Generalizing Skew Jensen Divergences and Bregman Divergences With Comparative Convexity

Article *in* IEEE Signal Processing Letters · June 2017

DOI: 10.1109/LSP.2017.2712195

2 authors:

Frank Nielsen
Professor

Richard Nock
Université des Antilles

Generalizing skew Jensen divergences and Bregman divergences with comparative convexity

Frank Nielsen, *Senior Member, IEEE* Richard Nock, *Non-member*

***Abstract*—Comparative convexity is a generalization of ordinary convexity based on abstract means instead of arithmetic means. We introduce the generalized skew Jensen divergences and their corresponding Bregman divergences with respect to comparative convexity. To illustrate those novel families of divergences, we consider the convexity induced by quasi-arithmetic means, and report explicit formula for the corresponding Bregman divergences. In particular, we show that those new Bregman divergences are equivalent to conformal ordinary Bregman divergences on monotone embeddings, and further state related results.**

***Index Terms*—Convexity, regular mean, quasi-arithmetic weighted mean, skew Jensen divergence, Bregman divergence, conformal divergence.**

I. Introduction

Let $F:\mathcal{X}\rightarrow\mathbb{R}$ be a real-valued function. Jensen [1] introduced the notion of convexity of F using the following *midpoint convex property*:

$$\frac{F(p)+F(q)}{2}\geq F\left(\frac{p+q}{2}\right),\quad\forall p,q\in\mathcal{X}.\tag{1}$$

A *continuous function* F obeying this midpoint convexity implies its *convexity property* [2]:

$$\forall p,q,\forall\lambda\in[0,1],\quad F(\lambda p+(1-\lambda)q)\leq\lambda F(p)+(1-\lambda)F(q).\tag{2}$$

When the inequality is strict for distinct points and $\lambda\in(0,1)$, this inequality defines the *strict convex* property of F. Note that a function satisfying only the midpoint convexity inequality may not be continuous [3], and hence not convex. Let $\mathcal{C}$ denote the class of strictly continuous and convex real-valued functions.

Convexity allows one to define classes of *dissimilarity measures* parameterized by functional generators. Burbea and Rao [4] studied the *Jensen difference* for $F\in\mathcal{C}$ as such a family of dissimilarity measures:

$$J_F(p,q):=\frac{F(p)+F(q)}{2}-F\left(\frac{p+q}{2}\right).\tag{3}$$

A dissimilarity $D(p,q)$ is *proper* iff $D(p,q)\geq 0$ with equality iff $p=q$. It follows from the strict midpoint convex property of F that J_F is proper. Nowadays, these Jensen differences are commonly called *Jensen Divergences* (JD), where a *divergence* is a *smooth* dissimilarity measure inducing a dual geometry [5]. One can further define the proper *skew Jensen divergences* for $\alpha\in(0,1)$, see [6], [7]:

$$J_{F,\alpha}(p:q):=(1-\alpha)F(p)+\alpha F(q)-F((1-\alpha)p+\alpha q),\tag{4}$$

with $J_{F,\alpha}(q:p)=J_{F,1-\alpha}(p:q)$. The ":" notation emphasizes the fact that the dissimilarity may be asymmetric. Another popular class of dissimilarities are the Bregman Divergences [8], [9] (BDs):

$$B_F(p:q):=F(p)-F(q)-(p-q)^\top\nabla F(q),\tag{5}$$

where ∇F denotes the gradient of F. Let $J'_{F,\alpha}(p:q):=\frac{1}{\alpha(1-\alpha)}J_{F,\alpha}(p:q)$ denote the scaled skew JDs. Then it was proved that BDs can be obtained as limit cases of skew JDs [6], [7]:

$$B_F(p:q)=\lim_{\alpha\rightarrow 1^-}J'_{F,\alpha}(p:q),\tag{6}$$

$$B_F(q:p)=\lim_{\alpha\rightarrow 0^+}J'_{F,1-\alpha}(p:q).\tag{7}$$

II. Jensen and Bregman divergences with comparative convexity

A. Comparative convexity

The notion of convexity can be generalized by observing that *two arithmetic means* $A(x,y)=\frac{x+y}{2}$ are used in Eq. 1: One in the *domain* of the function (*ie.*, $A(p,q)=\frac{p+q}{2}$), and the other one in the *codomain* of the function (*ie.*, $A(F(p),F(q))=\frac{F(p)+F(q)}{2}$). The branch of *comparative convexity* [2] studies classes $\mathcal{C}_{M,N}$ of (M,N)-strictly convex functions F that satisfies the following *generalized strict midpoint convex* inequality:

$$F\in\mathcal{C}_{M,N}\Leftrightarrow F(M(p,q))<N(F(p),F(q)),\quad\forall p,q\in\mathcal{X},\tag{8}$$

where M and N are two abstract means defined on the domain $\mathcal{X}$ and codomain $\mathbb{R}$, respectively. In the reminder, we shall assume F continuously differentiable.

