Sided and Symmetrized Bregman Centroids

Frank Nielsen, *Member, IEEE*, and Richard Nock, *Nonmember, IEEE*

***Abstract*—We generalize the notions of centroids (and barycenters) to the broad class of information-theoretic distortion measures called Bregman divergences. Bregman divergences form a rich and versatile family of distances that unifies quadratic Euclidean distances with various well-known statistical entropic measures. Since besides the squared Euclidean distance, Bregman divergences are asymmetric, we consider the *left-sided* and *right-sided* centroids and the *symmetrized* centroids as minimizers of average Bregman distortions. We prove that all three centroids are unique and give closed-form solutions for the sided centroids that are generalized means. Furthermore, we design a provably fast and efficient arbitrary close approximation algorithm for the symmetrized centroid based on its *exact* geometric characterization. The geometric approximation algorithm requires only to walk on a geodesic linking the two left/right sided centroids. We report on our implementation for computing entropic centers of image histogram clusters and entropic centers of multivariate normal distributions that are useful operations for processing multimedia information and retrieval. These experiments illustrate that our generic methods compare favorably with former limited *ad-hoc* methods.**

***Index Terms*—Information geometry, centroid, Bregman information, information radius, Legendre duality, Kullback-Leibler divergence, Bregman divergence, Bregman power divergence, Burbea-Rao divergence, Csiszár f-divergences.**

I. Introduction and motivations

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 [13], database images are processed offline during a *preprocessing* step by various feature extractors computing image characteristics such as color histograms or points of interest. These features are aggregated into *signature* vectors, say $\{p_i\}_i$, that represent handles to images. At query time, whenever an on-line query image is given, the system first computes its signature, and then search for the first, say h, best matches in the signature space. This image retrieval task requires to define an appropriate *similarity* (or *dissimilarity*) measure between any pair (p_i, p_j) of signatures. Designing an appropriate distance is tricky since the signature space is often heterogeneous (ie., cartesian product of feature spaces combining for examples various histograms with other geometric features) and the usual Euclidean distance or L_p-norms do not always make sense. For example, it has been shown better to use the information-theoretic relative entropy, known as the Kullback-Leibler divergence (or I-divergence for short), to measure the *oriented distance* between image histograms [13]. The definition of the Kullback-Leibler divergence [14] for two continuous probability densities[1] $p(x)$ and $q(x)$ is as follows:

$$\mathrm{KL}(p(x)||q(x)) = \int_x p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x.$$

The Kullback-Leibler divergence of statistical distributions $p(x)$ and $q(x)$ is called the *relative entropy* since it is equal to the cross-entropy of $p(x)$ and $q(x)$ minus the entropy $H(p(x)) = \int_x p(x) \log \frac{1}{p(x)} \mathrm{d}x$ of $p(x)$:

$$\mathrm{KL}(p(x)||q(x)) = H^{\times}(p(x)||q(x)) - H(p(x)) \geq 0$$

with the cross-entropy:

$$H^{\times}(p(x)||q(x)) = \int_x p(x) \log \frac{1}{q(x)} \mathrm{d}x$$

The Kullback-Leibler divergence represents the average loss (measured in bits if the logarithm's basis is 2) of using another code to encode a random variable X. The relative entropy can also be interpreted as the information gain achieved about X if p can be used instead of q (see [14] for various interpretations in information theory). For discrete random variables, the statistical Kullback-Leibler divergence on two real-valued d-dimensional probability vectors p and q encoding the histogram ditributions is defined [6] as:

$$\mathrm{KL}(p||q) = \sum_{i=1}^{d} p^{(i)} \log \frac{p^{(i)}}{q^{(i)}},$$

where $p^{(i)}$ and $q^{(i)}$ denote the d coordinates of probability vectors p and q, respectively (with both p, q belonging to the d-dimensional probability simplex $\mathbb{S}_d = \{(x^{(1)}, ..., x^{(d)}) \mid \sum_{i=1}^{d} x_i = 1 \text{ and } \forall i \; x_i > 0\}$, an open convex set). The $||$ in the notation $\mathrm{KL}(p||q)$ emphasizes that the distortion measure is not symmetric (ie., oriented distance), since we have $\mathrm{KL}(p||q) \neq \mathrm{KL}(q||p)$.

Notations: Throughout the paper, let $p_j, x_j, c_j, ...$ denote d-dimensional real-valued vectors of $\mathbb{R}^d$, and let $p_j^{(i)}, x_j^{(i)}, c_j^{(i)}, ..., 1 \leq i \leq d$ denote their coordinates. Sets $\mathcal{P}, \mathcal{C}_i, ...$ are denoted using calligraphic letters.

Efficiency is yet 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 beforehand to

F. Nielsen is with the Department of Fundamental Research of Sony Computer Science Laboratories, Inc., Tokyo, Japan, and the Computer Science Department (LIX) of École Polytechnique, Palaiseau, France. e-mail: Frank.Nielsen@acm.org

R. Nock is with the CEREGMIA Department, University of Antilles-Guyane, Martinique, France. e-mail: rnock@martinique.univ-ag.fr

Manuscript received November 2007, revised October 2008.

[1] A formal definition considers *probability measures* P and Q defined on a measurable space $(\mathcal{X}, \mathcal{A})$. These probability measures are assumed dominated by a σ-finite measure μ with respective densities $p = \frac{\mathrm{d}P}{\mathrm{d}\mu}$ and $q = \frac{\mathrm{d}Q}{\mathrm{d}\mu}$. The Kullback-Leibler divergence is then defined as $\mathrm{KL}(P||Q) = \int \frac{\mathrm{d}P}{\mathrm{d}\mu} \log(\frac{\mathrm{d}P}{\mathrm{d}\mu} / \frac{\mathrm{d}Q}{\mathrm{d}\mu}) \mathrm{d}\mu$. See [6] a recent study on information and divergences in statistics.

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 [15] proposed the k-means iterative clustering algorithm for solving vector quantization problems. Briefly, the k-means algorithm starts by choosing k seeds[2] 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 successive iterations goes below a prescribed threshold. Lloyd chose to minimize the *squared* Euclidean distance since the minimum average intra-cluster distance yields centroids, the *centers of mass* of the respective clusters. Lloyd [15] further proved that the iterative k-means algorithm *monotonically* converges to a *local* optima of the quadratic function loss (minimum variance loss):

$$\sum_{i=1}^{k}\sum_{p_j\in\mathcal{C}_i}^{n}||p_j-c_i||^2.$$

Cluster $\mathcal{C}_i$'s center c_i is defined by the following minimization problem

$$\begin{aligned} c_i &= \arg\min_{c}\sum_{p_j\in\mathcal{C}_i}||c-p_j||^2, &(1)\\ &= \arg\min_{c\in\mathbb{R}^d}\mathrm{AVG}_{L_2^2}(\mathcal{C}_i,c), &(2)\\ &= \frac{1}{|\mathcal{C}_i|}\sum_{p_j\in\mathcal{C}_i}p_j, &(3)\end{aligned}$$

where $|\mathcal{C}_i|$ denotes the cardinality of $\mathcal{C}_i$, and the c_i's and p_i's are real-valued d-dimensional vectors. That is, the minimum average squared distance of the cluster center to the cluster points is reached uniquely by the centroid: The center of mass of the cluster. Note that considering the Euclidean distance instead of the squared Euclidean distance yields another remarkable *center point* of the cluster called the Fermat-Weber point [18]. Although the Fermat-Weber point is also provably unique, it does not have closed-form solutions. It is thus interesting to ask oneself what other kinds of distances in Eq. 2 (besides the squared distance) yield simple closed-form solutions that are of interests for processing multimedia information. Half a century later, Banerjee et al. [19] showed in 2004 that the celebrated k-means algorithm *extends to* and remarkably *only* works [20] for a broad family of distortion measures called Bregman divergences [21], [22]. Let $\mathbb{R}^+$ denote the non-negative part of the real line: $\mathbb{R}^+=[0,+\infty)$. In this paper, we consider only Bregman divergences defined on vector points $p_i\in\mathbb{R}^d$ in fixed dimension.[3]

Bregman divergences D_F form a family of distortion measures that are defined by a strictly convex and differentiable generator function $F:\mathcal{X}\rightarrow\mathbb{R}^+$ on a convex domain

[2]Forgy's initialization [16] consists merely in choosing at random the seeds from the source vectors. Arthur and Vassilvitskii [17] proved that a better careful initialization yields expected guarantees on the clustering.

[3]See the concluding remarks in Section VI for extensions of Bregman divergences to matrices [23], [3], and recent functional extensions [24] of Bregman divergences.

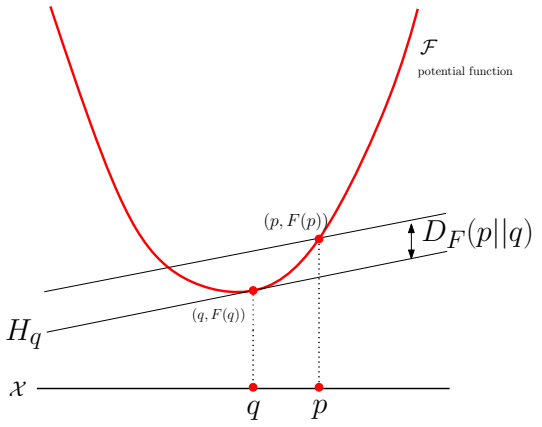

Fig. 1. Geometric interpretation of a univariate Bregman divergence. $D_F(.||q)$ is the vertical distance between the potential function plot $\mathcal{F}=\{(x,F(x))\mid x\in\mathcal{X}\}$ and the hyperplane H_q tangent to $\mathcal{F}$ at $(q,F(q))$.

$\mathrm{dom}F=\mathcal{X}\subseteq\mathbb{R}^d$ (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 (also commonly called the "dot" product):

$$<p,q>=\sum_{i=1}^{d}p^{(i)}q^{(i)}=p^Tq,$$

and $\nabla F(q)$ denotes the gradient of F at vector point q:

$$\nabla F(q)=\left[\frac{\partial F(q)}{\partial x^{(1)}},...,\frac{\partial F(q)}{\partial x^{(d)}}\right].$$

See Figure 1 for a geometric interpretation of Bregman divergences. Thus Bregman divergences define a *parameterized* family of distortions measures D_F that unify the squared Euclidean distance with the statistical Kullback-Leibler divergence:

- Namely, the squared Euclidean distance is a Bregman divergence in disguise obtained for the generator $F(x)=\sum_{i=1}^{d}(x^{(i)})^2$ that represents the paraboloid potential function (see Figure 1), or the quadratic loss on vector points in the k-means algorithm.
- The Kullback-Leibler divergence is yet another Bregman divergence in disguise obtained for the generator $F(x)=\sum_{i=1}^{d}x^{(i)}\log x^{(i)}$ that represents the negative Shannon entropy on probability vectors [14] (normalized unit length vectors lying on the d-dimensional probability simplex $\mathbb{S}^d$).

A Bregman divergence D_F is said *separable* [19], [25] if its generator can be obtained coordinate-wise from a univariate generator f as:

$$F(x)=\sum_{i=1}^{d}f(x^{(i)}).$$

Table I reports the generators of common univariate Bregman divergences (ie., divergences defined on scalars $x\in\mathbb{R}$ — $d=1$). Multivariate separable Bregman divergences defined on $x\in\mathbb{R}^d$ can be easily constructed piecewise from univariate

Domain $\mathcal{X}$	Function F	Gradient $\frac{F(x)}{\mathrm{d}x} = F'(x)$	Inverse gradient $(F'(x))^{-1}$	Divergence $D_F(p\|\|q)$
$\mathbb{R}$	Squared function x^2	$2x$	$\frac{x}{2}$	Squared loss $(p-q)^2$
$\mathbb{R}_+, \alpha \in \mathbb{N}$ $\alpha > 1$	Norm-like x^α	$\alpha x^{\alpha-1}$	$(\frac{x}{\alpha})^{\frac{1}{\alpha-1}}$	Norm-like $p^\alpha + (\alpha-1)q^\alpha - \alpha p q^{\alpha-1}$
$\mathbb{R}^+$	Unnormalized Shannon entropy $x\log x - x$	$\log x$	$\exp(x)$	Kullback-Leibler divergence (I-divergence) $p\log\frac{p}{q} - p + q$
$\mathbb{R}$	Exponential $\exp x$	$\exp x$	$\log x$	Exponential loss $\exp(p) - (p-q+1)\exp(q)$
$\mathbb{R}^+*$	Burg entropy $-\log x$	$-\frac{1}{x}$	$-\frac{1}{x}$	Itakura-Saito divergence $\frac{p}{q} - \log\frac{p}{q} - 1$
$[0,1]$	Bit entropy $x\log x + (1-x)\log(1-x)$	$\log\frac{x}{1-x}$	$\frac{\exp x}{1+\exp x}$	Logistic loss $p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$
$\mathbb{R}$	Dual bit entropy $\log(1+\exp x)$	$\frac{\exp x}{1+\exp x}$	$\log\frac{x}{1-x}$	Dual logistic loss $\log\frac{1+\exp p}{1+\exp q} - (p-q)\frac{\exp q}{1+\exp q}$
$[-1,1]$	Hellinger-like $-\sqrt{1-x^2}$	$\frac{x}{\sqrt{1-x^2}}$	$\frac{x}{\sqrt{1+x^2}}$	Hellinger-like $\frac{1-pq}{\sqrt{1-q^2}} - \sqrt{1-p^2}$

TABLE I
COMMON UNIVARIATE BREGMAN DIVERGENCES D_F USED FOR CREATING SEPARABLE BREGMAN DIVERGENCES.

