On the Centroids of Symmetrized Bregman Divergences —Extended Abstract—

Frank Nielsen
Sony Computer Science Laboratories, Inc
Fundamental Research Laboratory
3-14-13 Higashi Gotanda
141-0022 Shinagawa-Ku
Tokyo, Japan
Frank.Nielsen@acm.org

Richard Nock
Université des Antilles-Guyane
CEREGMIA
Campus de Schoelcher
BP 7209, 97275 Schoelcher
Martinique, France
rnock@martinique.univ-ag.org

21st November 2007

Abstract

In this paper, we generalize the notions of centroids and barycenters to the broad class of information-theoretic distortion measures called Bregman divergences. Bregman divergences are versatile, and unify quadratic geometric distances with various statistical entropic measures. 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. We report on our generic implementation for computing entropic centers of image clusters and entropic centers of multivariate normals, and compare our results with former *ad-hoc* methods.

Keywords: Centroid, Bregman divergence, Legendre duality.

Additional materials including C++ source codes, videos and Java^{TM} applets available at:
http://www.sonycsl.co.jp/person/nielsen/BregmanCentroids/

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 [15], 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 [15]. *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 [18] 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. [4] showed that the k-means algorithm *extends to* and *only* works for a broad family of distortion measures called Bregman divergences [8]. Bregman divergences D_F are parameterized families of distortion measures that are defined by a strictly convex and differentiable generator function $F : \mathcal{X} \to \mathbb{R}^+$ (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., $\nabla F(q) = \left[\frac{\partial F(q)}{\partial x^{(1)}}, ..., \frac{\partial F(q)}{\partial x^{(d)}}\right]$). Further, Teboulle [26] generalized this Bregman k-means algorithm in 2007 by considering both hard and soft *center-based* clustering algorithms designed for both Bregman [8] and Csiszár f-divergences [1, 12]. The fundamental underlying primitive for these *center-based* clustering algorithms is to find the intrinsic *best single representative* of a cluster. As mentionned 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 $F_Q(x) = x^T Q x$, $S_F(p; q) = \frac{D_F(p||q) + D_F(q||p)}{2}$ is *not* a Bregman divergence, see [20]. 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 [25, 19] and COSH centroids [10, 29] (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. These symmetrized centroids play a fundamental role in applications that require to handle symmetric information-theoretic distances.

1.1 Related work, contributions and paper organization

Prior work in the literature is sparse and disparate. We summarize below main references that will be concisely revisited in section 2 under our notational conventions. Ben-Tal et al. [7] 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 [6] compared in the 1-page paper the generalized/entropic mean values for two entropy-based classes of divergences: f-divergences [12] and Jensen-Shannon divergences [13]. The closest recent work to our study is Veldhuis' approximation method [27] for computing the symmetrical Kullback-Leibler centroid.

We summarize our contributions as follows:

- In section 2, we show that the two sided Bregman centroids c_R^F and c_L^F with respect to Bregman divergence D_F are *unique* and easily obtained as *generalized means* for the identity and ∇F functions, respectively. We extend Sibson' s notion of *information radius* [24] for these sided centroids, and show that they are both equal to the F-Jensen difference, a generalized Jensen-Shannon divergence [17] also known as Burbea-Rao divergences [9].

- Section 3 proceeds by first showing how to reduce the symmetrized $\min \mathrm{AVG}_F(\mathcal{P};c)$ optimization problem into a simpler system that depends only on the two sided centroids c_R^F and c_L^F. We then geometrically characterize *exactly* the symmetrized centroid as the intersection point of the geodesic linking the sided centroids with a new type of divergence bisector: the mixed-type bisector. This yields a simple and efficient dichotomic search procedure that provably converges fast to the exact symmetrized Bregman centroid.

- The symmetrized Kullback-Leibler divergence (J-divergence) and symmetrized Itakura-Saito divergence (COSH distance) are often used in sound/image applications, where our fast geodesic dichotomic walk algorithm converging to the unique symmetrized Bregman centroid comes in handy over former complex *adhoc* methods [19, 10, 25, 3, 23]. Section 4 considers *applications* of the generic geodesic-walk algorithm to two cases:

  - The symmetrized Kullback-Leibler for probability mass functions represented as d-dimensional points lying in the $(d-1)$-dimensional simplex S^d. These discrete distributions are handled as multinomials of the exponential families [20] with $d-1$ degrees of freedom. We instantiate the generic geodesic-walk algorithm for that setting, show how it compares favorably with the prior convex optimization work of Veldhuis [27, 3], and validate formally experimental remarks of Veldhuis.

– The symmetrized Kullback-Leibler of multivariate normal distributions. We describe the geodesic-walk for this particular *mixed-type* exponential family of multivariate normals, and explain the Legendre mixed-type vector/matrix dual convex conjugates defining the corresponding Bregman divergences. This yields a simple, fast and elegant geometric method compared to the former overly complex method of Myrvoll and Soong [19] that relies on solving Riccati matrix equations.

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 [2], it was again recently brought up to the attention of the machine learning community by Banerjee et al. [4] who proved that Lloyd's iterative k-means "centroid" clustering algorithm [18] 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 $\mathrm{AVG}_F(\mathcal{P}||q) = \sum_{i=1}^n \frac{1}{n}\left(F(p_i) - F(q) - < p_i - q, \nabla F(q) >\right)$. Subtracting and adding $F(\bar{p})$ to the right-hand side yields

$$\begin{aligned}
\mathrm{AVG}_F(\mathcal{P}, q) &= \left(\sum_{i=1}^n \frac{1}{n} F(p_i) - F(\bar{p})\right) + \left(F(\bar{p}) - F(q) - \sum_{i=1}^n \frac{1}{n} < p_i - q, \nabla F(q) >\right), \\
&= \left(\sum_{i=1}^n \frac{1}{n} F(p_i) - F(\bar{p})\right) + \left(F(\bar{p}) - F(q) - \left\langle \sum_{i=1}^n \frac{1}{n}(p_i - q), \nabla F(q) \right\rangle\right), \\
&= \left(\frac{1}{n}\sum_{i=1}^n F(p_i) - F(\bar{p})\right) + D_F(\bar{p}||q).
\end{aligned}$$

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 [24] 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] [9]. For $F =$

[1] In the paper [9], 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 $\pi_i = \frac{1}{n}$).

