Quasiconvex Jensen Divergences and Quasiconvex Bregman Divergences

Frank Nielsen[1](✉) and Gaëtan Hadjeres[2]

[1] Sony Computer Science Laboratories Inc., Tokyo, Japan
Frank.Nielsen@acm.org
[2] Sony Computer Science Laboratories, Paris, France
Gaetan.Hadjeres@sony.com

Abstract. We first introduce the class of strictly quasiconvex and strictly quasiconcave Jensen divergences which are asymmetric distances, and study some of their properties. We then define the strictly quasiconvex Bregman divergences as the limit case of scaled and skewed quasiconvex Jensen divergences, and report a simple closed-form formula which shows that these divergences are only pseudo-divergences at countably many inflection points of the quasiconvex generators. To remedy this problem, we propose the δ-averaged quasiconvex Bregman divergences which integrate the pseudo-divergences over a small neighborhood in order obtain a proper divergence. The formula of δ-averaged quasiconvex Bregman divergences extend even to non-differentiable strictly quasiconvex generators. These quasiconvex Bregman divergences between distinct elements have the property to always have one orientation finite while the reverse orientation is infinite. We show that these quasiconvex Bregman divergences can also be interpreted as limit cases of generalized skewed Jensen divergences with respect to comparative convexity by using power means. Finally, we illustrate how these quasiconvex Bregman divergences naturally appear as equivalent divergences for the Kullback-Leibler divergences between probability densities belonging to a same parametric family of distributions with nested density supports.

1 Introduction, Motivation, and Contributions

A *dissimilarity* $D(O, O')$ is a measure of the deviation of an object O' from a reference object O (i.e., $D_O(O') := D(O, O')$) which satisfies the following two basic properties:

- Non-negativity property: $D(O, O') \geq 0, \forall O, O'$
- Property of the indiscernibles: $D(O, O') = 0$ if and only if $O = O'$.

In other words, a dissimilarity $D(O, O')$ satisfies $D(O, O') \geq 0$ with equality if and only if $O = O'$. A *pseudo-dissimilarity* is a measure of deviation for which the non-negativity property holds but not necessarily the law of the indiscernibles [35]. The objects O and O' can be vectors, probability distributions, random variables, strings, graphs, etc. In general, a dissimilarity may not be

F. Barbaresco and F. Nielsen (Eds.): SPIGL 2020, PROMS 361, pp. 1–23, 2021.
https://doi.org/10.1007/978-3-030-77957-3_11

symmetric, i.e., we may potentially have $D(O, O') \neq D(O', O)$. In that case, the dissimilarity is said to be *oriented*, and we consider the following two reference orientations of the dissimilarity: the *forward ordinary dissimilarity* $D(O : O')$ and its associated *reverse dissimilarity* $D^r(O : O') := D(O' : O)$. Notice that we used the ':' notation instead of the usual comma delimiter ',' between the dissimilarity arguments to emphasize that the dissimilarity may be asymmetric. In the literature, a dissimilarity is also commonly called *a divergence* [3] although several additional meanings may be associated to this term like a dissimilarity between *probability distributions* instead of vectors (e.g., the Kullback-Leibler divergence [13] in information theory) or like a notion of smoothness (e.g., a C^3 contrast function in information geometry [3]). A dissimilarity may also be loosely called a *distance* although this may convey to mathematicians in some contexts the additional notion of a dissimilarity satisfying the metric axioms (non-negativity, law of the indiscernibles, symmetry and triangular inequality).

The *Bregman divergences* [10,11] were introduced in operations research, and are widely used nowadays in machine learning and information sciences. For a strictly convex and smooth generator F, called the *Bregman generator*, we define the corresponding Bregman divergence between parameter vectors θ and θ' as:

$$B_F(\theta : \theta') := F(\theta) - F(\theta') - (\theta - \theta')^\top \nabla F(\theta'). \tag{1}$$

Bregman divergences are always finite when θ and θ' both belong an open convex set Θ, and generalize many common distances [5], including the Kullback-Leibler (KL) divergence and the squared Euclidean and squared Mahalanobis distances. (Notice that although the Mahalanobis distance is a metric, the squared Mahalanobis distance is not a metric.) Furthermore, the KL divergence between two probability densities belonging to a same exponential family [5,6] amount to a *reverse Bregman divergence* between the corresponding parameters when setting the Bregman generator to be the cumulant function of the exponential family [4]. Moreover, a one-to-one correspondence (bijection) between regular exponential families [6] and the so-called class of "regular Bregman divergences" was reported in [5] and used for learning statistical mixtures showing that the expectation-maximization (EM) algorithm is equivalent to a Bregman clustering algorithm with soft membership. Bregman divergences have been extended to many non-vector data types like matrix arguments [36] or functional arguments [17].

In this chapter, we consider defining the notion of Jensen divergences [29] for strictly quasiconvex or strictly quasiconcave generators, and we investigate the induced notion of Bregman divergences. We term them *quasiconvex Bregman divergences* (and omit to prefix it by 'strictly' for sake of brevity). We then establish a connection between the KL divergence between parametric families of densities with nested density supports and these quasiconvex Bregman divergences.

We summarize our main contributions as follows:

- By using quasiconvex generators instead of convex generators (i.e., relaxing convex functions to quasiconvex functions), we define for a scalar $\alpha \in (0, 1)$ the α-skewed quasiconvex Jensen divergences (Definition 1) and derived thereof the quasiconvex Bregman divergences in the limit case of $\alpha \rightarrow 1$ or the reverse quasiconvex Bregman divergences when $\alpha \rightarrow 0$ (Definition 3 and Theorem 1). The quasiconvex Bregman divergences turn out to be only pseudo-divergences at inflection points of the generator. Since this can happen only at countably many points, we still loosely call them quasiconvex Bregman divergences. Then we integrate the quasiconvex Bregman (pseudo-) divergence over a small neighborhood and obtain a δ-averaged quasiconvex Bregman divergence in Sect. 3.2. The δ-averaged quasiconvex Bregman divergence are also well-defined for strictly quasiconvex but not differentiable generators. Quasiconvex Bregman divergences between distinct parameters always have one orientation finite while the reverse orientation evaluates to infinity.

- We show that quasiconvex Jensen divergences and quasiconvex Bregman divergences can be reinterpreted as generalized Jensen and Bregman divergences with *comparative convexity* [26,33] using power means in the limit case (Sect. 2.3 and Sect. 2.3).
- We exhibit some parametric families of probability distributions with strictly nested density supports such that the Kullback-Leibler divergences between them amount to equivalent quasiconvex Bregman divergences (Sect. 4).

The paper is organized as follows:

Section 2 defines the quasiconvex and quasiconcave difference distances by analogy to Jensen difference distances [38] (also called Burbea-Rao divergences [29]), study some of their properties, and show how to obtain them as generalized Jensen divergences [33] obtained from comparative convexity using power means. Henceforth their name: *quasiconvex Jensen divergences*. When the generator is quasilinear instead of quasiconvex, we call them *quasilinear Jensen divergences* (qln Jensen divergences for short). We then define the quasiconvex Bregman divergences in Sect. 3 as limit cases of scaled and skewed quasiconvex Jensen divergences, and report a closed-form formula which highlights the fact that one orientation of the distance is always finite while the reverse orientation is always infinite (for divergences between distinct elements). Since the quasiconvex Bregman divergences are only pseudo-divergences at inflection points, we define the δ-averaged quasiconvex Bregman divergences in Sect. 3.2. We also recover the formula by taking the limit case of power means Bregman divergences that were introduced using comparative convexity [33].

In Sect. 4, we consider the problem of finding a parametric family of probability distributions for which the Kullback-Leibler divergence amount to a quasiconvex Bregman divergence. We illustrate one example showing that nested supports of the densities ensure the property of having one orientation finite while the other one is infinite. Finally, Sect. 5 concludes this chapter and hints at applications perspectives of these quasiconvex Bregman divergences, including flat center-based clustering [31] and hierarchical clustering [27].

2 Divergences Based on Inequality Gaps of Quasiconvex or Quasiconcave Generators

2.1 Quasiconvex and Quasiconcave Difference Dissimilarities

In this work, a divergence or a distance $D(\theta : \theta')$ refers to a dissimilarity such that $D(\theta : \theta') \geq 0$ with equality iff. $\theta = \theta'$. A pseudo-divergence or pseudo-distance only satisfies the non-negativity property but not necessarily the law of the indiscernibles of the dissimilarities.

