The Bregman Chord Divergence

Frank Nielsen[1(✉)] and Richard Nock[2,3,4]

[1] Sony Computer Science Laboratories, Inc., Tokyo, Japan
Frank.Nielsen@acm.org
[2] Data61, Sydney, Australia
Richard.Nock@data61.csiro.au
[3] The Australian National University, Canberra, Australia
[4] The University of Sydney, Sydney, Australia

Abstract. Distances are fundamental primitives whose choice significantly impacts the performances of algorithms in applications. However selecting the most appropriate distance for a given task is an endeavor. Instead of testing one by one the entries of an ever-expanding dictionary of *ad hoc* distances, one rather prefers to consider parametric classes of distances that are exhaustively characterized by axioms derived from first principles. Bregman divergences are such a class. However fine-tuning a Bregman divergence is delicate since it requires to smoothly adjust a functional generator. In this work, we propose an extension of Bregman divergences called the Bregman chord divergences. This new class of distances bypasses the gradient calculations, uses two scalar parameters that can be easily tailored in applications, and generalizes asymptotically Bregman divergences.

Keywords: Csiszár's f-divergence · Bregman divergence · Jensen divergence · Skewed divergence

1 Introduction

Distances are at the heart of many signal processing tasks [6,14], and the performance of algorithms solving those tasks heavily depends on the chosen distances. Historically, many *ad hoc* distances have been proposed and empirically benchmarked on different tasks in order to improve the state-of-the-art performances. However, getting the most appropriate distance for a given task is often an endeavour. Thus principled *classes* of distances[1] have been proposed

[1] Here, we use the word distance to mean a *dissimilarity* (or a distortion, a deviance, a discrepancy, etc.), not necessarily a metric distance [14]. A distance between arguments θ_1 and θ_2 satisfies $D(\theta_1, \theta_2) \geq 0$ with equality if and only if $\theta_1 = \theta_2$.

F. Nielsen and F. Barbaresco (Eds.): GSI 2019, LNCS 11712, pp. 299–308, 2019.
https://doi.org/10.1007/978-3-030-26980-7_31

and studied. Among those generic classes of distances, three main generic classes have emerged:

- The *Bregman divergences* [5,7,22] defined for a strictly convex and differentiable generator $F \in \mathcal{B} : \Theta \rightarrow \mathbb{R}$ (where $\mathcal{B}$ denotes the class of strictly convex and differentiable functions defined modulo affine terms):

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

  measure the dissimilarity between *parameters* $\theta_1, \theta_2 \in \Theta$, where $\Theta \subset \mathbb{R}^d$ is a d-dimensional convex set. Bregman divergences have also been generalized to other types of objects like matrices [26].
- The Csiszár f-divergences [1,11,12] defined for a convex generator $f \in \mathcal{C}$ satisfying $f(1) = 0$ and strictly convex at 1:

$$I_f[p_1 : p_2] := \int_{\mathcal{X}} p_1(x) f\left(\frac{p_2(x)}{p_1(x)}\right) \mathrm{d}\mu(x) \geq f(1) = 0, \tag{2}$$

  measure the dissimilarity between *probability densities* p_1 and p_2 that are absolutely continuous with respect to a base measure μ (defined on a support $\mathcal{X}$).
- The Burbea-Rao divergences [9] also called Jensen differences or Jensen divergences because they rely on the Jensen's inequality [16] for a strictly convex function $F \in \mathcal{J} : \Theta \rightarrow \mathbb{R}$:

$$J_F(\theta_1, \theta_2) := \frac{F(\theta_1) + F(\theta_2)}{2} - F\left(\frac{\theta_1 + \theta_2}{2}\right) \geq 0, \tag{3}$$

  where θ_1 and θ_2 belong to a parameter space Θ.

These three fundamental classes of distances are *not* mutually exclusive, and their pairwise intersections (e.g., $\mathcal{B} \cap \mathcal{C}$ or $\mathcal{J} \cap \mathcal{C}$) have been studied in [2,17,27]. The ':' notation between arguments of distances emphasizes the potential asymmetry of distances (oriented distances with $D(\theta_1 : \theta_2) \neq D(\theta_2 : \theta_1)$), and the brackets surrounding distance arguments indicate that it is a *statistical distance* between probability densities, and not a distance between parameters. Using these notations, we express the Kullback-Leibler distance [10] (KL) as

$$\mathrm{KL}[p_1 : p_2] := \int p_1(x) \log \frac{p_1(x)}{p_2(x)} \mathrm{d}\mu(x). \tag{4}$$

