Corrigendum and addendum to: "Sided and symmetrized Bregman centroids" IEEE Transactions on Information Theory 55.6 (2009): 2882-2904.

Frank Nielsen Richard Nock

July 2020

Abstract

We correct and extend the results presented in [12].

1 Dissimilarities, dual centroids, and dual information radii

Let $D(P:Q)$ denote the *dissimilarity* between two points P and Q of a space $\mathbb{G}$ such that $D(P:Q) \geq 0$ with equality if and only if $P=Q$. By analogy with the notion of Fréchet barycenters in metric spaces [7], we define the *D-barycenters* or *D-centroid* $C_D(\mathcal{P})$ of a weighted point set $\mathcal{P}=\{P_1,\ldots,P_n\}$ with respect to D as

$$C_D(\mathcal{P}) := \arg\min_{X\in\mathbb{G}} \sum_{i=1}^n w_i D(P_i:X), \tag{1}$$

where $w_i>0$ and $\sum_{i=1}^n w_i = 1$ (i.e., w belongs to the $(n-1)$-dimensional standard simplex Δ_{n-1}). The centroids are special cases of barycenters obtained for the uniform weighting $w_i=\frac{1}{n}$. Notice that $C_D(\mathcal{P})$ is generally a subset of points of $\mathbb{G}$, and may not necessarily exist nor be unique. For example, the centroid of two antipodal points on the unit Euclidean sphere is a great circle. In Riemannian geometry, other notions of barycenters have been defined [1]: Karcher local barycenters, exponential barycenters, etc.

Since D may be asymmetric $D(P:Q)\neq D(Q:P)$ (oriented dissimilarity, hence the delimiter notation ":"), we define the *dual dissimilarity* $D^*(P:Q):=D(Q:P)$, and the *dual D-barycenter* or *left-sided D-barycenter*:

$$\begin{aligned} {C_D}^*(\mathcal{P}) &:= \arg\min_{X\in\mathbb{G}} \sum_{i=1}^n w_i D(X:P_i), & (2)\\ &= \arg\min_{X\in\mathbb{G}} \sum_{i=1}^n w_i D^*(P_i:X), & (3)\\ &= C_{D^*}(\mathcal{P}). & (4)\end{aligned}$$

Notice that the dual of the dual dissimilarity is the original (primal) dissimilarity: $D^{**}=D$ (involutive property of duality).

Let $C_D(\mathcal{P})$ be the primal D-barycenter (*right-sided D-barycenter*) and $C_D{}^*(\mathcal{P})$ be the dual D-barycenter (left-sided D-barycenter). The dual D-barycenter with respect to D amounts to the (primal) D^*-barycenter for the dual dissimilarity D^*. When D is the squared Euclidean distance, both primal and dual centroids coincide to the center of mass.

The (primal) *information radius* [13] is defined by

$$I_D(\mathcal{P}) := \sum_{i=1}^n w_i D(P_i : C), \quad C \in C_D(\mathcal{P}), \tag{5}$$

while the *dual information radius* is defined by

$$I_D{}^*(\mathcal{P}) := \sum_{i=1}^n w_i D(C : P_i), \quad C \in C_D{}^*(\mathcal{P}). \tag{6}$$

In general, we have $I_D{}^*(\mathcal{P}) \neq I_{D^*}(\mathcal{P})$ because the left-sided and right-sided centroids may not coincide. (They coincide by default when the dissimilarity is symmetric.) The information radius for the squared Euclidean distance represents the variance of the point set.

2 Bregman centroids and Bregman information

Let $F(\theta)$ be a strictly convex and differentiable real-valued function for $\theta \in \Theta$, where $\Theta \subset \mathbb{R}^D$ denotes the open parameter space. We define the *Bregman divergence* [6] with respect to generator F as:

$$B_F(\theta : \theta') := F(\theta) - F(\theta') - (\theta - \theta')^\top \nabla F(\theta'), \tag{7}$$

for $\theta, \theta' \in \Theta$.

