Fundamentals of Bregman Divergences

August 8, 2011

1 Introduction

Bregman divergences were introduced by Bregman in [Bre67] and have traditionally been studied by researchers in the optimization and numerical analysis communities. The book [CZ97] contains a detailed overview of Bregman divergences and their application in iterative optimization algorithms. Within machine learning, Bregman divergences first appear in work on iterative scaling [LPP97] and online learning, *e.g.* [GLS97, HW98, AW01] and were later identified with the k-means algorithm [BMDG05].

2 Definition and Motivation

In this section we explain what a bregman divergence is and why they are compelling objects of study. These divergences arise naturally from a surprisingly disparate collection of sources. We consider four. First, they are a key ingredient in iterative optimization schemes for convex optimization. Second, they have a deep connection to the exponential family of probability distributions. Third, they arise out of basic axioms in connection with maximum likelihood estimation. Finally, they satisfy certain mathematical properties that make them a central concept in information geometry.

2.1 Formal Definition

Definition 1. *Let $f : C \subset \mathbb{R}^D \to \mathbb{R}$ be a strictly convex function that is differentiable on the relative interior of C. The bregman divergence based on f is defined as*

$$d_f(x, y) \equiv f(x) - f(y) - \langle \nabla f(y), x - y \rangle.$$

Note that $d_f : C \times \mathrm{relint}(C) \longrightarrow \mathbb{R}_+$.

A bregman divergence measures the distance between f and its first-order taylor expansion. See figure 1. Table 1 lists some common bregman divergences.

It is often assumed that f is not only strictly convex but is also *Legendre*. A function $f : C \to \mathbb{R}$ is Legendre if

Table 1: Standard bregman divergences. S^D_{++} denotes the cone of positive definite $D \times D$ matrices.

Name	**Domain**	**Base function**	$d_f(x, y)$
ℓ_2^2	$\mathbb{R}^D$	$\frac{1}{2}\|x\|_2^2$	$\frac{1}{2}\|x-y\|_2^2$
Mahalanobis ($Q \succ 0$)	$\mathbb{R}^D$	$\frac{1}{2}x^\top Q x$	$\frac{1}{2}(x-y)^\top Q(x-y)$
KL-Divergence	$\mathbb{R}^D_+$	$\sum x_i \log x_i$	$\sum x_i \log \frac{x_i}{y_i} - x_i + y_i$
Itakura-Saito	$\mathbb{R}^D_+$	$-\sum \log x_i$	$\sum \left(\frac{x_i}{y_i} - \log \frac{x_i}{y_i} - 1\right)$
Exponential	$\mathbb{R}^D$	$\sum e^{x_i}$	$\sum e^{x_i} - (x_i - y_i + 1)e^{y_i}$
Bit Entropy	$[0,1]^D$	$\sum(x_i \log x_i + (1-x_i)\log(1-x_i))$	$\sum(x_i \log \frac{x_i}{y_i} + (1-x_i)\log(\frac{1-x_i}{1-y_i}))$
Hellinger	$[-1,1]^D$	$-\sum \sqrt{1-x_i^2}$	$\sum \frac{1-x_i y_i}{\sqrt{1-y_i^2}} - \sqrt{1-x_i^2}$
ℓ_p^p, $p \in [1,\infty]$	$\mathbb{R}^D$	$\|x\|_p^p$	$\sum \|x_i\|^p - p x_i \mathrm{sgn}(y_i)\|y_i\|^{p-1} + (p-1)\|y_i\|^p$
Log-det	S^D_{++}	$-\log\det X$	$\langle X, Y^{-1}\rangle - \log\det XY^{-1} - N$
von Neumann entropy	S^D_{++}	$\mathrm{tr}\,(X \log X - X)$	$\mathrm{tr}\,(X(\log X - \log Y) - X + Y)$

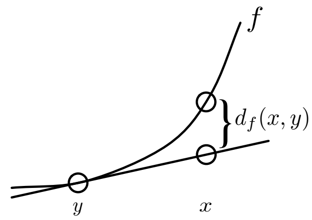

Figure 1: The bregman divergence between x and y.