Consider a function $Q : \Theta \subset \mathbb{R}^D \rightarrow \mathbb{R}$ which satisfies the following "Jensen-type" inequality [9] for any $\alpha \in (0, 1)$:

$$Q((\theta\theta')_\alpha) < \max\{Q(\theta), Q(\theta')\}, \quad \theta \neq \theta' \in \Theta \subset \mathbb{R}, \tag{2}$$

where $(\theta\theta')_\alpha := (1-\alpha)\theta + \alpha\theta'$ denotes the *weighted linear interpolation* of θ with θ', and Θ the parameter space. Function Q is said *strictly quasiconvex* [8,9,18,37] as it relaxes the strict convexity inequality:

$$Q((\theta\theta')_\alpha) < (1-\alpha)Q(\theta) + \alpha Q(\theta') \leq \max\{Q(\theta), Q(\theta')\}. \tag{3}$$

Let $\mathcal{Q}$ denote the space of such *strictly quasiconvex real-valued function*, and let $\mathcal{C}$ denote the space of strictly convex functions. We have $\mathcal{C} \subset \mathcal{Q}$: Any strictly convex function or any strictly increasing function is quasiconvex, but not necessarily the converse: Some examples of quasiconvex functions which are *not* convex are $Q(\theta) = \sqrt{\theta}$, $Q(\theta) = \theta^3$, $Q(\theta, \theta') = \log(\theta^2 + (\theta')^2)$, etc. Decreasing and then increasing functions are quasiconvex but may not be necessarily smooth. Some concave functions like $Q(\theta) = \log \theta$ are quasiconvex. The sum of quasiconvex functions are not necessarily quasiconvex (see also [7]). In the same spirit that function convexity can be reduced to set convexity via the epigraph representation of the function, a function Q is quasiconvex if the *level set* $L_\alpha := \{x : Q(x) \leq \alpha\}$ is (set) convex for all $\alpha \in \mathbb{R}$. When Q is univariate, a quasiconvex function is also commonly called *unimodal* (i.e., decreasing and then increasing function). Thus a multivariate quasiconvex function can be characterized as being unimodal along each line of its domain. Figure 1 displays some examples of quasiconvex functions with one function that fails to be quasiconvex. Notice that strictly monotonic functions which are *both* strictly quasiconvex and strictly quasiconcave are termed *strictly quasilinear*. The ceil function $\mathrm{ceil}(\theta) = \inf\{z \in \mathbb{Z} \ : \ z \geq \theta\}$ is an example of quasilinear function (idem for the floor function). Another example, are the linear fractional functions $Q_{a,b,c,d}(\theta) = \frac{a^\top\theta+b}{c^\top\theta+d}$ which are quasilinear functions on the domain $\Theta = \{\theta \ : \ c^\top\theta + d > 0\}$. We denote by $\mathcal{L} \subset \mathcal{Q}$ the set of strictly quasilinear functions, and by $\mathcal{H}$ the set of strictly quasiconcave functions.

Definition 1 (Quasiconvex difference distance). The *quasiconvex difference distance* (or qcvx distance for short) for $\alpha \in (0, 1)$ is defined as the inequality difference gap of Eq. 2

$$\begin{aligned} {}^{\mathrm{qcvx}}J_Q^\alpha(\theta : \theta') &:= \max\{Q(\theta), Q(\theta')\} - Q((\theta\theta')_\alpha) \geq 0, & (4)\\ &= \max\{Q(\theta), Q(\theta')\} - Q((1-\alpha)\theta + \alpha\theta'). & (5)\end{aligned}$$

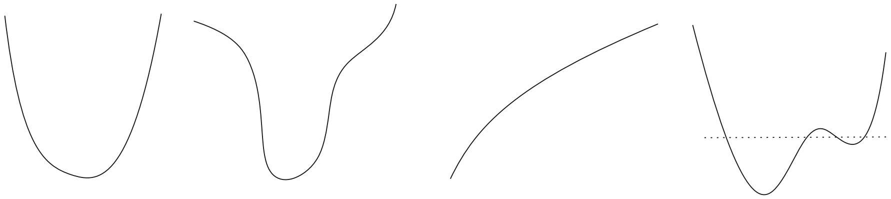

Fig. 1. The first three functions (from left to right) are quasiconvex because any level set is convex, but the last function is not quasiconvex because the dotted line intersects the function in four points (and therefore the level set is not convex). The first function is convex, the second function is quasiconvex but not convex (a chord may intersect the function in more than two points), the third function is monotonous and here concave (quasilinear)).

By definition, the quasiconvex difference distance is a dissimilarity satisfying ${}^{\mathrm{qcvx}}J_Q^{\alpha}(\theta : \theta') = 0$ iff. $\theta = \theta'$ when the generator Q is *strictly* quasiconvex (see Eq. 2).

Remark 1. Notice that we could also have defined a *log-ratio gap* [35] as a dissimilarity:

$$
{}^{\mathrm{qcvx}}JL_Q^{\alpha}(\theta : \theta') := -\log\left(\frac{Q((\theta\theta')_\alpha)}{\max\{Q(\theta), Q(\theta')\}}\right). \tag{6}
$$

However, in that case we should have required the extra condition that the generator does not vanish in the domain, i.e., $Q(\theta) \neq 0$ for any $\theta \in \Theta$.

Property 1. Let $a > 0$ and $b \in \mathbb{R}$, and define $Q_{a,b}(\theta) = aQ(\theta) + b$. Functions $Q_{a,b}$ are quasiconvex, and ${}^{\mathrm{qcvx}}J_{Q_{a,b}}^{\alpha}(\theta : \theta') = a\ {}^{\mathrm{qcvx}}J_Q^{\alpha}(\theta : \theta')$.

Similarly, we can characterize a *strictly quasiconcave real-valued function* $H \in \mathcal{H} : \Theta \subset \mathbb{R}^D \rightarrow \mathbb{R}$ by the following inequality for $\alpha \in (0, 1)$:

$$
H((\theta\theta')_\alpha) > \min\{H(\theta), H(\theta')\}, \quad \theta \neq \theta' \in \Theta \subset \mathbb{R}^D. \tag{7}
$$

This allows one to define the *quasiconcave difference distance* (or qccv distance for short):

Definition 2 (Quasiconcave difference distance). For H a quasiconcave function and $\alpha \in (0, 1)$, we define the quasiconcave distance as:

$$
\begin{aligned}
{}^{\mathrm{qccv}}J_H^{\alpha}(\theta : \theta') &:= H((\theta\theta')_\alpha) - \min\{H(\theta), H(\theta')\}, \qquad (8)\\
&= H((1-\alpha)\theta + \alpha\theta') - \min\{H(\theta), H(\theta')\} \qquad (9)
\end{aligned}
$$

Similarly, we have ${}^{\mathrm{qccv}}J_{H_{a,b}}^{\alpha}(\theta : \theta') = a\ {}^{\mathrm{qccv}}J_H^{\alpha}(\theta : \theta')$ for $a > 0$ and $b \in \mathbb{R}$.

Now, observe that for any $a, b \in \mathbb{R}$, we have[1] $\min\{a, b\} = -\max\{-a, -b\}$ (or equivalently $\max\{a, b\} = -\min\{-a, -b\}$). Thus it follows the following identity:

[1] Indeed, $\max\{a, b\} = \frac{a+b}{2} + \frac{1}{2}|b-a| = -(\frac{-a-b}{2} - \frac{1}{2}|b-a|) = -(\frac{-a-b}{2} - \frac{1}{2}|-b+a|) = -\min\{-a, -b\}$.

Property 2. A quasiconcave difference distance with quasiconcave generator H is equivalent to a quasiconvex difference distance for the quasiconvex generator $Q = -H$:

$$ {}^{\mathrm{qccv}}J_H^\alpha(\theta:\theta') = {}^{\mathrm{qcvx}}J_{-H}^\alpha(\theta:\theta'), \quad {}^{\mathrm{qcvx}}J_Q^\alpha(\theta:\theta') = {}^{\mathrm{qccv}}J_{-Q}^\alpha(\theta:\theta'). \tag{10}$$

Proof.

$$\begin{aligned} {}^{\mathrm{qccv}}J_H^\alpha(\theta:\theta') &= H((\theta\theta')_\alpha) - \min\{H(\theta), H(\theta')\}, && (11)\\ &= \max\{-H(\theta), -H(\theta')\} - (-H((\theta\theta')_\alpha)), && (12)\\ &= {}^{\mathrm{qcvx}}J_{-H}^\alpha(\theta:\theta'). && (13)\end{aligned}$$

□

Therefore, we consider without loss of generality quasiconvex difference distances in the reminder.

2.2 Relationship of Quasiconvex Difference Distances with Jensen Difference Distances

Since for any $a, b \in \mathbb{R}$, we have $\max(a,b) = \frac{a+b}{2} + \frac{1}{2}|b-a|$, $\min(a,b) = \frac{a+b}{2} - \frac{1}{2}|b-a|$ and $\max(a,b) - \min(a,b) = |b-a|$, we can rewrite Eq. 4 to get

$$\begin{aligned} {}^{\mathrm{qcvx}}J_Q^\alpha(\theta:\theta') &= \frac{Q(\theta)+Q(\theta')}{2} + \frac{1}{2}\left|Q(\theta)-Q(\theta')\right| - Q((\theta\theta')_\alpha), && (14)\\ &= \mathrm{eJ}_Q^\alpha(\theta:\theta') + \frac{1}{2}\left|Q(\theta)-Q(\theta')\right| + Q(\theta)\left(\alpha-\frac{1}{2}\right) + Q(\theta')\left(\frac{1}{2}-\alpha\right), && (15)\end{aligned}$$

where

$$\mathrm{eJ}_Q^\alpha(\theta,\theta') := (Q(\theta)Q(\theta'))_\alpha - Q\left((\theta\theta')_\alpha\right), \tag{16}$$

is called the *extended Jensen divergence*, a Jensen-type divergence *extended* to quasiconvex generators instead of ordinary convex generators.