Bregman divergences are canonical smooth dissimilarities of *dually flat space* in information geometry [2, 8]: That is, we can build a canonical Bregman divergence from any dually flat space, and a Bregman divergence yields a dually flat space [3]. In a dually flat space (or *Bregman manifold* [9]), the dissimilarity between two points P and Q is expressed by

$$D_F(P : Q) := B_F(\theta(P) : \theta(Q)), \tag{8}$$

where $\theta(\cdot)$ is a *global (affine) coordinate system* used to define the *potential function* $F(\theta)$, see [2, 8]. The dual divergence amounts to a dual Bregman divergence B_{F^*} as follows:

$$D_F^*(P : Q) = D(Q : P) = B_F(\theta(Q) : \theta(P)) = B_{F^*}(\eta(P) : \eta(Q)) = D_{F^*}(P : Q), \tag{9}$$

where F^* is the Legendre-Fenchel convex conjugate [9], and $\eta(\theta) = \nabla F(\theta)$ the dual affine global coordinate system [2, 8]. We can introduce the *Legendre-Fenchel divergence* from the dual potential functions F and F^* as follows:

$$A_F(\theta : \eta') := F(\theta) + F^*(\eta') - \theta^\top \eta' \geq 0 \tag{10}$$

with equality if and only if $\eta' = \nabla F(\theta)$, or equivalently $\theta = \nabla F^*(\eta')$.

Thus, in a Bregman manifold, we have the dual divergences that can be expressed using the dual coordinate systems either by Bregman divergences or by Legendre-Fenchel divergences as follows:

$$D_F(P : Q) = B_F(\theta(P) : \theta(Q)) = A_F(\theta(P) : \eta(Q)) =: D_F{}^*(Q : P), \tag{11}$$

$$D_F{}^*(P : Q) = B_{F^*}(\eta(P) : \eta(Q)) = A_{F^*}(\eta(P) : \theta(Q)) =: D_F(Q : P). \tag{12}$$

Theorem 1 *Theorem 3.1 and Theorem 3.2 of [12] Let $\theta_i = \theta(P_i)$ and $\eta_i = \eta(P_i)$ be the primal and dual coordinates of point P_i for $P_i \in \mathcal{P} = \{P_1, \ldots, P_n\}$. Let $\bar{\theta} = \sum_{i=1}^n w_i \theta_i$ and $\bar{\eta} = \sum_{i=1}^n w_i \eta_i$ denote the center of mass in the primal θ-coordinate system and dual η-coordinate system, respectively. The right-sided Bregman centroid $C_{D_F}(\mathcal{P})$ and the left-sided Bregman centroid ${C_{D_F}}^*(\mathcal{P})$ exist and are both unique, and we have $\theta(C_{D_F}(\mathcal{P})) = \bar{\theta}$ and $\eta({C_{D_F}}^*(\mathcal{P})) = \bar{\eta}$.*

Proof: We have

$$
\begin{aligned}
C_{D_F}(\mathcal{P}) &= \arg\min_{X \in \mathbb{G}} \sum_{i=1}^n w_i D_F(P_i : X), && (13)\\
&= \arg\min_{X \in \mathbb{G}} \sum_{i=1}^n w_i A_F(\theta_i : \eta(X)), && (14)\\
&= \arg\min_{X \in \mathbb{G}} E(X) = (\sum_{i=1}^n w_i F(\theta_i)) + F^*(\eta(X)) - \bar{\theta}^\top \eta(X). && (15)
\end{aligned}
$$

A point $X \in C_{D_F}(\mathcal{P})$ if and only if $\nabla_{\eta(X)} = 0$: $\nabla_\eta F^*(\eta(X)) = \bar{\theta}$. That is:

$$
\eta(X) = (\nabla F^*)^{-1}(\bar{\theta}) = (\nabla F^*)^{-1}(\sum_{i=1}^n w_i \nabla F^*(\eta_i)). \tag{16}
$$

The right-sided centroid is unique since the Hessian $\nabla^2_{\eta(X)} E(X)$ is $\nabla^2 F^*(\eta(X))$, and $\nabla^2 F^*$ is positive-definite (F^* is a strictly convex conjugate). The right-sided centroid is expressed in the θ-coordinate system as $\theta(C_{D_F}(\mathcal{P})) = (\nabla F^*)(\eta(C_{D_F}(\mathcal{P}))) = (\nabla F^*)((\nabla F^*)^{-1}(\bar{\theta})) = \bar{\theta}$.