- $C \neq \emptyset$ and its interior is convex;
- f is strictly convex and has continuous first partial derivatives; and
- if $x_1, x_2, \ldots \in C$ is a sequence converging to a boundary point of C, $\|\nabla f(x_n)\| \to \infty$.

These additional requirements aid in the analysis of certain iterative projection algorithms [CZ97].

The strict convexity of f implies that a bregman divergence is always non-negative and equal to zero only if its two arguments are identical. It is similarly easy to check that a bregman divergences is convex in its first argument, but not necessarily in its second.

2.2 Convex optimizaton

2.3 Exponential families

There is a remarkable connection between bregman divergences and exponential families. We sketch the basic details in this section.

Let ν be a σ-finite measure defined on the Borel subsets of $\mathbb{R}^D$. An exponential family consists of all distributions with probability density functions of the form

$$p_\eta(x) = \exp(\langle \eta, T(x) \rangle - G(\eta))\ \nu(x),$$

where $T(x)$ is a sufficient statistic. We focus on the case where $T(x) = x$; these are called *regular* exponential families. The set of all η such that

$$\int e^{\langle \eta, x \rangle} \nu(dx) < \infty$$

is called the *natural parameter space*. The function $G(\cdot)$ normalizes the probability distribution and is called the *log partition function*. It is equal to

$$G(\eta) = \log \int e^{\langle \eta, x \rangle} \nu(dx).$$

Many standard distributions are examples of regular exponential families. One example is the Bernoulli distribution, with pdf

$$p(x) = \begin{cases} \mu & \text{for } x = 1 \\ 1-\mu & \text{for } x = 0 \end{cases} = \mu^x (1-\mu)^{1-x}.$$

Let $\eta = \log \frac{\mu}{1-\mu}$. Then

$$p(x) = \exp(x \cdot \eta - \log(1 + e^\eta)),$$

so is an exponential family with log partition function $\log(1 + e^\eta)$. Notice that the natural parameter space is all of $\mathbb{R}$, whereas the expectation μ is confined to $[0, 1]$.

Another example is the univariate Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, with pdf

$$p(\omega) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(\omega-\mu)^2/2\sigma^2}.$$

Re-parametrizing with $x \equiv (\omega, \omega^2)$ and $\eta \equiv (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})$ puts $p(x)$ into standard exponential form.

A third example is the Poisson distribution, with pdf

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!},$$

which is defined over $\mathcal{N}$ for $\lambda \in \mathbb{R}_+$. Setting $\nu(x) = \frac{1}{x!}$ and $\eta = \log(\lambda)$ recovers the familiar exponential form.

There are many other examples, including the binomial and exponential distributions, multidimensional Gaussian distribution, and the multinomial distribution.

The log-partition function has many useful properties. Here we need only a couple of essential ones. First, the gradient of $G(\eta)$ is the mean of the corresponding exponential distribution:

$$\nabla G(\eta) = \int x p_\eta(x) \nu(dx).$$

Additionally, the Hessian of $G(\eta)$ is equal to covariance matrix of p_η. Because the gradient is equal to the mean, the gradient can be thought of as providing a link between the natural parameter space and the expectation parameter space. Another fact of the log-partition function is that it is strictly convex (assuming p_η is minimal), and in fact is Legendre [BMDG05]. Thus the convex conjugate is defined, and $\nabla G^*(y)$ provides the inverse mapping from the space of expectation parameters to the space of natural parameters.

Suppose we wish to measure how similar two members of an exponential family are. A natural way to do this is to compute the KL-divergence between them, and this is where the connection to bregman divergences becomes apparent. Let $p(x)$ and $q(x)$ be members of an exponential family with natural parameters η_1 and η_2 respectively. Then the KL-divergence between them is

$$\begin{aligned}\int p(x)\log\frac{p(x)}{q(x)}dx &= \int p(x)\log\frac{\exp(\langle\eta_1,x\rangle - G(\eta_1))}{\exp(\langle\eta_2,x\rangle - G(\eta_2))}dx \\ &= \int p(x)(\langle\eta_1-\eta_2,x\rangle + G(\eta_2) - G(\eta_1))dx \\ &= \langle\nabla G(\eta_1),\eta_1-\eta_2\rangle + G(\eta_2) - G(\eta_1), \quad \text{since } \nabla G(\eta) \text{ is the mean} \\ &= d_G(\eta_2,\eta_1).\end{aligned}$$