$-H = x \log x$ the negative Shannon entropy, J_F is known as the Jensen-Shannon divergence [17]: $\mathrm{JS}(\mathcal{P}) = H(\sum_{i=1}^n p_i) - \sum_{i=1}^n \frac{1}{n} H(p_i)$. 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). The Jensen-Shannon divergence can also be interpreted as the *noisy channel capacity* with two inputs giving output distributions P and Q [11]. Jensen-Shannon divergences are also useful for providing both lower and upper bounds for Bayes probability of error in decision problems [17].

2.2 Dual divergence and left-type centroid

Before characterizing the *left-type* sided Bregman centroid, we recall the fundamental duality of convex analysis: convex conjugation by Legendre transformation. We refer to [20] 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).

Corollary 2.3 *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 centers 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[2] that generalizes the usual arithmetic and

[2]Studied independently in 1930 by Kolmogorov and Nagumo, see [22]. A more detailed account is given in [16], Chapter 3.

geometric means. For a *strictly continuous* and *monotonous* function f, the *generalized mean* [22] 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 $f(x) = \frac{1}{x}$ (see appendix A). 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 [20]. 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 $||w|| = 1$):

Theorem 2.4 *Bregman divergences are in bijection with generalized means. The right-type barycenter $b_R^F(w)$ is independent of F and computed as the weighted arithmetic mean on the point set, a generalized mean for the identity function: $b_R^F(\mathcal{P}; w) = b_R(\mathcal{P}; w) = M(\mathcal{P}; x; w)$ with $M(\mathcal{P}; f; w) = f^{-1}(\sum_{i=1}^{n} w_i f(v_i))$. The left-type Bregman barycenter b_L^F is computed as a generalized mean on the point set for the gradient function: $b_L^F(\mathcal{P}) = M(\mathcal{P}; \nabla F; w)$. The information radius of sided barycenters is $\mathrm{JS}_F(\mathcal{P}; w) = \sum_{i=1}^{d} w_i F(p_i) - F(\sum_{i=1}^{d} w_i p_i)$.*

3 Symmetrized Bregman centroid

3.1 Revisiting the optimization problem

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 [27] that studied the *special case* of the symmetrized discrete Kullback-Leibler divergence, also known as J-divergence.

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

Proof We have previously shown that the right-type average divergence can be rewritten as $\mathrm{AVG}_F(\mathcal{P}||q) = \left(\sum_{i=1}^{n} \frac{1}{n} F(p_i) - F(\bar{p})\right) + D_F(\bar{p}||q)$. Using Legendre transformation, we have similarly $\mathrm{AVG}_F(q||\mathcal{P}) = \mathrm{AVG}_{F^*}(\mathcal{P}_F{}'||q') = (\sum_{i=1}^{n} \frac{1}{n} F^*(p_i') - F^*(\bar{p'})) + D_{F^*}(\bar{p_F'}||q_F')$. But $D_{F^*}(\bar{p_F'}||q_F') = D_{F^{**}}(\nabla F^* \circ \nabla F(q)||\nabla F^*(\sum_{i=1}^{n} \nabla F(p_i))) = D_F(q||c_L^F)$ since $F^{**} = F$, $\nabla F^* = \nabla F^{-1}$ and $\nabla F^* \circ \nabla F(q) = q$ from Legendre duality. Combining these two sum averages, it comes that minimizing $\arg\min_{c\in\mathcal{X}} \frac{1}{2}\left(\mathrm{AVG}_F(\mathcal{P}||q) + \mathrm{AVG}_F(q||\mathcal{P})\right)$ boils down to minimizing $\arg\min_{q\in\mathcal{X}} D_F(c_R^F||q) + D_F(q||c_L^F)$, after removing all terms independent of q. The solution is

unique since the optimization problem $\arg\min_{q\in\mathcal{X}} D_F(c_R^F||q) + D_F(q||c_L^F)$ can be itself rewritten as $\arg\min_{q\in\mathcal{X}} D_{F^*}(\nabla F(q)||\nabla F(c_R^F)) + D_F(q||c_L^F)$, where $\nabla F(q)$ is monotonous and $D_F(\cdot||\cdot)$ and $D_{F^*}(\cdot||\cdot)$ are both convex in the first argument (but not necessarily in the second). Therefore the optimization problem is convex and admits a unique solution.

3.2 Geometric characterization

We now characterize the exact geometric location of the symmetrized Bregman centroid by introducing a new type of bisector[3] called the mixed-type bisector:

Theorem 3.2 *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: $M_F(c_R^F, c_L^F) = \{x \in \mathcal{X} \mid D_F(c_R^F||x) = D_F(x||c_L^F)\}$.*

Proof. First, let us prove by contradiction that q necessarily belongs to the geodesic $\Gamma(c_R^F, c_L^F)$. Assume q does not belong to that geodesic and consider the point $q_\perp$ that is the *Bregman perpendicular projection* of q onto the (convex) geodesic [20]: $q_\perp = \arg\min_{t\in\Gamma(c_R^F,c_L^F)} D_F(t||q)$ as depicted in Figure 1. Using *Bregman Pythagoras' theorem*[4] twice (see [20]), we have: $D_F(c_R^F||q) \geq D_F(c_R||q_\perp) + D_F(q_\perp||q)$ and $D_F(q||c_L^F) \geq D_F(q||q_\perp) + D_F(q_\perp||C_L^F)$. Thus, we get $D_F(c_R^F||q) + D_F(q||c_L^F) \geq D_F(c_R^F||q_\perp) + D_F(q_\perp||c_L^F) + (D_F(q_\perp||q) + D_F(q||q_\perp))$. But since $D_F(q_\perp||q) + D_F(q||q_\perp) > 0$, we reach the contradiction since $D_F(c_R^F||q_\perp) + D_F(q_\perp||c_L^F) < D_F(c_R^F||q) + D_F(q||c_L^F)$. Therefore q necessarily belongs to the geodesic $\Gamma(c_R^F, c_L^F)$. Second, let us show that q necessarily belongs to the mixed-type bisector. Assume it is not the case. Then $D_F(c_R^F||q) \neq D_F(q||c_L^F)$ and suppose without loss of generality that $D_F(c_R^F||q) > D_F(q||c_L^F)$. Let $\Delta = D_F(c_R^F||q) - D_F(q||c_L^F) > 0$ and $l_0 = D_F(q||c_L^F)$ so that $D_F(c_R^F||q) + D_F(q||c_L^F) = 2l_0 + \Delta$. Now move q on the geodesic towards c_R^F by an amount such that $= D_F(q||c_L^F) \leq l_0 + \frac{1}{2}\Delta$. Clearly, $D_F(c_R^F||q) < l_0$ and $D_F(c_R^F||q) + D_F(q||c_L^F) < 2l_0 + \frac{1}{2}\Delta$ contradicting the fact that q was not on the mixed-type bisector.