Bregman divergences. The generalized quadratic distances[4] $||p-q||_Q^2 = (p-q)^T Q(p-q)$ defined for a $d \times d$ positive definite matrix Q are the only symmetric Bregman divergences[5] obtained from the non-separable generator $F(x) = x^T Qx$, see [25], [23].

Thus in Barnerjee et al. [19], the original k-means algorithm is extended into a meta-algorithm, called the Bregman k-means, that works for any given Bregman divergence. Furthermore, Barnerjee et al. [20], [19] proved the property that the mean is the *minimizer* of the expected Bregman divergence. The fundamental underlying primitive for these *center-based* clustering algorithms is to find the intrinsic *best single representative* of a cluster with respect to a distance function $d(\cdot,\cdot)$. As mentioned above, the centroid of a point set $\mathcal{P} = \{p_1, ..., p_n\}$ (with $\mathcal{P} \subset \mathcal{X}$) is defined as the optimizer of the *minimum average distance*:

$$c = \arg\min_p \frac{1}{n}\sum_i d(p, p_i).$$

For oriented distance functions such as aforementioned Bregman divergences that are not necessarily symmetric, we thus need to distinguish *sided* and *symmetrized* centroids as follows:

$$\begin{aligned} c_R^F &= \arg\min_{c\in\mathcal{X}} \frac{1}{n}\sum_{i=1}^n D_F(p_i||\boxed{\mathrm{c}}), \\ c_L^F &= \arg\min_{c\in\mathcal{X}} \frac{1}{n}\sum_{i=1}^n D_F(\boxed{\mathrm{c}}||p_i), \end{aligned}$$

$$c^F = \arg\min_{c\in\mathcal{X}} \frac{1}{n}\sum_{i=1}^n \frac{D_F(p_i||\boxed{\mathrm{c}}) + D_F(\boxed{\mathrm{c}}||p_i)}{2}.$$

The first right-type and left-type centroids c_R^F and c_L^F are called *sided centroids* (with the superscript L standing for left, and R for right), 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 [25] for a proof. Since the three centroids coincide with the center of mass for symmetric Bregman divergences (generalized quadratic distances), we consider in the remainder asymmetric Bregman divergences. For a given point set $\mathcal{P} = \{p_1, ..., p_n\}$, we write for short the minimum averages as:

$$\mathrm{AVG}_F(\mathcal{P}||c) = \frac{1}{n}\sum_{i=1}^n D_F(p_i||c), \quad (4)$$

$$\mathrm{AVG}_F(c||\mathcal{P}) = \frac{1}{n}\sum_{i=1}^n D_F(c||p_i), \quad (5)$$

$$\mathrm{AVG}_F(c;\mathcal{P}) = \frac{1}{n}\sum_{i=1}^n S_F(c;p_i) = \mathrm{AVG}_F(\mathcal{P};c), \quad (6)$$

so that we get respectively the three kinds of centroids as:

$$c_R^F = \arg\min_{c\in\mathcal{X}} \mathrm{AVG}_F(\mathcal{P}||c), \quad (7)$$

$$c_L^F = \arg\min_{c\in\mathcal{X}} \mathrm{AVG}_F(c||\mathcal{P}), \quad (8)$$

$$c^F = \arg\min_{c\in\mathcal{X}} \mathrm{AVG}_F(\mathcal{P};c). \quad (9)$$

We use the semi-colon ";" notation[6] in symmetrized divergence $S_F(c;p_i)$ and average mean $\mathrm{AVG}_F(\mathcal{P};c)$ to indicate that it is symmetric: $S_F(c;p_i) = S_F(p_i;c)$ and

[4]The squared Mahalanobis distance is a generalized quadratic distance obtained by choosing matrix Q as the inverse of the variance-covariance matrix [25].

[5]Note that the quadratic form of distances $||p-q||_Q^2 = (p-q)^T Q(p-q)$ amounts to compute the squared Euclidean distance on transformed points with the mapping $x \mapsto Lx$, where L is the triangular matrix of Cholesky decomposition $Q = L^T L$ since $||p-q||_Q^2 = (p-q)^T L^T L(p-q) = ||Lp - Lq||^2$.

[6]We reserve the comma notation "," in divergences to stress out the metric property.

$\mathrm{AVG}_F(\mathcal{P};c) = \mathrm{AVG}_F(c;\mathcal{P})$. The Jensen-Shannon divergence [26], [27] (symmetrized Kullback-Leibler divergence obtained for $F(x) = \sum_{i=1}^{d} x^{(i)} \log x^{(i)}$, the negative Shannon entropy) and COSH centroids [28], [29] (symmetrized Itakura-Saito divergence S_F, obtained for the Burg entropy [19], [30]: $F(x) = -\sum_{i=1}^{d} \log x^{(i)}$) are certainly the most famous symmetrized Bregman centroids, widely used in image and sound processing. These symmetrized centroids play a fundamental role in information retrieval (IR) applications that require to handle symmetric information-theoretic distances. Note that Bregman divergences can also be assembled block-wise for processing multimedia information and retrieval combining both auditory and visual signals. Table II presents a table of common Bregman divergences (or symmetrized Bregman divergences) in action for processing multimedia signals in real-world applications. This table is by no means exhaustive. Banerjee et al. [19] proved a bijection between regular exponential families and a corresponding subclass of Bregman divergences called regular Bregman divergences. They experimentally showed that clustering exponential families with the corresponding Bregman divergences yields better results. This exponential family/Bregman divergence bijection indicates why some Bregman divergences are better suited than others. For example in sound processing, the speech power spectra can be modeled by exponential family densities of the form $\lambda e^{-\lambda x}$ whose corresponding associated regular Bregman divergence is no less than the Itakura-Saito divergence. We refer the reader to the first comprehensive "Dictionary of distances" [9] (especially, chapter 21 dealing with "Image and Audio Distances") for further hints and explanations for which divergence is useful for which applications.

A. Kullback-Leibler divergence of exponential families as Bregman divergences

In statistics, exponential families [19], [25] represent a large class of popular discrete and continuous distributions with prominent members such as Bernoulli, multinomial, beta, gamma, normal, Rayleigh, Laplacian, Poisson, Wishart, etc. just to name a few. The probability mass/density functions of exponential families are parametric distributions that can be written using the following canonical decomposition:

$$p(x|\theta) = \exp\{\langle \theta, t(x)\rangle - F(\theta) + C(x)\},$$

where $t(x)$ denotes the *sufficient statistics* and θ represents the *natural parameters*. Since $\log \int_x p(x|\theta)\mathrm{d}x = \log 1 = 0$, we have $F(\theta) = \log \int_x \exp\{\langle \theta, t(x)\rangle + C(x)\}\mathrm{d}x$. F is called the *log normalizer function* and fully characterizes the exponential family $\mathcal{E}_F$. Term $C(x)$ ensures density normalization.

It turns out that the Kullback-Leibler divergence of distributions $p(x|\theta_p)$ and $p(x|\theta_q)$ belonging to the same exponential family $\mathcal{E}_F$ is equivalent to the Bregman divergence D_F for the log normalizer function on swapped natural parameters:

$$\mathrm{KL}(p(x|\theta_p)||p(x|\theta_q)) = D_F(\theta_q||\theta_p)$$

See [25] for a proof. Thus a left-sided/right-sided/symmetrized Kullback-Leibler centroid on a set of distributions of the same exponential family is a corresponding right-sided/left-sided/symmetrized Bregman centroid on a set of vectors of the natural space $\mathcal{X}$.

B. Properties of sided and symmetrized centroids

In practice, once the proper Bregman divergence is chosen, we still need to choose between the left-sided, right-sided or symmetrized centroid. These centroids exhibit different characteristics that help choose the proper centroid for the given application. Without loss of generality[7], consider the most prominent asymmetric Bregman divergence: The Kullback-Leibler divergence. Furthermore, for illustrative purposes, consider a set of n normal distributions $\{\mathcal{N}_1, ..., \mathcal{N}_n\}$. Each normal distribution $\mathcal{N}_i$ has probability density function $p_i(x|\mu_i, \sigma_i^2)$ (pdf. for short):

$$p_i(x|\mu_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x-\mu_i)}{2{\sigma_i}^2}\right)$$

that can be modeled by a corresponding 2D point $p_i = (\mu_i, {\sigma_i}^2)$ of mean μ_i and variance σ^2 in parameter space $\mathcal{X} = \mathbb{R} \times \mathbb{R}_*^+$. The Kullback-Leibler divergence between two normals has the following closed-form solution[8]:

$$\mathrm{KL}(p(x|\mu_p, \sigma_p^2)||p(x|\mu_q, \sigma_q^2)) = \frac{1}{2}\left(2\log\frac{\sigma_q}{\sigma_p} + \frac{\sigma_p^2}{\sigma_q^2} + \frac{(\mu_q-\mu_p)^2}{\sigma_q^2} - 1\right).$$

Observe that the closed-form formula is computed for 2D points $p_i = (p_i^{(1)} = \mu_i, p_i^{(2)} = \sigma_i^2)$ in the parameter space $\mathcal{X}$. For identical normal variances $\sigma_p^2 = \sigma_q^2$ the Kullback-Leibler divergence amounts to a weighted squared Euclidean distance.

Figure 2 displays an example of left/right sided and symmetrized centroids of normals for a set that consists of two normals: $\mathcal{N}_1 = \mathcal{N}(-4, 2^2 = 4)$ and $\mathcal{N}_2 = \mathcal{N}(5, 0.8^2 = 0.64)$. We observe the following properties:

- The Kullback-Leibler right-sided centroid is *"zero-avoiding"* so that its corresponding density function tries to cover the support of all input normals,
- The Kullback-Leibler left-sided centroid is *"zero-forcing"* so that it focuses on the highest mass mode normal.

That zero-avoiding/zero-forcing terminology is related to the description of Minka [11] (pages 3-4) that considered Gaussian mixture simplification of a 2-component Gaussian mixture to a single Gaussian component. The Kullback-Leibler left-sided centroid prefers to better represent only the highest-mode individual of the set while the right-sided centroid prefers to stretch over all individuals. Following yet another terminology of Winn and Bishop [31], we observe when modeling the "mean" probability density function that the Kullback-Leibler left-sided centroid exhibits an *exclusive* behavior (ignore modes of the set to select the highest one)

[7] Indeed, as shown earlier, Bregman divergences can be interpreted as equivalent Kullback-Leibler divergences on corresponding parametric exponential families in statistics by swapping the argument order [19], [25].

[8] The Kullback-Leibler divergence of normals is equivalent to a Bregman divergence for a corresponding generator F by swapping argument order. See [19], [25].

while the Kullback-Leibler right-sided centroid displays an inclusive property.

To get a mathematical flavor of these zero-forcing/zero-avoiding behaviors, consider without loss of generality[9] the Kullback-Leibler divergence on finite discrete set of distributions (ie., multinomial distributions with d outcomes). The right-sided centroid is the minimizer $c_R = \arg_c \min \frac{1}{n}\sum_{i=1}^n \mathrm{KL}(p_i||c)$. That is, we seek for the d-dimensional probability vector c that minimizes $\min \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^d p_i^{(j)} \log \frac{p_i^{(j)}}{c^{(j)}}$. Thus, intuitively whenever $p_i^{(j)} \neq 0$, the minimization process ought to choose $c^{(j)} \neq 0$. Otherwise, setting $c^{(j)} = 0$ yields $p_i \log \frac{p_i^{(j)}}{c^{(j)}} \rightarrow \infty$ (ie., the Kullback-Leibler divergence is unbounded). That is, the right-sided Kullback-Leibler centroid (that is a left-sided Bregman centroid) is zero-avoiding. Note that this minimization is equivalent to maximizing the average cross-entropies $\frac{1}{n}\sum_{i=1}^n \sum_{j=1}^d p_i^{(j)} \log c^{(j)}$, and thus the right-sided Kullback-Leibler centroid c is zero-avoiding for all $p_i^{(j)} \neq 0$.

Similarly, the left-sided Kullback-Leibler centroid $c_L = \arg_c \min \frac{1}{n}\sum_{i=1}^n \mathrm{KL}(c||p_i)$ is obtained by minimizing $\min \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^d c_i^{(j)} \log \frac{c_i^{(j)}}{p_i^{(j)}}$. This minimization is zero-forcing since whenever there exists a $p_i^{(j)} = 0$, the minimization tasks chooses to set $c^{(j)} = 0$. That means that the right-sided Bregman centroid (a left-sided Kullback-Leibler divergence in disguise) is zero-forcing.

The symmetrized Kullback-Leibler centroid is defined as the minimizer of the Jensen-Shannon divergence (which has always finite value). That is, the symmetrized centroid minimizes the *total divergence* to the *average probability density* $m(x) = \frac{p(x)+q(x)}{2}$ as follows:

$$c = \arg\min_{c\in\mathcal{X}} \frac{1}{2}\mathrm{KL}(p(x)||m(x)) + \frac{1}{2}\mathrm{KL}(q(x)||m(x)).$$

Therefore the symmetrized centroid strikes a balance between the two zero-forcing and zero-avoiding properties with respect to the mean distribution.

C. 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 III under our notational conventions. Ben-Tal et al. [32] 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 [33] compared in the 1-page paper the generalized/entropic mean values for two entropy-based classes of divergences: f-divergences [34] and Jensen-Shannon divergences [35]. The closest recent work to our study is Veldhuis' approximation method [36], [37] for computing the symmetrical Kullback-Leibler centroid.

[9]As explained by Banerjee et al. [19], [25], the Kullback-Leibler divergence of distributions of the same exponential families is a Bregman divergence on the natural parameters of these distributions obtained by swapping the order of the arguments. Arbitrary probability measures can be approximated by multinomial distributions that belong to the exponential family.