Property 3 (Upperbounded the extended Jensen divergence by ${}^{\mathrm{qcvx}}J_Q^\alpha$). We have:

$$\mathrm{eJ}_Q^\alpha(\theta:\theta') \le {}^{\mathrm{qcvx}}J_Q^\alpha(\theta:\theta') \tag{17}$$

since $(Q(\theta)Q(\theta'))_\alpha \le \max\{Q(\theta), Q(\theta')\}$. In particular, when $Q = F$ is strictly convex, we have $0 \le J_F^\alpha(\theta:\theta') \le {}^{\mathrm{qcvx}}J_F^\alpha(\theta:\theta')$.

Notice that $\mathrm{eJ}_Q^\alpha(\theta,\theta') \ge 0$ when Q is strictly convex, but may be negative when only quasiconvex. For example, $Q(\theta) = \log\theta$ is a quasiconvex and concave function, and therefore $\mathrm{eJ}_Q^\alpha(\theta,\theta') \le 0$.

When $\alpha = \frac{1}{2}$, we get the following identity:

Property 4 (Regularization of extended Jensen divergences).

$$
\begin{aligned}
{}^{\mathrm{qcvx}}J_Q(\theta : \theta') &= \frac{Q(\theta)+Q(\theta')}{2} + \frac{1}{2}|Q(\theta)-Q(\theta')| - Q\left(\frac{\theta+\theta'}{2}\right), \qquad (18)\\
&= \mathrm{eJ}_Q(\theta,\theta') + \frac{1}{2}|Q(\theta)-Q(\theta')|, \qquad (19)
\end{aligned}
$$

where

$$\mathrm{eJ}_Q(\theta,\theta') := \frac{Q(\theta)+Q(\theta')}{2} - Q\left(\frac{\theta+\theta'}{2}\right), \tag{20}$$

is an *extension* of the Jensen divergence [12,38] to a quasiconvex generator Q.

Thus when the generator is convex, we can interpret the quasiconvex divergence as a ℓ_1-regularization of the ordinary Jensen divergence. When the generator Q is not convex, beware that $\mathrm{eJ}_Q(\theta,\theta')$ may be negative but we always have $\mathrm{eJ}_Q(\theta,\theta') \geq -\frac{1}{2}|Q(\theta)-Q(\theta')|$.

Similarly, when the generator H is strictly quasiconcave, we rewrite the quasiconvex difference distance as

$$
\begin{aligned}
{}^{\mathrm{qccv}}J_H(\theta : \theta') &= H\left(\frac{\theta+\theta'}{2}\right) - \frac{H(\theta)+H(\theta')}{2} + \frac{1}{2}|H(\theta)-H(\theta')|, \qquad (21)\\
&= \mathrm{eJ}_{-H}(\theta,\theta') + \frac{1}{2}|H(\theta)-H(\theta')|. \qquad (22)
\end{aligned}
$$

2.3 Quasiconvex Difference Distances from the Viewpoint of Comparative Convexity

In [33], a generalization of the skewed Jensen divergences with respect to comparative convexity [26] is obtained using a pair of weighted means. A *mean* between two reals x and y belonging to an interval $I \subset \mathbb{R}$ is a bivariate function $M(x,y)$ such that

$$\min\{x,y\} \leq M(x,y) \leq \max\{x,y\}. \tag{23}$$

That is, a mean satisfies the *in-betweenness property* (see [26], p. 328). A *weighted mean* M_α for $\alpha \in [0,1]$ can always be built from a mean by using the unique dyadic expansions of real numbers [26]. That is, we define a weighted mean $M(x,y;w) := M(x,y;w,1-w)$ for a weight $w \in [0,1]$ as follows: Set $M(x,y;1,0) = x$, $M(x,y;0,1) = y$, and $M(x,y;\frac{1}{2},\frac{1}{2}) := M(x,y)$. Then we define the weighted means $M\left(x,y;\frac{3}{4},\frac{1}{4}\right) := M(M(x,y),x)$ and $M\left(x,y;\frac{1}{4},\frac{3}{4}\right) := M(M(x,y),y)$, and so on by induction for dyadic numbers with finite binary representations. Thus we can define the weighted mean $M(p,q;w)$ for $w = \frac{i}{2^k}$ with $i \in \{0,\ldots,2^k\}$. For example, when $k=3$, we get the following weighted means:

$$M\left(p,q;\frac{0}{8}=0\right)=q$$

$$M\left(p,q;\frac{1}{8}\right)=M(M(M(p,q),q),q)$$

$$M\left(p,q;\frac{2}{8}=\frac{1}{4}\right)=M(M(p,q),q)$$

$$M\left(p,q;\frac{3}{8}\right)=M(M(M(p,q),p),q)$$

$$M\left(p,q;\frac{4}{8}=\frac{1}{2}\right)=M(p,q)$$

$$M\left(p,q;\frac{5}{8}\right)=M(M(M(p,q),q),p)$$

$$M\left(p,q;\frac{6}{8}=\frac{3}{4}\right)=M(M(p,q),p)$$

$$M\left(p,q;\frac{7}{8}\right)=M(M(M(p,q),p),p)$$

$$M\left(p,q;\frac{8}{8}=1\right)=p$$

Let $w=\sum_{i=1}^{\infty}\frac{d_i}{2^i}$ be the unique dyadic expansion of the real $w\in(0,1)$ where the d_i's are binary digits (i.e., $d_i\in\{0,1\}$). Finally, we define the *weighted mean* $M(x,y,w,1-w)$ of two positive reals p and q for a real weight $w\in(0,1)$ as

$$M(x,y,w,1-w):=\lim_{n\to\infty}M\left(x,y,\sum_{i=1}^{n}\frac{d_i}{2^i},1-\sum_{i=1}^{n}\frac{d_i}{2^i}\right). \quad (24)$$

Consider two *weighted means* M_α and N_α.

A function F is said (M,N) *convex* if and only if we have

$$N_\alpha(F(\theta),F(\theta'))\geq F(M_\alpha(\theta,\theta')),\quad \theta,\theta'\in\Theta. \quad (25)$$

We recover the ordinary convexity (Jensen's midpoint convexity) when $M_\alpha=N_\alpha=A_\alpha$, where $A_\alpha(x,y)=(1-\alpha)x+\alpha y$ is the weighted arithmetic mean.

We can define the α-skewed (M,N)-*Jensen divergence* as:

$$J_{F,\alpha}^{M,N}(\theta:\theta'):=N_\alpha(F(\theta),F(\theta'))-F(M_\alpha(\theta,\theta')). \quad (26)$$

By definition, $J_{F,\alpha}^{M,N}(\theta:\theta')\geq 0$ when F is a (M,N)-strictly convex function.

A *quasi-arithmetic mean* [26] is defined for a continuous strictly increasing function $f:I\subset\mathbb{R}\to J\subset\mathbb{R}$ as:

$$M_f(p,q):=f^{-1}\left(\frac{f(p)+f(q)}{2}\right). \quad (27)$$

These quasi-arithmetic means are also called Kolmogorov-Nagumo-de Finetti means [14,22,25]. Without loss of generality, we assume strictly increasing functions instead of monotonic functions since $M_{-f} = M_f$. By choosing $f(x) = x$, $f(x) = \log x$ or $f(x) = \frac{1}{x}$, we recover the Pythagorean arithmetic, geometric, and harmonic means, respectively. Choosing $f_{\mathrm{LSE}}(x) = \exp(x)$ (with $f_{\mathrm{LSE}}^{-1}(x) = \log(x)$), we get a mean related to the log-sum-exp function LSE [34]: $M_{f_{\mathrm{LSE}}}(p,q) = \log \frac{e^p+e^q}{2} = \mathrm{LSE}(p,q) - \log 2$, where

$$\max\{p,q\} < \mathrm{LSE}(p,q) := \log(e^p + e^q) \leq \max\{p,q\} + \log 2.$$

Now, consider the family of *power means* for $x, y > 0$:

$$P_0(x,y) := \sqrt{xy}, \tag{28}$$

and

$$P_\delta(x,y) := \left(\frac{x^\delta + y^\delta}{2}\right)^{\frac{1}{\delta}}, \quad \delta \neq 0. \tag{29}$$

These means fall in the class of quasi-arithmetic means obtained for $f_\delta(x) = x^\delta$ for $\delta \neq 0$ with $I = J = (0,\infty)$, and include in the limit cases the maximum and minimum values: $\lim_{\delta\to+\infty} P_\delta(a,b) = \max\{a,b\}$ and $\lim_{\delta\to-\infty} P_\delta(a,b) = \min\{a,b\}$.

The *power mean Jensen divergence* [33] is defined as a special case of the (M,N)-Jensen divergence by:

$$J_F^{P_\delta}(\theta:\theta') := J_F^{A,P_\delta}(\theta:\theta') = P_\delta(F(\theta),F(\theta')) - F((\theta\theta')_\alpha), \tag{30}$$

for a (A, P_δ) strictly convex generator F.

Let us now observe that the quasiconvex difference distance is a *limit case* of power mean Jensen divergences:

Property 5 (${}^{\mathrm{qcvx}}J_Q$ *as a limit case of power mean Jensen divergences).* We have

$${}^{\mathrm{qcvx}}J_Q(\theta:\theta') = \lim_{\delta\to\infty} J_Q^{P_\delta}(\theta:\theta'). \tag{31}$$

Notice that a strictly quasiconvex function Q is interpreted as a $(A,\max)$-strictly convex function in comparative convexity, a limit case of (A,P_δ)-convexity when $\delta \to \infty$. From now on, we term the quasiconvex difference distance the *quasiconvex Jensen divergence*.

