Bregman sided and symmetrized centroids

Frank Nielsen
École Polytechnique
Sony Computer Science Laboratories
France
nielsen@lix.polytechnique.fr

Richard Nock
CEREGMIA
University of Antilles-Guyane
France
rnock@martinique.univ-ag.fr

Abstract

We generalize the notions of centroids and barycenters to the broad class of information-theoretic distortion measures called Bregman divergences. Because Bregman divergences are typically asymmetric, we consider both the left-sided and right-sided centroids and the symmetrized centroids, and prove that all three are unique. We give closed-form solutions for the sided centroids that are generalized means, and design a provably fast and efficient approximation algorithm for the symmetrized centroid based on its exact geometric characterization that requires solely to walk on the geodesic linking the two sided centroids.

1. Introduction

Content-based multimedia retrieval applications with their prominent image retrieval systems (CBIRs) are very popular nowadays with the broad availability of massive digital multimedia libraries. CBIR systems spurred an intensive line of research for better *ad-hoc* feature extractions and effective yet accurate geometric clustering techniques. In a typical CBIR system [5], database images are processed offline during a preprocessing step by various feature extractors computing image characteristics such as color histograms. These features are aggregated into signature vectors that represent handles to images. Then given an online query image, the system first computes its signature, and search for the first, say h, best matches in the signature space. This requires to define an appropriate *similarity measure* between pairs of signatures. Designing an appropriate distance is tricky since the signature space is often heterogeneous (ie., cartesian product of feature spaces) and the usual Euclidean distance or L_p-norms do not always make sense. For example, it is better to use the information-theoretic relative entropy, known as the Kullback-Leibler divergence, to measure the *oriented distance* between image histograms. *Efficiency* is another key issue of CBIR systems since we do not want to compute the similarity measure (query,image) for each image in the database. We rather want to prealably *cluster* the signatures efficiently during the preprocessing stage for fast retrieval of the best matches given query signature points. A first seminal work by Lloyd in 1957 [1] proposed the k-means iterative clustering algorithm. In short, k-means starts by choosing k seeds for cluster centers, associate to each point its "closest" cluster "center," update the various cluster centers, and reiterate until either convergence is met or the difference of the "loss function" between any two sucessive iterations goes below a prescribed threshold. Lloyd choosed the *squared* Euclidean distance since the minimum average intracluster distance yields centroids, the centers of mass of the respective clusters, and further proved that k-means *monotonically* converges to a *local* optima. Cluster C_i's center c_i is defined by the minimization problem $c_i = \arg\min_c \sum_{p_j \in C_i} ||cp_j||^2 = \frac{1}{|C_i|} \sum_{p_j \in C_i} p_j \stackrel{\text{def}}{=} \arg\min_c \mathrm{AVG}_{L_2^2}(C_i, c)$, where $|C_i|$ denotes the cardinality of C_i. Half a century later, Banerjee et al. [1] showed that the k-means algorithm *extends to* and *only* works for a broad family of distortion measures called Bregman divergences. Bregman divergences D_F are parameterized families of distortion measures that are defined by a strictly convex and differentiable generator function $F : \mathcal{X} \rightarrow \mathbb{R}^+$ (with $\dim \mathcal{X} = d$) as $D_F(p||q) = F(p) - F(q) - < p - q, \nabla F(q) >$, where $< \cdot, \cdot >$ denotes the inner product ($< p, q >= \sum_{i=1}^d p^{(i)} q^{(i)} = p^T q$) and $\nabla F(q)$ the gradient at point q (ie., $\nabla F(q) = \left[\frac{\partial F(q)}{\partial x^{(1)}}, ..., \frac{\partial F(q)}{\partial x^{(d)}}\right]$). The fundamental underlying primitive for these *center-based* clustering algorithms is to find the intrinsic *best single representative* of a cluster. As mentionned

