A Generalization of K-Means Clustering Using Bregman Divergences

Julie Zhang

June 2020

Contents

1 Introduction **2**

2 Preliminaries **3**
2.1 Convexity 3
2.2 Bregman Divergences 6

3 Bregman Hard Clustering **9**
3.1 Bregman Information 9
3.2 The Hard Clustering Algorithm 13

4 Relation to Exponential Families **16**
4.1 Legendre Duality 16
4.2 Exponential Families 19
4.3 The Bijection 20

5 Bregman Soft Clustering **28**
5.1 The Expectation-Maximization Algorithm 28
5.2 The Soft Clustering Algorithm 30
5.3 Clustering in the Second Argument 32

6 Conclusion **34**

References **35**

1 Introduction

With the rise of machine learning algorithms, one of the big focuses is on unsupervised learning, where we aim to learn some inherent patterns of unlabeled data. The goal of these clustering algorithms is simple, partition a set of n observations into $k \leq n$ sets that minimize some measure of variance, the most common being within cluster variance. That is, minimize the distance between each point and the nearest cluster center. This minimization problem is an NP-hard problem, so a 2-step iterative approach is taken to find a local minimum.

First, we random choose k points to be our cluster centers, they can be data points or not. Then, for each iteration t, we will update the k cluster centers in the following two steps.

1. Cluster Assignment: Find the cluster center that each observation is closest to and assign it to the corresponding cluster.

2. Cluster Update: Update the cluster center to be the mean of all points in that cluster. If the cluster is empty, no update is made.

Terminate the algorithm when there is no update to the cluster centers.

In many cases, squared Euclidean distance is used as the distance measure because the algorithm is guaranteed to converge to a minimum in a finite number of iterations. Euclidean distance is also a natural measure of distance between points. This is the classic K-Means algorithm, one of the simplest clustering algorithms. However, the Euclidean distance measurement may not be accurate of the true data generating process and yield results that do not make sense or are not meaningful. There is a natural question that arises: are there other distances measures that we can use in the above algorithm that will still guarantee convergence? The answer is yes. There are a collection of functions known as Bregman divergences and they are the only such functions.

In this paper, we will introduce a class of distortion functions known as Bregman divergences and analyze parametric clustering algorithms based off of these Bregman divergences. There are two types of clustering, named hard clustering and soft clustering. Hard clustering assigns each data point to exactly one cluster. In soft clustering, each data point is assigned a probability of belonging to a cluster, and therefore points can belong to multiple clusters. The hard clustering algorithm presented is a direct generalization of the K-Means algorithm that will be shown to converge. In addition, we establish a relationship between regular exponential families and Bregman divergences to develop an efficient EM scheme for learning mixtures of exponential family distributions that leads to a simple soft clustering algorithm. The primary source for this exposition is the paper *Clustering with Bregman Divergences* by Banerjee, Dhillon, Ghosh, and Merugu [2].

2 Preliminaries

In this section, we present some results on convex analysis as preparation. We then define the Bregman divergence and present some of its useful properties.

2.1 Convexity

We begin with a brief discussion of convexity of sets and functions.

Definition 2.1. A subset S of $\mathbb{R}^d$ is said to be **convex** if $(1-\lambda)x+\lambda y \in S$ for all $x, y \in S, \lambda \in [0,1]$. That is, the closed line segment between x and y lies in S. A **convex combination** of $x_1, ..., x_n \in S$ is $\sum_{j=1}^{n} \lambda_j x_j$ for $\lambda_j \geq 0, \sum_{j=1}^{n} \lambda_j = 1$.

Theorem 2.1. A set $S \subset \mathbb{R}^d$ is convex if and only if S contains all the convex combinations of elements in S.

Proof. We prove by induction on n, the number of elements in the convex combination. By definition, a set is convex if and only if it contains all the convex combinations with $n = 2$. Take $n > 2$. Suppose that S is closed under taking all convex combinations of fewer than n vectors. Let $x = \sum_{j=1}^{n} \lambda_j x_j$ be a convex combination. There must exist $\lambda_j \neq 1$ and we relabel to call it λ_1. Let $y = \sum_{j=2}^{n} \lambda'_j x_j$, where $\lambda'_j = \frac{\lambda_j}{1-\lambda_1}$. Then $\sum_{j=2}^{n} \lambda'_j = 1$ and y is a convex combination of $n-1$ elements. Hence $y \in S$ by the inductive hypothesis. Since $x = \lambda_1 x_1 + (1-\lambda_1)y$, we have $x \in S$. □

Definition 2.2. The intersection of all the convex sets containing $C \subset \mathbb{R}^d$ is called the **convex hull** of S. We denote this by $\operatorname{co}(S)$. It can be shown that $\operatorname{co}(S)$ is the set of all convex combinations of elements in S.

Definition 2.3. The **relative interior** of a convex set is

$$\operatorname{ri}(S) = \{x \in S | \ \forall\, y \in S, \ \exists\, \lambda > 1 : \lambda x + (1-\lambda)y \in S\}$$

Definition 2.4. Let $f : S \to \mathbb{R}$, where $S \subset \mathbb{R}^d$ is a convex set. f is a **convex function** on S if

$$f((1-\lambda)x+\lambda y) \leq (1-\lambda)f(x) + \lambda f(y)$$

for all $x, y \in S, \lambda \in [0,1]$. If the above inequality is strict for $\lambda \in (0,1)$, f is called **strictly convex**.

Proposition 2.2. A function $f : \mathbb{R}^d \to \mathbb{R}$ is (strictly) convex if and only if the function $g : \mathbb{R} \to \mathbb{R}$ given by $g(t) = f(x+ty)$ is (strictly) convex as a function of t for all $x \in \operatorname{dom}(f), y \in \mathbb{R}^d$ where $\operatorname{dom}(f)$ is the domain of f. The domain of g is $\{t : x + ty \in \operatorname{dom}(f)\}$.

Proof. If f is convex, and $t_1, t_2 \in \operatorname{dom}(g), c \in [0,1]$, we have

$$\begin{aligned} g((1-c)t_1 + ct_2) &= f\Big(x + ((1-c)t_1 + ct_2)y\Big) \\ &= f\Big((1-c)(x+t_1 y) + c(x + t_2 y)\Big) \\ &\leq (1-c)f(x+t_1 y) + cf(x+t_2 y) \\ &= (1-c)g(t_1) + cg(t_2) \end{aligned}$$

To prove the other direction, let $x, y \in \operatorname{dom}(f)$ and define $g(t) = f(x + t(y-x))$. We assume that all such functions are convex. For $\lambda \in [0,1]$, we have

$$\begin{aligned} g(\lambda) &\leq (1-\lambda)g(0) + \lambda g(1) \\ f(x+\lambda(y-x)) &\leq (1-\lambda)f(x) + \lambda f(y) \end{aligned}$$

Hence f is convex.

Notice that if we replace all $\leq$ with $<$, we obtain the result for strict convexity. □

Theorem 2.3. Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable over an open domain. The following are equivalent.

(a) f is convex.

(b) $f(y) - f(x) \geq \nabla f(x)^T (y - x)$ for all $x, y \in \mathrm{dom}(f)$.

(c) $\nabla^2 f(x)$ is positive semi-definite for all $x \in \mathrm{dom}(f)$.

Proof. We will prove (a) $\Leftrightarrow$ (b) and (b) $\Leftrightarrow$ (c). If f is convex, then for all $x, y \in \mathrm{dom}(f), \lambda \in [0, 1]$,

$$\begin{aligned} f((1-\lambda)x + \lambda y) &\leq (1-\lambda)f(x) + \lambda f(y) \\ f(x + \lambda(y - x)) &\leq f(x) + \lambda(f(y) - f(x)) \\ f(y) - f(x) &\geq \frac{f(x + \lambda(y - x)) - f(x)}{\lambda} \end{aligned}$$

Taking $\lambda \searrow 0$, we have $f(y) - f(x) \geq \nabla f(x)^T (y - x)$.

Suppose $f(y) - f(x) \geq \nabla f(x)^T (y - x)$ for all $x, y \in \mathrm{dom}(f)$. Take any $x, y \in \mathrm{dom}(f), \lambda \in [0, 1]$ and set $z = (1 - \lambda)x + \lambda y$. Then

$$\begin{aligned} f(x) &\geq f(z) + \nabla f(z)^T (x - z) \\ f(y) &\geq f(z) + \nabla f(z)^T (y - z) \\ (1-\lambda)f(x) + \lambda f(y) &\geq f(z) + \nabla f(x)^T ((1-\lambda)x + \lambda y - z) \\ &= f(z) \\ &= f((1-\lambda)x + \lambda y) \end{aligned}$$

Hence f is convex. This concludes (a) $\Leftrightarrow$ (b).

We first prove (b) $\Leftrightarrow$ (c) in the case $n = 1$. Suppose (b) holds for $x, y \in \mathrm{dom}(f), x < y$. Then

$$\begin{aligned} f(y) &\geq f(x) + f'(x)(y - x) \\ f(x) &\geq f(y) + f'(y)(x - y) \\ \Rightarrow f'(x)(y - x) &\leq f(y) - f(x) \leq f'(y)(y - x) \\ \frac{f'(y) - f'(x)}{y - x} &\geq 0 \end{aligned}$$

Taking $y \to x$, we have $f''(x) \geq 0$. Thus (c) holds.

To show the other direction, suppose $f''(x) \geq 0$ for all $x \in \mathrm{dom}(f)$. Let $x, y \in \mathrm{dom}(f)$, and WLOG $x < y$. By Taylor's Theorem, for some $c \in [x, y]$

$$\begin{aligned} f(y) &= f(x) + f'(x)(y - x) + \frac{f''(c)}{2}(y - x)^2 \\ \Rightarrow f(y) &\geq f(x) + f'(x)(y - x) \end{aligned}$$

For general $d > 1$, recall Proposition 2.2: $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if $g(t) = f(x_0 + tx)$ for all $x_0 \in \mathrm{dom}(f), x \in \mathbb{R}^d$. By above, this happens if and only if $g''(t) = x^T \nabla^2 f(x_0 + tx) x \geq 0$. Hence f is convex if and only if $\nabla^2 f(x)$ is positive semi-definite. □

The above result does not hold verbatim for strictly convex functions, as we will see in the following theorems.

Theorem 2.4. Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable over an open domain. Then f is strictly convex if and only if $f(y) - f(x) > \nabla f(x)^T (y - x)$ for all $x, y \in \mathrm{dom}(f)$.

Proof. Suppose $f(y) - f(x) > \nabla f(x)^T(y-x)$ for all $x, y \in \operatorname{dom}(f)$. By an analogous proof in the convex case above, f is strictly convex.

Suppose f is strictly convex. Then f is convex, and we must have $f(y) - f(x) \geq \nabla f(x)^T(y-x)$ for all $x, y \in \operatorname{dom}(f)$. Suppose that for the sake of contradiction, there exists $x \neq y$ such that $f(y) = f(x) + \nabla f(x)^T(y-x)$. Let $g(t) = f(x + t(y-x)) - f(x) - t\nabla f(x)^T(y-x)$. For all $t_1, t_2 \in [0,1], c \in [0,1]$, by strict convexity of f, we have

$$\begin{aligned} g((1-c)t_1 + ct_2) &= f(x + ((1-c)t_1 + ct_2)(y-x)) - f(x) - ((1-c)t_1 + ct_2)\nabla f(x)^T(y-x) \\ &= f\Big((1-c)(x + t_1(y-x)) + c(x + t_2(y-x))\Big) - f(x) - ((1-c)t_1 + ct_2)\nabla f(x)^T(y-x) \\ &< (1-c)f(x + t_1(y-x)) + cf(x + t_2(y-x)) - f(x) - ((1-c)t_1 + ct_2)\nabla f(x)^T(y-x) \\ &= (1-c)g(t_1) + cg(t_2) \end{aligned}$$

Hence g is strictly convex. Notice that, $g(0) = g(1) = 0$ and so

$$g(t) = g((1-t)\cdot 0 + t \cdot 1) < (1-t)g(0) + tg(1) = 0$$

However, we also have $g'(t) = \nabla f(x + t(y-x))^T(y-x) - \nabla f(x)^T(y-x)$ so $g'(0) = 0$. By the strict convexity of g, we have $g(t) \geq g(0) + g'(0)t = g(0)$ for all $t \in [0,1]$. Thus we have reached a contradiction, and we cannot have $f(y) = f(x) + \nabla f(x)^T(y-x)$ for some $x \neq y$. Thus $f(y) - f(x) > \nabla f(x)^T(y-x)$ for all $x, y \in \operatorname{dom}(f)$. □

Proposition 2.5. Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable over an open domain. If $\nabla^2 f(x)$ is positive definite for all $x \in \operatorname{dom}(f)$, then f is strictly convex.

Proof. Suppose $\nabla^2 f(x)$ is positive definite for all $x \in \operatorname{dom}(f)$. Let $x, y \in \operatorname{dom}(f)$, and WLOG $x < y$. By Taylor's Theorem, for some $c \in [x, y]$,

$$\begin{aligned} f(y) &= f(x) + f'(x)(y-x) + (y-x)^T \frac{\nabla^2 f(c)}{2}(y-x) \\ \Rightarrow f(y) &> f(x) + f'(x)(y-x) \end{aligned}$$

Therefore f is strictly convex. □

Remark. Unlike the case of convex functions, the converse of Proposition 2.5 does not hold. Consider $f(x) = x^4$ on $\mathbb{R}$. Then $f''(x) = 12x^2$ is not positive definite on $\mathbb{R}$, but $f(x)$ is strictly convex: for $x \neq y \in \mathbb{R}$,

$$\begin{aligned} f(y) - f(x) - f'(x)(y-x) &= y^4 - x^4 - 4x^3(y-x) \\ &= y^4 - x^4 - 4x^3y + 4x^4 \\ &= 3x^4 - 4x^3y + y^4 \\ &= (x-y)^2(2x^2 + (x+y)^2) \\ &> 0 \end{aligned}$$

Example 2.1. Examples of convex functions are $e^x, -\log x$ on $\mathbb{R}$, $x^a, a \geq 1$ or $a \leq 0$ on $x > 0$, and $x \log x$ on $x > 0$. Examples of multivariable convex functions include affine functions, $f(x) = a^T x + b$ on $\mathbb{R}$ for any $a \in \mathbb{R}^d, b \in \mathbb{R}$, norms, and real-valued linear transformations.