The KL distance/divergence between two members p_{θ_1} and p_{θ_2} of a parametric family $\mathcal{F}$ of distributions amount to a parameter divergence

$$\mathrm{KL}_{\mathcal{F}}(\theta_1 : \theta_2) := \mathrm{KL}[p_{\theta_1} : p_{\theta_2}]. \tag{5}$$

For example, the KL statistical distance between two probability densities belonging to the same exponential family or the same mixture family amounts to a (parameter) Bregman divergence [3,23]. When p_1 and p_2 are finite discrete distributions of the d-dimensional probability simplex Δ_d, we have $\mathrm{KL}_{\Delta_d}(p_1 : p_2) = \mathrm{KL}[p_1 : p_2]$. This explains why sometimes we can handle loosely distances between discrete distributions as both a parameter distance and a statistical distance. For example, the KL distance between two discrete distributions is a Bregman divergence $B_{F_{\mathrm{KL}}}$ for $F_{\mathrm{KL}}(x) = \sum_{i=1}^{d} x_i \log x_i$ (Shannon negentropy) for $x \in \Theta = \Delta_d$. Extending $\Theta = \Delta_d$ to positive measures $\Theta = \mathbb{R}_+^d$, this Bregman divergence $B_{F_{\mathrm{KL}}}$ yields the extended KL distance:

$$\mathrm{eKL}[p : q] = \sum_{i=1}^{d} p_i \log \frac{p_i}{q_i} + q_i - p_i. \tag{6}$$

Notice that the KL divergence of 4 between non-probability positive distributions may yield potential negativity of the measure (e.g., Example 2.1 of [28] and [8]). This case also happens when doing Monte Carlo stochastic integrations of the KL divergence integral.

Whenever using a functionally parameterized distance in applications, we need to choose the most appropriate functional generator, ideally from first principles [3, 4, 13]. For example, Non-negative Matrix Factorization (NMF) for audio source separation or music transcription from the signal power spectrogram can be done by selecting the Itakura-Saito divergence [15][2] that satisfies the requirement of being *scale invariant*:

$$B_{F_{\mathrm{IS}}}(\lambda\theta : \lambda\theta') = B_{F_{\mathrm{IS}}}(\theta : \theta') = \sum_i \left(\frac{\theta_i}{\theta'_i} - \log \frac{\theta_i}{\theta'_i} - 1 \right), \tag{7}$$

for any $\lambda > 0$. When no such first principles can be easily stated for a task [13], we are left by choosing manually or by cross-validation a generator. Notice that the convex combinations of Csiszár generators is a Csiszár generator (idem for Bregman divergences): $\sum_{i=1}^{d} \lambda_i I_{f_i} = I_{\sum_i i=1^d \lambda_i f_i}$ for λ belonging to the standard $(d-1)$-dimensional standard simplex Δ_d.

In this work, we propose a novel class of distances, termed *Bregman chord divergences*. A Bregman chord divergence is parameterized by a Bregman generator and two scalar parameters which make it easy to fine-tune in applications, and matches asymptotically the ordinary Bregman divergence.

The paper is organized as follows: In Sect. 2, we describe the skewed Jensen divergence, show how to bi-skew any distance by using two scalars, and report on the Jensen chord divergence [20]. In Sect. 3, we first introduce the univariate Bregman chord divergence, and then extend its definition to the multivariate case, in Sect. 4. Finally, we conclude in Sect. 5.

2 Geometric Design of Skewed Divergences

We can geometrically *design* divergences from convexity gap properties of the graph plot of the generator. For example, the Jensen divergence $J_F(\theta_1 : \theta_2)$ of

[2] A Bregman divergence for the Burg negentropy $F_{\mathrm{IS}}(x) = -\sum_i \log x_i$.

Eq. 3 is visualized as the ordinate (vertical) gap between the midpoint of the line segment $[(\theta_1, F(\theta_1)); (\theta_2, F(\theta_2))]$ and the point $(\frac{\theta_1+\theta_2}{2}, F(\frac{\theta_1+\theta_2}{2}))$. The non-negativity property of the Jensen divergence follows from the Jensen's midpoint convex inequality [16]. Instead of taking the midpoint $\bar{\theta} = \frac{\theta_1+\theta_2}{2}$, we can take *any* interior point $(\theta_1\theta_2)_\alpha := (1-\alpha)\theta_1 + \alpha\theta_2$, and get the skewed α-Jensen divergence (for any $\alpha \in (0,1)$):

$$J_F^\alpha(\theta_1 : \theta_2) := (F(\theta_1)F(\theta_2))_\alpha - F((\theta_1\theta_2)_\alpha) \geq 0. \tag{8}$$

A remarkable fact is that the scaled α-Jensen divergence $\frac{1}{\alpha}J_F^\alpha(\theta_1 : \theta_2)$ tends asymptotically to the reverse Bregman divergence $B_F(\theta_2 : \theta_1)$ when $\alpha \to 0$, see [21,30].

