Quasi-arithmetic centers, quasi-arithmetic mixtures, and the Jensen-Shannon ∇-divergences

Frank Nielsen[1[0000-0001-5728-0726]]

Sony Computer Science Laboratories Inc, Tokyo, Japan.

Abstract. We first explain how the information geometry of Bregman manifolds brings a natural generalization of scalar quasi-arithmetic means that we term quasi-arithmetic centers. We study the invariance and equivariance properties of quasi-arithmetic centers from the viewpoint of the Fenchel-Young canonical divergences. Second, we consider statistical quasi-arithmetic mixtures and define generalizations of the Jensen-Shannon divergence according to geodesics induced by affine connections.

Keywords: Legendre-type function · quasi-arithmetic means · co-monotonicity · information geometry · statistical mixtures · Jensen-Shannon divergence.

1 Introduction

Let $\Delta_{n-1} = \{(w_1,\ldots,w_n) \ : \ w_i \geq 0, \sum_i w_i = 1\} \subset \mathbb{R}^d$ denotes the closed $(n-1)$-dimensional standard simplex sitting in $\mathbb{R}^n$, ∂ be the set boundary operator, and $\Delta_{n-1}^\circ = \Delta_{n-1} \backslash \partial\Delta_{n-1}$ the open standard simplex. Weighted quasi-arithmetic means [12] (QAMs) generalize the ordinary weighted arithmetic mean $A(x_1,\ldots,x_n;w) = \sum_i w_i x_i$ as follows:

Definition 1 (Weighted quasi-arithmetic mean (1930's)). *Let $f : I \subset \mathbb{R} \to \mathbb{R}$ be a strictly monotone and differentiable real-valued function. The weighted quasi-arithmetic mean (QAM) $M_f(x_1,\ldots,x_n;w)$ between n scalars $x_1,\ldots,x_n \in I \subset \mathbb{R}$ with respect to a normalized weight vector $w \in \Delta_{n-1}$, is defined by*

$$M_f(x_1,\ldots,x_n;w) := f^{-1}\left(\sum_{i=1}^n w_i f(x_i)\right).$$

Let us write for short $M_f(x_1,\ldots,x_n) := M_f(x_1,\ldots,x_n;\frac{1}{n},\ldots,\frac{1}{n})$, and $M_{f,\alpha}(x,y) := M_f(x,y;\alpha,1-\alpha)$ for $\alpha \in [0,1]$, the weighted bivariate QAM. A QAM satisfies the in-betweenness property:

$$\min\{x_1,\ldots,x_n\} \leq M_f(x_1,\ldots,x_n;w) \leq \max\{x_1,\ldots,x_n\},$$

and we have [16] $M_g(x,y) = M_f(x,y)$ if and only if $g(t) = \lambda f(t) + c$ for $\lambda \in \mathbb{R}\backslash\{0\}$ and $c \in \mathbb{R}$. The power means $M_p(x,y) := M_{f_p}(x,y)$ are obtained for the following continuous family of QAM generators indexed by $p \in \mathbb{R}$:

$$f_p(t) = \begin{cases} \frac{t^p-1}{p}, & p \in \mathbb{R}\backslash\{0\}, \\ \log(t), & p = 0. \end{cases}, \quad f_p^{-1}(t) = \begin{cases} (1+tp)^{\frac{1}{p}}, & p \in \mathbb{R}\backslash\{0\}, \\ \exp(t), & p = 0. \end{cases},$$

Special cases of the power means are the harmonic mean ($H = M_{-1}$), the geometric mean ($G = M_0$), the arithmetic mean ($A = M_1$), and the quadratic mean also called root mean square ($Q = M_2$). A QAM is said positively homogeneous if and only if $M_f(\lambda x, \lambda y) = \lambda\, M_f(x, y)$ for all $\lambda > 0$. The power means M_p are the only positively homogeneous QAMs [12].

In Section 2, we define a generalization of quasi-arithmetic means called quasi-arithmetic centers (Definition 3) induced by a Legendre-type function. We show that the gradient maps of convex conjugate functions are co-monotone (Proposition 1). We then study their invariance and equivariance properties (Proposition 2). In Section 4, we define quasi-arithmetic mixtures (Definition 4), show their connections to geodesics, and define a generalization of the Jensen-Shannon divergence with respect to affine connections (Definition 5).

2 Quasi-arithmetic centers and information geometry

2.1 Quasi-arithmetic centers

To generalize scalar QAMs to other non-scalar types such as vectors or matrices, we face two difficulties:

1. we need to ensure that the generator $G : \mathbb{X} \to \mathbb{R}$ admits a global inverse[1] G^{-1}, and
2. we would like the smooth function G to bear a generalization of monotonicity of univariate functions.