An *abstract mean* $M(p,q)$ aggregates two values to produce an intermediate quantity that satisfies the *innerness property* [10]:

$$\min\{x,y\}\leq M(x,y)\leq\max\{x,y\}.\tag{9}$$

There are many families of means. For example, the family of *power means* P_δ (Hölder means [11]) is defined by:

$$P_\delta(x,y)=\left(\frac{x^\delta+y^\delta}{2}\right)^{\frac{1}{\delta}},\tag{10}$$

Frank Nielsen is with École Polytechnique (France) and Sony Computer Science Laboratories Inc (Japan). Frank.Nielsen@acm.org

Richard Nock is with Data61, the Australian National University, and the University of Sydney, Australia. Richard.Nock@data61.csiro.au

The arithmetic, harmonic and quadratic means are obtained for $\delta = 1$, $\delta = -1$, and $\delta = 2$, respectively. To get a continuous family of power means for $\delta \in \mathbb{R}$, we define for $\delta = 0$, $P_0(x,y) = \sqrt{xy}$, the geometric mean. Notice that power means satisfy the innerness property, and include in the limit cases the minimum and maximum values: $\lim_{\delta\rightarrow -\infty} P_\delta(x,y) = \min\{x,y\}$ and $\lim_{\delta\rightarrow\infty} P_\delta(x,y) = \max\{x,y\}$. Moreover, the power means are ordered, $P_\delta(x,y) \leq P_{\delta'}(x,y)$ for $\delta' \geq \delta$, a property generalizing the well-known inequality of arithmetic and geometric means [10].

There are many ways to define *parametric* family of means [10]: For example, Appendix B presents the Stolarksy, Lehmer and Gini means, with the Gini means including the power means. Means can also be parameterized by monotone functions: Let us cite the quasi-arithmetic means [12]–[14], the Lagrange means [15], the Cauchy means [16], etc.

B. Generalized skew Jensen divergences

We shall introduce *univariate* divergences for $\mathcal{X} \subset \mathbb{R}$ in the remainder. Multivariate divergences for $\mathcal{X} \subset \mathbb{R}^d$ can be built from univariate divergences component-wise.

Definition 1 (Comparative Convexity Jensen Divergence): The Comparative Convexity Jensen Divergence (ccJD) is defined for a midpoint (M,N)-strictly convex function $F : I \subset \mathbb{R} \rightarrow \mathbb{R}$ by:

$$J_F^{M,N}(p,q) {:=} N(F(p),F(q))) - F(M(p,q)) \qquad (11)$$

It follows from the strict midpoint (M,N)-convexity that the ccJDs are proper: $J_F^{M,N}(p,q) \geq 0$ with equality iff $p = q$.

To define generalized skew Jensen divergences, we need (i) to consider *weighted means* (see [2], p. 3 for the generic construction of a weighted mean), and (ii) to ensure that the divergence is proper. This restrict weighted means to be *regular*:

Definition 2 (Regular mean): A mean M is said *regular* if it is (i) symmetric ($M(p,q) = M(q,p)$), (ii) continuous, (iii) increasing in each variable, and (iv) homogeneous ($M(\lambda p, \lambda q) = \lambda M(p,q), \forall\lambda > 0$).

Power means are regular: They belong to a broader family of regular means, the quasi-arithmetic means. A *quasi-arithmetic mean* is defined for a continuous and strictly increasing function $f : I \subset \mathbb{R} \rightarrow J \subset \mathbb{R}$ as:

$$M_f(p,q) {:=} f^{-1}\left(\frac{f(p)+f(q)}{2}\right). \qquad (12)$$

These means are also called Kolmogorov-Nagumo-de Finetti means [12]–[14]. By choosing $f(x) = x$, $f(x) = \log x$ or $f(x) = \frac{1}{x}$, we obtain the Pythagorean arithmetic, geometric, and harmonic (power) means, respectively. A quasi-arithmetic weighted mean is defined by $M_f(p,q;1-\alpha,\alpha) {:=} f^{-1}\left((1-\alpha)f(p)+\alpha f(q)\right)$ for $\alpha \in [0,1]$. Let $M_\alpha(p,q) {:=} M(p,q;1-\alpha,\alpha)$ denote a shortcut for a weighted regular mean.

