arXiv:1803.07225v1 [cs.LG] 20 Mar 2018

Monte Carlo Information Geometry: The dually flat case

Frank Nielsen[1]* Gaëtan Hadjeres[2]†

[1] Sony Computer Science Laboratories Inc, Tokyo, Japan
[2] Sony Computer Science Laboratories, Paris, France

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.

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 [7, 1] in §1.1, define the statistical manifolds of exponential families and mixture families in §1.2, and discuss about the computational tractability of Bregman algorithms in dually flat spaces in §1.3.

1.1 Dually flat space: Bregman geometry

A smooth (potentially asymmetric) distance $D(\cdot,\cdot)$ is called a *divergence* in information geometry [7, 1], and induces a differential-geometric dualistic structure [15, 2, 7, 1]. 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) [3] $B_F(\cdot,\cdot)$ 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. We use the standard notational convention of information geometry [7, 1]: $\partial_i :=: \frac{\partial}{\partial \theta^i}$

*E-mail: `Frank.Nielsen@acm.org`

†E-mail: `gaetan.hadjeres@sony.com`

to indicate a contravariant vector [16] $\theta = (\theta^i)_i$. (The :=: symbol means it is a notational convention equality, like $\sum_{i=1}^k x_i$:=: $x_1 + \ldots x_k$. It differs from $a := b$ that denotes the symbol by of a quantity equality by definition.)

The Legendre-Fenchel transformation [27] :

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

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 convert between these two dual coordinates as follows:

$$\eta = \eta(\theta) = \nabla F(\theta) = (\partial_i F(\theta))_i, \tag{3}$$

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

with reciprocal gradients $\nabla F^* := (\nabla F)^{-1}$. We used the notational convention ∂^i :=: $\frac{\partial}{\partial \eta_i}$ that indicates the covariant vector [16] $\eta = (\eta_i)_i$.

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 [48]:

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

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

and 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, we have:

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

$$= \eta(v)^\top G^*(\eta) \eta(w), \tag{8}$$

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 $\{{e^*}^i\}_i$ is a dual basis [16] 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 [12]:

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

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 [52] (with $\{e_i\}_i$ and $\{{e^*}^i\}_i$ forming a dual basis [16] 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:

$$\frac{\partial \eta_i}{\partial \theta^j} = g_{ij}(\theta) = \langle e_i, e_j \rangle, \tag{11}$$

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

$$\tag{13}$$

The convex conjugate functions $F(\theta)$ and $F^*(\eta)$ are called *dual potential functions*, and define the global metric [48].

Table 1 summarizes the differential-geometric structures of dually flat spaces. Since Bregman divergences are *canonical divergences* of dually flat spaces [1], the geometry of dually flat spaces is also referred to the *Bregman geometry* [14] in the literature.

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: 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^*.

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 [3] of Bregman generators:

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 [3], minimum enclosing balls [36], hypothesis testing [28] and Chernoff information [29], Voronoi diagrams [31, 5], proximity data-structures [42, 43], 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:

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

$$(PQ)^*_\lambda = \left\{ \begin{array}{l} \eta((PQ)^*_\lambda) = \eta(P) + \lambda(\eta(Q) - \eta(P)), \\ \theta((PQ)^*_\lambda) = \nabla F^*(\eta((PQ)^*_\lambda)) = \nabla F^*(\nabla F(\theta(P)) + \lambda(\nabla F(\theta(Q)) - \nabla F(\theta(P)))) \end{array} \right. \tag{15}$$

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 [3, 38, 19] 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 review quickly the dualistic structure of statistical manifolds [22].

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 (wrt) a base measure μ defined on the support $\mathcal{X}$, is defined as:

$$I(P : Q) = \int_{x \in \mathcal{X}} I_1\left(p(x) : q(x)\right) \mathrm{d}\mu(x). \tag{16}$$

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)	×	✓	✓	✓

Table 2: Some fundamental Bregman clustering algorithms [3, 38, 19] (of the Bregman algorithmic toolbox) that illustrate which closed-form are required to be run in practice.

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:

$$\mathrm{KL}(P:Q) := \int_{x\in\mathcal{X}} \mathrm{kl}(p(x):q(x))\mathrm{d}\mu(x), \tag{17}$$

with corresponding scalar divergence:

$$\mathrm{kl}(x:y) := x\log\frac{x}{y}. \tag{18}$$

The KL divergence between P and Q is also called the *relative entropy* [10] 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:

$$\begin{aligned}
\mathrm{KL}(P:Q) &= h^\times(P:Q)-h(P), && (19)\\
h^\times(P:Q) &:= \int_{x\in\mathcal{X}} p(x)\log\frac{1}{q(x)}\mathrm{d}\mu(x), && (20)\\
h(P) &:= \int_{x\in\mathcal{X}} p(x)\log\frac{1}{p(x)}\mathrm{d}\mu(x) = h^\times(P:P). && (21)
\end{aligned}$$

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 [7, 1], 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 $D(\cdot:\cdot)$ [15, 2, 7, 1] as follows:

$$\begin{aligned}
g_{ij}(x) &= \left.\frac{\partial^2}{\partial x_i\partial x_j}D(x:y)\right|_{y=x}, && (22)\\
\Gamma_{ijk}(x) &= -\left.\frac{\partial^3}{\partial x_i\partial x_j\partial y_k}D(x:y)\right|_{y=x}. && (23)
\end{aligned}$$

The *dual divergence* $D^*(p:q):=D(q:p)$ highlights the *reference duality* [52], and the dual connection ∇^* is induced by the dual divergence $D^*(\cdot:\cdot)$ (∇^* is defined by $\Gamma^*_{ijk}(x)=-\frac{\partial^3}{\partial x_i\partial x_j\partial y_k}D^*(x:y)\Big|_{y=x}$). 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.

	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 $g = (g_{ij})_{ij}$	$g_{ij}(\theta) = \partial_i \partial_j F(\theta)$ $g = \mathrm{Var}[t(X)]$	$g_{ij}(\eta) = \int_{\mathcal{X}} \frac{f_i(x) f_j(x)}{m(x;\eta)} \mathrm{d}\mu(x)$
		$g_{ij}(\eta) = -\partial_i \partial_j h(\eta)$
Christoffel symbol	$\Gamma_{ij,k} = \frac{1}{2} \partial_i \partial_j \partial_k F(\theta)$	$\Gamma_{ij,k} = -\frac{1}{2} \int_{\mathcal{X}} \frac{f_i(x) f_j(x) f_k(x)}{m^2(x;\eta)} \mathrm{d}\mu(x)$
Entropy	$-F^*(\eta)$	$-F^*(\eta)$
Kullback-Leibler divergence	$B_F(\theta_2 : \theta_1)$ $= B_{F^*}(\eta_1 : \eta_2)$	$B_{F^*}(\eta_1 : \eta_2)$ $= B_F(\theta_2 : \theta_1)$