We consider a well-behaved class $\mathcal{F}$ of non-scalar functions G (i.e., vector or matrix functions) which admits global inverse functions G^{-1} belonging to the same class $\mathcal{F}$: Namely, we consider the gradient maps of Legendre-type functions where Legendre-type functions are defined as follows:

Definition 2 (Legendre type function [24]). *(Θ, F) is of Legendre type if the function $F : \Theta \subset \mathbb{X} \to \mathbb{R}$ is strictly convex and differentiable with $\Theta \neq \emptyset$ an open convex set and*

$$\lim_{\lambda \to 0} \frac{d}{\mathrm{d}\lambda} F(\lambda\theta + (1-\lambda)\bar{\theta}) = -\infty, \quad \forall \theta \in \Theta, \forall \bar{\theta} \in \partial\Theta. \tag{1}$$

Legendre-type functions $F(\Theta)$ admits a convex conjugate $F^*(\eta)$ of Legendre type via the Legendre transform (Theorem 1 [24]):

$$F^*(\eta) = \left\langle \nabla F^{-1}(\eta), \eta \right\rangle - F(\nabla F^{-1}(\eta)),$$

[1] The inverse function theorem [10, 11] in multivariable calculus states only the local existence of an inverse continuously differentiable function G^{-1} for a multivariate function G provided that the Jacobian matrix of G is not singular

where $\langle\theta,\eta\rangle$ denotes the inner product in $\mathbb{X}$ (e.g., Euclidean inner product $\langle\theta,\eta\rangle=\theta^\top\eta$ for $\mathbb{X}=\mathbb{R}^d$, the Hilbert-Schmidt inner product $\langle A,B\rangle:=\mathrm{tr}(AB^\top)$ where $\mathrm{tr}(\cdot)$ denotes the matrix trace for $\mathbb{X}=\mathrm{Mat}_{d,d}(\mathbb{R})$, etc.), and $\eta\in H$ with H the image of the gradient map $\nabla F:\Theta\to H$. Moreover, we have $\nabla F^*=(\nabla F)^{-1}$ and $\nabla F=(\nabla F^*)^{-1}$, i.e., gradient maps of conjugate functions are reciprocal to each others.

The gradient of a strictly convex function of Legendre type exhibit a generalization of the notion of monotonicity of univariate functions: A function $G:\mathbb{X}\to\mathbb{R}$ is said strictly increasing co-monotone if

$$\forall\theta_1,\theta_2\in\mathbb{X},\theta_1\neq\theta_2,\quad \langle\theta_1-\theta_2,G(\theta_1)-G(\theta_2)\rangle>0.$$

and strictly decreasing co-monotone if $-G$ is strictly increasing co-monotone.

Proposition 1 (Gradient co-monotonicity [25]). *The gradient functions $\nabla F(\theta)$ and $\nabla F^*(\eta)$ of the Legendre-type convex conjugates F and F^* in $\mathcal{F}$ are strictly increasing co-monotone functions.*

Proof. We have to prove that

$$\langle\theta_2-\theta_1,\nabla F(\theta_2)-\nabla F(\theta_1)\rangle>0,\quad\forall\theta_1\neq\theta_2\in\Theta \tag{2}$$

$$\langle\eta_2-\eta_1,\nabla F^*(\eta_2)-\nabla F^*(\eta_1)\rangle>0,\quad\forall\eta_1\neq\eta_2\in H \tag{3}$$

The inequalities follow by interpreting the terms of the left-hand-side of Eq. 2 and Eq. 3 as Jeffreys-symmetrization [17] of the dual Bregman divergences [9] B_F and B_{F^*}:

$$\begin{aligned}B_F(\theta_1:\theta_2)&=F(\theta_1)-F(\theta_2)-\langle\theta_1-\theta_2,\nabla F(\theta_2)\rangle\geq 0,\\ B_{F^*}(\eta_1:\eta_2)&=F^*(\eta_1)-F^*(\eta_2)-\langle\eta_1-\eta_2,\nabla F^*(\eta_2)\rangle\geq 0,\end{aligned}$$

where the first equality holds if and only if $\theta_1=\theta_2$ and the second inequality holds iff $\eta_1=\eta_2$. Indeed, we have the following Jeffreys-symmetrization of the dual Bregman divergences:

$$\begin{aligned}B_F(\theta_1:\theta_2)+B_F(\theta_2:\theta_1)&=\langle\theta_2-\theta_1,\nabla F(\theta_2)-\nabla F(\theta_1)\rangle>0,\quad\forall\theta_1\neq\theta_2\\ B_{F^*}(\eta_1:\eta_2)+B_{F^*}(\eta_2:\eta_1)&=\langle\eta_2-\eta_1,\nabla F^*(\eta_2)-\nabla F^*(\eta_1)\rangle>0,\quad\forall\eta_1\neq\eta_2\end{aligned}$$