A *continuous* function F satisfying the midpoint (M,N)-convex property for regular means M and N is (M,N)-*convex* (Theorem A of [2]):

$$N_\alpha(F(p),F(q)) \geq F(M_\alpha(p,q)), \forall p,q \in \mathcal{X}, \forall\alpha \in [0,1]. \qquad (13)$$

Thus we can define a proper divergence for a strictly (M,N)-convex function when considering regular weighted means:

Definition 3 (Comparative Convexity skew Jensen Divergence): The Comparative Convexity skew α-Jensen Divergence (ccsJD) is defined for a strictly (M,N)-convex function $F \in \mathcal{C}_{M,N} : I \rightarrow \mathbb{R}$ by:

$$J_{F,\alpha}^{M,N}(p:q) {:=} N_\alpha(F(p),F(q)) - F(M_\alpha(p,q)), \qquad (14)$$

where M and N are regular weighted means, and $\alpha \in (0,1)$. Thus Eq.4 can be interpreted as $J_{F,\alpha}(p:q) = A_\alpha(F(p),F(q)) - F(A_\alpha(p,q)) = J_{F,\alpha}^{A,A}(p:q)$, where $A_\alpha(x,y) = (1-\alpha)x + \alpha y$ denotes the weighted arithmetic mean. For regular weighted means, we have $J_{F,\alpha}^{M,N}(q,p) = J_{F,1-\alpha}^{M,N}(p:q)$ since the weighted means satisfy $M_\alpha(p,q) = M_{1-\alpha}(q,p)$. This generalized ccsJD can be extended to a positively weighted set of values by defining a notion of *diversity* [4], [17] as:

Definition 4 (Comparative Convexity Jensen Diversity Index): Let $\{(w_i,x_i)\}_{i=1}^n$ be a set of n positive weighted values so that $\sum_{i=1}^n w_i = 1$. Then the *Jensen diversity index* with respect to the strict (M,N)-convexity of a function F for regular weighted means is:

$$\begin{aligned} J_F^{M,N}&(x_1,\ldots,x_n;w_1,\ldots,w_n) {:=} \\ &N(\{(F(x_i),w_i)\}_i) - F(M(\{(x_i,w_i)\}_i)). \end{aligned} \qquad (15)$$

When both means M and N are set to the arithmetic mean, this diversity index has also been called the *Bregman information* [18] in the context of k-means clustering.

C. Generalized Bregman divergences

By analogy to the ordinary setting, let us define the (M,N)-Bregman divergence as the limit case of a scaled skew (M,N)-ccsJDs. Let ${J'}_{F,\alpha}^{M,N}(p:q) = \frac{1}{\alpha(1-\alpha)} J_{F,\alpha}^{M,N}(p:q)$.

Definition 5 ((M,N)-Bregman divergence): For regular weighted means M and N, the (M,N)-Bregman divergence is defined for a strictly (M,N)-convex function $F : I \rightarrow \mathbb{R}$ by

$$B_F^{M,N}(p:q) \quad {:=} \quad \lim_{\alpha\rightarrow 0^+} {J'}_{F,\alpha}^{M,N}(p:q). \qquad (16)$$

It follows from the symmetry ${J'}_{F,\alpha}(p:q) = {J'}_{F,1-\alpha}(q:p)$ that we get the *reverse Bregman divergence* as:

$$B_F^{M,N}(q:p) = \lim_{\alpha\rightarrow 1^-} {J'}_{F,\alpha}^{M,N}(p:q). \qquad (17)$$

Note that a generalization of Bregman divergences has also been studied by Petz [19] to get generalized quantum relative entropies when considering the arithmetic weighted means: Petz defined the Bregman divergence between two points p and q of a convex set C sitting in a Banach space for a given function $F : C \to \mathcal{B}(\mathcal{H})$ (Banach space induced by a Hilbert space $\mathcal{H}$) as:

$$B_F(p:q) := F(p) - F(q) - \lim_{\alpha \to 0^+} \frac{1}{\alpha}(F(q+\alpha(p-q)) - F(q)). \tag{18}$$

This last equation can be rewritten in our framework as $B_F(p : q) = \lim_{\alpha\to 1^-} \frac{1}{1-\alpha} J_{F,\alpha}^{A,A}(p,q)$ (A referring to the Arithmetic mean, see Section II-A).