The proof for the left-sided centroid is similar, and we have $\theta({C_{D_F}}^*(\mathcal{P})) = (\nabla F)^{-1}(\bar{\eta}) = (\nabla F)^{-1}(\sum_{i=1}^n w_i \nabla F(\theta_i))$ so that ${C_{D_F}}^*(\mathcal{P})$ expressed in the η-coordinate system is $\bar{\eta}$. $\square$

To summarize, we have:

	θ-coordinate system	η-coordinate system
Right-sided centroid $C_{D_F}(\mathcal{P})$	$\bar{\theta} = \sum_{i=1}^n w_i \theta_i$	$(\nabla F^*)^{-1}(\sum_{i=1}^n w_i \nabla F^*(\eta_i))$
Left-sided centroid ${C_{D_F}}^*(\mathcal{P})$	$(\nabla F)^{-1}(\sum_{i=1}^n w_i \nabla F(\theta_i))$	$\bar{\eta} = \sum_{i=1}^n w_i \eta_i$

In term of Bregman divergences, the right-sided Bregman centroid is the center of mass [4]. The Bregman information radius is called *Bregman information* in [4]. It was shown in [11, 5] that the only *symmetrized Bregman divergences* are squared Mahalanobis divergences. Thus the left-sided centroid and right-sided Bregman centroids coincide only for squared Mahalanobis divergences, and the dual Bregman information radii differ in the general case.

Corollary 1 *Correct Corollary 3.3 of [12] The information radius $I_{D_F}(\mathcal{P}) = J_F(\theta_1, \ldots, \theta_n; w_1, \ldots, w_n)$ where J_F denotes the* Jensen diversity index *[10]:*

$$
J_F(\theta_1, \ldots, \theta_n; w_1, \ldots, w_n) := \sum_{i=1}^n w_i F(\theta_i) - F\left(\sum_{i=1}^n w_i \theta_i\right) \geq 0. \tag{17}
$$

The dual information radius ${I_{D_F}}^*(\mathcal{P}) = I_{D_F^*}(\mathcal{P}) = J_{F^*}(\eta_1, \ldots, \eta_n; w_1, \ldots, w_n)$ differs from the primal information radius except when D_F is a squared Mahalanobis divergence.

Thus we have:

$$
\begin{aligned}
I_{D_F}(\mathcal{P}) &= \sum_{i=1}^{n} w_i F(\theta_i) - F\left(\sum_{i=1}^{n} w_i \theta_i\right), &(18)\\
I_{D_F^*}(\mathcal{P}) &= \sum_{i=1}^{n} w_i F^*(\eta_i) - F^*\left(\sum_{i=1}^{n} w_i \eta_i\right). &(19)
\end{aligned}
$$

Example 1 *When $F(\theta)=\frac{1}{2}\theta^\top Q\theta$ for a positive-definite matrix $Q\succ 0$, we have the convex conjugate $F^*(\eta)=\frac{1}{2}\eta^\top Q^{-1}\eta$ (with $Q^{-1}\succ 0$). We have $\eta_i=Q^{-1}\theta_i$ and $\eta_i=Q\theta_i$. It follows that $\bar\theta=\sum_{i=1}^n w_i\theta_i=Q^{-1}\bar\eta$ and $\bar\eta=\sum_{i=1}^n w_i\eta_i=Q\bar\theta$. Thus we check that the information radii coincide when dealing with squared Mahalanobis Bregman divergences:*

$$
\begin{aligned}
I_{D_F}(\mathcal{P}) &= \sum_{i=1}^{n} w_i \frac{1}{2}\theta_i^\top Q\theta_i - \frac{1}{2}\bar\theta^\top Q\bar\theta, &(20)\\
&= \sum_{i=1}^{n} w_i \frac{1}{2}(Q^{-1}\eta_i)^\top Q(Q^{-1}\eta_i) - \frac{1}{2}(Q^{-1}\bar\eta)^\top Q(Q^{-1}\bar\eta), &(21)\\
&= \sum_{i=1}^{n} w_i \eta_i^\top Q^{-1}\eta_i - \frac{1}{2}\bar\eta Q^{-1}\bar\eta, &(22)\\
&= I_{D_{F^*}}(\mathcal{P}) = I_{D_F^*}(\mathcal{P}). &(23)
\end{aligned}
$$

Let $Q=LL^\top$ be the Cholesky decomposition of a positive-definite matrix $Q\succ 0$. It is well-known that the Mahalanobis distance amounts to the Euclidean distance on affinely transformed points:

$$
\begin{aligned}
M_Q^2(\theta,\theta') &= \Delta\theta^\top Q\Delta\theta, &(24)\\
&= \Delta\theta^\top LL^\top\Delta\theta, &(25)\\
&= M_I^2(L^\top\theta, L^\top\theta') = \|L^\top\theta - L^\top\theta'\|^2. &(26)
\end{aligned}
$$

