta)$. We have the partial first derivatives:

$$\partial^i g_x(\eta) = (p_i(x) - p_0(x))(1 + \log m(x;\eta)), \tag{61}$$

and the partial second derivatives:

$$\partial^i\partial^j g_x(\eta) = \frac{(p_i(x) - p_0(x))(p_j(x) - p_0(x))}{m(x;\eta)}, \tag{62}$$

so that

$$\partial^i\partial^j\tilde{G}_{\mathcal{S}}(\eta) = \frac{1}{m}\sum_{l=1}^{m}\frac{1}{q(x_l)}\frac{(p_i(x_l) - p_0(x_l))(p_j(x_l) - p_0(x_l))}{m(x_l;\eta)}. \tag{63}$$

Theorem 2 (Monte Carlo Mixture Family Function is a Bregman generator) *The Monte Carlo multivariate function $\tilde{G}_{\mathcal{S}}(\eta)$ is always convex and twice continuously differentiable, and strictly convex almost surely.*

Proof Consider the D-dimensional vector:

$$v_l = \begin{bmatrix} \frac{p_1(x_l)-p_0(x_l)}{\sqrt{q(x_l)m(x_l;\eta)}} \\ \vdots \\ \frac{p_D(x_l)-p_0(x_l)}{\sqrt{q(x_l)m(x_l;\eta)}} \end{bmatrix}. \tag{64}$$

Then we rewrite the Monte Carlo generator $\tilde{G}_{\mathcal{S}}(\eta)$ as:

$$\partial^i\partial^j\tilde{G}_{\mathcal{S}}(\eta) = \frac{1}{m}\sum_{l=1}^{m} v_l v_l^\top. \tag{65}$$

Since $v_l v_l^\top$ is always a symmetric positive semidefinite matrix of rank one, we conclude that $\tilde{G}_{\mathcal{S}}(\eta)$ is a symmetric positive semidefinite matrix when $m < D$ (rank deficient) and a symmetric positive definite matrix (full rank) almost surely when $m \geq D$.

3 Monte Carlo Information Geometry of Exponential Families

We follow the same outline as for mixture familes: Sect. 3.1 first describes the univariate case. It is then followed by the general multivariate case in Sect. 3.1.

3.1 MCIG of Order-1 Exponential Family

We consider the order-1 exponential family of parametric densities with respect to a base measure μ:

$$\mathcal{E} := \{p(x;\theta) = \exp(t(x)\theta - F(\theta) + k(x)) : \theta \in \Theta\}, \tag{66}$$

where Θ is the natural parameter space, such that the log-normalizer/cumulant function [1] is

$$F(\theta) = \log\left(\int \exp(t(x)\theta + k(x))\mathrm{d}\mu(x)\right). \tag{67}$$

The sufficient statistic function $t(x)$ and 1 are linearly independent [8].

We perform Monte Carlo stochastic integration by sampling a set $\mathcal{S} = \{x_1, \ldots, x_m\}$ of m iid variates from a proposal distribution $q(x)$ to get:

$$F(\theta) \simeq \tilde{F}^{\dagger}_{\mathcal{S}}(\theta) := \log\left(\frac{1}{m}\sum_{i=1}^{m}\frac{1}{q(x_i)}\exp(t(x_i)\theta + k(x_i))\right). \tag{68}$$

Without loss of generality, assume that x_1 is the element that minimizes the sufficient statistic $t(x)$ among the elements of $\mathcal{S}$, so that $a_i = t(x_i) - t(x_1) \geq 0$ for all $x_i \in \mathcal{S}$.

Let us factorize $\frac{1}{q(x_1)}\exp(t(x_1)\theta + k(x_1))$ in Eq. 68 and remove an affine term from the generator $\tilde{F}_{\mathcal{S}}(\theta)$ to get the equivalent generator (see Property 1):

$$\tilde{F}^{\dagger}_{\mathcal{S}}(\theta) \equiv \tilde{F}_{\mathcal{S}}(\theta), \tag{69}$$

$$\tilde{F}_{\mathcal{S}}(\theta) = \log\left(1 + \sum_{i=2}^{m}\exp((t(x_i) - t(x_1))\theta + k(x_i) - k(x_1) - \log q(x_i) + \log q(x_1))\right), \tag{70}$$

$$= \log\left(1 + \sum_{i=2}^{m}\exp(a_i\theta + b_i)\right), \tag{71}$$

$$:= \mathrm{lse}^{+}_{0}(a_2\theta + b_2, \ldots, a_m\theta + b_m), \tag{72}$$

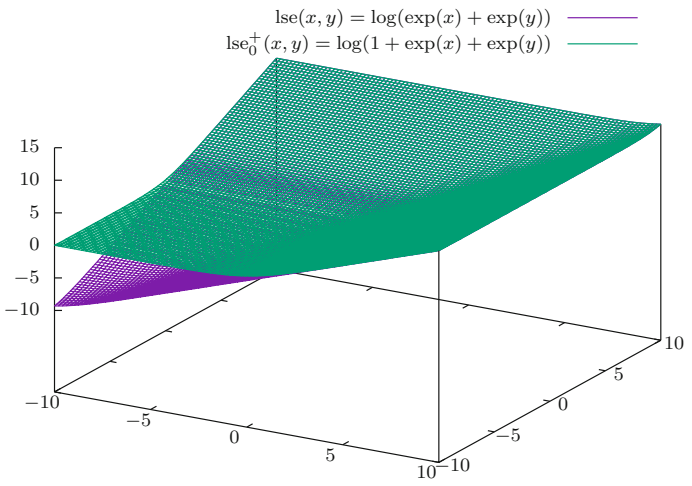

Fig. 5 Graph plots of the lse and lse_0^+ functions: The lse function (violet) is only convex while the lse_0^+ function (green) is always guaranteed to be strictly convex

with $a_2, \ldots, a_m > 0$ and $b_i = k(x_i) - k(x_1) - \log q(x_i) + \log q(x_1)$. Function $\mathrm{lse}_0^+(x_1, \ldots, x_m) = \mathrm{lse}(0, x_1, \ldots, x_m)$ is the log-sum-exp function [20, 47] $\mathrm{lse}(x_1, \ldots, x_m) = \log \sum_{i=1}^n \exp(x_i)$ with an additional argument set to zero.