In order to have well-defined (M,N)-Bregman divergences, we need to prove that (1) the limits of Eq. 16 and Eq. 17 exist, and (2) that those divergences are proper: $B_F^{M,N}(q:p) \geq 0$ with equality iff $p = q$.

III. Quasi-arithmetic Bregman divergences

For a strictly continuously monotone function, let $M_{\gamma,\alpha}(x,y) = \gamma^{-1}((1-\alpha)\gamma(x)+\alpha\gamma(y))$ denote the weighted quasi-arithmetic mean.

A. A direct formula

By definition, a function $F \in \mathcal{C}_{M_\rho,M_\tau}$ is (ρ,τ)-convex iff $M_\tau(F(p),F(q))) \geq F(M_\rho(p,q))$. This midpoint (ρ,τ)-convexity property with the continuity of F yields the more general definition of (ρ,τ)-convexity $M_{\tau,\alpha}(F(p),F(q))) \geq F(M_{\rho,\alpha}(p,q)), \quad \alpha \in [0,1]$ (See after eq. (12) for the definitions of $M_{\tau,\alpha}$ and $M_{\rho,\alpha}$, for continuous and monotonic functions τ and ρ, respectively). Let us study the generalized Bregman Divergences $B_F^{\rho,\tau}$ obtained when taking the limit:

$$B_F^{\rho,\tau}(q:p) := \lim_{\alpha\to 0} \frac{M_{\tau,\alpha}(F(p),F(q))) - F\left(M_{\rho,\alpha}(p,q)\right)}{\alpha(1-\alpha)}. \tag{19}$$

We state the generalized Bregman divergence formula obtained with respect to quasi-arithmetic comparative convexity:

Theorem 1 (Quasi-arithmetic Bregman divergences): Let $F : I \subset \mathbb{R} \to \mathbb{R}$ be a real-valued strictly (ρ,τ)-convex function defined on an interval I for two strictly monotone and differentiable functions ρ and τ (with ρ' and τ' the respective derivatives). The Quasi-Arithmetic Bregman divergence (QABD) induced by the comparative convexity is:

$$\begin{aligned} B_F^{\rho,\tau}(p:q) &= \frac{\tau(F(p)) - \tau(F(q))}{\tau'(F(q))} - \frac{\rho(p)-\rho(q)}{\rho'(q)}F'(q), \\ &= \kappa_\tau(F(q):F(p)) - \kappa_\rho(q:p)F'(q), \end{aligned} \tag{20}$$

where primes denote derivatives and

$$\kappa_\gamma(x:y) = \frac{\gamma(y)-\gamma(x)}{\gamma'(x)}. \tag{21}$$

Proof:

By taking the first-order Taylor expansion of $\tau^{-1}(x)$ at x_0, we get $\tau^{-1}(x) \simeq_{x_0} \tau^{-1}(x_0) + (x-x_0)(\tau^{-1})'(x_0)$. Using the property of the derivative of an inverse function, $(\tau^{-1})'(x) = \frac{1}{(\tau'(\tau^{-1})(x))}$, it follows that the first-order Taylor expansion of $\tau^{-1}(x)$ is $\tau^{-1}(x) \simeq \tau^{-1}(x_0) + (x-x_0)\frac{1}{(\tau'(\tau^{-1})(x_0))}$. Plugging $x_0 = \tau(p)$ and $x = \tau(p)+\alpha(\tau(q)-\tau(p))$, we get a first-order approximation of the weighted quasi-arithmetic mean $M_{\tau,\alpha}$ when $\alpha \to 0$:

$$M_{\tau,\alpha}(p,q) \simeq p + \frac{\alpha(\tau(q)-\tau(p))}{\tau'(p)}. \tag{22}$$

For example, when $\tau(x) = x$ (*ie.*, arithmetic mean), we have $A_\alpha(p,q) \simeq p + \alpha(q-p)$, when $\tau(x) = \log x$ (*ie.*, geometric mean), we obtain $G_\alpha(p,q) \simeq p + \alpha p \log\frac{q}{p}$, and when $\tau(x) = \frac{1}{x}$ (*ie.*, harmonic mean) we get $H_\alpha(p,q) \simeq p + \alpha(p - \frac{p^2}{q})$. For the regular power means, we have $P_\alpha(p,q) \simeq p + \alpha\frac{q^\delta - p^\delta}{\delta p^{\delta-1}}$. These are first-order weighted mean approximations obtained for small values of α.