By measuring the ordinate gap between two non-crossing upper and lower chords anchored at the generator graph plot, we can extend the α-Jensen divergences to a tri-parametric family of Jensen chord divergences [20]:

$$J_F^{\alpha,\beta,\gamma}(\theta : \theta') := (F(\theta)F(\theta'))_\gamma - (F((\theta\theta')_\alpha)F((\theta\theta')_\beta))_{\frac{\gamma-\alpha}{\beta-\alpha}}, \tag{9}$$

with $\alpha, \beta \in [0,1]$ and $\gamma \in [\alpha, \beta]$. The α-Jensen divergence is recovered when $\alpha = \beta = \gamma$ (Fig. 1).

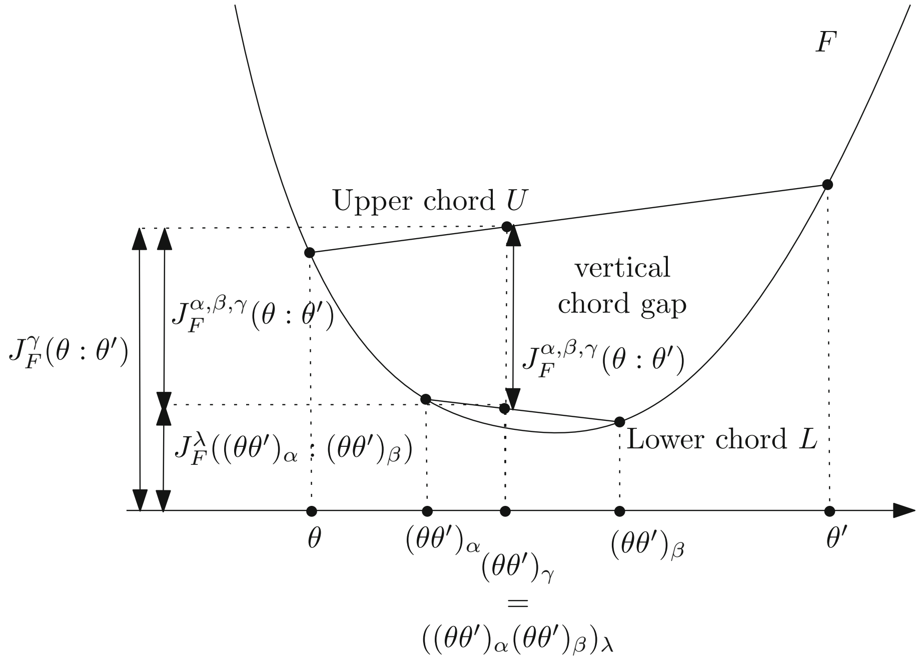
F
Upper chord U
vertical
chord gap
$J_F^{\alpha,\beta,\gamma}(\theta : \theta')$
$J_F^{\alpha,\beta,\gamma}(\theta : \theta')$
$J_F^\gamma(\theta : \theta')$
$J_F^\lambda((\theta\theta')_\alpha : (\theta\theta')_\beta)$
Lower chord L
θ
$(\theta\theta')_\alpha$
$(\theta\theta')_\gamma$
$=$
$((\theta\theta')_\alpha(\theta\theta')_\beta)_\lambda$
$(\theta\theta')_\beta$
θ'

Fig. 1. The Jensen chord gap divergence.

For any given distance $D : \Theta \times \Theta \rightarrow \mathbb{R}_+$ (with convex parameter space Θ), we can bi-skew the distance by considering two scalars $\gamma, \delta \in \mathbb{R}$ (with $\delta \neq \gamma$) as:

$$D_{\gamma,\delta}(\theta_1 : \theta_2) := D((\theta_1\theta_2)_\gamma : (\theta_1\theta_2)_\delta). \tag{10}$$

Clearly, $(\theta_1\theta_2)_\gamma = (\theta_1\theta_2)_\delta$ if and only if $(\delta-\gamma)(\theta_1-\theta_2) = 0$. That is, if (i) $\theta_1 = \theta_2$ or if (ii) $\delta = \gamma$. Since by definition $\delta \neq \gamma$, we have $D_{\gamma,\delta}(\theta_1 : \theta_2) = 0$ if and only if $\theta_1 = \theta_2$. Notice that both $(\theta_1\theta_2)_\gamma = (1-\gamma)\theta_1 + \gamma\theta_2$ and $(\theta_1\theta_2)_\delta = (1-\delta)\theta_1 + \delta\theta_2$ should belong to the parameter space Θ. A sufficient condition is to ensure that $\gamma, \delta \in [0, 1]$ so that both $(\theta_1\theta_2)_\gamma \in \Theta$ and $(\theta_1\theta_2)_\delta \in \Theta$. When $\Theta = \mathbb{R}^d$, we may further consider any $\gamma, \delta \in \mathbb{R}$.