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

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:*

$$\mathcal{E}_{t,\mu} := \{p(x;\theta) \propto \exp(\langle t(x), \theta\rangle)\}_\theta, \tag{24}$$

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

The densities are normalized by the cumulant function F:

$$F(\theta) := \log\left(\int_{x\in\mathcal{X}} \exp(\langle t(x), \theta\rangle) \mathrm{d}\mu(x)\right), \tag{25}$$

so that:

$$p(x;\theta) = \exp(\langle t(x), \theta\rangle - F(\theta)). \tag{26}$$

Function F is a Bregman generator on the natural parameter space:

$$\Theta := \left\{\theta : \int_{x\in\mathcal{X}} \exp(\langle t(x), \theta\rangle) \mathrm{d}\mu(x) < \infty\right\}. \tag{27}$$

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 [33]:

$$\mathcal{E}_{t,k,\nu} := \{p(x;\theta) \propto \exp(\langle t(x), \theta\rangle + k(x)) \,:\, \theta \in \Theta\}. \tag{28}$$

We call 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:

$$\begin{aligned} \mathrm{KL}(p(x;\theta_1) : p(x;\theta_2)) &= B_F(\theta_2 : \theta_1), & (29)\\ &= B_{F^*}(\eta_1 : \eta_2), & (30) \end{aligned}$$

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 family of probability distributions are the mixture families:

Definition 3 (Mixture family) *Given a set of k prescribed statistical distributions $p_0(x), \dots, p_{k-1}(x)$, 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 [40, 41]:*

$$\mathcal{M} := \left\{ m(x;\eta) = \sum_{i=1}^{k-1} \eta_i p_i(x) + (1 - \sum_{i=1}^{k-1} \eta_i) p_0(x) \,:\, \eta_i > 0, \sum_{i=1}^{k-1} \eta_i < 1 \right\}. \tag{31}$$

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

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

$$\mathrm{KL}(m(x;\eta) : m(x;\eta')) = B_G(\eta : \eta'), \tag{32}$$

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

$$G(\eta) = -h(m(x;\eta)) = \int_{x\in\mathcal{X}} m(x;\eta) \log m(x;\eta) \mathrm{d}\mu(x). \tag{33}$$

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 [7, 1]. One reason comes from the fact that the KL divergence between two mixtures amounts to a BD on their respective parameters (Eq. 32) while the KL divergence between exponential family distributions is equivalent to a BD on the swapped order of their respective parameters (Eq. 29). 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. 30.

1.2.3 Cauchy family manifold (CFM)

This example is merely given just to emphasize that probability families may neither be exponential nor mixture families.

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

$$p(x;\mu,\sigma) = \frac{1}{\pi\sigma\left(1 + \left(\frac{(x-\mu}{\sigma}\right)^2\right)}. \tag{34}$$

The space of all Cauchy distributions:

$$\mathcal{C} = \{p(x;\mu,\sigma) \,:\, \mu\in\mathbb{R}, \sigma > 0\}. \tag{35}$$

is a location-scale family [21]. 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 tensor* and ∇^e and ∇^m are the exponential and mixture connections, respectively. These connections are dual to each others, see [7].

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

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 [49]
Type 4	not closed-form	not closed-form	Polynomial exponential family [39]
Type 5	not analytic	not analytic	mixture family

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

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 [33] has sufficient statistic vector $t(x) = (x, x^2)$ so that its log-normalizer is

$$F(\theta) = \log\left(\int_{-\infty}^{+\infty} \exp(\theta_1 x + \theta_2 x^2)\mathrm{d}x\right). \tag{36}$$

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:

$$F(\theta) = \log\left(\int \exp(\theta_1 x + \theta_2 x^2)\mathrm{d}x\right) = -\frac{\theta_1^2}{4\theta_2} + \frac{1}{2}\log\frac{\pi}{-\theta_2}. \tag{37}$$

This is in accordance with the direct canonical decomposition [33] 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 $x \in (0, 1)$:

$$p(x; \alpha, \beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \tag{38}$$

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.[1]

[1] 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.

Type 3. This type of families has discrete support $\mathcal{X}$ that requires an exponential time to compute the log-normalizer. For example, consider the Ising models [18, 8, 4]: 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:

$$p(x;\theta) = \exp\left(\sum_{v\in V} \theta_v x_v + \sum_{(v,w)\in E} \theta_{vw} x_v x_w - F(\theta)\right). \tag{39}$$

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:

$$F(\theta) = \log\left(\sum_{(x_v)_v\in\{0,1\}^{|V|}} \left(\exp \sum_{v\in V} \theta_v x_v + \sum_{(v,w)\in E} \theta_{vw} x_v x_w\right)\right). \tag{40}$$

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

Type 4. This type of families has provably the Bregman generator that is not available in closed-form. For example, this is the case of the *Polynomial Exponential Family* [9, 39] (PEF) that 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:

$$\mathcal{E}_{t(x),\mu} = \left\{ p(x;\theta) = \exp\left(\sum_{i=1}^{D} \theta_i x^i - F(\theta)\right) \ : \theta \in \Theta \right\}. \tag{41}$$

(Beware that here, $x^i = \mathrm{Pow}(x, i) := \underbrace{x \times \ldots \times x}_{i \text{ times}}$ denotes the i-th 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. 25 does not admit a closed form, but is analytic. For example, choosing $t(x) = x^8$, we have:

$$F(\theta) = \log \int_{-\infty}^{\infty} \exp(\theta x^8)\mathrm{d}x = \log 2 + \log \Gamma(9/8) - \frac{1}{8}\log(-\theta), \tag{42}$$

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 [40]. It is in fact even worse, as the Shannon entropy of mixtures is not analytic [51].

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

In [11], 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 [47].

1.4 Paper organization

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

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 $p_0(x), \dots, p_{k-1}(x)$, 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 [40]:

$$\mathcal{M} := \left\{ m(x;\eta) = \sum_{i=1}^{k-1} \eta_i p_i(x) + (1 - \sum_{i=1}^{k-1} \eta_i) p_0(x) : \eta_i > 0, \sum_{i=1}^{k-1} \eta_i < 1 \right\}. \tag{43}$$

The differential-geometric structure of $\mathcal{M}$ is well studied in information geometry [7, 1] (although much less than for the exponential families), where it is known that:

$$\mathrm{KL}(m(x;\eta) : m(x;\eta')) = B_G(\eta : \eta'), \tag{44}$$

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

$$G(\eta) = -h(m(x;\eta)) = \int_{x\in\mathcal{X}} m(x;\eta) \log m(x;\eta) \mathrm{d}\mu(x). \tag{45}$$

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:

$$F(\theta) = G^*(\theta) = -\int_{x\in\mathcal{X}} p_0(x) \log m(x;\eta) \mathrm{d}\mu(x) = h^\times(p_0(x) : m(x;\eta)), \tag{46}$$

interpretable as the cross-entropy of $p_0(x)$ with the mixture $m(x;\eta)$ [40].