3 Bregman Divergences for Quasiconvex Generators

3.1 Quasiconvex Bregman Divergences as Limit Cases of Quasiconvex Jensen Divergences

Recall that for a strictly quasiconvex generator Q, define the *α-skewed quasiconvex distance* for $\alpha \in (0,1)$ as

$${}^{\mathrm{qcvx}}J_Q^\alpha(\theta:\theta') := \max\{Q(\theta),Q(\theta')\} - Q((\theta\theta')_\alpha). \tag{32}$$

We have

$$ {}^{\mathrm{qcvx}}J_Q^{\alpha}(\theta:\theta') \geq 0, \tag{33}$$

with equality if and only if $\theta=\theta'$. Notice that we do not require smoothness [20] of Q, and ${}^{\mathrm{qcvx}}J_Q = {}^{\mathrm{qcvx}}J_Q^{\frac{1}{2}}$ is symmetric. For an asymmetric divergence $D(\theta:\theta')$, denote $D^r(\theta:\theta')=D(\theta':\theta)$ the *reverse divergence*.

By analogy to Bregman divergences [5] being interpreted as *limit cases* of scaled and skewed Jensen divergences [29,41]:

$$\lim_{\alpha\rightarrow 1^-} \frac{1}{\alpha(1-\alpha)} J_F^{\alpha}(\theta:\theta') = B_F(\theta:\theta'), \tag{34}$$

$$\lim_{\alpha\rightarrow 0^+} \frac{1}{\alpha(1-\alpha)} J_F^{\alpha}(\theta:\theta') = B_F^r(\theta:\theta') = B_F(\theta':\theta). \tag{35}$$

Let us define the following divergence:

Definition 3 (Quasiconvex Bregman pseudo-divergence). For a strictly quasiconvex generator $Q\in\mathcal{L}$, we define the quasiconvex Bregman pseudo-divergence as

$$ {}^{\mathrm{qcvx}}B_Q(\theta:\theta') := \lim_{\alpha\rightarrow 1^-} \frac{1}{\alpha(1-\alpha)} {}^{\mathrm{qcvx}}J_Q^{\alpha}(\theta:\theta'). \tag{36}$$

As it will be shown below, we get only a pseudo-divergence in the limit case.

Theorem 1 (Formula for the quasiconvex Bregman pseudo-divergence). *For a strictly quasiconvex and differentiable generator Q, the quasiconvex Bregman pseudo-divergence is*

$$ {}^{\mathrm{qcvx}}B_Q(\theta:\theta') = \begin{cases} -(\theta-\theta')^\top \nabla Q(\theta') & \textit{if } Q(\theta)\leq Q(\theta') \\ +\infty & \textit{otherwise (i.e., } Q(\theta)>Q(\theta')\textit{).}\end{cases} \tag{37}$$

Proof. By definition, we have

$$ {}^{\mathrm{qcvx}}B_Q(\theta:\theta') = \lim_{\alpha\rightarrow 1^-} \frac{1}{\alpha(1-\alpha)} \left(\max\{Q(\theta),Q(\theta')\} - Q((\theta\theta')_\alpha)\right).$$

Applying a first-order Taylor expansion to $Q\left((\theta\theta')_\alpha\right)$, we get

$$ Q\left((\theta\theta')_\alpha)\right) \simeq_{\alpha\rightarrow 1} Q(\theta') - (1-\alpha)(\theta-\theta')^\top \nabla Q(\theta'). \tag{38}$$

Thus we have

$$ \begin{aligned} &{}^{\mathrm{qcvx}}B_Q(\theta:\theta') \\ &= \lim_{\alpha\rightarrow 1^-} \frac{1}{\alpha(1-\alpha)} \Big(\max\{Q(\theta),Q(\theta')\} - Q(\theta') - (1-\alpha)(\theta-\theta')^\top \nabla Q(\theta')\Big). \end{aligned} \tag{39}$$

Consider the following two cases:

- Case $\max\{Q(\theta), Q(\theta')\} = Q(\theta')$: That is, $Q(\theta') \geq Q(\theta)$. Then it follows that

$$
\begin{aligned}
{}^{\mathrm{qcvx}}B_Q(\theta : \theta') &= \lim_{\alpha \to 1^-} \frac{1}{\alpha(1-\alpha)} \left(-(1-\alpha)(\theta - \theta')^\top \nabla Q(\theta')\right), \qquad (40)\\
&= -(\theta - \theta')^\top \nabla Q(\theta'). \qquad (41)
\end{aligned}
$$

- Case $\max\{Q(\theta), Q(\theta')\} = Q(\theta)$: That is, $Q(\theta) \geq Q(\theta')$. Then we have

$$
{}^{\mathrm{qcvx}}B_Q(\theta : \theta') = \lim_{\alpha \to 1^-} \frac{1}{\alpha(1-\alpha)} \left(Q(\theta) - Q(\theta') - (1-\alpha)(\theta - \theta')^\top \nabla Q(\theta')\right).
$$

We have $\lim_{\alpha \to 1^-} Q(\theta) - Q(\theta') - (1-\alpha)(\theta - \theta')^\top \nabla Q(\theta') = Q(\theta) - Q(\theta') = \Delta_Q(\theta : \theta')$ that is finite and different from 0 when $\theta \neq \theta'$, and therefore $\lim_{\alpha \to 1^-} \frac{1}{\alpha(1-\alpha)} \Delta_Q(\theta : \theta') = +\infty$.

Let us now prove the axiom of non-negativity and disprove the law of the indiscernibles at inflection points for the quasiconvex Bregman pseudo-divergences.

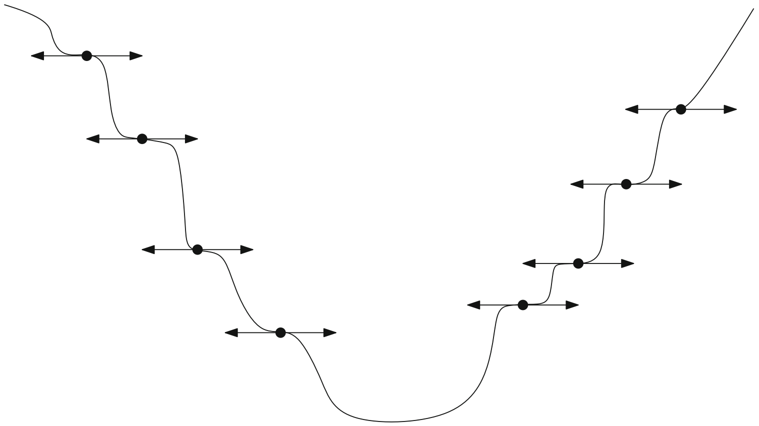

Fig. 2. An example of a strictly quasiconvex function Q with (countably) many inflection points (at locations θ_i's) for which the derivative vanishes $Q'(\theta_i) = 0$ and the second derivative Q'' changes sign at the θ_i's.

- Law of the indiscernibles: Clearly, ${}^{\mathrm{qcvx}}B_Q(\theta : \theta) = 0$ for all $\theta \in \Theta$. So consider $\theta \neq \theta'$, and ${}^{\mathrm{qcvx}}B_Q(\theta : \theta') = -\nabla Q(\theta')^\top (\theta - \theta') = 0$ for $Q(\theta') \geq Q(\theta)$. It is enough to consider the 1D case, by considering the divergence restricted to the line passing through θ and θ' intersected by the domain Θ. We may have countably many inflection points θ' for which $Q'(\theta') = 0$. At those inflection points, we may find $\theta \neq \theta'$ such that ${}^{\mathrm{qcvx}}B_Q(\theta : \theta') = 0$. Thus the quasiconvex Bregman divergence *does not* satisfy the law of the indiscernibles. Figure 2

displays an example of such a quasiconvex function with a few inflection points.
For example, consider the strictly quasiconvex generator $Q(x) = x^3$, with $\theta < 0$ and $\theta' = 0$. We have:

$$ {}^{\mathrm{qcvx}}J_Q^\alpha(\theta : \theta') = \max\{Q(\theta), Q(\theta')\} - Q((1-\alpha)\theta + \alpha\theta') = -(1-\alpha)^3\theta^3 > 0. \tag{42}$$

Defining the corresponding quasiconvex Bregman divergence by taking the limit of scaled quasiconvex Jensen divergence yields

$$ {}^{\mathrm{qcvx}}B_Q \lim_{\alpha\to 1} \frac{1}{\alpha(1-\alpha)} {}^{\mathrm{qcvx}}J_Q^\alpha(\theta : \theta') = \lim_{\alpha\to 1^-} -\frac{(1-\alpha)^2}{\alpha}\theta^3 = 0. \tag{43}$$

Thus the quasiconvex Bregman divergence is only a pseudo-divergence at countably many inflection points. Section 3.2 will overcome this problem by introducing the δ-averaged quasiconvex Bregman divergence.