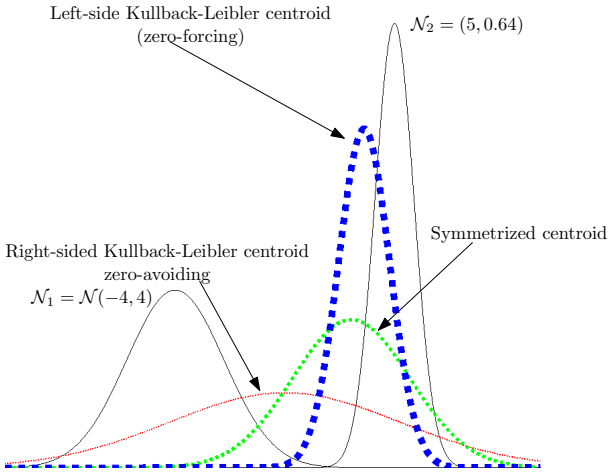

Fig. 2. Visualizing the fundamental properties of the left-sided, the right-sided and the symmetrized centroids (with $\mathcal{N}_1 = \mathcal{N}(-4, 2^2 = 4)$ and $\mathcal{N}_2 = \mathcal{N}(5, 0.8^2 = 0.64)$): The right-sided centroid (thin dashed red line) is zero-avoiding and tries to cover the support of both normals. The left-sided centroid (thick dashed blue line) is zero-forcing and focuses on the highest mode (smallest variance). The symmetrized centroid (medium dashed green line) exhibits a trade-off between these two zero-avoiding/zero-forcing properties.

We summarize our contributions as follows:

- In section III, 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 characterize Sibson' s notion of *information radius* [38] for these sided centroids, and show that they are both equal to the F-Jensen difference, a generalized Jensen-Shannon divergence [39] also known as Burbea-Rao divergences [40].
- Section IV proceeds by first showing how to reduce the symmetrized $\min \mathrm{AVG}_F(c;\mathcal{P})$ 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 [27], [28], [26], [41], [42]. Section V 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

Divergence name	Formula	Generator $F(x)$ for D_F	Examples of application domains
(squared) Mahalanobis	$M_A(p;q) = (p-q)^T A(p-q)$ (gen. quadratic loss, A semi-positive definite matrix)	$F(x) = x^T Ax$ (operations research)	Facility locations
Kullback-Leibler	$\mathrm{KL}(p\|\|q) = \sum_{i=1}^d p^{(i)} \log \frac{p^{(i)}}{q^{(i)}}$ (negative Shannon entropy)	$H(x) = \sum_i x^{(i)} \log x^{(i)}$	Statistical analysis
Jensen-Shannon	$\mathrm{JS}(p;q) = \sum_{i=1}^d (p^{(i)} - q^{(i)}) \log \frac{p^{(i)}}{q^{(i)}}$	symmetrized Kullback-Leibler	Image retrieval
Itakura-Saito	$\mathrm{IS}(p\|\|q) = \sum_{i=1}^d \left(\frac{p^{(i)}}{q^{(i)}} - \log \frac{p^{(i)}}{q^{(i)}} - 1 \right)$ (Burg entropy)	$B(x) = -\sum_i \log x^{(i)}$	Sound processing
COSH	$\mathrm{COSH}(p;q) = \frac{1}{2} \left(\sum_{i=1}^d \left(\frac{p^{(i)}}{q^{(i)}} + \frac{q^{(i)}}{p^{(i)}} \right) \right) - d$ $\mathrm{COSH}(p;q) = \frac{1}{2} \sum_{i=1}^d \left(\sqrt{\frac{p^{(i)}}{q^{(i)}}} - \sqrt{\frac{q^{(i)}}{p^{(i)}}} \right)^2$	symmetrized Itakura-Saito	Sound retrieval

TABLE II
BREGMAN OR SYMMETRIZED BREGMAN DIVERGENCES WITH CORRESPONDING CORE APPLICATION DOMAINS.

the exponential families [25] 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 [36], [37], [41], 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 [27] that relies on solving Riccati matrix equations.

But first, we start in Section II by introducing the dually flat space construction from an arbitrary convex function. This section may be skimmed through at first reading since it is devoting to define the sided Bregman centroids under the framework of dually flat spaces of information geometry.

II. GEOMETRY UNDERLYING BREGMAN DIVERGENCES: DUALLY FLAT MANIFOLDS

We concisely review the construction of dually flat manifolds from convex functions. This construction lies at the very heart of information geometry [43]. A full description of this construction is presented in the comprehensive survey chapter of Amari [10] (see also [44], [45]). Information geometry [43] originally emerged from the studies of *invariant properties* of a manifold of probability distributions $\mathcal{D}$, say the manifold of univariate normal distributions:

$$\mathcal{D} = \{p(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mid \mu \in \mathbb{R}, \sigma \in \mathbb{R}_*^+\}.$$

Information geometry relies on differential geometry and in particular on the sophisticated notion of *affine* connections[10] (pioneered by Cartan [46]) whose explanation is beyond the scope of this paper [43]. We rather describe the three most fundamental items of dually flat manifolds:

- The fundamental convex duality and the dual coordinate systems arising from Legendre transformation, and
- The generalized Pythagorean relation, and
- The notion of Bregman projection.

These descriptions will enlighten geometrically the results of the paper. The point is to show that Bregman divergences form the *canonical distances* of dually flat manifolds arising when studying family of probability distributions. Those flat geometries nicely generalize the familiar Euclidean geometry. Furthermore, these flat geometries reveal a fundamental geometric duality that is hidden when dealing with the regular Euclidean geometry.

A. Riemannian metric associated to a convex function

Consider a smooth real-valued convex function $F(\theta)$ defined in an open set $\mathcal{X}$ of $\mathbb{R}^d$, where θ denotes a fixed coordinate system. Notice that the notion of function convexity depends on the considered coordinate system θ:

$$F\left((1-\lambda)\theta_1 + \lambda\theta_2\right) \leq (1-\lambda)F(\theta_1) + \lambda F(\theta_2).$$

The second derivatives of the function F form its Hessian matrix $\nabla^2 F = (g_{ij})$ that is a positive definite matrix[11] depending on its position θ:

$$\nabla^2 F(\theta) = (g_{ij}(\theta)) = (\partial_i \partial_j F(\theta)) \succ 0,$$

where $\partial_i = \frac{\partial}{\partial \theta^{(i)}}$ and $\theta = (\theta^{(1)}, ..., \theta^{(d)})$. For two infinitesimally nearby points θ and $\theta + \mathrm{d}\theta$, define the *square* of their distance by

$$\mathrm{d}s^2 = < \mathrm{d}\theta, \mathrm{d}\theta > = \sum_{i,j} g_{ij}(\theta) \mathrm{d}\theta^{(i)} \mathrm{d}\theta^{(j)},$$

where $< \mathrm{d}\theta, \mathrm{d}\theta >$ denote the inner product. A manifold with such an infinitesimal distance is called a Riemannian manifold, and matrix $g = (g_{ij})$ is called the Riemannian metric. Observe that $\mathrm{d}s^2$ is obtained from the second-order term of the Taylor expansion of $F(\theta + \mathrm{d}\theta)$:

[10]Connections relate the vector tangent spaces for infinitesimal displacements on the manifold. A riemannian connection (also called Levi-Civita connection) is such that parallel transport gives an isometry between the tangent planes. To contrast with, an affine connection uses an affine transformation.

[11]A matrix M is positive definite iff. for all x we have $x^T Mx > 0$. We write $M \succ 0$ to denote the positive-definiteness of the matrix M.

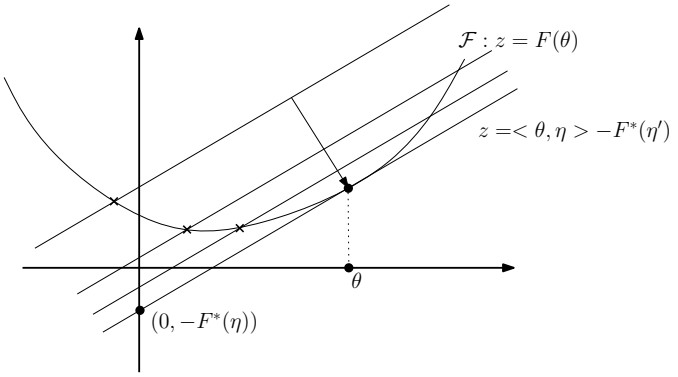

Fig. 3. Legendre transformation of a strictly convex function F: The z-intercept $(0, -F^*(\eta))$ of the tangent hyperplane $H_\theta : z =< \eta, \theta > -F^*(\eta)$ of the potential function defines the value of the Legendre transform F^* for the dual coordinate η. Any hyperplane passing through another point of the potential function and parallel to H_θ necessarily intersects the z-axis above $-F^*(\eta)$.

$$F(\theta + \mathrm{d}\theta) = F(\theta) + \sum_i \partial_i F(\theta)\mathrm{d}\theta^{(i)} + \frac{1}{2}\sum_{i,j} g_{ij}(\theta)\mathrm{d}\theta^{(i)}\mathrm{d}\theta^{(j)}.$$

A *geodesic* Γ_{PQ} of manifold $\mathcal{D}$ is defined by the *straight line* connecting two points P and Q (with respective coordinates $\theta_P = \theta(P)$ and $\theta_Q = \theta(Q)$ in the θ-coordinate system):

$$\Gamma_{PQ} = \{X(\lambda), \lambda \in [0,1] \mid \theta_{X(\lambda)} = (1-\lambda)\theta_P + \lambda\theta_Q\}.$$

When $F(\theta) = \frac{1}{2}\sum_i {\theta^{(i)}}^2$ is the paraboloid function, we have $g_{ij} = \delta_{ij}$ the Krönecker symbol:

$$\delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

and the geometry is Euclidean because of the implied squared distance $\mathrm{d}s^2 = \sum_i {\mathrm{d}\theta^{(i)}}^2$. In order to retrieve the global geometrical structure of the manifold, we need the geometry to be independent of the choice of the coordinate system. The following section reveals that the θ-coordinate system admits a dual η-coordinate system.

B. Convex duality and dual coordinate systems from Legendre transformation

Consider the gradient $\nabla F(\theta) = \eta$ defined by the following partial derivatives:

$$\eta^{(i)} = \frac{\partial}{\partial\theta^{(i)}}F(\theta).$$

There is a one-to-one correspondence [10] between θ and η so that we can use η as another coordinate system. The transformation mapping θ to η (with $\theta^{(i)}$ mutually reciprocal to $\eta^{(i)}$) is the Legendre transformation [43] defined for any convex function F as follows:

$$F^*(\eta) = \max_{\theta\in\mathcal{X}}\{< \theta, \eta > -F(\theta)\}.$$

Figure 3 visually depicts the Legendre transformation. (The drawing illustrates why the Legendre transformation is also sometimes loosely called the "slope transformation.")

Table III displays two examples of Legendre transformation. (For the geometry of exponential families in statistics, the primal θ-coordinate system is called the natural coordinate system and the dual η-coordinate system is called the expectation or moment coordinate system.) The dual convex conjugates F and F^* are called potential functions (or contrast functions) and satisfy the following fundamental equality:

$$F(\theta) + F^*(\eta) - < \theta, \eta >= 0.$$

The inverse transformation $\eta \rightarrow \theta$ is given by the gradient of F^*:

$$\theta = \nabla F^*(\eta),$$

with $\theta^{(i)} = \frac{\partial}{\partial\eta^{(i)}}F^*(\theta)$. That is, θ and η are *coupled* and form a *dual coordinate system* of the geometry implied by a pair of Legendre convex function (F, F^*). The dual Riemannian metric associated with F^* is

$$g^*_{ij} = \frac{\partial^2}{\partial\eta^{(i)}\partial\eta^{(j)}}F^*(\eta),$$

and we have the remarkable property that

$$(g^*_{ij}) = (g_{ij})^{-1}$$

That is, Riemannian metric (g^*_{ij}) is the inverse matrix of the Riemannian metric (g_{ij}). It follows from the construction that these two metrics are geometrically the same [10], as we have identical infinitesimal lengths:

$$\sum g_{ij}\mathrm{d}\theta^{(i)}\mathrm{d}\theta^{(j)} = \sum g^*_{ij}\mathrm{d}\eta^{(i)}\mathrm{d}\eta^{(j)}.$$

C. Bregman divergences from the dual coordinate systems

A distortion measure, called divergence, between two points P and Q of the geometric manifold (either indexed by θ or η coordinate system) is defined as:

$$D_F(P||Q) = F(\theta_P) + F^*(\eta_Q) - < \theta_P, \eta_Q >,$$

with $< \theta_P, \eta_Q >= \theta_P^T\eta_Q = \sum\theta^{(i)}\eta^{(i)}$. We have $D(P||Q) \geq 0$. Changing the role of P and Q, or θ and η, we get the dual divergence:

$$D_{F^*}(P||Q) = F^*(\eta_P) + F(\theta_Q) - < \eta_P, \theta_Q >,$$

so that

$$D_F(P||Q) = D_{F^*}(Q||P).$$

When Q is close to P, we write $Q = P + \mathrm{d}P$ and get the squared Riemannian distance as:

$$\begin{aligned} D(P||Q) &= D(P||P + \mathrm{d}P) \\ &= \frac{1}{2}\sum g_{ij}\mathrm{d}\theta^{(i)}\mathrm{d}\theta^{(j)} = \frac{1}{2}\sum g^*_{ij}\mathrm{d}\eta^{(i)}\mathrm{d}\eta^{(j)}. \end{aligned}$$

In particular, this squared Riemmanian approximation means that the canonical divergence does *not* satisfy[12] the triangle inequality. Next, we show that we get a remarkable generalization of Pythagoras' theorem.

[12]Indeed, notice that the *squared* Euclidean distance obtained from the paraboloid function does not satisfy the triangle inequality.

