Bregman Divergences and Triangle Inequality

Sreangsu Acharyya * Arindam Banerjee † Daniel Boley†

Abstract

While Bregman divergences have been used for clustering and embedding problems in recent years, the facts that they are asymmetric and do not satisfy triangle inequality have been a major concern. In this paper, we investigate the relationship between two families of symmetrized Bregman divergences and metrics that satisfy the triangle inequality. The first family can be derived from any well-behaved convex function. The second family generalizes the Jensen-Shannon divergence, and can only be derived from convex functions with certain conditional positive definiteness structure. We interpret the required structure in terms of cumulants of infinitely divisible distributions, and related results in harmonic analysis. We investigate kmeans-type clustering problems using both families of symmetrized divergences, and give efficient algorithms for the same.

1 Introduction

Recent years have seen interest in going beyond Euclidean distances for a variety of data mining problems. One important development is to use Bregman divergences [6, 2]. Bregman divergences are a general class of distortion functions, which include squared Euclidean distance, KL-divergence, Itakura-Saito distance, etc., as special cases. Indeed, such a divergence can be generated from any (differentiable) convex function.

As examined by Banerjee et al., [2] Bregman divergences may be considered a generalization of squared Euclidean distance because of many shared properties. A crucial property that is not shared is that, the square root of a Bregman divergence is not necessarily a a metric. In particular, Bregman divergences are not symmetric, and do not satisfy the triangle inequality. As a result, data-structures and algorithms [15] that exploit these properties for scalability lay beyond reach of methods that use Bregman divergences [7]. There have been recent notable attempts to investigate symmetry [23], however, they do not satisfy triangle inequality. In this paper, we investigate two families of symmetrized Bregman divergences that do.

The first family of symmetrizations investigated, called Generalized Symmetrized Bregman (GSB) divergence, can be derived from any well behaved convex function. We present necessary and sufficient conditions under which a GSB divergence is the square of a metric. Further, we show that they can be isometrically embedded in a finite dimensional Euclidean space. The second family, called Jensen-Bregman (JB) divergences, generalizes the Jensen-Shannon divergence [9]. The second family is obtained by direct symmetrization of Bregman divergences obtained from a special class of convex functions. In particular, we show that JB divergences are squares of a metric only when the associated convex functions are conditionally positive definite (CPD).

We relate CPD functions with cumulants of infinitely divisible distributions, and related results in harmonic analysis [5, 4, 19]. In the process, we develop a powerful and flexible method for constructing metrics from convex functions. This technique proves metric properties of some well known divergences. Chen et al., [10] study the same symmetrization and identify necessary and sufficient conditions for triangle inequality, but for univariate convex functions. In comparison, our work generalizes to multivariate convex functions, explains the connection with CPD functions and infinitely divisible measures, shows how the divergence may be embedded in a Hilbert space, provides recipes for creating such functions, and develops algorithms intimately tied with the properties of these divergences.

Both families of divergences considered in this paper lead to Hilbert space embeddable metrics and permit development of efficient algorithms for clustering and search. We believe this family would be of interest to practioners because of the computational advantages of triangle inequality. Indeed, Cherian et al. [11] have recently reported excellent performance on the task of similarity based image search using a matrix extension.

A second advantage is the non-linearity associated with such metrics, which is similar to but more general than kernel methods. Recall that squared Euclidean distance, the basis of many data-mining algorithms, comes with some well known limitations. k-means clustering, for example, leads to piece-wise linear cluster boundaries that partition the space into polyhedrons.

*Department of Electrical and Computer Engineering, University of Texas at Austin.

†Department of Computer Science and Engineering, University of Minnesota, Twin Cities.

Such a partitioning severely restricts the applicability of k-means to clustering problems which require non-linear partitions, possibly based on data density and related factors. Spectral graph partitioning does not have this limitation, but can be computationally demanding. A popular solution is to introduce non-linearity using the "kernel trick" [29] which depend on positive (semi)definite (PD) functions. In our work, an analogous role is played by the larger class of CPD functions. Indeed, a contribution of this paper is to show that the well established PD kernel (algorithmic) machinery [28] may be re-used for this CPD class as well.

We show how to generate kernel matrices that are isometric with respect to JB divergences. Such kernel matrices may then be substituted transparently for problems involving JB divergences. As a result, clustering using JB divergences can be converted into a kernel k-means problem based on such derived kernels. One downside of a kernel k-means algorithm is that each iteration is quadratic in the number of data points, and hence can be slow for large datasets. Interestingly, we show that clustering using JB divergences can be solved using a variational algorithm, with linear complexity per iteration. Thus, we get the power of kernel k-means with a (per-iteration) computational cost of k-means.

The rest of the paper is organized as follows: In Section 2, we review some necessary technical background. In Section 3, we introduce Generalized Symmetrized Bregman (GSB) divergences, which can be obtained from any well behaved convex function, and study their metric properties. In Section 4, we discuss Jensen-Bregman (JB) divergences and their metric properties. We present and analyze clustering algorithms based on both GSB and JB divergences in Section 5, and conclude in Section 6.

2 Background

In this section we review Bregman divergences, positive and conditionally positive definite characterization of kernel functions, and infinitely divisible distributions.

2.1 Bregman Divergence Let ϕ be a convex function of Legendre type, i.e., ϕ is a closed proper convex function, and if $\Theta = \mathrm{int}(\mathrm{dom}(\phi)) \subseteq \mathbb{R}^d$, then Θ is non-empty, ϕ is strictly convex and differentiable in Θ, and $\forall \theta \in \mathrm{bd}(\Theta)$, $\lim_{\theta \to \theta_b} ||\nabla\phi(\theta)|| \to \infty$, where $\nabla\phi(\theta)$ is the gradient of ϕ at θ [25]. For any $x \in \mathrm{dom}(\phi), y \in \mathrm{int}(\mathrm{dom}(\phi)) = \Theta$, the Bregman divergence [8] corresponding to ϕ is defined as

$$d_\phi(x, y) = \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle \ . \tag{2.1}$$

It is easy to show that $d_\phi(x, y) \geq 0$ and $d_\phi(x, y) = 0$ iff $x = y$. Let $\psi = \phi^*$ be the conjugate function of ϕ, i.e., $\psi(t) = \sup_{x \in \mathrm{dom}(\phi)} \{\langle x, t\rangle - \phi(x)\}$.