- Non-negativity follows from a classic theorem of quasiconvex analysis which reports a first-order condition for a function to be quasiconvex[2]: A C^1 function $Q : \Theta \subset \mathbb{R}^D \to \mathbb{R}$ is quasiconvex iff. the following property holds (see Theorem 21.14 of [39] and §3.4.3 of [9]):

$$Q(\theta') \geq Q(\theta) \Rightarrow \nabla Q(\theta')(\theta - \theta') \leq 0. \tag{44}$$

That is equivalent to $\nabla Q(\theta')^\top(\theta - \theta') \leq 0$ or ${}^{\mathrm{qln}}\mathrm{B}_Q(\theta : \theta') = -\nabla Q(\theta')^\top(\theta - \theta') \geq 0$.
Notice that when $Q = F$ is strictly convex and differentiable, then the property also follows from the non-negativity of the corresponding Bregman divergence $B_F(\theta : \theta') \geq 0$ and $F(\theta') \geq F(\theta)$:

$$F(\theta) - F(\theta') - (\theta - \theta')^\top \nabla F(\theta') \geq 0, \tag{45}$$

$$\underbrace{-(\theta - \theta')^\top \nabla F(\theta')}_{{}^{\mathrm{qcvx}}B_F(\theta:\theta')} \geq F(\theta') - F(\theta) \geq 0. \tag{46}$$

□

Notice that $-(\theta - \theta')^\top \nabla Q(\theta') = (\theta' - \theta)^\top \nabla Q(\theta') \geq 0$ when $Q(\theta) \leq Q(\theta')$. Figure 3 illustrates the quasiconvex Bregman divergence for a strictly quasiconvex generator which is strictly concave and has no inflection point.

An interesting property is that if ${}^{\mathrm{qcvx}}B_Q(\theta : \theta') < \infty$ for $\theta \neq \theta'$ then necessarily ${}^{\mathrm{qcvx}}B_Q(\theta' : \theta) = \infty$, and vice-versa (when both parameters are not at inflection points). The forward ${}^{\mathrm{qcvx}}B_Q$ and reverse ${}^{\mathrm{qcvx}}B_Q^r$ quasiconvex Bregman

[2] By analogy to a classic second-order condition for a strictly convex and differentiable function F to be convex: To have its Hessian ∇^2 positive-definite (Alexandrov's theorem). Similarly, the first-order condition for convexity of a function states that a differentiable function F with convex domain is convex iff. $F(\theta) \geq F(\theta') + (\theta - \theta')^\top \nabla F(\theta')$ from which we recover the Bregman divergence: $B_F(\theta : \theta') = F(\theta) - F(\theta') - (\theta - \theta')^\top \nabla F(\theta') \geq 0$.

pseudo-divergences are both finite only when $Q(\theta) = Q(\theta')$ and then we have ${}^{\mathrm{qcvx}}B_Q(\theta : \theta) = 0$ or when one parameter is an inflection point.

Moreover, we have the following decomposition for a quasiconvex function $Q \in \mathcal{Q}$:

$$\mathrm{eB}_Q(\theta : \theta') = Q(\theta) - Q(\theta') + {}^{\mathrm{qcvx}}B_Q(\theta : \theta'), \tag{47}$$

when $Q(\theta) \leq Q(\theta')$, where eB_Q stands for the *extended Bregman divergence*, i.e., the Bregman divergence extended to a quasiconvex generator.

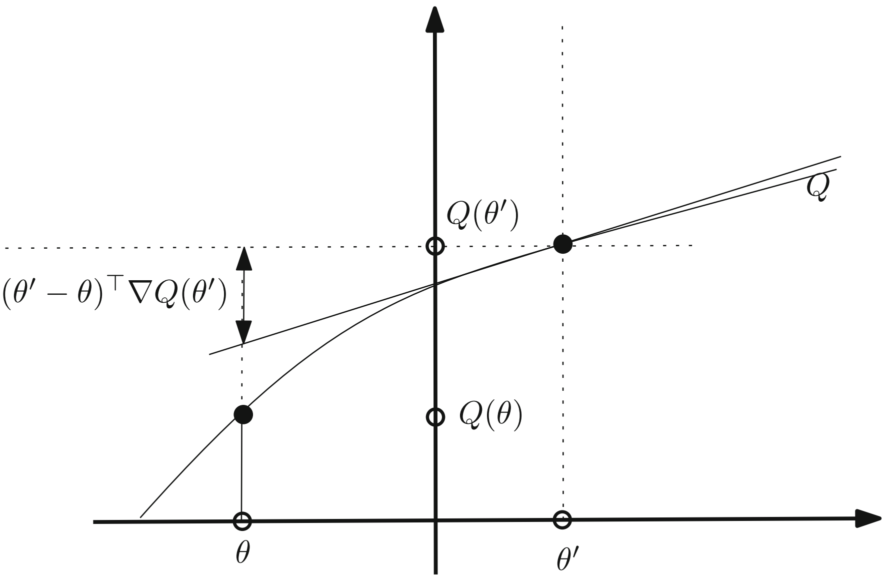

Fig. 3. Illustration of the quasiconvex Bregman divergence for a strictly quasilinear function Q chosen to be concave (e.g. logarithmic type).

Remark 2 (Separability/non-separability of generators and divergences). When the D-dimensional generator Q is *separable*, i.e., $Q(\theta) = \sum_{i=1}^D Q_i(\theta_i)$ where $\theta = (\theta_1, \ldots, \theta_D)$ and the Q_i's are differentiable and quasiconvex univariate functions, the quasiconvex Bregman divergence rewrites as

$$ {}^{\mathrm{qcvx}}B_Q(\theta : \theta') = \begin{cases} -\sum_{i=1}^D (\theta_i - \theta'_i) Q'_i(\theta'_i) & \text{if } Q(\theta) \leq Q(\theta') \\ +\infty & \text{otherwise } (Q(\theta) > Q(\theta')). \end{cases} \tag{48}$$

Notice that the condition for the quasiconvex Bregman divergence to be infinite is $Q(\theta) > Q(\theta')$, and not that there exists one index $i \in \{1, \ldots, D\}$ such that $Q_i(\theta_i) > Q(\theta'_i)$. Thus, we have ${}^{\mathrm{qcvx}}B_Q(\theta : \theta') \neq \sum_{i=1}^D {}^{\mathrm{qln}}\mathrm{B}_{Q_i}(\theta_i : \theta'_i)$. This is to contrast with Bregman divergences for which the separability of the generator $F(\theta) = \sum_{i=1}^D F_i(\theta_i)$ yields the separability of the divergence: $B_F(\theta : \theta') = \sum_{i=1}^D B_{F_i}(\theta_i : \theta'_i)$.

3.2 The δ-averaged Quasiconvex Bregman Divergence

We shall overcome the problem of indiscernability for quasiconvex Bregman pseudo-divergences:

$$^{\mathrm{qcvx}}B_Q(\theta:\theta') = (\theta'-\theta)Q(\theta') \quad \text{for} \quad Q(\theta') \geq Q(\theta). \tag{49}$$

Since the number of inflection points is at most countable for a strictly quasiconvex generator Q, we can integrate over a neighborhood of the parameters to obtain a strictly positive divergence when $\theta' \neq \theta$.

Given a prescribed parameter $\delta > 0$ (chosen to be small), we introduce the *δ-averaged quasiconvex Bregman divergence* $^{\mathrm{qcvx}}B_Q^\delta$ via the following definition:

$$^{\mathrm{qcvx}}B_Q^\delta(\theta,\theta') := \frac{1}{\delta}\int_0^\delta {}^{\mathrm{qcvx}}B_Q(\theta+u:\theta'+u)\mathrm{d}u. \tag{50}$$

This divergence is infinite if there exists $u \in]0,\delta[$ such that $Q(\theta+u) > Q(\theta'+u)$. This δ-averaged quasiconvex Bregman divergence now satisfies the law of the indiscernables.

When Q is differentiable, we obtain:

$$^{\mathrm{qcvx}}B_Q^\delta(\theta,\theta') := \frac{1}{\delta}\int_0^\delta (\theta'-\theta)Q'(\theta'+u)du = (\theta'-\theta)\left(\frac{Q(\theta'+\delta)-Q(\theta')}{\delta}\right), \tag{51}$$

when $Q(\theta'+u) \geq Q(\theta+u) \quad \forall 0 \leq u < \delta$.

Since Q is strictly quasiconvex, we can prove that the condition

$$Q(\theta'+u) \geq Q(\theta+u) \quad \forall 0 \leq u < \delta \tag{52}$$

is in fact equivalent to

$$Q(\theta') \geq Q(\theta) \quad \text{and} \quad Q(\theta'+\delta) \geq Q(\theta+\delta). \tag{53}$$

In 1D, a quasiconvex function is a decreasing then increasing function (i.e., a unimodal function). It is trivial that $Q(\theta') \geq Q(\theta)$ and $Q(\theta'+\delta) \geq Q(\theta+\delta)$ implies that $Q(\theta'+u) \geq Q(\theta+u) \quad \forall 0 \leq u < \delta$ for monotonic functions. The only remaining case is when θ' lies on the decreasing part and θ on the increasing part. If $\theta'+\delta$ also lies on the decreasing part, then since Q is decreasing between θ' and $\theta'+\delta$ we have

$$Q(\theta') \geq Q(\theta'+u) \geq Q(\theta'+\delta) \quad \forall 0 \leq u < \delta, \tag{54}$$

and, similarly,

$$Q(\theta) \leq Q(\theta+u) \leq Q(\theta+\delta) \quad \forall 0 \leq u < \delta, \tag{55}$$

so that

$$Q(\theta'+u) \geq Q(\theta+u) \quad \forall 0 \leq u < \delta. \tag{56}$$

Finally, if $\theta'+\delta$ lies on the increasing part of Q, since $\theta > \theta'$, we have $\theta'+\delta < \theta+\delta$ so $Q(\theta'+\delta) < Q(\theta+\delta)$ which is a contradiction.