We present two useful corollaries.

Corollary 2.6. Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable. Then x_0 is a global minimizer of f if and only if $\nabla f(x_0) = 0$.

Proof. Since $\nabla f(x_0) = 0$ is a necessary condition for x_0 to be a global minimum, it suffices to show that if $\nabla f(x_0) = 0$, then x_0 is a global minimizer. By Theorem 2.3, for any $y \in \mathbb{R}^d$,

$$f(y) \geq f(x_0) + \nabla f(x_0)^T(y - x_0) = f(x_0)$$

Hence x_0 is a global minimizer.

If f is strictly convex, then $\nabla f(x_0) = 0$ implies x_0 is the unique global minimizer of $f(x)$, since $f(y) > f(x_0) + \nabla f(x_0)^T(y - x_0) = f(x_0)$. □

Corollary 2.7. Let $f : \mathbb{R} \to \mathbb{R}$ be twice differentiable. Then f is convex if and only if f' is increasing.

Proof. This follows directly from the second order conditions that $f''(x) \geq 0$. The assumption on f can be simplified to once differentiable, but the proof is not algebraically involved and will not be presented here. □

We conclude with a result due to Sierpinski.

Theorem 2.8. Let $f : S \to \mathbb{R}$ be continuous, where $S \subset \mathbb{R}^d$ is a convex set. If for all $x, y \in S$,

$$f\left(\frac{x+y}{2}\right) \leq \frac{f(x) + f(y)}{2} \tag{$*$}$$

then f is convex. We call a function f satisfying $(*)$ a **midpoint convex function**.

Proof. Since the condition for convexity holds for $\lambda = 1/2$, it holds for all dyadic numbers $\lambda \in (0, 1)$ by induction. The dyadic numbers are dense in $[0, 1]$, so by continuity, f is convex. □

There are many more interesting properties of convex functions that we will not discuss. For more on convex sets and functions, see Chapter 1 of Rockafellar's *Convex Analysis* [11].

2.2 Bregman Divergences

We are now ready to define the Bregman divergence, the basis of this exposition. This was first introduced by Bregman in 1967 to solve problems in linear and convex programming [5].

Definition 2.5. Let $\phi : S \to \mathbb{R}$ be a strictly convex function defined on a convex set $S \subset \mathbb{R}^d$ such that ϕ is differentiable on $\text{ri}(S) \neq \emptyset$. The **Bregman divergence** $d_\phi : S \times \text{ri}(S) \to [0, \infty)$ is defined as

$$d_\phi(x, y) = \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y) \rangle$$

Example 2.2. Squared Euclidean distance.

Let $\phi(x) = \|x\|^2$ on $\mathbb{R}^d$. It can be verified that ϕ is strictly convex using the Cauchy-Schwarz inequality. The corresponding Bregman divergence is

$$\begin{aligned} d_\phi(x, y) &= \|x\|^2 - \|y\|^2 - \langle x - y, \nabla\phi(y) \rangle \\ &= \|x\|^2 - \|y\|^2 - \langle x - y, 2y \rangle \\ &= \|x\|^2 - \|y\|^2 - 2\langle x, y \rangle + 2\|y\|^2 \\ &= \|x - y\|^2 \end{aligned}$$

This is the squared Euclidean distance.

Example 2.3. Kullback-Leibler Divergence.

The Kullback–Leibler divergence (also called relative entropy) is a measure of how one probability distribution is different from a second, reference probability distribution. We define the KL divergence between two random variables P, Q to be $E_P\left(\log \frac{p(x)}{q(x)}\right)$, where $p(x), q(x)$ are the probability density functions (pdfs) of P, Q respectively, where P, Q can be discrete or continuous. This is commonly used in information theory. Let $\mathbf{p} = (p_1, ..., p_d)$ be a discrete probability distribution: $\sum_{j=1}^d p_j = 1$. Define the negative entropy by $\phi(\mathbf{p}) = \sum_{j=1}^d p_j \log p_j$. This function is strictly convex since $x \log x$ is strictly convex using the 2nd order conditions. The corresponding Bregman divergence is

$$\begin{aligned}
d_\phi(\mathbf{p}, \mathbf{q}) &= \sum_{j=1}^d p_j \log p_j - \sum_{j=1}^d q_j \log q_j - \langle \mathbf{p} - \mathbf{q}, \nabla \phi(\mathbf{q}) \rangle \\
&= \sum_{j=1}^d p_j \log p_j - \sum_{j=1}^d q_j \log q_j - \sum_{j=1}^d (p_j - q_j)(\log q_j + 1) \\
&= \sum_{j=1}^d p_j \log\left(\frac{p_j}{q_j}\right) - \sum_{j=1}^d (p_j - q_j) \\
&= \sum_{j=1}^d p_j \log\left(\frac{p_j}{q_j}\right) \\
&= KL(\mathbf{p} \| \mathbf{q})
\end{aligned}$$

Example 2.4. Mahalanobis distance.

The Mahalanobis distance is used to calculate the distance from a point to a distribution, or to calculate a dissimilarity measure between two random vectors of the same distribution with covariance matrix A, which is assumed to be positive definite. Let $\phi(x) = x^T A x, x \in \mathbb{R}^d$ and A be positive definite. Then ϕ is strictly convex since the Hessian of ϕ is A. The corresponding Bregman divergence is the Mahalanobis distance.

$$\begin{aligned}
d_\phi(x, y) &= x^T A x - y^T A y - \langle (x - y), 2Ay \rangle \\
&= x^T A x - y^T A y - 2x^T A y + 2y^T A y \\
&= (x - y)^T A (x - y)
\end{aligned}$$

Bregman divergences have many useful properties, and we will prove several in the following proposition.

Proposition 2.9. Let d_ϕ be a Bregman divergence corresponding to $\phi : S \to \mathbb{R}$.

(a) **Non-negativity**: $d_\phi(x, y) \geq 0$ and equality holds if and only if $x = y$.

(b) **Convexity**: d_ϕ is convex in the first argument.

(c) **Linearity**: Bregman divergence as a map $\phi \mapsto d_\phi$ is linear.

(d) **Linear Separation**: The locus of all points $x \in S$ that are equidistant from two points $\mu_1, \mu_2 \in \text{ri}(S)$ is a subset of a hyperplane where distance is calculated by a Bregman divergence.

(e) **Equivalence Classes**: If $\phi(x) = \phi_0(x) + \langle b, x \rangle + c$ where $b \in \mathbb{R}^d, c \in \mathbb{R}$, then $d_\phi(x, y) = d_{\phi_0}(x, y)$ for all $x \in S, y \in \text{ri}(S)$.

Proof. (a) Since ϕ is strictly convex, for $x \neq y$, we have

$$\phi(x) - \phi(y) > \langle x - y, \nabla \phi(y) \rangle$$

which means $d_\phi(x, y) \geq 0$ for $x \neq y$. Therefore d_ϕ is nonnegative and $d_\phi(x, y) = 0$ if and only if $x = y$.

(b) Fix $y \in \mathrm{ri}(S)$ and let $x_1 \neq x_2 \in S$. Then

$$\begin{aligned}
d_\phi(x_1, y) - d_\phi(x_2, y) &= \phi(x_1) - \phi(x_2) - \langle x_1 - y, \nabla\phi(y)\rangle + \langle x_2 - y, \nabla\phi(y)\rangle \\
&= \phi(x_1) - \phi(x_2) - \langle x_1 - x_2, \nabla\phi(y)\rangle \\
\nabla d_\phi(x_2, y) &= \nabla\phi(x_2) - \nabla\phi(y) \\
d_\phi(x_1, y) - d_\phi(x_2, y) - \langle x_1 - x_2, \nabla d_\phi(x_2, y)\rangle &= \phi(x_1) - \phi(x_2) - \langle x_1 - x_2, \nabla\phi(y)\rangle \\
&\quad - \langle x_1 - x_2, \nabla\phi(x_2) - \nabla\phi(y)\rangle \\
&= \phi(x_1) - \phi(x_2) - \langle x_1 - x_2, \nabla\phi(x_2)\rangle \\
&\geq 0
\end{aligned}$$

The last equation holds by the strict convexity of ϕ. Thus d_ϕ is convex in the first argument.

(c) By the linearity of the inner product and gradient operators, we see for $c \geq 0$ and strictly convex functions ϕ_1, ϕ_2,

$$\begin{aligned}
d_{\phi_1+\phi_2}(x, y) &= d_{\phi_1}(x, y) + d_{\phi_2}(x, y) \\
d_{c\phi_1}(x, y) &= c d_{\phi_1}(x, y)
\end{aligned}$$

(d) Fix $\mu_1, \mu_2 \in \mathrm{ri}(S)$. Then all points $x \in S$ satisfying $d_\phi(x, \mu_1) = d_\phi(x, \mu_2)$ satisfies

$$\begin{aligned}
d_\phi(x, \mu_1) &= d_\phi(x, \mu_2) \\
\phi(x) - \phi(\mu_1) - \langle x - \mu_1, \nabla\phi(\mu_1)\rangle &= \phi(x) - \phi(\mu_2) - \langle x - \mu_2, \nabla\phi(\mu_2)\rangle \\
\langle x, \nabla\phi(\mu_2) - \nabla\phi(\mu_1)\rangle &= \phi(\mu_1) - \langle \mu_1, \nabla\phi(\mu_1)\rangle - \phi(\mu_2) + \langle \mu_2, \nabla\phi(\mu_2)\rangle
\end{aligned}$$

Since the right-hand side is a constant, the set of $x \in S$ is a subset of a hyperplane.

(e) We have

$$\begin{aligned}
d_\phi(x, y) &= \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle \\
&= \phi_0(x) + \langle x, b\rangle + c - \phi_0(y) - \langle y, b\rangle - c - \langle x - y, \nabla\phi_0(y) + b\rangle \\
&= \phi_0(x) - \phi_0(y) - \langle x - y, \nabla\phi_0(y)\rangle + \langle x, b\rangle - \langle y - b\rangle - \langle x - y, b\rangle \\
&= d_{\phi_0}(x, y)
\end{aligned}$$

Hence we can partition the space of strictly convex, differentiable functions on a convex set S into equivalence classes of the form $[\phi_0] = \{\phi : d_\phi(x, y) = d_{\phi_0}(x, y), x \in S, y \in \mathrm{ri}(S)\}$.

□

Example 2.5. Bregman divergences are not necessarily convex in the second arguments. Consider the function $\phi : (0, \infty) \to \mathbb{R}, \phi(x) = x^3$. Then $\phi(x)$ is strictly convex because $\phi''(x) = 6x > 0$ on $(0, \infty)$. The Bregman divergence is $d_\phi(x, y) = x^3 - y^3 - 3(x - y)y^2$. If we fix $x \in (0, \infty)$, and consider d_ϕ only as a function of y, then $d''_\phi(x, y) = 12y - 6x$. We do not have $d''_\phi(x, y) \geq 0$ for all y, so d_ϕ is not convex in the second argument.

3 Bregman Hard Clustering

We introduce a concept known as the Bregman information of a random variable. Then, we show the Bregman hard clustering problem is equivalent to an optimization problem of minimizing the loss in Bregman information and present the generalized version of the K-Means algorithm.

3.1 Bregman Information

Let X be a random variable that takes values on a finite set $\mathcal{X} = \{x_i\}_{i=1}^n \subset S \subset \mathbb{R}^d$ for some convex set S following a discrete probability measure ν. Let d_ϕ be a Bregman divergence corresponding to the strictly convex function ϕ on S.

Definition 3.1. The **Bregman information** of the random variable X for the Bregman divergence d_ϕ is

$$I_\phi(X) = \min_{s \in \mathrm{ri}(S)} E_\nu(d_\phi(X, s)) = \min_{s \in \mathrm{ri}(S)} \sum_{i=1}^n \nu_i d_\phi(x_i, s)$$

The optimal vector s that achieves the minimal value of $I_\phi(X)$ is called the **Bregman representative**, or the representative of X.

Quite surprisingly, the Bregman representative is uniquely determined and does not depend on d_ϕ.

Theorem 3.1. Let X be a random variable that takes values on a finite set $\mathcal{X} = \{x_i\}_{i=1}^n \subset S \subset \mathbb{R}^d$ for some convex S following a discrete probability measure ν such that $\mu = E_\nu(X) \in \mathrm{ri}(S)$. Let $d_\phi : S \times \mathrm{ri}(S) \to [0, \infty)$ be a Bregman divergence. The problem $\min_{s \in \mathrm{ri}(S)} E_\nu(d_\phi(x, s))$ has a unique minimizer given by $s = \mu = E_\nu(X)$.