Since ϕ is a convex function of Legendre type, it follows [25] that ψ will also be a convex function of Legendre type. Further, if $\Theta^* = \mathrm{int}(\mathrm{dom}(\psi))$, then the gradient function $\nabla\phi : \Theta \mapsto \Theta^*$ is a one-to-one function from the open set Θ to the open set Θ^*. Further, the gradient functions $\nabla\phi, \nabla\psi$ are continuous, and $\nabla\psi = (\nabla\phi)^{-1}$. As a consequence, for any $x \in \Theta$, there is a unique $t \in \Theta^*$ such that they are Legendre transforms of each other, i.e., $t = \nabla\phi(x)$ and $x = \nabla\psi(t)$. As appropriate, we will denote the conjugate of x as t_x, or the conjugate of t as x_t.

2.2 Conditionally Positive Definite Kernels A real valued function $C(\cdot, \cdot) : \mathcal{S} \times \mathcal{S} \mapsto \mathbb{R}$ is called a *conditionally positive definite* (CPD) kernel[1] if for any positive integer n and any choice of n elements $x_{1 \leq i \leq n} \in \mathcal{S}$ and a choice of n reals $u_i \in \mathbb{R}$ such that $\sum_i u_i = 0$, the following inequality $\sum_{i,j=0}^{n} u_i u_j C(x_i, x_j) \geq 0$ holds. The kernels for which the inequality holds for any choice of u_i are called positive (semi)definite (PD). All PD kernels are CPD, but the converse is not true.

The following is a striking result related to CPD kernels, PD kernels and metric on a Hilbert space that will be used in our analysis:

Theorem 1 ([27]) *Let $C(\cdot, \cdot)$ be a function on a topological set $\mathcal{S} \times \mathcal{S}$. If $\mathcal{S}$ is* separable *then there exists a Hilbert space $\mathcal{H}$ of real-valued functions on S, and a mapping $\Phi : S \mapsto \mathcal{H}$ such that*

$$\|\Phi(x) - \Phi(y)\|^2 = -C(x, y) + \frac{1}{2}(C(x, x) + C(y, y)) \tag{2.2}$$

if and only if $C(\cdot, \cdot)$ is a CPD kernel or equivalently $K(\cdot, \cdot) = \exp(-\beta C(\cdot, \cdot))$ is a PD kernel for any $\beta > 0$.[2]

One should take special note of the fact that this condition is both necessary as well as sufficient.

2.3 Infinitely Divisible Distributions For a probability measure μ, let μ^n denote the n-fold convolution of the probability measure with itself, i.e., $\mu^n = \mu * \mu * \cdots * \mu$ (n times). A probability measure μ on $\mathbb{R}^d$ is infinitely divisible if, for any positive integer n, there is a probability measure μ_n on $\mathbb{R}^d$ such that $\mu = \mu_n^n$, i.e., μ is the n-fold convolution of some other measure μ_n, for all $n \in \mathbb{N}$. While the definition of infinitely divisible distributions is based on the n-fold convolution μ^n, the β-fold convolution μ^β is well defined and infinitely divisible for any $\beta \geq 0$ [26, Lemma 7.9]. For the purposes of our analysis, we need the following result concerning

[1] In the literature [4, 18], $-C$ is often called negative definite.
[2] Note that the exponentiation here is element-wise.

the characteristic function of infinitely divisible distributions. The result follows directly from Theorem 8.1 and Corollary 8.3 of [26].

Theorem 2 ([26]) *Let μ be an infinitely divisible distribution with characteristic function $F(z)$. Then, the characteristic function of μ^β is $F(z)^\beta$. Conversely, the characteristic function $F(z)^\beta$ uniquely corresponds to the distribution μ^β.*

3 Symmetrized Bregman Divergences and Metrics

For any $x, y \in \Theta$, the symmetrized Bregman divergence corresponding to ϕ is given by

$$d_\phi^{sym}(x,y) = d_\phi(x,y) + d_\phi(y,x) \ . \tag{3.3}$$

In general, this symmetrized Bregman divergence is not related to metrics. We propose a natural generalization of the symmetrized Bregman divergence, that we call Generalized Symmetrized Bregman (GSB) divergence. For $x, y \in \Theta$, the GSB divergence corresponding to ϕ is defined as

$$\begin{aligned} d_\phi^{gsb}(x,y) &= d_\phi^{sym}(x,y) + \frac{\alpha}{2}\|x-y\|^2 + \frac{\beta}{2}||t_x - t_y||^2 \\ &= d_\phi(x,y) + d_\phi(y,x) + \frac{\alpha}{2}\|x-y\|^2 + \frac{\beta}{2}||t_x - t_y||^2 \ , \end{aligned}$$

where $\alpha, \beta \geq 0$ are constants and t_x and t_y are $\nabla\phi(x)$ and $\nabla\phi(y)$ respectively . It is easy to see that the GSB divergence is symmetric, i.e., $d_\phi^{gsb}(x,y) = d_\phi^{gsb}(y,x)$.

We are now ready to state the main result connecting Generalized Symmetrized Bregman (GSB) divergences and metrics.

Theorem 3 *Let ϕ be any convex function of Legendre type. Then, $\sqrt{d_\phi^{gsb}(x,y)}$ is a metric iff $\alpha\beta \geq 1$.*[3]

Proof. Let $\alpha\beta \geq 1$. The kernels $C_1(x,y) = (\sqrt{\alpha}x + \frac{1}{\sqrt{\alpha}}t_x)^T(\sqrt{\alpha}y + \frac{1}{\sqrt{\alpha}}t_y)$ and $C_2(x,y) = t_x^T t_y$, being inner products, are CPD. From [18, Lemma 2], recall that $C(x,y)$ is CPD iff $C(x,y) + f(x) + f(y)$ is CPD, so that adding functions of only x or y does not affect the CPD property. Further, since CPD kernels are closed under convex combinations,

$$\begin{aligned} C(x,y) &= \frac{1}{2}C_1(x,y) + \frac{\alpha\beta - 1}{\alpha}C_2(x,y) - x^T t_x - y^T t_y \\ &= \frac{\alpha}{2}x^T y + \frac{\beta}{2}t_x^T t_y - [d_\phi(x,y) + d_\phi(y,x)] \end{aligned}$$

is CPD. It is straightforward to see that $d_\phi^{gsb}(x,y) = -C(x,y) + (C(x,x) + C(y,y))/2$ so that, following Theorem 1, $\sqrt{d_\phi^{gsb}}$ is a metric.

[3] By 'only if,' we mean $\exists\phi$ for which $\alpha\beta \geq 1$ is necessary.