Thus the entropy between two members of an exponential family is the bregman divergence based on the log-partition function between their natural parameters. Moreover, we can write the pdf of any regular exponential family in terms of a bregman divergence:

$$\begin{aligned}p_\eta(x) &= \exp(\langle x,\eta\rangle - G(\eta))\nu(x) \\ &= \exp(G^*(\eta') + \langle x-\eta',\eta\rangle - G^*(x) + G^*(x))\nu(x) \\ &= \exp(-d_{G^*}(x,\eta') + G^*(x))\nu(x).\end{aligned}$$

Since η' is the mean of p_η, this equality shows that the pdf is a function of the divergence between x and the mean. Similarly,

$$p_\eta(x) = \exp(-d_G(\eta,x') + G^*(x))\nu(x),$$

where $x' \equiv \nabla G^*(x)$.

The connection between bregman divergences and exponential families goes a bit deeper. Here, we showed that for exponential family, there is a related bregman divergence. In [BMDG05], it is shown that there is a one-to-one correspondence between the regular exponential families and the so-called regular bregman divergences—bregman divergences defined over a restricted class of convex base functions.

3 Geometry

3.1 Pythagorean theorem

The pythagorean theorem is a key property of bregman divergences that governs the geometry of bregman projections. This theorem underlies some of the bregman geometry we develop in chapter **??**, even though it is not explicitly used.

Theorem 1. *Let C be a closed convex set and let $y = \mathrm{argmin}_{y\in C} d_f(y,z)$ be the bregman projection of z onto C. Further suppose that $x \in C$. Then*

$$d_f(x,z) \geq d_f(x,y) + d_f(y,z).$$

Proof. Define $G(x) = d_f(x, y) - d_f(x, z)$. Then

$$\begin{aligned} G(x) &= f(x) - f(y) - \langle \nabla f(y), x - y\rangle - (f(x) - f(z) - \langle \nabla f(z), x - z\rangle) \\ &= f(z) + \langle \nabla f(z), x - z\rangle - f(y) - \langle \nabla f(y), x - y\rangle, \end{aligned}$$

so G is linear in x. Moreover $G(y) = d_f(y, y) - d_f(y, z) = -d_f(y, z)$. Let $\alpha \in [0, 1]$ and $x_\alpha = \alpha x + (1 - \alpha)y$.

$$\begin{aligned} G(x_\alpha) = G(\alpha x + (1 - \alpha)y) &= \alpha G(x) + (1 - \alpha)G(y) \\ &= \alpha(d_f(x, y) - d_f(x, z) + d_f(y, z)) - d_f(y, z). \end{aligned}$$

Also,

$$G(x_\alpha) = d_f(x_\alpha, y) - d_f(x_\alpha, z),$$

so, when $\alpha > 0$,

$$d_f(x, y) + d_f(y, z) - d_f(x, z) = \frac{d_f(y, z) + d_f(x_\alpha, y) - d_f(x_\alpha, z)}{\alpha}.$$

We show that the right hand side is less than or equal to zero. Since $x_\alpha \in C$ and y is the projection of z onto C, $d_f(y, z) \leq d_f(x_\alpha, z)$. Thus we only need to show that $\frac{d_f(x_\alpha, y)}{\alpha} \leq 0$ for some α. We do this by taking the limit as $\alpha \to 0$ from the right:

$$\begin{aligned} \lim_{\alpha\to 0} \frac{d_f(x_\alpha, y)}{\alpha} &= \lim_{\alpha\to 0} \frac{f(\alpha x + (1 - \alpha)y) - f(y) - \langle \nabla f(y), \alpha x - \alpha y\rangle}{\alpha} \\ &= \lim_{\alpha\to 0} \frac{f(\alpha x + (1 - \alpha)y) - f(y)}{\alpha} - \lim_{\alpha\to 0} \frac{\langle \nabla f(y), \alpha x + (1 - \alpha)y - y\rangle}{\alpha} \\ &= \langle \nabla f(y), x - y\rangle - \lim_{\alpha\to 0} \langle \nabla f(y), x - y\rangle \quad \text{(by l'Hôpitals rule)} \\ &= 0. \end{aligned}$$

□

The previous inequality becomes an equality if C is an affine space.