Proof. First, the assumption that $E_\nu(X) \in \mathrm{ri}(S)$ is not restrictive because we have $E_\nu(X) \notin \mathrm{ri}(S)$ if and only if the convex hull of $\mathcal{X}$ is a subset of the boundary of S. This is not possible, so we have $E_\nu(X) \in \mathrm{ri}(S)$. Denote the objective function by $J_\phi(s) = \sum_{i=1}^n \nu_i d_\phi(x_i, s)$. Since $\mu \in \mathrm{ri}(S)$, the objective function is well-defined at μ. For all $s \in \mathrm{ri}(S)$, we have

$$\begin{aligned}
J_\phi(s) - J_\phi(\mu) &= \sum_{i=1}^n \nu_i d_\phi(x_i, s) - \sum_{i=1}^n \nu_i d_\phi(x_i, \mu) \\
&= \sum_{i=1}^n \nu_i \Big(\phi(\mu) - \phi(s) - \langle x_i - s, \nabla\phi(s) \rangle + \langle x_i - \mu, \nabla\phi(\mu) \rangle \Big) \\
&= \phi(\mu) - \phi(s) - \left\langle \sum_{i=1}^n \nu_i x_i - s, \nabla\phi(s) \right\rangle + \left\langle \sum_{i=1}^n \nu_i x_i - \mu, \nabla\phi(\mu) \right\rangle \\
&= \phi(\mu) - \phi(s) - \langle \mu - s, \nabla\phi(s) \rangle + \langle 0, \nabla\phi(\mu) \rangle \\
&= d_\phi(\mu, s) \\
&\geq 0
\end{aligned}$$

We have equality if and only if $\mu = s$ by Proposition 2.5(i). Hence μ is the unique minimizer of J_ϕ. □

The converse of Theorem 3.1 is also true, as stated in the next theorem. This is a very powerful result because Bregman divergences are now the only types of functions that satisfy the property in Theorem 3.1. We will use this in order to generalize K-Means algorithm.

Theorem 3.2. Let $F : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ be such that $F(x, x) = 0$ for all x. Assume all second-order partial derivatives $\frac{\partial^2 F}{\partial x_i \partial_j}, 1 \leq i, j \leq d$ are continuous. For all sets $S \subset \mathbb{R}^d$ and all probability measures μ over S, if the random variable X takes values in S following μ such that $y = E_\nu(X)$ is the unique minimizer of $E_\nu(F(X, y))$ over all $y \in \mathbb{R}^d$, that is if

$$\arg\min_{y\in\mathbb{R}^d} E_\nu(F(X, y)) = E_\nu(X) \tag{$*$}$$

then $F(x, y)$ is a Bregman divergence, i.e., $F(X, y) = d_\phi(x, y)$ for some strictly convex, differentiable function $\phi : \mathbb{R}^d \to \mathbb{R}$.

Proof. We will give the proof for the one-dimensional case $F : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$, where it suffices to suppose F_x, F_y are continuous. The multidimensional case requires more machinery and can be found in [3].

We first show that $F(x, y) = d_\phi(x, y)$ for a convex function ϕ and then prove ϕ must be strictly convex. We will give a constructive argument for a particular choice of S, ν. Let $S = \{a, b\} \subset \mathbb{R}, p \in [0, 1], q = 1 - p$. Define $\nu(\{a\}\}) = p, \nu(\{b\}) = q$. This is a probability measure over S. Then $E_\nu(X) = pa + qb$ and by $(*)$, we have

$$\begin{aligned} pF(a, y) + qF(b, y) &= E(F(X, y)) \\ &\geq E(F(X, E(X)) \\ &= pF(a, pa + qb) + qF(b, pa + qb) \end{aligned}$$

Let $g(y) = pF(a, y)+qF(b, y)$. Then we have equality in the above statements if and only if $y = y_0 = pa+qb$. Therefore, setting $p = \frac{y_0-b}{a-b}$, we have

$$\begin{aligned} 0 &= g'(y_0) \\ &= pF_y(a, y_0) + qF_y(b, y_0) \\ &= \frac{y_0 - b}{a - b}F_y(a, y_0) + \frac{a - y_0}{a - b}F_y(b, y_0) \\ \frac{F_y(a, y_0)}{y_0 - a} &= \frac{F_y(b, y_0)}{y_0 - b} \end{aligned}$$

Since a, b, p were chosen arbitrarily, $\frac{F_y(x,y)}{y-x}$ is independent of x and we can write $F_y(x, y) = (y - x)H(y)$ for some continuous function $H(y)$.

Define $\phi(y) = \int_0^y \int_0^t H(s)ds$. Then ϕ is twice differentiable and satisfies

$$\phi(0) = 0, \qquad \phi'(y) = \int_0^y H(s)ds \Rightarrow \phi'(0) = 0, \qquad \phi''(y) = H(y)$$

Using integration by parts, we have

$$\begin{aligned} F(x, y) - F(x, x) &= \int_x^y (t - x)H(t)dt \\ &= \Big((t - x)\int_0^t H(s)ds\Big)\Big|_{t=x}^{t=y} - \int_x^y \int_0^t H(s)dsdt \\ &= (y - x)\int_0^y H(s)ds - (\phi(y) - \phi(x)) \\ &= \phi(x) - \phi(y) - \phi'(y)(x - y) \end{aligned}$$

Since $F(x, x) = 0$, we see $F(x, y) = d_\phi(x, y)$ for some ϕ, and since $F(x, y) \geq 0$, ϕ is convex.

Suppose ϕ is not strictly convex. Then there exists $a < b \in \mathbb{R}$ such that $\phi(b) = \phi(a) + \phi'(a)(b - a)$, that is $\phi'(a) = \frac{\phi(b)-\phi(a)}{b-a}$. In addition, we know $\phi(a) \geq \phi(b) + \phi'(b)(a - b)$, so $\phi'(b) \leq \frac{\phi(a)-\phi(b)}{a-b} = \phi'(a)$. However, by Corollary 2.7, we know ϕ' must be increasing, so $\phi'(a) = \phi'(b) = \phi'(y)$ for all $y \in [a, b]$. Now take any $y \in [a, b]$. Since $d_\phi(x, y) \geq 0$, we know $E(d_\phi(X, y)) \geq 0$. If we find two $y \in [a, b]$ such that $E(d_\phi(x, y)) = 0$, then we are contradicting the assumption of uniqueness, and hence ϕ must be strictly convex. We have

$$\begin{aligned}
E(d_\phi(X, y)) &= E(\phi(X)) - \phi(y) - \phi'(y)E(X) + \phi'(y)y \\
&= \frac{\phi(a) + \phi(b)}{2} - \phi(y) - \phi'(y)\left(\frac{a+b}{2} - y\right) \\
&= \frac{\phi(a) + \phi(b)}{2} - \phi(y) - \left(\frac{\phi(b) - \phi(a)}{b-a}\right)\left(\frac{a+b}{2} - y\right) \\
&= \frac{(\phi(a) + \phi(b))(b-a) - 2\phi(y)(b-a) - (\phi(b) - \phi(a))(a+b) + 2y(\phi(b) - \phi(a))}{2(b-a)} \\
&= \frac{(\phi(a) - \phi(y))b + (\phi(y) - \phi(b))a + (\phi(b) - \phi(a))y}{b-a}
\end{aligned}$$

Then we see $E(d_\phi(X, a)) = E(d_\phi(X, b)) = 0$. Hence the uniqueness assumption is not satisfied, and ϕ must be strictly convex. Therefore, $F(x, y)$ must be a Bregman divergence. □

Remark. We assumed that $F(x, x) = 0$ for all $x \in \mathbb{R}^d$. This assumption is not restrictive because if $F(x, y)$ is a function satisfying $\arg\min_{y\in\mathbb{R}^d} E(F(X, y)) = E_\nu(X)$, then $G(x, y) = F(x, y) - F(x, x)$ also satisfies this property and $G(x, x) = 0$.

We now give the formal definition of the Bregman information.

Definition 3.2. Let X be a random variable that takes values on a finite set $\mathcal{X} = \{x_i\}_{i=1}^n \subset S \subset \mathbb{R}^d$ for some convex S following a discrete probability measure ν. Let $\mu = E_\nu(X) = \sum_{i=1}^n \nu_i x_i \in \text{ri}(S)$. Let $d_\phi : S \times \text{ri}(S) \to [0, \infty)$ be a Bregman divergence. The **Bregman information** of the random variable X for the Bregman divergence d_ϕ is

$$I_\phi(X) = E_\nu(d_\phi(X, \mu)) = \sum_{i=1}^n \nu_i d_\phi(x_i, \mu)$$

We work through a few examples for calculating Bregman Information.

Example 3.1. Variance and the squared Euclidean distance.

Let $\mathcal{X} = \{x_1, ..., x_n\} \subset \mathbb{R}^d$ and consider the uniform measure, $\nu_i = 1/n$ over $\mathcal{X}$. Let d_ϕ be the squared Euclidean distance. Then the Bregman information is

$$I_\phi(X) = \sum_{i=1}^n \nu_i d_\phi(x_i, \mu) = \frac{1}{n}\sum_{i=1}^n \|x_i - \mu\|^2$$

This is the sample variance.

Example 3.2. Mutual information and the KL-Divergence.

Let U, V be two discrete random variables with joint distribution $\{p(u_i, v_j) : i = 1, ..., n, j = 1, ..., m\}$. The mutual information between U, V is defined by

$$I(U; V) = \sum_{i=1}^n \sum_{j=1}^m p(u_i, v_j) \log \frac{p(u_i, v_j)}{p(u_i)p(v_j)}$$

Expanding the terms, we have

$$\begin{aligned} I(U;V) &= \sum_{i=1}^{n} p(u_i) \sum_{j=1}^{m} p(v_j|u_i) \log \frac{p(v_j|u_i)}{p(v_j)} \\ &= \sum_{i=1}^{n} p(u_i) KL\Big(p(V|u_i) \| p(V)\Big) \end{aligned}$$

Define a random variable Z_u that takes values in the set of probability distributions $\mathcal{Z}_u = \{p(V|u_i)\}_{i=1}^{n}$ following probability measure $\{\nu_i\}_{i=1}^{n} = \{p(u_i)\}_{i=1}^{n}$ over $\mathcal{Z}_u$. The mean of Z_u is

$$\mu = E_\mu(p(V|u)) = \sum p(u_i)p(V|u_i) = \sum_{i=1}^{n} p(u_i, V) = p(V)$$

If d_ϕ is the KL-Divergence, then the Bregman information of Z_u is

$$I_\phi(Z_u) = \sum_{i=1}^{n} \nu_i d_\phi(p(V|u_i), \mu) = \sum_{i=1}^{n} p(u_i) KL\Big(p(V|u_i) \| p(V)\Big) = I(U;V)$$

Therefore, the mutual information of U, V is the Bregman information of Z_u. Similarly, we can show

$$I(U;V) = \sum_{j=1}^{m} p(v_i) KL\Big(p(U|v_j) \| p(U)\Big) = I_\phi(Z_v)$$

where Z_v is a random variable that takes values in the set of probability distributions $\mathcal{Z}_v = \{p(U|v_i)\}_{j=1}^{m}$ following probability measure $\{\nu_j\}_{j=1}^{m} = \{p(v_j)\}_{j=1}^{m}$ over $\mathcal{Z}_v$.

Example 3.3. We can interpret the Bregman information as the difference in the values of Jensen's Inequality.

Proposition 3.3. Jensen's Inequality.

For any convex and differentiable function f and random variable X,

$$E(f(X)) \geq f(E(X))$$

Proof. By Theorem 2.3, we have

$$\begin{aligned} f(X) &\geq f(E(X)) + (X - E(X))f'(E(X)) \\ E(f(X)) &\geq E\Big(f(E(X)) + (X - E(X))f'(E(X))\Big) \\ &= f(E(X)) + f'(E(X))E\Big(X - E(X)\Big) \\ &= f(E(X)) \end{aligned}$$

□

Now, for a Bregman information $I_\phi(X)$, we have

$$\begin{aligned} I_\phi(X) &= E\Big(d_\phi(X, E(X))\Big) \\ &= E\left(\phi(X) - \phi(E(X)) - \langle X - E(X), \nabla\phi(E(X))\rangle\right) \\ &= E(\phi(X)) - \phi(E(X)) - E\left(\langle X - E(X), \nabla\phi(E(X))\rangle\right) \\ &= E(\phi(X)) - \phi(E(X)) \\ &\geq 0 \end{aligned}$$

3.2 The Hard Clustering Algorithm

Using the results in Theorem 3.1 and Theorem 3.2, we begin to formulate the hard clustering problem with Bregman divergences. If X is a random variable over a set $\mathcal{X} = \{x_1, ..., x_n\}$ following probability measure ν with large Bregman information, it makes sense to split $\mathcal{X}$ into smaller sets each with its own Bregman representative and find another random variable M that serves as an appropriate representation, or quantization of X. More specifically, let $\{\mathcal{X}_h\}_{h=1}^k$ be k disjoint partitions of $\mathcal{X}$, $\mathcal{M} = \{\mu_h = \sum_{x_i \in \mathcal{X}_h} \frac{\nu_i x_i}{\pi_h}\}_{h=1}^k$ be the set of representatives, and $\pi = \{\pi_h = \sum_{x_i \in \mathcal{X}_h} \nu_i\}_{h=1}^k$ be the induced probability measure on $\mathcal{M}$. The induced variable M takes values in $\mathcal{M}$ following π. We can think of cluster representatives μ_h as the "cluster centers".

We can measure the quality of the quantization M in two ways. The first way is to calculate the expected Bregman divergence between X and M. Since M is a deterministic function of X, we have

$$\begin{aligned} E_{X,M}(d_\phi(X, M)) = E_X(d_\phi(X, M)) &= \sum_{h=1}^k \sum_{x_i \in \mathcal{X}_h} \nu_i d_\phi(x_i, \mu_h) \\ &= \sum_{h=1}^k \pi_h \sum_{x_i \in \mathcal{X}_h} \frac{\nu_i}{\pi_h} d_\phi(x_i, \mu_h) \\ &= E_\pi\Big(I_\phi(X_h)\Big) \end{aligned}$$

where X_h is the random variable with values in $\mathcal{X}_h$ following a probability distribution ν_i/π_h. We see this quantity is equal to the expected Bregman information of the partitions.

