Meet the Bregman Divergences

If you've read theoretical papers in machine learning then you've likely seen the term "Bregman divergences" thrown about and might be wondering what they are and what the fuss is about. As with most mathematical topics, the Wikipedia page on Bregman divergences is heavy on formalism and light on context, which is fine as a reference but not ideal if you are reading about something for the first time.

What I hope to do in this post is gently introduce you to the Bregman divergences, point out some of their interesting properties, and highlight one result that I found surprising and I believe is underappreciated. Roughly speaking, the surprising result[1] – due to Banerjee, Gou, and Wang in 2005 – is the following:

> If you have some abstract way of measuring the "distance" between any two points and, for any choice of distribution over points the mean point minimises the average distance to all the others, then your distance measure must be a Bregman divergence.

Interest piqued? Good, let's get started.

A Geometric Look at Squared Euclidean Distance

Most high school students have met at least one member of the Bregman divergence family: the squared Euclidean distance (SED). As the name suggests, this is just the square of the standard Euclidean distance between the two points. That is, given two points n-dimensional points $x, y \in \mathbb{R}^n$, the SED between them is simply:

$$d^2(x,y) := \sum_{i=1}^{n} (x_i - y_i)^2.$$

For the rest of the post, I will use the terms *distance* and *divergence* loosely and interchangeably to mean some non-negative valued function of two points that measures "how far apart" they are. The SED above is clearly a distance in this sense.[2]

Now, by introducing a little notation we can re-express the above distance in a curious way. We will use $\langle x, y \rangle := \sum_{i=1}^{n} x_i \, y_i$ to denote the *inner product* between x and y and use $\|x\| := \sqrt{\langle x, x \rangle}$ to denote the inner product's associated *norm*. Using these definitions, the linearity of inner products, and a little manipulation, we get

$$d^2(x,y) = \|x-y\|^2 = \langle x-y, x-y \rangle = \|x\|^2 - \|y\|^2 - \langle 2y, x-y \rangle.$$

You may rightly ask, "Why on earth would I want to write it like that?" Well, this form of the SED lends itself to a particularly nice geometrical interpretation if we notice that the derivative of $\|y\|^2$ is $2y$. Now, if you squint at the term $\|y\|^2 + \langle 2y, x-y \rangle$ you'll notice it is the value of the tangent line to $\|y\|^2$ at y evaluated at x (it is clearly equal to $\|y\|^2$ when $x = y$ and changes linearly in x as x deviates from y). This means the whole expression is just the difference between the function $f(x) = \|x\|^2$ at x and the value of f's tangent at y evaluated at x. That is,

$$d^2(x, y) = f(x) - \underbrace{(f(y) + \langle \nabla f(y), x - y \rangle)}_{\text{Tangent of } f \text{ at } y}$$

The interactive Figure 1 shows this interpretation in the one dimensional case. The SED between x and y is the vertical distance between the black curve and the red tangent line, where both are evaluated at the point x (shown on the horizontal axis).