Logistic loss/binary relative entropy		
$F(\theta) = \log(1+\exp\theta)$	$D_F(\theta\|\|\theta') = \log\frac{1+\exp\theta}{1+\exp\theta'} - (\theta-\theta')\frac{\exp\theta'}{1+\exp\theta'}$	$\nabla F(\theta) = \frac{\exp\theta}{1+\exp\theta} = \eta$
$F^*(\eta) = \eta\log\eta + (1-\eta)\log(1-\eta)$	$D_{F^*}(\eta'\|\|\eta) = \eta'\log\frac{\eta'}{\eta} + (1-\eta)\log\frac{1-\eta'}{1-\eta}$	$\nabla F^*(\eta) = \log\frac{\eta}{1-\eta} = \theta$
Exponential loss/Unnormalized Shannon entropy		
$F(\theta) = \exp\theta$	$D_F(\theta\|\|\theta') = \exp\theta - \exp\theta' - (\theta-\theta')\exp\theta'$	$\nabla F(\theta) = \exp\theta = \eta$
$F^*(\eta) = \eta\log\eta - \eta$	$D_{F^*}(\eta'\|\|\eta) = \eta'\log\frac{\eta'}{\eta} + \eta - \eta'$	$\nabla F^*(\eta) = \log\eta = \theta$

TABLE III
TWO EXAMPLES OF LEGENDRE TRANSFORMATIONS WITH THEIR ASSOCIATED DUAL PARAMETERIZATIONS.

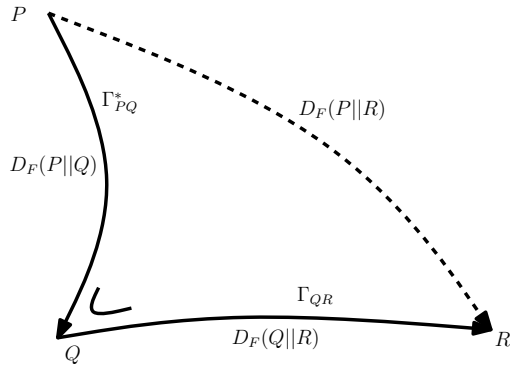

Fig. 4. Illustrating the generalized Pythagorean theorem: For $\Gamma^*_{PQ} \perp \Gamma_{QR}$, we have $D_F(P||R) = D_F(P||Q) + D_F(Q||R)$.

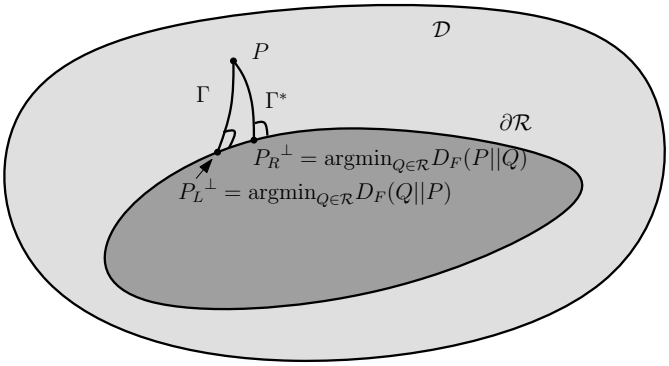

Fig. 5. Illustrating the sided Bregman projections $P_R{}^{\perp}$ and $P_L{}^{\perp}$ of a point $P \in \mathcal{X}$ for a convex region $\mathcal{R}$: The dual geodesic Γ^* connecting P to $P_R{}^{\perp}$ and the geodesic Γ connecting P to $P_L{}^{\perp}$ are orthogonal to the boundary $\partial\mathcal{R}$.

D. Generalized Pythagoras' theorem

Consider two curves $\theta(t)$ and $\theta'(t)$ parameterized by a scalar t in the θ-coordinate system, and assume w.l.o.g that these curves intersect at $t = 0$: $\theta(0) = \theta'(0)$. Using the dual coordinate system η, we similarly have $\eta(0) = \eta'(0)$. The tangent vector of a curve $\theta(t)$ is at t is the vector:

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = \left(\frac{\mathrm{d}\theta^{(1)}(t)}{\mathrm{d}t}, ..., \frac{\mathrm{d}\theta^{(n)}(t)}{\mathrm{d}t}\right)$$

of derivatives with respect to t. The two curves are said to be *orthogonal* at the intersection point when their inner product vanishes:

$$< \frac{\mathrm{d}\theta}{\mathrm{d}t}, \frac{\mathrm{d}\theta'}{\mathrm{d}t} >= \sum g_{ij}\frac{\mathrm{d}\theta^{(i)}}{\mathrm{d}t}\frac{\mathrm{d}\theta'^{(i)}}{\mathrm{d}t} = 0.$$

Using the two coordinate systems, this is equivalent to

$$< \frac{\mathrm{d}\theta}{\mathrm{d}t}, \frac{\mathrm{d}\eta'}{\mathrm{d}t} >= 0.$$

Dually flat manifolds exhibit a generalized Pythagoras' theorem:

Theorem 2.1 (Generalized Pythagoras' theorem [43]): When the dual geodesic Γ^*_{PQ} connecting P and Q is orthogonal to the geodesic Γ_{QR} connecting Q and R (see Figure 4), we have: $D_F(P||R) = D_F(P||Q) + D_F(Q||R)$, or dually $D_{F^*}(P||R) = D_{F^*}(P||Q) + D_{F^*}(Q||R)$.

Notice that when we consider the paraboloid convex function $F(\theta) = \sum_i (\theta^{(i)})^2$, the metric $(g_{ij}) = (g^*_{ij})$ is the identity matrix and therefore the primal/dual geodesics coincide.

E. Dual convexity and sided Bregman projections

We say that a region $\mathcal{R}$ is *convex* (or θ-convex) when the *geodesic* connecting any two points $P, Q \in \mathcal{R}$ is fully contained in $\mathcal{R}$. That is,

$$\forall P, Q \in \mathcal{X}, \lambda \in [0,1],\ (1-\lambda)\theta_P + \lambda\theta_Q \in \mathcal{R}.$$

Similarly, region $\mathcal{R}$ is said *dual convex* (or η-convex) when the *dual geodesic* connecting any two points $P, Q \in \mathcal{R}$ is fully contained in $\mathcal{R}$:

$$\forall P, Q \in \mathcal{X}, \lambda \in [0,1],\ (1-\lambda)\eta_P + \lambda\eta_Q \in \mathcal{R}.$$

Let $P_R{}^{\perp} \in \mathcal{R}$ be the point that minimizes $D_F(P||Q)$ for $Q \in \mathcal{R}$, and $P_L{}^{\perp} \in \mathcal{R}$ be the point that minimizes $D_{F^*}(P||Q) = D_F(Q||P)$ for $Q \in \mathcal{R} \subset \mathcal{X}$. $P_L{}^{\perp}$ is called the Bregman projection[13] and $P_R{}^{\perp}$ the dual Bregman projection.

We have the following projection theorem [43], [10] illustrated in Figure 5:

Theorem 2.2 ([43], [10]): When $\mathcal{R}$ is convex, $P_R{}^{\perp}$ is unique and the dual geodesic Γ^* connecting P to $P_R{}^{\perp}$ is orthogonal to the boundary of $\mathcal{R}$. Similarly, when $\mathcal{R}$ is dual convex, $P_L{}^{\perp}$ is unique and the geodesic Γ connecting P to $P_L{}^{\perp}$ is orthogonal to the boundary of $\mathcal{R}$.

[13]In information geometry [43], $P_R{}^{\perp}$ is called the reverse I-projection or the dual geodesic projection. Dually, $P_L{}^{\perp}$ is called the I-projection or geodesic projection.

F. Geometry of symmetrized Bregman divergences

As mentionned in the introduction, the symmetrized Bregman divergence S_F is typically *not* a Bregman divergence[14] because the convexity argument may fail as reported in [25]. Therefore the underlying geometry of symmetrized Bregman divergence does not fit the dually flat manifolds presented above. However, the symmetrized Bregman divergence can be interpreted using the framework of Csiszár f-divergence [34] (also called Ali-Silvey divergence [47]). In particular the geometry implied by the symmetrized Kullback-Leibler divergence is *not* flat anymore [48], [44]. We refer to the work of Vos [48] for explanations.

We now turn to the study of sided and symmetrized Bregman centroids. In the remainder, we consider computing either in the θ or η coordinate system. It shall be clear that all following results may be dually interpreted using the coupled dual coordinate system or the dual Legendre convex conjugate.

III. The Sided Bregman centroid

A. 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 [43], it was again recently brought up to the attention of the machine learning community by Banerjee et al. [19] who proved that Lloyd's iterative k-means "centroid" clustering algorithm [15] generalizes to the class of Bregman divergences. We state the result and give the proof for completeness and familiarizing us with notations.

Theorem 3.1: 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 characterizing for the relative entropy the notion introduced by Sibson [38] for probability measures), is for a point set $\mathcal{P} \subset \mathcal{X}$:

$$\begin{aligned}
\mathcal{P} &= \{p_1, ..., p_n\} \subset \mathbb{R}^d \mapsto \mathbb{R}^+ \\
\mathrm{JS}_F(\mathcal{P}) &= \frac{1}{n}\sum_{i=1}^{n} F(p_i) - F(\bar{p}) \geq 0,
\end{aligned}$$

which bears the name of the F-Jensen difference[15] [40]. For $F(x) = -H(x) = \sum_{i=1}^d x^{(i)} \log x^{(i)}$ the negative Shannon entropy, JS_F is known as the Jensen-Shannon divergence [39]:

$$\mathrm{JS}(\mathcal{P}) = H(\frac{1}{n}\sum_{i=1}^{n} p_i) - \sum_{i=1}^{n} \frac{1}{n} H(p_i).$$

For a multinomial distribution with d outcomes, the Shannon entropy can also be interpreted as an index of *diversity* [40] of the distribution. The Jensen difference $\mathrm{JS}(p; q) = H(\frac{p+q}{2}) - \frac{H(p)+H(q)}{2}$ is therefore a difference of diversity: Namely, the diversity of the mixed distribution $\frac{p+q}{2}$ minus the average diversity of the source distributions. Following Burbea and Rao [40], the Jensen-Shannon divergence can naturally be extended to a mixture of n distributions with a vector of *a priori* weights w as follows:

$$\mathrm{JS}(\mathcal{P}, w) = H(\sum_{i=1}^{n} w_i p_i) \sum_{i=1}^{n} w_i H(p_i).$$

[14] Besides the class of symmetric quadratic distances that also bears the name of Mahalanobis distances [25].

[15] In the paper [40], it is used for strictly concave function $H(x) = -F(x)$ 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 $\pi_i = \frac{1}{n}$).

It follows from the concavity of Shannon entropy H that $\mathrm{JS}(\mathcal{P}, w) \geq 0$. This generalized Jensen difference is the same as the mutual information [40]. See also the related definition of Jensen-Tsallis divergence [49] for nonextensive Tsallis entropies. Thus the minimization score of the right-sided Bregman centroid is the information radius of the population, a measure of diversity. Note that the information radius is always bounded. Banerjee et al. [19] called the information radius the *Bregman information* (and the sided centroids the best Bregman representatives). It is remarkable to notice that for the squared generator, the information radius turns out to be the sample variance $\frac{1}{n}\sum_{i=1}^{n}||p_i - c_R^F|| = \frac{1}{n}\sum_{i=1}^{n}||p_i - \bar{p}||$. For the Kullback-Leibler Bregman divergence, the information radius can be interpreted as the mutual information [19] p. 1711.

The information retrieval criterion $\mathrm{JS}(P;Q)$ is continuously connected with the classical statistical Bayesian criterion $e(P;Q)$ as shown by Liese and Vajda [6] using the notion of Arimoto entropies [50], [51], where $e(P;Q)$ denote the error of the Bayesian identification of an object from the set of two objects having distributions P and Q.

B. Dual divergence and left-type centroid

Using the Legendre convex conjugation twice, we get the following (dual) theorem for the left-sided Bregman centroid:

Theorem 3.2: The left-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

$$c_R^{F^*}(\mathcal{P}_F{}') = \frac{1}{n}\sum_{i=1}^{n}\nabla F(p_i).$$

That is, we get

$$c_L^F = (\nabla F)^{-1}(\sum_{i=1}^{n}\frac{1}{n}\nabla F(p_i)) = (\nabla F)^{-1}(\bar{p'}).$$

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 3.3: 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.

C. Centers and barycenters as generalized means

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[16] that generalizes the usual arithmetic and geometric means. For a *strictly continuous* and *monotonous* function f, the *generalized mean* [52], [12], [8] of a sequence $\mathcal{V}$ of n real positive numbers $V = \{v_1, ..., v_n\}$ is defined as

$$M_f(\mathcal{V}) = f^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}f(v_i)\right).$$

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 Table IV). 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 $\max_i v_i$:

$$\min_{i\in\{1,...,n\}} x_i \leq M_f(\mathcal{V}) \leq \max_{i\in\{1,...,n\}} x_i.$$

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 $p \to \pm\infty$.

Generalized means can be extended to weighted means using an *a priori* normalized weight vector w (with $\forall i, w_i \geq 0$ and $\sum_{i=1}^{n} w_i = 1$):

$$M_f(\mathcal{V}; w) = f^{-1}\left(\sum_{i=1}^{n} w_i f(v_i)\right).$$

By default, we consider the uniform distribution so that $w_i = \frac{1}{n}\forall i \in \{1, ..., n\}$. These means can also be naturally extended to d-dimensional positive vectors $\mathcal{P} = \{p_1, ..., p_n\}$ (with $\forall i, p_i \in (\mathbb{R}_+)^d$) following the Eq. 10. For example, the arithmetic mean of a set of positive vector points $\mathcal{P}$ (obtained with generator $f(x) = Ix = x$, where I is the $d \times d$ identity matrix) is its center of mass:

$$M_f(\mathcal{P}) = \frac{1}{n}\sum_{i=1}^{d} p_i.$$

(In fact, choosing $f(x) = Qx$ for any positive-definite matrix Q yields the center of mass.) In the remainder, we consider generalized means on vectors although these notions have been interestingly extended to a broader setting like matrices. See for example the axiomatic approach of Petz and Temesi [8] that defines means[17] on matrices using the framework of operator means via operator monotone functions.