Thus, the condition (52) for the finiteness of the integral (51) can only be checked at the endpoints (Eq. (53)).

Using the same reasoning as before, we can double check that the rhs. of (51) is indeed positive if conditions (Eq. (53)) are verified.

Also, the rhs. of (51) can also serve as the definition of the ${}^{\mathrm{qcvx}}B_Q^{\delta}$ divergences, even when the strictly quasiconvex function Q is not differentiable:

Definition 4 (δ-averaged quasiconvex Bregman divergence). For a prescribed $\delta > 0$ and a strictly quasiconvex generator Q not necessarily differentiable, the δ-averaged quasiconvex Bregman divergence is defined by

$$
\begin{aligned}
&{}^{\mathrm{qcvx}}B_Q^{\delta}(\theta, \theta') \\
&:= \begin{cases} \frac{1}{\delta}(\theta' - \theta)\left(Q(\theta' + \delta) - Q(\theta')\right) & \text{if } Q(\theta' + u) \geq Q(\theta + u) \quad \forall 0 \leq u < \delta \\ +\infty & \text{otherwise} \end{cases}
\end{aligned} \tag{57}
$$

Let us report some examples of δ-averaged quasiconvex Bregman divergences:

- $Q(x) = x$.
$$ {}^{\mathrm{qcvx}}B_Q(\theta : \theta') = (\theta' - \theta)\frac{\theta' + \delta - \theta'}{\delta} = \theta' - \theta, $$
when $\theta' \geq \theta$, or $+\infty$ otherwise.
- $Q(x) = x^2$.
$$ {}^{\mathrm{qcvx}}B_Q(\theta : \theta') = (\theta' - \theta)(2\theta' + \delta), $$
when $|\theta'| \geq |\theta|$ and $\theta'^2 - \theta^2 + 2\delta(\theta' - \theta) \geq 0$, or $+\infty$ otherwise.
- $Q(x) = x^3$.
$$ {}^{\mathrm{qcvx}}B_Q(\theta : \theta') = (\theta' - \theta)\frac{((\theta' + \delta)^3 + \theta'^3)}{\delta}, $$
when $\theta' \geq \theta$, or $+\infty$ otherwise. When $\theta' = 0$, we now have
$$ {}^{\mathrm{qcvx}}B_Q(\theta : \theta') = -\delta^2\theta > 0 \quad \forall \theta < 0. $$

These examples show that the second condition of (53) is only useful for non-monotonic functions.

3.3 Multivariate Quasiconvex Generators Q

The construction of the preceding section also applies when Q is multivariate.

We suppose for now that the quasiconvex function Q is differentiable so that we have
$$ {}^{\mathrm{qcvx}}B_Q(\theta : \theta') = (\theta' - \theta)^{\top}\nabla Q(\theta'). \tag{58} $$

Let us fix $\delta > 0$, and define the δ-averaged quasiconvex Bregman divergence ${}^{\mathrm{qcvx}}B_Q^{\delta}$ when Q is multivariate as:

$$ {}^{\mathrm{qcvx}}B_Q^{\delta}(\theta : \theta') := \frac{1}{\delta}\int_0^{\delta} {}^{\mathrm{qcvx}}B_Q\left(\theta + u(\theta' - \theta) : \theta' + u(\theta' - \theta)\right) \mathrm{d}u, \tag{59} $$

for $\theta' \neq \theta$ and $Q(\theta') \geq Q(\theta)$.

Using Eq. 58, we obtain:

$$ {}^{\mathrm{qcvx}}B_Q^{\delta}(\theta:\theta') = \frac{1}{\delta}\left(Q\left(\theta'+\delta(\theta'-\theta)\right) - Q(\theta')\right). \tag{60} $$

As this expression does not involves the derivatives of Q, we can use Eq. 60 to define the δ-averaged quasiconvex Bregman divergence in the case where the quasiconvex function Q is not differentiable.

Let us report one example of quasiconvex Bregman divergence for a bivariate generator: Let $Q(\theta) = \max\left\{\theta_1^3, \theta_2^3\right\}$.

Q is quasiconvex as it is the maximum of two quasiconvex functions [1]. Let $\theta' = (0,0)$. We have $\nabla Q(\theta') = (0,0)$ so that ${}^{\mathrm{qcvx}}B_Q(\theta':\theta) = 0$ for all θ such that $Q(\theta) < Q(\theta') = 0$.

Now, considering the δ-averaged quasiconvex Bregman divergence, we obtain

$$ {}^{\mathrm{qcvx}}B_Q(\theta':\theta) = \frac{1}{\delta}\max\left\{(-\delta\theta_1)^3, (-\delta\theta_2)^3\right\} = -\delta^2\min\left\{\theta_1^3,\theta_2^3\right\} > 0, \tag{61} $$

since $\max\left\{\theta_1^3,\theta_2^3\right\} = Q(\theta) < 0$ implies that θ_1 and θ_2 are strictly negative.

3.4 Quasiconvex Bregman Divergences as Limit Cases of Power Mean Bregman Divergences

For sake of simplicity, consider scalar divergences below. In [33], the (M,N)-*Bregman divergence* is defined as the limit case:

$$ \begin{aligned} B_F^{M,N}(p:q) &= \lim_{\alpha\rightarrow 1^-}\frac{1}{\alpha(1-\alpha)}J_{F,\alpha}^{M,N}(p:q), && (62)\\ &= \lim_{\alpha\rightarrow 1^-}\frac{1}{\alpha(1-\alpha)}\left(N_\alpha(F(p),F(q))) - F(M_\alpha(p,q))\right). && (63) \end{aligned} $$

In particular, the univariate *power mean Bregman divergences* are obtained by taking the power means, yielding the following formula:

$$ B_F^{\delta_1,\delta_2}(p:q) = \frac{F^{\delta_2}(p)-F^{\delta_2}(q)}{\delta_2 F^{\delta_2-1}(q)} - \frac{p^{\delta_1}-q^{\delta_1}}{\delta_1 q^{\delta_1-1}}F'(q). \tag{64} $$

Let $\delta_2 = r$ and $\delta_1 = 1$. Then we get the subfamily of r-power Bregman divergences:

$$ \begin{aligned} B_F^r(\theta:\theta') &= \frac{F^r(\theta)-F^r(\theta')}{rF^{r-1}(\theta')} - (\theta-\theta')F'(\theta'), && (65)\\ &= = \frac{F^r(\theta)}{rF^{r-1}(\theta')} - \frac{F(\theta')}{r} - (\theta-\theta')F'(\theta'). && (66) \end{aligned} $$

In Eq. 66, when $F(\theta) > F(\theta')$ then we have $\lim_{r\rightarrow\infty} B_F^r(\theta:\theta') = \infty$ since $\left(\frac{F^r(\theta)}{F^{r-1}(\theta')}\right)$ diverges. Otherwise ${}^{\mathrm{qln}}\mathrm{B}_F(\theta:\theta') = \lim_{r\rightarrow\infty} B_F^r(\theta:\theta') = -(\theta-\theta')F'(\theta')$ since $\lim_{r\rightarrow}\frac{F(\theta')}{r} = 0$ (because $|F(\theta')| < \infty$).

When $r \to \infty$, the power mean operator P_r tends to the maximum operator: $\lim_{r\to\infty} P_r(a,b) = \max\{a,b\}$, and the (A, P_δ)-Bregman divergence tends to the quasiconvex Bregman pseudo-divergence.

3.5 Some Illustrating Examples of Quasiconvex Bregman Divergences

We concisely report two univariate quasiconvex scalar Bregman divergences:

- For $Q(\theta) = \theta$ with $\theta \in \mathbb{R}$, we have

$$^{\mathrm{qcvx}}J_Q^\alpha(\theta : \theta') = \max\{\theta, \theta'\} - (1-\alpha)\theta + \alpha\theta'.$$

  We consider the two cases for calculating the limit $^{\mathrm{qcvx}}B_Q(\theta : \theta') = \lim_{\alpha\to 1^-} \frac{1}{\alpha(1-\alpha)} {}^{\mathrm{qcvx}}J_Q^\alpha(\theta : \theta')$:
  - When $\theta' \geq \theta$:

$$\lim_{\alpha\to 1^-} \frac{1}{\alpha(1-\alpha)} {}^{\mathrm{qcvx}}J_Q^\alpha(\theta : \theta') = \lim_{\alpha\to 1^-} \frac{1}{\alpha(1-\alpha)}(-(1-\alpha)\theta + (1-\alpha)\theta') = \theta' - \theta \geq 0.$$

  - When $\theta > \theta'$:

$$\begin{aligned}\lim_{\alpha\to 1^-} \frac{1}{\alpha(1-\alpha)} {}^{\mathrm{qcvx}}J_Q^\alpha(\theta : \theta') &= \lim_{\alpha\to 1^-} \frac{1}{\alpha(1-\alpha)}(\theta - (1-\alpha)\theta - \alpha\theta'),\\ &= \lim_{\alpha\to 1^-} \frac{1}{1-\alpha}(\theta - \theta') = +\infty.\end{aligned}$$