Now, consider the comparative convexity skew Jensen Divergence defined by $J_{F,\alpha}^{\tau,\rho}(p:q) = M_{\tau,\alpha}(F(p),F(q)) - F(M_{\rho,\alpha}(p,q))$, and apply a first-order Taylor expansion to get $F(M_{\rho,\alpha}(p,q))) \simeq F\left(p + \frac{\alpha(\rho(q)-\rho(p))}{\rho'(p)}\right) \simeq F(p) + \frac{\alpha(\tau(q)-\tau(p))}{\tau'(p)}F'(p)$. Thus it follows that the Bregman divergence for quasi-arithmetic comparative convexity is $B_F^{\rho,\tau}(q:p) = \lim_{\alpha\to 0} {J'}_\alpha^{\tau,\rho}(p:q) = \frac{\tau(F(q))-\tau(F(p))}{\tau'(F(p))} - \frac{\rho(q)-\rho(p)}{\rho'(p)}F'(p)$, and the reverse Bregman divergence $B_F^{\rho,\tau}(p:q) = \lim_{\alpha\to 1}\frac{1}{\alpha(1-\alpha)}J_\alpha^{\tau,\rho}(p:q) = \lim_{\alpha\to 0}\frac{1}{\alpha(1-\alpha)}J_\alpha^{\tau,\rho}(q:p)$. $\square$

Since power means are regular quasi-arithmetic means, we get the following family of *power mean Bregman divergences*:

Corollary 1 (Power Mean Bregman Divergences): For $\delta_1,\delta_2 \in \mathbb{R}\backslash\{0\}$ with $F \in \mathcal{C}_{P_{\delta_1},P_{\delta_2}}$, we have the family of Power Mean Bregman Divergences (PMBDs):

$$B_F^{\delta_1,\delta_2}(p:q) = \frac{F^{\delta_2}(p) - F^{\delta_2}(q)}{\delta_2 F^{\delta_2-1}(q)} - \frac{p^{\delta_1}-q^{\delta_1}}{\delta_1 q^{\delta_1-1}}F'(q) \tag{23}$$

A sanity check for $\delta_1 = \delta_2 = 1$ let us recover the ordinary Bregman divergence.

B. Quasi-arithmetic Bregman divergences are proper

Appendix A proves that a function $F \in \mathcal{C}_{\rho,\tau}$ iff $G = F_{\rho,\tau} = \tau \circ F \circ \rho^{-1} \in \mathcal{C}$. We still need to prove that QABDs are proper: $B_F^{\rho,\tau}(p:q) \geq 0$ with equality iff $p = q$. Defining the ordinary Bregman divergence on the convex generator $G(x) = \tau(F(\rho^{-1}(x)))$ for a (ρ,τ)-convex function with $G'(x) = \tau(F(\rho^{-1}(x)))' = \frac{1}{(\rho'(\rho^{-1})(x))}F'(\rho^{-1}(x))\tau'(F(\rho^{-1}(x)))$, we get an ordinary Bregman divergence that is, in general, *different* from the generalized quasi-arithmetic Bregman divergence $B_F^{\rho,\tau}$: $B_G(p:q) \neq B_F^{\rho,\tau}(p:q)$ with:

$$\begin{aligned} B_G(p:q) &= \tau(F(\rho^{-1}(p))) - \tau(F(\rho^{-1}(q))) \\ &\quad -(p-q)\frac{F'(\rho^{-1}(q))\tau'(F(\rho^{-1}(q)))}{(\rho'(\rho^{-1})(q))} \end{aligned} \tag{24}$$

A sanity check shows that $B_G(p:q) = B_F^{\rho,\tau}(p:q)$ when $\rho(x) = \tau(x) = x$ (since we have the derivatives $\rho'(x) = \tau'(x) = 1$). Let us notice the following remarkable identity:

$$B_F^{\rho,\tau}(p:q) = \frac{1}{\tau'(F(q))}B_G(\rho(p):\rho(q)). \tag{25}$$

This identity allows us to prove that QABDs are proper divergences.