[16] Studied independently in 1930 by Kolmogorov and Nagumo, see [52]. A more detailed account is given in [53], Chapter 3.

[17] Following [8], the geometric mean of two positive matrices A and B is found as $A^{\frac{1}{2}}(A^{-\frac{1}{2}}BA^{-\frac{1}{2}})^{\frac{1}{2}}A^{\frac{1}{2}}$.

These generalized (vector) 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 [25]. This affine invariant property *transposes* to generalized means as an offset/scaling invariant property:

$$M_f(\mathcal{P}) = M_{Af+b}(\mathcal{P}),$$

for any invertible matrix A and vector b.

Although we have considered centroids for simplicity to show the relationship between Bregman centroids and generalized means (ie., uniform weight distribution on the input set $\mathcal{P}$), our 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 3.4: Bregman divergences are in bijection with generalized means. The right-sided barycenter $r^F(w)$ is independent of F and computed as the weighted arithmetic mean on the vector point set, a generalized mean for the identity function: $r^F(\mathcal{P}; w) = r(\mathcal{P}; w) = M_x(\mathcal{P}; w)$ with $M_f(\mathcal{P}; w) = f^{-1}(\sum_{i=1}^n w_i f(v_i))$. The left-sided Bregman barycenter $L^F(w)$ is computed as a generalized mean on the point set for the gradient function ∇F: $L^F(\mathcal{P}) = M_{\nabla F}(\mathcal{P}; w)$. The information radius of sided Bregman barycenters is defined by the *Jensen divergence* of the mixture of vectors: $\mathrm{BR}_F(\mathcal{P}; w) = \sum_{i=1}^d w_i F(p_i) - F(\sum_{i=1}^d w_i p_i)$.

The seminal paper of Burbea and Rao [40] considered multinomial distributions in d-dimensional real vector spaces where a J-divergence measure is by means of an arbitrary *separable* entropic function (Eq. 13 of [40]). It is interesting to note that Rényi [54] also made use of generalized means for defining entropies H_α of order α.

A (weighted) mean is said *homogeneous* if and only if we have for any non-negative scalar factor $\lambda \geq 0$:

$$M_f(\lambda\mathcal{P}; w) = \lambda M_f(\mathcal{P}; w).$$

It is well-known [53], [12] that a generalized mean is homogeneous (or linear scale free) if and only if the generator function f belongs to the family $\{f_\alpha\}_\alpha$ (for $\alpha \in \mathbb{R}$) of functions defined by:

$$f_\alpha(x) = \begin{cases} x^{\frac{1-\alpha}{2}} & \alpha \neq 1, \\ \log x & \alpha = 1 \end{cases}$$

For $\alpha = 1$, we get $f_1(x) = \log x$. This function is modulo a constant the f-means related to the Kullback-Leibler divergence, since we have:

$$(x \log x)' \equiv_{ax+b} \log x.$$

D. Dominance relationships of sided centroid coordinates

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

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 3.5: 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

$$\begin{aligned} c_L^F &= (\nabla F)^{-1}(\frac{1}{n}\sum_{i=1}^n \nabla F(p_i)), && (10) \\ &\leq (\nabla F)^{-1}(\nabla F(\frac{1}{n}\sum_{i=1}^n p_i)), && (11) \\ &= \frac{1}{n}\sum_{i=1}^n p_i = c_R^F. && (12) \end{aligned}$$

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.

IV. Symmetrized Bregman centroid

A. Revisiting the optimization problem

For asymmetric Bregman divergences, the symmetrized Bregman centroid is defined by the following optimization problem

$$\begin{aligned} c^F &= \arg\min_{c \in \mathcal{X}} \sum_{i=1}^n \frac{D_F(c||p_i) + D_F(p_i||c)}{2}, \\ &= \arg\min_{c \in \mathcal{X}} \mathrm{AVG}(\mathcal{P}; c). \end{aligned}$$

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

Lemma 4.1: The symmetrized Bregman centroid c^F is unique and obtained by minimizing $\min_{q \in \mathcal{X}} D_F(c_R^F||q) + D_F(q||c_L^F)$: $c^F = \arg\min_{q \in \mathcal{X}} D_F(c_R^F||q) + D_F(q||c_L^F)$.

Bregman divergence D_F (entropy/loss function 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 $\frac{n}{\sum_{j=1}^n \frac{1}{x_j}}$

TABLE IV
BIJECTION BETWEEN BREGMAN DIVERGENCES AND GENERALIZED f-MEANS EXPLICITED FOR THE PYTHAGORAS' MEANS.

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

$$\begin{aligned} \mathrm{AVG}_F(q||\mathcal{P}) &= \mathrm{AVG}_{F^*}({\mathcal{P}_F}'||q'), \\ &= (\sum_{i=1}^n \frac{1}{n} F^*(p_i') - F^*(\bar{p'})) + D_{F^*}(\bar{p_F'}||q_F'). \end{aligned}$$

But

$$\begin{aligned} D_{F^*}(\bar{p_F'}||q_F') &= D_{F^{**}}(\nabla F^* \circ \nabla F(q)||\nabla F^*(\sum_{i=1}^n \nabla F(p_i))), \\ &= D_F(q||c_L^F), \end{aligned}$$

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. ■

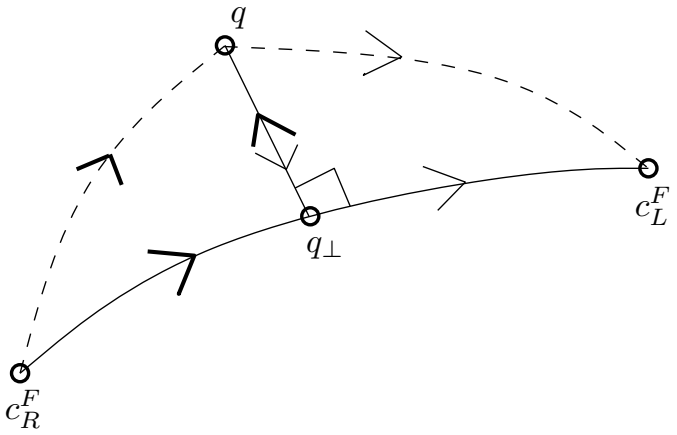

Fig. 7. The symmetrized Bregman centroid necessarily lies on the geodesic passing through the two sided centroids c_R^F and c_L^F.

B. Geometric characterization

We now characterize the exact geometric location of the symmetrized Bregman centroid by introducing a new type of bisector[18] called the mixed-type bisector:

Theorem 4.2: 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 [25]:

$$q_\perp = \arg \min_{t \in \Gamma(c_R^F, c_L^F)} D_F(t||q)$$

[18]See [25] for the affine/curved and symmetrized bisectors studied in the context of Bregman Voronoi diagrams.

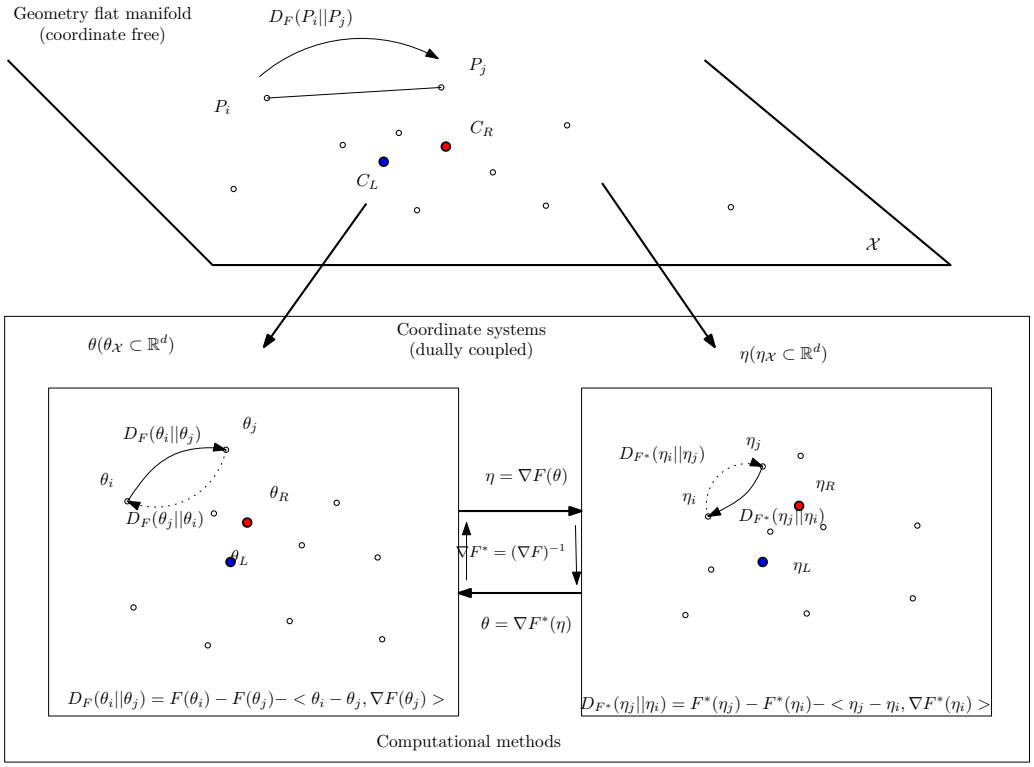

Fig. 6. Interpretation of the sided Bregman centroids on the dually flat manifold.

as depicted in Figure 7. Using *Bregman Pythagoras' theorem*[19] twice (see [25]), 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. ■

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:

Corollary 4.3: 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 8 displays the mixed-type bisector, and sided and symmetrized Bregman centroids for the extended[20] Kullback-Leibler (eKL) and Itakura-Saito (IS) divergences.

[19]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 [19].

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

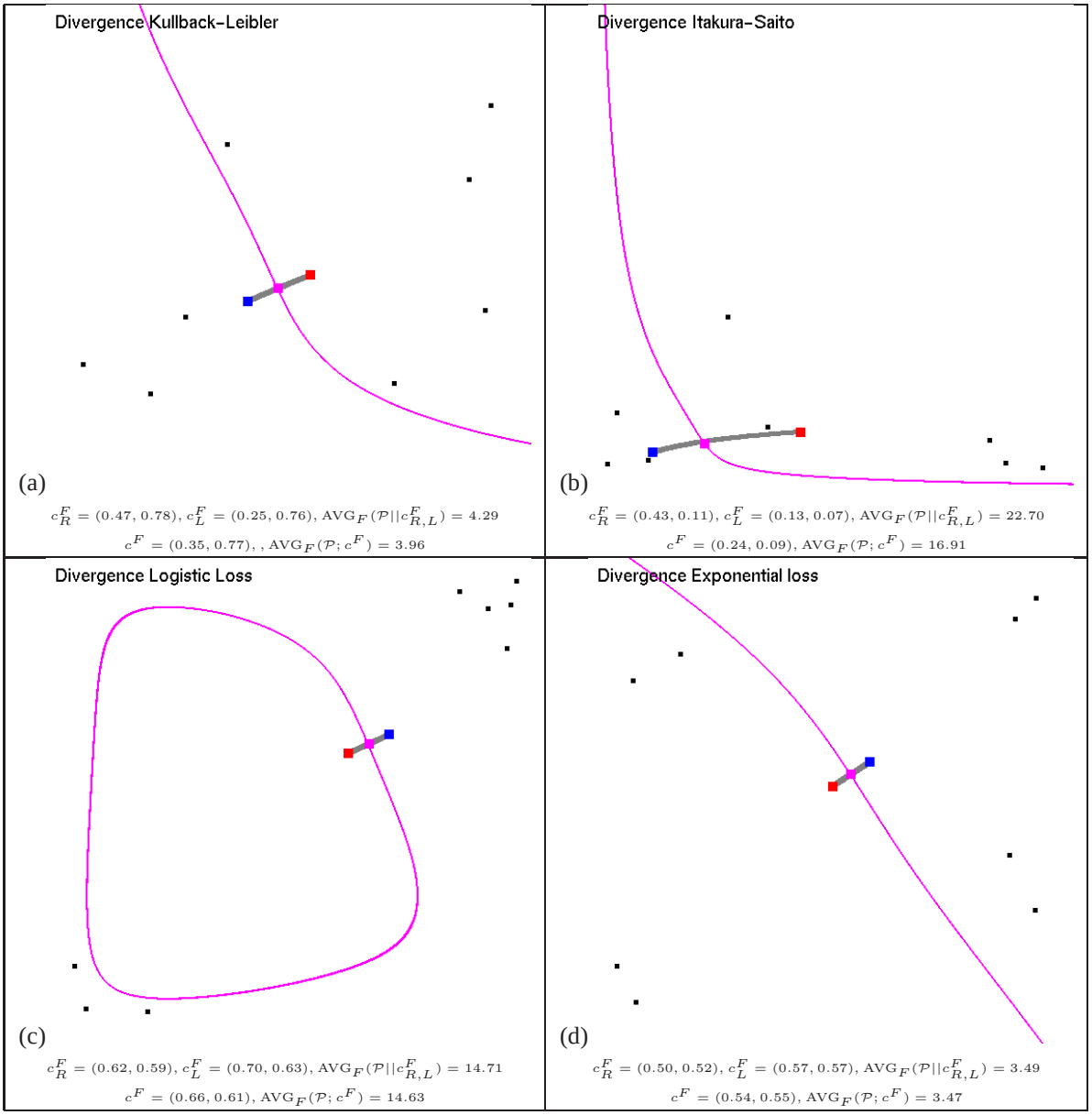

Fig. 8. Bregman centroids for (a) the extended Kullback-Leibler, (b) Itakura-Saito, (c) Logistic , and (d) exponential losses (divergences) on the open square $\mathcal{X} =]0,1[^2$. Right- and left-sided, and symmetrized centroids are displayed respectively as thick red, blue and purple points. The geodesic linking the right-type centroid to the left-type one is shown in grey, and the mixed-type bisector is displayed in purple.