3 The Scalar Bregman Chord Divergence

Let $F : \Theta \subset \mathbb{R} \rightarrow \mathbb{R}$ be a univariate Bregman generator with open convex domain Θ, and denote by $\mathcal{F} = \{(\theta, F(\theta))\}_\theta$ its graph. Let us rewrite the ordinary univariate Bregman divergence [7] of Eq. 1 as follows:

$$B_F(\theta_1 : \theta_2) = F(\theta_1) - T_{\theta_2}(\theta_1), \tag{11}$$

where $y = T_\theta(\omega)$ denotes the equation of the tangent line of F at θ:

$$T_\theta(\omega) := F(\theta) + (\omega - \theta)F'(\theta), \tag{12}$$

Let $\mathcal{T}_\theta = \{(\theta, T_\theta(\omega)) \ : \ \theta \in \Theta\}$ denote the graph of that tangent line. Line $\mathcal{T}_\theta$ is tangent to curve $\mathcal{F}$ at point $P_\theta := (\theta, F(\theta))$. Graphically speaking, the Bregman divergence is interpreted as the *ordinate gap* (gap vertical) between the point $P_{\theta_1} = (\theta_1, F(\theta_1)) \in \mathcal{F}$ and the point of $(\theta_1, T_{\theta_2}(\theta_1)) \in \mathcal{T}_\theta$, as depicted in Fig. 2.

Now let us observe that we may *relax* the tangent line $\mathcal{T}_{\theta_2}$ to a *chord line* (or secant) $\mathcal{C}^{\alpha,\beta}_{\theta_1,\theta_2} = \mathcal{C}_{(\theta_1\theta_2)_\alpha,(\theta_1\theta_2)_\beta}$ passing through the points $((\theta_1\theta_2)_\alpha, F((\theta_1\theta_2)_\alpha))$ and $((\theta_1\theta_2)_\beta, F((\theta_1\theta_2)_\beta))$ for $\alpha, \beta \in (0, 1)$ with $\alpha \neq \beta$ (with corresponding Cartesian equation $C_{(\theta_1\theta_2)_\alpha,(\theta_1\theta_2)_\beta}$), and still get a non-negative vertical gap between $(\theta_1, F(\theta_1))$ and $(\theta_1, C_{(\theta_1\theta_2)_\alpha,(\theta_1\theta_2)_\beta}(\theta_1))$ (because any line intersects a convex body in at most two points). By construction, this vertical gap is smaller than the gap measured by the ordinary Bregman divergence. This yields the Bregman chord divergence ($\alpha, \beta \in (0, 1]$, $\alpha \neq \beta$):

$$B_F^{\alpha,\beta}(\theta_1 : \theta_2) := F(\theta_1) - C_F^{(\theta_1\theta_2)_\alpha,(\theta_1\theta_2)_\beta}(\theta_1) \leq B_F(\theta_1 : \theta_2), \tag{13}$$

illustrated in Fig. 3. By expanding the chord equation and massaging the equation, we get the following formula:

$$\begin{aligned} B_F^{\alpha,\beta}(\theta_1 : \theta_2) &:= F(\theta_1) - \Delta_F^{\alpha,\beta}(\theta_1, \theta_2)(\theta_1 - (\theta_1\theta_2)_\alpha) - F((\theta_1\theta_2)_\alpha), \qquad (14)\\ &= F(\theta_1) - F\left((\theta_1\theta_2)_\alpha\right) + \frac{\alpha\left\{F\left((\theta_1\theta_2)_\alpha\right) - F\left((\theta_1\theta_2)_\beta\right)\right\}}{\beta - \alpha}, \end{aligned}$$

where

$$\Delta_F^{\alpha,\beta}(\theta_1, \theta_2) := \frac{F((\theta_1\theta_2)_\alpha) - F((\theta_1\theta_2)_\beta)}{(\theta_1\theta_2)_\alpha - (\theta_1\theta_2)_\beta} \tag{15}$$