Let us notice that the lse_0^+ function is always *strictly convex* while the lse function is only convex[4] [7], p. 74. Figure 5 displays the graph plots of the lse and lse_0^+ functions. Let us clarify this point with a usual exponential family: The binomial family. The binomial distribution is a categorical distribution with $D = 1$ (and 2 bins). We have $F(\theta) = \log(1 + \exp(\theta)) = \mathrm{lse}(0, \theta) := \mathrm{lse}_0^+(\theta)$. We check the strict convexity of $F(\theta)$: $F'(\theta) = \frac{e^\theta}{1+e^\theta}$ and $F''(\theta) = \frac{e^\theta}{(1+e^\theta)^2} > 0$.

We write for short $\mathrm{lse}_0^+(x) = \mathrm{lse}_0^+(x_1, \ldots, x_d)$ for a d-dimensional vector x.

Theorem 3 (lse_0^+ is a Bregman generator) *Multivariate function* $\mathrm{lse}_0^+(x)$ *is a Bregman generator.*

Proof is deferred to Appendix 7.

Lemma 2 (Univariate Monte Carlo Exponential Family Function is a Bregman generator) *Almost surely, the univariate function* $\tilde{F}_S(\theta)$ *is a Bregman generator.*

Proof The first derivative is:

$$\eta = \tilde{F}'_S(\theta) = \frac{\sum_{i=2}^m a_i \exp(a_i\theta + b_i)}{1 + \sum_{i=2}^m \exp(a_i\theta + b_i)} \geq 0, \tag{73}$$

[4]Function lse can be interpreted as a vector function, and is C^2, convex but not strictly convex on $\mathbb{R}^m$. For example, lse is affine on lines since $\mathrm{lse}(x + \lambda 1) = \mathrm{lse}(x) + \lambda$ (or equivalently $\mathrm{lse}(x_1, \ldots, x_m) = \lambda + \mathrm{lse}(x_1 - \lambda, \ldots, x_m - \lambda)$). It is affine only on lines passing through the origin.

and is strictly greater than 0 when there exists at least two elements with distinct sufficient statistics (i.e., $t(x_i) \neq t(x_j)$) so that at least one $a_i > 0$.

The second derivative is:

$$\tilde{F}''_S(\theta) = \frac{\left(\sum_{i=2}^m a_i^2 \exp(a_i\theta + b_i)\right)\left(1 + \sum_{i=2}^m \exp(a_i\theta + b_i)\right) - \left(\sum_{i=2}^m a_i \exp(a_i\theta + b_i)\right)^2}{(1 + \sum_{i=2}^m \exp(a_i\theta + b_i))^2} =: \frac{\text{Num}}{\text{Den}} \quad (74)$$

For each value of $\theta \in \Theta$, we shall prove that $\tilde{F}''_S(\theta) > 0$. Let $c_i = c_i(\theta) = \exp(a_i\theta + b_i) > 0$ for short (θ being fixed, we omit it in the c_i notation in the calculus derivation). Consider the numerator Num since the denominator Den is a non-zero square, hence strictly positive. We have:

$$\text{Num} > \left(\sum_{i=2}^m a_i^2 c_i\right)\left(\sum_{i=2}^m c_i\right) - \left(\sum_{i=2}^m a_i c_i\right)^2, \quad (75)$$

$$\text{Num} > \sum_{ij} a_i^2 c_i c_j - \sum_i a_i^2 c_i^2 - 2\sum_{i<j} a_i a_j c_i c_j, \quad (76)$$

$$\text{Num} > \sum_{i=j} a_i^2 c_i^2 + \sum_{i\neq j} a_i^2 c_i c_j - \sum_i a_i^2 c_i^2 - 2\sum_{i<j} a_i a_j c_i c_j, \quad (77)$$

$$\text{Num} > \sum_{i<j} a_i^2 c_i c_j + \sum_{i>j} a_i^2 c_i c_j - 2\sum_{i<j} a_i a_j c_i c_j, \quad (78)$$

$$\text{Num} > \sum_{i<j} a_i^2 c_i c_j + \sum_{i<j} a_j^2 c_i c_j - 2\sum_{i<j} a_i a_j c_i c_j, \quad (79)$$

$$\text{Num} > \sum_{i<j} (a_i^2 + a_j^2 - 2a_i a_j) c_i c_j, \quad (80)$$

$$\text{Num} > \sum_{i<j} (a_i - a_j)^2 c_i c_j > 0. \quad (81)$$

Therefore the numerator is strictly positive if at least two a_i's are distinct.

Thus we add the following assumption:

Assumption 3 *(AEF1D)* For all $y \in \text{dom}(t)$, $E_q[1_{t(x)=y}] = 0$.

To recap, the MCEFF of the MCIG of uni-order family has the following characteristics:

Monte Carlo Mixture Family Generator 1D:

$$\tilde{F}_{\mathcal{S}}(\theta) = \mathrm{lse}_0^+(a_2\theta + b_2, \ldots, a_m\theta + b_m), \tag{82}$$

$$a_i = t(x_i) - t(x_1), \tag{83}$$

$$b_i = k(x_i) - k(x_1) - \log q(x_i) + \log q(x_1), \tag{84}$$

$$\tilde{F}'_{\mathcal{S}}(\theta) = \frac{\sum_{i=2}^m a_i \exp(a_i\theta + b_i)}{1 + \sum_{i=2}^m \exp(a_i\theta + b_i)} =: \eta, \tag{85}$$

$$\tilde{F}''_{\mathcal{S}}(\theta) = \frac{\left(\sum_{i=2}^m a_i^2 \exp(a_i\theta + b_i)\right)\left(1 + \sum_{i=2}^m \exp(a_i\theta + b_i)\right) - \left(\sum_{i=2}^m a_i \exp(a_i\theta + b_i)\right)^2}{(1 + \sum_{i=2}^m \exp(a_i\theta + b_i))^2} \tag{86}$$

3.2 The general D-Order Case

The difference of sufficient statistics $a_i = t(x_i) - t(x_1)$ is now a vector of dimension D:

$$a_i = \begin{bmatrix} a_i^1 \\ \vdots \\ a_i^D \end{bmatrix}. \tag{87}$$

We replace the scalar multiplication $a_i\theta$ by an inner product $\langle a_i, \theta \rangle$ in Eq. 72, and let $c_i(\theta) = \exp(\langle a_i, \theta \rangle + b_i)$ with $b_i = k(x_i) - k(x_1) - \log q(x_i) + \log q(x_1)$. Then the Monte Carlo Exponential Family Function (MCEFF) writes concisely as:

$$\tilde{F}_{\mathcal{S}}(\theta) = \log\left(1 + \sum_{l=2}^m c_l(\theta)\right), \tag{88}$$

$$:= \mathrm{lse}_0^+(c_2(\theta), \ldots, c_m(\theta)), \tag{89}$$

Theorem 4 (Monte Carlo Exponential Family Function is a Bregman Generator) *Almost surely, the function $\tilde{F}_{\mathcal{S}}(\theta)$ is a proper Bregman generator.*