  Thus we have the following quasiconvex Bregman divergence: $^{\mathrm{qcvx}}B_Q(\theta : \theta') = \theta' - \theta$ for $\theta' \geq \theta$ and $+\infty$ when $\theta' < \theta$.
- When $Q(\theta) = \log\theta$, we have $Q'(\theta) = \frac{1}{\theta}$ and $^{\mathrm{qcvx}}B_Q(\theta : \theta') = 1 - \frac{\theta}{\theta'}$ for $\log\theta' \geq \log\theta$ (i.e. $\theta' \geq \theta$) and $+\infty$ when $\theta' < \theta$.
- For $Q(\theta) = \sqrt{\theta}$ and $\theta \in \Theta = (0, \infty)$, we have $Q'(\theta) = \frac{1}{2\sqrt{\theta}}$ and $^{\mathrm{qcvx}}B_Q(\theta : \theta') = \frac{1}{2}\left(\sqrt{\theta'} - \frac{\theta}{\sqrt{\theta'}}\right)$ for $\sqrt{\theta'} \geq \sqrt{\theta}$ (i.e., $\theta' \geq \theta$), and $+\infty$ when $\theta' < \theta$.

4 Statistical Divergences, Parametric Families of Distributions and Equivalent Parameter Divergences

Consider a probability space $(\mathcal{X}, \mathcal{F}, \mu)$ with $\mathcal{X}$, $\mathcal{F}$, and μ denoting the sample space, the σ-algebra and the positive measure, respectively. The most celebrated *statistical divergence* between two densities $p_\theta \ll \mu$ and $p_{\theta'} \ll \mu$ absolutely continuous with respect to a measure μ is the Kullback-Leibler (KL) divergence (also called *relative entropy* [13]), defined by:

$$\mathrm{KL}[p : q] = \begin{cases} \int_{x\in\mathcal{X}} p(x) \log\frac{p(x)}{q(x)} \mathrm{d}\mu(x), & \mathrm{supp}(p) \subset \mathrm{supp}(q), \\ +\infty, & \mathrm{supp}(p) \not\subset \mathrm{supp}(q). \end{cases}, \qquad (67)$$

where $\mathrm{supp}(p) = \{x \in \mathbb{R} \ : \ p(x) > 0\}$ denotes the *support* of a distribution $p(x)$, and $\log\frac{0}{0} = 0$ by convention. Thus the KL divergence is said unbounded in general.[3]

In general, a statistical divergence between densities belonging to the same parametric family $\mathcal{P} = \{p_\theta\}_\theta$ of mutually absolutely continuous densities is equivalent to a corresponding *parameter divergence* B:

$$B(\theta : \theta') := D[p_\theta : p_{\theta'}]. \tag{68}$$

For example, when $\mathcal{P} = \{p_\theta(x) = \exp(x^\top\theta - F(\theta))\mathrm{d}\mu(x)\}_\theta$ is an exponential family [5,6,24] on a probability space $(\mathcal{X}, \mathcal{F}, \mu)$, then the Kullback-Leibler divergence between two densities of the exponential family (e.g., two Gaussians distributions belonging to the Gaussian exponential family) amount to a *reverse* Bregman divergence [5] for the Bregman generator set to the cumulant function $F(\theta) = \log\int\exp(x^\top\theta)\mathrm{d}\mu(x)$:

$$\mathrm{KL}[p_\theta : p_{\theta'}] = B(\theta : \theta') = {B_F}^r(\theta : \theta') = B_F(\theta' : \theta). \tag{69}$$

Banerjee et al. [5] proved a one-to-one correspondence (*bijection*) between regular natural exponential families and so-called *regular* Bregman divergences. Note that since the Ali-Silvey-Csiszár's f-divergence [2,3] (including the KL divergence) is invariant to one-to-one smooth mapping $m(x)$ of the sample space x, the same Bregman divergence equivalent to the KL divergence can be obtained for different exponential families where $y = m(x)$. For example, the KL divergence between two normal distributions or two "equivalent" log-normal distributions is the same (using the mapping $y = \log x$). This can be also noticed by the matching of their cumulant function: $F_{\mathrm{normal}}(\theta) = F_{\mathrm{lognormal}}(\theta)$.

Quasiconvex Bregman divergences have the interesting property to be finite for one orientation and infinite for the other orientation. Thus to find an example of parametric family of distributions which the KL divergence amount to a quasiconvex Bregman divergence, we shall consider parametric distributions with *nested supports* (or *nested densities*), so that one orientation of the KL divergence will be finite while the other is will be equal to infinity.

For example, consider the family of univariate uniform densities ($D = 1$):

$$p_\theta(x) = 1_{0<x<e^\theta}\ e^{-\theta}, \tag{70}$$

where 1_A denotes the indicator function of A. We have $\mathrm{supp}(p_{\theta'}) \subset \mathrm{supp}(p_\theta)$ for $0 < \theta' \leq \theta$. Then we have

$$\mathrm{KL}[p_\theta : p_{\theta'}] = \begin{cases} \theta' - \theta = {}^{\mathrm{qln}}\mathrm{B}_Q(\theta : \theta') & 0 < \theta \leq \theta', \\ +\infty & \theta' > \theta. \end{cases}, \tag{71}$$

for $Q(\omega) = \omega$.

Notice that the family $\mathcal{P} = \{p_\theta\}$ is not an exponential family since the family has not a fixed support. A truncated exponential family with fixed truncation

[3] The Jensen-Shannon divergence [28] is a particular symmetrization of the KL divergence which is always bounded, and may accept densities with different supports.

parameters yields an exponential family which may neither be regular nor steep (e.g., the singly truncated normal distributions [15]).

Now, consider the parametric family $\{q_\theta\}_\theta$ of nested densities:

$$q_\theta(x) = 1_{0<x<e^\theta} \alpha \frac{x^{\alpha-1}}{e^{\theta\alpha}}, \tag{72}$$

for a prescribed $\alpha > 1$. After a short calculation (or using a computer algebra system as reported in Sect. 6), we find that

$$\mathrm{KL}[q_\theta : q_{\theta'}] = \begin{cases} \alpha(\theta' - \theta) = {}^{\mathrm{qln}}\mathrm{B}_Q(\theta : \theta') & \theta' \geq \theta > 0, \\ +\infty & \theta' < \theta. \end{cases}, \tag{73}$$

for $Q(\omega) = \omega$. Thus we have built several parametric families of nested densities that up to a scaling factor yields the same quasiconvex Bregman divergence.

For parametric densities belonging to the same exponential family, it is known that the Bhattacharrya distance amount to a Jensen divergence [29]. For an exponential family $p_\theta(x) = \exp(\theta^\top x - F(\theta))\mathrm{d}\mu(x)$ with cumulant function F, the cross-entropy between two densities [30] is

$$h(p_\theta : p_{\theta'}) = \int -p_\theta(x) \log p_{\theta'}(x) \mathrm{d}\mu(x) = F(\theta') - (\theta')^\top \nabla F(\theta), \tag{74}$$

and the entropy is

$$h(p_\theta) = h(p_\theta : p_\theta) = F(\theta) - \theta^\top \nabla F(\theta). \tag{75}$$

Since $\mathrm{KL}(p_\theta : p_{\theta'}) = B_F(\theta' : \theta) = F(\theta') - F(\theta) - (\theta' - \theta)^\top \nabla F(\theta)$, when $F(\theta') \leq F(\theta)$, we have $-(\theta' - \theta)^\top \nabla F(\theta) = {}^{\mathrm{qln}}\mathrm{B}_F(\theta' : \theta)$, and it follows that

$${}^{\mathrm{qln}}\mathrm{B}_F(\theta' : \theta) = \mathrm{KL}(p_\theta : p_{\theta'}) + F(\theta) - F(\theta'), \quad F(\theta') \leq F(\theta). \tag{76}$$

The Wasserstein distance between two nested univariate distributions has been studied in [23] with applications to Bayesian statistics to study the influence of the prior distribution in the posterior distribution in the finite sample size setting.

5 Conclusion and Perspectives

In this chapter, we have introduced two novel families of distortions between vector parameters: The quasiconvex Jensen divergences and the quasiconvex Bregman divergences. We showed that the quasiconvex Jensen divergences measuring the difference gaps of the quasiconvex inequalities can be interpreted as a ℓ_1-regularized ordinary Jensen divergence. We noticed that any quasiconcave Jensen divergence amounts to an equivalent quasiconvex Jensen divergence for the negative generator. We then derived the quasiconvex Bregman pseudo-divergences as limit cases of scaled and skewed quasiconvex Jensen divergences for strictly quasiconvex generators. The quasiconvex Bregman pseudo-divergences is a pseudo-divergence only at countably many inflection points of the generators. We thus propose to define the δ-averaged quasiconvex Bregman divergences by integrating the pseudo-divergence over a small neighborhood. This yields a formula (Eq. 57) that can be used as the definition of the quasiconvex Bregman divergence even for non-differentiable strictly quasiconvex generators. We also showed how to derive again the result of the quasiconvex Bregman pseudo-divergences using comparative convexity [33] using the limit case of power means. A key property of the quasiconvex Bregman divergences between distinct elements is that they are necessarily finite on one orientation and infinite for the reverse orientation. Finally, we showed how some of these quasiconvex Bregman divergences can be obtained from the Kullback-Leibler divergence between probability densities belonging to the same parametric family of distributions with nested density support. We can retrieve the Bregman pseudo-divergences and quasiconvex Bregman pseudo-divergences from first-order convexity and quasiconvexity conditions, as illustrated in Table 1. Additional conditions on the generators ensure that the pseudo-divergences are proper divergences and satisfy the law of the indiscernibles (i.e., strict convexity and differentiability for Bregman divergences and strict quasiconvexity without inflection points for the quasiconvex Bregman divergences).