We prove the necessity of $\alpha\beta \geq 1$ using a specific convex function, $\phi(x) = xs\log_2 x - (\log_2 e)sx$ where $t_x = s\log_2 x, s > 0$. Let $r = \sqrt{\beta/\alpha}$ We will choose three specific scalars x, y, z and set $A = \sqrt{d_\phi^{gsb}(x,y)}$, $B = \sqrt{d_\phi^{gsb}(y,z)}$, and $C = \sqrt{d_\phi^{gsb}(x,z)}$. The triangle inequality is $A + B \geq C$, which through a direct calculation yields $C^4 + (A^2 - B^2)^2 - 2C^2(A^2 + B^2) \leq 0$. Now we use specific positive values for A, B, C, choosing $x = rs$,$y = 2rs$, and $z = 4rs$. The corresponding conjugate 't' values are $t_x = s(0 + \log_2 rs)$, $t_y = s(1 + \log_2 rs)$, and $t_z = s(2 + \log_2 rs)$. Computing A^2, B^2, C^2 and plugging them back in the inequality above we have $C^4 + (A^2 - B^2)^2 - 2C^2(A^2 + B^2) = 4r^2s^4(1 - (r\alpha)^2) \leq 0$, which implies $r\alpha \geq 1$. Since $r = \sqrt{\beta/\alpha}$, we note that $\alpha\beta \geq 1$ is necessary. This completes the proof. ∎

Since $\sqrt{d_\phi^{gsb}}$ is a metric generated from a CPD kernel, from [27, 18] it follows that $(\Theta, \sqrt{d_\phi^{gsb}})$ can be isometrically embedded in ℓ_2, the space of square integrable functions. For GSB divergences, multiple exact finite-dimensional metric embeddings can be obtained in closed form. We give two such examples below.

Lemma 1 *$(\Theta, \sqrt{d_\phi^{gsb}})$ can be isometrically embedded in $\mathbb{R}^{2d}$ using two different maps:*

$$f_1(x) = \begin{pmatrix} \sqrt{\alpha}x + \frac{1}{\sqrt{\alpha}}t_x \\ \sqrt{\frac{\alpha\beta - 1}{\alpha}}t_x \end{pmatrix} \quad \textit{and} \quad f_2(x) = \begin{pmatrix} \frac{1}{\sqrt{\beta}}x + \sqrt{\beta}t_x \\ \sqrt{\frac{\alpha\beta - 1}{\beta}}x \end{pmatrix}$$

Proof. A direct calculation shows that $d_\phi^{gsb}(x,y) = ||f_i(x) - f_i(y)||^2, i = 1, 2.$ ∎

A substantial generalization of the proposed GSB divergences can be made to Mahalanobis-type metrics. In particular, for symmetric positive definite matrices A, B, we consider the GSB divergence

$$\begin{aligned} D_\phi^{gsb}(\mathbf{x},\mathbf{y}) \ &= d_\phi(\mathbf{x},\mathbf{y}) + d_\phi(\mathbf{y},\mathbf{x}) + \frac{1}{2}(\mathbf{x}-\mathbf{y})^T A(\mathbf{x}-\mathbf{y}) \\ &\quad + \frac{1}{2}(\mathbf{t_x} - \mathbf{t_y})^T B(\mathbf{t_x} - \mathbf{t_y}) \ . \end{aligned} \tag{3.4}$$

Theorem 4 *Let ϕ be any convex function of Legendre type. Then, $D_\phi^{gsb}(\mathbf{x},\mathbf{y})$ is a metric iff $(AB - I)$ is positive semi-definite.*

The proof is similar to that of Theorem 3, and is skipped due to lack of space. Theorem 4 is expected to be useful in the context of metric learning, a topic we do not explore in the current paper.

Example 1.A For $\phi(x) = x\log x - x$, with $\alpha, \beta = 1$,

$$\sqrt{d_\phi^{gsb}(x,y)} = |x-y| + \log\left(\frac{\max(x,y)}{\min(x,y)}\right) ,$$

is a metric. Similarly for $\phi(x) = -\log x$, with $\alpha, \beta = 1$,

$$\sqrt{d_\phi^{gsb}(x,y)} = |x-y|\left(1+\frac{1}{xy}\right).$$

4 Jensen Bregman Divergences and Metrics

In this section, we investigate a different symmetrization. Let ϕ be a convex function of Legendre type and $d_\phi(\cdot,\cdot)$ be the corresponding Bregman divergence. Then, for $x, y \in \Theta$, we define the Jensen Bregman as

(4.5)
$$\begin{aligned}\Delta_\phi(x,y) &\triangleq \frac{1}{2}d_\phi\left(x, \frac{x+y}{2}\right) + \frac{1}{2}d_\phi\left(y, \frac{x+y}{2}\right)\\ &= \frac{1}{2}\phi(x) + \frac{1}{2}\phi(y) - \phi\left(\frac{x+y}{2}\right) .\end{aligned}$$

Strict convexity of ϕ and the Jensen's inequality together ensure that $\Delta_\phi(x,y) \geq 0$ and $\Delta_\phi(x,y) = 0$ if and only if $x = y$.

Note that apart from triangle inequality, all properties of a metric are satisfied by Jensen Bregman generated from any strictly convex function ϕ. We show that $\Delta_\phi(x,y)$ is the square of a metric when the kernel $\phi(x+y)$, induced by the convex function ϕ, is CPD [3].

Lemma 2 *$\sqrt{\Delta_\phi(\cdot,\cdot)}$ is a metric iff $\phi(x+y)$ is a conditionally positive definite (CPD) kernel. In this case, there exists a a Hilbert space $\mathcal{H}$ of real-valued functions on Θ, and a mapping $\Phi : \Theta \mapsto \mathcal{H}$, such that $\sqrt{\Delta_\phi(x,y)} = ||\Phi(x) - \Phi(y)||$.*

Proof. Let $\{c_i\}_{i=1}^n$ be a set of real numbers s.t. $\sum_i c_i = 0$ and $\{x_i\}_{i=1}^n$ be any set of points $x_i \in \Theta$. Then, if $\phi(x+y)$ is CPD

$$\sum_{i,j=1}^n c_i c_j \Delta_\phi(x_i,x_j) = -\sum_{i,j=1}^n c_i c_j \phi\left(\frac{x_i+x_j}{2}\right) \leq 0 ,$$

where the other terms vanish since

$$\sum_{i,j} c_i c_j \phi(x_i) = \left(\sum_i c_i\phi(x_i)\right)\left(\sum_j c_j\right) = 0 ,$$

and so on. Since $-\Delta_\phi(x,y)$ is CPD, from Theorem 1 it follows that $\sqrt{\Delta_\phi(\cdot,\cdot)}$ is a metric with an isometric embedding $\phi(x)$.