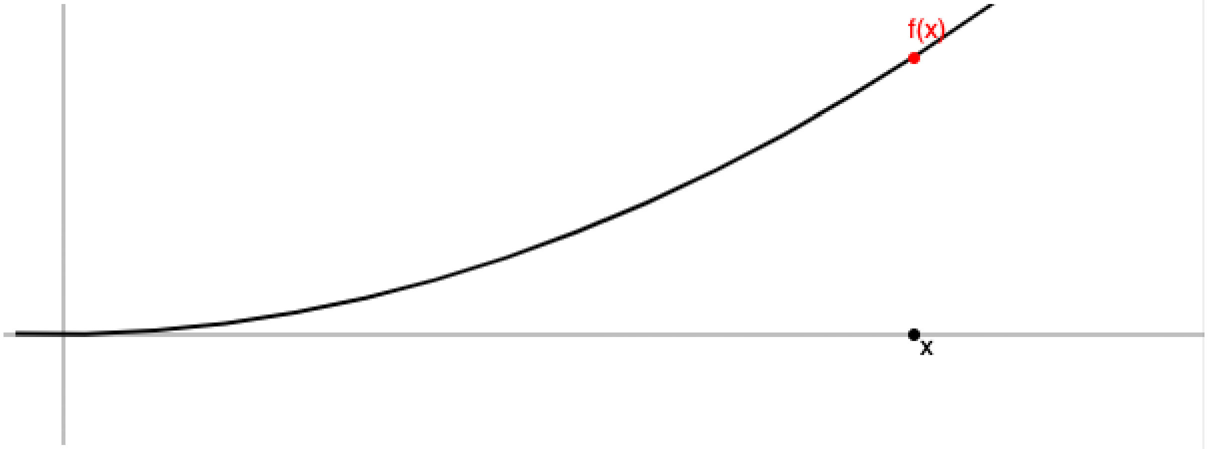

Figure 1: A geometric interpretation of the squared Euclidean distance. **Hover** to move x; **Click** to place a tangent line at y. The black curve is $\|x\|^2$. The red line is the tangent to $\|x\|^2$ at y. The length of the blue line is the value of $d^2(x, y)$.

Convexity *über alles*

So far, the only thing we have wanted from our distance measure d^2 is that it be *non-negative* for all possible choices of points x and y. Viewed geometrically, this is equivalent to the function f always sitting above its tangent. That is, for d^2 to be a "sensible" distance we require that

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle, \quad \text{for all } x, y \in \mathbb{R}^n.$$

However, if you pick up any book on convex analysis (*e.g.*, Boyd & Vandeberghe) you will see results saying that the above condition is equivalent to (suitably differentiable) functions f being *convex*. This means that we can a derive a distance measure d_f that has a similar structure to the squared Euclidean distance by simply choosing a convex function f and defining

$$\boxed{d_f(x, y) := f(x) - f(y) - \langle \nabla f(y), x - y \rangle \geq 0.}$$

Distances defined like this are precisely the **Bregman divergences** and the convexity of f guarantees they are non-negative for all $x, y \in \mathbb{R}^n$. These were first introduced by L.M. Bregman in his 1967 paper and first given the name "Bregman distances" by Censor and Lent in 1981 (see the references below).

There are obviously many convex function you can choose to build a Bregman divergence, but one of the things that makes it a good generalisation is that this class of distances already includes several existing distance measures. For example, the Mahalanobis distances, which are usually defined in terms of a matrix $A \in \mathbb{R}^{n \times n}$ can be generated as a Bregman divergence from the "distorted" squared norm $f_A(x) = \frac{1}{2} x^\top A x$, which reduces to the usual squared norm when A is the identity.

Perhaps more importantly, the famous Kullback-Leibler (KL) divergence can be expressed as a Bregman divergence using the convex function $f_{KL}(p) = \sum_{i=1}^n p_i \log p_i$ (i.e., the negative Shannon entropy) defined over $p \in \mathbb{R}^n$ with $\sum_{i=1}^n p_i = 1$. Figure 2 shows an interactive rendering of the KL divergence as a Bregman divergence.