Proof We have the gradient of first-order partial derivatives:

$$\eta_i = \partial_i \tilde{F}_{\mathcal{S}}(\theta) = \frac{\sum_{l=2}^m a_l^i c_l(\theta)}{1 + \sum_{l=2}^m c_l(\theta)}, \tag{90}$$

and the Hessian matrix of second-order partial derivatives:

$$\partial_i \partial_j \tilde{F}_{\mathcal{S}}(\theta) = \frac{(\sum_{l=2}^m a_l^i a_l^j c_l(\theta))(1+\sum_{l=2}^m c_l(\theta)) - (\sum_{l=2}^m a_l^i c_l(\theta))(\sum_{l=2}^m a_l^j c_l(\theta))}{(1+\sum_{l=2}^m c_l(\theta))^2} =: \frac{\mathrm{Num}}{\mathrm{Den}}. \tag{91}$$

Let us prove that the Hessian matrix $\nabla^2 \tilde{F}_{\mathcal{S}}(\theta) = (\partial_i \partial_j \tilde{F}_{\mathcal{S}}(\theta))_{ij}$ is always symmetric positive semi-definite, and symmetric positive definite almost surely.

Indeed, we have:

$$\mathrm{Num} = \underbrace{\sum_k a_k^i a_k^j c_k}_{:=D} + \underbrace{\sum_{k,l} a_k^i a_k^j c_k c_l - \sum_{k,l} a_k^i c_k a_l^j c_l}_{:=E}. \tag{92}$$

Let us rewrite D as $D = CA^\top A$ with $C = \mathrm{diag}(c_1, \ldots, c_D)$. It follows that matrix D is symmetric positive definite. Let us prove that matrix E is also SPD:

$$E \overset{\star}{=} \sum_{k<l} a_k^i a_k^j c_k c_l + \sum_{l<k} a_k^i z_k^j c_k c_l - \sum_{k<l} a_k^i a_l^j c_k c_l - \sum_{l<k} a_k^i a_l^j c_k c_l, \tag{93}$$

$$\overset{\star\star}{=} \sum_{k<l} \left(a_k^i a_k^j + a_l^i a_l^j - a_k^i a_l^j - a_l^i a_k^j \right) c_k c_l, \tag{94}$$

$$= \sum_{k<l} (a_k^i - a_l^i)(a_k^j - a_l^j) c_k c_l. \tag{95}$$

$\star$: The terms $l = k$ vanish

$\star\star$: After a change of variable $l \leftrightarrow k$ in the second and fourth sums of Eq. 93.

Thus Eq. 95 can be rewritten as $(a_k - a_l)(a_k - a_l)^\top c_k c_l$ where $a_k = \begin{bmatrix} a_k^1 \\ \vdots \\ a_k^D \end{bmatrix}$. It follows that E is a positively weighted sum of rank-1 symmetric positive semi-definite matrices, and is therefore symmetric positive semi-definite.

We want $y^T E y > 0$ for all $y \neq 0 \in \mathbb{R}^D$. Suppose that there exists $y \neq 0 \in \mathbb{R}^D$ such that $y^T E y = 0$. Noting that $a_k^i - a_l^i = t_i(x_k) - t_i(x_l)$, we can write this as

$$\sum_{k<l} \left(\sum_i y_i c_i (t_i(x_k) - t_i(x_l)) \sum_j y_j c_j (t_j(x_k) - t_j(x_l)) \right) = 0, \tag{96}$$

which implies

$$\sum_i y_i c_i \left(t_i(x_k) - t_i(x_l)\right) \sum_j y_j c_j \left(t_j(x_k) - t_j(x_l)\right) = 0, \quad \forall k < l, \tag{97}$$

since each of these terms is non negative. In particular, we have the existence of a $y \neq 0 \in \mathbb{R}^D$ such that

$$\sum_i y_i t_i(x_k) = \sum_i y_i t_i(x_l), \quad \forall y \neq 0, \quad \forall k < l. \tag{98}$$

To get almost surely a Monte Carlo Bregman generator, we introduce the following assumption:

Assumption 4 (*AEF*) The sufficient statistics (t_i) verify that for all $\lambda \neq 0$ and all $y \in dom(\sum_i \lambda_i t_i)$:

$$E_q\left[1_{\sum_i \lambda_i t_i(x)=y}\right] = 0.$$

4 Application to Clustering

In this section, we demonstrate the practical use of MCIG to cluster a set of mixtures in Sect. 4.1, and consider in Sect. 4.2 parallel calculations/aggregations of Monte Carlo Exponential/Mixture Functions.

4.1 Clustering Mixtures on the Mixture Family Manifold

Consider clustering a set of n mixtures $m(x;\eta_1), \ldots, m(x;\eta_n)$ of the mixture family manifold. Prior work considered clustering the mixture components (e.g., Gaussian components) to simplify mixtures by using the Bregman k-means [14, 37]. This prior work can be interpreted as a Gaussian component quantization procedure.

Here, we address the different problem of clustering the mixtures themselves, not their components.

Since $\mathrm{KL}(m(x;\eta_i) : m(x;\eta_j)) = B_G(\eta_i : \eta_j)$ for $G(\eta) = -h(m(x;\eta))$ (Shannon information), we may approximate the KL divergence from the MC Bregman Divergence (MCBD) $\tilde{G}_{\mathcal{S}}$ as follows:

$$\mathrm{KL}(m(x;\eta_i) : m(x;\eta_j)) = B_G(\eta_i : \eta_j), \tag{99}$$

$$\simeq B_{\tilde{G}_{\mathcal{S}}}(\eta_i : \eta_j). \tag{100}$$

One advantage of using a MCIG is that all divergence computations $B_{\tilde{G}_{\mathcal{S}}}$ performed during the execution of a Bregman algorithm are consistent by reusing the same variates of $\mathcal{S}$. In particular, this also guarantees to always have nonnegative estimated KL divergences.