For the 'only if' part, let $\sqrt{\Delta_\phi((,)x,y)}$ be a metric. From Theorem 1, we know that there exists a CPD kernel $C(x,y)$ such that $\Delta_\phi(x,y) = -C(x,y) + \frac{1}{2}(C(x,x)+C(y,y))$, so that

$$\phi\left(\frac{x+y}{2}\right) = C(x,y) - \frac{1}{2}(C(x,x)+C(y,y)) + \frac{1}{2}(\phi(x)+\phi(y)) .$$

Then, for $c_i, [i]_1^n$ with $\sum_i c_i = 0$, and $x_i, [i]_1^n \in \Theta$, we have

$$\sum_{i,j} c_i c_j \phi\left(\frac{x_i+x_j}{2}\right) = \sum_{i,j} c_i c_j C(x_i,x_j) \geq 0 ,$$

so that ϕ is a CPD kernel. ∎

While the above result can be useful, and CPD functions are indeed convex, it is unclear as to which convex functions ϕ will lead to CPD kernels and as a consequence impart metric property to $\sqrt{\Delta_\phi(x,y)}$. Hence it is crucially important to exactly characterize the class of convex functions that lead to CPD kernels. This is obtained by the following result [3, 4, 5].

Theorem 5 *The additive kernel $k(x,y) = g(x+y)$ is conditionally positive definite, if and only if $g(x) = \log\int_r \exp(\langle x,r\rangle)d\mu(r)$ for a uniquely determined infinitely divisible measure μ.*

The theorem implies that $g(x)$ has to be the cumulant or log-partition function of any infinitely divisible distribution. From a harmonic analysis perspective $g(x)$ is the log of the (multivariate) Laplace transform of such a distribution [4, 5].

Proof. For the 'if' part, we know that μ is infinitely divisible, and, by Devinatz's theorem [12] $G(s) = \int_r \exp(\langle r,s\rangle)d\mu(r)$ is the moment generating function of μ with $g(s) = \log G(s)$ being the cumulant, and $\mathcal{S} = \mathrm{dom}(G)$. Let $F(t) = \int_r \exp(i\langle r,t\rangle)d\mu(r)$ be the characteristic function. Now, the characteristic function of the base measure $dm(r) = \exp(\langle r,s\rangle)d\mu(r)$ is given by $H(t) = \int_r \exp(i\langle t,r\rangle)dm(r) = \int_r \exp(i\langle r,t-is\rangle)d\mu(r)$, where the integral is convergent and analytic as a function of $(t-is), t \in \mathbb{R}^n, s \in \mathcal{S}$ [21, Theorem 2.7.1]. In other words, $F(t)$ has an analytic extension to $\mathbb{R}^n - i\mathcal{S} \subset \mathbb{C}^n$. As a result, following [14] $G(s) = F(-is)$. Since μ is infinitely divisible, the characteristic function of μ^β, the β-fold convolution of μ where $\beta \geq 0$, is simply $F_\beta(t) = F(t)^\beta$, from Theorem 2. Let $G_\beta(s) = \int_r \exp(\langle r,s\rangle)d\mu^\beta(r)$ be the corresponding Laplace transform. Now, the characteristic function of the base measure $dm^\beta(r) = \exp(\langle r,s,\rangle)d\mu^\beta(r)$ is given by $H_\beta(t) = \int_r \exp(i\langle t,r\rangle)dm^\beta(r) = \int_r \exp(i\langle r,t-is\rangle)d\mu^\beta(r)$, where, as before, the integral is convergent and analytic as a function of $(t-is), t \in \mathbb{R}^n, s \in \mathcal{S}$. In other words, $F_\beta(t)$ has an analytic extension to

$\mathbb{R}^n - i\mathcal{S} \in \mathbb{C}^d$. As a result, following Theorem 2, $G_\beta(s) = F_\beta(-is) = F(-is)^\beta = G(s)^\beta$. Since $G_\beta(s)$ is the Laplace transform of a probability measure, following [12] $G(x+y)^\beta$ is positive semi-definite $\forall \beta \geq 0$ so that $g(x+y) = \log G(x+y)$ is CPD from Theorem 1.

For the 'only if' part, since $G(x+y)^\beta = exp(g(x+y))$ is a PD kernel following Theorem 1, from Devinatz's theorem [12] it follows that $\forall \beta > 0$, there exists a non-negative measure μ_β such that $G(s)^\beta = \int_x \exp(\langle x, s\rangle) d\mu_\beta(x)$. Then, from [14], it follows that the characteristic function $F_\beta(t)$ of μ_β can be obtained using a simple plug-in $F_\beta(t) = G(it)^\beta$. Choosing $\beta = 1$ and $\beta = n$, we note that $F_n(t) = G(it)^n = F(t)^n$ so that the characteristic function of μ_n is the n-fold product of that of μ. This holds for all n, μ. ∎

Since the cumulant is always convex, we have

Corollary 1 *$\sqrt{\Delta_\phi(\cdot,\cdot)}$ is a metric if ϕ is the cumulant of an infinitely divisible distribution.*

We now focus on constructing examples of such convex functions. It is well known that the characteristic function $F_\mu(t)$ of infinitely divisible measures μ on $\mathbb{R}^n$ can be expressed in closed form by the *Levy-Khintchine* (L-K) formula [26, 20]. A careful analysis based on a recent result [14] shows that one can obtain the moment-generating function $L_\mu(s) = F_\mu(-is)$,[4] and further the cumulant function $\phi(s) = \log L_\mu(s)$ is given by

(4.6)
$$\phi(s) = \frac{1}{2}\langle s, As\rangle + \langle \gamma, s\rangle + \int_r (e^{\langle s,r\rangle} - 1 - \langle s, r\rangle \mathbb{1}_D(r)) d\nu(r)$$

where A is a $d \times d$ positive definite matrix, $\gamma \in \mathbb{R}^d$, and ν is a Levy measure on $\mathbb{R}^d$ satisfying $\nu(0) = 0$ and $\int_r \min(||r||^2, 1) d\nu(r) < \infty$. In fact, for any choice of the triplet (A, γ, ν), the corresponding $\phi(s)$ will be such that $\phi(x+y)$ is a CPD kernel. We illustrate the utility of the above characterization by showing the JB divergence corresponding to $\phi(x) = -\log x$ is a metric.

Lemma 3 *For $\phi(x) = -\log x$, $\sqrt{\Delta_\phi(x,y)}$ is a metric for $x, y \in \mathbb{R}_{++}$, where*

$$\Delta_\phi(x,y) = \log\left(\frac{\frac{x+y}{2}}{\sqrt{xy}}\right) . \tag{4.7}$$

In other words, the log of the ratio of the arithmetic and geometric mean of two positive numbers satisfy the triangle inequality.

[4] There are several variants of the L-K formula, we are using one from [26] to illustrate the point.