C. 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 (bisection search) 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[21] of Veldhuis [36]: 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

$$D_F(c_R^F||q_h) - D_F(q_h||c_L^R),$$

- **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

[21] Veldhuis' method [36] 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. [41] set the number of steps of the outer loop and inner loops to ten and five iterations each, respectively. Appendix A provides a synopsis of Veldhuis' method.

experimentally by Veldhuis [36] for the Kullback-Leibler divergence that this midpoint provides "experimentally" a good approximation, let us emphasize that is *not true* in general, as depicted in Figure 8(b) for the Itakura-Saito divergence.

Theorem 4.4: 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 [55] (See Lemma 3 of [55]).

V. Applications of the dichotomic geodesic-walk algorithm

A. Bregman power symmetrized divergences

In sound processing, the Itakura-Saito divergence is often used as the *de facto* distortion measure for comparing two spectra envelopes [29]. That is, a set of discrete all-pole model coefficients are first extracted so that the distance between any two sound spectra is later measured at the harmonic peaks $x^{(i)}$, for $i \in \{1, ..., d\}$ — see [29]. It turns out that the Itakura-Saito divergence on d-dimensional real-valued probability vectors:

$$\mathrm{IS}(p||q) = \sum_{i=1}^{d} \left(\frac{p^{(i)}}{q^{(i)}} - \log \frac{p^{(i)}}{q^{(i)}} - 1 \right) = D_F(p||q),$$

is yet another separable Bregman divergence in disguise obtained for the strictly convex generator function

$$F(x) = -\sum_{i=1}^{d} \log x^{(i)},$$

where function $F(x)$ is commonly called the Burg entropy. Wei and Gibson [29] showed that the least-mean square on the COSH distance:

$$\mathrm{COSH}(p;q) = \frac{\mathrm{IS}(p||q) + \mathrm{IS}(q||p)}{2},$$

the symmetrized Itakura-Saito divergence, yields better[22] and smoother discrete all-pole spectral modeling results than by using the Itakura-Saito divergence. Moreover, in some applications such as in concatenative speech synthesis, the COSH distance is considered for minimizing artifacts in speech diphone synthesis. However, one may also consider alternatively the symmetrized Kullback-Leibler distance for the same task by choosing different feature extractors [26]. Interestingly, both the Itakura-Saito and the Kullback-Leibler divergences can be *encapsulated* into a common parameterized family of distortions measures D_{F_α}, generated by the following set of strictly convex and differentiable *power function* generators:

$$F_\alpha : \mathcal{X} \subset (\mathbb{R}_*^+)^d \mapsto \mathbb{R}^+$$

$$F_\alpha(x) = \begin{cases} \sum_{i=1}^d x^{(i)} - \log x^{(i)} - 1 & \alpha = 0 \\ \sum_{i=1}^d \frac{1}{\alpha(1-\alpha)}(-(x^{(i)})^\alpha + \alpha^{(i)} - \alpha + 1) & \alpha \in (0,1) \\ \sum_{i=1}^d x^{(i)} \log x^{(i)} - x^{(i)} + 1 & \alpha = 1 \end{cases}$$

That family of power generators F_α (with F_0 and F_1 the limits for $\alpha \to 0$ and $\alpha \to 1$) yields the corresponding family of *Bregman power divergences* D_{F_α} for real-valued d-dimensional probability vectors p and q:

$$\begin{aligned} D_{F_0}(p||q) &= \sum_{i=1}^{d} \left(\log \frac{q^{(i)}}{p^{(i)}} + \frac{p^{(i)}}{q^{(i)}} - 1 \right), \\ &= \sum_{i=1}^{d} \left(\frac{p^{(i)}}{q^{(i)}} - \log \frac{p^{(i)}}{q^{(i)}} \right) - d. \end{aligned}$$

$$D_{F_\alpha}(p||q) = \frac{1}{\alpha(1-\alpha)} \sum_{i=1}^{d} \left((q^{(i)})^\alpha - (p^{(i)})^\alpha + \alpha (q^{(i)})^{\alpha-1}(p^{(i)} - q^{(i)}) \right),$$

for $\alpha \in (0,1)$.

$$D_{F_1}(p||q) = \sum_{i=1}^{d} p^{(i)} \log \frac{p^{(i)}}{q^{(i)}} - p^{(i)} + q^{(i)}.$$

The Itakura-Saito (D_{F_0}) and extended[23] Kullback-Leibler (D_{F_1}) divergences represent the two extremities of the generic family that is *axiomatically* justificated as the notion of projection in least-mean square problems [35]. This parametric family of Bregman divergences D_{F_α} are the symmetrized Bregman-Csiszár power divergence is defined

$$S_{F_\alpha}(p;q) = \frac{D_{F_\alpha}(p||q) + D_{F_\alpha}(q||p)}{2},$$

$$S_{F_\alpha}(p;q) = \sum_{i=1}^{d} \frac{1}{1-\alpha} ({q^{(i)}}^{\alpha-1}(p^{(i)} - q^{(i)}) + {p^{(i)}}^{\alpha-1}(q^{(i)} - p^{(i)})),$$

[22] Refer to Fig. 2 and Fig. 3 of [29]. It is said that "...the COSH distance measure is the best criterion measure..." (*dixit*)

[23] Defined over the positive orthant of unnormalized probability density functions. Considering the extended Kullback-Leibler measure makes a huge difference from the practical point of view since the left-type centroid C_L^F always falls inside the domain. This is not anymore true if we consider the probability $(d-1)$-dimensional probability simplex $\mathcal{S}^d$ where the left-type centroid C_L^F falls outside $\mathcal{S}^d$, and need to the projected back onto $\mathcal{S}^d$ using a Kullback-Leibler (Bregman) projection. See Pelletier [56] for details. We show how to bypass this problem in the next section by considering discrete distribution as multinomials with $d-1$ degrees of freedom.

for $0 < \alpha < 1$. Since our generic symmetrized Bregman centroid procedure allows to compute the centroid for any Bregman divergence, we can also obviously apply it for this important parameterized family. This is all the more important for distance learning algorithms [57] that seek for the best distance representative (ie., the best α value) to perform[24] a given task. Note that except for the class of generalized quadratic distance with generators $F_Q(x) = x^T Qx$ for a positive definite matrix $Q \succ 0$, the symmetrized Bregman divergences are not of Bregman type [25], [32].

We now consider parameteric family of distributions which admit a canonical decomposition of their probability density functions. We start from the non-parametric probability mass functions that are in fact parametric multinomials in disguise.

Historically, Read and Cressie [4], [6] considered that family of power generators for studying properties of the corresponding family of Csiszár's $I_{F_\alpha}(p||q)$ *power divergences* of order $\alpha \in \mathbb{R}$. Lafferty [58] investigated the Legendra transform properties of these Bregman power divergences D_{F_α}. Csiszár [35] proved that these divergences arise naturally from axiomatic characterizations (Eq. (3.7) of [35]). Notice that Csiszár and Bregman power divergences differ unless $\alpha = 1$, the Kullback-Leibler divergence.

B. 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[25]. 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 an index $i \in \{1, ..., d\}$ such that $q^{(i)} = 0$ and $p^{(i)}$ is non-zero. 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 [25] 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 [25] in statistics for which have the important property that

$$\mathrm{KL}(p(\theta_p)||q(\theta_q)) = D_F(\theta_q||\theta_p),$$

see [25]. 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 [25], it comes out that the natural parameters associated with the sufficient statistics $t(x)$ are

$$\begin{aligned} \theta^{(i)} &= \log \frac{q^{(i)}}{q^{(d)}}, \\ &= \log \frac{q^{(i)}}{1 - \sum_{j=1}^{d-1} q^{(j)}} \end{aligned}$$

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 $x_1, ..., x_n$. 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 [59]. 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^* are dual convex conjugates obtained by the Legendre transformation that maps both domains and functions:

$$(\mathcal{X}_F, F) \longleftrightarrow (\mathcal{X}_{F*}, F^*).$$

[24]Being more efficient while keeping accuracy is a key issue of search engines as mentioned in the introduction.

[25]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.

It follows by construction from the Legendre transformation that the gradients of these F and F^* functions are *reciprocal* to each other:

$$\nabla F^* = \nabla F^{-1}, \qquad \nabla F = (\nabla F^*)^{-1}.$$

This yields one method to deduce the convex conjugate F^* from the gradient ∇F as the integral primitive of the inverse of the gradient of F:

$$F^* = \int (\nabla F)^{-1},$$

We get the inverse $(\nabla F)^{-1}$ of the gradient ∇F as

$$\begin{aligned}(\nabla F)^{-1}(\eta) &= \left(\log \frac{\eta^{(i)}}{1-\sum_{j=1}^{d-1}\eta^{(j)}}\right)_i, \\ &= \theta.\end{aligned}$$

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[26] function

$$F^*(\eta) = \eta\log\eta + (1-\eta)\log(1-\eta).$$

Further, reinterpreting F^* as the log normalizer of an exponential family distribution, we get the Dirichlet distribution, which is precisely the *conjugate prior* [60] of multinomial distributions used in prior-posterior Bayesian updating estimation procedures. We summarize the chain of duality as follows:

$$\begin{aligned}\mathrm{KL}(p^F||q^F) = D_F(\theta_q||\theta_p) = \\ D_{F^*}(\eta_p||\eta_q) = \mathrm{KL}(q^{F^*}||p^{F^*}),\end{aligned}$$

where p^F indicate that the density function p^F follows the distribution of the exponential family $\mathcal{E}_F$ with log normalizer F.

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:

$$q^{(d)} = \frac{1}{1+\sum_{j=1}^{d-1}(\exp\theta^{(j)})}$$

and

$$q^{(i)} = \frac{\exp\theta^{(i)}}{\sum_{j=1}^{d-1}(1+\exp\theta^{(j)})}$$

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, $q^{(d)}$.

[26]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*.

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

$$F(\theta) = \log(1+\sum_{i=1}^{d-1}\exp\theta^{(i)}).$$

The algorithm is summarized in Figure 9. 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 [36] that "... the assumption that the SKL centroid is a linear combination of the arithmetic and normalized geometric mean must be rejected." Appendix A displays Veldhuis' method for reference.

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.

The plots of Figure 10 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 *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).

Jensen-Shannon divergence (Table II) does not only play an important role in image processing. In fact, it is also related to some prominent approaches to supervised classification throughout its continuous connection with classification-calibrated surrogates [61]. More precisely, we have [6]:

$$\mathrm{JS}(p;q) = \lim_{\alpha\to 1} D_{G_\alpha}(p;q)\ , \tag{13}$$

with:

$$G_\alpha(x) = \frac{\left(x^{1/\alpha}+1\right)^\alpha - 2^\alpha}{2(1-\alpha)}\ , x \in \mathbb{R}_{+*}, \alpha \in (0,1) \tag{14}$$

Bregman Divergences D_{G_α} are called Arimoto divergences. Most notably, we have in addition to (13):

$$G_0 = \lim_{\alpha\to 0} G_\alpha = \max\left\{0, \frac{x-1}{2}\right\}\ , \tag{15}$$

$$G_{1/2} = \sqrt{1+x^2} - \sqrt{2}\ . \tag{16}$$

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
$\forall i\ q_i^{(d)} = \frac{1}{1+\sum_{j=1}^{d-1} \exp \theta_i^{(j)}}$
$\forall i\ \forall k\ q_i^{(k)} = \frac{\exp \theta_i^{(k)}}{1+\sum_{j=1}^{d-1} \exp \theta_i^{(j)}}$

Fig. 9. Synopsis of our symmetrized Kullback-Leibler centroid for discrete distributions. The algorithm first converts the probability mass functions into multinomials of the exponential families, and then perform a dichotomic walk on the geodesic linking the sided Kullback-Leibler centroids.

Since Bregman divergences are not affected by linear terms, one can replace (15) and (16) respectively by $G'_0 = G_0 + (1-x)/2$ and $G'_{1/2} = G_{1/2} - x + \sqrt{2}$ while guaranteeing $D_{G_0} = D_{G'_0}$ and $D_{G_{1/2}} = D_{G'_{1/2}}$. These two new generators are remarkable: the former leads to Hinge loss, while the latter brings Matsushita's loss [61], two *classification calibrated* surrogates, functions that carry appealing properties for supervised learning [62]. Moreover, throughout a duality between real-valued classification and density estimation which calls to the Arimoto divergence and convex duality [61], the first one becomes the popular empirical risk, while the second becomes Schapire-Singer's renown Z criterion for boosting pioneered by Matsushita [61], [63]. Thus, Arimoto divergences make a continuous connection between Jensen-Shannon divergence and the empirical risk, throughout classification calibrated surrogates. Without going in depth, this is interesting as any Bregman (symmetrized) centroid defines, from the classification standpoint, some optimal constant estimation of class labels for a huge set of proper scoring rules [61].