Theorem 2. *Let C be an affine space and let $y = \operatorname{argmin}_{y\in C} d_f(y, z)$ be the bregman projection of z onto C. Let $x \in C$. Then*

$$d_f(x, z) = d_f(x, y) + d_f(y, z).$$

Proof. Let x' be any point in C, and set λ to satisfy $y = \lambda x + (1 - \lambda)x'$. Define $x_\alpha = \alpha x + (1 - \alpha)x'$, and let $G(x) = d_f(x, y) - d_f(x, z)$. Since $G(x_\alpha) = d_f(x_\alpha, y) - d_f(x_\alpha, z)$,

$$\frac{\partial G(x_\alpha)}{\partial \alpha} = \frac{\partial}{\partial \alpha} d_f(x_\alpha, y) - \frac{\partial}{\partial \alpha} d_f(x_\alpha, z). \tag{1}$$

Note that $x_\alpha \in C$; since y is the projection of z onto C and $x_\alpha = y$ when $\alpha = \lambda$, $d_f(x_\alpha, z)$ is minimized at $\alpha = \lambda$. Moreover, $d_f(x_\alpha, y)$ is also minimized at $\alpha = \lambda$ (it is zero there). Thus, both divergences on the right of (1) are minimized at $\alpha = \lambda$, implying $\frac{\partial G(x_\alpha)}{\partial \alpha}\Big|_{\alpha=\lambda} = 0$.

Using the linearity of G (see the proof of theorem 1), we can expand

$$\begin{aligned} G(x_\alpha) &= d_f(x_\alpha, y) - d_f(x_\alpha, z) \\ &= \alpha(d_f(x, y) - d_f(x, z)) + (1-\alpha)(d_f(x', y) - d_f(x', z)). \end{aligned}$$

Differentiating with respect to α, we get

$$\frac{\partial G(x_\alpha)}{\partial \alpha} = d_f(x, y) - d_f(x, z) - d_f(x', y) + d_f(x', z).$$

At $\alpha = \lambda$, $\frac{\partial G(x_\alpha)}{\partial \alpha} = 0$, so

$$d_f(x, y) - d_f(x, z) - d_f(x', y) + d_f(x', z) = 0.$$

This equality is true for all $x' \in C$, so picking $x' = y$, we get

$$d_f(x, z) = d_f(x, y) + d_f(y, z).$$

□

3.2 Conjugacy

Conjugacy is a central duality notion used in convex analysis [Roc70].

Definition 2. *The* convex conjugate *of a function* $f : \mathbb{R}^d \to \mathbb{R}$ *is*

$$f^*(x) = \sup_y \{\langle x, y\rangle - f(y)\}.$$

The definition of a conjugate function immediately yields the *Fenchel-Young inequality*:

$$f(x) + f^*(y) \geq \langle x, y\rangle.$$

If f is differentiable, we can compute f^* by setting the derivative to zero:

$$\begin{aligned} &\nabla_y (\langle x, y\rangle - f(y)) = x - \nabla f(y) = 0 \\ &\implies x = \nabla f(y) \\ &\implies y = (\nabla f)^{-1}(x). \end{aligned}$$

Thus

$$f^*(x) = \langle x, (\nabla f)^{-1}(x)\rangle - f((\nabla f)^{-1}(x)). \qquad (2)$$

A crucial property of the conjugate function is that its gradient is an inverse to the gradient of the original function. To keep notation clean, define $h(x) \equiv \nabla f(x)$. Then, following the above derivation, we have

$$\begin{aligned} f^*(x) &= \langle x, h^{-1}(x)\rangle - f(h^{-1}(x)); \text{ implying that} \\ \nabla f^*(x) &= h^{-1}(x) + \nabla h^{-1}(x) \cdot x - \nabla h^{-1}(x) \nabla f(h^{-1}(x)) \\ &= h^{-1}(x) \quad (\text{since } \nabla f(x) = h(x)). \end{aligned}$$

Thus ∇f and ∇f^* are inverses of one-another.