The traditional way to MC estimate the KL divergence is to consider the MC stochastic integration of the extended Kullback–Leibler divergence [4]:

$$\widehat{\mathrm{eKL}}_m(p:q) := \frac{1}{m}\sum_{i=1}^{m}\left(\log\frac{p(x_i)}{q(x_i)} + \frac{q(x_i)}{p(x_i)} - 1\right), \tag{101}$$

for $x_1, \ldots, x_m \sim p(x)$. Indeed, if we just used the MC KL estimator:

$$\widehat{\mathrm{KL}}_m(p:q) := \frac{1}{m}\sum_{i=1}^{m}\log\frac{p(x_i)}{q(x_i)}, \tag{102}$$

we may endup with negative values to our estimated KL, depending on the sample variates! This never happens for eKL which is a statistical divergence for the scalar divergence $\mathrm{ekl}(p:q) = p\log\frac{p}{q} + q - p \geq 0$.

Bregman k-means [4, 22] can be applied using either the sided or their symmetrized centroid [40]: The right-sided centroid is always the center of mass of the parameters. The left-sided centroid requires to compute $F'(\theta)$ and its reciprocal inverse function $(F'(\theta))^{-1}$ (wlog, assuming $D = 1$ for simplicity[5]). Although $F'(\theta)$ is available in closed form (and define the dual parameter θ):

$$\tilde{G}'_{\mathcal{S}}(\eta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{q(x_i)}\left(p_1(x_i) - p_0(x)\right)\left(1 + \log m(x;\eta)\right) = \theta, \tag{103}$$

the dual parameter of $(\mathcal{M}, G)$ cannot be written as a simple function $\eta = F^{*\prime}(\eta)$. Notice that $\theta = \tilde{G}'_{\mathcal{S}}(\eta)$ is an increasing function of η and that the inverting operation can be performed numerically. Indeed, we can compute $\eta = (\tilde{G}'_{\mathcal{S}})^{-1}(\theta) = \tilde{G}^*_{\mathcal{S}}(\theta)$ using a numerical scheme (e.g., bisection search).

The symmetric Jeffreys divergence is:

$$\begin{aligned}
J(m(x;\eta_i) : m(x;\eta_j)) &= \mathrm{KL}(m(x;\eta_i) : m(x;\eta_j)) + \mathrm{KL}(m(x;\eta_j) : m(x;\eta_i)), && (104)\\
&= B_G(\eta_i : \eta_j) + B_G(\eta_j : \eta_i), && (105)\\
&= B_G(\eta_i : \eta_j) + B_{G^*}(\theta_i : \theta_j), && (106)\\
&= \langle \Delta\theta_{ij}, \Delta\eta_{ij}\rangle, && (107)
\end{aligned}$$

where $\Delta\theta_{ij} = \theta_i - \theta_j$ and $\Delta\eta_{ij} = \eta_i - \eta_j$.

We may approximate the J divergence by considering the Monte Carlo Bregman generator in Eq. 105:

$$J(m(x;\eta_i) : m(x;\eta_j)) \simeq B_{\tilde{G}_{\mathcal{S}}}(\eta_i : \eta_j) + B_{\tilde{G}_{\mathcal{S}}}(\eta_j : \eta_i). \tag{108}$$

We can then apply the technique of mixed Bregman clustering [49] that considers two centers per cluster. Moreover a fast probabilistic initialization, called *mixed Bregman k-means++* [49], allows one to guarantee a good initialization with high probability (without computing centroids but requiring to compute divergences).

[5] Otherwise, we need to consider monotone operator theory [25] to invert $\nabla F(\theta)$.

Another technique to bypass the computation of the gradient $\nabla \tilde{G}_{\mathcal{S}}$ in the BD consists in taking the scaled skew α-Jensen divergence [35] for an infinitesimal value of α. Indeed, we have the α-Jensen divergence defined by:

$$J_F^{\alpha}(p:q) = (1-\alpha)F(p) + \alpha F(q) - F((1-\alpha)p + \alpha q), \tag{109}$$

and asymptotically this skewed Jensen divergences yield the sided Bregman divergences [35] as follows:

$$\lim_{\alpha \to 0^+} \frac{J_F^{\alpha}(p:q)}{\alpha} = B_F(q:p), \tag{110}$$

$$\lim_{\alpha \to 1^-} \frac{J_F^{\alpha}(p:q)}{1-\alpha} = B_F(p:q), \tag{111}$$

Thus we have for small values of $\alpha > 0$ (say, $\alpha = 0.001$):

$$J(m(x;\eta_i) : m(x;\eta_j)) = B_G(\eta_i : \eta_j) + B_G(\eta_j : \eta_i), \tag{112}$$

$$\simeq \frac{1}{\alpha} J^{\alpha}_{\tilde{G}_{\mathcal{S}}}(\eta_i : \eta_j) + \frac{1}{1-\alpha} J^{1-\alpha}_{\tilde{G}_{\mathcal{S}}}(\eta_i : \eta_j). \tag{113}$$

The last equation Eq.113 is the symmetrized skew Jensen divergence studied in [29].

Figure 6 plots the result of a 2-cluster clustering wrt the Jeffreys' divergence for a set of $n = 8$ mixtures.

4.2 Parallelizing Information Geometry

We can distribute the Monte Carlo information geometry either on a multicore machine with l cores with shared memory or on a cluster of l machines with distributed memory, or even consider hybrid architectures.

Let $(M, \tilde{F}_{\mathcal{S}_1}), \ldots, (M, \tilde{F}_{\mathcal{S}_l})$ be a set of l information-geometric manifolds obtained from iid sample sets $\mathcal{S}_1, \ldots, \mathcal{S}_l$. Let $\oplus_{i=1}^{s} \mathcal{S}_i$ be a partition of $\mathcal{S}$.