Proof. It is well known [26] that the Gamma distribution

$$p(x;\alpha,\beta) = x^{\alpha-1}\frac{\beta^\alpha e^{-\beta x}}{\Gamma(\alpha)} = e^{\beta(-x)-\alpha(-\log\beta)}\frac{x^{\alpha-1}}{\Gamma(\alpha)} ,$$

where $\alpha, \beta > 0$ is infinitely divisible. From an infinite divisibility perspective, the scale parameter α denotes the number (amount) of convolutions of the measure with itself. For a fixed α, we note that the cumulant function $\phi(\beta) = -\log\beta$. Then, since $\phi(x) = -\log x$ satisfies the condition of Theorem 5, from Corollary 1 it follows that $\sqrt{\Delta_\phi(\cdot,\cdot)}$ is a metric. ∎

While the above characterization is exhaustive and theoretically appealing, it does not give a way to construct such functions. We now describe two approaches to construct such convex functions. The first family of functions is based on a special class of infinitely divisible measures called stable measures [17, 24] which have the following stability property: if a set of i.i.d. random variables have a stable distribution, then a linear combination of these variables will have the same distribution possibly with different shift and scale parameters. A complete characterization of the characteristic function $F(t)$ of such measures is given as follows:

Theorem 6 ([17]) *For a distribution μ on $\mathbb{R}$ to be stable it is necessary and sufficient that its characteristic function $F(t)$ satisfies*

(4.8)
$$\log F(t) = \begin{cases} i\gamma t - |ct|^\alpha\left(1 - i\beta\frac{t}{|t|}\tan(\pi\alpha/2)\right) & \alpha \neq 1 \\ i\gamma t - |ct|\left(1 + i\beta\frac{t}{|t|}\frac{2}{\pi}\log(|t|)\right) & \alpha = 1 \end{cases} ,$$

for the ranges $-1 \leq \beta \leq 1$, $c \geq 0$, $0 < \alpha \leq 2$, $\gamma \in \mathbb{R}$.

The corresponding closed forms for their measures are known only for a few examples such as Gaussian ($\alpha = 2$) and Cauchy($\alpha = 1$, $\beta = 1$). Since the $F(s)$ has an analytic extension to the complex plane, following [14], one can construct a family of suitable convex functions $\phi(s)$ based on the following plug-in procedure $\phi(s) = \log L(s) = \log F(-is)$. We illustrate the utility of such a construction using the following result.

Lemma 4 *For $\phi(x) = x\log x$, $\sqrt{\Delta_\phi(x,y)}$ is a metric.*

Proof. From Theorem 4.8 for $\alpha = 1$, setting $\gamma = c = \frac{\pi}{2}, \beta = -1$, based on the plug-in procedure from [14], for $s > 0$ we get

$$\phi(s) = \log F(-is) = \gamma s - cs - cs\beta\frac{2}{\pi}\log s = s\log s .$$

Since all stable distributions are infinitely divisible, $\phi(x+y)$ is CPD, and hence $\sqrt{\Delta_\phi(x,y)}$ is a metric. ∎

An important consequence of the above result is an elementary proof of the fact that the Jensen-Shannon divergence is the square of a metric [9, 16].

Corollary 2 *Let $p, q \in \mathbb{R}^d_+$ and $m = (p+q)/2$. If $I(\cdot||\cdot)$ denotes the I-divergence, $\Delta(p,q) = \frac{1}{2}I(p||m) + \frac{1}{2}I(q||m)$ is the square of a metric.*

Proof. A direct calculation shows that for $\phi(x) = x \log x$ $\Delta(p,q) = \sum_{j=1}^{d} \Delta_\phi(p_j, q_j)$. The result follows from the fact that the sum of CPD kernels is always CPD. ∎

The second approach to construct CPD functions is based an alternative characterization of infinitely divisible distributions. This method provides a connection between the necessary and sufficient conditions stated in lemma 2 and those identified by Chen et al. recently [10] for the special case of univariate convex functions. [5]

Theorem 7 ([22]) *Let μ be a distribution in $\mathbb{R}$ such that $L_\mu(\theta) = \int_x \exp(\langle\theta, x\rangle) d\mu(x)$ and $\Theta = \{\theta : L(\theta) < \infty\}$ is non empty. Then μ is infinitely divisible if and only if there exists a unique positive measure ρ such that for all $\theta \in \Theta$*

$$\frac{\partial^2}{\partial\theta^2} \log L_\mu(\theta) = L_\rho(\theta) = \int_x \exp(\langle\theta, x\rangle) d\rho(x).$$

Now, given any positive measure ρ such that $L_\rho(\theta)$ is doubly integrable to give $\phi(\theta)$, and $\phi(\theta)$ is the cumulant function of some measure, then $\sqrt{\Delta_\phi(\cdot,\cdot)}$ will be a metric. We illustrate the point with an example.

Example 1 The Laplace transform of the unit ramp $\rho(x) = xu(x)$ is given by $L_\rho(s) = 1/s^2$. A double integral gives the function $\phi(s) = -\log s$ (ignoring affine terms). Since $\phi(s) = -\log s$ is the cumulant of the Gamma distribution with $\alpha = 1$, the corresponding $\sqrt{\Delta_\phi(\cdot,\cdot)}$ is a metric.

Example 2 We give an alternative proof of the fact that the Jensen-Shannon divergence is square of a metric. Consider the moment generating function of a Gamma distribution $L(s) = \left(\frac{\lambda}{\lambda - s}\right)^\tau$. Integrating twice we obtain $\iint L(s)d(s) = \lambda^\tau \frac{s^{2-\tau}}{(2-\tau)(1-\tau)} + As$. The interesting case is when $\tau = \{1, 2\}$. $\lim_{\tau\to 1} \frac{s^{2-\tau}}{(2-\tau)(1-\tau)} = \lim_{\tau\to 1} \frac{s}{2-\tau}\frac{s^{1-\tau}}{1-\tau} = s\log(s)$ because the limit of second term is $\log(s)$. Taking $A = 0$ and $\tau = 1$ or 2, we obtain that $\phi(s) = s\log(s)$. Since it has already been shown (see Lemma 4) that $s \log s$ is the log cumulant of a stable measure, it is conditionally positive definite.

Another way of constructing these metrics is to derive new ones from known CPD functions by applying transformations that preserve the metric property. A place to look for such transformations are convexity preserving ones, however it is not sufficient to preserve convexity alone. The property of infinitely divisibility too needs to be preserved. We list a couple of examples.
(1) Conic combination of known CPD functions, i.e. functions of the form $\sum_i \alpha_i \phi_i(\cdot)$ where $\alpha_i > 0$ are CPD. Furthermore if $\phi_n = \sum_{i=0}^{n} \alpha_i \phi_i(\cdot)$ is such that $\lim_{n\to\infty} \phi_n(\cdot) = \psi(\cdot)$, then $\psi(\cdot)$ is also CPD.
(2) Exponentiation, i,e, $\exp(-\beta\phi(\cdot))$ for $\beta > 0$. The resulting function is positive definite and hence CPD.