Theorem 2 (QABDs are proper): The quasi-arithmetic Bregman divergences are proper divergences.

Proof: B_G is a proper ordinary BD, $\tau' > 0$ a positive function since τ is a strictly increasing function, and $\rho(p) = \rho(q)$ iff $p = q$ since ρ is strictly monotonous. It follows that $\frac{1}{\tau'(F(q))}B_G(\rho(p):\rho(q)) \geq 0$ with equality iff $p = q$. $\square$

C. Conformal Bregman divergences on monotone embeddings

A closer look at Eq. 25 allows one to interpret the QABDs $B_F^{\rho,\tau}(p:q)$ as conformal divergences. A conformal divergence [5], [20], [21] $D_\kappa(p:q)$ of a divergence $D(p:q)$ is defined by a positive conformal factor function κ as follows: $D_\kappa(p:q)=\kappa(q)D(p:q)$. An example of Bregman conformal divergence is the total Bregman divergence [22] with $\kappa(q)=\frac{1}{\sqrt{1+\|\nabla F(q)\|^2}}$.

Property 1 (QABDs as conformal BDs): The quasi-arithmetic Bregman divergence $B_F^{\rho,\tau}(p:q)$ amounts to compute an ordinary Bregman conformal divergence in the ρ-embedded space:

$$B_F^{\rho,\tau}(p:q)=\kappa(\rho(q))B_G(\rho(p):\rho(q)), \tag{26}$$

with conformal factor $\kappa(x)=\frac{1}{\tau'(F(\rho^{-1}(x)))}>0$.

D. (ρ,τ)-Jensen-Bregman divergences

In [7], the Jensen divergence J_F was interpreted as a *Jensen-Bregman divergence* defined by:

$$\mathrm{JB}_F(p,q)=\frac{B_F\left(p:\frac{p+q}{2}\right)+B_F\left(q:\frac{p+q}{2}\right)}{2}=\mathrm{JB}_F(q,p). \tag{27}$$

The *Jensen-Shannon divergence* [23] is a Jensen-Bregman divergence for the Shannon information function $F(x)=\sum_{i=1}^d x_i\log x_i$, the negative Shannon entropy: $F(x)=-H(x)$. It turns out that $\mathrm{JB}_F(p,q)=J_F(p,q)$. This identity comes from the fact that the terms $p-\frac{p+q}{2}=\frac{p-q}{2}$ and $q-\frac{p+q}{2}=\frac{q-p}{2}=-\frac{p-q}{2}$ being multiplied by $F'(\frac{p+q}{2})$ cancel out. Similarly, we can define the generalized *Quasi-Arithmetic Jensen-Bregman Divergences* (QAJBDs) as:

$$\mathrm{JB}_F^{\rho,\tau}(p,q)=\frac{B_F^{\rho,\tau}\left(p:M_\rho(p,q)\right)+B_F^{\rho,\tau}\left(q:M_\rho(p,q)\right)}{2}. \tag{28}$$

Consider $\tau=\mathrm{id}$, the identity function. Since $\rho(M_\rho(p,q))=\frac{\rho(p)+\rho(q)}{2}$, and $\rho(p)-\rho(M_\rho(p,q))=\frac{\rho(p)-\rho(q)}{2}=-(\rho(q)-\rho(M_\rho(p,q))$ we get the following identity:

$$\mathrm{JB}_F^{\rho,\mathrm{id}}(p,q)=\frac{F(p)+F(q)}{2}-F(M_\rho(p,q))=J_F^{\rho,\mathrm{id}}(p,q). \tag{29}$$

IV. Concluding remarks

We have introduced generalized (M,N)-Bregman divergences as limit of scaled skew (M,N)-Jensen divergences for regular M and N means. Regular means include power means, quasi-arithmetic means, Stolarsky means, etc. But not all means are regular: For example, the Lehmer mean $L_2(x,y)=\frac{x^2+y^2}{x+y}$ is not increasing and therefore not regular (see Appendix B). We reported closed-form expression for quasi-arithmetic (ρ,τ)-Bregman divergences, prove that those divergences are proper, and show that they can be interpreted as conformal ordinary Bregman divergences on a monotone embedding [24]. This latter observation further let us extend usual Bregman divergence results to quasi-arithmetic Bregman divergences (*eg.*, conformal Bregman k-means [22], conformal Bregman Voronoi diagrams [25]).