4.2.1 Multicore Architectures

On a multicore architecture, we may evaluate the mixture family Bregman divergence $B_{\tilde{G}_{\mathcal{S}}}(\eta : \eta')$ by evaluating $B_{\tilde{G}_{\mathcal{S}_i}}(\theta : \theta')$, and using the compositionality rule of Bregman generators in BDs (Property 2) with:

$$\tilde{G}_{\mathcal{S}}(\theta) = \sum_{i=1}^{l} \frac{|\mathcal{S}_i|}{|\mathcal{S}|} \tilde{G}_{\mathcal{S}_i}(\eta). \tag{114}$$

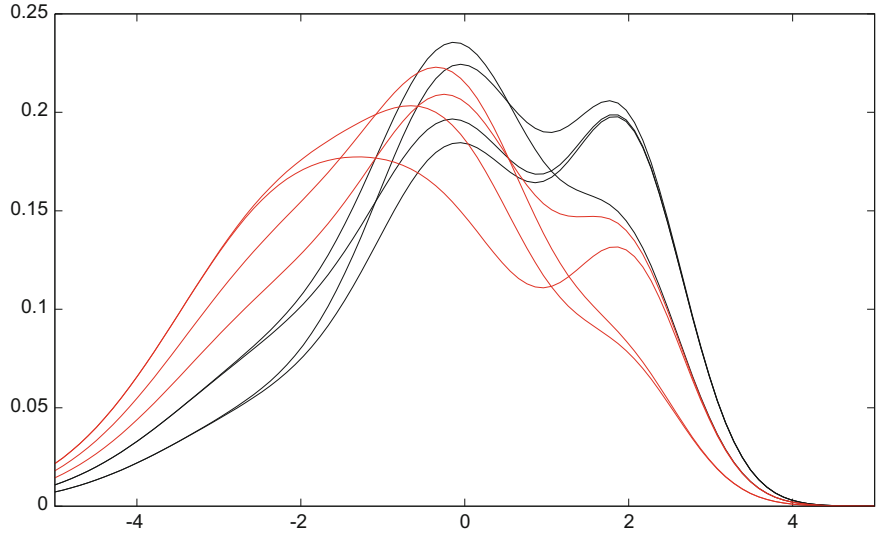

Fig. 6 Clustering a set of $n = 8$ statistical mixtures of order $D = 2$ with $K = 2$ clusters: Each mixture is represented by a 2D point on the mixture family manifold. The Kullback–Leibler divergence is equivalent to an integral-based Bregman divergence that is computationally untractable: The Bregman generator is stochastically approximated by Monte Carlo sampling

That is, $\tilde{G}_{\mathcal{S}}(\eta)$ is the *arithmetic weighted mean* of the mixture sub-generators.

For the exponential families, recall that we have:

$$\tilde{F}_{\mathcal{S}}(\theta) = \log\left(\sum_{i=1}^{s} \frac{|\mathcal{S}_i|}{|\mathcal{S}|} \exp(\tilde{F}_{\mathcal{S}_i})\right). \tag{115}$$

That is, $\tilde{F}_{\mathcal{S}}(\theta)$ can be interpreted as an *exponential mean* (quasi-arithmetic mean, called f-mean [35] for the monotonically increasing function $f(x) = \exp(x)$) of the sub-generators. Thus we can perform the computation of the MC Bregman generators on multi-core architectures easily with a MapReduce strategy [33].

Fact 1 (MapReduce evaluation of MC Bregman generators) *The MCMF or MCEF functions can be computed in parallel using a quasi-arithmetic mean MapReduce operation.*

4.2.2 Cluster Architectures

Since the MC Bregman generators can be interpreted as random variables $\tilde{G}_m(\theta)$ and $\tilde{F}_m(\theta)$, we may obtain robust estimate [51] by carrying the calculations on l MCIGs on a cluster architecture, and then integrate those l geometries.

Given a sequence of matching parameters $\theta_1 \in (M, \tilde{F}_{s_1}), \ldots, \theta_l \in (M, \tilde{F}_{s_l})$, we aggregate these parameters by doing the *KL-averaging* method [26]. This amounts to compute a sided centroid for θ.

5 Information-Geometric Structures Induced by Statistical Separable Divergences

In this section, we consider Monte Carlo sampling to define a (tractable) statistical divergence that approximates another (untractable) statistical divergence, and uses this MC statistical divergence to define an information-geometric manifold.

The core structure of information geometry [1] is a manifold M equipped with a pair of dual connections, ∇ and ∇^* coupled to the metric tensor g: (M, g, ∇, ∇^*). In terms of differential geometry, the definition of this coupling is expressed as

$$X\langle Y, Z\rangle_g = \langle \nabla_X Y, Z\rangle_g + \langle Y, \nabla_X^* Z\rangle_g, \tag{116}$$

where X, Y and Z are smooth vector fields on M. The coupling of connections to the metric tensor means that the dual parallel transport is compatible with the metric:

$$\langle u, v\rangle_{c(0)} = \left\langle \prod_{c(0)\to c(t)}^{\nabla} u, \prod_{c(0)\to c(t)}^{\nabla^*} v \right\rangle_{c(t)}, \tag{117}$$

where c is a smooth curve (parallel transport is path dependent, except for dually flat connections). The notation $\prod_{c(0)\to c(t)}^{\nabla} u$ means that vector $u \in T_{c(0)} = T_p$ is parallel transported along smooth curve c to tangent plane $T_{c(t)}$ with respect to the affine connection ∇. From this (M, g, ∇, ∇^*) structure, a statistical manifold [24] (M, g, C) can be defined, where $C(X, Y, Z) = \langle \nabla_X Y - \nabla_X^* Y, Z\rangle$ is a totally symmetric cubic tensor, termed the Amari–Chentsov cubic tensor. It follows a one-parameter family of dual connections [1] (with ∇^0 being the Levi-Civita metric connection): $(M, g, \nabla^{-\alpha}, \nabla^{\alpha})$ so that if connection ∇^{α} has constant curvature κ then its dual connection has also the same curvature. Furthermore, one can build [1, 16, 17] a pair of dual connections coupled to a metric from any smooth divergence D: $(M, {}^D g, {}^D\nabla, {}^D\nabla^*)$. Figure 7 summarizes the fundamental structures of parametric information geometry and their relationships.