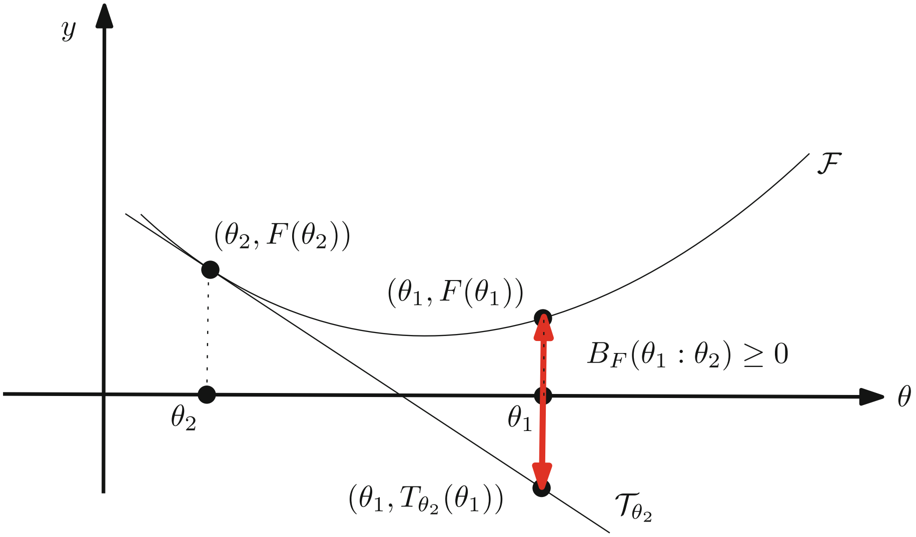
y
$\mathcal{F}$
$(\theta_2, F(\theta_2))$
$(\theta_1, F(\theta_1))$
$B_F(\theta_1 : \theta_2) \geq 0$
θ
θ_2
θ_1
$(\theta_1, T_{\theta_2}(\theta_1))$
$\mathcal{T}_{\theta_2}$

Fig. 2. Bregman divergence as the vertical gap between the generator graph $\mathcal{F}$ and the tangent line $\mathcal{T}_{\theta_2}$ at θ_2.

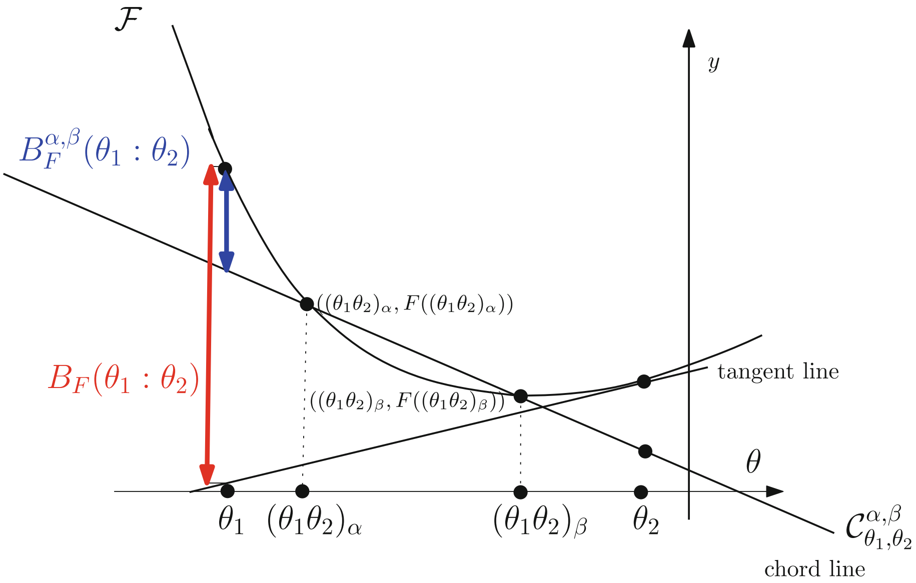
$\mathcal{F}$
y
$B_F^{\alpha,\beta}(\theta_1 : \theta_2)$
$((\theta_1\theta_2)_\alpha, F((\theta_1\theta_2)_\alpha))$
$B_F(\theta_1 : \theta_2)$
tangent line
$((\theta_1\theta_2)_\beta, F((\theta_1\theta_2)_\beta))$
θ
θ_1
$(\theta_1\theta_2)_\alpha$
$(\theta_1\theta_2)_\beta$
θ_2
$\mathcal{C}^{\alpha,\beta}_{\theta_1,\theta_2}$
chord line

Fig. 3. The Bregman chord divergence $B_F^{\alpha,\beta}(\theta_1 : \theta_2)$.

is the slope of the chord, and since $(\theta_1\theta_2)_\alpha - (\theta_1\theta_2)_\beta = (\beta-\alpha)(\theta_1-\theta_2)$ and $\theta_1 - (\theta_1\theta_2)_\alpha = \alpha(\theta_1-\theta_2)$.