C. 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

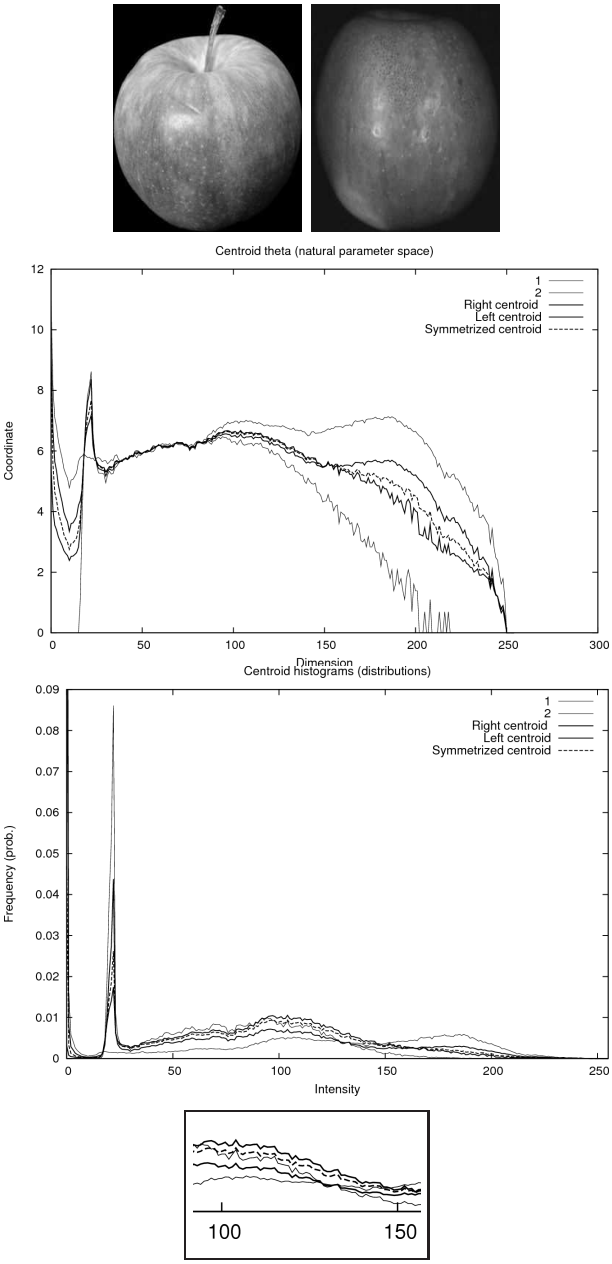

Fig. 10. Centroids of image histograms with respect to the relative entropy. The symmetrized centroid distribution is above both source distributions for intensity range $[100-145]$, but this is never the case in the natural parameter space.

Σ is given by $\Pr(X = x) = p(x; \mu, \Sigma)$ with:

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{\det \Sigma}} \exp\left(-\frac{(x-\mu)^T \Sigma^{-1} (x-\mu)}{2} \right).$$

It is certainly the engineer's favorite family of distributions that nevertheless becomes intricate 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

$$\begin{aligned}\tilde{\Theta} &= (\theta, \Theta),\\ &= (\Sigma^{-1}\mu, \frac{1}{2}\Sigma^{-1}).\end{aligned}$$

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 [44], [43]). 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:

$$\begin{aligned}\mathrm{KL}(N_p||N_q) &= D_F(\tilde{\Theta}_q||\tilde{\Theta}_p),\\ &= F(\tilde{\Theta}_q) - F(\tilde{\Theta}_p) - < (\tilde{\Theta}_q - \tilde{\Theta}_p), \nabla F(\tilde{\Theta}_p) > .\end{aligned}$$

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 [44] 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 [43]. Yoshizawa and Tanabe [44] carried out computations that yield the dual natural/expectation coordinate systems arising from the canonical decompotion of the density function $p(x; \mu, \Sigma)$:

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

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

$$\begin{aligned}\tilde{H} &= \nabla_{\tilde{\Theta}} F(\tilde{\Theta}),\\ &= \begin{pmatrix} \nabla_{\tilde{\Theta}} F(\theta) \\ \nabla_{\tilde{\Theta}} F(\Theta) \end{pmatrix},\\ &= \begin{pmatrix} \frac{1}{2}\Theta^{-1}\theta \\ -\frac{1}{2}\Theta^{-1} - \frac{1}{4}(\Theta^{-1}\theta)(\Theta^{-1}\theta)^T \end{pmatrix},\\ &= \begin{pmatrix} \mu \\ -(\Sigma + \mu\mu^T) \end{pmatrix}\end{aligned}$$

and

$$\begin{aligned}\tilde{\Theta} &= \nabla_{\tilde{H}} F^*(\tilde{H}),\\ &= \begin{pmatrix} \nabla_{\tilde{H}} F^*(\eta) \\ \nabla_{\tilde{H}} F^*(H) \end{pmatrix},\\ &= \begin{pmatrix} -(H + \eta\eta^T)^{-1}\eta \\ -\frac{1}{2}(H + \eta\eta^T)^{-1} \end{pmatrix},\\ &= \begin{pmatrix} \Sigma^{-1}\mu \\ \frac{1}{2}\Sigma^{-1} \end{pmatrix}.\end{aligned}$$

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 [64], [65]. Nock et al. [66] proposed the framework of mixed Bregman divergences to manipulate implicitly and efficiently symmetrized Bregman centroids by pairs of left/right sided centroids. Myrvoll and Soong [27] 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" [42]. 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.

We report on our implementation for bivariate normal distributions[27] (see Figure 11). 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

$$c_L^F(\mathcal{P}) = (\nabla F)^{-1}(\sum_{i=1}^n \frac{1}{n}\nabla F(p_i))$$

[27]Random multivariate distributions are computed as follows: The mean coordinates μ has independent uniform random distribution in $[0, 1]$, and the variance-covariance matrix Σ is obtained from a Wishart distribution obtained as $\Sigma = AA^T$ where A is a triangular matrix with entries sampled from a standard normal distribution.

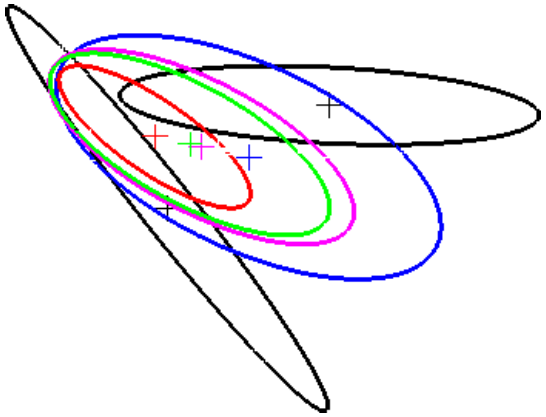

Fig. 11. Entropic sided and symmetrized centroids of bivariate normal distributions. The two input bivariate normals are $m_0 = (0.34029138065736869, 0.26130947813348798)$, $S_0 = \begin{bmatrix} 0.43668091668767117 & -0.42663095837289156 \\ -0.42663095837289161 & 0.63899446830332574) \end{bmatrix}$, and $m_1 = (0.95591075380718404, 0.6544489172032838)$, $S_1 = \begin{bmatrix} 0.79712692342719804 & -0.033060250957646142 \\ -0.033060250957646142 & 0.14609813043797121 \end{bmatrix}$. The right, left and symmetrized centroids are respectively given as $m_R = (0.29050997932657774, 0.53527112890397821)$, $S_R = \begin{bmatrix} 0.33728018979019664 & -0.13844874409795613 \\ -0.13844874409795613 & 0.2321103610207193 \end{bmatrix}$ $m_L^F = (0.64810106723227623, 0.45787919766838603)$, $S_L^F = \begin{bmatrix} 0.71165072320677747 & -0.16933954090511438 \\ -0.16933954090511441 & 0.43118595400867693 \end{bmatrix}$, and $\mathbf{m}^F = (0.42475123207621085, 0.5062178606510539)$, $\mathbf{S}^F = \begin{bmatrix} 0.50780328118070528 & -0.15653432651371618 \\ -0.15653432651371618 & 0.30824860232457035 \end{bmatrix}$. The geodesic half-length bound is found as $m_{\frac{1}{2}} = (0.46930552327942698, 0.49657516328618234)$ with $S_{\frac{1}{2}} = \begin{bmatrix} 0.55643330303588234 & -0.16081280872294987 \\ -0.1608128087229499 & 0.33314553526979185 \end{bmatrix}$. The information radii are 0.83419372149741644 (for the left/right), 0.64099815325721565 (symmetrized) and 0.6525069280087431 (geodesic point with $\lambda = \frac{1}{2}$).

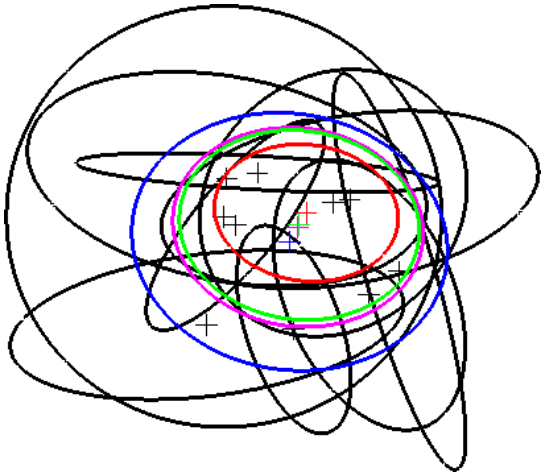

Fig. 12. Entropic centroids for a set of ten bivariate normals: The figure displays the entropic sided and symmetrized centroids (points in 5D shown on the 2D plane using centered ellipsoids). The right-sided centroid, left-sided centroid and symmetrized centroid are rasterized in red, blue and green, respectively. The magenta ellipsoid depicts the point on the geodesic linking the sided centroids for $\lambda = \frac{1}{2}$: This yields a fast approximation of the symmetrized centroid.

coincides with that of the paper [64]. Furthermore, we would like to stress out that our method extends to *arbitrary* entropic centroids of members of the same exponential family.

The Figure 11 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.

VI. Concluding remarks and discussion

In this paper, we have considered and shown that the two sided and the symmetrized Bregman centroids are unique. The right-type centroid is independent of the considered divergence and always coincide with the center of mass of the point set. The left-type centroid is a generalized mean which admits the same Jensen-Shannon information radius as the right-type centroid. The symmetrized Bregman centroid is geometrically characterized as the unique intersection point of the geodesic linking the sided centroids with the mixed-type bisector, and can be approximated efficiently by a simple dichotomic walk. The symmetrized centroid can thus also be interpreted as a generalized mean on the two sided centroids. This work extends straightforwardly to barycenters [56] as well by considering a normalized weight distribution w with $||w|| = 1$. For example, the left-type sided barycenter for weight w is defined as

$$b_L^F(\mathcal{P}; w) = \arg\min_{c \in \mathcal{X}} \frac{1}{n} \sum_{i=1}^{n} w_i D_F(c||p_i),$$

is a ∇F-mean for weight vector w, and has information radius $\mathrm{JS}_F(\mathcal{P}; w)$. Computing the symmetrized Bregman centroids of multinomials (ie., the SKL centroid of histograms, see also [67]) was successfully used for segmenting online music flows [68]. Choosing the most appropriate distortion measure to define a "center" and minimize nearest neighbor queries is an important issue of contents-based multimedia retrieval systems. Spellman et al. [69] carried out preliminary experiments to emphasize on the fact that the MinMax KL center is computationally more efficient than the centroid for nearest neighbor queries. The Bregman-Csiszár one-parameter family of α-divergences may further provide a flexible framework for tuning individually the "appropriate" distance function in each cluster. Note that since the mixture of exponential families is not an exponential family (eg., the family of Gaussian mixtures is not an exponential family), our method does not allow to compute the centroid of Gaussian mixtures [70]. However, since the *product* of exponential families is an exponential family, we can compute the entropic centroids of theses product distributions.

Finally, although Bregman divergences are an important family of information-theoretic distance measures, there are by no means covering the full spectrum of distances. Csiszár f-divergences [34] which includes the Bhattacharyya distance is also another major family of parametrized distances that intersects with the family of Bregman divergences only for the Kullback-Leibler representative. It would be interested to study the properties of f-divergence centroids and barycenters. Amari [12] fully characterized the centroids with respect to α-divergences, a 1-parameter family of Csiszár divergences

parametrized by generators f_α. Namely, Amari proved [12] that the α-means which are the generalized means for the corresponding f_α generator minimize the average sum with respect to the α-divergence. Rigazio et al. [71] presented another work in that direction by approximating the Bhattacharyya centroid of multivariate normals with diagonal covariance matrices using an iterative converging algorithm. The Kullback-Leibler divergence is the only common divergence member of Bregman and Csiszár families. Johnson and Sinanovic [72] presented a symmetric resistor-average distance that does not belong to the family of f-divergences by averaging two Kullback-Leibler distance using an harmonic mean for which it would be interesting to compute the centroid too. Teboulle [65] generalized this Bregman k-means algorithm in 2007 by considering both hard and soft *center-based* clustering algorithms designed for both Bregman [21] and Csiszár f-divergences [47], [34].

Although we have considered in this paper Bregman divergences defined on a space $\mathcal{X} \subset \mathbb{R}^d$, Bregman divergences can also be extended to handle other elements such as Hermitian matrices [3]. See also the work on functional Bregman divergences [24] that extends vector Bregman divergences to measure spaces using Fréchet derivatives. Finally, observe that for any given Bregman divergence $D_F(p||q)$ used on a *finite* vector set $\mathcal{P}$, it is always possible to "metrize" this distortion measure, by first symmetrizing it as $S_F(p;q) = \frac{D_F(p||q)+D_F(q||p)}{2}$, and then finding the largest exponent $\alpha > 0$ such that the triangle inequality on triplets of vectors p_i, p_j and p_k of $\mathcal{P}$ is satisfied:

$$S_F^\alpha(p_i,p_k) \leq S_F^\alpha(p_i,p_j) + S_F^\alpha(p_j,p_k)\ \forall p_i,p_j,p_k \in \mathcal{P}.$$

See [51] for related work on metric divergences.