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 interpretable both as either a mixture family and an exponential family, see [1]), $G(\eta)$ provably does not admit a closed form [51] (i.e., meaning that the definite integral of Eq. 33 does not admit a simple formula using common standard functions). Thus the dually-flat geometry $(\mathcal{M}, G)$ is a theoretical construction that cannot be explicitly used by Bregman algorithms.

One way to tackle the lack of closed form of Eq. 33, is to approximate 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 $x_1, \dots, x_m \sim p(x)$, and use the the following MC KL estimator:

$$\widehat{\mathrm{KL}}_m(p : q) := \frac{1}{m} \sum_{i=1}^{m} \log \frac{p(x_i)}{q(x_i)}. \tag{47}$$

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 variates). 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 [31] of a set of n mixtures belonging to a mixture family manifold (say, with $D = 2$) using the algorithm explained in [31]: 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 as its output: 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 identically and independently distributed (iid) random sample $\mathcal{S}$ of prescribed size m.

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:

$$m(x;\eta) = \eta p_1(x) + (1-\eta)p_0(x) = p_0(x) + \eta(p_1(x) - p_0(x)), \tag{48}$$

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 [7].

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* $q(x)$ (defined over the same support $\mathcal{X}$, and independent of η). We approximate the Bregman generator $G(\eta)$ using Monte Carlo stochastic integration with importance sampling as follows:

$$G(\eta) \simeq \tilde{G}_\mathcal{S}(\eta) := \frac{1}{m}\sum_{i=1}^{m} \frac{1}{q(x_i)} m(x_i;\eta) \log m(x_i;\eta). \tag{49}$$

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 basic 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 the finite positively weighted sum of strictly convex functions.

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

$$g_x(\eta)' = m_x(\eta)'(\log m_x(\eta) + 1), \tag{50}$$

$$g_x(\eta)'' = m_x(\eta)''(\log m_x(\eta) + 1) + \frac{(m_x(\eta)')^2}{m_x(\eta)}. \tag{51}$$

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

$$g_x(\eta)'' = \frac{(p_1(x) - p_0(x))^2}{m_x(\eta)}. \tag{52}$$

Thus it follows that:

$$\tilde{G}_\mathcal{S}''(\eta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{q(x_i)}\frac{(p_1(x_i) - p_0(x_i))^2}{m(x_i;\eta)} \geq 0. \tag{53}$$

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

$$\mu_q(\mathcal{D}) := \int_{x\in\mathcal{X}} 1_{[p_0(x)=p_1(x)]} q(x) \mathrm{d}\mu(x) = 0. \tag{54}$$

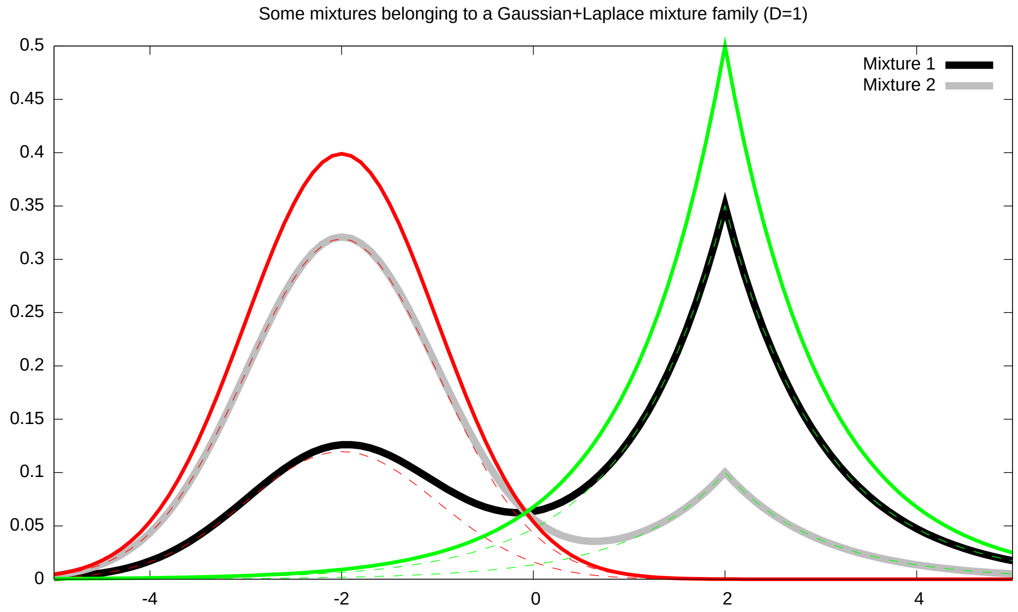

Figure 1: Example of a mixture family of order $D = 1$ ($k = 2$): $p_0(x) \sim$ Gaussian$(-2, 1)$ (red) and $p_1(x) \sim$ Laplace$(2, 1)$ (green). The two mixtures are $m_1(x) = m(x; \eta_1)$ (black) with $\eta_1 = 0.7$ and $m_2(x) = m(x; \eta_2)$ (grey) with $\eta_2 = 0.2$. Weighted component distributions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

Assumption 1 (AMF1D) *We assume that $p_0(x)$ and $p_1(x)$ are linearly independent (non-singular statistical model, see [7]), 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:

Monte Carlo Mixture Family Generator 1D:

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

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

$$\tilde{G}''_{\mathcal{S}}(\eta) = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{q(x_i)} \frac{(p_1(x_i)-p_0(x_i))^2}{m(x_i;\eta)}. \tag{57}$$

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.

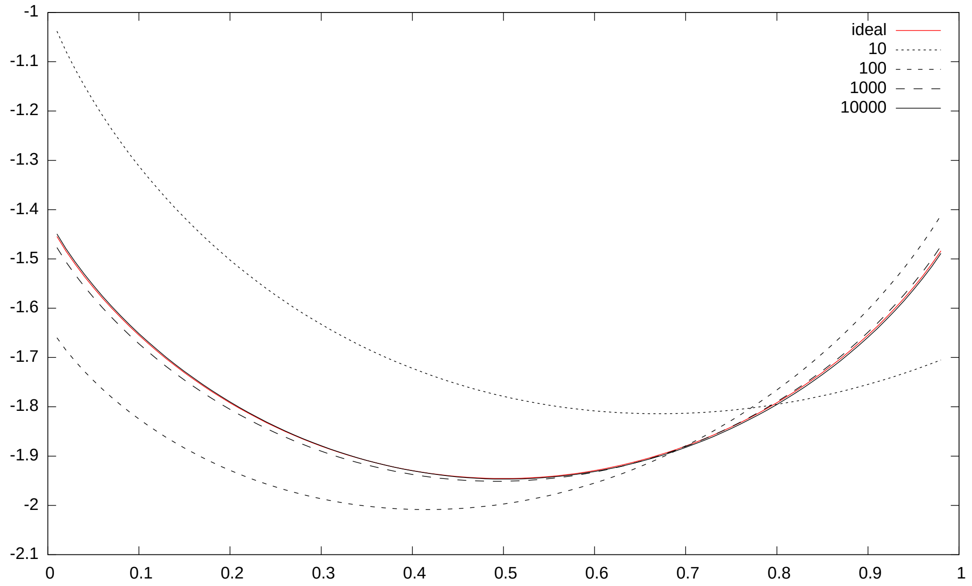

Figure 2: A series $G_{\mathcal{S}}(\eta)$ of Bregman Monte Carlo Mixture Family generators (for $m = |\mathcal{S}| \in \{10, 100, 1000, 10000\}$) approximating the untractable ideal negentropy generator $G(\eta) = -h(m(x;\eta))$ (red) of a mixture family with prescribed Gaussian distributions $m(x;\eta)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

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.