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

above, the centroid of a point set $\mathcal{P} = \{p_1, ..., p_n\}$ is defined as the optimizer of the *minimum average distance*: $c = \arg\min_c \frac{1}{n}\sum_i d(c, p_i)$. For oriented distance functions such as Bregman divergences that are not necessarily symmetric, we thus distinguish *sided* and *symmetrized* centroids as follows: $c_R^F = \arg\min_{c\in\mathcal{X}} \frac{1}{n}\sum_{i=1}^n D_F(p_i||\boxed{c})$, $c_L^F = \arg\min_{c\in\mathcal{X}} \frac{1}{n}\sum_{i=1}^n D_F(\boxed{c}||p_i)$, and $c^F = \arg\min_{c\in\mathcal{X}} \frac{1}{n}\sum_{i=1}^n \frac{D_F(p_i||\boxed{c})+D_F(\boxed{c}||p_i)}{2}$. The first right-type and left-type centroids c_R^F and c_L^F are called *sided centroids*, and the third type centroid c^F is called the *symmetrized* Bregman centroid. Except for the class of generalized quadratic distances with generator $F_Q(x) = x^T Qx$, $S_F(p;q) = \frac{D_F(p||q)+D_F(q||p)}{2}$ is *not* a Bregman divergence, see [6]. Since the three centroids coincide with the center of mass for symmetric Bregman divergences, we consider in the remainder asymmetric Bregman divergences. We write for short $\mathrm{AVG}_F(\mathcal{P}||c) = \frac{1}{n}\sum_{i=1}^n D_F(p_i||c)$, $\mathrm{AVG}_F(c||\mathcal{P}) = \frac{1}{n}\sum_{i=1}^n D_F(c||p_i)$ and $\mathrm{AVG}_F(c;\mathcal{P}) = \frac{1}{n}\sum_{i=1}^n S_F(c;p_i)$, so that we get respectively $c_R^F = \arg\min_{c\in\mathcal{X}} \mathrm{AVG}_F(\mathcal{P}||c), c_L^F = \arg\min_{c\in\mathcal{X}} \mathrm{AVG}_F(c||\mathcal{P})$ and $c^F = \arg\min_{c\in\mathcal{X}} \mathrm{AVG}_F(\mathcal{P};c)$. The symmetrized Kullback-Leibler and COSH centroids(symmetrized Itakura-Saito divergence obtained for $F(x) = -\log x$, the Burg entropy) are certainly the most famous symmetrized Bregman centroids, widely used in image and sound processing.

Prior work in the literature is sparse and disparate: Ben-Tal et al. [3] studied *entropic means* as the minimum average optimization for various distortion measures such as the f-divergences and Bregman divergences. Their study is limited to the sided left-type (generalized means) centroids. Basseville and Cardoso [2] compared in the 1-page paper the generalized/entropic mean values for two entropy-based classes of divergences: f-divergences and Jensen-Shannon divergences [4]. The closest recent work to our study is Veldhuis' approximation method [8] for computing the symmetrical Kullback-Leibler centroid.

2. Sided Bregman centroids

2.1. Right-type centroid

We first prove that the right-type centroid c_R^F is *independent* of the considered Bregman divergence D_F: $c_F(\mathcal{P}) = \bar{p} = \frac{1}{n}\sum_{i=1}^n p_i$ is always the center of mass. Although this result is well-known in disguise in information geometry, it was again recently brought up to the attention of the machine learning community by Banerjee et al. [1] who proved that Lloyd's iterative k-means "centroid" clustering algorithm generalizes to the class of Bregman divergences. We state the result and give the proof for completeness and familizaring us with notations.