ACKNOWLEDGEMENTS.

We gratefully thank Professor Lev M. Bregman [21], Professor Marc Teboulle [32], Professor Baba Vemuri [73] and Professor Liese [6] for email correspondences and for sending us their seminal papers. We also thank Guillaume Aradilla for sharing with us his C++ implementation [41] of Velduis' algorithm [36] for comparison. We are indebted to the anonymous referees for their valuable and constructive comments on our previous draft, especially for pointing out to us the referece [12]. We also thank Dr. Sylvain Boltz and Dr. Arshia Cont for informal discussions related to the content of this article. This work was supported in part by the French research agency ANR GAIA under grant 07-BLAN-0328-04 and in part by DIGITEO GAS under grant 2008-16D.

ADDITIONAL MATERIALS

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

APPENDIX

Synopsis of Veldhuis' and the generic geodesic-walk methods

Figure 13 summarizes the Veldhuis' J-divergence centroid convex programming method [36].

Veldhuis' 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:
$\forall k\ \breve{q}^{(k)} = \frac{\tilde{q}^{(k)}}{\sum_{i=1}^d \tilde{q}_i}$ 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:
$\forall k\ c^{(k)} = x^{(k)}\breve{q}^{(k)}\exp\alpha$

Fig. 13. Veldhuis' approximation algorithm for the J-divergence (symmetrized Kullback-Leibler divergence).

REFERENCES

[1] H. H. Bauschke and J. M. Borwein, "Legendre functions and the method of random Bregman projections," *Journal of Convex Analysis*, vol. 4, no. 1, pp. 27–67, 1997.
[2] F. Nielsen and R. Nock, "On the smallest enclosing information disk," *Inf. Process. Lett.*, vol. 105, no. 3, pp. 93–97, 2008.
[3] "Quantum voronoi diagrams and Holevo channel capacity for 1-qubit quantum states," in *IEEE International Symposium on Information Theory (ISIT)*, E.-H. Y. F. Kschischang, Ed. IEEE CS Press, 2008, pp. 96–100.
[4] T. R. C. Read and N. A. C. Cressie, *Goodness-of-Fit Statistics for Discrete Multivariate Data*, ser. Springer Series in Statistics. New York, NY: Springer-Verlag, 1988.
[5] F. Liese and I. Vajda, *Convex Statistical Distances*, 1987.
[6] ——, "On divergences and informations in statistics and information theory," *IEEE Transactions on Information Theory*, vol. 52, no. 10, pp. 4394–4412, 2006.
[7] S.-I. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, no. 2, pp. 251–276, 1998.
[8] D. Petz and R. Temesi, "Means of positive numbers and matrices," *SIAM J. Matrix Anal. Appl.*, vol. 27, no. 3, pp. 712–720, 2005.
[9] M.-M. Deza and E. Deza, *Dictionary of distances*. Elsevier, 2006.
[10] S.-I. Amari, "Information geometry and its applications: Convex function and dually flat manifold," in *Emerging Trends in Visual Computing (ETVC)*, ser. Lecture Notes in Statistics, F. Nielsen, Ed. Springer, 2008.
[11] T. Minka, "Divergence measures and message passing," Microsoft Research, Tech. Rep. MSR-TR-2005-173, 2005.
[12] S.-I. Amari, "Integration of stochastic models by minimizing α-divergence," *Neural Comput.*, vol. 19, no. 10, pp. 2780–2796, 2007.
[13] M. N. Do and M. Vetterli, "Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance," *IEEE Transactions on Image Processing*, vol. 11, no. 2, pp. 146–158, 2002.
[14] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 2006, (Wiley Series in Telecommunications and Signal Processing).

[15] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–136, 1982, first published in 1957 in a Technical Note of Bell Laboratories.

[16] E. Forgy, "Cluster analysis of multivariate data: efficiency versus interpretability of classifications," *Biometrics*, vol. 21, pp. 768–780, 1965.

[17] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms (SODA)*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[18] P. Carmi, S. Har-Peled, and M. J. Katz, "On the fermat-weber center of a convex object," *Comput. Geom.*, vol. 32, no. 3, pp. 188–195, 2005.

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

[20] A. Banerjee, X. Guo, and H. Wang, "On the optimality of conditional expectation as a bregman predictor," *IEEE Transactions on Information Theory*, vol. 51, no. 7, pp. 2664–2669, 2005.

[21] 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*, vol. 7, pp. 200–217, 1967.

[22] Y. A. Censor and S. A. Zenios, *Parallel Optimization: Theory, Algorithms and Applications*. Oxford University Press, 1997.

[23] I. S. Dhillon and J. A. Tropp, "Matrix nearness problems with bregman divergences," *SIAM J. Matrix Anal. Appl.*, vol. 29, no. 4, pp. 1120–1146, 2007.

[24] B. A. Frigyik, S. Srivastava, and M. R. Gupta, "Functional Bregman divergence," in *IEEE International Symposium on Information Theory (ISIT)*, E.-H. Y. F. Kschischang, Ed. IEEE CS Press, 2008, pp. 1681 – 1684.

[25] F. Nielsen, J.-D. Boissonnat, and R. Nock, "Bregman Voronoi diagrams: Properties, algorithms and applications," September 2007, extended abstract appeared in ACM-SIAM SODA 2007. INRIA Technical Report RR-6154.

[26] 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)*. Washington, DC, USA: IEEE Computer Society, 2001, pp. 837–840.

[27] T. A. Myrvoll and F. K. Soong, "On divergence-based clustering of normal distributions and its application to HMM adaptation," in *Proceedings of EuroSpeech 2003*, Geneva, Switzerland, 2003, pp. 1517–1520.

[28] B. A. Carlson and M. A. Clements, "A computationally compact divergence measure for speech processing," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 13, no. 12, pp. 1255–1260, 1991.

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

[30] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Comput.*, pp. 159–195, 2008.

[31] J. M. Winn and C. M. Bishop, "Variational message passing," *Journal of Machine Learning Research*, vol. 6, pp. 661–694, 2005.

[32] A. Ben-Tal, A. Charnes, and M. Teboulle, "Entropic means," *Journal of Mathematical Analysis and Applications*, pp. 537–551, 1989.

[33] M. Basseville and J.-F. Cardoso, "On entropies, divergences and mean values," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Whistler, Ca., September 1995, pp. 330–330.

[34] I. Csiszár, "Information type measures of difference of probability distributions and indirect observations," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 299–318, 1967.

[35] ——, "Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems," *The Annals of Statistics*, vol. 19, no. 4, pp. 2032–2066, 1991, http://projecteuclid.org/.

[36] R. N. J. Veldhuis, "The centroid of the symmetrical Kullback-Leibler distance," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 96–99, March 2002.

[37] R. Veldhuis and E. Klabbers, "On the computation of the Kullback-Leibler measure for spectral distances," in *IEEE transactions on speech and audio processing*, vol. 11, no. 1, 2003, pp. 100–103.

[38] R. Sibson, "Information radius," *Probability Theory and Related Fields*, vol. 14, no. 2, pp. 149–160, 1969.

[39] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information Theory (TIT)*, vol. 37, no. 1, pp. 145–151, 1991.

[40] J. Burbea and C. R. Rao, "On the convexity of some divergence measures based on entropy functions," *IEEE Transactions on Information Theory*, vol. 28, no. 3, pp. 489–495, 1982.

[41] 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)*, vol. 4, 2007, pp. 657–660.

[42] K. Shinoda and C. H. Lee, "A structural Bayes approach to speaker adaptation," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 3, pp. 276–287, 2001.

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

[44] 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*, vol. 35, no. 1, pp. 113–137, 1999.

[45] J. Zhang, "Divergence function, duality, and convex analysis," *Neural Comput.*, vol. 16, no. 1, pp. 159–195, 2004.

[46] E. Cartan, *Geometry of Riemannian spaces*. Math Sci Press, 1983.

[47] 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*, vol. 28, no. Series B, pp. 131–142, 1966.

[48] P. Vos, "Geometry of f-divergence," *Annals of the Institute of Statistical Mathematics*, vol. 43, no. 3, pp. 515–537, 1991.

[49] A. F. T. Martins, M. A. T. Figueiredo, P. M. Q. Aguiar, N. A. Smith, and E. P. Xing, "Nonextensive entropic kernels," in *International Conference on Machine Learning (ICML)*, 2008, pp. 640–647.

[50] S. Arimoto, "Information-theoretical considerations on estimation problems," *Information and Control*, vol. 19, no. 3, pp. 181–194, 1971.

[51] F. Osterreicher and I. Vajda, "A new class of metric divergences on probability spaces and its applicability in statistics," *Annals of the Institute of Statistical Mathematics*, vol. 55, no. 3, pp. 639–653, September 2003. [Online]. Available: http://ideas.repec.org/a/spr/aistmt/v55y2003i3p639-653.html

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

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

[54] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Statist. and Prob.*, vol. 1. Berkeley: Univ. Calif. Press, 1961, pp. 547–561.

[55] R. Nock and F. Nielsen, "Fitting the smallest enclosing Bregman ball," in *16th European Conference on Machine Learning (ECML)*, vol. Volume 3720/2005, 2005, pp. 649–656, lecture Notes in Computer Science.

[56] B. Pelletier, "Informative barycentres in statistics," *Annals of the Institute of Statistical Mathematics*, vol. 57, no. 4, pp. 767–780, 2005.

[57] T. Hertz, A. Bar-Hillel, and D. Weinshall, "Learning distance functions for image retrieval," in *Proceedings of IEEEComputer Vision and Pattern Recognition (CVPR)*, vol. 2, 2004, pp. II–570–II–577 Vol.2.

[58] J. Lafferty, "Additive models, boosting, and inference for generalized divergences," in *Proceedings 12th Conference on Computational Learning Theory (COLT)*. New York, NY, USA: ACM Press, 1999, pp. 125–133. [Online]. Available: http://dx.doi.org/10.1145/307400.307422

[59] O. E. Barndorff-Nielsen, *Parametric statistical models and likelihood*, ser. Lecture Notes in Statistics. New York: Springer-Verlag, 1988, vol. 50.

[60] A. Boratynska, "Stability of Bayesian inference in exponential families," *Statistics and Probability Letters*, no. 2, pp. 173–178, December 1997.

[61] R. Nock and F. Nielsen, "Bregman divergences and surrogates for learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009.

[62] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, "Convexity, classification, and risk bounds," *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 138–156, 2006.

[63] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997.

[64] J. V. Davis and I. S. Dhillon, "Differential entropic clustering of multivariate Gaussians." in *Neural Information Processing Systems (NIPS)*, B. Scholkopf, J. Platt, and T. Hoffman, Eds. MIT Press, 2006, pp. 337–344.

[65] M. Teboulle, "A unified continuous optimization framework for center-based clustering methods," *Journal of Machine Learning Research*, vol. 8, pp. 65–102, 2007.

[66] R. Nock, P. Luosto, and J. Kivinen, "Mixed bregman clustering with approximation guarantees," in *European Conference on Machine Learning and Knowledge Discovery in Databases, ECML/PKDD*, 2008, pp. 154–169.
[67] F. Nielsen, P. Piro, and M. Barlaud, "Tailored bregman ball trees for effective nearest neighbors," in *European Workshop on Computational Geometry*, 2009.
[68] A. Cont, "Modeling musical anticipation: From the time of music to the music of time," Ph.D. dissertation, University of Paris 6 and University of California in San Diego, October 2008.
[69] E. Spellman, B. C. Vemuri, and M. Rao, "Using the KL-center for efficient and accurate retrieval of distributions arising from texture images," in *Proceedings IEEEComputer Vision and Pattern Recognition (CVPR)*. Washington, DC, USA: IEEE Computer Society, 2005, pp. 111–116.
[70] J. Goldberger, S. Gordon, and H. Greenspan, "An efficient image similarity measure based on approximations of KL-divergence between two Gaussian mixtures," in *Proceedings IEEE International Conference on Computer Vision (CVPR)*, 2003, pp. 487–493 vol.1.
[71] L. Rigazio, B. Tsakam, and J.-C. Junqua, "An optimal Bhattacharyya centroid algorithm for gaussian clustering with applications in automatic speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, April 2000, pp. 1599–1602.
[72] D. Johnson and S. Sinanovic, "Symmetrizing the Kullback-Leibler distance," 2001, technical report Rice University.
[73] Z. Wang and B. C. Vemuri, "DTI segmentation using an information theoretic tensor dissimilarity measure," *IEEE Transactions on Medical Imaging*, vol. 24, no. 10, pp. 1267–1277, 2005.

PLACE PHOTO HERE

Frank Nielsen received the BSc and MSc degrees from Ecole Normale Superieure (ENS) of Lyon (France) in 1992 and 1994, respectively. He defended his PhD thesis on adaptive computational geometry prepared at INRIA Sophia-Antipolis (France) under the supervision of Professor J.-D. Boissonnat in 1996. As a civil servant of the University of Nice (France), he gave lectures at the engineering schools ESSI and ISIA (Ecole des Mines). In 1997, he served in the army as a scientific member in the computer science laboratory of Ecole Polytechnique. In 1998, he joined Sony Computer Science Laboratories Inc., Tokyo (Japan), as a researcher. His current research interests include geometry, vision, graphics, learning, and optimization.

PLACE PHOTO HERE

Richard Nock received the agronomical engineering degree from the Ecole Nationale Superieure Agronomique de Montpellier, France (1993), the PhD degree in computer science (1998), and an accreditation to lead research (HDR, 2002) from the University of Montpellier II, France. Since 1998, he has been a faculty member at the Universite Antilles-Guyane in Guadeloupe and in Martinique, where his primary research interests include machine learning, data mining, computational complexity, and image processing