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

$$\mathrm{Var}_q\left[\frac{m(x;\eta)\log m(x;\eta)}{q(x)}\right] < \infty. \tag{58}$$

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:

$$q(x) = \frac{1}{m}\sum_{i=0}^{D} p_i(x). \tag{59}$$

Notice that the Monte Carlo Mixture Family Function is a random variable (rv) 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 $m = 10$.

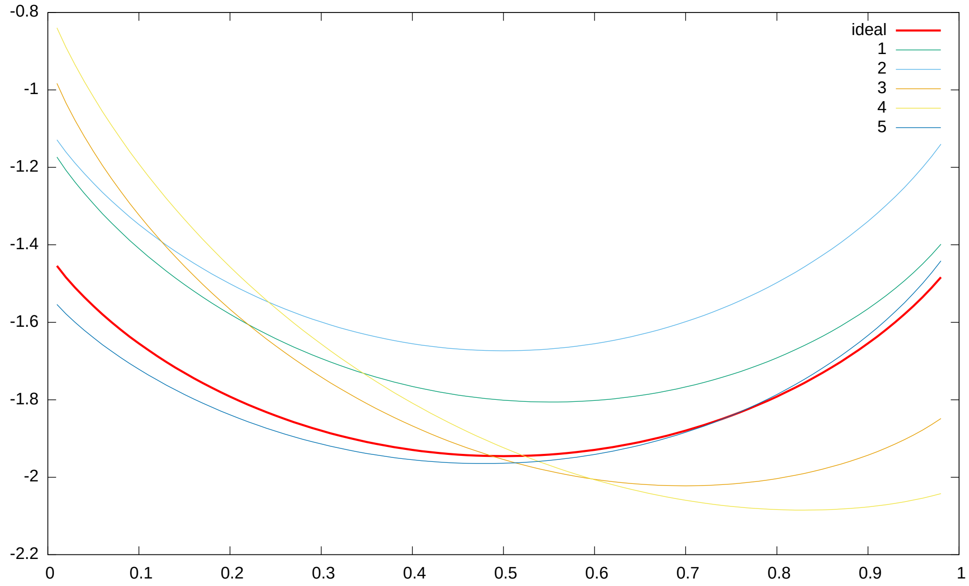

Figure 3: The Monte Carlo Mixture Family Generator $\hat{G}_{10}$ (MCMFG) considered as a random variable: Here, we show five realizations (i.e., $\mathcal{S}_1, \ldots, \mathcal{S}_5$) of the randomized generator for $m = 5$. The ideal generator is plot in thick red.

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 [7].

We further strengthen conditions on the prescribed distributions as follows:

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

$$\mathrm{sup}_{B\in\mathcal{B}} \left\{ \mu_q(B) : \exists \lambda \neq (0), \sum_{i\neq j} \lambda_i \left(p_i|_B - p_j|_B \right) = 0 \right\} = 0, \quad \forall j, \tag{60}$$

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:

$$\tilde{G}_{\mathcal{S}}(\eta) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{q(x_i)} m(x_i;\eta) \log m(x_i;\eta). \tag{61}$$

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:

$$\partial^i g_x(\eta) = (p_i(x) - p_0(x))(1 + \log m(x;\eta)), \tag{62}$$

and the partial second derivatives:

$$\partial^i \partial^j g_x(\eta) = \frac{(p_i(x) - p_0(x))(p_j(x) - p_0(x))}{m(x;\eta)}, \tag{63}$$

so that

$$\partial^i \partial^j \tilde{G}_{\mathcal{S}}(\eta) = \frac{1}{m} \sum_{l=1}^{m} \frac{1}{q(x_l)} \frac{(p_i(x_l) - p_0(x_l))(p_j(x_l) - p_0(x_l))}{m(x_l;\eta)}. \tag{64}$$

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:

$$v_l = \begin{bmatrix} \frac{p_1(x_l) - p_0(x_l)}{\sqrt{q(x_l) m(x_l;\eta)}} \\ \vdots \\ \frac{p_D(x_l) - p_0(x_l)}{\sqrt{q(x_l) m(x_l;\eta)}} \end{bmatrix}. \tag{65}$$

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

$$\partial^i \partial^j \tilde{G}_{\mathcal{S}}(\eta) = \frac{1}{m} \sum_{l=1}^{m} v_l v_l^\top. \tag{66}$$

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: §3.1 first describes the univariate case. It is then followed by the general multivariate case in 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:

$$\mathcal{E} := \{p(x;\theta) = \exp(t(x)\theta - F(\theta) + k(x)) : \theta \in \Theta\}, \tag{67}$$

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

$$F(\theta) = \log\left(\int \exp(t(x)\theta + k(x)) \mathrm{d}\mu(x)\right). \tag{68}$$

The sufficient statistic function $t(x)$ and 1 are linearly independent [7].

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:

$$F(\theta) \simeq \tilde{F}_{\mathcal{S}}^{\dagger}(\theta) := \log\left(\frac{1}{m} \sum_{i=1}^{m} \frac{1}{q(x_i)} \exp(t(x_i)\theta + k(x_i))\right). \tag{69}$$