Theorem 2.1 *The right-type sided Bregman centroid c_R^F of a set $\mathcal{P}$ of n points p_1, ...p_n, defined as the minimizer for the average right divergence $c_R^F = \arg\min_c \sum_{i=1}^n \frac{1}{n} D_F(p_i||c) = \arg\min_c \mathrm{AVG}_F(\mathcal{P}||c)$, is unique, independent of the selected divergence D_F, and coincides with the center of mass $c_R^F = c_R = \bar{p} = \frac{1}{n}\sum_{i=1}^n p_i$.*

Proof: For a given point q, the right-type average divergence is defined as $\mathrm{AVG}_F(\mathcal{P}||q) = \sum_{i=1}^n \frac{1}{n} D_F(p_i||q)$. Expanding the terms $D_F(p_i||q)$'s using the definition of Bregman divergence, we get $\mathrm{AVG}_F(\mathcal{P}||q) = \sum_{i=1}^n \frac{1}{n}(F(p_i) - F(q) - < p_i - q, \nabla F(q) >)$. Subtracting and adding $F(\bar{p})$ to the right-hand side yields $\mathrm{AVG}_F(\mathcal{P}, q) = \left(\sum_{i=1}^n \frac{1}{n} F(p_i) - F(\bar{p})\right) + \left(F(\bar{p}) - F(q) - \sum_{i=1}^n \frac{1}{n} < p_i - q, \nabla F(q) >\right) = \left(\sum_{i=1}^n \frac{1}{n} F(p_i) - F(\bar{p})\right) + \left(F(\bar{p}) - F(q) - \left\langle \sum_{i=1}^n \frac{1}{n}(p_i - q), \nabla F(q)\right\rangle\right) = \left(\frac{1}{n}\sum_{i=1}^n F(p_i) - F(\bar{p})\right) + D_F(\bar{p}||q)$.

Observe that since $\sum_{i=1}^n \frac{1}{n} F(p_i) - F(\bar{p})$ is *independent* of q, minimizing $\mathrm{AVG}_F(\mathcal{P}||q)$ is equivalent to minimizing $D_F(\bar{p}||q)$. Using the fact that Bregman divergences $D_F(p||q)$ are non-negative, $D_F(p||q) \geq 0$, and equal to zero *if and only if* $p = q$, we conclude that $c_R^F = \arg\min_q \mathrm{AVG}_F(\mathcal{P}||q) = \bar{p}$, namely the center of mass of the point set. The minimization remainder, representing the "information radius" (by generalizing the notion introduced by Sibson [7] for the relative entropy), is $\mathrm{JS}_F(\mathcal{P}) = \frac{1}{n}\sum_{i=1}^n F(p_i) - F(\bar{p}) \geq 0$, which bears the name of the F-Jensen difference[1] [4]. For $F = -H = x\log x$ the negative Shannon entropy, J_F is known as the Jensen-Shannon divergence: $\mathrm{JS}(\mathcal{P}) = H(\sum_{i=1}^n p_i) - \sum_{i=1}^n \frac{1}{n} H(p_i)$. The Jensen-Shannon divergence is also known as half of the Jeffreys divergence (JD): $\mathrm{JS}(P;Q) = \frac{1}{2}\mathrm{JD}(P;Q)$, and can be interpreted as the *expected information gain* when discovering which probability distribution is drawn from (either P or Q).

2.2 Dual divergence and left-type centroid

Before characterizing the *left-type* sided Bregman centroid, we recall the fundamental duality of convex

[1] In the paper [4], it is used for strictly concave function $H = -F$ on a weight distribution vector π: $J_\pi(p_1, ..., p_n) = H(\sum_{i=1}^n \pi_i p_i) - \sum_{i=1}^n \pi_i H(p_i)$. Here, we consider uniform weighting distribution $\pi = u$ (with $\pi_i = \frac{1}{n}$).