The second way is to calculate the **loss in Bregman information** due to the quantization, defined by

$$L_\phi(M) = I_\phi(X) - I_\phi(M)$$

Since the choice of k is not clear, different quantization M's result from different k. If $k = 1$, then with probability 1 we pick $E_\nu(X)$ and the loss is $I_\phi(X)$. If $k = n$, then $M = X$ and the loss is 0. The following theorem states these two quantities are the same.

Theorem 3.4. Let X be a random variable that takes values in $\mathcal{X} = \{x_i\}_{i=1}^n \subset S \subset \mathbb{R}^d$ following the positive probability measure ν. Let $\{\mathcal{X}_h\}_{h=1}^k$ be a partitioning of $\mathcal{X}$, where $1 \leq k \leq n$ and let $\pi_h = \sum_{x_i \in \mathcal{X}_h} \nu_i$ be the induced probability measure π on the partitions. For $h = 1, ..., k$, let X_h be the random variable that takes values in $\mathcal{X}_h$ following ν_i/π_h for $x_i \in \mathcal{X}_h$. Let $\mathcal{M} = \{\mu_h\}_{h=1}^k, \mu_h \in \text{ri}(S)$ be the set of representatives of $\{X_h\}_{h=1}^k$ and M be a random variable that takes values in $\mathcal{M}$ following π. Then

$$L_\phi(M) = I_\phi(X) - I_\phi(M) = E_\pi\Big(I_\phi(X_h)\Big) = \sum_{h=1}^k \pi_h \sum_{x_i \in \mathcal{X}_h} \frac{\nu_i}{\pi_h} d_\phi(x_i, \mu_h)$$

Proof. Recall $\mu_h = \sum_{x_i \in \mathcal{X}_h} \frac{\nu_i x_i}{\pi_h}$. We can directly calculate.

$$\begin{aligned} I_\phi(X) &= \sum_{i=1}^n \nu_i d_\phi(x_i, \mu) \\ &= \sum_{h=1}^k \sum_{x_i \in \mathcal{X}_h} \nu_i \Big(\phi(x_i) - \phi(\mu) - \langle x_i - \mu, \nabla\phi(\mu)\rangle\Big) \\ &= \sum_{h=1}^k \sum_{x_i \in \mathcal{X}_h} \nu_i \Big(\phi(x_i) - \phi(\mu_h) - \langle x_i - \mu_h, \nabla\phi(\mu_h)\rangle + \langle x_i - \mu_h, \nabla\phi(\mu_h)\rangle\Big) + \\ &\quad \sum_{h=1}^k \sum_{x_i \in \mathcal{X}_h} \nu_i \Big(\phi(\mu_h) - \phi(\mu) - \langle x_i - \mu_h, \nabla\phi(\mu)\rangle - \langle \mu_h - \mu, \nabla\phi(\mu)\rangle\Big) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{h=1}^{k} \sum_{x_i \in \mathcal{X}_h} \nu_i \Big(d_\phi(x_i, \mu_h) + d_\phi(\mu_h, \mu) + \langle x_i - \mu_h, \nabla\phi(\mu_h) - \nabla\phi(\mu) \rangle \Big) \\
&= \sum_{h=1}^{k} \pi_h \sum_{x_i \in \mathcal{X}_h} \frac{\nu_i}{\pi_h} d_\phi(x_i, \mu_h) + \sum_{h=1}^{k} \sum_{x_i \in \mathcal{X}_h} \nu_i d_\phi(\mu_h, \mu) + \\
&\quad \sum_{h=1}^{k} \pi_h \sum_{x_i \in \mathcal{X}_h} \frac{\nu_i}{\pi_h} \langle x_i - \mu_h, \nabla\phi(\mu_h) - \nabla\phi(\mu) \rangle \\
&= \sum_{h=1}^{k} \pi_h I_\phi(X_h) + \sum_{h=1}^{k} \pi_h d_\phi(\mu_h, \mu) + \sum_{h=1}^{k} \pi_h \left\langle \sum_{x_i \in \mathcal{X}_h} \frac{\nu_i (x_i - \mu_h)}{\pi_h}, \phi(\mu_h) - \nabla\phi(\mu) \right\rangle \\
&= E_\pi \Big(I_\phi(X_h) \Big) + I_\phi(M) + \sum_{h=1}^{k} \pi_h \langle \mu_h - \mu_h, \phi(\mu_h) - \nabla\phi(\mu) \rangle
\end{aligned}
$$

Therefore, we exactly have $I_\phi(X) - I_\phi(M) = E_\pi\Big(I_\phi(X_h)\Big)$. □

Since $I_\phi(X)$ is the total Bregman information, we can interpret $I_\phi(M)$ as the between cluster Bregman information and hence $L_\phi(M)$ is the within cluster Bregman information. Analaysis of variance (ANOVA) is the special case with this variance as the Bregman information (Example 3.1). We now define the Bregman hard clustering problem as the problem of finding a partitioning of $\mathcal{X}$ (finding a random variable M) to minimize $L_\phi(M) = I_\phi(X) - I_\phi(M)$, the loss in Bregman information.

Algorithm 1: Bregman Hard Clustering

```
Data: A set $\mathcal{X} = \{x_i\}_{i=1}^n \subset S \subset \mathbb{R}^d$; a probability measure $\nu$ over $\mathcal{X}$; a Bregman divergence
      $d_\phi : S \times \mathrm{ri}(S) \to \mathbb{R}$; a natural number $k$ (the number of clusters)
Result: A local minimizer $M^*$ of $L_\phi(M)$ where $\mathcal{M} = \{\mu_h\}_{h=1}^k$ and $M$ is the induced random
        variable; a partitioning $\{\mathcal{X}_h\}_{h=1}^k$ of $\mathcal{X}$

Choose the representatives: Initialize $\mathcal{M} = \{\mu_h\}_{h=1}^k$ with $\mu_h \in \mathrm{ri}(S)$.
opt ← False
while not opt do
    "The Assignment Step"
    $\mathcal{X}_h \leftarrow \emptyset, 1 \le h \le k$
    for i=1 to n do
        $h = h^*(x_i) \leftarrow \arg\min_{h'} d_\phi(x_i, \mu_{h'})$
        $\mathcal{X}_h \leftarrow \mathcal{X}_h \cup \{x_i\}$
    end
    for h=1 to k do
        "The Estimation Step"
        $\pi_h \leftarrow \sum_{x_i \in \mathcal{X}_h} \nu_i$
        $\mu_h \leftarrow \frac{1}{\pi_h} \sum_{x_i \in \mathcal{X}_h} \nu_i x_i$
        $\mathcal{M}^* \leftarrow \{\mu_h\}_{h=1}^k$
    end
    if $\mathcal{M}^* == \mathcal{M}$ then
        opt ← True
    else
        $\mathcal{M} \leftarrow \mathcal{M}^*$
    end
end
return $\mathcal{M}^*, \{\mathcal{X}_h\}_{h=1}^k$
```

Theorem 3.5. The Bregman hard clustering algorithm monotonically decreases the loss function $L_\phi(M)$ and hence produces a local minimizer of $\min_M L_\phi(M)$. In addition, it terminates in a finite number of iterations to a partition that is locally optimal.

Proof. At iteration t, let $\{\mathcal{X}_h^{(t)}\}_{h=1}^k$ be the partitioning of $\mathcal{X}$ and $\mathcal{M}^{(t)} = \{\mu_h^{(t)}\}_{h=1}^k$ be the corresponding set of cluster representatives. Recalling the definition of $h^*(x_i)$ in the assignment step, we have

$$L_\phi(M^{(t)}) = \sum_{h=1}^k \sum_{x_i \in \mathcal{X}_h^{(t)}} \nu_i d_\phi(x_u, \mu_h^{(t)}) \geq \sum_{h=1}^k \sum_{x_i \in \mathcal{X}_h^{(t)}} \nu_i d_\phi(x_u, \mu_{h^*(x_i)}^{(t)})$$

By the estimation step, we have

$$\sum_{h=1}^k \sum_{x_i \in \mathcal{X}_h^{(t)}} \nu_i d_\phi(x_u, \mu_{h^*(x_i)}^{(t)}) \geq \sum_{h=1}^k \sum_{x_i \in \mathcal{X}_h^{(t+1)}} \nu_i d_\phi(x_u, \mu_h^{(t+1)}) = L_\phi(M^{(t+1)})$$

Thus $L_\phi(M)$ is monotonically decreasing. If we have equality at any step, this implies the clusters did not change and hence the algorithm will terminate.

Since the number of distinct partitioning of n objects into k clusters is finite and the algorithm objective monotonically decreases, the algorithm must terminate in a finite number of iterations to a solution that is locally optimal. □

Remark. We note some useful properties of the Bregman hard clustering algorithm.

(a) Linear separators: By Proposition 2.5(d), the partitions induced by the Bregman hard clustering algorithm are separated by hyperplanes.

(b) Scalability: The computational complexity of the algorithm scales linearly with n and k, making it appropriate for large clustering problems.

(c) Mixed data types: Since the Bregman divergence is linear, this algorithm is very applicable to mixed data types. One can choose d_ϕ corresponding to a convex combination of convex functions that are each appropriate for a subset of the features.

In light of Theorems 3.1 and 3.2, we see the hard clustering algorithm with cluster centroids as the optimal representatives works if and only if the distance measure taken is a Bregman divergence. This holds because the mean, or the expectation, is the best predictor only for Bregman divergences. This special family of functions becomes the only possible extension of the K-Means algorithm. However, there are similar clustering algorithms that one can use with distance metrics such as the L^1 norm, but the optimal cluster representative will not be the mean. For example, the K-Medians algorithm uses the median as cluster representative and the L^1 norm as the distance metric, and this algorithm can be shown to converge to a local minimum.

4 Relation to Exponential Families

In this section, we will present some results on Legendre duality and introduce exponential families, an important class of probability distributions that are widely studied in statistics. We will establish an one-to-one correspondence between regular exponential families and regular Bregman divergences. This will provide the basis for the soft clustering algorithm presented in Section 5. We assume the reader has some familiarity with measure theory.

4.1 Legendre Duality

We present some more results on convex functions, including the epigraph and the subgradient, and introduce the Legendre function to establish a notion of Legendre duality.

Definition 4.1. Let $f : \mathbb{R}^d \to \mathbb{R}$. The **epigraph** of f is defined to be

$$\mathrm{epi}(f) = \{(x, y) : x \in \mathbb{R}^d, y \in \mathbb{R}, f(x) \leq y\}$$

Proposition 4.1. Let $S \subset \mathbb{R}^d$ be a convex set. A function $f : S \to \mathbb{R}$ is convex if and only if its epigraph is convex.

Proof. Suppose f is convex. Let $(x_1, y_1), (x_2, y_2) \in \mathrm{epi}(f), \lambda \in (0, 1)$ and let $(x', y') = (1 - \lambda)(x_1, y_1) + \lambda(x_2, y_2)$. Then

$$\begin{aligned} y' &= (1-\lambda)y_1 + \lambda y_2 \\ &\geq (1-\lambda)f(x_1) + \lambda f(x_2) \\ &\geq f\Big((1-\lambda)x_1 + \lambda x_2)\Big) \\ &= f(x') \end{aligned}$$

Hence $\mathrm{epi}(f)$ is convex.

Suppose $\mathrm{epi}(f)$ is convex. Let $x_1, x_2 \in S, \lambda \in (0, 1)$. The points $(x_1, f(x_1)), (x_2, f(x_2))$ in the epigraph of f, so $(1 - \lambda)(x_1, f(x_1)) + \lambda(x_2, f(x_2)) = \Big((1 - \lambda)x_1 + \lambda x_2, (1 - \lambda)f(x_1) + \lambda f(x_2)\Big) \in \mathrm{epi}(f)$. Therefore, $f\Big((1 - \lambda)x_1 + \lambda x_2\Big) \leq (1 - \lambda)f(x_1) + \lambda f(x_2)$ and f is convex. □

Definition 4.2. A convex function f is **proper** if $\mathrm{dom}(\psi)$ is nonempty. A convex function is **closed** if it is lower semi-continuous, that is, all of its sublevel sets, $\{x \in \mathrm{dom}(f) : f(x) \leq y\}$ for $y \in \mathbb{R}$, are closed. This definition is also equivalent to $\mathrm{epi}(f)$ being closed.

Definition 4.3. Let ψ be a real-valued function on $\mathbb{R}^d$. Its **conjugate function** ψ^* is given by

$$\psi^*(t) = \sup_{\theta \in \mathrm{dom}(\psi)} \Big\{\langle t, \theta\rangle - \psi(\theta)\Big\}$$

Proposition 4.2. The conjugate function ψ^* is a convex function and $\psi(x) \geq \psi^{**}(x)$.