□

Definition 3 (Quasi-arithmetic centers, QACs)). *Let $F:\Theta\to\mathbb{R}$ be a strictly convex and smooth real-valued function of Legendre-type in $\mathcal{F}$. The weighted quasi-arithmetic average of $\theta_1,\ldots,\theta_n$ and $w\in\Delta_{n-1}$ is defined by the gradient map ∇F as follows:*

$$M_{\nabla F}(\theta_1,\ldots,\theta_n;w):=\nabla F^{-1}\left(\sum_i w_i\nabla F(\theta_i)\right), \tag{4}$$

$$=\nabla F^*\left(\sum_i w_i\nabla F(\theta_i)\right), \tag{5}$$

where $\nabla F^*=(\nabla F)^{-1}$ is the gradient map of the Legendre transform F^* of F.

We recover the usual definition of scalar QAMs M_f (Definition 1) when $F(t) = \int_a^t f(u)\mathrm{d}u$ for a strictly increasing or strictly decreasing and continuous function f: $M_f = M_{F'}$ (with $f^{-1} = (F')^{-1}$). Notice that we only need to consider F to be strictly convex or strictly concave and smooth to define a multivariate QAM since $M_{\nabla F} = M_{-\nabla F}$.

Example 1 (Matrix example). Consider the strictly convex function [8] $F : \mathrm{Sym}_{++}(d) \to \mathbb{R}$ with $F(\theta) = -\log\det(\theta)$, where $\det(\cdot)$ denotes the matrix determinant. Function $F(\theta)$ is strictly convex and differentiable [8] on the domain of d-dimensional symmetric positive-definite matrices $\mathrm{Sym}_{++}(d)$ (open convex cone). We have

$$\begin{aligned}
F(\theta) &= -\log\det(\theta), \\
\nabla F(\theta) &= -\theta^{-1} =: \eta(\theta), \\
\nabla F^{-1}(\eta) &= -\eta^{-1} =: \theta(\eta) \\
F^*(\eta) &= \langle \theta(\eta), \eta \rangle - F(\theta(\eta)) = -d - \log\det(-\eta),
\end{aligned}$$

where the dual parameter η belongs to the d-dimensional negative-definite matrix domain, and the inner matrix product is the Hilbert-Schmidt inner product $\langle A, B \rangle := \mathrm{tr}(AB^\top)$, where $\mathrm{tr}(\cdot)$ denotes the matrix trace. It follows that

$$M_{\nabla F}(\theta_1, \theta_2) = 2(\theta_1^{-1} + \theta_2^{-1})^{-1},$$

is the matrix harmonic mean [1] generalizing the scalar harmonic mean $H(a, b) = \frac{2ab}{a+b}$ for $a, b > 0$. Other examples of matrix means are reported in [7].

2.2 Quasi-arithmetic barycenters and dual geodesics

A Bregman generator $F : \Theta \to \mathbb{R}$ induces a dually flat space [4]

$$(\Theta, g(\theta) = \nabla_\theta^2 F(\theta), \nabla, \nabla^*)$$

that we call a Bregman manifold (Hessian manifold with a global chart), where ∇ is the flat connection with Christoffel symbols $\Gamma_{ijk}(\theta) = 0$ and ∇^* is the dual connection with respect to g such that $\Gamma^{*ijk}(\eta) = 0$.

In a Bregman manifold, the primal geodesics $\gamma_\nabla(P, Q; t)$ are obtained as line segments in the θ-coordinate system (because the Christoffel symbols of the connection ∇ vanishes in the θ-coordinate system) while the dual geodesics $\gamma_{\nabla^*}(P, Q; t)$ are line segments in the η-coordinate system (because the Christoffel symbols of the dual connection ∇^* vanishes in the η-coordinate system). The dual geodesics define interpolation schemes $(PQ)^\nabla(t) = \gamma_\nabla(P, Q; t)$ and $(PQ)^{\nabla^*}(t) = \gamma_{\nabla^*}(P, Q; t)$ between input points P and Q with $P = \gamma_\nabla(P, Q; 0) = \gamma_{\nabla^*}(P, Q; 0)$ and $Q = \gamma_\nabla(P, Q; 1) = \gamma_{\nabla^*}(P, Q; 1)$ when t ranges in $[0, 1]$. We express the coordinates of the interpolated points on γ_∇ and γ_{∇^*} using quasi-arithmetic averages as follows:

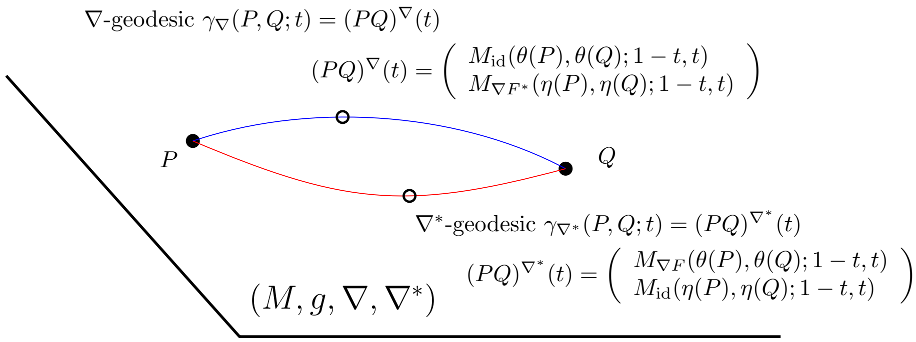

Fig. 1. The points on dual geodesics in a dually flat spaces have dual coordinates expressed with quasi-arithmetic averages.

$$(PQ)^\nabla(t) = \gamma_\nabla(P,Q;t) = \begin{bmatrix} M_{\mathrm{id}}(\theta(P),\theta(Q);1-t,t) \\ M_{\nabla F^*}(\eta(P),\eta(Q);1-t,t) \end{bmatrix}, \tag{6}$$

$$(PQ)^{\nabla^*}(t) = \gamma_{\nabla^*}(P,Q;t) = \begin{bmatrix} M_{\nabla F}(\theta(P),\theta(Q);1-t,t) \\ M_{\mathrm{id}}(\eta(P),\eta(Q);1-t,t) \end{bmatrix}, \tag{7}$$

where id denotes the identity mapping. See Figure 1.

Quasi-arithmetic centers were also used by a geodesic bisection algorithm to approximate the circumcenter of the minimum enclosing balls with respect to the canonical divergence in Bregman manifolds in [21], and for defining the Riemannian center of mass between two symmetric positive-definite matrices with respect to the trace metric in [15]. See also [22, 23].

3 Invariance and equivariance properties

A dually flat manifold [4] (M,g,∇,∇^*) has a canonical divergence [2] D_{∇,∇^*} which can be expressed either as a primal Bregman divergence in the ∇-affine coordinate system θ (using the convex potential function $F(\theta)$) or as a dual Bregman divergence in the ∇^*-affine coordinate system η (using the convex conjugate potential function $F^*(\eta)$), or as dual Fenchel-Young divergences [18] using the mixed coordinate systems θ and η. The dually flat manifold (M,g,∇,∇^*) (a particular case of Hessian manifolds [26] which admit a global coordinate system) is thus characterized by $(\theta,F(\theta);\eta,F^*(\eta))$ which we shall denote by $(M,g,\nabla,\nabla^*) \leftarrow \mathrm{DFS}(\theta,F(\theta);\eta,F^*(\eta))$ (or in short $(M,g,\nabla,\nabla^*) \leftarrow (\Theta,F(\theta))$). However, the choices of parameters θ and η and potential functions $F(\theta)$ and $F^*(\eta)$ are *not unique* since they can be chosen up to affine reparameterizations and additive affine terms [4]: $(M,g,\nabla,\nabla^*) \leftarrow \mathrm{DFS}([\theta,F(\theta);\eta,F^*(\eta)])$ where $[\cdot]$ denotes the equivalence class that has been called purposely the affine Legendre invariance in [14]:

First, consider changing the potential function $F(\theta)$ by adding an affine term: $\bar{F}(\theta) = F(\theta) + \langle c, \theta\rangle + d$. We have $\nabla\bar{F}(\theta) = \nabla F(\theta) + c = \bar{\eta}$. Inverting $\nabla\bar{F}(x) = \nabla F(x) + c = y$, we get $\nabla\bar{F}^{-1}(y) = \nabla F(y - c)$. We check that $B_F(\theta_1 : \theta_2) = B_{\bar{F}}(\theta_1 : \theta_2) = D_{\nabla,\nabla^*}(P_1 : P_2)$ with $\theta(P_1) =: \theta_1$ and $\theta(P_2) =: \theta_2$. It is indeed well-known that Bregman divergences modulo affine terms coincide [5]. For the quasi-arithmetic averages $M_{\nabla\bar{F}}$ and $M_{\nabla F}$, we thus obtain the following invariance property:

$$M_{\nabla\bar{F}}(\theta_1, \ldots; \theta_n; w) = M_{\nabla F}(\theta_1, \ldots; \theta_n; w).$$