Appendix A
Quasi-arithmetic to ordinary convexity criterion

To check whether a function F is (M,N)-convex or not when using quasi-arithmetic means M_ρ and M_τ, we use an equivalent test to ordinary convexity as follows:

Lemma 1 ((ρ,τ)-convexity $\leftrightarrow$ ordinary convexity [26]): Let $\rho: I\rightarrow\mathbb{R}$ and $\tau: J\rightarrow\mathbb{R}$ be two continuous and strictly monotone real-valued functions with τ increasing, then function $F: I\rightarrow J$ is (ρ,τ)-convex iff function $G=F_{\rho,\tau}=\tau\circ F\circ\rho^{-1}$ is (ordinary) convex on $\rho(I)$.

Proof: Let us rewrite the (ρ,τ)-convexity midpoint inequality as follows:

$$\begin{aligned} F(M_\rho(x,y)) &\leq M_\tau(F(x),F(y)),\\ F\left(\rho^{-1}\left(\frac{\rho(x)+\rho(y)}{2}\right)\right) &\leq \tau^{-1}\left(\frac{\tau(F(x))+\tau(F(y))}{2}\right), \end{aligned}$$

Since τ is strictly increasing, we have:

$$(\tau\circ F\circ\rho^{-1})\left(\frac{\rho(x)+\rho(y)}{2}\right)\leq\frac{(\tau\circ F)(x)+(\tau\circ F)(y)}{2}. \tag{30}$$

Let $u=\rho(x)$ and $v=\rho(y)$ so that $x=\rho^{-1}(u)$ and $y=\rho^{-1}(v)$ (with $u,v\in\rho(I)$). Then it comes that:

$$(\tau\circ F\circ\rho^{-1})\left(\frac{u+v}{2}\right)\leq\frac{(\tau\circ F\circ\rho^{-1})(u)+(\tau\circ F\circ\rho^{-1})(v)}{2}. \tag{31}$$

This last inequality is precisely the ordinary midpoint convexity inequality for function $G=F_{\rho,\tau}=\tau\circ F\circ\rho^{-1}$. Thus a function F is (ρ,τ)-convex iff $G=\tau\circ F\circ\rho^{-1}$ is ordinary convex, and vice-versa. $\square$

Appendix B
Lehmer, Gini, and Stolarsky means

The weighted *Lehmer mean* [27] of order δ is defined for $\delta\in\mathbb{R}$ as: $L_\delta(x_1,\ldots,x_n;w_1,\ldots,w_n)=\frac{\sum_{i=1}^n w_i x_i^{\delta+1}}{\sum_{i=1}^n w_i x_i^\delta}$. The Lehmer means intersect with the Hölder means only for the arithmetic, geometric and harmonic means. The family of Lehmer barycentric means can further be encapsulated into the family of *Gini means*: $G_{\delta_1,\delta_2}(x_1,\ldots,x_n;w_1,\ldots,w_n)=\left(\frac{\sum_{i=1}^n w_i x_i^{\delta_1}}{\sum_{i=1}^n w_i x_i^{\delta_2}}\right)^{\frac{1}{\delta_1-\delta_2}}$ when $\delta_1\neq\delta_2$, and $G_{\delta_1,\delta_2}(x_1,\ldots,x_n;w_1,\ldots,w_n)=\left(\prod_{i=1}^n x_i^{w_i x_i^\delta}\right)^{\frac{1}{\sum_{i=1}^n w_i x_i^\delta}}$ when $\delta_1=\delta_2=\delta$. Those families of Gini and Lehmer means are homogeneous means: $G_{\delta_1,\delta_2}(\lambda x_1,\ldots,\lambda x_n;w_1,\ldots,w_n)=\lambda G_{\delta_1,\delta_2}(x_1,\ldots,x_n;w_1,\ldots,w_n)$ for any $\lambda>0$. The family of Gini means include the power means: $G_{0,\delta}=P_\delta$ for $\delta\leq 0$ and $G_{\delta,0}=P_\delta$ for $\delta\geq 0$. The Lehmer and Gini means are not always regular since L_2 is not regular. The Stolarsky regular means are not quasi-arithmetic means nor mean-value means [10], and are defined as follows:

$$S_p(x,y)=\left(\frac{x^p-y^p}{p(x-y)}\right)^{\frac{1}{p-1}},\quad p\notin\{0,1\}. \tag{32}$$

In limit cases, the Stolarsky family of means yields the logarithmic mean (L) when $p\rightarrow 0$: $L(x,y)=\frac{y-x}{\log y-\log x}$, and the *identric mean* (I) when $p\rightarrow 1$: $I(x,y)=\left(\frac{y^y}{x^x}\right)^{\frac{1}{y-x}}$.