Proof. Let $y, z \in \mathrm{dom}(\psi^*)$ and $\lambda \in (0, 1)$. Then

$$\begin{aligned} \psi^*\Big((1-\lambda)y + \lambda z\Big) &= \sup_{x \in \mathrm{dom}(\psi)} \Big\{\langle (1-\lambda)y + \lambda z, x\rangle - \psi(x)\Big\} \\ &\leq (1-\lambda) \sup_{x \in \mathrm{dom}(\psi)} \Big\{\langle y, x\rangle - \psi(x)\Big\} + \lambda \sup_{x \in \mathrm{dom}(\psi)} \Big\{\langle z, x\rangle - \psi(x)\Big\} \\ &= (1-\lambda)\psi^*(y) + \lambda\psi^*(z) \end{aligned}$$

Hence ψ^* is always convex. For the second part, for $x \in \operatorname{dom}(\psi), y \in \operatorname{dom}(\psi^*)$,

$$\begin{aligned} \psi^*(y) &\geq \langle y, x\rangle - \psi(x) \\ \psi(x) &\geq \langle y, x\rangle - \psi^*(y) \\ \Rightarrow \psi(x) &\geq \sup_{y \in \operatorname{dom}(\psi^*)} \Big\{ \langle y, x\rangle - \psi^*(y) \Big\} \\ &= \psi^{**}(x) \end{aligned}$$

This shows $\psi(x) \geq \psi^{**}(x)$ for all x. □

Theorem 4.3. If ψ is a proper, closed, convex function, then ψ^* is too and $\psi^{**} = \psi$.

Proof. Suppose ψ is proper, closed and convex. Then ψ^* is proper and convex, so it remains to show ψ^* is closed. We will show that all sublevel sets $L_\lambda = \{y : \psi^*(y) \leq \lambda\}$ are closed. Let $\{y_n\}$ be a sequence in L_λ such that $y_n \to y \in \operatorname{dom}(\psi^*)$. Then for all $x \in \operatorname{dom}(\psi)$ and for all n, we have

$$\begin{aligned} &\langle y_n, x\rangle - \psi(x) \leq \psi^*(y_n) \leq \lambda \\ \Rightarrow &\langle y, x\rangle - \psi(x) \leq \lambda \\ \Rightarrow &\psi^*(y) = \sup_{x \in \operatorname{dom}(\psi)} \{\langle y, x\rangle - \psi(x)\} \leq \lambda \end{aligned}$$

Thus $y \in L_\lambda$ and L_λ is closed.

We prove by contradiction that $\psi = \psi^{**}$. We showed in Proposition 4.2 that $\psi(x) \geq \psi^{**}(x)$, so suppose there exists x such that $\psi^{**}(x) < \psi(x)$. Since ψ is closed and convex, $\operatorname{epi}(\psi)$ is closed and convex. By the Hyperplane Separation Theorem, there exists a hyperplane in $\mathbb{R}^{n+1}$ that strictly separates $(x, \psi^{**}(x))$ from $\operatorname{epi}(\psi)$. This hyperplane cannot be vertical by the shape of the epigraph, so we can normalize the normal vector of the hyperplane to be 1 in the vertical component. Therefore, there exists $\epsilon > 0$ and $y \in \mathbb{R}^n$, the non-vertical component, such that for all $z \in \mathbb{R}^n$,

$$\begin{aligned} \psi(z) + \epsilon &\geq \langle y, z - x\rangle + \psi^{**}(x) \\ \langle y, x\rangle - \psi^{**}(x) + \epsilon &\geq \langle y, z\rangle - \psi(z) \end{aligned}$$

Taking the supremum over z, we have

$$\begin{aligned} \langle y, x\rangle - \psi^{**}(x) + \epsilon &\geq \psi^*(y) \\ \langle y, x\rangle - \psi^*(y) + \epsilon &\geq \psi^{**}(x) \\ &= \sup_w \{\langle w, x\rangle - \psi^*(w)\} \end{aligned}$$

This a contradiction to the definition of supremum. Thus we have $\psi = \psi^{**}$. □

To prove our main result of Legendre duality, we introduce the concept of subgradient for convex functions. Subgradients are a generalization of the derivative to non-differentiable convex functions.

Definition 4.4. Let $\psi : S \to \mathbb{R}$ be a convex function, where $S \subset \mathbb{R}^n$ is a convex set. A vector $v \in \mathbb{R}^n$ is a **subgradient** for $x_0 \in S$ for all $x \in S$, we have

$$\psi(x) \geq \langle v, x - x_0\rangle + \psi(x_0)$$

We will let $\partial\psi(x)$ denote the set of subgradients of ψ at x.

Notice that ψ is differentiable if and only if $\partial\psi(x) = \{v\}$ contains one element, namely $v = \nabla\psi(x)$. In addition, a point x_0 is a minimizer of a convex function ψ if and only if ψ is subdifferentiable at x_0 and $0 \in \partial\psi(x_0)$. This follows directly from the fact that $\psi(x) \geq \psi(x_0)$ for all x. If ψ is differentiable, this reduces to the case that we know, $\nabla\psi(x_0) = 0$.

Theorem 4.4. If ψ is a proper, closed, convex function, then the following are equivalent.

(a) $y \in \partial\psi(x)$.

(b) $x \in \partial\psi^*(y)$.

(c) $\langle x, y\rangle = \psi(x) + \psi^*(y)$.

Proof. We will show (a)⇒(c)⇒(b)⇒(a).

If $y \in \partial\psi(x)$, then by the closed and convexity of ϕ,

$$\psi^*(y) = \sup_{v\in\text{dom}(\psi)} \Big\{\langle y, v\rangle - \psi(v)\Big\} = \max_{v\in\text{dom}(\psi)} \Big\{\langle y, v\rangle - \psi(v)\Big\}$$

We know that $v^* = x$ is the global minimizer if and only if $0 \in \partial(\langle y, x\rangle - \psi(x)) = y - \partial\psi(x)$, which is equivalent to $y \in \partial\psi(x)$. Therefore $\psi^*(y) = \langle x, y\rangle - \psi(x)$ and $\langle x, y\rangle = \psi(x) + \psi^*(y)$.

Suppose $\langle x, y\rangle = \psi(x) + \psi^*(y)$. For any u, we have

$$\begin{aligned}\psi^*(u) &= \sup_{v\in\text{dom}(\psi)} \Big\{\langle v, u\rangle - \psi(v)\Big\}\\ &\geq \langle x, u\rangle - \psi(x)\\ &= \langle u - y, x\rangle + \langle y, x\rangle - \psi(x)\\ &= \langle u - y, x\rangle + \psi^*(y)\end{aligned}$$

Hence $x \in \partial\psi^*(y)$ by definition of subgradient.

To show the last implication, since $\psi(x) = \psi^{**}(x)$, we repeat the argument in the first implication and get

$$x \in \partial\psi^*(y) \Rightarrow y \in \partial\psi^{**}(x) = \partial\psi(x)$$

Thus the claim is true. □

Definition 4.5. Let ψ be a proper, closed, convex function with $\Theta = \text{int}(\text{dom}(\psi))$. The pair (Θ, ψ) is called a **convex function of Legendre type** or a **Legendre function** if the following hold.

(a) Θ is nonempty.

(b) ψ is strictly convex and differentiable on θ.

(c) For all $\theta_b \in \text{bd}(\Theta)$, $\lim_{\theta\to\theta_b} \|\nabla\psi(\theta)\| \to \infty$ where $\theta \in \Theta$.

Proposition 4.5. If ψ is proper, closed, convex, then ψ^* is differentiable and

$$\nabla\psi^*(t) = \arg\max_{x\in\text{dom}(\psi)} \{\langle t, x\rangle - \psi(x)\}$$

Proof. By Theorem 4.4, x maximizes $\langle x, y\rangle - \psi(x)$ if and only if $y \in \partial\psi(x)$, which is equivalent to $x \in \partial\psi^*(y)$. Since ψ is strictly convex and differentiable on $\Theta = \text{int}(\text{dom}(\psi))$, there is a unique minimizer, so $\partial\psi^*(y) = \{\nabla\psi^*(y)\} = \{x\}$. hence ψ^* is differentiable and $\nabla\psi^*(y) = \arg\max_{x\in\text{dom}(\psi)}\{\langle y, x\rangle - \psi(x)\}$. In addition, $\psi^*(t)$ attains its supremum at the unique θ satisfying $t = \nabla\psi(\theta)$. □

The results presented in this section are the setup for this main theorem on Legendre duality. This theorem will be used to establish the one-to-one correspondence of the natural parameter and expectation of exponential families in Section 4.2. The gist of the argument is mostly given above. The full proof is given in Section 26 of [11].

Theorem 4.6. Let ψ be a real-valued proper, closed, convex, differentiable, function with conjugate function ψ^*. Let $\Theta = \mathrm{int}(\mathrm{dom}(\psi))$ and $\Theta^* = \mathrm{int}(\mathrm{dom}(\psi^*))$. If (Θ, ψ) is a Legendre function, then

(a) (Θ^*, ψ^*) is a Legendre function and (Θ, ψ) and (Θ^*, ψ^*) are called Legendre duals of each other.

(b) The gradient function $\nabla\psi : \Theta \to \Theta^*$ is an injective function from the open convex set Θ onto the open convex set Θ^*.

(c) The gradient functions $\nabla\psi, \nabla\psi^*$ are continuous and

$$\nabla\psi^*(t) = (\nabla\psi)^{-1}(t) = \arg \max_{x \in \mathrm{dom}(\psi)} \{\langle t, x\rangle - \psi(x)\}$$

4.2 Exponential Families

Let $(\Omega, \mathcal{B})$ be a measurable space and let $t : \Omega \to T \subset \mathbb{R}^d$ be measurable. We may have T discrete. Let $p_0 : T \to \mathbb{R}^+$ be any function such that $dP_0(\omega) = p_0(t(\omega))dt(\omega)$ is a measure on $(\Omega, \mathcal{B})$ and $\int_{\omega\in\Omega} dP_0(\omega) < \infty$. If T is a discrete set, $dt(\omega)$ is the counting measure and P_0 is absolutely continuous with respect to the counting measure. If T is not discrete, then P_0 is absolutely continuous with respect to the Lebesgue measure $dt(\omega)$. We have that $t(w)$ is a random variable from $(\Omega, \mathcal{B}, P_0)$ to $(T, \sigma(T))$, where $\sigma(T)$ is the $\sigma-$algebra generated by T. Let Θ be the set of all $\theta \in \mathbb{R}^d$ such that

$$0 < \int_{\omega\in\Omega} e^{\langle\theta, t(\omega)\rangle} dP_0(\omega) = \int_{\omega\in\Omega} e^{\langle\theta, t(\omega)\rangle} p_0(t(\omega))dt(\omega) < \infty$$

Then, it is possible to define a function $\psi : \Theta \to \mathbb{R}$ such that

$$\psi(\theta) = \log\left(\int_{\omega\in\Omega} e^{\langle\theta, t(\omega)\rangle} dP_0(\omega)\right)$$

Definition 4.6. A family of probability distributions F_ψ parametrized by $\theta \in \Theta \subset \mathbb{R}^d$ where the probability density functions with respect to the measure $dt(\omega)$ can expressed in the form

$$f(\omega, \theta) = e^{\langle\theta, t(\omega)\rangle - \psi(\theta)} p_0(t(\omega))$$

is called an **exponential family** with natural statistic $t(\omega)$, natural parameter θ, and natural parameter space Θ. It is clear $f(\omega, \theta)$ of this form are indeed probability density functions with respect to $dt(\omega)$, since $\int_{\omega\in\Omega} f(\omega, \theta)dt(\omega) = 1$. Notice $e^{-\psi(\theta)}$ is the normalizing constant.

If the components of $t(\omega)$ are affinely independent, that is there exists a nonzero $a \in \mathbb{R}^d$ such that $P_0(\{\omega : \langle t(\omega), a\rangle = c\}) = 1$, for all $\omega \in \Omega$, then this representation is said to be **minimal**. For a minimal representation, there exists a unique probability density function $f(\omega, \theta)$ for each $\theta \in \Theta$ and we call F_ψ a **full exponential family** of order d.

If the natural parameter space Θ is open, then we call F_ψ a **regular exponential family**.

Letting $x = t(\omega)$, the probability density function $g(x, \theta)$ with respect to the measure dx given by $g(x, \theta) = e^{\langle\theta, x\rangle - \psi(\theta)} p_0(x)$ has the property that $\frac{f(\omega,\theta)}{g(x,\theta)}$ is $\theta-$free. It can be shown that x is a minimally sufficient statistic for the family F_ψ [4].

Example 4.1. Gaussian Distributions.

The probability density functions for the one-dimensional Gaussian distribution is given by

$$f(\omega, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\omega-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\omega^2 - 2\omega\mu)}{2\sigma^2}} e^{-\frac{\mu^2}{2\sigma^2}}$$

The natural statistic is given by $x = (\omega, \omega^2)$ and the corresponding natural parameter is $\theta = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})$. In addition, θ is minimally sufficient.

It is more convenient to work with the minimal natural sufficient statistic x, so we modify the definition of exponential families to be in terms of x.

Definition 4.7. Let $F_\psi = \{p_{\psi,\theta} | \theta \in \Theta = \text{int}(\Theta) = \text{dom}(\psi) \subset \mathbb{R}^d\}$ be a multivariate parametric family of distributions. We call F_ψ a **regular exponential family** if each probability density is of the form

$$p_{\psi,\theta}(x) = e^{\langle x,\theta\rangle - \psi(\theta)} p_0(x)$$

for all $x \in \mathbb{R}^d$ where x is a minimal sufficient statistic for the family.

We call the function $\psi(\theta)$ the **log partition function** or the **cumulant function** corresponding to the exponential family. It is uniquely determined up to an additive constant term.

We now present a result without proof regarding the cumulant function of a regular exponential family. The result of convexity follows as a result of Hölder's Inequality and the fact that x is minimal sufficient implies the strict convexity [4].

Proposition 4.7. Let ψ be the cumulant function of a regular exponential family with natural parameter space $\theta = \text{dom}(\psi)$. Then ψ is a proper, closed, convex function with $\text{int}(\Theta) = \Theta$ and (Θ, ψ) is a convex function of Legendre type.