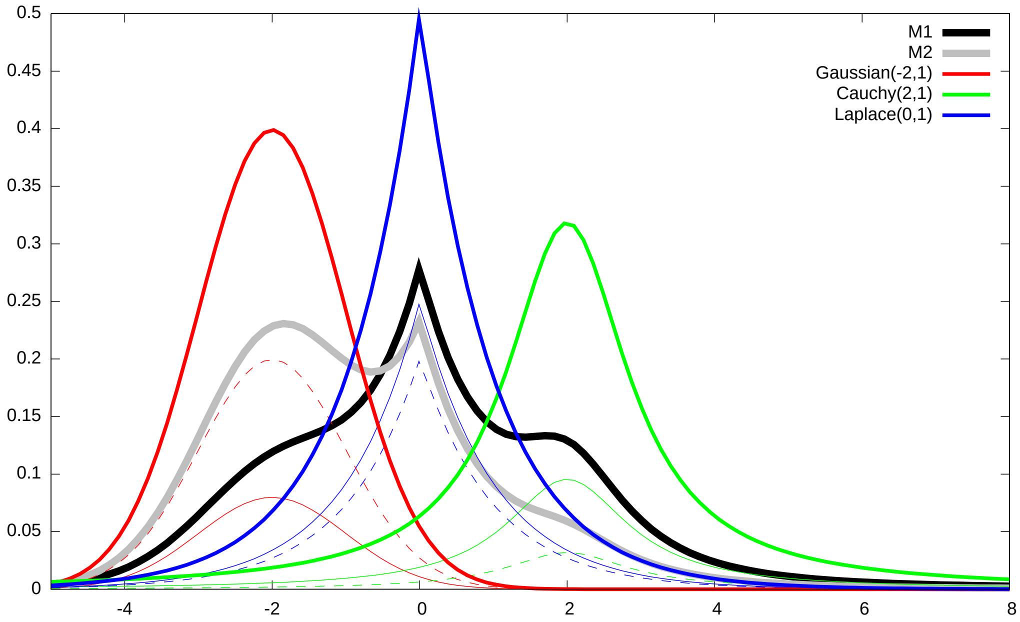

Figure 4: Example of a mixture family of order $D = 2$ ($k = 3$): $p_0(x) \sim \text{Gaussian}(-2, 1)$ (red), $p_1(x) \sim \text{Laplace}(0, 1)$ (blue) and $p_2(x) \sim \text{Cauchy}(2, 1)$ (green). The two mixtures are $m_1(x) = m(x; \eta_1)$ (black) with $\eta_1 = (0.3, 0.5)$ and $m_2(x) = m(x; \eta)$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

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. 69 and remove an affine term from the generator $\tilde{F}_{\mathcal{S}}(\theta)$ to get the equivalent generator (see Property 1):

$$\begin{aligned} \tilde{F}^{\dagger}_{\mathcal{S}}(\theta) &\equiv \tilde{F}_{\mathcal{S}}(\theta), && (70) \\ \tilde{F}_{\mathcal{S}}(\theta) &= \log\left(1 + \sum_{i=2}^{m} \exp((t(x_i) - t(x_1))\theta + k(x_i) - k(x_1) - \log q(x_i) + \log q(x_1))\right), && (71) \\ &= \log\left(1 + \sum_{i=2}^{m} \exp(a_i\theta + b_i)\right), && (72) \\ &:= \mathrm{lse}^{+}_{0}(a_2\theta + b_2, \ldots, a_m\theta + b_m), && (73) \end{aligned}$$

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 [44, 17] $\mathrm{lse}(x_1, \ldots, x_m) = \log \sum_{i=1}^{n} \exp(x_i)$ 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[2] [6], 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.

[2] 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.

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

Proof is deferred to Appendix A.

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

Proof. The first derivative is:

$$\eta = \tilde{F}'_{\mathcal{S}}(\theta) = \frac{\sum_{i=2}^m a_i \exp(a_i\theta + b_i)}{1 + \sum_{i=2}^m \exp(a_i\theta + b_i)} \geq 0, \tag{74}$$

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:

$$\tilde{F}''_{\mathcal{S}}(\theta) = \frac{\left(\sum_{i=2}^m a_i^2 \exp(a_i\theta + b_i)\right)\left(1 + \sum_{i=2}^m \exp(a_i\theta + b_i)\right) - \left(\sum_{i=2}^m a_i \exp(a_i\theta + b_i)\right)^2}{\left(1 + \sum_{i=2}^m \exp(a_i\theta + b_i)\right)^2} =: \frac{\mathrm{Num}}{\mathrm{Den}} \tag{75}$$

For each value of $\theta \in \Theta$, we shall prove that $\tilde{F}''_{\mathcal{S}}(\theta) > 0$. 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:

$$\mathrm{Num} > \left(\sum_{i=2}^m a_i^2 c_i\right)\left(\sum_{i=2}^m c_i\right) - \left(\sum_{i=2}^m a_i c_i\right)^2, \tag{76}$$

$$\mathrm{Num} > \sum_{ij} a_i^2 c_i c_j - \sum_i a_i^2 c_i^2 - 2\sum_{i<j} a_i a_j c_i c_j, \tag{77}$$

$$\mathrm{Num} > \sum_{i=j} a_i^2 c_i^2 + \sum_{i\neq j} a_i^2 c_i c_j - \sum_i a_i^2 c_i^2 - 2\sum_{i<j} a_i a_j c_i c_j, \tag{78}$$

$$\mathrm{Num} > \sum_{i<j} a_i^2 c_i c_j + \sum_{i>j} a_i^2 c_i c_j - 2\sum_{i<j} a_i a_j c_i c_j, \tag{79}$$

$$\mathrm{Num} > \sum_{i<j} a_i^2 c_i c_j + \sum_{i<j} a_j^2 c_i c_j - 2\sum_{i<j} a_i a_j c_i c_j, \tag{80}$$

$$\mathrm{Num} > \sum_{i<j} (a_i^2 + a_j^2 - 2a_i a_j) c_i c_j, \tag{81}$$

$$\mathrm{Num} > \sum_{i<j} (a_i - a_j)^2 c_i c_j > 0. \tag{82}$$

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 $y \in \mathrm{dom}(t)$, $E_q[1_{t(x)=y}] = 0$.*

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

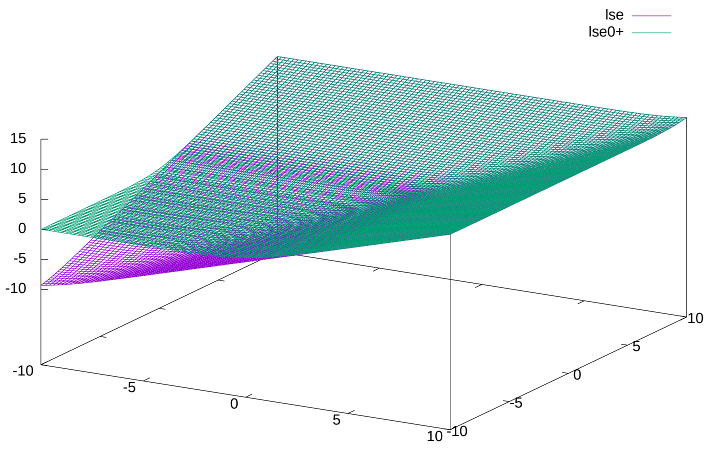

Figure 5: Graph plots of the lse and lse_0^+ functions: The lse function (violet) is only convex while the lse_0^+ function (green) is always guaranteed to be strictly convex.