Second, consider an affine change of coordinates $\bar{\theta} = A\theta + b$ for $A \in \mathrm{GL}(d)$ and $b \in \mathbb{R}^d$, and define the potential function $\bar{F}(\bar{\theta})$ such that $\bar{F}(\bar{\theta}) = F(\theta)$. We have $\theta = A^{-1}(\bar{\theta} - b)$ and $\bar{F}(x) = F(A^{-1}(x - b))$. It follows that

$$\nabla\bar{F}(x) = (A^{-1})^\top \nabla F(A^{-1}(x - b)),$$

and we check that $B_{\bar{F}(\overline{\theta_1}:\overline{\theta_2})} = B_F(\theta_1 : \theta_2)$:

$$\begin{aligned}
B_{\bar{F}(\bar{F}(\overline{\theta_1}:\overline{\theta_2})} &= \bar{F}(\overline{\theta_1}) - \bar{F}(\overline{\theta_2}) - \left\langle \overline{\theta_1} - \overline{\theta_2}, \nabla\bar{F}(\overline{\theta_2})\right\rangle, \\
&= F(\theta_1) - F(\theta_2) - (A(\theta_1 - \theta_2))^\top (A^{-1})^\top \nabla F(\theta_2), \\
&= F(\theta_1) - F(\theta_2) - (\theta_1 - \theta_2)^\top \underbrace{A^\top (A^{-1})^\top}_{(A^{-1}A)^\top = I} \nabla F(\theta_2) = B_F(\theta_1 : \theta_2).
\end{aligned}$$

This highlights the invariance that $D_{\nabla,\nabla^*}(P_1 : P_2) = B_F(\theta_1 : \theta_2) = B_{\bar{F}(\bar{\theta}_1:\bar{\theta}_2)}$, i.e., the canonical divergence does not change under a reparameterization of the ∇-affine coordinate system. For the induced quasi-arithmetic averages $M_{\nabla\bar{F}}$ and $M_{\nabla F}$, we have $\nabla\bar{F}(x) = (A^{-1})^\top \nabla F(A^{-1}(x - b)) = y$, we calculate

$$x = \nabla\bar{F}(x)^{-1}(y) = A\,\nabla\bar{F}^{-1}(((A^{-1})^\top)^{-1} y) + b,$$

and we have

$$\begin{aligned}
M_{\nabla\bar{F}}(\bar{\theta}_1, \ldots, \bar{\theta}_n; w) &:= \nabla\bar{F}^{-1}(\sum_i w_i \nabla\bar{F}(\bar{\theta}_i)), \\
&= (\nabla\bar{F})^{-1}\left((A^{-1})^\top \sum_i w_i \nabla F(\theta_i)\right), \\
&= A\,\nabla F^{-1}\left(\underbrace{((A^{-1})^\top)^{-1}(A^{-1})^\top}_{=I} \sum_i w_i \nabla F(\theta_i)\right) + b, \\
M_{\nabla\bar{F}}(\bar{\theta}_1, \ldots, \bar{\theta}_n; w) &= A\, M_{\nabla F}(\theta_1, \ldots, \theta_n; w) + b
\end{aligned}$$

More generally, we may define $\bar{F}(\bar{\theta}) = F(A\theta + b) + \langle c, \theta\rangle + d$ and get via Legendre transformation $\bar{F}^*(\bar{\eta}) = F^*(A^*\eta + b^*) + \langle c^*, \eta\rangle + d^*$ (with A^*, b^*, c^* and d^* expressed using A, b, c and d since these parameters are linked by the Legendre transformation).

Third, the canonical divergences should be considered relative divergences (and not absolute divergences), and defined according to a prescribed arbitrary "unit" $\lambda > 0$. Thus we can scale the canonical divergence by $\lambda > 0$, i.e., $D_{\lambda,\nabla,\nabla^*} := \lambda D_{\nabla,\nabla^*}$. We have $D_{\lambda,\nabla,\nabla^*}(P_1 : P_2) = \lambda B_F(\theta_1 : \theta_2) = \lambda B_{F^*}(\eta_2 : \eta_1)$, and $\lambda B_F(\theta_1 : \theta_2) = B_{\lambda F}(\theta_1 : \theta_2)$ (and $\nabla \lambda F = \lambda \nabla F$). We check the scale invariance of quasi-arithmetic averages: $M_{\lambda \nabla F} = M_{\nabla F}$.