Notice the symmetry $B_F^{\alpha,\beta}(\theta_1:\theta_2) = B_F^{\beta,\alpha}(\theta_1:\theta_2)$. We have

$$\lim_{\alpha\to 1,\beta\to 1} B_F^{\alpha,\beta}(\theta_1:\theta_2) = B_F(\theta_1:\theta_2). \tag{16}$$

When $\alpha\to\beta$, the Bregman chord divergences yields a subfamily of *Bregman tangent divergences*:

$$B_F^{\alpha}(\theta_1:\theta_2) = \lim_{\beta\to\alpha} B_F^{\alpha,\beta}(\theta_1:\theta_2) \leq B_F(\theta_1:\theta_2). \tag{17}$$

We consider the tangent line $\mathcal{T}_{(\theta_1\theta_2)_\alpha}$ at $(\theta_1\theta_2)_\alpha$ and measure the ordinate gap at θ_1 between the function plot and this tangent line:

$$\begin{aligned} B_F^{\alpha}(\theta_1:\theta_2) &:= F(\theta_1) - F\left((\theta_1\theta_2)_\alpha\right) - \left(\theta_1 - (\theta_1\theta_2)_\alpha\right)^\top \nabla F\left((\theta_1\theta_2)_\alpha\right), \\ &= F(\theta_1) - F\left((\theta_1\theta_2)_\alpha\right) - \alpha(\theta_1-\theta_2)^\top \nabla F\left((\theta_1\theta_2)_\alpha\right), \end{aligned} \tag{18}$$

for $\alpha\in(0,1]$. The ordinary Bregman divergence is recovered when $\alpha=1$. Notice that the *mean value theorem* yields $\Delta_F^{\alpha,\beta}(\theta_1,\theta_2) = F'(\xi)$ for $\xi\in(\theta_1,\theta_2)$. Thus $B_F^{\alpha,\beta}(\theta_1:\theta_2) = B_F^{\xi}(\theta_1:\theta_2)$ for $\xi\in(\theta_1,\theta_2)$. Letting $\beta=1$ and $\alpha=1-\epsilon$ (for small values of $1>\epsilon>0$), we can approximate the ordinary Bregman divergence by the Bregman chord divergence without requiring to compute the gradient:

$$B_F(\theta_1:\theta_2) \simeq_{\epsilon\to 0} B_F^{1-\epsilon,1}(\theta_1:\theta_2). \tag{19}$$

4 The Multivariate Bregman Chord Divergence

When the generator is separable [3], i.e., $F(x)=\sum_i F_i(x_i)$ for univariate generators F_i, we extend easily the Bregman chord divergence as:

$$B_F^{\alpha,\beta}(\theta:\theta') = \sum_i B_{F_i}^{\alpha,\beta}(\theta_i:\theta_i'). \tag{20}$$

Otherwise, we have to carefully define the notion of "slope" for the multivariate case. An example of such a non-separable multivariate generator is the Legendre dual of the Shannon negentropy: The log-sum-exp function [24,25]:

$$F(\theta) = \log(1+\sum_i e^{\theta_i}). \tag{21}$$

Given a multivariate (non-separable) Bregman generator $F(\theta)$ with $\Theta\subseteq\mathbb{R}^D$ and two prescribed distinct parameters θ_1 and θ_2, consider the following univariate function, for $\lambda\in\mathbb{R}$:

$$F_{\theta_1,\theta_2}(\lambda) := F\left((1-\lambda)\theta_1 + \lambda\theta_2\right) = F\left(\theta_1 + \lambda(\theta_2-\theta_1)\right), \tag{22}$$

with $F_{\theta_1,\theta_2}(0) = F(\theta_1)$ and $F_{\theta_1,\theta_2}(1) = F(\theta_2)$.

The functions $\{F_{\theta_1,\theta_2}\}_{\theta_1\neq\theta_2}$ are strictly convex and differentiable univariate Bregman generators.

Proof. To prove the strict convexity of a univariate function G, we need to show that for any $\alpha \in (0,1)$, we have $G((1-\alpha)x+\alpha y) < (1-\alpha)G(x)+\alpha G(y)$.

$$\begin{aligned}
F_{\theta_1,\theta_2}((1-\alpha)\lambda_1+\alpha\lambda_2) &= F\left(\theta_1+((1-\alpha)\lambda_1+\alpha\lambda_2)(\theta_2-\theta_1)\right),\\
&= F((1-\alpha)(\lambda_1(\theta_2-\theta_1)+\theta_1)+\alpha((\lambda_2(\theta_2-\theta_1)+\theta_1))),\\
&< (1-\alpha)F(\lambda_1(\theta_2-\theta_1)+\theta_1)+\alpha F((\lambda_2(\theta_2-\theta_1)+\theta_1)),\\
&< (1-\alpha)F_{\theta_1,\theta_2}(\lambda_1)+\alpha F_{\theta_1,\theta_2}(\lambda_2).
\end{aligned}$$