Monte Carlo Mixture Family Generator 1D:

$$\begin{aligned}
\tilde{F}_{\mathcal{S}}(\theta) &= \mathrm{lse}_0^+(a_2\theta+b_2,\ldots,a_m\theta+b_m), && (83)\\
& a_i = t(x_i)-t(x_1), && (84)\\
& b_i = k(x_i)-k(x_1)-\log q(x_i)+\log q(x_1), && (85)\\
\tilde{F}'_{\mathcal{S}}(\theta) &= \frac{\sum_{i=2}^m a_i\exp(a_i\theta+b_i)}{1+\sum_{i=2}^m \exp(a_i\theta+b_i)} =: \eta, && (86)\\
\tilde{F}''_{\mathcal{S}}(\theta) &= \frac{\left(\sum_{i=2}^m a_i^2\exp(a_i\theta+b_i)\right)\left(1+\sum_{i=2}^m\exp(a_i\theta+b_i)\right)-\left(\sum_{i=2}^m a_i\exp(a_i\theta+b_i)\right)^2}{(1+\sum_{i=2}^m\exp(a_i\theta+b_i))^2} && (87)
\end{aligned}$$

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:

$$a_i = \begin{bmatrix} a_i^1 \\ \vdots \\ a_i^D \end{bmatrix}. \qquad (88)$$

We replace the scalar multiplication $a_i\theta$ by an inner product $\langle a_i,\theta\rangle$ in Eq. 73, 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:

$$\begin{aligned}
\tilde{F}_{\mathcal{S}}(\theta) &= \log\left(1+\sum_{l=2}^m c_l(\theta)\right), && (89)\\
&:= \mathrm{lse}_0^+(c_2(\theta),\ldots,c_m(\theta)), && (90)
\end{aligned}$$

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:

$$\eta_i = \partial_i \tilde{F}_{\mathcal{S}}(\theta) = \frac{\sum_{l=2}^m a_l^i c_l(\theta)}{1 + \sum_{l=2}^m c_l(\theta)}, \tag{91}$$

and the Hessian matrix of second-order partial derivatives:

$$\partial_i \partial_j \tilde{F}_{\mathcal{S}}(\theta) = \frac{(\sum_{l=2}^m a_l^i a_l^j c_l(\theta))(1 + \sum_{l=2}^m c_l(\theta)) - (\sum_{l=2}^m a_l^i c_l(\theta))(\sum_{l=2}^m a_l^j c_l(\theta))}{(1 + \sum_{l=2}^m c_l(\theta))^2} =: \frac{\mathrm{Num}}{\mathrm{Den}}. \tag{92}$$

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:

$$\mathrm{Num} = \underbrace{\sum_k a_k^i a_k^j c_k}_{:=D} + \underbrace{\sum_{k,l} a_k^i a_k^j c_k c_l - \sum_{k,l} a_k^i c_k a_l^j c_l}_{:=E}. \tag{93}$$

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:

$$\begin{aligned}
E &\overset{\star}{=} \sum_{k<l} a_k^i a_k^j c_k c_l + \sum_{l<k} a_k^i z_k^j c_k c_l - \sum_{k<l} a_k^i a_l^j c_k c_l - \sum_{l<k} a_k^i a_l^j c_k c_l, & (94)\\
&\overset{\star\star}{=} \sum_{k<l} \left(a_k^i a_k^j + a_l^i a_l^j - a_k^i a_l^j - a_l^i a_k^j \right) c_k c_l, & (95)\\
&= \sum_{k<l} (a_k^i - a_l^i)(a_k^j - a_l^j) c_k c_l. & (96)
\end{aligned}$$

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

$\star\star$: After a change of variable $l \leftrightarrow k$ in the second and fourth sums of Eq. 94.

Thus Eq. 96 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

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

which implies

$$\sum_i y_i c_i \left(t_i(x_k) - t_i(x_l) \right) \sum_j y_j c_j \left(t_j(x_k) - t_j(x_l) \right) = 0, \quad \forall k < l, \tag{98}$$

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

$$\sum_i y_i t_i(x_k) = \sum_i y_i t_i(x_l), \quad \forall y \neq 0, \quad \forall k < l. \tag{99}$$

□

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 §4.1, and consider in §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 [13, 34]. This 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:

$$\begin{aligned}\mathrm{KL}(m(x;\eta_i):m(x;\eta_j)) &= B_G(\eta_i:\eta_j), &(100)\\ &\simeq B_{\tilde{G}_{\mathcal{S}}}(\eta_i:\eta_j). &(101)\end{aligned}$$

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 [3]:

$$\widehat{\mathrm{eKL}}_m(p:q) := \frac{1}{m}\sum_{i=1}^{m}\left(\log\frac{p(x_i)}{q(x_i)}+\frac{q(x_i)}{p(x_i)}-1\right), \qquad (102)$$

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

$$\widehat{\mathrm{KL}}_m(p:q) := \frac{1}{m}\sum_{i=1}^{m}\log\frac{p(x_i)}{q(x_i)}, \qquad (103)$$

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

Bregman k-means [3, 20] can be applied using either the sided o ther symmetrized centroid [37]: 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[3]). Although $F'(\theta)$ is available in closed form (and define the dual parameter θ):

$$\tilde{G}'_{\mathcal{S}}(\eta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{q(x_i)}\left(p_1(x_i)-p_0(x)\right)\left(1+\log m(x;\eta)\right)=\theta, \qquad (104)$$

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 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:

$$\begin{aligned}J(m(x;\eta_i):m(x;\eta_j)) &= \mathrm{KL}(m(x;\eta_i):m(x;\eta_j))+\mathrm{KL}(m(x;\eta_j):m(x;\eta_i)), &(105)\\ &= B_G(\eta_i:\eta_j)+B_G(\eta_j:\eta_i), &(106)\\ &= B_G(\eta_i:\eta_j)+B_{G^*}(\theta_i:\theta_j), &(107)\\ &= \langle\Delta\theta_{ij},\Delta\eta_{ij}\rangle, &(108)\end{aligned}$$

[3] Otherwise, we need to consider monotone operator theory [23] to invert $\nabla F(\theta)$.

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. 106:

$$J(m(x;\eta_i) : m(x;\eta_j)) \simeq B_{\tilde{G}_{\mathcal{S}}}(\eta_i : \eta_j) + B_{\tilde{G}_{\mathcal{S}}}(\eta_j : \eta_i). \tag{109}$$

We can then apply the technique of mixed Bregman clustering [45] that considers two centers per cluster. Moreover a fast probabilistic initialization, called *mixed Bregman k-means++* [45], allows one to guarantee a good initialization with high probability (without computing centroids but requiring to compute divergences).

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 [32] for an infinitesimal value of α. Indeed, we have the α-Jensen divergence defined by:

$$J_F^\alpha(p:q) = (1-\alpha)F(p) + \alpha F(q) - F((1-\alpha)p + \alpha q), \tag{110}$$

and asymptotically this skewed Jensen divergences yield the sided Bregman divergences [32] as follows:

$$\begin{aligned} \lim_{\alpha\rightarrow 0^+} \frac{J_F^\alpha(p:q)}{\alpha} &= B_F(q:p), & (111)\\ \lim_{\alpha\rightarrow 1^-} \frac{J_F^\alpha(p:q)}{1-\alpha} &= B_F(p:q), & (112)\end{aligned}$$

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

$$\begin{aligned} J(m(x;\eta_i) : m(x;\eta_j)) &= B_G(\eta_i : \eta_j) + B_G(\eta_j : \eta_i), & (113)\\ &\simeq \frac{1}{\alpha} J_{\tilde{G}_{\mathcal{S}}}^\alpha(\eta_i : \eta_j) + \frac{1}{1-\alpha} J_{\tilde{G}_{\mathcal{S}}}^{1-\alpha}(\eta_i : \eta_j). & (114)\end{aligned}$$

The last equation Eq.114 is the symmetrized skew Jensen divergence studied in [26].