Applying these transformation we can show:

Example 3 For $\lambda, l, \tau, \sigma^2 \geq 0$ the following are CPD functions: $-\log\left(1 - \frac{\sigma^2}{2\lambda}s^2\right)$, $-\log\left(1 - \frac{l}{\lambda}(e^s - 1)\right)$, $-\log\left(1 + \frac{|s|}{\lambda}\right)$, $l\left(e^{-|s|} - 1\right)$. These can be derived by the composition of infinitely divisible distributions with distribution on positive integers.

5 Clustering Algorithms

In this section, we consider clustering problems based on GSB and JB divergences. We focus on the k-means family of clustering problems, and demonstrate that algorithms with provable guarantees can be established for all cases, based on existing results [1].

5.1 Clustering with GSB Divergences We first consider the k-means problem using any GSB divergence, where given n data points $\mathcal{X} = \{x_i, [i]_1^n, \in \mathbb{R}^d\}$, the goal is to find a clustering $\mathcal{C}$ and a corresponding set of k means $\mathcal{M} = \{\mu_h, [h]_1^k \in \mathbb{R}^d\}$ such that

$$J_1(\mathcal{C}) = \sum_{i=1}^{n} \min_h \; d_\phi^{gsb}(x_i, \mu_h) \tag{5.9}$$

is minimized. Generalizing the elegant kmeans++ algorithm [1], we propose GSB++, which has the same structure, while using GSB divergences instead of squared Euclidean distance. GSB++ builds the initial cluster centers sequentially following probabilistic farthest-first. At any point during the initialization, if M is the current set of means, let $D(x) = \min_{\mu\in M} d_\phi^{gsb}(x, \mu)$. With this notation, GSB++ is presented as Algorithm 1.

Lemma 5 *Let $\mathcal{C}_{GSB}$ be the final clustering obtained from GSB++, and let J_1^* be the optimal value of the*

[5] Chen et al. [10] identify the necessary and sufficient condition for triangle inequality to be $\frac{\partial^2}{\partial\theta^2} L_\rho(\theta) \geq 0$ which is equivalent to the convexity criteria on $L_\rho(\theta)$ above.

Algorithm 1 GSB++

Choose initial cluster center μ_1 uniformly at random from $\mathcal{X}$.
Choose the next center $\mu_h = x \in \mathcal{X}, h = 2, \ldots, k$ with probability $\frac{D(x)}{\sum_{x \in \mathcal{X}} D(x)}$. Where $D(x)$ is the maximum GSB distance of x from the previous clusters.
Let $f(x_i)$ be any finite-dimensional isometric embedding as in Lemma 1.
repeat
For $i = 1, \ldots, n$, assign x_i to C_{h^*} if $h^* = \text{argmin}_h \, ||f(x_i) - f(\mu_h)||$
For $h = 1, \ldots, k$, compute $f(\mu_h) = \frac{1}{|C_h|} \sum_{x \in C_h} f(x)$
until *convergence*

objective function. Then

$$E[J_1(\mathcal{C}_{GSB})] \leq 8(\log k + 2) J_1^* \ .$$

Proof. Note that the original kmeans++ argument goes through if the squared Euclidean distance $||x - y||^2$ is replaced by any kernelized distance $K(x,x) + K(y,y) - 2K(x,y) = ||\Phi(x) - \Phi(y)||^2$. Kernel-kmeans++ follows directly from kmeans++. From Lemma 1, since $d_\phi^{gsb}(x,y) = ||f(x) - f(y)||^2$, the initialization in GSB++ itself guarantees being within $8(\log k + 2)$-times the optimal. The iterative steps can only improve the objective function, maintaining the guarantee. ∎

5.2 Clustering with JB Divergences The k-means clustering problem with JB divergence can be posed as one of obtaining a clustering $\mathcal{C}$ so as to minimize the objective function

$$J_\Delta(\mathcal{C}) = \sum_{i=1}^{n} \min_h \Delta_\phi(x_i, \mu_h) \ . \tag{5.10}$$

The above problem has two important challenges. First, for a given cluster $\mathcal{C}_h$, the problem of mean estimation, i.e., $\min_{\mu_h} \sum_{x_i \in \mathcal{C}_h} \Delta_\phi(x_i, \mu_h)$ does not have a closed form solution. A brute force approach would be to solve the following non-linear equation iteratively:

$$\nabla\phi(\mu) = \sum_i \nabla\phi\left(\frac{x_i + \mu}{2}\right) \ ,$$

but that can be computationally problematic. Second, although we know that $\sqrt{\Delta_\phi(x,y)} = \|\Phi(x) - \Phi(y)\|$ for some Φ, we do not know $\Phi(x)$ or $K(x,y) = \langle \Phi(x), \Phi(y) \rangle$ is. In particular, even though we have

$$\Delta_\phi(x,y) = \frac{1}{2}\phi(x) + \frac{1}{2}\phi(y) - \phi((x+y)/2) \tag{5.11}$$

$$= K(x,x) + K(y,y) - 2K(x,y) \ , \tag{5.12}$$

we cannot conclude that $K(x,y) = \frac{1}{2}\phi((x+y)/2)$, since ϕ need not be positive (semi)definite.

To solve the clustering problem, we propose two algorithms: (i)`Kernel-JB++` (Algorithm 2) and (ii) `Variational-JB++` (Algorithm 3). Both avoid computing the cluster means μ_h explicitly using different techniques. `Kernel-JB++` uses updates similar to kernel kmeans [13], but rather than requiring a kernel as its input, requires a specification of a JB divergence. On the other hand, `Variational-JB++` uses a variational characterization of the mean to avoid computing the cluster mean. The advantage that variational JB clustering has over Kernel-JB++ is that it updates are linear in the number of data-points, whereas for kernel-JB++ they are quadratic. Thus `Variational-JB++` is more suitable for large scale problems.