Let us consider a separable statistical divergence:

$$D[p : q] := \int d(p(x) : q(x))\mathrm{d}\mu(x), \tag{118}$$

where $d(x : y)$ is a scalar divergence. For example, the f-divergences [1] are obtained for $i_f(x : y) = xf(x/y)$:

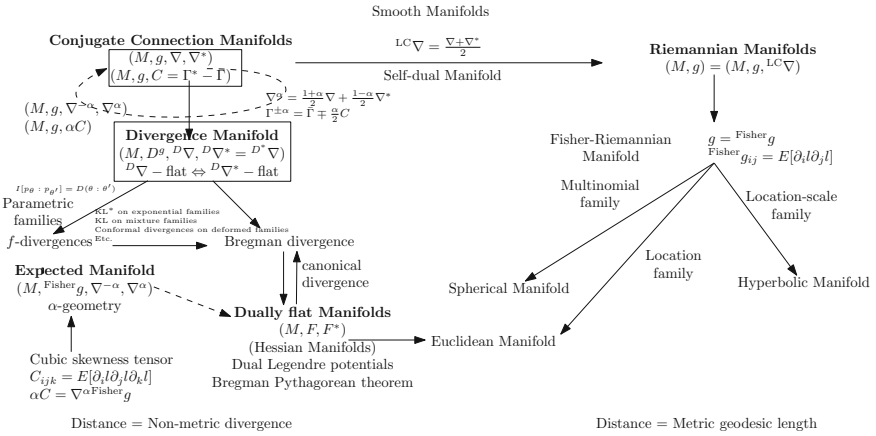

Fig. 7 The web of fundamental information-geometric structures. An arrow $a \rightarrow b$ means that geometric structure b is a special case of the (meta-)structure a

$$I_f[p : q] := \int p(x) f\left(\frac{p(x)}{q(x)}\right) \mathrm{d}\mu(x) = \int i_f(p(x) : q(x)) \mathrm{d}\mu(x), \tag{119}$$

The f-divergences are the only statistical separable divergences that satisfy the information monotonicity property [1]. On a parametric family of distributions $\{p_\theta\}$, the statistical f-divergences amount to equivalent parameter divergences:

$$D_f(\theta_1 : \theta_2) := I_f[p_{\theta_1} : p_{\theta_2}] \tag{120}$$

The information-geometric structure induced by this (parameter) divergence is $(M, {}^{D_f}g, {}^{D_f}\nabla, {}^{D_f}\nabla^*)$, and the dual connections correspond to the expected α-connections[1] for f-divergences.

It may happen that D_f, although well-defined, may not be available in closed form. In that case, we approximate the divergence by Monte Carlo stochastic integration by drawing a set $\mathcal{S}_m = \{x_1, \ldots, x_m\}$ of m iid variates from p_{θ_1}:

$$\tilde{D}_{\mathcal{S}_m}(\theta_1 : \theta_2) := \frac{1}{m} \sum_{i=1}^{m} \frac{1}{p_{\theta_1}(x_i)} d(p_{\theta_1}(x_i) : p_{\theta_2}(x_i)). \tag{121}$$

We need to assert that $\tilde{D}_{\mathcal{S}_m}$ is a smooth divergence: The smoothness of the divergence $\tilde{D}_{\mathcal{S}_m}$ follows from the smoothness divergence of the corresponding scalar divergence d. Then we need to guarantee that $\tilde{D}_{\mathcal{S}_m}(\theta_1 : \theta_2) = 0$ iff $\theta_1 = \theta_2$. Since $d(p_{\theta_1}(x) : p_{\theta_2}(x)) = 0$ if and only if $p_{\theta_1}(x) = p_{\theta_2}(x)$, we need to assert that with high probability $p_{\theta_1}(x) \neq p_{\theta_2}(x)$ when $\theta_1 \neq \theta_2$. Let $I = \max_{\theta_1, \theta_2} \mu(\{p_{\theta_1}(x) = p_{\theta_2}(x), x \in \mathcal{X}\})$. When $I = 0$, then almost surely $\tilde{D}_{\mathcal{S}}$ is a divergence. This condition holds when the probability densities intersect in at most a finite number of points. It

follows the corresponding information-geometric structure $(M, {}^{\tilde{D}_{S_m}}g, {}^{\tilde{D}_{S_m}}\nabla, {}^{\tilde{D}_{S_m}}\nabla^*)$ (with its associated one-family of α-connections) such that asymptotically, we have:

$$\lim_{m\to\infty} (M, {}^{\tilde{D}_{S_m}}g, {}^{\tilde{D}_{S_m}}\nabla, {}^{\tilde{D}_{S_m}}\nabla^*) = (M, {}^{D}g, {}^{D}\nabla, {}^{D}\nabla^*), \tag{122}$$

as desired.

Let us quickly report two examples to illustrate these divergence-based sequences of information-geometric structures:

- Polynomial Exponential Families (PEFs) of order D with the γ-divergences [50]: Let us notice that we do not need to normalize the PEF distributions in order to sample variates, and that the γ-divergence D_γ is a projective divergence [48] (invariant by positive rescaling of the distributions) which tends to the KL divergence when $\gamma \to 0$. Since the densities of any two distinct PEF distributions of order D intersect in at most $D + 1$ points, we check that $I = 0$. Thus for $\gamma \to 0$ and $m \to \infty$, we tend to a dually flat manifold. As an application, we can consider clustering these PEFs on a MCIG manifold.
- Consider a mixture family $\{m_\eta(x) = (1 - \eta)p_1(x) + \eta p_2(x), \eta \in (0, 1)\}$ of order $D = 1$ for the two mixture component distributions p_1 and p_2, linearly independent. We have $m_{\eta_1}(x) = m_{\eta_2}(x)$ iff $p_1(x) = p_2(x)$ (holds only for this particular case of $D = 1$). Assume $I = 0$ for the component distributions, then we obtain a sequence of Monte Carlo information-geometric structures that tend asymptotically to the dually flat mixture manifold.

In the later case, we consider the MCIG manifold for a 1D mixture manifold with respect to an arbitrary divergence. Notice that the divergence-based MCIG for the exponential/mixture manifold may not be flat for KL. In Sect. 2.2, we took the different approach of approximating the negative differential entropy via Monte-Carlo, ensuring that all sequence of MCIG manifolds are dually flat.