Definition 4.8. Let X be a $d-$dimensional real random vector distributed according to a regular exponential family density $p_{\psi,\theta}$ specified with natural parameter $\theta \in \Theta$. Define the **expectation** μ, or the expectation of X with respect to $p_{\psi,\theta}$, to be

$$\mu = \mu(\theta) = E_{p_{\psi,\theta}}(X) = \int_{\mathbb{R}^d} x p_{\psi,\theta}(x) dx$$

Notice that μ is a $d-$dimensional real vector as well.

Using the theory of Legendre duality, we look at the relationship between the natural parameter θ and the expectation μ. If we differentiate $\int p_{\psi,\theta}(x)dx = 1$ with respect to θ, we have

$$0 = \int (x - \nabla\psi(\theta)) p_{\psi,\theta}(x) dx \Rightarrow \mu(\theta) = \nabla\psi(\theta)$$

Let $\phi(\mu) = \psi^*(\mu) = \sup_{\theta\in\Theta}\{\langle\mu,\theta\rangle - \psi(\theta)\}$. Since (Θ, ψ) is a Legendre function, the pair (Θ, ψ), $(\text{int}(\text{dom}(\phi)), \phi)$ are Legendre duals of each other. Hence the mappings between $\text{int}(\text{dom}(\phi))$ and Θ are given by $\mu(\theta) = \nabla\psi(\theta)$ and $\theta(\mu) = \nabla\phi(\mu)$ and we can write the conjugate function of ψ as

$$\phi(\mu) = \langle\theta(\mu), \mu\rangle - \psi(\theta(\mu))$$

for all $\mu \in \text{int}(\text{dom}(\phi))$. This duality illustrates the simple relation between μ, θ.

4.3 The Bijection

Our main goal is to show **there is a bijection between regular exponential families and regular Bregman divergences**, a subset of the Bregman divergences. First, we show is there exists a unique Bregman divergence corresponding to every regular exponential family distribution. We can think of this as a one-to-one mapping.

Theorem 4.8. Let $p_{\psi,\theta}$ be the probability density function of a regular exponential family distribution. Let ϕ be the conjugate function of ψ such that $(\text{int}(\text{dom}(\phi)), \phi)$ is the Legendre dual of (Θ, ψ). Let $\theta \in \Theta$ be the natural parameter and $\mu \in \text{int}(\text{dom}(\phi))$ be the corresponding expectation. Let d_ϕ be the Bregman divergence derived from ϕ. Then $p_{\psi,\theta}$ can be uniquely expressed as

$$p_{\psi,\theta}(x) = e^{-d_\phi(x,\mu)} b_\phi(x)$$

for all $x \in \text{dom}(\phi)$ where $b_\phi : \text{dom}(\phi) \to \mathbb{R}_+$ is a uniquely determined function.

Proof. For all $x \in \text{dom}(\phi)$, we have

$$\begin{aligned}\langle x, \theta\rangle - \psi(\theta) &= \langle \mu, \theta\rangle - \psi(\theta) + \langle x - \mu, \theta\rangle \\ &= \phi(\mu) + \langle x - \mu, \nabla\phi(\mu)\rangle \\ &= -d_\phi(x, \mu) + \phi(x) \\ p_{\psi,\theta}(x) &= e^{\langle x,\theta\rangle - \psi(\theta)} p_0(x) \\ &= e^{-d_\phi(x,\mu)} b_\phi(x)\end{aligned}$$

where $b_\phi(x) = e^{\phi(x)} p_0(x)$.

Since $p_{\psi,\theta}$ uniquely determines ψ up to a constant, the expectation $\mu = \nabla\psi(\theta)$ corresponding to θ is uniquely determined, and hence the corresponding conjugate functions ϕ are unique up to an additive constant term. Therefore $d_\phi(x, \mu)$ is uniquely determined by Proposition 2.9(e). The Legendre duality of ϕ, ψ imply that no two different exponential families will correspond to the same Bregman divergence. Thus, we can conclude $b_\phi(x) = e^{d_\phi(x,\mu)} p_{\psi,\theta}(x)$ is uniquely determined as well. □

We acknowledge that the above theorem is only useful if it is true for all x that we can sample from $p_{\psi,\theta}(x)$. To make this notion concrete, we introduce the following definition.

Definition 4.9. We say x_0 can be sampled from $p_{\psi,\theta}(x)$ if for all I such that $x_0 \in I, \int_I dx > 0$, we have $\int_I dP_0(x) > 0$, where P_0 was defined earlier in this section. We define I_ψ to be the set of instances that can be sampled from $p_{\psi,\theta}$.

Indeed, it can be shown that $I_\psi \subset \text{dom}(\phi)$, where ϕ is the conjugate of ψ, and hence the theorem is useful. We omit the proof here but the interested reader can see Theorem 9.1 in [4]. We give an example of a case where I_ψ and $\text{int}(\text{dom}(\phi))$ are disjoint, so this extra step is necessary to acknowledge.

Example 4.2. Let X be a Bernoulli random variable where $P(X = 1) = q, P(X = 0) = 1 - q$ for some $q \in [0, 1]$. The instance space for X is $I_\psi = \{0, 1\}$. We find the cumulant function of X.

$$\begin{aligned}p(x; q) &= q^x (1 - q)^{1-x} \\ &= e^{x \log q + (1-x)\log(1-q)} \\ &= e^{x \log\left(\frac{q}{1-q}\right) - \log\left(\frac{1}{1-q}\right)}\end{aligned}$$

Let $\theta = \log \frac{q}{1-q}$, this is the natural parameter. Then $\frac{1}{1-q} = e^\theta + 1$ and

$$p(x; \theta) = e^{x\theta - \log(1+e^\theta)}$$

Hence the cumulant function is $\psi(\theta) = \log(1+e^\theta)$ and the expectation is $\mu = q$. Then the conjugate function of ψ is

$$\begin{aligned}\phi(\mu) &= \langle\theta(\mu), \mu\rangle - \psi(\theta(\mu)) \\ &= \mu \log\left(\frac{\mu}{1-\mu}\right) - \log\log\left(1 + e^{\log\left(\frac{\mu}{1-\mu}\right)}\right) \\ &= \mu\log\mu - \mu\log(1-\mu) - \log\left(1 + \frac{\mu}{1-\mu}\right) \\ &= \mu\log\mu - (1-\mu)\log(1-\mu)\end{aligned}$$

for $\mu \in (0, 1)$. Taking limits as $\mu \to 0$ and $\mu \to 1$, we have $\phi(\mu) = 0$ for $\mu = 0$ and $\mu = 1$ since ϕ is a closed function. Hence the domain of ϕ is $[0, 1]$ and the parameter space of μ is $\text{int}(\text{dom}(\phi)) = (0, 1)$. Hence we see I_ψ and $\text{int}(\text{dom}(\phi))$ are disjoint but $I_\psi \subset \text{dom}(\phi)$.

We now define the regular Bregman divergence, a class of Bregman divergences with another smoothness condition. More results on exponentially convex functions can be found in [1].

Definition 4.10. A function $F : \Theta \to (0, \infty), \Theta \subset \mathbb{R}^d$ is called **exponentially convex** if f is continuous and the kernel $K_f(\alpha, \beta) = f(\alpha + \beta), \alpha + \beta \in \Theta$ satisfies

$$\sum_{i=1}^{n} \sum_{j=1}^{n} K_f(\theta_i, \theta_j) u_i \overline{u}_j \geq 0$$

for all $\{\theta_1, ..., \theta_n\} \subset \Theta$ with $\theta_i + \theta_j \in \Theta$ and $u_i \in \mathbb{C}$. That is, the kernel is positive semi-definite.

Proposition 4.9. An exponentially convex function is convex. The logarithm of an exponentially convex function is convex.

Proof. For $n = 1$, we have $f(\theta) \geq 0$ for all $\theta \in \Theta$. For $n = 2$, we have

$$f(\theta_1)u_1^2 + 2f\left(\frac{\theta_1 + \theta_2}{2}\right) u_1\overline{u}_2 + f(\theta_2)\overline{u}_2^2 \geq 0$$

Setting $u_1 = -1, u_2 = 1$, we have

$$f\left(\frac{\theta_1 + \theta_2}{2}\right) \leq \frac{f(\theta_1) + f(\theta_2)}{2}$$

By Theorem 2.8, f is midpoint convex so it is convex.

In addition,

$$\begin{aligned} 0 &\leq f(\theta_1)u_1^2 + 2f\left(\frac{\theta_1 + \theta_2}{2}\right) u_1\overline{u}_2 + f(\theta_2)\overline{u}_2^2 \\ &= (\sqrt{f(\theta_1)}u_1 + \sqrt{f(\theta_2)}\overline{u}_2)^2 + 2f\left(\frac{\theta_1 + \theta_2}{2}\right) u_1\overline{u}_2 - 2\sqrt{f(\theta_1)f(\theta_2)}u_1\overline{u}_2 \end{aligned}$$

If $f(\theta_1) = f(\theta_2) = 0$, then by above, $f\left(\frac{\theta_1+\theta_2}{2}\right) = 0 = \sqrt{f(\theta_1)f(\theta_2)}$. Now, we can suppose $f(\theta_1) \neq 0$. Set $u_2 = 1$ and $u_1 = -\sqrt{\frac{f(\theta_2)}{f(\theta_1)}}$. Then

$$\begin{aligned} 0 &\leq 2f\left(\frac{\theta_1 + \theta_2}{2}\right)\left(-\sqrt{\frac{f(\theta_2)}{f(\theta_1)}}\right) - 2\sqrt{f(\theta_1)f(\theta_2)}\left(-\sqrt{\frac{f(\theta_2)}{f(\theta_1)}}\right) \\ f\left(\frac{\theta_1 + \theta_2}{2}\right) &\leq \sqrt{f(\theta_1)f(\theta_2)} \\ \log\left(f\left(\frac{\theta_1 + \theta_2}{2}\right)\right) &\leq \frac{\log(f(\theta_1)) + \log(f(\theta_2))}{2} \end{aligned}$$

Hence $\log(f)$ is also convex by Theorem 2.8. □

Definition 4.11. Let $f : \Theta \to (0, \infty)$ be an exponentially convex function such that Θ is open and $\psi(\theta) = \log(f(\theta))$ is strictly convex. Let ϕ be the conjugate function of ψ. Then we call the Bregman divergence d_ϕ a **regular Bregman divergence**.

We will use the following theorem from [7] along with a technical lemma to establish our main result.

Theorem 4.10. Let $\Theta \subset \mathbb{R}^d$ be an open convex set. A necessary and sufficient condition that there exists a unique, bounded, non-negative measure ν such that $f : \Theta \to (0, \infty)$ can be represented as $f(\theta) = \int_{\mathbb{R}^d} e^{\langle x, \theta \rangle} d\nu(x)$ is that f is exponentially convex.

Lemma 4.11. Let ψ be the cumulant of an exponential family with base measure P_0 and natural parameter space $\Theta \subset \mathbb{R}^d$. Then if P_0 is concentrated on an affine subspace of $\mathbb{R}^d$, then ψ is not strictly convex.

Proof. Suppose P_0 is concentrated on an affine subspace $S = \{x \in \mathbb{R}^d : \langle x, b\rangle = c\}$ for some $c \in \mathbb{R}, b \in \mathbb{R}^d$. Let $I = \{\theta : \theta = \alpha b, \alpha \in \mathbb{R}\}$. Then for any $\theta = \alpha b \in I$, we have $\langle x, \theta\rangle = \alpha c$ for all $x \in S$. Then the cumulant function is

$$\begin{aligned}\psi(\theta) &= \log\left(\int_{\mathbb{R}^d} e^{\langle x,\theta\rangle} dP_0(x)\right)\\ &= \log\left(\int_S e^{\alpha c} dP_0(x)\right)\\ &= \log\left(e^{\alpha c} P_0(S)\right)\\ &= \langle y, \theta\rangle + \log(P_0(S))\end{aligned}$$

for any $y \in S$. Since affine functions are convex but not strictly convex, ψ is not strictly convex. □

Now, we are set to prove the main theorem of this exposition.

Theorem 4.12. There is a bijection between regular exponential families and regular Bregman divergences.

Proof. In Theorem 4.8, we showed there is a unique Bregman divergence corresponding to every regular exponential family. Hence, for the one-to-one direction, it remains to show this Bregman divergence is regular. Then, we show that for every regular Bregman divergence, there exists a unique regular exponential family.

Let $\mathcal{F}_\psi$ be a regular exponential family with cumulant function ψ and natural parameter space Θ. Then there exists a non-negative bounded measure ν such that for all $\theta \in \Theta$,

$$\begin{aligned}1 &= \int_{\mathbb{R}^d} e^{\langle x,\theta\rangle - \psi(\theta)} d\nu(x)\\ e^{\psi(\theta)} &= \int_{\mathbb{R}^d} e^{\langle x,\theta\rangle} d\nu(x)\end{aligned}$$

Hence by Theorem 4.10, $e^{\psi(\theta)}$ is an exponentially convex function on Θ. Furthermore, by Proposition 4.7, ψ is strictly convex. Therefore, the Bregman divergence d_ϕ, where ϕ is the conjugate of ψ is a regular Bregman divergence.