Let $x' \equiv \nabla f(x)$ and $y' \equiv \nabla f(y)$. Then plugging x' and y' into (2), we get the convenient relations

$$\begin{aligned} f^*(x') &= \langle x', x\rangle - f(x) \quad \text{and} \\ f^*(y') &= \langle y', y\rangle - f(y). \end{aligned}$$

Let us now consider the bregman divergence associated with f in light of the conjugate function. We show that $d_{f^*}(y', x') = d_f(x, y)$.

$$\begin{aligned} d_{f^*}(y', x') &= f^*(y') - f^*(x') - \langle \nabla f^*(x'), y' - x'\rangle \\ &= \langle y', y\rangle - f(y) - \langle x, x'\rangle + f(x) - \langle \nabla f^*(x'), y' - x'\rangle \\ &= f(x) - f(y) - \langle x', x\rangle + \langle y', y\rangle - \langle x, y' - x'\rangle \quad (\text{since } x = \nabla f^*(x')) \\ &= f(x) - f(y) - \langle \nabla f(y), x - y\rangle \\ &= d_f(x, y). \end{aligned}$$

This equality is useful because it permits algorithms designed to optimize over the first argument to work over the second argument. It provides an elegant way to sidestep the asymmetry of bregman divergences. See section **??** for an example.

Another easily derived equality is

$$d_f(x, y) = f(x) + f^*(y') - \langle x, y'\rangle. \tag{3}$$

This equality reveals another interpretation interpretation of a bregman divergence: it is a measurement of the slackness in the Fenchel-Young inequality.

3.3 Centroidal property

Bregman divergences satisfy a centroidal property that makes them amenable to k-means style clustering [BMDG05]. This property states that the average divergence from a point to a set can be decomposed as the divergence from the point to the set's centroid plus the average divergence from the set to the centroid.

Theorem 3. *Let $S \subset \mathbb{R}^d$ be a set of size n and let $d_f(S, x)$ denote $\sum_{s \in S} d_f(s, x)$. Let $\mu = \frac{1}{n}\sum_{s \in S} s$ be the mean of S. Then for all $x \in \mathbb{R}^d$,*

$$d_f(S, x) = d_f(S, \mu) + n d_f(\mu, x).$$

Proof. Expanding the right side, we have

$$\begin{aligned}
&d_f(S,\mu) + n \cdot d_f(\mu, x)\\
&\quad = \sum_s f(s) - nf(\mu) - \sum_s \langle \nabla f(\mu), s-\mu\rangle + nf(\mu) - nf(x) - n\langle \nabla f(x), \mu - x\rangle\\
&\quad = \sum_s (f(s) - f(x)) - \sum_s \langle \nabla f(\mu), s-\mu\rangle - n\langle \nabla f(x), \mu - x\rangle\\
&\quad = \sum_s (f(s) - f(x)) - \left\langle \nabla f(\mu), \sum_s s - n\mu \right\rangle - \langle \nabla f(x), n\mu - nx\rangle\\
&\quad = \sum_s (f(s) - f(x)) - \left\langle \nabla f(x), \sum_s (s - x) \right\rangle\\
&\quad = \sum_s (f(s) - f(x) - \langle \nabla f(x), s - x\rangle)\\
&\quad = d_f(S, x).
\end{aligned}$$

□

3.4 Other technical results

One convenient property of bregman divergences is that they can by composed. Suppose that $\{f_i\}$ are strictly convex functions and $\{\alpha_i\}$ is a set of non-negative numbers. Then $g \equiv \sum \alpha_i f_i$ is also a strictly convex function. Thus we can define a bregman divergence based on g

$$\begin{aligned}
d_g(x,y) &= \sum \alpha_i f_i(x) - \sum \alpha_i f_i(y) - \langle \nabla \left[\sum \alpha_i f_i(y)\right], x - y\rangle\\
&= \sum \alpha_i d_{f_i}(x,y).
\end{aligned}$$