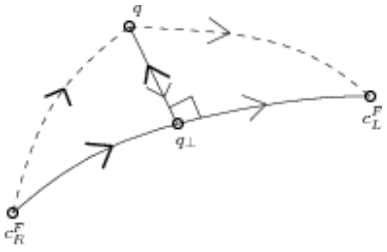


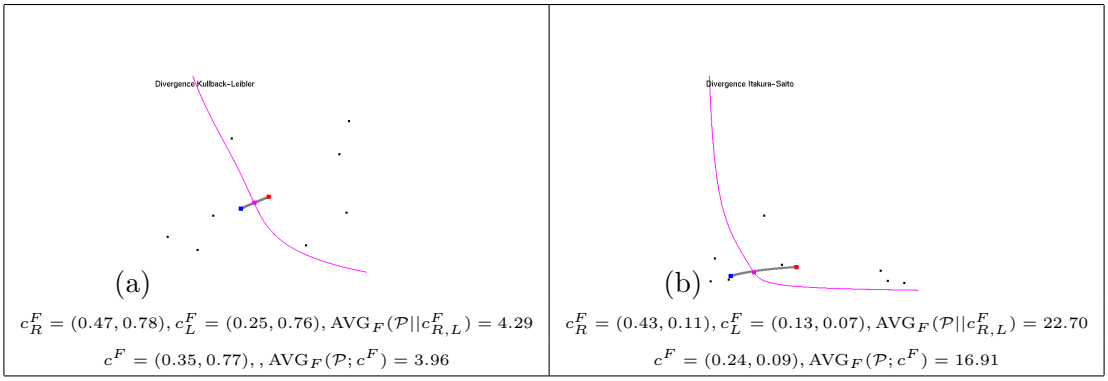
Figure 1: The symmetrized Bregman centroid necessarily lies on the geodesic passing through the two sided centroids c_R^F and c_L^F.

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).

Using the fact that the symmetrized Bregman centroid necessarily lies on the geodesic linking the two sided centroids c_R^F and c_L^F, we get the following corollary:

[3] See [20] for the affine/curved and symmetrized bisectors studied in the context of Bregman Voronoi diagrams.

[4] Bregman Pythagoras' theorem is also called the generalized Pythagoras' theorem, and is stated as follows: $D_F(p||q) \geq D(p||P_\Omega(q)) + D_F(P_\Omega(q)||q)$ where $P_\Omega(q) = \arg\min_{\omega\in\Omega} D_F(\omega||q)$ is the Bregman projection of q onto a convex set Ω, see [4].



(a) $c_R^F = (0.47, 0.78), c_L^F = (0.25, 0.76), \mathrm{AVG}_F(\mathcal{P}||c_{R,L}^F) = 4.29$
$c^F = (0.35, 0.77), , \mathrm{AVG}_F(\mathcal{P}; c^F) = 3.96$

(b) $c_R^F = (0.43, 0.11), c_L^F = (0.13, 0.07), \mathrm{AVG}_F(\mathcal{P}||c_{R,L}^F) = 22.70$
$c^F = (0.24, 0.09), \mathrm{AVG}_F(\mathcal{P}; c^F) = 16.91$

Figure 2: Bregman centroids for (a) the extended Kullback-Leibler and (b) Itakura-Saito divergences on the open square $\mathcal{X} =]0, 1[^2$. Right-sided and left-sided, and symmetrized centroids are displayed respectively as red, blue and purple points. The geodesic linking the right-sided centroid to the left-sided one is shown in grey, and the mixed-type bisector is displayed in purple.

Corollary 3.3 *The symmetrized Bregman divergence minimization problem is both lower and upper bounded as follows:* $\mathrm{JS}_F(\mathcal{P}) \leq \mathrm{AVG}_F(\mathcal{P}; c^F) \leq D_F(c_R^F||c_L^F)$.

Figure 2 displays the mixed-type bisector, and sided and symmetrized Bregman centroids for the extended[5] Kullback-Leibler (eKL) and Itakura-Saito (IS) divergences.

3.3 A simple geodesic-walk dichotomic approximation algorithm

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[6] of Veldhuis [27]: 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 [27] for the Kullback-Leibler divergence that this midpoint provides

[5] We relax the probability distributions to belong to the positive orthant $\mathbb{R}_+^d$ (ie., unnormalized probability mass function) instead of the open simplex $\mathcal{S}^d$.

[6] Veldhuis' method [27] is based on the general purpose Lagrangian multiplier method with a normalization step. It requires to set up one threshold for the outer loop and two prescribed thresholds for the inner loops. For example, Aradilla et al. [3] set the number of steps of the outer loop and inner loops to ten and five iterations each, respectively. Appendix B provides a synopsis of Veldhuis' method.

"experimentally" a good approximation, let us emphasize that is *not true* in general, as depicted in Figure 2(b) for the Itakura-Saito divergence.

Theorem 3.4 *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 $\Gamma(c_R^F, c_L^F)$ by mimicking the analysis in [21] (See Lemma 3 of [21]).