6 Conclusion and Perspectives

In this work, we proposed a new type of *randomized information-geometric structure* to cope with computationally untractable information-geometric structures (types 4 and 5 in the classification of Table 4): Namely, the Monte Carlo Information Geometry [38] (MCIG). MCIG performs stochastic integration of the ideal but computationally intractable definite integral-based Bregman generator (e.g. Eq. 32 for mixture family) for mixture family and Eq. 24 for exponential family). We proved that the MC Bregman generators for the mixture family and the exponential family are almost surely strictly convex and differentiable (Theorem 2 and Theorem 4, respectively), and therefore yield a computationally tractable information-geometric structure (Type 2 in the classification of Table 4). Thus we can get a series of *consistent* and *computationally-friendly* information-geometric structures that tend asymptotically

Table 4 A smooth and strictly convex function F induces a dually flat structure: We classify those structures according to their computational tractability properties

Type	F	∇F^*	Example
Type 1	Closed-form	Closed-form	Gaussian (exponential) family
Type 2	Closed-form	Not closed-form	Beta (exponential) family
Type 3	Comp. intractable	Not closed-form	Ising family [54]
Type 4	Not closed-form	Not closed-form	Polynomial exponential family [42]
Type 5	Not analytic	Not analytic	Mixture family

to the untractable ideal information geometry. We have demonstrated the usefulness of our technique for a basic Bregman k-means clustering technique: Clustering statistical mixtures on a mixture family manifold. Although the MCIG structures are computationally convenient, we do not have in closed-form ∇F^* (nor F^*) because our Bregman generators are the sum of basic generators whose gradients are the sum of elementary gradients that cannot be inverted easily.[6] This step requires a numerical or symbolic technique [25].

We note that in the recent work of [27], Matsuzoe et al. defined a sequence of statistical manifolds relying on a sequential structure of escort expectations for non-exponential type statistical models.

Codes for reproducible results are available at:

https://franknielsen.github.io/MCIG/

7 Function $\mathrm{lse}_0^+(x)$ is a Bregman Generator

We give the proof of Theorem 3:

Proof Since $\mathrm{lse}_0^+(x_1,\ldots,x_d) = \log\left(1+\sum_{i=1}^d \exp(x_i)\right)$ is twice continuously differentiable, it suffices to prove that $\nabla^2 \mathrm{lse}_0^+(x) \succ 0$. We have:

$$\partial_i \mathrm{lse}_0^+(x) = \frac{e^{x_i}}{1+\sum_k e^{x_k}}, \tag{123}$$

$$\partial_j \partial_i \mathrm{lse}_0^+(x) \overset{j\neq i}{=} \frac{-e^{x_i}e^{x_j}}{(1+\sum_k e^{x_k})^2}, \tag{124}$$

$$\partial_i \partial_i \mathrm{lse}_0^+(x) = \frac{e^{x_i}(1+\sum_k e^{x_k}) - e^{x_i}e^{x_j}}{(1+\sum_k e^{x_k})^2}. \tag{125}$$

[6] The Legendre conjugate of an infimal convolution of elementary functions is the sum of the elementary conjugate functions.

It follows that the Hessian $(\partial_j \partial_i \mathrm{lse}_0^+(x))_{ij}$ is a diagonally dominant matrix since:

$$e^{x_i}\left(1+\sum_k e^{x_k}\right) = e^{x_i} + e^{x_i}\sum_k e^{x_k} > \sum_{j\neq i}\left|-e^{x_i}e^{x_j}\right| = e^{x_i}\sum_{j\neq i} e^{x_j}. \quad (126)$$

To conclude that the Hessian matrix is SPD, we use Gershgorin circle theorem [55] to bound the spectrum of a square matrix: The eigenvalues of the Hessian matrix are thus real and fall inside a disk of center $(e^{x_i}(1+\sum_k e^{x_k}))_i$ and radius $e^{x_i}\sum_{j\neq i} e^{x_j}$. Therefore all eigenvalues are positive, and the Hessian matrix is positive definite.

For $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, we have:

$$\nabla \mathrm{lse}(x) = \sigma(x), \quad (127)$$

where $\sigma(x)$ is the *softmax* function:

$$\sigma(x) := \left(\frac{e^{x_i}}{\sum_{k=1}^d e^{x_k}}\right)_{i\in\{1,\ldots,d\}}. \quad (128)$$

By analogy, we may define for $x \in \mathbb{R}^d$:

$$\sigma_0^+(x) := \left(\frac{e^{x_i}}{1+\sum_k e^{x_k}}\right)_{i\in\{1,\ldots,d\}}, \quad (129)$$

so that $\nabla \mathrm{lse}_0^+(x) = \sigma_0^+(x)$.

References

1. Amari, S.: Information Geometry and Its Applications. Applied Mathematical Sciences. Springer, Japan (2016)
2. Amari, Si, Cichocki, A.: Information geometry of divergence functions. Bull. Polish Acad. Sci.: Tech. Sci. **58**(1), 183–195 (2010)
3. Azoury, K.S., Warmuth, M.K.: Relative loss bounds for on-line density estimation with the exponential family of distributions. Mach. Learn. **43**(3), 211–246 (2001)
4. Banerjee, A., Merugu, S., Dhillon, I.S., Ghosh, J.: Clustering with Bregman divergences. J. Mach. Learn. Res. **6**(Oct), 1705–1749 (2005)
5. Bhattacharya, B.B., Mukherjee, S., et al.: Inference in Ising models. Bernoulli **24**(1), 493–525 (2018)
6. Boissonnat, J.D., Nielsen, F., Nock, R.: Bregman Voronoi diagrams. Discret. Comput. Geom. **44**(2), 281–307 (2010)
7. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
8. Calin, O., Udriste, C.: Geometric Modeling in Probability and Statistics. Mathematics and Statistics. Springer International Publishing, Berlin (2014)