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:

$$\tilde{G}_{\mathcal{S}}(\theta) = \sum_{i=1}^{l} \frac{|\mathcal{S}_i|}{|\mathcal{S}|} \tilde{G}_{\mathcal{S}_i}(\eta). \tag{115}$$

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:

$$\tilde{F}_{\mathcal{S}}(\theta) = \log\left(\sum_{i=1}^{s} \frac{|\mathcal{S}_i|}{|\mathcal{S}|} \exp(\tilde{F}_{\mathcal{S}_i})\right). \tag{116}$$

That is, $\tilde{F}_{\mathcal{S}}(\theta)$ can be interpreted as an *exponential mean* (quasi-arithmetic mean, called f-mean [32] 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 [30].

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.*

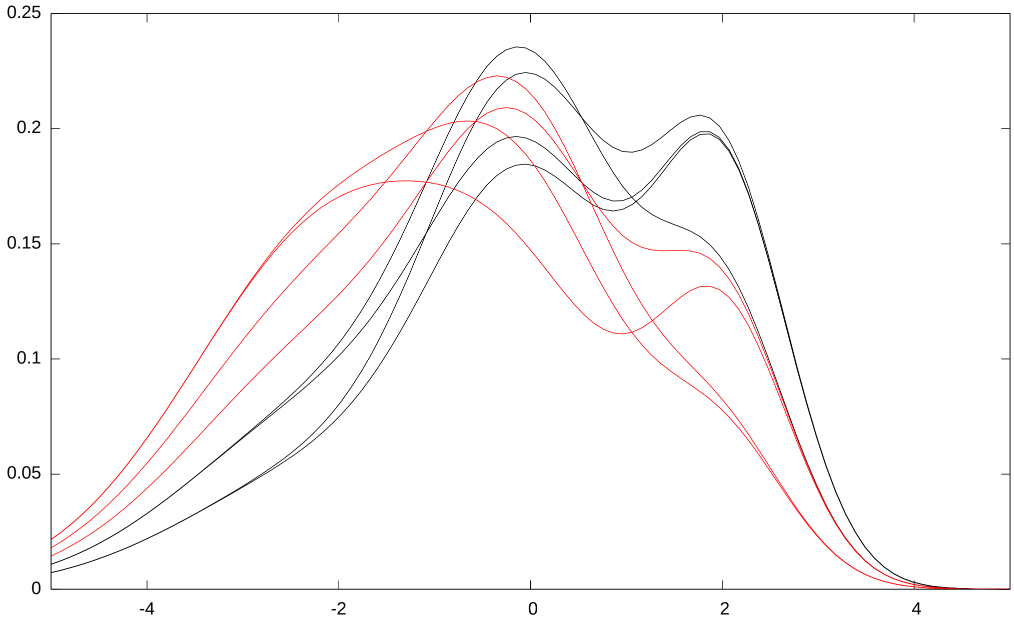

Figure 6: Clustering a set of $n = 8$ statistical mixtures of order $D = 2$ with $K = 2$ clusters: Each mixture is represented by a 2D point on the mixture family manifold. The Kullback-Leibler divergence is equivalent to an integral-based Bregman divergence that is computationally untractable: The Bregman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

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 [46] 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 [24]. This amounts to compute a sided centroid for θ.

5 Conclusion and perspectives

In this work, we have 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 (MCIG). MCIG performs stochastic integration of the ideal but computationally intractable definite integral-based Bregman generator (e.g. Eq 33 for mixture family) for mixture family and Eq 25 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 yields a computational 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 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 is the sum of elementary gradients that cannot be inverted easily. This step requires a numerical or symbolic technique [23].

We note that in the recent work of [25], Matsuzoe et al. defined a sequence of statistical manifolds relying on a sequential structure of escort expectations for non-exponential type statistical models.

In a forthcoming work [35], we address the more general case of the Monte Carlo information-geometric structure of a generic statistical manifold of a parametric family of distributions induced by an arbitrary statistical f-divergence. That is, we consider a statistical divergence $D(p(x;\theta_1):p(x;\theta_2)) = \int_{x\in\mathcal{X}} D_1(p(x;\theta_1):p(x;\theta_2))\mathrm{d}\mu(x)$ (where $D_1(\cdot:\cdot)$ is a univariate divergence), and study the information-geometric structure $(\mathcal{M}, g_{\tilde{D}}, \nabla_{\tilde{D}}, \nabla^*_{\tilde{D}})$ induced by the Monte Carlo stochastic approximation of the divergence with m iid samples x_i's: $\tilde{D}(p(x;\theta_1):p(x;\theta_2)) := \frac{1}{m}\sum_{i=1}^m \frac{1}{q(x_i)} D_1(p(x_i;\theta_1):p(x_i;\theta_2))$.

Codes for reproducible results are available at:

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

References

[1] S. Amari. *Information Geometry and Its Applications*. Applied Mathematical Sciences. Springer Japan, 2016.

[2] Shun-ichi Amari and Andrzej Cichocki. Information geometry of divergence functions. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 58(1):183–195, 2010.

[3] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of machine learning research*, 6(Oct):1705–1749, 2005.

[4] Bhaswar B Bhattacharya, Sumit Mukherjee, et al. Inference in Ising models. *Bernoulli*, 24(1):493–525, 2018.

[5] Jean-Daniel Boissonnat, Frank Nielsen, and Richard Nock. Bregman Voronoi diagrams. *Discrete & Computational Geometry*, 44(2):281–307, 2010.

[6] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[7] Ovidiu Calin and Constantin Udriste. *Geometric Modeling in Probability and Statistics*. Mathematics and Statistics. Springer International Publishing, 2014.

[8] Barry A Cipra. The Ising model is NP-complete. *SIAM News*, 33(6):1–3, 2000.

[9] Loren Cobb, Peter Koppstein, and Neng Hsin Chen. Estimation and moment recursion relations for multimodal distributions of the exponential family. *Journal of the American Statistical Association*, 78(381):124–130, 1983.

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

[11] Frank Critchley and Paul Marriott. Computational information geometry in statistics: theory and practice. *Entropy*, 16(5):2454–2471, 2014.

[12] Jean-Pierre Crouzeix. A relationship between the second derivatives of a convex function and of its conjugate. *Mathematical Programming*, 13(1):364–365, 1977.

[13] Jason V Davis and Inderjit S Dhillon. Differential entropic clustering of multivariate gaussians. In *Advances in Neural Information Processing Systems*, pages 337–344, 2007.

[14] A Philip Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59(1):77–93, 2007.