4 Applications of the dichotomic geodesic-walk algorithm

4.1 Revisiting the centroid of symmetrized Kullback-Leibler divergence

Consider a random variable Q on d events $\Omega = \{\Omega_1, ..., \Omega_d\}$, called the sample space. Its associated discrete distribution q (with $\Pr(Q = \Omega_i) = q^{(i)}$) belongs to the topologically *open* $(d-1)$-dimensional probability simplex $\mathcal{S}^d$ of $\mathbb{R}_+^d$: $\sum_{i=1}^d q^{(i)} = 1$ and $\forall i \in \{1, ..., d\}$ $q_i > 0$. Distributions q arise often in practice from image intensity histograms[7]. To measure the distance between two discrete distributions p and q, we use the Kullback-Leibler divergence also known as relative entropy or discrimination information: $\mathrm{KL}(p||q) = \sum_{i=1}^d p^{(i)} \log \frac{p^{(i)}}{q^{(i)}}$. Note that this information measure is unbounded whenever there exists $q^{(i)} = 0$ for a non-zero $q^{(i)} > 0$. But since we assumed that both p and q belongs to the open probability simplex $\mathcal{S}^d$, this case does not occur in our setting: $0 \leq \mathrm{KL}(p||q) < \infty$ with left-hand side equality if and only if $p = q$. The symmetrized KL divergence $\frac{1}{2}(\mathrm{KL}(p||q) + \mathrm{KL}(q||p))$ is also called J-divergence or SKL divergence, for short.

The random variable Q can also be interpreted as a regular exponential family member [20] in statistics of order $d-1$, generalizing the Bernoulli random variable. Namely, Q is a *multinomial* random variable indexed by a $(d-1)$-dimensional *parameter vector* θ_q. These multinomial distributions belong to the broad class of exponential families [20] in statistics for which have the important property that $\mathrm{KL}(p(\theta_p)||q(\theta_q)) = D_F(\theta_q||\theta_p)$, see [20]. That is, this property allows us to bypass the fastidious integral computations of Kullback-Leibler divergences and replace it by a simple gradient derivatives for probability distributions belonging to the *same* exponential families. From the canonical decomposition $\exp(< \theta, t(x) > -F(\theta) + C(x))$ of exponential families [20], it comes out that the natural parameters associated with the sufficient statistics $t(x)$ are $\theta^{(i)} = \log \frac{q^{(i)}}{q^{(d)}} = \log \frac{q^{(i)}}{1 - \sum_{j=1}^{d-1} q^{(j)}}$ since $q^{(d)} = 1 - \sum_{j=1}^{d-1} q^{(j)}$. The natural parameter space is the topologically open $\mathbb{R}^{d-1}$. The log normalizer is $F(\theta) = \log(1 + \sum_{i=1}^{d-1} \exp \theta^{(i)})$, called the multivariate *logistic entropy*. It follows that the gradient is $\nabla F(\theta) = \eta = (\eta_i)_i$ with $\eta_i = \frac{\exp \theta^{(i)}}{1 + \sum_{j=1}^{d-1} \exp \theta^{(j)}}$ and yields the *dual parameterization* of the expectation parameters: $\eta = \nabla_\theta F(\theta)$. The expectation parameters play an important role in practice for infering the distributions from identically and independently distributed observations

[7] To ensure to all bins of the histograms are non-void, we add a small quantity ϵ to each bin, and normalize to unit. This is the same as considering the random variable $Q + \epsilon U$ where U is a unit random variable.

$x_1, ..., x_n$. Indeed, the maximum likelihood estimator of exponential families is simply given by the center of mass of the sufficient statistics computed on the observations: $\hat{\eta} = \frac{1}{n}\sum_{i=1}^{n} t(x_i)$, see [5]. Observe in this case that the log normalizer function is not separable ($F(x) \neq \sum_{i=1}^{d-1} f_i(x^{(i)})$). The function F and $F^* = \int \nabla^{-1} F$ are convex conjugates obtained by the Legendre transformation that maps both domains and functions $(\mathcal{X}_F, F) \longleftrightarrow (\mathcal{X}_{F*}, F^*)$. We get the inverse $\nabla^{-1} F = (\nabla F)^{-1}$ of the gradient ∇F as $\nabla^{-1} F(\eta) = \left(\log \frac{\eta^{(i)}}{1-\sum_{j=1}^{d-1} \eta^{(j)}} \right)_i = \theta$. Thus it comes that the Legendre convex conjugate is $F^*(\eta) = \left(\sum_{i=1}^{d-1} \eta^{(i)} \log \eta^{(i)} \right) + (1 - \sum_{i=1}^{d-1} \eta^{(i)}) \log(1 - \sum_{i=1}^{d-1} \eta^{(i)})$, the *d-ary entropy*. Observe that for $d = 2$, this yields the usual bit entropy[8] function $F^*(\eta) = \eta \log \eta + (1-\eta) \log(1-\eta)$.

To convert back from the multinomial $(d-1)$-order natural parameters θ to discrete d-bin normalized probability mass functions (eg., histograms) $\Lambda \in \mathcal{S}^d$, we use the following mapping: $q^{(d)} = \frac{1}{1+\sum_{j=1}^{d-1}(1+\exp \theta^{(j)})}$ and $q^{(i)} = \frac{\exp \theta^{(i)}}{1+\sum_{j=1}^{d-1}(1+\exp \theta^{(j)})}$ for all $i \in \{1, ..., d-1\}$. This gives a *valid* (*ie.*, normalized) distribution $q \in \mathcal{S}^d$ for *any* $\theta \in \mathbb{R}^{d-1}$. Note that the coefficients in θ may be either positive or negative depending on the ratio of the probability of the ith event with the last one, $q^{(d)}$.

As mentioned above, it turns out that the Kullback-Leibler measure can be computed from the Bregman divergence associated to the multinomial by *swapping* arguments: $\mathrm{KL}(p||q) = D_F(\theta_q||\theta_p)$, where the Bregman divergence $D_F(\theta_q||\theta_p) = F(\theta_q) - F(\theta_p) - < \theta_q - \theta_p, \nabla F(\theta_p) >$ is defined for the strictly convex ($\nabla^2 F > 0$) and diffentiable log normalizer $F(\theta) = \log(1 + \sum_{i=1}^{d-1} \exp \theta^{(i)})$. We implemented the geodesic-walk approximation algorithm for that context, and observed in practice that the SKL centroid deviates much (20% or more in information radius) from the "middle" point of the geodesic ($\lambda = \frac{1}{2}$), thus reflecting the asymmetry of the underlying space. Further, note that our geodesic-walk algorithm *proves* the *empirical remark* of Veldhuis [27] that "... the assumption that the SKL centroid is a linear combination of the arithmetic and normalized geometric mean must be rejected." Appendix B displays side by side Veldhuis' and the geodesic-walk methods for reference, and appendix C report on the sided and symmetrized Bregman centroids of two probability mass functions obtained from intensity histograms of `apple` images. Observe that the symmetrized centroid distribution *may be above* both source distributions, but this is *never* the case in the natural parameter domain since the two sided centroids are generalized means, and that the symmetrized centroid belongs to the geodesic linking these two centroids (ie., a barycenter mean of the two sided centroids).