analysis: convex conjugation by Legendre transformation. We refer to [6] for detailed explanations that we concisely summarize here as follows: Any Bregman generator function F admits a *dual* Bregman generator function $G = F^*$ via the Legendre transformation $G(y) = \sup_{x\in\mathcal{X}}\{< y, x > -F(x)\}$. The supremum is reached at the *unique* point where the gradient of $G(x) =< y, x > -F(x)$ vanishes, that is when $y = \nabla F(x)$. Writing $\mathcal{X}'_F$ for the *gradient space* $\{x' = \nabla F(x) | x \in \mathcal{X}\}$, the convex conjugate $G = F^*$ of F is the function $\mathcal{X}'_F \subset \mathbb{R}^d \to \mathbb{R}$ defined by $F^*(x') =< x, x' > -F(x)$. It follows from Legendre transformation that *any* Bregman divergence D_F admits a *dual* Bregman divergence D_{F^*} related to D_F as follows: $D_F(p||q) = F(p) + F^*(\nabla F(q)) - < p, \nabla F(q) >= F(p) + F^*(q') - < p, q' >= D_{F^*}(q'||p')$. Using the convex conjugation twice, we get the following (dual) theorem for the left-type Bregman centroid:

Theorem 2.2 *The left-type sided Bregman centroid c_L^F, defined as the minimizer for the average left divergence $c_L^F = \arg\min_{c\in\mathcal{X}} \mathrm{AVG}_L^F(c||\mathcal{P})$, is the unique point $c_L^F \in \mathcal{X}$ such that $c_L^F = (\nabla F)^{-1}(\bar{p'}) = (\nabla F)^{-1}(\sum_{i=1}^n \nabla F(p_i))$, where $\bar{p'} = c_R^{F^*}(\mathcal{P}_F{}')$ is the center of mass for the gradient point set $\mathcal{P}_F{}' = \{p'_i = \nabla F(p_i) \mid p_i \in \mathcal{P}\}$.*

Proof: Using the dual Bregman divergence D_{F^*} induced by the convex conjugate F^* of F, we observe that the left-type centroid $c_L^F = \arg\min_{c\in\mathcal{X}} \mathrm{AVG}_F(c||\mathcal{P})$ is obtained *equivalently* by minimizing the dual right-type centroid problem on the gradient point set: $\arg\min'_{c'\in\mathcal{X}} \mathrm{AVG}_{F^*}(\mathcal{P}_F{}'||c')$, where we recall that $p' = \nabla F(p)$ and $\mathcal{P}_F{}' = \{\nabla F(p_1), ..., \nabla F(p_n)\}$ denote the gradient point set. Thus the left-type Bregman centroid c_L^F is computed as the *reciprocal gradient* of the center of mass of the gradient point set $c_R^{F^*}(\mathcal{P}_F{}') = \frac{1}{n}\sum_{i=1}^n \nabla F(p_i) : c_L^F = (\nabla F)^{-1}(\sum_{i=1}^n \frac{1}{n}\nabla F(p_i)) = (\nabla F)^{-1}(\bar{p'})$. It follows that the left-type Bregman centroid is *unique*.

Observe that the duality also proves that the information radius for the left-type centroid is the *same* F-Jensen difference (Jensen-Shannon divergence for the convex entropic function F). The information radius equality $\mathrm{AVG}_F(\mathcal{P}||c_R^F) = \mathrm{AVG}_F(c_L^F||\mathcal{P}) = \mathrm{JS}_F(\mathcal{P}) = \frac{1}{n}\sum_{i=1}^n F(p_i) - F(\bar{p}) > 0$ is the F-Jensen-Shannon divergence for the uniform weight distribution.

2.3 Generalized means and barycenters

We show that both sided centroids are generalized means also called quasi-arithmetic or f-means. We first recall the basic definition of generalized means that generalizes the usual arithmetic and geometric means. For a *strictly continuous* and *monotonous* function f, the *generalized mean* [6] of a sequence $\mathcal{V}$ of n real numbers $V = \{v_1, ..., v_n\}$ is defined as $M(\mathcal{V}; f) = f^{-1}(\frac{1}{n}\sum_{i=1}^n f(v_i))$. The generalized means include the Pythagoras' arithmetic, geometric, and harmonic means, obtained respectively for functions $f(x) = x$, $f(x) = \log x$ and $f(x) = \frac{1}{x}$. Note that since f is injective, its reciprocal function f^{-1} is properly defined. Further, since f is monotonous, it is noticed that the generalized mean is necessarily bounded between the *extremal set* elements $\min_i v_i$ and $\max_i v_i$: $\min_i x_i \leq M(\mathcal{V}; f) \leq \max_i x_i$. In fact, finding these minimum and maximum set elements can be treated themselves as a special generalized power mean, another generalized mean for $f(x) = x^p$ in the limit case $p \to \pm\infty$.