For the other direction, let d_ϕ be a regular Bregman divergence and ψ the conjugate of ϕ. By Definition 4.9, ψ is strictly convex and $\Theta = \text{dom}(\psi)$ is an open set. In addition, $e^{\psi(\theta)}$ is exponentially convex. Therefore, Theorem 4.10 implies there exists a unique nonnegative bounded measure ν such that

$$e^{\psi(\theta)} = \int_{\mathbb{R}^d} e^{\langle x,\theta\rangle} d\nu(x)$$

Choose $b \in \Theta$ such that

$$e^{\psi(b)} = \int_{\mathbb{R}^d} e^{\langle x,b\rangle} d\nu(x)$$

Then $dP_0(x) = e^{\langle x,b\rangle - \psi(b)} d\nu(x)$ is a probability density function. We notice that

$$\begin{aligned}\int_{\mathbb{R}^d} e^{\langle x,\theta\rangle} dP_0(x) &= \frac{\int_{\mathbb{R}^d} e^{\langle x,\theta+b\rangle} d\nu(x)}{e^{\psi(b)}}\\ &= e^{\psi(\theta+b)-\psi(b)}\end{aligned}$$

Therefore the set of $\theta \in \mathbb{R}^d$ such that $\int_{\mathbb{R}^d} e^{\langle x,\theta\rangle} dP_0(x) < \infty$ is the set $\{\theta \in \mathbb{R}^d : \theta + b \in \Theta\}$ and for any such $\theta + b \in \Theta$, we have

$$\int_{\mathbb{R}^d} e^{\langle x,\theta+b\rangle - \psi(\theta+b)} d\nu(x) = 1$$

This implies the exponential family $\mathcal{F}_\psi$ consisting of the densities $p_{\psi,\theta}(x) = e^{\langle x,\theta\rangle - \psi(\theta)}$ with respect to the measure ν has natural parameter space Θ and cumulant function $\psi(\theta)$.

By Lemma 4.11, P_0 is not concentrated on an affine subspace of $\mathbb{R}^d$, which means x is full and a minimal statistic. In addition, Θ is open, so $\mathcal{F}_\psi$ is a regular exponential family. To show the family is unique, we note that for any Bregman divergence d_ϕ, the generating function can be $\phi_0(x) = \phi(x) + \langle x, a\rangle + c$ for $a \in \mathbb{R}^d, c \in \mathbb{R}$. The corresponding conjugate function ψ_0 of ϕ_0 is

$$\begin{aligned}
\psi_0(\theta) &= \sup_x \Big\{\langle\theta, x\rangle - \phi_0(x)\Big\} \\
&= \sup_x \Big\{\langle\theta, x\rangle - \phi(x) - \langle a, x\rangle - c\Big\} \\
&= \sup_x \Big\{\langle\theta - a, x\rangle - \phi(x)\Big\} - c \\
&= \psi(\theta - a) - c
\end{aligned}$$

Hence the corresponding cumulant functions differs only by a constant. Since cumulant functions are unique up to additive constant, the exponential family $\mathcal{F}_\psi$ is unique. □

We give three examples of this bijection. In all these examples, the expectation can be calculated in the regular manner of integration, but we calculate it in using duality results.

Example 4.3. We look at the $d-$dimensional Gaussian distribution $N(a, \sigma^2 I_d)$.

We express the probability density function in the canonical form of an exponential family.

$$\begin{aligned}
p(x; a) &= (2\pi\sigma^2)^{-\frac{d}{2}} \exp\left(-\frac{1}{2\sigma^2}\|x - a\|^2\right) \\
&= (2\pi\sigma^2)^{-\frac{d}{2}} \exp\left(\langle x, \frac{a}{\sigma^2}\rangle\right) \exp\left(-\frac{1}{2\sigma^2}\|a\|^2\right) \exp\left(-\frac{1}{2\sigma^2}\|x\|^2\right) \\
&= e^{\theta - \psi(\theta)} p_0(x) \\
\theta &= \frac{a}{\sigma^2} \\
\psi(\theta) &= \frac{\sigma^2}{2}\|\theta\|^2 \\
p_0(x) &= \exp\left(-\frac{1}{2\sigma^2}\|x\|^2\right)(2\pi\sigma^2)^{-\frac{d}{2}}
\end{aligned}$$

The expectation is $\mu = \nabla\psi(\theta) = \sigma^2\theta = a$. The Legendre dual ϕ of ψ is

$$\begin{aligned}
\phi(\mu) &= \langle\mu, \theta\rangle - \psi(\theta) \\
&= \langle\mu, \mu/\sigma^2\rangle - \frac{\sigma^2}{2}\|\theta\|^2 \\
&= \frac{\|\mu\|^2}{\sigma^2} - \frac{\sigma^2}{2}\frac{\|\mu\|^2}{\sigma^4} \\
&= \frac{\|\mu\|^2}{2\sigma^2}
\end{aligned}$$

The Bregman divergence d_ϕ is

$$\begin{aligned} d_\phi(x,\mu) &= \phi(x) - \phi(\mu) - \langle x-\mu, \nabla\phi(\mu)\rangle \\ &= \frac{\|x\|^2}{2\sigma^2} - \frac{\|\mu\|^2}{2\sigma^2} - \langle x-\mu, \mu/\sigma^2\rangle \\ &= \frac{\|x-\mu\|^2}{2\sigma^2} \end{aligned}$$

The function $b_\phi(x)$ is $b_\phi(x) = e^{\phi(x)}p_0(x) = (2\pi\sigma^2)^{-d/2}$. Thus, we see that $p_{\psi,\theta}(x) = e^{-d_\phi(x,\mu)}b_\phi(x)$ like in Theorem 4.8.

Example 4.4. Another example of an exponential family is aptly the Exponential distribution, with probability distribution function $p(x;\lambda) = \lambda e^{-\lambda x}, \lambda > 0, x \geq 0$. We have

$$\begin{aligned} p(x;\lambda) &= e^{x(-\lambda)-(-\log\lambda)} \\ &= e^{x\theta-\psi(\theta)}p_0(x) \\ \theta &= -\lambda \\ \psi(\theta) &= -\log\lambda = -\log(-\theta) \\ p_0(x) &= 1 \end{aligned}$$

The expectation is $\mu = \psi'(\theta) = -\frac{1}{\theta} = \frac{1}{\lambda}$. The Legendre dual ϕ of ψ is

$$\phi(\mu) = \mu\theta - \psi(\theta) = (-1/\theta)\theta + \log(-\theta) = -1 - \log\mu$$

The Bregman divergence d_ϕ is

$$\begin{aligned} d_\phi(x,\mu) &= \phi(x) - \phi(\mu) - (x-\mu)\phi'(\mu) \\ &= (-1-\log x) - (-1-\log\mu) - (x-\mu)(-1/\mu) \\ &= \frac{x}{\mu} - \log\left(\frac{x}{\mu}\right) - 1 \end{aligned}$$

The function $b_\phi(x)$ is $b_\phi(x) = e^{\phi(x)}p_0(x) = e^{-1-\log x} = \frac{1}{ex}$. Thus, we see that

$$\begin{aligned} e^{-d_\phi(x,\mu)}b_\phi(x) &= e^{-\frac{x}{\mu}} \cdot \frac{x}{\mu} \cdot e \cdot \frac{1}{ex} \\ &= e^{-\frac{x}{\mu}}\frac{1}{\mu} \\ &= e^{-\lambda x}\lambda \\ &= p(x;\lambda) \end{aligned}$$

Example 4.5. The last example we consider is a discrete distribution: the multinomial. The multinomial distribution is a generalization of the binomial distribution. We can think of the multinomial distribution as modeling N rolls of a $d-$sided unfair die. The density function is

$$p(\mathbf{x};\mathbf{q}) = \frac{N!}{\prod_{j=1}^{d} x_j!}\prod_{j=1}^{d} q_j^{x_j}$$

where $\mathbf{x} = (x_1, ..., x_{d_1}) \in \mathbb{Z}_+^{d-1}, \sum_{j=1}^{d} x_j = N$ represent the frequencies of events, and $\mathbf{q} = (q_1, ..., q_{d-1}), q_j \geq 0, \sum_{j=1}^{d} q_j = 1$ represent the probabilities of events. Note the subscript runs up to $d-1$ only, since x_d, q_d can be uniquely determined by $(x_1, ..., x_{d-1}), (q_1, ..., q_{d-1})$. We show that we can express the multinomial as the density of an exponential distribution in $\mathbf{x} = (x_1, ..., x_{d-1}), \boldsymbol{\theta} = (\log(\frac{q_1}{q_d}), ..., \log(\frac{q_{d-1}}{q_d}))$. Let $p_0(\mathbf{x}) = \frac{N!}{\prod_{j=1}^{d} x_j!}$.

$$\begin{aligned}
p(\mathbf{x}, \mathbf{q}) &= \exp\left(\sum_{j=1}^{d} x_j \log q_j\right) p_0(x) \\
&= \exp\left(\sum_{j=1}^{d-1} x_j \log q_j + (N - \sum_{j=1}^{d-1} x_j) \log q_d\right) p_0(x) \\
&= \exp\left(\sum_{j=1}^{d-1} x_j \log\left(\frac{q_j}{q_d}\right) + N \log q_d\right) p_0(x) \\
&= \exp\left(\langle \mathbf{x}, \boldsymbol{\theta}\rangle - N \log\left(\frac{1}{q_d}\right)\right) p_0(x) \\
&= \exp\left(\langle \mathbf{x}, \boldsymbol{\theta}\rangle - N \log\left(\sum_{j=1}^{d} \frac{q_j}{q_d}\right)\right) p_0(x) \\
&= \exp\left(\langle \mathbf{x}, \boldsymbol{\theta}\rangle - N \log\left(1 + \sum_{j=1}^{d-1} e^{\theta_j}\right)\right) p_0(x)
\end{aligned}$$

The cumulant function is $\psi(\boldsymbol{\theta}) = -N \log q_d = N \log\left(1 + \sum_{j=1}^{d-1} e^{\theta_j}\right)$. The expectation μ is

$$\boldsymbol{\mu} = \nabla\psi(\boldsymbol{\theta}) = \left(\frac{Ne^{\theta_1}}{1 + \sum_{j=1}^{d-1} e^{\theta_j}}, ..., \frac{Ne^{\theta_{d-1}}}{1 + \sum_{j=1}^{d-1} e^{\theta_j}}\right) = N\mathbf{q}$$

This is very expected, since the jth expectation is the probability of the jth event multiplied by the number of trials: $\mu_j = Nq_j$. The Legendre dual ϕ of ψ is

$$\begin{aligned}
\phi(\boldsymbol{\mu}) &= \langle \boldsymbol{\mu}, \boldsymbol{\theta}\rangle - \psi(\boldsymbol{\theta}) \\
&= \sum_{j=1}^{d-1} Nq_j \log\left(\frac{q_j}{q_d}\right) + N \log q_d \\
&= \sum_{j=1}^{d} Nq_j \log(q_j) \\
&= N \sum_{j=1}^{d} \left(\frac{\mu_j}{N}\right) \log\left(\frac{\mu_j}{N}\right)
\end{aligned}$$

Notice this is a multiple of the negative entropy for a discrete probability distribution given by $\{\mu_j/N\}_{j=1}^{d}$

(see Example 2.3). Then, the corresponding Bregman divergence d_ϕ is

$$\begin{aligned}
d_\phi(\mathbf{x}, \boldsymbol{\mu}) &= \phi(\mathbf{x}) - \phi(\boldsymbol{\mu}) - \langle \mathbf{x} - \boldsymbol{\mu}, \nabla\phi(\boldsymbol{\mu})\rangle \\
&= N\sum_{j=1}^{d} \left(\frac{x_j}{N}\right)\log\left(\frac{x_j}{N}\right) - N\sum_{j=1}^{d} \left(\frac{\mu_j}{N}\right)\log\left(\frac{\mu_j}{N}\right) - \sum_{j=1}^{d}(x_j - \mu_j)\left(1 + \log\left(\frac{\mu_j}{N}\right)\right) \\
&= N\sum_{j=1}^{d} \left(\frac{x_j}{N}\right)\log\left(\frac{x_j}{N}\right) - \left(\frac{x_j}{N}\right)\log\left(\frac{\mu_j}{N}\right) \\
&= N\sum_{j=1}^{d} \left(\frac{x_j}{N}\right)\log\left(\frac{x_j/N}{\mu_j/N}\right)
\end{aligned}$$

As expected, this is a multiple of the KL-divergence. The function $b_\phi(x)$ is

$$\begin{aligned}
b_\phi(\mathbf{x}) &= e^{\phi(\mathbf{x})}p_0(x) \\
&= \exp\left(\sum_{j=1}^{d} x_j \log\left(\frac{x_j}{N}\right)\right)\frac{N!}{\prod_{j=1}^{d} x_j!} \\
&= \frac{\prod_{j=1}^{d} x_j^{x_j}}{N^N}\frac{N!}{\prod_{j=1}^{d} x_j!}
\end{aligned}$$

Indeed, we have

$$\begin{aligned}
e^{-d_\phi(\mathbf{x}, \boldsymbol{\mu})}b_\phi(\mathbf{x}) &= \prod_{j=1}^{d} \frac{\mu_j^{x^j}}{x_j^{x_j}}\frac{\prod_{j=1}^{d} x_j^{x_j}}{N^N}\frac{N!}{\prod_{j=1}^{d} x_j!} \\
&= \prod_{j=1}^{d} q_j^{x_j}\frac{N!}{\prod_{j=1}^{d} x_j!} \\
&= p_{\psi,\boldsymbol{\theta}}(\mathbf{x})
\end{aligned}$$

5 Bregman Soft Clustering

As mentioned in the introduction, soft clustering does not assign each data point to a cluster, but rather **each data point is assigned a probability belonging to each of the clusters**, and potentially belong to more than one cluster. The probabilities also indicate some sort of degree to which data points belong to the clusters. For example, points at the center of cluster will have higher probability than those on the boundary. The collection of probabilities is called a **soft partition**. In this section, we will formulate the Bregman soft clustering problem as a parameter estimation problem for mixture models which can be solved by the Expectation-Maximization (EM) algorithm. In addition, we will see that the hard clustering algorithm is a special case of the soft clustering algorithm.