Computing the centroid of a set of image histograms, a center robust to outliers, allows one to design novel applications in information retrieval and image processing. For example, we can perform *simultaneous contrast* image enhancement by first computing the histogram centroid of a *group* of pictures, and then performing histogram normalization to that same reference histogram.

4.2 Entropic means of multivariate normal distributions

The probability density function of an arbitary d-variate normal $\mathcal{N}(\mu, \Sigma)$ with mean μ and variance-covariance matrix Σ is given by $\Pr(X = x) = p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{\det \Sigma}} \exp\left(-\frac{(x-\mu)^T \Sigma^{-1} (x-\mu)}{2} \right)$. It is certainly the engineer's favorite family of distributions that nevertheless becomes intricate

[8] This generalizes the 1D case of Kullback-Leibler's Bernoulli divergence: $F(x) = \log(1 + \exp x)$ is the *logistic entropy*, $F'(x) = \frac{\exp x}{1+\exp x}$ and $F'^{-1} = \log \frac{x}{1-x}$, and $F^*(x) = x \log x + (1-x) \log(1-x)$, is the dual *bit entropy*.

to use as dimension goes beyond 3D. The density function can be rewritten into the canonical decomposition to yield an exponential family of order $D = \frac{d(d+3)}{2}$ (the mean vector and the positive definite matrix Σ^{-1} accounting respectively for d and $\frac{d(d+1)}{2}$ parameters). The sufficient statistics is *stacked* onto a two-part D-dimensional vector $\tilde{x} = (x, -\frac{1}{2}xx^T)$ associated with the natural parameter $\tilde{\Theta} = (\theta, \Theta) = (\Sigma^{-1}\mu, \frac{1}{2}\Sigma^{-1})$. Accordingly, the source parameter are denoted by $\tilde{\Lambda} = (\mu, \Sigma)$. The log normalizer specifying the exponential family is $F(\tilde{\Theta}) = \frac{1}{4}\mathrm{Tr}(\Theta^{-1}\theta\theta^T) - \frac{1}{2}\log\det\Theta + \frac{d}{2}\log\pi$ (see [30, 2]). To compute the Kullback-Leibler divergence of two normal distributions $N_p = \mathcal{N}(\mu_p, \Sigma_p)$ and $N_q = \mathcal{N}(\mu_q, \Sigma_q)$, we use the Bregman divergence as follows: $\mathrm{KL}(N_p||N_q) = D_F(\tilde{\Theta}_q||\tilde{\Theta}_p) = F(\tilde{\Theta}_q) - F(\tilde{\Theta}_p) - < (\tilde{\Theta}_q - \tilde{\Theta}_p), \nabla F(\tilde{\Theta}_p) >$. The inner product $< \tilde{\Theta}_p, \tilde{\Theta}_q >$ is a *composite* inner product obtained as the sum of inner products of vectors and matrices: $< \tilde{\Theta}_p, \tilde{\Theta}_q > = < \Theta_p, \Theta_q > + < \theta_p, \theta_q >$. For matrices, the inner product $< \Theta_p, \Theta_q >$ is defined by the trace of the matrix product $\Theta_p\Theta_q^T$: $< \Theta_p, \Theta_q > = \mathrm{Tr}(\Theta_p\Theta_q^T)$. In this setting, however, computing the gradient, inverse gradient and finding the Legendre convex conjugates are quite involved operations. Yoshizawa and Tanabe [30] investigated in a unifying framework the differential geometries of the families of probability distributions of *arbitrary* multivariate normals from both the viewpoint of Riemannian geometry relying on the corresponding Fisher information metric, and from the viewpoint of Kullback-Leibler information, yielding the classic torsion-free flat shape geometry with dual affine connections [2]. Yoshizawa and Tanabe [30] carried out computations that yield the dual natural/expectation coordinate systems arising from the canonical decompotion of the density function $p(x; \mu, \Sigma)$:

$$\tilde{H} = \begin{pmatrix} \eta = \mu \\ H = -(\Sigma + \mu\mu^T) \end{pmatrix} \Longleftrightarrow \tilde{\Lambda} = \begin{pmatrix} \lambda = \mu \\ \Lambda = \Sigma \end{pmatrix} \Longleftrightarrow \tilde{\Theta} = \begin{pmatrix} \theta = \Sigma^{-1}\mu \\ \Theta = \frac{1}{2}\Sigma^{-1} \end{pmatrix}$$

The strictly convex and differentiable dual Bregman generator functions (ie., potential functions in information geometry) are $F(\tilde{\Theta}) = \frac{1}{4}\mathrm{Tr}(\Theta^{-1}\theta\theta^T) - \frac{1}{2}\log\det\Theta + \frac{d}{2}\log\pi$, and $F^*(\tilde{H}) = -\frac{1}{2}\log(1 + \eta^T H^{-1}\eta) - \frac{1}{2}\log\det(-H) - \frac{d}{2}\log(2\pi e)$ defined respectively both on the topologically open space $\mathbb{R}^d \times \mathrm{Cone}_d^-$. Note that removing constant terms does not change the Bregman divergences. The $\tilde{H} \Leftrightarrow \tilde{\Theta}$ coordinate transformations obtained from the Legendre transformation (with $(\nabla F)^{-1} = \nabla F^*$) are given by $\tilde{H} = \nabla_{\tilde{\Theta}} F(\tilde{\Theta}) = \begin{pmatrix} \nabla_{\tilde{\Theta}} F(\theta) \\ \nabla_{\tilde{\Theta}} F(\Theta) \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\Theta^{-1}\theta \\ -\frac{1}{2}\Theta^{-1} - \frac{1}{4}(\Theta^{-1}\theta)(\Theta^{-1}\theta)^T \end{pmatrix} = \begin{pmatrix} \mu \\ -(\Sigma + \mu\mu^T) \end{pmatrix}$ and $\tilde{\Theta} = \nabla_{\tilde{H}} F^*(\tilde{H}) = \begin{pmatrix} \nabla_{\tilde{H}} F^*(\eta) \\ \nabla_{\tilde{H}} F^*(H) \end{pmatrix} = \begin{pmatrix} -(H + \eta\eta^T)^{-1}\eta \\ -\frac{1}{2}(H + \eta\eta^T)^{-1} \end{pmatrix} = \begin{pmatrix} \Sigma^{-1}\mu \\ \frac{1}{2}\Sigma^{-1} \end{pmatrix}$. These formula simplifies significantly when we restrict ourselves to diagonal-only variance-covariance matrices Σ_i, spherical normals $\Sigma_i = \sigma_i I$, or univariate normals $\mathcal{N}(\mu_i, \sigma_i)$.