These generalized means highlight a bijection: Bregman divergence $D_F \leftrightarrow \nabla F$-means. The one-to-one mapping holds because Bregman generator functions F are strictly convex and differentiable functions chosen up to an affine term [6]. This affine invariant property *transposes* to generalized means as an offset/scaling invariant property: $M(\mathcal{S}; f) = M(\mathcal{S}; af + b)\ \forall a \in \mathbb{R}_*^+$ and $\forall b \in \mathbb{R}$. Although we have considered centroids for simplicity (ie., uniform weight distribution on the input set $\mathcal{P}$), this approach generalizes straightforwardly to *barycenters* defined as solutions of minimum average optimization problems for arbitrary unit weight vector w ($\forall i,\ w_i \geq 0$ with $||w|| = 1$).

3. Symmetrized Bregman centroid

For asymmetric Bregman divergences, the symmetrized Bregman centroid is defined by the following optimization problem $c^F = \arg\min_{c\in\mathcal{X}} \sum_{i=1}^n \frac{D_F(c||p_i) + D_F(p_i||c)}{2} = \arg\min_{c\in\mathcal{X}} \mathrm{AVG}(\mathcal{P}; c)$. We simplify this optimization problem to another *constant-size* system relying only the right-type and left-type sided centroids, c_R^F and c_L^F, respectively. This will prove that the symmetrized Bregman centroid is uniquely defined as the zeroing argument of a sided centroid function by generalizing the approach of Veldhuis [8] that studied the *special case* of the symmetrized discrete Kullback-Leibler divergence, also known as J-divergence.

Theorem 3.1 (Proof in [6]) *The symmetrized Bregman centroid c^F is unique and obtained by minimizing $\min_{q\in\mathcal{X}} D_F(c_R^F||q) + D_F(q||c_L^F)$: $c^F = \arg\min_{q\in\mathcal{X}} D_F(c_R^F||q) + D_F(q||c_L^F)$.*

We now characterize the exact geometric location of the symmetrized Bregman centroid by introducing a

new type of bisector called the mixed-type bisector:

Theorem 3.2 (Proof in [6]) *The symmetrized Bregman centroid c^F is uniquely defined as the minimizer of $D_F(c_R^F||q) + D_F(q||c_L^F)$. It is defined geometrically as $c^F = \Gamma_F(c_R^F, c_L^F) \cap M_F(c_R^F, c_L^F)$, where $\Gamma_F(c_R^F, c_L^F) = \{(\nabla F)^{-1}((1-\lambda)\nabla F(c_R^F) + \lambda \nabla F(c_L^F)) \mid \lambda \in [0,1]\}$ is the geodesic linking c_R^F to c_L^F, and $M_F(c_R^F, c_L^F)$ is the mixed-type Bregman bisector: $M_F(c_R^F, c_L^F) = \{x \in \mathcal{X} \mid D_F(c_R^F||x) = D_F(x||c_L^F)\}$.*

The equation of the mixed-type bisector $M_F(p,q)$ is neither linear in x nor in $x' = \nabla F(x)$ (nor in $\tilde{x} = (x, x')$) because of the term $F(x)$, and can thus only be manipulated implicitly in the remainder: $M_F(p,q) = \{x \in \mathcal{X} \mid F(p) - F(q) - 2F(x) - < p, x' > + < x, x' > + < x, q' > - < q, q' >= 0\}$. The mixed-type bisector is not necessarily connected (eg., extended Kullback-Leibler divergence), and yields the full space $\mathcal{X}$ for symmetric Bregman divergences (ie., generalized quadratic distances).