5.2.1 Kernel Kmeans for JB Clustering Recall that the mapping $\Phi : \Theta \mapsto \mathcal{H}$ induced by $\Delta_\phi(\cdot,\cdot)$ is not known, neither is the kernel $K(x,y) = \langle \Phi(x), \Phi(y) \rangle$. However, we know that the PD kernel $K(x,y)$ is isometric to the CPD kernel $C(x,y) = \frac{1}{2}\phi((x+y)/2)$ so that $\Delta_\phi(x,y) = K(x,x) + K(y,y) - 2K(x,y) = \frac{1}{2}\phi(x) + \frac{1}{2}\phi(y) - \phi((x+y)/2)$. Now, we note that one can run kernel kmeans [13] without knowing K by simply using any other PD kernel $\tilde{K}$ which is isometric to K, i.e., $K(x,x) + K(y,y) - 2K(x,y) = \tilde{K}(x,x) + \tilde{K}(y,y) - 2\tilde{K}(x,y)$. In particular, from the same initialization, kernel kmeans using K and $\tilde{K}$ will lead to the same final clustering. However, we cannot run kernel kmeans directly with the CPD kernel $C(x,y) = \frac{1}{2}\phi((x+y)/2)$ since kernel kmeans require the kernel to be PD. Hence, we focus on the question: is it possible to derive a isometric PD kernel $\tilde{K}$ from a CPD kernel C? The following remarkable result answers precisely this question:

Theorem 8 *Let $\mathcal{S}$ be a non-empty set. Let $C(\cdot,\cdot) : \mathcal{S} \times \mathcal{S} \mapsto \mathbb{R}$ be any symmetric CPD kernel and a kernel $K(\cdot,\cdot) : \mathcal{S} \times \mathcal{S} \mapsto R$ be defined by $K(x,y) = C(x,y) - C(x,a) - C(y,a) + C(a,a)$ for some fixed $a \in \mathcal{S}$. Then K is positive semidefinite and isometric w.r.t C.*

Proof. Berg et al., [4] proves that $C(x,y) - C(x,a) - C(y,a) + C(a,a)$ is positive semidefinite if C is CPD. It easily verified that $K(x,y)$, defined as above, leads to an isometric distance, by substituting the value of $K(x,y)$ in the expression $K(x,x) + K(y,y) - 2K(x,y)$. ∎

The point a acts as the origin in the sense that $K(x,a) = 0, \forall x$ and its choice is arbitrary. In our current context, for any $a \in \Theta$, we note that

$$\tilde{K}(x,y) = \frac{1}{2}\left\{\phi(\frac{x+y}{2}) - \phi(\frac{x+a}{2}) - \phi(\frac{y+a}{2}) + \phi(a)\right\} \tag{5.13}$$

Algorithm 2 Kernel-JB++

Choose initial cluster center μ_1 uniformly at random from $\mathcal{X}$.
Choose the next center $\mu_h = x \in \mathcal{X}, h = 2, \ldots, k$ with probability $\frac{\Delta(x)}{\sum_{x \in \mathcal{X}} \Delta(x)}$. Where $\Delta(x)$ is the maximum JB distance of x from the previous clusters.
Let $C^{(0)}$ be the initial clustering, $t = 0$
Compute $n \times n$ kernel matrix $\tilde{K}$ using (5.13)
Run kernel kmeans [13] till convergence with $\tilde{K}$ initialized with $C^{(0)}$

is positive definite and isometric to both C and K so that $\Delta_\phi(x, y) = \tilde{K}(x, x) + \tilde{K}(y, y) - 2\tilde{K}(x, y)$. Based on the above construction, we propose KernelJB++ (Algorithm 2) which uses $\tilde{K}$ as the kernel for kernel kmeans [13]. The algorithm chooses the initial clusters based on probabilistic farthest-first. At any point during the initialization, if M is the current set of cluster means, let $\Delta(x) = \min_{\mu \in M} \Delta_\phi(x, \mu)$.

Lemma 6 *Let $\mathcal{C}_{KJB}$ be the final clustering obtained from Kernel-JB++, and let J^*_Δ be the optimal value of the objective function. Then*

$$E[J_\Delta(\mathcal{C}_{KJB})] \leq 8(\log k + 2)J^*_\Delta \ .$$

Proof. First, note that the initialization leads to a clustering $C^{(0)}$ which satisfies the bound since the kmeans++ argument goes through for kernel-kmeans, even if we do not know the kernel $K(x, y)$ but have a way of computing the squared distance $\|\Phi(x) - \Phi(y)\|^2 = K(x, x) + K(y, y) - 2K(x, y)$. For JB++, the squared distance is simply $\Delta_\phi(x, y)$. ∎

5.2.2 Variational Kmeans for JB Clustering

For every iteration the kernel kmeans algorithm is $O(n^2)$ which can be slow for large datasets. We now present an alternative which is similar to traditional kmeans where each iteration is $O(n)$. We start with a variational characterization of $\Delta_\phi(x, y)$ in terms of

(5.14)
$$L_{\mathcal{C}_h}(\{s_i\}, \mu) = \frac{1}{2} \sum_{x_i \in \mathcal{C}_h} d_\phi(x_i, s_i) + \frac{1}{2} \sum_{x_i \in \mathcal{C}_h} d_\phi(\mu, s_i) \ .$$

Lemma 7 *For any clustering $\mathcal{C}_h$, and any $\mu, s_i \in \Theta$, we have $\sum_{x_i \in \mathcal{C}_h} \Delta_\phi(x_i, \mu) \leq L_{\mathcal{C}_h}(\{s_i\}, \mu)$. Further, $\sum_{x_i \in \mathcal{C}_h} \Delta_\phi(x_i, \mu) = \min_{\{s_i\}} L_{\mathcal{C}_h}(\{s_i\}, \mu)$.*

Proof. For any $x_i, \mu, s_i \in \Theta$, following [2, Proposition 1], $\frac{1}{2} d_\phi(x_i, s_i) + \frac{1}{2} d_\phi(\mu, s_i)$ is minimized by $s_i = \frac{1}{2}(x_i + \mu)$, and the minimum is $\Delta_\phi(x_i, \mu)$. Summing over all $x_i \in \mathcal{C}_h$, we have $\sum_{x_i \in \mathcal{C}_h} \Delta_\phi(x_i, \mu) \leq L_{\mathcal{C}_h}(\{s_i\}, \mu)$. Noting that the inequality holds with equality for $s_i = \frac{1}{2}(x_i + \mu)$ completes the proof. ∎

Algorithm 3 Variational-JB++

Choose initial cluster center $\mu_1^{(0)}$ uniformly at random from $\mathcal{X}$,
Choose the next center $\mu_h^{(0)} = x \in \mathcal{X}, h = 2, \ldots, k$ with probability $\frac{\Delta(x)}{\sum_{x \in \mathcal{X}} \Delta(x)}$. Set $t = 0$
repeat
For $i = 1, \ldots, n$, assign x_i to C_{h^*} if $h^* = \text{argmin}_h \Delta_\phi\left(x_i, \mu_h^{(t)}\right)$ (5.11)
For $h = 1, \ldots, k$, update

(5.15) $$s_i^{(t+1)} = \frac{x_j + \mu_h^{(t)}}{2} \ , \qquad \forall x_i \in C_h$$

(5.16) $$\mu_h^{(t+1)} = \nabla\phi^{-1}\left(\frac{\sum_{x_i \in \mathcal{C}_h} \nabla\phi(s_i^{(t+1)})}{|\mathcal{C}_h|}\right)$$

$$t = t + 1$$

until *convergence*

Recall that a key challenge in running kmeans-type iterations for JB divergences was that the optimal cluster prototype cannot be computed in closed form. The above result gives a variational approach to computing the cluster prototypes. We present Variational-JB++ in Algorithm 3 based on this approach. In particular, `Variational-JB++` uses the same initialization strategy as kmeans++, and then uses one-pass variational updates over individual s_i and μ_h for each cluster, which is guaranteed to improve the objective till convergence.