Computing the symmetrized Kullback-Leibler centroid of a set of normals (Gaussians) is an essential operation for clustering sets of multivariate normal distributions using center-based k-means algorithm [14, 26]. Myrvoll and Soong [19] described the use of multivariate normal clustering in automatic speech recognition. They derived a numerical local algorithm for computing the multivariate normal centroid by solving iteratively Riccati matrix equations, initializing the solution to the so-called "expectation centroid" [23]. Their method is a complex and costly since it also involves solving for eigensystems. In comparison, our geometric geodesic dichotomic walk procedure for computing the entropic centroid, a Bregman symmetrized centroid, yields an extremely fast and simple algorithm with *guaranteed* performance.

Acknowledgements.

We gratefully thank Professors Lev M. Bregman [8], Marc Teboulle [7], and Baba Vemuri [28] for email correspondences and sending us printed copies of seminal papers. We also thank Guillaume Aradilla for sharing with us his experience [3] concerning Veldhuis's algorithm [27].

References

[1] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society*, 28(Series B):131–142, 1966.

[2] S.-I. Amari and N. Nagaoka. *Methods of Information Geometry*. Oxford University Press, 2000. ISBN-10:0821805312.

[3] G. Aradilla, J. Vepa, and H. Bourlard. An acoustic model based on Kullback-Leibler divergence for posterior features. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 657–660, 2007.

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

[5] Ole E. Barndorff-Nielsen. *Parametric statistical models and likelihood*, volume 50 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1988.

[6] Michèle Basseville and Jean-François Cardoso. On entropies, divergences and mean values. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, pages 330–330, Whistler, Ca., September 1995.

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

[8] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217, 1967.

[9] 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.

[10] Beth A. Carlson and Mark A. Clements. A computationally compact divergence measure for speech processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 13(12):1255–1260, 1991.

[11] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2006. (Wiley Series in Telecommunications and Signal Processing).

[12] Imre Csiszár. Information type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.

[13] Imre Csiszár. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *The Annals of Statistics*, 19(4):2032–2066, 1991.

[14] Jason V. Davis and Inderjit S. Dhillon. Differential entropic clustering of multivariate Gaussians. In Bernhard Scholkopf, John Platt, and Thomas Hoffman, editors, *Neural Information Processing Systems (NIPS)*, pages 337–344. MIT Press, 2006.

[15] 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.

[16] G. H. Hardy, J. E. Littlewood, and G. Polya. *Inequalities*. Cambridge University Press, Cambridge, England, 1967.

[17] J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory (TIT)*, 37(1):145–151, 1991.

[18] Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–136, 1982. First published in 1957 in a Technical Note of Bell Laboratories.

[19] Tor André Myrvoll and Frank K. Soong. On divergence-based clustering of normal distributions and its application to HMM adaptation. In *Proceedings of EuroSpeech 2003*, pages 1517–1520, Geneva, Switzerland.

[20] Frank Nielsen, Jean-Daniel Boissonnat, and Richard Nock. Bregman Voronoi diagrams: Properties, algorithms and applications, September 2007. Extended abstract appeared in ACM-SIAM SODA 2007. INRIA Technical Report RR-6154.

[21] Richard Nock and Frank Nielsen. Fitting the smallest enclosing Bregman ball. In *16th European Conference on Machine Learning (ECML)*, volume Volume 3720/2005, pages 649–656, 2005. Lecture Notes in Computer Science.

[22] E. Porcu, J. Mateu, and G. Christakos. Quasi-arithmetic means of covariance functions with potential applications to space-time data, 2006. arXiv:math/0611275.

[23] K. Shinoda and C. H. Lee. A structural Bayes approach to speaker adaptation. *Speech and Audio Processing, IEEE Transactions on*, 9(3):276–287, 2001.

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

[25] Y. Stylianou and A. K. Syrdal. Perceptual and objective detection of discontinuities in concatenative speech synthesis. In *Proceedings IEEE Acoustics, Speech, and Signal Processing (ICASSP)*, pages 837–840, Washington, DC, USA, 2001. IEEE Computer Society.

[26] Marc Teboulle. A unified continuous optimization framework for center-based clustering methods. *Journal of Machine Learning Research*, 8:65–102, 2007.

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

[28] Zhizhou Wang and Baba C. Vemuri. DTI segmentation using an information theoretic tensor dissimilarity measure. *IEEE Transactions on Medical Imaging*, 24(10):1267–1277, 2005.

[29] Bo Wei and Jerry D. Gibson. Comparison of distance measures in discrete spectral modeling. In *Proc. 9th DSP Workshop & 1st Signal Processing Education Workshop*, 2000.

[30] S. Yoshizawa and K. Tanabe. Dual differential geometry associated with Kullback-Leibler information on the Gaussian distributions and its 2-parameter deformations. *SUT Journal of Mathematics*, 35(1):113–137, 1999.