Then we define the multivariate Bregman chord divergence by applying the definition of the univariate Bregman chord divergence on these families of univariate Bregman generators:

$$B_F^{\alpha,\beta}(\theta_1:\theta_2) := B_{F_{\theta_1,\theta_2}}^{\alpha,\beta}(0:1), \tag{23}$$

Since $(01)_\alpha = \alpha$ and $(01)_\beta = \beta$, we get:

$$\begin{aligned}
B_F^{\alpha,\beta}(\theta_1:\theta_2) &= F_{\theta_1,\theta_2}(0) + \frac{\alpha(F_{\theta_1,\theta_2}(\alpha)-F_{\theta_1,\theta_2}(\beta))}{\beta-\alpha} - F_{\theta_1,\theta_2}(\alpha),\\
&= F(\theta_1) - F\left((\theta_1\theta_2)_\alpha\right) - \frac{\alpha\left(F\left((\theta_1\theta_2)_\beta\right)-F\left((\theta_1\theta_2)_\alpha\right)\right)}{\beta-\alpha},
\end{aligned}$$

in accordance with the univariate case. Since $(\theta_1\theta_2)_\beta = (\theta_1\theta_2)_\alpha - (\beta-\alpha)(\theta_2-\theta_1)$, we have the first-order Taylor expansion

$$F\left((\theta_1\theta_2)_\beta\right) \simeq_{\beta\simeq\alpha} F\left((\theta_1\theta_2)_\alpha\right) - (\beta-\alpha)(\theta_2-\theta_1)^\top \nabla F\left((\theta_1\theta_2)_\alpha\right). \tag{24}$$

Therefore, we have:

$$\frac{\alpha\left(F\left((\theta_1\theta_2)_\beta\right)-F\left((\theta_1\theta_2)_\alpha\right)\right)}{\beta-\alpha} \simeq -\alpha(\theta_2-\theta_1)^\top \nabla F\left((\theta_1\theta_2)_\alpha\right). \tag{25}$$

This proves that

$$\lim_{\beta\rightarrow\alpha} B_F^{\alpha,\beta}(\theta_1:\theta_2) = B_F^\alpha(\theta_1:\theta_2). \tag{26}$$

Notice that the Bregman chord divergence does *not* require to compute the gradient ∇F The "slope term" in the definition is reminiscent to the q-derivative [18] (quantum/discrete derivatives). However the (p,q)-derivatives [18] are defined with respect to a *single* reference point while the chord definition requires *two* reference points.

5 Conclusion

In this paper, we geometrically designed a new class of distances using a Bregman generator and two additional scalar parameters, termed the *Bregman chord divergence*, and its one-parametric subfamily, the *Bregman tangent divergences* that includes the ordinary Bregman divergence. This generalization allows one

to easily fine-tune Bregman divergences in applications by adjusting smoothly one or two (scalar) knobs. Moreover, by choosing $\alpha = 1 - \epsilon$ and $\beta = 1$ for small $\epsilon > 0$, the Bregman chord divergence $B_F^{1-\epsilon,1}(\theta_1 : \theta_2)$ lower bounds closely the Bregman divergence $B_F(\theta_1 : \theta_2)$ without requiring to compute the gradient (a different approximation without gradient is $\frac{1}{\epsilon} J_F^{\epsilon}(\theta_2 : \theta_1)$). We expect that this new class of distances brings further improvements in signal processing and information fusion applications [29] (e.g., by tuning $B_{F_{\mathrm{KL}}}^{\alpha,\beta}$ or $B_{F_{\mathrm{IS}}}^{\alpha,\beta}$). While the Bregman chord divergence defines an ordinate gap on the exterior of the epigraph, the Jensen chord divergence [20] defines the gap inside the epigraph of the generator. In future work, the dualistic information-geometric structure induced by the Bregman chord divergences shall be investigated from the viewpoint of gauge theory [19] and in contrast with the dually flat structures of Bregman manifolds [3].
Source code in Java$^{\mathrm{TM}}$ is available for reproducible research.[3]