We plan to consider applications of these novel divergences in clustering: We note that the generic k-means++ probabilistic seeding analysis reported in [32] does not apply because of the forward/reverse infinite property of these quasiconvex Bregman divergences. We may consider discrete k-means, k-center (with the minimum enclosing ball obtained from quasiconvex programming [1, 16, 19, 21] when $k = 1$), and quasiconvex Bregman hierarchical clustering [40].

Table 1. Bregman divergence and Bregman quasidivergence with their relationship to first-order convexity and quasiconvexity.

	First-order condition	Pseudo-divergence/condition for divergence
Convexity	$F(\theta) \geq F(\theta') + (\theta - \theta')^\top \nabla F(\theta')$	$B_F(\theta : \theta') = F(\theta) - F(\theta') + (\theta - \theta')^\top \nabla F(\theta')$
of F	Divergence when F strictly convex and differentiable	
Quasiconvexity	$Q(\theta) \leq Q(\theta') \Rightarrow (\theta - \theta')^\top \nabla Q(\theta') \leq 0$	$\begin{cases} -(\theta - \theta')^\top \nabla Q(\theta') & \text{if } Q(\theta) \leq Q(\theta') \\ +\infty & \text{otherwise.} \end{cases}$
of Q	Divergence when Q strictly quasiconvex with no inflection point	

6 Calculations Using a Computer Algebra System

Using the computer algebra system (CAS) MAXIMA[4], we report a snippet code for the calculation of the Kullback-Leibler divergence for nested probability densities.

[4] Freely downloadable at http://maxima.sourceforge.net/.

```
assume(alpha>1);
assume(theta>0);
p(x,theta):=alpha*(x**(alpha-1))/(exp(theta*alpha));
integrate(p(x,theta),x,0,exp(theta));
assume(thetap>theta);
/* Kullback-Leibler divergence */
integrate(p(x,theta)*log(p(x,theta)/p(x,thetap)),x,0,exp(theta));
```

References

1. Agrawal, A., Boyd, S.: Disciplined quasiconvex programming. arXiv preprint arXiv:1905.00562 (2019)
2. Ali, S.M., Silvey, S.D.: A general class of coefficients of divergence of one distribution from another. J. Roy. Stat. Soc.: Ser. B (Methodol.) **28**(1), 131–142 (1966)
3. Amari, S.: Information Geometry and Its Applications, vol. 194. Springer, Tokyo (2016)
4. Azoury, K.S., Warmuth, M.K.: Relative loss bounds for on-line density estimation with the exponential family of distributions. Mach. Learn. **43**(3), 211–246 (2001)
5. Banerjee, A., Merugu, S., Dhillon, I.S., Ghosh, J.: Clustering with Bregman divergences. J. Mach. Learn. Res. **6**(Oct), 1705–1749 (2005)
6. Barndorff-Nielsen, O.: Information and Exponential Families in Statistical Theory. Wiley, Hoboken (2014)
7. Baryshnikov, Y., Ghrist, R.: Unimodal category and topological statistics. In: Proceedings of Nonlinear Theory and Its Applications (NOLTA) (2011)
8. Bereanu, B.: Quasi-convexity, strictly quasi-convexity and pseudo-convexity of composite objective functions. Revue française d'automatique informatique recherche opérationnelle. Mathématique **6**(R1), 15–26 (1972)
9. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
10. Brègman, L.M.: Finding the common point of convex sets by the method of successive projection. Dokl. Akad. Nauk SSSR **162**(3), 487–490 (1965). (in Russian)
11. 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 Comput. Math. Math. Phys. **7**(3), 200–217 (1967)
12. Burbea, J., Rao, C.R.: Entropy differential metric, distance and divergence measures in probability spaces: a unified approach. J. Multivar. Anal. **12**(4), 575–596 (1982)
13. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, Hoboken (2012)
14. De Finetti, B.: Sul concetto di media. Istituto italiano degli attuari **3**, 369–396 (1931)
15. Del Castillo, J.: The singly truncated normal distribution: a non-steep exponential family. Ann. Inst. Stat. Math. **46**(1), 57–66 (1994)
16. Eppstein, D.: Quasiconvex programming. Comb. Comput. Geom. **52**(287–331), 3 (2005)
17. Frigyik, B.A., Srivastava, S., Gupta, M.R.: Functional Bregman divergence. In: 2008 IEEE International Symposium on Information Theory, pp. 1681–1685. IEEE (2008)
18. Greenberg, H.J., Pierskalla, W.P.: A review of quasi-convex functions. Oper. Res. **19**(7), 1553–1570 (1971)

19. Hazan, E., Levy, K., Shalev-Shwartz, S.: Beyond convexity: stochastic quasi-convex optimization. In: Advances in Neural Information Processing Systems, pp. 1594–1602 (2015)
20. Iyer, R., Bilmes, J.A.: Submodular-Bregman and the Lovász-Bregman divergences with applications. In: Advances in Neural Information Processing Systems, pp. 2933–2941 (2012)
21. Ke, Q., Kanade, T.: Quasiconvex optimization for robust geometric reconstruction. IEEE Trans. Pattern Anal. Mach. Intell. **29**(10), 1834–1847 (2007)
22. Kolmogorov, A.N.: Sur la notion de moyenne. Acad. Naz. Lincei Mem. Cl. Sci. His. Mat. Natur. Sez. **12**, 388–391 (1930)
23. Ley, C., Reinert, G., Swan, Y., et al.: Distances between nested densities and a measure of the impact of the prior in Bayesian statistics. Ann. Appl. Probab. **27**(1), 216–241 (2017)
24. Miao, W., Hahn, M.: Existence of maximum likelihood estimates for multi-dimensional exponential families. Scand. J. Stat. **24**(3), 371–386 (1997)
25. Nagumo, M.: Über eine Klasse der Mittelwerte. Jpn. J. Math. Trans. Abs. **7**, 71–79 (1930)
26. Niculescu, C.P., Persson, L.E.: Convex Functions and Their Applications: A Contemporary Approach, 2nd edn. Springer, Cham (2018)
27. Nielsen, F.: Hierarchical clustering. In: Introduction to HPC with MPI for Data Science, pp. 195–211. Springer, Cham (2016)
28. Nielsen, F.: On the Jensen-Shannon symmetrization of distances relying on abstract means. Entropy **21**(5), 485 (2019)
29. Nielsen, F., Boltz, S.: The Burbea-Rao and Bhattacharyya centroids. IEEE Trans. Inf. Theory **57**(8), 5455–5466 (2011)
30. Nielsen, F., Nock, R.: Entropies and cross-entropies of exponential families. In: 2010 IEEE International Conference on Image Processing, pp. 3621–3624. IEEE (2010)
31. Nielsen, F., Nock, R.: Further heuristics for k-means: the merge-and-split heuristic and the (k, l)-means. arXiv:1406.6314 (2014)
32. Nielsen, F., Nock, R.: Total Jensen divergences: definition, properties and clustering. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2016–2020. IEEE (2015)
33. Nielsen, F., Nock, R.: Generalizing skew Jensen divergences and Bregman divergences with comparative convexity. IEEE Signal Process. Lett. **24**(8), 1123–1127 (2017)
34. Nielsen, F., Sun, K.: Guaranteed bounds on information-theoretic measures of univariate mixtures using piecewise log-sum-exp inequalities. Entropy **18**(12), 442 (2016)
35. Nielsen, F., Sun, K., Marchand-Maillet, S.: On Hölder projective divergences. Entropy **19**(3), 122 (2017)
36. Nock, R., Magdalou, B., Briys, E., Nielsen, F.: Mining matrix data with Bregman matrix divergences for portfolio selection. In: Matrix Information Geometry, pp. 373–402. Springer, Heidelberg (2013)
37. Penot, J.P.: Glimpses upon quasiconvex analysis. In: ESAIM Proceedings, vol. 20, pp. 170–194. EDP Sciences (2007)
38. Rao, C., Nayak, T.: Cross entropy, dissimilarity measures, and characterizations of quadratic entropy. IEEE Trans. Inf. Theory **31**(5), 589–593 (1985)
39. Simon, C.P., Blume, L., et al.: Mathematics for Economists, vol. 7. Norton, New York (1994)

40. Telgarsky, M., Dasgupta, S.: Agglomerative Bregman clustering. In: Proceedings of the 29th International Conference on International Conference on Machine Learning, pp. 1011–1018. Omnipress (2012)
41. Zhang, J.: Divergence function, duality, and convex analysis. Neural Comput. **16**(1), 159–195 (2004)