A Dominance relationships of sided centroid coordinates

The table below illustrates the bijection between Bregman divergences and generalized f-means for the Pythagoras' means (ie., extend to separable Bregman divergences):

Bregman divergence D_F	F	$\longleftrightarrow$	$f = F'$	$f^{-1} = (F')^{-1}$	f-mean (Generalized means)
Squared Euclidean distance (half squared loss)	$\frac{1}{2}x^2$	$\longleftrightarrow$	x	x	Arithmetic mean $\sum_{j=1}^n \frac{1}{n} x_j$
Kullback-Leibler divergence (Ext. neg. Shannon entropy)	$x \log x - x$	$\longleftrightarrow$	$\log x$	$\exp x$	Geometric mean $(\prod_{j=1}^n x_j)^{\frac{1}{n}}$
Itakura-Saito divergence (Burg entropy)	$-\log x$	$\longleftrightarrow$	$-\frac{1}{x}$	$-\frac{1}{x}$	Harmonic mean $\frac{n}{\sum_{j=1}^n \frac{1}{x_j}}$

We give a characterization of the coordinates $c_R^{F(i)}$ of the right-type average centroid (center of mass) with respect to those of the left-type average centroid, the $c_L^{F(i)}$ coordinates.

Corollary

Provided that ∇F is convex (e.g., Kullback-Leibler divergence), we have $c_R^{F(i)} \geq c_L^{F(i)}$ for all $i \in \{1, ..., d\}$. Similarly, for concave gradient function (e.g., exponential loss), we have $c_R^{F(i)} \leq c_L^{F(i)}$ for all $i \in \{1, ..., d\}$.

Proof Assume ∇F is convex and apply Jensen's inequality to $\frac{1}{n}\sum_{i=1}^n \nabla F(p_i)$. Consider for simplicity without loss of generality 1D functions. We have $\frac{1}{n}\sum_{i=1}^n \nabla F(p_i) \leq \nabla F(\frac{1}{n}\sum_{i=1}^n p_i)$. Because $(\nabla F)^{-1}$ is a monotonous function, we get $c_L^F = (\nabla F)^{-1}(\frac{1}{n}\sum_{i=1}^n \nabla F(p_i)) \leq (\nabla F)^{-1}(\nabla F(\frac{1}{n}\sum_{i=1}^n p_i)) = \frac{1}{n}\sum_{i=1}^n p_i = c_R^F$. Thus we conclude that $c_R^{F(i)} \geq c_L^{F(i)}\ \forall i \in \{1, ..., d\}$ for convex ∇F (proof performed coordinatewise). For concave ∇F functions (i.e., dual divergences of ∇F-convex primal divergences), we simply reverse the inequality (e.g., the exponential loss dual of the Kullback-Leibler divergence).

Note that Bregman divergences D_F may neither have their gradient ∇F convex nor concave. The bit entropy $F(x) = x \log x + (1-x)\log(1-x)$ yielding the logistic loss D_F is such an example. In that case, we cannot *a priori* order the coordinates of c_R^F and c_L^F.

This dominance relationship can be verified for the plot in natural parameter space of Appendix C.

B Synopsis of Veldhuis' and the generic geodesic-walk methods

The table below provides a side-by-side comparison of Veldhuis' J-divergence centroid convex programming method [27] with our generic symmetrized Bregman centroid (entropic means) geodesic-walk instantiated for the Kullback-Leibler divergence.

Veldhuis' algorithm	Geodesic-walk algorithm

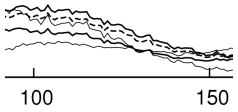
INPUT:
n discrete distributions $q_1, ..., q_n$ of $\mathcal{S}^d$ with
$\forall\, i \in \{1, ..., n\}\ q_i = (q_i^{(1)}, ..., q_i^{(d)})$.
 	
INITIALIZATION
 	
Arithmetic mean:
$\forall k\ \bar{q}^{(k)} = \frac{1}{n} \sum_{i=1}^n q_i^{(k)}$
Geometric normalized mean:
$\forall k\ \breve{q}^{(k)} = \frac{\bar{q}_i^{(k)}}{\sum_{i=1}^d \bar{q}_i}$ with $\forall k\ \tilde{q}^{(k)} = \left(\prod_{i=1}^n q_i^{(k)}\right)^{\frac{1}{n}}$
$\alpha = -1$
 	
MAIN LOOP:
For 1 to 10
$\forall k\ y^{(k)} = \frac{\bar{q}^{(k)}}{\breve{q}^{(k)} \exp \alpha}$
$\forall k\ x^{(k)} = 1$
 	
INNER LOOP 1:
For 1 to 5
$\forall k\ x^{(k)} \leftarrow x^{(k)} - \frac{x^{(k)} \log x^{(k)} - y^{(k)}}{\log x^{(k)} + 1}$
 	
INNER LOOP 2:
For 1 to 5
$\alpha \leftarrow \alpha - \frac{(\sum_{i=1}^d x^{(k)} \breve{q}^{(k)} \exp \alpha) - 1}{\sum_{i=1}^d x^{(k)} \breve{q}^{(k)} \exp \alpha}$
 	
CENTROID:
$\forall k\ c^{(k)} = x^{(k)} \breve{q}^{(k)} \exp \alpha$
	INPUT:
	n discrete distributions $q_1, ..., q_n$ of $\mathcal{S}^d$ with
	$\forall i \in \{1, ..., n\}\ q_i = (q_i^{(1)}, ..., q_i^{(d)})$
	
	CONVERSION:
	Probability mass function $\rightarrow$ multinomial
	$\forall i\ \forall k\ \theta_i^{(k)} = \log \frac{q_i^{(k)}}{1 - \sum_{i=1}^{d-1} q_i^{(j)}}$
	$F(\theta) = \log(1 + \sum_{j=1}^{d-1} \exp \theta^{(j)})$
	$\nabla F(\theta) = \left(\frac{\exp \theta^{(i)}}{1 + \sum_{j=1}^{d-1} \exp \theta^{(j)}} \right)_{i \in \{1, ..., d-1\}}$
	$(\nabla F)^{-1(\eta)} = \left(\log \frac{\eta^{(i)}}{1 - \sum_{i=1}^{d-1} \eta^{(j)}} \right)_{i \in \{1, ..., d-1\}}$
	
	INITIALIZATION:
	Arithmetic mean: $\theta_R^F = \frac{1}{n} \sum_{i=1}^n \theta_i$
	∇F-mean: $\theta_L^F = \nabla F^{-1}(\frac{1}{n} \sum_{i=1}^n \nabla F(\theta_i))$
	$\lambda_m = 0, \lambda_M = 1$
	