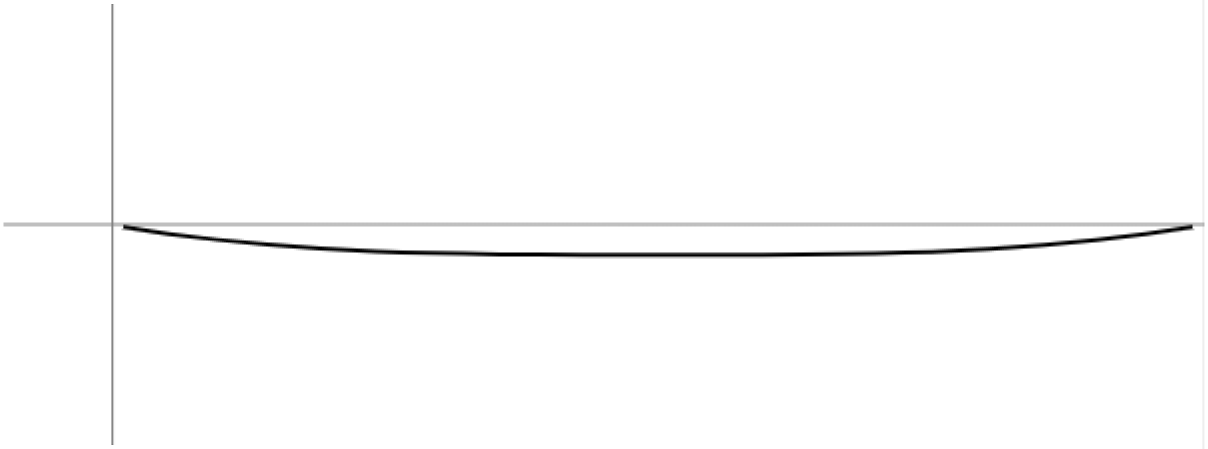

Figure 2: A geometric interpretation of the KL divergence. **Hover** to move x; **Click** to place y. The black curve is $f_{KL}(x) = x \ln x + (1-x) \ln(1-x)$. The red line is the tangent at y. The length of the blue line is the value of $KL(x, y)$.

What's the point?

So we've pulled apart and put back together squared Euclidean distance and come up with a generalisation that covers at least two other important distance measures. What? That's not enough for you?

One of the main reasons Bregman divergences are studied in machine learning are their close ties with convexity. Convex functions are general enough to be broadly applicable but have just enough structure for us to say interesting things about them. Because Bregman divergences all measure the gap between a convex function and its tangents we can obtain general results about all of them by applying the rich collection of geometric results from convex analysis.

For example, we already established that $d_f(x,y) \geq 0$ for all $x, y \in \mathbb{R}^n$ via the convexity of f. Other readily obtainable facts about Bregman divergences include:

- **Convexity** in the first argument: *i.e.*, $x \mapsto d_f(x,y)$ is convex for all y.
- **Linearity**: $d_{\alpha f + \beta g} = \alpha d_f + \beta d_f$ for all convex f and g and positive constants α and β.
- **Affine invariance**: $d_{f+g} = d_f$ for all convex f and affine g (*i.e.*, $g(x) = Ax + c$ for constant matrix A and vector c)
- The **Bregman projection** onto a convex set $C \subseteq \mathbb{R}^n$ given by $y' = \arg\min_{x \in C} d_f(x,y)$ is unique.
- A **generalised Pythagorean theorem** holds: for convex $C \in \mathbb{R}^n$ and for all $x \in C$ and $y \in \mathbb{R}^n$ we have $d_f(x,y) \geq d_f(x,y') + d_f(y',y)$ where y' is the Bregman projection of y, and equality holds when the convex set C defining the projection y' is affine.
- **Duality**: $d_f(x,y) = d_{f^*}(\nabla f(y), \nabla f(x))$ for all x, y where f^* is the convex conjugate to f

If you are analysing something that has a convex or concave function in it then, chances are you're going to bump into a Bregman divergence eventually. If that's the case, the above properties give you a lot of avenues for understanding what's going on and providing some geometric intuition.

Bregman Divergence iff The Mean is a Minimiser

As promised, I wanted to present what I found to be a suprising result about Bregman divergences: that they are characterised by how they are minimised.

More precisely, and paraphrasing their result slightly, Banerjee et al. (2005) proved the following:

> Suppose $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a continuous, non-negative function such that $d(x,x) = 0$ for all $x \in \mathbb{R}^n$ and $\frac{\partial}{\partial x_i \partial y_j} d(x,y)$ are continuous for $1 \leq, i, j \leq n$. If, for all random variables X with values in $\mathbb{R}^n$, the mean $\mathbb{E}[X]$ is the unique minimiser of $y \mapsto \mathbb{E}[d(X,y)]$ then there exists a strictly convex and differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ such that $d = d_f$.

Their proof is quite technical but hinges on showing that the minimisation property is enough to show that the derivative of $d(x,y)$ with respect to y is linear in x. More specifically, the minimisation property ensures that we can find functions $H_{ij} : \mathbb{R}^n \to \mathbb{R}$ so that for each y_i the partial derivative $\frac{\partial}{\partial y_i} d(x,y) = \sum_{j=1}^{n} H_{ij}(y)(y_j - x_j)$. The result follows, more or less, by integrating this and checking that the resulting generating function is strictly convex.

As I noted in the introduction, when I first saw this result a couple of years ago I was quite surprised. At first glance, the definition of Bregman divergences in terms of convex functions leaves a lot of scope for their behaviour and seems to have little to do with means of random variables. However, I've since spent a lot of time thinking about convex functions and now know they are intimiately related to expectations — indeed, Jensen's inequality characterises convex functions in terms of an inequality involving means. That said, I still think it is an impressive result that deserves more attention.