Proposition 2 (Invariance and equivariance of QACs). *Let $F(\theta)$ be a function of Legendre type. Then $\bar{F}(\bar{\theta}) := \lambda(F(A\theta+b)+\langle c,\theta\rangle+d)$ for $A \in \mathrm{GL}(d)$, $b, c \in \mathbb{R}^d$, $d \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}_{>0}$ is a Legendre-type function, and we have*

$$M_{\nabla \bar{F}} = A\, M_{\nabla F} + b.$$

This proposition generalizes the invariance property of scalar QAMs, and untangles the role of scale $\lambda > 0$ from the other invariance roles brought by the Legendre transformation.

Consider the Mahalanobis divergence Δ^2 (i.e., the squared Mahalanobis distance Δ) as a Bregman divergence obtained for the quadratic form generator $F_Q(\theta) = \frac{1}{2}\theta^\top Q\theta + c\theta + \kappa$ for a symmetric positive-definite $d \times d$ matrix Q, $c \in \mathbb{R}^d$ and $\kappa \in \mathbb{R}$. We have:

$$\Delta^2(\theta_1, \theta_2) = B_{F_Q}(\theta_1 : \theta_2) = \frac{1}{2}(\theta_2 - \theta_1)^\top\, Q\, (\theta_2 - \theta_1).$$

When $Q = I$, the identity matrix, the Mahalanobis divergence coincides with the Euclidean divergence[2] (i.e., the squared Euclidean distance). The Legendre convex conjugate is

$$F^*(\eta) = \frac{1}{2}\eta^\top Q^{-1}\eta = F_{Q^{-1}}(\eta),$$

and we have $\eta = \nabla F_Q(\theta) = Q\theta$ and $\theta = \nabla F_Q^*(\eta) = Q^{-1}\eta$. Thus we get the following dual quasi-arithmetic averages:

$$M_{\nabla F_Q}(\theta_1, \ldots, \theta_n; w) = Q^{-1}\left(\sum_{i=1}^n w_i Q\theta_i\right) = \sum_{i=1}^n w_i\theta_i = M_{\mathrm{id}}(\theta_1, \ldots, \theta_n; w),$$
$$M_{\nabla F_Q^*}(\eta_1, \ldots, \eta_n; w) = Q\left(\sum_{i=1}^n w_i Q^{-1}\eta_i\right) = M_{\mathrm{id}}(\eta_1, \ldots, \eta_n; w).$$

The dual quasi-arithmetic centers $M_{\nabla F_Q}$ and $M_{\nabla F_Q^*}$ induced by a Mahalanobis Bregman generator F_Q coincide since $M_{\nabla F_Q} = M_{\nabla F_Q^*} = M_{\mathrm{id}}$. This means geometrically that the left-sided and right-sided centroids of the underlying canonical divergences match. The average $M_{\nabla F_Q}(\theta_1, \ldots, \theta_n; w)$ expresses the centroid $C = \bar{C}_R = \bar{C}_L$ in the θ-coordinate system ($\theta(C) = \underline{\theta}$) and the average $M_{\nabla F_Q^*}(\eta_1, \ldots, \eta_n; w)$ expresses the same centroid in the η-coordinate system ($\eta(C) = \underline{\eta}$). In that case of self-dual flat Euclidean geometry, there is an affine

[2] The squared Euclidean/Mahalanobis divergence are not metric distances since they fail the triangle inequality.

transformation relating the θ- and η-coordinate systems:$\eta = Q\theta$ and $\theta = Q^{-1}\eta$. As we shall see this is because the underlying geometry is self-dual Euclidean flat space $(M, g_{\mathrm{Euclidean}}, \nabla_{\mathrm{Euclidean}}, \nabla^*_{\mathrm{Euclidean}} = \nabla_{\mathrm{Euclidean}})$ and that both dual connections coincide with the Euclidean connection (i.e., the Levi-Civita connection of the Euclidean metric). In this particular case, the dual coordinate systems are just related by affine transformations.