Conversely, we can transform the Euclidean distance as an equivalent Mahalanobis distance on affinely transformed points:

$$M_Q((L^\top)^{-1}\theta,(L^\top)^{-1}\theta') = M_I(\theta,\theta') = \|\theta-\theta'\|.$$

Thus the Euclidean distance can be rewritten as the following equivalent Mahalanobis distances:

$$M_{Q_2}((L_2^\top)^{-1}\theta,(L_2^\top)^{-1}\theta') = M_{Q_1}((L_1^\top)^{-1}\theta,(L_1^\top)^{-1}\theta') = \|\theta-\theta'\| = M_I(\theta,\theta')$$

It follows that we can transform one Mahalanobis distance M_{Q_2} into another Mahalanobis distance M_{Q_1} by a linear transformation:

$$M_{Q_2}(\theta,\theta') = M_{Q_1}((L_1^\top)^{-1}L_2^\top\theta,(L_1^\top)^{-1}L_2^\top\theta').$$

Observe that when $Q_1=I$, we have $L_1=I$, and we recover $M_{Q_2}(\theta,\theta')=M_I(L_2^\top\theta,L_2^\top\theta')=\|L_2^\top\theta-L_2^\top\theta'\|$, as expected.

For any lower triangular matrix, we have $(L^{-1})^\top=(L^\top)^{-1}$.

Let $L_{12} = L_2 \left((L_1^\top)^{-1} \right)^\top$. Notice that $L_{12} = L_2 L_1^{-1}$. Therefore we have $M_{Q_2}(\theta, \theta') = M_{Q_1}(L_{12}^\top \theta, L_{12}^\top \theta')$.

Another short proof consists in writing for symmetric positive-definite (SPD) matrix $Q = L^\top L \succ 0$ that

$$M_Q(\theta_1, \theta_2) = M_I(L^\top \theta_1, L^\top \theta_2) \Leftrightarrow M_I(\theta_1, \theta_2) = M_Q((L^\top)^{-1}\theta_1, ((L^\top)^{-1}\theta_2).$$

Then we have for two SPD matrices $Q_1 = L_1^\top L_1 \succ 0$ and $Q_2 = L_2^\top L_2 \succ 0$:

$$M_{Q_1}(\theta_1, \theta_2) = M_I(L_1^\top \theta_1, L_1^\top \theta_2) = M_{Q_2}((L_2^\top)^{-1} L_1^\top \theta_1, (L_2^\top)^{-1} L_1^\top \theta_2).$$

Thus we have

$$M_{Q_1}(\theta_1, \theta_2) = M_{Q_2}((L_2^\top)^{-1} L_1^\top \theta_1, (L_2^\top)^{-1} L_1^\top \theta_2).$$

3 The symmetrized Bregman centroids

Acknowledgments: We kindly thank Professor Gavin Brown of the University of Manchester (UK) for communications on this topic.

References

[1] Adil Ahidar-Coutrix, Thibaut Le Gouic, and Quentin Paris. Convergence rates for empirical barycenters in metric spaces: curvature, convexity and extendable geodesics. *Probability Theory and Related Fields*, pages 1–46, 2019.

[2] Shun-ichi Amari. *Information Geometry and Its Applications*. Applied Mathematical Sciences. Springer Japan, 2016.

[3] Shun-ichi Amari and Andrzej Cichocki. Information geometry of divergence functions. *Bulletin of the polish academy of sciences. Technical sciences*, 58(1):183–195, 2010.

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

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

[6] Lev 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*, 7(3):200–217, 1967.

[7] Maurice Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'institut Henri Poincaré*, 10(4):215–310, 1948.

[8] Frank Nielsen. An elementary introduction to information geometry. Technical report, arXiv preprint, 2018. arxiv:1808.08271.

[9] Frank Nielsen. On geodesic triangles with right angles in a dually flat space. Technical report, arXiv preprint, 2019. arXiv:1910.03935.

[10] Frank Nielsen. On a generalization of the Jensen–Shannon divergence and the Jensen–Shannon centroid. *Entropy*, 22(2):221, 2020.

[11] Frank Nielsen, Jean-Daniel Boissonnat, and Richard Nock. Bregman Voronoi diagrams: Properties, algorithms and applications. *arXiv preprint arXiv:0709.2196*, 2007.

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

[13] Robin Sibson. Information radius. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 14(2):149–160, 1969.