References and Further Reading

I could write several blog posts of this length touching on the many applications of Bregman divergences in machine learning. However, I'd prefer it if someone else did them instead so I'll just leave these links here for you. No pressure.

The details of main result I discussed above can be found here:

- Banerjee, A. and Gou, X., and Wang, H., *On the Optimality of Conditional Expectation as a Bregman Predictor*, IEEE Trans. on Information Theory, Vol. 51 (7), 2005.

Another paper by Arindam Banerjee and colleagues that I like connects exponential families with Bregman divergences. Its Appendix A has a nice summary of some properites of Bregman divergences:

- Banerjee, A. and Merugu, S. and Dhillon, I.S. and Ghosh, J., *Clustering with Bregman Divergences*, Journal of Machine Learning Research, 2005.

Another concise summary of Bregman divergences along with their application to the analysis of online convex optimisation algorithms can be found in the following lecture notes by Sasha Rakhlin:

- Rakhlin, A., *Lecture Notes on Online Learning* (Draft), 2009.

Earlier applications of Bregman divergences to learning theory include these two papers on boosting:

- Lafferty, J., *Additive Models, Boosting, and Inference for Generalized Divergences*, COLT, 1999.
- Kivinen, J. and Warmuth, M.K., *Boosting as Entropy Projection*, COLT, 1999.

I first encountered Bregman divergences in the context of proper losses in a paper by Andreas Buja and others. It shows that the *regret* of a prediction (*i.e.*, the loss of predicting q when the true distribution of outcomes is p) is the Bregman divergence between p and q:

- Buja, A. and Stuetzle, W. and Shen, Y., *Loss Functions for Binary Class Probability Estimation and Classification*, Tech. Report, University of Pennsylvania, 2005.

Bob and I subsequently spent some time looking at properness and various relationships between risk and divergences, including Bregman divergences:

- Reid, M.D. and Williamson, R.C., *Information, Divergence and Risk for Binary Experiments*, JMLR, 2011.
- Reid, M.D. and Williamson, R.C, *Surrogate Regret Bounds for Proper Losses*, ICML 2009.

Similar observations to that last paper were also made in the following paper:

- Nock, R. and Nielsen, F., *Bregman Divergences and Surrogates for Learning*, IEEE Trans. on PAMI, 2009

More recently, Jake and Raf showed that all proper scoring rules for linear properties are Bregman divergences:

- Abernethy, J.D. and Frongillo, R.M., *A Characterization of Scoring Rules for Linear Properties*, COLT, 2012.

If infinite dimensional spaces are your thing and you want your favourite Bregman divergence to work there you're going to have to learn about the Fréchet derivative. I'd recommend starting here:

- Frigyik, B.A. and Srivastava, S. and Gupta, M.R., *Functional Bregman Divergences and Bayesian Estimation of Distributions*, Trans. on Information Theory, 2008.

Finally, for some historical context, you can track down the original paper by Bregman, as well as the paper by Censor & Lent that gave the divergences his name:

- Bregman, L.M., *The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming*, USSR Comp. Math. and Math. Physics 7 (3): 200–217, 1967.
- Censor, Y. and Lent, A., *An iterative row-action method for interval convex programming*, Journal of Optimization Theory and Applications, Vol. 34 (3), pp 321—353, 1981

… and that's only the tip of the iceberg.

If there are some other elegant, original, or striking uses of Bregman divergences in the machine learning literature that I haven't listed above, please feel free to add them in the comments.

1\.
The converse holds as well but that is much easier to show.↩

2\.
However, it is not a *metric* – the usual mathematical definition of a distance – as it does not satisfy the triangle inequality. As we will see, this is true of Bregman divergences in general.↩

MARK REID

JUNE 4, 2013
Canberra, Australia

Important Update

When you log in with Disqus, we process personal data to facilitate your authentication and posting of comments. We also store the comments you post and those comments are immediately viewable and searchable by anyone around the world.

Please access our **Privacy Policy** to learn what personal data Disqus collects and your choices about how it is used. All users of our service are also subject to our **Terms of Service**.

Proceed