4 Quasi-arithmetic mixtures and Jensen-Shannon-type divergences

Consider a quasi-arithmetic mean M_f and n probability distributions $P_1, \ldots, P_n$ all dominated by a measure μ, and denote by $p_1 = \frac{\mathrm{d}P_1}{\mathrm{d}\mu}, \ldots, p_n = \frac{\mathrm{d}P_n}{\mathrm{d}\mu}$ their Radon-Nikodym derivatives. Let us define *statistical* M_f*-mixtures* of $p_1, \ldots, p_n$:

Definition 4. *The M_f-mixture of n densities $p_1, \ldots, p_n$ weighted by $w \in \Delta_n^\circ$ is defined by*

$$(p_1, \ldots, p_n; w)^{M_f}(x) := \frac{M_f(p_1(x), \ldots, p_n(x); w)}{\int M_f(p_1(x), \ldots, p_n(x); w)\mathrm{d}\mu(x)}.$$

The quasi-arithmetic mixture (QAMIX) $(p_1, \ldots, p_n; w)^{M_f}$ generalizes the ordinary statistical mixture $\sum_{i=1}^d w_i p_i(x)$ when $f(t) = t$ and $M_f = A$ is the arithmetic mean. A statistical M_f-mixture can be interpreted as the M_f-integration of its weighted component densities, the densities p_i. The power mixtures $(p_1, \ldots, p_n; w)^{M_p}(x)$ (including the ordinary and geometric mixtures) are called α-mixtures in [3] with $\alpha(p) = 1 - 2p$ (or equivalently $p = \frac{1-\alpha}{2}$). A nice characterization of the α-mixtures is that these mixtures are the *density centroids* of the weighted mixture components with respect to the α-divergences [3] (proven by calculus of variation):

$$(p_1, \ldots, p_n; w)^{M_\alpha} = \arg\min_p \sum_i w_i D_\alpha(p_i, p),$$

where D_α denotes the $\alpha-$divergences [4, 20]. See also the entropic means defined according to f-divergences [6]. M_f-mixtures can also been used to define a generalization of the Jensen-Shannon divergence [17] between densities p and q as follows:

$$D_{\mathrm{JS}}^{M_f}(p, q) := \frac{1}{2}\left(D_{\mathrm{KL}}(p : (pq)^{M_f}) + D_{\mathrm{KL}}(q : (pq)^{M_f})\right) \geq 0, \tag{8}$$

where $D_{\mathrm{KL}}(p : q) = \int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}\mu(x)$ is the Kullback-Leibler divergence, and $(pq)^{M_f} := (p, q; \frac{1}{2}, \frac{1}{2})^{M_f}$. The ordinary JSD is recovered when $f(t) = t$ and $M_f = A$:

$$D_{\mathrm{JS}}(p, q) = \frac{1}{2}\left(D_{\mathrm{KL}}\left(p : \frac{p+q}{2}\right) + D_{\mathrm{KL}}\left(q : \frac{p+q}{2}\right)\right).$$

In general, we may consider quasi-arithmetic paths between densities on the space $\mathcal{P}$ of probability density functions with a common support all dominated by a reference measure. On $\mathcal{P}$, we can build a parametric statistical model called a M_f-mixture family of order n as follows:

$$\mathcal{F}^{M_f}_{p_0,p_1,\ldots,p_n} := \left\{(p_0,p_1,\ldots,p_n;(\theta,1))^{M_f} \ : \ \theta\in\Delta_n^\circ\right\}.$$

In particular, power q-paths have been investigated in [13] with applications in annealing importance sampling and other Monte Carlo methods.

To conclude, let us give a geometric definition of a generalization of the Jensen-Shannon divergence on $\mathcal{P}$ according to an arbitrary affine connection [4, 27] ∇:

Definition 5 (Affine connection-based ∇-Jensen-Shannon divergence). *Let ∇ be an affine connection on the space of densities $\mathcal{P}$, and $\gamma_\nabla(p,q;t)$ the geodesic linking density $p=\gamma_\nabla(p,q;0)$ to density $q=\gamma_\nabla(p,q;1)$. Then the ∇-Jensen-Shannon divergence is defined by:*

$$D^{\mathrm{JS}}_\nabla(p,q) := \frac{1}{2}\left(D_{\mathrm{KL}}\left(p:\gamma_\nabla\left(p,q;\frac{1}{2}\right)\right)+D_{\mathrm{KL}}\left(q:\gamma_\nabla\left(p,q;\frac{1}{2}\right)\right)\right). \quad (9)$$

When $\nabla=\nabla^m$ is chosen as the mixture connection [4], we end up with the ordinary Jensen-Shannon divergence since $\gamma_{\nabla^m}(p,q;\frac{1}{2})=\frac{p+q}{2}$. When $\nabla=\nabla^e$, the exponential connection, we get the geometric Jensen-Shannon divergence [17] since $\gamma_{\nabla^e}(p,q;\frac{1}{2})=(pq)^G$ is a statistical geometric mixture. We may consider the α-connections [4] ∇^α of parametric or non-parametric statistical models, and skew the geometric Jensen-Shannon divergence to define the β-skewed ∇^α-JSD:

$$D^{\mathrm{JS}}_{\nabla^\alpha,\beta}(p,q)=\beta\, D_{\mathrm{KL}}(p:\gamma_{\nabla^\alpha}(p,q;\beta))+(1-\beta)\, D_{\mathrm{KL}}(q:\gamma_{\nabla^\alpha}(p,q;\beta)). \quad (10)$$

A longer technical report of this work is available [19].