[15] Shinto Eguchi. Geometry of minimum contrast. *Hiroshima Mathematical Journal*, 22(3):631–647, 1992.

[16] Daniel A Fleisch. *A student's guide to vectors and tensors*. Cambridge University Press, 2011.

[17] B. Gao and L. Pavel. On the Properties of the Softmax Function with Application in Game Theory and Reinforcement Learning. *ArXiv e-prints*, April 2017.

[18] Stuart Geman and Christine Graffigne. Markov random field image models and their applications to computer vision. In *Proceedings of the international congress of mathematicians*, volume 1, page 2, 1986.

[19] Allan Grønlund, Kasper Green Larsen, Alexander Mathiasen, and Jesper Sindahl Nielsen. Fast exact k-means, k-medians and Bregman divergence clustering in 1D. *CoRR*, abs/1701.07204, 2017.

[20] Allan Grønlund, Kasper Green Larsen, Alexander Mathiasen, and Jesper Sindahl Nielsen. Fast exact k-means, k-medians and Bregman divergence clustering in 1d. *arXiv preprint arXiv:1701.07204*, 2017.

[21] Robert E. Kass and Paul W. Vos. *Geometrical Foundations of Asymptotic Inference*. Wiley-Interscience, July 1997. Fisher-Rao metric of location-scale family is hyperbolic (and can be diagonalized), pages 192–193.

[22] Steffen L. Lauritzen. Statistical manifolds. *Differential Geometry in Statistical Inference*, page 164, 1987.

[23] Florian Lauster, D Russell Luke, and Matthew K Tam. Symbolic computation with monotone operators. *Set-Valued and Variational Analysis*, pages 1–16, 2017.

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

[25] Hiroshi Matsuzoe, Antonio M Scarfone, and Tatsuaki Wada. A sequential structure of statistical manifolds on deformed exponential family. In *International Conference on Geometric Science of Information*, pages 223–230. Springer, 2017.

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

[27] Frank Nielsen. Legendre transformation and information geometry, 2010.

[28] Frank Nielsen. Hypothesis testing, information divergence and computational geometry. In *Geometric Science of Information*, pages 241–248. Springer, 2013.

[29] Frank Nielsen. An information-geometric characterization of Chernoff information. *IEEE Signal Processing Letters*, 20(3):269–272, 2013.

[30] Frank Nielsen. *Introduction to HPC with MPI for Data Science*. Undergraduate Topics in Computer Science. Springer, 2016.

[31] Frank Nielsen, Jean-Daniel Boissonnat, and Richard Nock. On Bregman Voronoi diagrams. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 746–755. Society for Industrial and Applied Mathematics, 2007.

[32] Frank Nielsen and Sylvain Boltz. The Burbea-Rao and Bhattacharyya centroids. *IEEE Transactions on Information Theory*, 57(8):5455–5466, 2011.

[33] Frank Nielsen and Vincent Garcia. Statistical exponential families: A digest with flash cards. *arXiv preprint arXiv:0911.4863*, 2009.

[34] Frank Nielsen, Vincent Garcia, and Richard Nock. Simplifying Gaussian mixture models via entropic quantization. In *17th European Conference on Signal Processing (EUSIPCO)*, pages 2012–2016. IEEE, 2009.

[35] Frank Nielsen and Gaëtan Hadjeres. Monte Carlo information geometry: The generic case of statistical manifolds. *preprint*, 2018.

[36] Frank Nielsen and Richard Nock. On the smallest enclosing information disk. *Information Processing Letters*, 105(3):93–97, 2008.

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

[38] Frank Nielsen and Richard Nock. Optimal interval clustering: Application to Bregman clustering and statistical mixture learning. *IEEE Signal Processing Letters*, 21(10):1289–1292, 2014.

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

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

[41] Frank Nielsen and Richard Nock. On the geometric of mixtures of prescribed distributions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[42] Frank Nielsen, Paolo Piro, and Michel Barlaud. Bregman vantage point trees for efficient nearest neighbor queries. In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pages 878–881. IEEE, 2009.

[43] Frank Nielsen, Paolo Piro, and Michel Barlaud. Tailored Bregman ball trees for effective nearest neighbors. In *Proceedings of the 25th European Workshop on Computational Geometry (EuroCG)*, pages 29–32, 2009.

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

[45] Richard Nock, Panu Luosto, and Jyrki Kivinen. Mixed Bregman clustering with approximation guarantees. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 154–169. Springer, 2008.

[46] Bruno Pelletier. Informative barycentres in statistics. *Annals of the Institute of Statistical Mathematics*, 57(4):767–780, 2005.

[47] Christian P Robert. *Monte Carlo methods*. Wiley Online Library, 2004.

[48] Hirohiko Shima. *The geometry of Hessian structures*. World Scientific, 2007.

[49] Yichuan Tang and Ruslan R Salakhutdinov. Learning stochastic feedforward neural networks. In *Advances in Neural Information Processing Systems*, pages 530–538, 2013.

[50] Richard S Varga. *Geršgorin and his circles*, volume 36. Springer Science & Business Media, 2010.

[51] Sumio Watanabe, Keisuke Yamazaki, and Miki Aoyagi. Kullback information of normal mixture is not an analytic function. *technical report of IEICE (in Japanese)*, (2004-0):41–46, 2004.

[52] Jun Zhang. Reference duality and representation duality in information geometry. In *AIP Conference Proceedings*, volume 1641, pages 130–146. AIP, 2015.

A $\mathrm{lse}_0^+(x)$ 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:

$$\begin{aligned}
\partial_i \mathrm{lse}_0^+(x) &= \frac{e^{x_i}}{1+\sum_k e^{x_k}}, && (117)\\
\partial_j\partial_i \mathrm{lse}_0^+(x) &\overset{j\neq i}{=} \frac{-e^{x_i}e^{x_j}}{(1+\sum_k e^{x_k})^2}, && (118)\\
\partial_i\partial_i \mathrm{lse}_0^+(x) &= \frac{e^{x_i}(1+\sum_k e^{x_k}) - e^{x_i}e^{x_j}}{(1+\sum_k e^{x_k})^2}. && (119)
\end{aligned}$$

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

$$e^{x_i}\left(1+\sum_k e^{x_k}\right) = e^{x_i} + e^{x_i}\sum_k e^{x_k} > \sum_{j\neq i} |-e^{x_i}e^{x_j}| = e^{x_i}\sum_{j\neq i} e^{x_j}. \tag{120}$$

To conclude that the Hessian matrix is SPD, we use Gershgorin circle theorem [50] 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:

$$\nabla \mathrm{lse}(x) = \sigma(x), \tag{121}$$

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

$$\sigma(x) := \left(\frac{e^{x_i}}{\sum_{k=1}^d e^{x_k}}\right)_{i\in\{1,\ldots,d\}}. \tag{122}$$

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

$$\sigma_0^+(x) := \left(\frac{e^{x_i}}{1+\sum_k e^{x_k}}\right)_{i\in\{1,\ldots,d\}}, \tag{123}$$

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