9. Cipra, B.A.: The Ising model is NP-complete. SIAM News **33**(6), 1–3 (2000)
10. Cobb, L., Koppstein, P., Chen, N.H.: Estimation and moment recursion relations for multimodal distributions of the exponential family. J. Am. Stat. Assoc. **78**(381), 124–130 (1983)
11. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, New York (2012)
12. Critchley, F., Marriott, P.: Computational information geometry in statistics: theory and practice. Entropy **16**(5), 2454–2471 (2014)
13. Crouzeix, J.P.: A relationship between the second derivatives of a convex function and of its conjugate. Math. Programm. **13**(1), 364–365 (1977)
14. Davis, J.V., Dhillon, I.S.: Differential entropic clustering of multivariate gaussians. In: Advances in Neural Information Processing Systems, pp. 337–344 (2007)
15. Dawid, A.P.: The geometry of proper scoring rules. Ann. Inst. Stat. Math. **59**(1), 77–93 (2007)
16. Eguchi, S.: Second order efficiency of minimum contrast estimators in a curved exponential family. Ann. Stat. **11**, 793–803 (1983)
17. Eguchi, S.: Geometry of minimum contrast. Hiroshima Math. J. **22**(3), 631–647 (1992)
18. Fleisch, D.A.: A Student's Guide to Vectors and Tensors. Cambridge University Press, Cambridge (2011)
19. Frongillo, R., Reid, M.D.: Convex foundations for generalized MaxEnt models. In: AIP Conference Proceedings, vol. 1636, pp. 11–16. AIP (2014)
20. Gao, B., Pavel, L.: On the Properties of the Softmax Function with Application in Game Theory and Reinforcement Learning. ArXiv e-prints (2017)
21. Geman, S., Graffigne, C.: Markov random field image models and their applications to computer vision. In: Proceedings of the International Congress of Mathematicians, vol. 1, p. 2 (1986)
22. Grønlund, A., Larsen, K.G., Mathiasen, A., Nielsen, J.S.: Fast exact k-means, k-medians and Bregman divergence clustering in 1D (2017). arXiv:1701.07204
23. Kass, R.E., Vos, P.W.: Geometrical Foundations of Asymptotic Inference. Fisher-Rao metric of location-scale family is hyperbolic (and can be diagonalized), pp. 192–193. Wiley-Interscience (1997)
24. Lauritzen, S.L.: Statistical manifolds. Differential Geometry in Statistical Inference, p. 164 (1987)
25. Lauster, F., Luke, D.R., Tam, M.K.: Symbolic computation with monotone operators. Set-Valued and Variational Analysis, pp. 1–16 (2017)
26. Liu, Q., Ihler, A.T.: Distributed estimation, information loss and exponential families. In: Advances in Neural Information Processing Systems, pp. 1098–1106 (2014)
27. Matsuzoe, H., Scarfone, A.M., Wada, T.: A sequential structure of statistical manifolds on deformed exponential family. In: International Conference on Geometric Science of Information, pp. 223–230. Springer (2017)
28. Mitchell, A.F.S.: Statistical manifolds of univariate elliptic distributions. International Statistical Review/Revue Internationale de Statistique, pp. 1–16 (1988)
29. Nielsen, F.: A family of statistical symmetric divergences based on Jensen's inequality (2010). arXiv:1009.4004
30. Nielsen, F.: Legendre transformation and information geometry (2010)
31. Nielsen, F.: Hypothesis testing, information divergence and computational geometry. In: Geometric Science of Information, pp. 241–248. Springer (2013)
32. Nielsen, F.: An information-geometric characterization of Chernoff information. IEEE Signal Process. Lett. **20**(3), 269–272 (2013)
33. Nielsen, F.: Introduction to HPC with MPI for Data Science. Undergraduate Topics in Computer Science. Springer (2016). https://doi.org/10.1007/978-3-319-21903-5. https://doi.org/10.1007/978-3-319-21903-5
34. Nielsen, F., Boissonnat, J.D., Nock, R.: On Bregman Voronoi diagrams. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, pp. 746–755 (2007)
35. Nielsen, F., Boltz, S.: The Burbea-Rao and Bhattacharyya centroids. IEEE Trans. Inf. Theory **57**(8), 5455–5466 (2011)

36. Nielsen, F., Garcia, V.: Statistical exponential families: a digest with flash cards (2009) arXiv:0911.4863
37. Nielsen, F., Garcia, V., Nock, R.: Simplifying Gaussian mixture models via entropic quantization. In: 17th European Conference on Signal Processing (EUSIPCO), pp. 2012–2016. IEEE (2009)
38. Nielsen, F., Hadjeres, G.: Monte Carlo information geometry: the dually flat case (2018). CoRR arXiv:1803.07225
39. Nielsen, F., Nock, R.: On the smallest enclosing information disk. Inf. Process. Lett. **105**(3), 93–97 (2008)
40. Nielsen, F., Nock, R.: Sided and symmetrized Bregman centroids. IEEE Trans. Inf. Theory **55**(6), 2882–2904 (2009)
41. Nielsen, F., Nock, R.: Optimal interval clustering: application to Bregman clustering and statistical mixture learning. IEEE Signal Process. Lett. **21**(10), 1289–1292 (2014)
42. Nielsen, F., Nock, R.: Patch matching with polynomial exponential families and projective divergences. In: International Conference on Similarity Search and Applications, pp. 109–116. Springer (2016)
43. Nielsen, F., Nock, R.: On w-mixtures: finite convex combinations of prescribed component distributions (2017). CoRR arXiv:1708.00568
44. Nielsen, F., Nock, R.: On the geometric of mixtures of prescribed distributions. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2018)
45. Nielsen, F., Piro, P., Barlaud, M.: Bregman vantage point trees for efficient nearest neighbor queries. In: 2009 IEEE International Conference on Multimedia and Expo, ICME 2009, pp. 878–881. IEEE (2009)
46. Nielsen, F., Piro, P., Barlaud, M.: Tailored Bregman ball trees for effective nearest neighbors. In: Proceedings of the 25th European Workshop on Computational Geometry (EuroCG), pp. 29–32 (2009)
47. Nielsen, F., Sun, K.: Guaranteed bounds on information-theoretic measures of univariate mixtures using piecewise log-sum-exp inequalities. Entropy **18**(12), 442 (2016)
48. Nielsen, F., Sun, K., Marchand-Maillet, S.: On hölder projective divergences. Entropy **19**(3), 122 (2017)
49. Nock, R., Luosto, P., Kivinen, J.: Mixed Bregman clustering with approximation guarantees. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 154–169. Springer (2008)
50. Notsu, A., Komori, O., Eguchi, S.: Spontaneous clustering via minimum γ-divergence. Neural Comput. **26**(2), 421–448 (2014)
51. Pelletier, B.: Informative barycentres in statistics. Ann. Inst. Stat. Math. **57**(4), 767–780 (2005)
52. Robert, C.P.: Monte Carlo methods. Wiley Online Library (2004)
53. Shima, H.: The Geometry of Hessian Structures. World Scientific, Singapore (2007)
54. Tang, Y., Salakhutdinov, R.R.: Learning stochastic feedforward neural networks. In: Advances in Neural Information Processing Systems, pp. 530–538 (2013)
55. Varga, R.S.: Geršgorin and His Circles, vol. 36. Springer Science & Business Media, Berlin (2010)
56. Watanabe, S., Yamazaki, K., Aoyagi, M.: Kullback information of normal mixture is not an analytic function. Technical report of IEICE (in Japanese) (2004-0), pp. 41–46 (2004)
57. Zhang, J.: Reference duality and representation duality in information geometry. In: AIP Conference Proceedings, vol. 1641, pp. 130–146. AIP (2015)