References

1. Alić, M., Mond, B., Pečarić, J., Volenec, V.: The arithmetic-geometric-harmonic-mean and related matrix inequalities. Linear Algebra and its Applications **264**, 55–62 (1997)
2. Amari, S.i.: Differential-geometrical methods in statistics. Lecture Notes on Statistics **28**, 1 (1985)
3. Amari, S.i.: Integration of stochastic models by minimizing α-divergence. Neural computation **19**(10), 2780–2796 (2007)
4. Amari, S.i.: Information Geometry and Its Applications. Applied Mathematical Sciences, Springer Japan (2016)
5. Banerjee, A., Merugu, S., Dhillon, I.S., Ghosh, J., Lafferty, J.: Clustering with Bregman divergences. Journal of machine learning research **6**(10) (2005)
6. Ben-Tal, A., Charnes, A., Teboulle, M.: Entropic means. Journal of Mathematical Analysis and Applications **139**(2), 537–551 (1989)

7. Bhatia, R., Gaubert, S., Jain, T.: Matrix versions of the Hellinger distance. Letters in Mathematical Physics **109**(8), 1777–1804 (2019)
8. Boyd, S., Boyd, S.P., Vandenberghe, L.: Convex optimization. Cambridge university press (2004)
9. Bregman, L.M.: The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR computational mathematics and mathematical physics **7**(3), 200–217 (1967)
10. Clarke, F.: On the inverse function theorem. Pacific Journal of Mathematics **64**(1), 97–102 (1976)
11. Dontchev, A.L., Rockafellar, R.T., Rockafellar, R.T.: Implicit functions and solution mappings: A view from variational analysis, vol. 11. Springer (2009)
12. Hardy, G.H., Littlewood, J.E., Pólya, G., Pólya, G.: Inequalities. Cambridge university press (1952)
13. Masrani, V., Brekelmans, R., Bui, T., Nielsen, F., Galstyan, A., Ver Steeg, G., Wood, F.: q-paths: Generalizing the geometric annealing path using power means. In: Uncertainty in Artificial Intelligence. pp. 1938–1947. PMLR (2021)
14. Nakajima, N., Ohmoto, T.: The dually flat structure for singular models. Information Geometry **4**(1), 31–64 (2021)
15. Nakamura, Y.: Algorithms associated with arithmetic, geometric and harmonic means and integrable systems. Journal of computational and applied mathematics **131**(1-2), 161–174 (2001)
16. Niculescu, C., Persson, L.E.: Convex functions and their applications, vol. 23. Springer (2006)
17. Nielsen, F.: On the Jensen–Shannon symmetrization of distances relying on abstract means. Entropy **21**(5), 485 (2019)
18. Nielsen, F.: Statistical Divergences between Densities of Truncated Exponential Families with Nested Supports: Duo Bregman and Duo Jensen Divergences. Entropy **24**(3), 421 (2022)
19. Nielsen, F.: Beyond scalar quasi-arithmetic means: Quasi-arithmetic averages and quasi-arithmetic mixtures in information geometry (2023)
20. Nielsen, F., Nock, R., Amari, S.i.: On clustering histograms with k-means by using mixed α-divergences. Entropy **16**(6), 3273–3301 (2014)
21. Nock, R., Nielsen, F.: Fitting the smallest enclosing Bregman ball. In: European Conference on Machine Learning. pp. 649–656. Springer (2005)
22. Ohara, A.: Geodesics for dual connections and means on symmetric cones. Integral Equations and Operator Theory **50**, 537–548 (2004)
23. Pálfia, M.: Classification of affine matrix means. arXiv preprint arXiv:1208.5603 (2012)
24. Rockafellar, R.T.: Conjugates and Legendre transforms of convex functions. Canadian Journal of Mathematics **19**, 200–205 (1967)
25. Rockafellar, R.T.: Convex analysis, vol. 11. Princeton university press (1997)
26. Shima, H., Yagi, K.: Geometry of Hessian manifolds. Differential geometry and its applications **7**(3), 277–290 (1997)
27. Zhang, J.: Nonparametric information geometry: From divergence function to referential-representational biduality on statistical manifolds. Entropy **15**(12), 5384–5418 (2013)