# Symmetrized Bregman Divergences and Metrics

Arindam Banerjee Daniel Boley
Univ of Minnesota, Twin Cities
banerjee,boley@cs.umn.edu

Sreangsu Acharyya
Univ of Texas at Austin
srean@lans.ece.utexas.edu

While Bregman divergences [3] have been used for several machine learning problems in recent years, the facts that they are asymmetric and does not satisfy triangle inequality have been a major limitation. In this paper, we investigate the relationship between two families of symmetrized Bregman divergences and metrics, which satisfy the triangle inequality. Further, we investigate kmeans-type clustering problems using both families of symmetrized divergences, and give efficient algorithms for the same.

The first family, called Generalized Symmetrized Bregman (GSB) divergences, can be derived from any well-behaved convex function. In particular, if $\phi$ is a convex function of Legendre type [5], the GSB divergence can be defined as:

$$D_\phi^{gsb}(x,y) = d_\phi(x,y) + d_\phi(y,x) + \frac{1}{2}(x-y)^T A(x-y) + \frac{1}{2}(t_x - t_y)^T B(t_x - t_y) \ , \tag{1}$$

where $A, B$ are positive definite matrices and $t_x, t_y$ denote the Legendre conjugates of $x, y$ respectively, i.e., $t_x = \nabla\phi(x), t_y = \nabla\phi(y)$. Then, we show that $(AB - I)$ being positive semi-definite is a sufficient condition for the GSB divergence $D_\phi^{gsb}(\cdot,\cdot)$ to be the square of a metric for any $\phi$. In addition, there are convex functions $\phi$ for which the condition is necessary as well. Further, we show that the metric derived from GSB divergences can be isometrically in a finite dimensional space [7].

The second family, called Jensen Bregman (JB) divergences, generalizes the Jensen-Shannon divergence [4]. In particular, for a convex function $\phi$ of Legendre type, the JB divergence is defined as

$$\Delta_\phi(x,y) \triangleq \frac{1}{2} d_\phi\left(x, \frac{x+y}{2}\right) + \frac{1}{2} d_\phi\left(y, \frac{x+y}{2}\right) = \frac{1}{2}\phi(x) + \frac{1}{2}\phi(y) - \phi\left(\frac{x+y}{2}\right) \ . \tag{2}$$

While JB divergences are non-negative, symmetric, and zero only when $x = y$, they are not squares of metrics in general. We show that square root of JB divergences give a metric only when the associated convex functions have a certain conditional positive definiteness (CPD) structure. Using results from harmonic analysis of infinitely divisible distributions [6], we show that the desired CPD structure is present in a convex function if and only if it is the cumulant function of an infinitely divisible distribution. Among other things, our result leads to elementary proofs of the metric properties of Jensen-Shannon divergence [4] and the log-AM-GM divergence. Unlike the GSB divergence, the JB divergences need not have a finite dimensional isometric embedding. However, using a remarkable result from harmonic analysis [7], we can show that metrics derived from the JB divergence can be isometrically embedded in a Hilbert space, whose kernel is intimately related to the convex function generating the divergence.

Motivated by recent advances in clustering with Bregman divergences [2] as well as practical algorithms for the kmeans problem with provable guarantees [1], we investigate kmeans-type problems using GSB and JB divergences. We show that the kmeans++ argument [1] readily generalizes to all GSB divergences with the exact same guarantees. For JB divergences, we outline two different approaches to iteratively optimize the clustering objective while maintaining the kmeans++ argument. The first approach makes use of the following fact: while clustering with JB divergences cannot be directly reduced to kernel kmeans, we can derive an isometric kernel using the convex function which can be used as a surrogate while maintaining the theoretical guarantees. The second approach is based on a variational approximation to a kmeans-type iterative optimization that maintains the guarantees as well.

In addition to uncovering precise connections between symmetrized Bregman divergences and metrics, our analysis brings to light several important results from the harmonic analysis literature that may play an important role in other areas of machine learning.

## References


[1] D. Arthur and S. Vassilvitskii. `k-means++`: The Advantages of Careful Seeding. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035, 2007.

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

[3] Y. Censor and S. Zenios. *Parallel Optimization: Theory, Algorithms, and Applications*. Oxford University Press, 1998.

[4] D. M. Endres and J. E. Schindelin. A new metric for probability distributions. *IEEE Transactions of Information Theory*, 49(7):1858–1860, 2003.

[5] R. T. Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics. Princeton University Press, 1970.

[6] K. Sato. *Levy Processes and Infinitely Divisible Distributions*. Cambridge University Press, 1999.

[7] I. J. Schoenberg. Metric spaces and positive definite functions. *Transactions of American Mathematical Society*, 44(3):522–536, 1938.


**Topic: Learning algorithms**
**Preference: Oral/Poster**