Similarly, one can define a D-dimensional multivariate bregman by summing D one-dimensional bregman divergences coordinate-wise. These properties are useful for working with mixed-type data. For example, a point x might have a histogram component of, say, the first 10 coordinates, and a vector coordinate of the next 5 coordinates. A reasonable way to compute the distance between two such points would be to measure the KL-divergence between the first 10 coordinates, and the ℓ_2^2 distance between the last 5. Any machinery defined for arbitrary bregman divergences, such as the fast NN algorithms developed in this dissertation, will be able to handle this mixed-type divergence as it is itself a bregman divergence.

In the proof of theorem 1, we showed that the difference $G(x) = d_f(x,y) - d_f(x,z)$ is linear in x:

$$\begin{aligned}
G(x) &= f(x) - f(y) - \langle \nabla f(y), x - y\rangle - (f(x) - f(z) - \langle f(z), x - z\rangle)\\
&= f(z) + \langle \nabla f(z), x - z\rangle - f(y) - \langle \nabla f(y), x - y\rangle.
\end{aligned}$$

In particular, this means that the set of points equidistant from y and z lie on a hyperplane.

Theorem 4. *The set of points* $\{x : d_f(x, y) = d_f(x, z)\}$ *is an affine set.*

Bregman divergences do not in general satisfy the triangle inequality, but do satisfy the three point property

$$d_f(x, y) + d_f(y, z) = d_f(x, z) + \langle \nabla f(z) - \nabla f(y), x - y \rangle.$$

Finally, symmetrized bregman divergences have a particularly simple form:

$$\begin{aligned} d_f(x, y) + d_f(y, x) &= f(x) - f(y) - \langle \nabla f(y), x - y \rangle + f(y) - f(x) - \langle \nabla f(x), y - x \rangle \\ &= \langle \nabla f(x) - \nabla f(y), x - y \rangle. \end{aligned}$$

4 Work related to bregman proximity search

This dissertation presents the first algorithms and data structures for bregman proximity search, but there has been some related work. [NBN07] explores the geometric properties of bregman voronoi diagrams. Voronoi diagrams are related to NN search, but do not lead to practical algorithms. [GIM07] contains results on sketching bregman (and other) divergences. Sketching is related to dimensionality reduction, which is the basis for many NN schemes.

We are aware of only one NN speedup scheme for KL-divergences [SV05]. The results in this paper are quite limited: experiments were conducted on only one dataset and the speedup is less than 3x. Moreover, there appears to be a significant technical flaw in the derivation of their data structure. In particular, they cite the pythagorean theorem as an *equality* for projection onto an arbitrary convex set, whereas it is actually an *inequality*.

References

[AW01] Katy Azoury and Manfred Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, 2001.

[BMDG05] Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *JMLR*, Oct 2005.

[Bre67] L.M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

[CZ97] Yair Censor and Stavros Zenios. *Parallel optimization: theory, algorithms, and applications*. Oxford University Press, 1997.

[GIM07] Sudipto Guha, Piotr Indyk, and Andrew McGregor. Sketching information divergences. In *COLT*, 2007.

[GLS97] Adam Grove, Nick Littlestone, and Dale Schuurmans. General convergence results for linear discriminant updates. *Machine Learning*, pages 171–183, 1997.

[HW98] Mark Herbster and Manfred K. Warmuth. Tracking the best regressor. In *COLT*, pages 24–31. ACM Press, 1998.

[LPP97] John Lafferty, Stephen Della Pietra, and Vincent Della Pietra. Statistical learning algorithms based on bregman distances. In *Proceedings of the Canadian Workshop on Information Theory*, 1997.

[NBN07] Frank Nielsen, Jean-Daniel Boissonnat, and Richard Nock. On bregman voronoi diagrams. In *SODA*, pages 746–755, 2007.

[Roc70] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[SV05] Eric Spellman and Baba Vemuri. Efficient shape indexing using an information theoretic representation. In *International Conference on Image and Video Retrieval*, 2005.