References

1. 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)
2. Amari, S.I.: α-divergence is unique, belonging to both f-divergence and Bregman divergence classes. IEEE Trans. Inf. Theory **55**(11), 4925–4931 (2009)
3. Amari, S.: Information Geometry and Its Applications. AMS, vol. 194. Springer, Tokyo (2016). https://doi.org/10.1007/978-4-431-55978-8
4. Banerjee, A., Guo, X., Wang, H.: On the optimality of conditional expectation as a Bregman predictor. IEEE Trans. Inf. Theory **51**(7), 2664–2669 (2005)
5. Banerjee, A., Merugu, S., Dhillon, I.S., Ghosh, J.: Clustering with Bregman divergences. J. Mach. Learn. Res. **6**(Oct), 1705–1749 (2005)
6. Basseville, M.: Divergence measures for statistical data processing: an annotated bibliography. Sig. Process. **93**(4), 621–633 (2013)
7. 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)
8. Broniatowski, M., Stummer, W.: Some universal insights on divergences for statistics, machine learning and artificial intelligence. In: Nielsen, F. (ed.) Geometric Structures of Information. SCT, pp. 149–211. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-02520-5_8
9. Burbea, J., Rao, C.: On the convexity of some divergence measures based on entropy functions. IEEE Trans. Inf. Theory **28**(3), 489–495 (1982)
10. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, New York (2012)
11. Csiszár, I.: Eine infonnationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizitlit von Markoffschen Ketten. Magyar Tudományos Akadémia - MAT **8**, 85–108 (1963)
12. Csiszár, I.: Information-type measures of difference of probability distributions and indirect observation. Studia Scientiarum Mathematicarum Hungarica **2**, 229–318 (1967)

[3] https://franknielsen.github.io/~nielsen/BregmanChordDivergence/.

13. Csiszár, I.: Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. Ann. Stat. **19**(4), 2032–2066 (1991). https://doi.org/10.1007/978-1-4613-0071-7
14. Deza, M.M., Deza, E.: Encyclopedia of distances. In: Deza, M.M., Deza, E. (eds.) Encyclopedia of Distances, pp. 1–583. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-00234-2_1
15. Févotte, C.: Majorization-minimization algorithm for smooth Itakura-Saito nonnegative matrix factorization. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1980–1983. IEEE (2011)
16. Jensen, J.L.W.V.: Sur les fonctions convexes et les inégalités entre les valeurs moyennes. Acta Math. **30**(1), 175–193 (1906)
17. Jiao, J., Courtade, T.A., No, A., Venkat, K., Weissman, T.: Information measures: the curious case of the binary alphabet. IEEE Trans. Inf. Theory **60**(12), 7616–7626 (2014)
18. Kac, V., Cheung, P.: Quantum Calculus. Springer, New York (2001)
19. Naudts, J., Zhang, J.: Rho-tau embedding and gauge freedom in information geometry. Inf. Geom. **1**(1), 79–115 (2018)
20. Nielsen, F.: The chord gap divergence and a generalization of the Bhattacharyya distance. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2276–2280, April 2018. https://doi.org/10.1109/ICASSP.2018.8462244
21. Nielsen, F., Boltz, S.: The Burbea-Rao and Bhattacharyya centroids. IEEE Trans. Inf. Theory **57**(8), 5455–5466 (2011)
22. Nielsen, F., Nock, R.: Sided and symmetrized Bregman centroids. IEEE Trans. Inf. Theory **55**(6), 2882–2904 (2009)
23. Nielsen, F., Nock, R.: On the geometry of mixtures of prescribed distributions. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2861–2865. IEEE (2018)
24. Nielsen, F., Sun, K.: Guaranteed bounds on information-theoretic measures of univariate mixtures using piecewise log-sum-exp inequalities. Entropy **18**(12), 442 (2016)
25. Nielsen, F., Sun, K.: Guaranteed bounds on the Kullback-Leibler divergence of univariate mixtures. IEEE Signal Process. Lett. **23**(11), 1543–1546 (2016)
26. Nock, R., Magdalou, B., Briys, E., Nielsen, F.: Mining matrix data with Bregman matrix divergences for portfolio selection. In: Nielsen, F., Bhatia, R. (eds.) Matrix Information Geometry, pp. 373–402. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-30232-9_15
27. Pardo, M.C., Vajda, I.: About distances of discrete distributions satisfying the data processing theorem of information theory. IEEE Trans. Inf. Theory **43**(4), 1288–1293 (1997)
28. Stummer, W., Vajda, I.: On divergences of finite measures and their applicability in statistics and information theory. Statistics **44**(2), 169–187 (2010)
29. Üney, M., Houssineau, J., Delande, E., Julier, S.J., Clark, D.E.: Fusion of finite set distributions: pointwise consistency and global cardinality. CoRR abs/1802.06220 (2018). http://arxiv.org/abs/1802.06220
30. Zhang, J.: Divergence function, duality, and convex analysis. Neural Comput. **16**(1), 159–195 (2004)