Lemma 8 *Let $\mathcal{C}_{VJB}$ be the final clustering obtained from Variational-JB++, and let J^*_Δ be the optimal value of the objective. Then $E[J_\Delta(\mathcal{C}_{VJB})] \leq 8(\log k + 2)J^*_\Delta$.*

Proof. As before, the initialization leads to a clustering $\mathcal{C}^{(0)}$ which satisfies the bound. We need to show that the iterative updates give a non-increasing objective function. Since the cluster assignment step improves the objective, we focus on the variational step. For any cluster $\mathcal{C}_h$, $\sum_{x_i \in \mathcal{C}_h} \Delta_\phi\left(x_i, \mu_h^{(t+1)}\right) \overset{(a)}{=} L_{\mathcal{C}_h}\left(\{s_i^{(t+1)}\}, \mu_h^{(t+1)}\right)$
$\overset{(b)}{\leq} L_{\mathcal{C}_h}\left(\{s_i^{(t+1)}\}, \mu_h^{(t)}\right) \overset{(c)}{\leq} L_{\mathcal{C}_h}\left(\{s_i^{(t)}\}, \mu_h^{(t)}\right)$
$\overset{(d)}{=} \sum_{x_i \in \mathcal{C}_h} \Delta_\phi\left(x_i, \mu_h^{(t)}\right)$ where (a), (d) follow from Lemma 7, (b) follows since that $\mu^{(t+1)}$ minimizes $L_{\mathcal{C}}(\{s_i^{(t+1)}\}, \mu)$, and (c) follows since $s_i^{(t+1)}$ minimizes $L_{\mathcal{C}_h}(\{s_i\}, \mu^{(t)})$. ∎

6 Conclusion

In this paper we introduce two families of Hilbert space embeddable metrics that can be generated from

Bregman divergences. On one hand these give the practitioner a fertile ground to search for metrics well suited for their application. On the other hand the finite embedabilty of the GSB family, and the kernel isometry property of the JB family allows them to be seamlessly incorporated into existing data mining algorithms. In addition, the variational representation of Bregman divergence allows the representation of non-linear cluster boundaries but with a linear complexity per iteration. On the theory side it establishes a connection between Bregman divergences and results concerning cumulants of infinitely divisible measures.

Acknowledgements: The first author thanks Suriya Gunasekar for suggesting improvements. We thank the anonymous reviewers for valuable comments and pointers. The research was supported in part by NSF CAREER award IIS-0953274, and NSF grants IIS-0916750, IIS-0812183, and IIS-1029711, and NASA grant NNX12AQ39A.

References

[1] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 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] A. Banerjee and N. Srivastava. Conditionally positive definite kernels and infinitely divisible distributions. Technical Report TR 08-034, Dept of CS&E, University of Minnesota, 2008.

[4] C. Berg, J. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*. Springer-Verlag, 1984.

[5] C. Berg and G. Forst. *Potential Theory on Locally Compact Abelian Groups*. Springer, 1975.

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

[7] L. Cayton. Fast nearest neighbor retrieval for Bregman divergences. In *Proceedings of the 25th International Conference on Machine Learning*, 2008.

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

[9] K. Chaudhuri and A. McGregor. Finding metric structure in information theoretic clustering. In *Proceedings of the 21st Annual Conference on Learning Theory*, 2008.

[10] P. Chen, Y. Chen, and M. Rao. Metrics defined by bregman divergences. *Communications in Mathematical Sciences*, 6(4):915–926, 2008.

[11] A. Cherian, S. Sra, A. Banerjee, and N. Papanikolopoulos. Efficient similarity search for covariance matrices via the jensen-bregman logdet divergence. In *ICCV*, pages 2399–2406, 2011.

[12] A. Devinatz. The representation of functions as Laplace-Stieltjes integrals. *Duke Mathematical Journal*, 24:481–498, 1955.

[13] I. Dhillon, Y. Guan, and B. Kulis. Kernel k-means, spectral clustering and normalized cuts. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004.

[14] W. Ehm, M. G. Genton, and T. Gneiting. Stationary covariances associated with exponentially convex functions. *Bernoulli*, 9(4):607–615, 2003.

[15] C. Elkan. Using the triangle inequality to accelerate k-means. In *ICML*, 2003.

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

[17] B. V. Gnedenko and A. N. Kolmogorov. *Limit distributions for sums of independent random variables*. Addison-Wesley, 1954.

[18] D. Haussler. Convolution kernels on discrete structures. Technical Report Technical Report UCSC-CRL-99-10, UC Sanra Cruz, 1999.

[19] M. Hein and O. Bousquet. Hilbertian metrics and positive definite kernels on probability measures. In *AISTATS*, 2005.

[20] S. I. Karpushev. Conditionally positive-definite functions on locally compact groups and the Levy-Khinchin formula. *Journal of Mathematical Sciences*, 28:489–498, 1985.

[21] E. Lehmann and J. Romano. *Testing Statistical Hypothesis*. Springer, 2005.

[22] G. Letac. *Lectures on natural exponential families and their variance functions*, volume vol. 50 of Monographias de Mathematica. Instituto de Mathematica pura a aplicada, 1992.

[23] F. Nielsen and R. Nock. Sided and symmetrized bregman centroids. *IEEE Transactions on Information Theory*, 55(6):2882–2904, 2009.

[24] J. P. Nolan. *Stable distributions - models for heavy tailed data*. Boston: Birkhauser, 2009. In progress, Chapter 1 online at academic2.american.edu/jpnolan.

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

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

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

[28] B. Scholkopf. The kernel trick for distances. In *NIPS*, 2000.

[29] B. Scholkopf and A. Smola. *Learning With Kernels*. MIT Press, 2001.