The exact geometric characterization of the symmetrized Bregman centroid provides us a simple method to approximately converge to c^F: Namely, we perform a dichotomic walk on the geodesic linking the sided centroids c_R^F and c_L^F. This dichotomic search yields a novel efficient algorithm that enables us to solve for *arbitrary* symmetrized Bregman centroids, beyond the former Kullback-Leibler case of Veldhuis [8]: We initially consider $\lambda \in [\lambda_m = 0, \lambda_M = 1]$ and repeat the following steps until $\lambda_M - \lambda_m \leq \epsilon$, for $\epsilon > 0$ a *prescribed* precision threshold:

Geodesic walk. Compute interval midpoint $\lambda_h = \frac{\lambda_m + \lambda_M}{2}$ and corresponding geodesic point
$q_h = (\nabla F)^{-1}((1-\lambda_h)\nabla F(c_R^F) + \lambda_h \nabla F(c_L^F))$,

Mixed-type bisector side. Evaluate the sign of $D_F(c_R^F||q_h) - D_F(q_h||c_L^R)$, and

Dichotomy. Branch on $[\lambda_h, \lambda_M]$ if the sign is negative, or on $[\lambda_m, \lambda_h]$ otherwise.

Note that *any* point on the geodesic (including the midpoint $q_{\frac{1}{2}}$) or on the mixed-type bisector provides an upperbound $\mathrm{AVG}_F(\mathcal{P}; q_h)$ on the minimization task. Although it was noted experimentally by Veldhuis [8] for the Kullback-Leibler divergence that this midpoint provides "experimentally" a good approximation, let us emphasize that is *not true* in general [6].

Theorem 3.3 *The symmetrized Bregman centroid can be approximated within a prescribed precision by a simple dichotomic walk on the geodesic $\Gamma(c_R^F, c_L^F)$ helped by the mixed-type bisector $M_F(c_R^F, c_L^F)$. In general, symmetrized Bregman centroids do not admit closed-form solutions.*

In practice, we can control the stopping criterion ϵ by taking the difference $W_F(q) = D_F(c_R^F||q) - D_F(q||c_L^R)$ between two successive iterations since it monotonically decreases. The number of iterations can also be theoretically upper-bounded as a function of ϵ using the maximum value of the Hessian $h_F = \max_{x \in \Gamma(c_R^F, c_L^F)} ||H_F(x)||^2$ along the geodesic $\Gamma(c_R^F, c_L^F)$, see [6]. In [6],we also consider *entropic centroids* defined with respect to the Kullback-Leibler divergence, and show how to compute the sided and symmetrized entropic centroids for (1) a set of histograms, and (2) for a set of multivariate normals. See `www.sonycsl.co.jp/person/nielsen/BregmanCentroids/`

References

[1] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

[2] Michèle Basseville and Jean-François Cardoso. On entropies, divergences and mean values. In *Proceedings of the IEEE International Symposium on Information Theory* , pp. 330–330, 1995.

[3] Aharon Ben-Tal, Abraham Charnes, and Marc Teboulle. Entropic means. *Journal of Mathematical Analysis and Applications*, pp. 537–551, 1989.

[4] Jacob Burbea and C. Radhakrishna Rao. On the convexity of some divergence measures based on entropy functions. *IEEE Transactions on Information Theory*, 28(3):489–495, 1982.

[5] Minh N. Do and Martin Vetterli. Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance. *IEEE Transactions on Image Processing*, 11(2):146–158, 2002.

[6] Nielsen, F., Nock, R., 2007. On the Centroids of Symmetrized Bregman Divergences, arXiv:0711.3242.

[7] R. Sibson. Information radius. *Probability Theory and Related Fields*, 14(2):149–160, 1969.

[8] R. N. J. Veldhuis. The centroid of the symmetrical Kullback-Leibler distance. *IEEE Signal Processing Letters*, 9(3):96–99, 2002.