REFERENCES

[1] J. L. W. V. Jensen, "Sur les fonctions convexes et les inégalités entre les valeurs moyennes," *Acta mathematica*, vol. 30, no. 1, pp. 175–193, 1906.

[2] C. P. Niculescu and L.-E. Persson, *Convex functions and their applications: A contemporary approach*. Springer Science & Business Media, 2006.

[3] G. Maksa and Z. Páles, "Convexity with respect to families of means," *Aequationes mathematicae*, vol. 89, no. 1, pp. 161–167, 2015. [Online]. Available: http://dx.doi.org/10.1007/s00010-014-0281-7

[4] J. Burbea and C. Rao, "On the convexity of some divergence measures based on entropy functions," *IEEE Transactions on Information Theory*, vol. 28, no. 3, pp. 489–495, 1982.

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

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

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

[8] L. M. Bregman, "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*, vol. 7, no. 3, pp. 200–217, 1967.

[9] F. Nielsen and R. Nock, "Sided and symmetrized Bregman centroids," *IEEE Transactions on Information Theory*, vol. 55, no. 6, pp. 2882–2904, 2009.

[10] P. S. Bullen, D. S. Mitrinovic, and M. Vasic, *Means and their Inequalities*. Springer Science & Business Media, 2013, vol. 31.

[11] O. L. Holder, "Über einen Mittelwertssatz," *Nachr. Akad. Wiss. Gottingen Math.-Phys. Kl.*, pp. 38–47, 1889.

[12] A. N. Kolmogorov, "Sur la notion de moyenne," vol. 12, p. 38891, 1930, acad. Naz. Lincei Mem. Cl. Sci. His. Mat. Natur. Sez.

[13] M. Nagumo, "Über eine klasse der mittelwerte," in *Japanese journal of mathematics: transactions and abstracts*, vol. 7, no. 0. The Mathematical Society of Japan, 1930, pp. 71–79.

[14] B. De Finetti, "Sul concetto di media," vol. 3, p. 36996, 1931, istituto italiano degli attuari.

[15] L. R. Berrone and J. Moro, "Lagrangian means," *Aequationes Mathematicae*, vol. 55, no. 3, pp. 217–226, 1998.

[16] J. Matkowski, "On weighted extensions of Cauchy's means," *Journal of mathematical analysis and applications*, vol. 319, no. 1, pp. 215–227, 2006.

[17] C. R. Rao, "Diversity and dissimilarity coefficients: a unified approach," *Theoretical population biology*, vol. 21, no. 1, pp. 24–43, 1982.

[18] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *Journal of machine learning research*, vol. 6, no. Oct, pp. 1705–1749, 2005.

[19] D. Petz, "Bregman divergence as relative operator entropy," *Acta Mathematica Hungarica*, vol. 116, no. 1-2, pp. 127–131, 2007.

[20] R. Nock, F. Nielsen, and S. Amari, "On conformal divergences and their population minimizers," *IEEE Trans. Information Theory*, vol. 62, no. 1, pp. 527–538, 2016.

[21] F. Nielsen and R. Nock, "Total Jensen divergences: definition, properties and clustering," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 2016–2020.

[22] B. C. Vemuri, M. Liu, S.-i. Amari, and F. Nielsen, "Total Bregman divergence and its applications to DTI analysis," *IEEE Transactions on medical imaging*, vol. 30, no. 2, pp. 475–483, 2011.

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

[24] J. Zhang, "Nonparametric information geometry: From divergence function to referential-representational biduality on statistical manifolds," *Entropy*, vol. 15, no. 12, pp. 5384–5418, 2013.

[25] J.-D. Boissonnat, F. Nielsen, and R. Nock, "Bregman Voronoi diagrams," *Discrete & Computational Geometry*, vol. 44, no. 2, pp. 281–307, 2010.

[26] J. Aczél, "A generalization of the notion of convex functions," *Det Kongelige Norske Videnskabers Selskabs Forhandlinger, Trondheim*, vol. 19, no. 24, pp. 87–90, 1947.

[27] G. Beliakov, H. B. Sola, and T. C. Sánchez, *A practical guide to averaging functions*. Springer, 2015, vol. 329.