	GEODESIC DICHOTOMIC WALK:
	While $\lambda_M - \lambda_m >$ precision do
	$\lambda = \frac{\lambda_m + \lambda_M}{2}$
	$\theta = (\nabla F)^{-1}((1 - \lambda) \nabla F(c_R^F) + \lambda \nabla F(c_L^F))$
	if $D_F(c_R^F \| \theta) > D_F(\theta \| c_L^F)$ then
	$\lambda_M = \lambda$
	else
	$\lambda_m = \lambda$
	
	CONVERSION:
	Multinomial $\rightarrow$ Probability mass function
	$\forall i\ q_i^{(d)} = \frac{1}{1 + \sum_{j=1}^{d-1} (1 + \exp \theta_i^{(j)})}$
	$\forall i\ \forall k\ q_i^{(k)} = \frac{\exp \theta_i^{(k)}}{1 + \sum_{j=1}^{d-1} (1 + \exp \theta_i^{(j)})}$

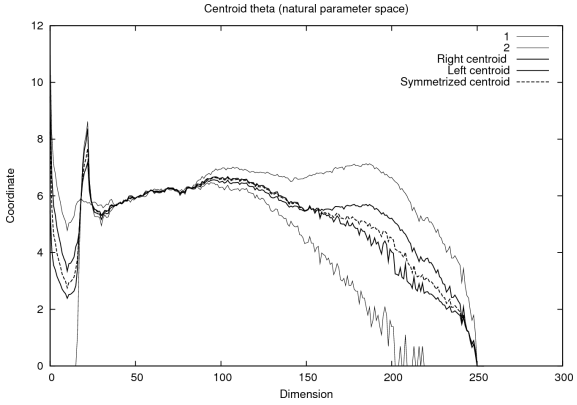
Both C++ source codes with cross-check validations are available at

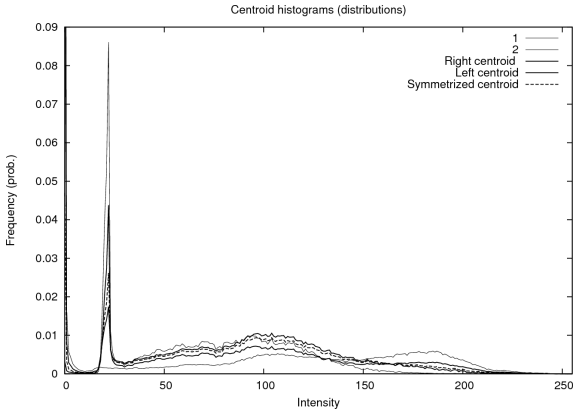
`http://www.sonycsl.co.jp/person/nielsen/BregmanCentroids/`

C Image histogram centroids with respect to the relative entropy

The plots below show the Kullback-Leibler sided and symmetrized centroids on two distributions taken as the intensity histograms of the `apple` images shown below. Observe that the symmetrized centroid distribution is *above* both source distributions for intensity range $[100 - 145]$, but this is *never* the case in the natural parameter space due to the property of generalized means.



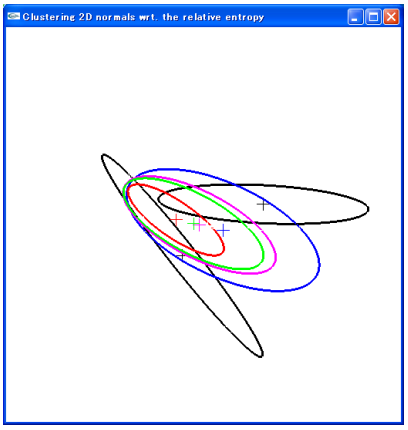




D Entropic sided and symmetrized centroids of bivariate normal distributions

We report on our implementation for multivariate normal distributions below. Observe that the right-type Kullback-Leibler centroid is a left-type Bregman centroid for the log normalizer of the exponential family. Our method allowed us to verify that the simple generalized ∇F-mean formula $c_L^F(\mathcal{P}) = (\nabla F)^{-1}(\sum_{i=1}^n \frac{1}{n}\nabla F(p_i))$ coincides with that of the NIPS*06 paper [14]. Furthermore, we would like to stress out that our method extends to *arbitrary* entropic centroids of members of the same exponential family.

The figure below plots the entropic right- and left-sided and the symmetrized centroids in red, blue and green respectively for a set that consists of two bivariate normals ($D = \frac{d(d+3)}{2} = 5$). The geodesic midpoint interpolant (obtained for $\lambda = \frac{1}{2}$) is very close to the symmetrized centroid, and shown in magenta.



$m_0 = (0.34029138065736869, 0.26130947813348798)$,
$S_0 = \begin{bmatrix} 0.43668091668767117 & -0.42663095837289156 \\ -0.42663095837289161 & 0.63899446830332574) \end{bmatrix}$
$m_1 = (0.95591075380718404, 0.6544489172032838)$,
$S_1 = \begin{bmatrix} 0.79712692342719804 & -0.033060250957646142 \\ -0.033060250957646142 & 0.14609813043797121 \end{bmatrix}$

$m_R = (0.29050997932657774, 0.53527112890397821)$,
$S_R = \begin{bmatrix} 0.33728018979019664 & -0.13844874409795613 \\ -0.13844874409795613 & 0.2321103610207193 \end{bmatrix}$
$m_L^F = (0.64810106723227623, 0.45787919766838603)$,
$S_L^F = \begin{bmatrix} 0.71165072320677747 & -0.16933954090511438 \\ -0.16933954090511441 & 0.43118595400867693 \end{bmatrix}$
$\mathbf{m}^F = (0.42475123207621085, 0.5062178606510539)$,
$\mathbf{S}^F = \begin{bmatrix} 0.50780328118070528 & -0.15653432651371618 \\ -0.15653432651371618 & 0.30824860232457035 \end{bmatrix}$
$m_{\frac{1}{2}} = (0.46930552327942698, 0.49657516328618234)$,
$S_{\frac{1}{2}} = \begin{bmatrix} 0.55643330303588234 & -0.16081280872294987 \\ -0.1608128087229499 & 0.33314553526979185 \end{bmatrix}$.

Information radius:

- right, left: 0.83419372149741644
- **symmetrized**: 0.64099815325721565
- geodesic $\lambda = \frac{1}{2}$: 0.6525069280087431

We give other pictorial results below for $n = 2$ and $n = 10$ bivariate normals, respectively.