5.1 The Expectation-Maximization Algorithm

First, we give a short introduction to the EM algorithm, an iterative method used to find maximum likelihood estimators (MLEs) when the model depends on latent variables, or missing variables. These latent variables may be unobservable or observable but missing from the dataset. We have the following data structure: X is the complete dataset, Y is observed, and Z is missing. We write

$$X = (Y, Z) \sim f_\theta(y, z)$$

to indicate that X is distributed with pdf $f_\theta(y, z)$, where θ is the unknown parameter of interest. We seek the MLE $\hat{\theta}$ of θ based on only the observed data Y:

$$\hat{\theta} = \hat{\theta}(y) = \arg\max_\theta f_\theta(y) = \arg\max_\theta \int f_\theta(y, z)dz$$

The integral may be intractable and hard to compute. The EM algorithm iteratively applies two steps.

- **First E-Step:** Let $\hat{\theta}_0 = \hat{\theta}_0(y)$ be an initial estimate or guess. Compute the expectation of the log-likelihood function of θ with respect to the current estimate:

$$E_{\hat{\theta}_0}\left[\log\left(\frac{f_\theta(Y, Z)}{f_{\hat{\theta}_0}(Y, Z)}\middle| Y = y\right)\right] \equiv J(\theta|\hat{\theta}_0(y), y)$$

- **First M-Step:** Find

$$\hat{\theta}_1 = \hat{\theta}_1(y) = \arg\max_\theta J(\theta|\hat{\theta}_0(y), y)$$

  Notice that $J(\hat{\theta}_1|\hat{\theta}_0, y) \geq J(\hat{\theta}_0|\hat{\theta}_0, y) = 0$.

- **Further Steps:** For $k = 1, 2, ...$, repeat the E-step and M-step using $\hat{\theta}_k$ to compute $\hat{\theta}_{k+1}$.

Theorem 5.1. The likelihood increases at each iteration of the EM algorithm: $f_{\hat{\theta}_{k+1}}(y) \geq f_{\hat{\theta}_k}(y)$.

Proof. We investigate $\log\left(\frac{f_{\hat{\theta}_{k+1}}(y)}{f_{\hat{\theta}_k}(y)}\right)$.

$$\begin{aligned}
\log\left(\frac{f_{\hat{\theta}_{k+1}}(y)}{f_{\hat{\theta}_k}(y)}\right) &= E_{\hat{\theta}_k}\left[\log\left(\frac{f_{\hat{\theta}_{k+1}}(Y)}{f_{\hat{\theta}_k}(Y)}\middle| Y = y\right)\right] \\
&= E_{\hat{\theta}_k}\left[\log\left(\frac{f_{\hat{\theta}_{k+1}}(Y, Z)}{f_{\hat{\theta}_k}(Y, Z)}\middle| Y = y\right)\right] - E_{\hat{\theta}_k}\left[\log\left(\frac{f_{\hat{\theta}_{k+1}}(Z|Y)}{f_{\hat{\theta}_k}(Z|Y)}\middle| Y = y\right)\right] \\
&= J(\hat{\theta}_{k+1}|\hat{\theta}_k, y) + E_{\hat{\theta}_k}\left[\log\left(\frac{f_{\hat{\theta}_k}(Z|Y)}{f_{\hat{\theta}_{k+1}}(Z|Y)}\middle| Y = y\right)\right] \\
&\geq 0
\end{aligned}$$

The first term is nonnegative by definition. The second term is a conditional KL divergence $KL_{Z|Y}(f_{\hat{\theta}_k}, f_{\hat{\theta}_{k+1}})$ and so it is nonnegative (recall Example 2.3). Hence $f_{\hat{\theta}_{k+1}}(y) \geq f_{\hat{\theta}_k}(y)$. □

This shows the EM algorithm, if convergent, will converge to a local maximizer of the likelihood. However it is not guaranteed to converge, and when it converges, the limit $\hat{\theta}_k \to \hat{\theta}$ may not converge to the true MLE. The theory behind the EM algorithm is well-studied. For more on the EM algorithm and its properties, see [6], [8].

In certain cases such as missing data in an exponential family, there is a simple representation of the estimates $\hat{\theta}_k$ [10]. Suppose the complete data $X = (Y, Z)$ has the canonical exponential family form

$$f_\theta(y, z) = p_0(y, z)a(\theta)e^{\langle T(y,z), \theta\rangle}$$

where $T(y, z)$ is a sufficient statistic for $\theta \in \mathbb{R}^d$. Then

$$\begin{aligned} \log\left(\frac{f_\theta(y, z)}{f_{\hat{\theta}_0}(y, z)}\right) &= \log\left(\frac{a(\theta)}{a(\theta_0)}\right) + (\theta - \hat{\theta}_0)^T T(y, z) \\ J(\theta|\hat{\theta}_0, y) &= \log\left(\frac{a(\theta)}{a(\theta_0)}\right) + (\theta - \hat{\theta}_0)^T E_{\hat{\theta}_0}\Big(T(Y, Z)|Y = y\Big) \\ &= \log\left(\frac{f_\theta(\hat{T}_1(y))}{f_{\hat{\theta}_0}(\hat{T}_1(y))}\right) \end{aligned}$$

where $\hat{T}_1(y) = E_{\hat{\theta}_0}\Big(T(Y, Z)|Y = y\Big)$. This is just the complete log-likelihood ratio of Y based on the value $\hat{T}_1$. Hence, the $(k + 1)$st E-step computes

$$\hat{T}_{k+1} = \hat{T}_{k+1}(y) = E_{\hat{\theta}_k}\Big(T(Y, Z)|Y = y\Big)$$

The $(k + 1)$st M-step maximizes the complete log-likelihood ratio

$$\hat{\theta}_{k+1} = \arg\max_\theta \log\left(\frac{f_\theta(\hat{T}_{k+1}(y))}{f_{\hat{\theta}_0}(\hat{T}_{k+1}(y))}\right)$$

This is a very simple computation and illustrates the elegance of exponential families.

Our use case for the EM algorithm is in finite mixture models. It is a hierarchical model where random variables $X_1, ..., X_n$ are sampled from a **mixture probability density function**, defined by

$$f(x) = \sum_{k=1}^{K} \pi_k g_{\theta_k}(x)$$

where $\{g_{\theta_k}(x)\}_{k=1}^K$ are a set of pdfs belonging to the same parametric family with parameter θ_k and $\{\pi_k\}_{k=1}^K$ are mixture weights with $\pi_k \geq 0, \sum_{k=1}^K \pi_k = 1$. Let $Z_{i,k}, i \in \{1, ..., n\}, k \in \{1, ..., K\}$ be the indicator variable such that $Z_{i,k} = 1$ if X_i was drawn from component k; this is our latent variable. The mixture weight $\pi_k = P(Z_{i,k} = 1)$ is the probability that X_i belongs to mixture component k, that is, X_i was sampled from pdf $g_{\theta_k}(x)$. We say that X_i is sampled from a **mixture model**. The complete set of parameters for maximum likelihood estimation is $\Gamma = \{\pi_1, ..., \pi_K, \theta_1, ..., \theta_K\}$.

We use the EM algorithm to estimate these parameters. In the E-step, the expected log-likelihood gives us estimates for the probability that X_i was sampled from mixture k, which we calculate using Bayes' Rule. This gives a soft partitioning for the soft clustering that we desire.

Algorithm 2: EM for MLE of Mixture Models

```
Data: A set $\mathcal{X} = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$; a natural number $K$ (the number of clusters)
Result: A local minimizer $\Gamma^*$ of $L_{\mathcal{X}}(\Gamma) = \prod_{i=1}^n \sum_{k=1}^K \pi_k p_{\psi,\theta_k}(x_i)$ where $\Gamma = \{\theta_k, \pi_k\}_{k=1}^K$; a soft
        partitioning $\{\{p(k|x_i)\}_{k=1}^K\}_{i=1}^n$, where $p(k|x_i)$ is the probability of $x_i$ belonging to
        mixture component $k$.

Initialize $\Gamma = \{\theta_k, \pi_k\}_{k=1}^K$ with $\theta_k \in \Theta, \pi_k \geq 0, \sum_{k=1}^K \pi_k = 1$.
opt ← False
while not opt do
    "The E-Step"
    for i=1 to n do
        for k=1 to K do
            $p(k|x_i) \leftarrow \frac{\pi_k p_{\psi,\theta_k}(x_i)}{\sum_{j=1}^K \pi_j p_{\psi,\theta_j}(x_i)}$
        end
    end
    "The M-Step"
    for k=1 to K do
        $\pi_k \leftarrow \frac{1}{n}\sum_{i=1}^n p(k|x_i)$
        $\theta_k \leftarrow \arg\max_\theta \sum_{i=1}^n \log(p_{\psi,\theta}(x_i))p(k|x_i)$
        $\Gamma^* \leftarrow \{\theta_k, \pi_k\}_{k=1}^K$
    end
    if $\Gamma^* == \Gamma$ then
        opt ← True
    else
        $\Gamma \leftarrow \Gamma^*$
    end
end
return $\Gamma^*, \{\{p(k|x_i)\}_{k=1}^K\}_{i=1}^n$
```

5.2 The Soft Clustering Algorithm

Let $X_1, ..., X_n$ be independently and identically distributed (iid) sample drawn from a single exponential family $X \sim p_{\psi,\theta}(x)$, we know the problem of maximum likelihood estimation of θ is equivalent to minimizing negative log-likelihood. By Theorem 4.8, minimizing the negative log-likelihood is equivalent to maximizing the corresponding expected Bregman divergence, because

$$-\log(p_{\psi,\theta}(x)) = -d_\phi(x,\mu) + \log(b_\phi(x))$$

By Theorem 3.1, the optimal distribution has $\mu = E(X)$ as the expectation, that is, the MLE $\hat{\theta}$ of θ satisfies $E(X) = \nabla\psi(\hat{\theta})$. In addition, the minimum negative log-likelihood of X under an exponential family with cumulant function ψ is $I_\phi(X)$, the Bregman information of X (up to an additive constant), where ϕ is the the Legendre conjugate of ψ.

If $X_1, ..., X_n$ are an iid sample from a mixture model of K densities from the same exponential family, then we know from Algorithm 2, we can obtain the soft clustering. Therefore, for regular Bregman divergences, we can define the Bregman soft clustering problem as learning the maximum likelihood parameters $\Gamma = \{\pi_k, \theta_k\}_{k=1}^K \equiv \{\pi_k, \mu_k\}_{k=1}^K$ of a mixture model

$$f(x|\Gamma) = \sum_{k=1}^K \pi_k e^{-d_\phi(x,\mu_k)} b_\phi(x)$$

where we used Theorem 4.8 to express the regular exponential family in the desired form. With this viewpoint, we obtain a simplified version of Algorithm 2 which is the Bregman soft clustering algorithm. In particular, the M-step becomes very straightforward to solve when the corresponding Bregman divergence d_ϕ is known. Hence, in some situations, it may be easier to use regular Bregman divergences for mixture models instead of coming up with an appropriate exponential family.

Algorithm 3: Bregman Soft Clustering

```
Data: A set $\mathcal{X} = \{x_i\}_{i=1}^n \subset S \subset \mathbb{R}^d$; a Bregman divergence $d_\phi : S \times \text{ri}(S) \to \mathbb{R}$; a natural
      number $K$ (the number of clusters)
Result: A local minimizer $\Gamma^*$ of $L_\phi(\Gamma) = \prod_{i=1}^n \sum_{k=1}^K \pi_k e^{-d_\phi(x_i,\mu_k)} b_\phi(x_i)$ where
        $\Gamma = \{\mu_k, \pi_k\}_{k=1}^K$; a soft partitioning $\{\{p(k|x_i)\}_{k=1}^K\}_{i=1}^n$, where $p(k|x_i)$ is the probability
        of $x_i$ belonging to cluster $\#k$.

Initialize $\Gamma = \{\theta_k, \pi_k\}_{k=1}^K$ with $\mu_k \in \text{ri}(S), \pi_k \geq 0, \sum_{k=1}^K \pi_k = 1$.
opt ← False
while not opt do
    "The E-Step"
    for i=1 to n do
        for k=1 to K do
            $p(k|x_i) \leftarrow \frac{\pi_k \exp(-d_\phi(x_i,\mu_k))}{\sum_{j=1}^K \pi_j \exp(-d_\phi(x_i,\mu_j))}$
        end
    end
    "The M- Step"
    for k=1 to K do
        $\pi_k \leftarrow \frac{1}{n} \sum_{i=1}^n p(k|x_i)$
        $\mu_k \leftarrow \frac{\sum_{i=1}^n p(k|x_i) x_i}{\sum_{i=1}^n p(k|x_i)}$
        $\Gamma^* \leftarrow \{\mu_k, \pi_k\}_{k=1}^K$
    end
    if $\Gamma^* == \Gamma$ then
        opt ← True
    else
        $\Gamma \leftarrow \Gamma^*$
    end
end
return $\Gamma^*, \{\{p(k|x_i)\}_{k=1}^K\}_{i=1}^n$
```

Notice the E-step in Algorithm 3 is the same as in Algorithm 2, since the $b_\phi(x_i)$ cancels in the numerator and denominator. We now prove that the M-steps in Algorithm 2 and 3 are equivalent for regular Bregman divergences and exponential families, that is, the M-step in Algorithm 3 is correct.

Theorem 5.2. For a mixture model with density

$$p(x|\Gamma) = \sum_{k=1}^K \pi_k e^{-d_\phi(x,\mu_k)} b_\phi(x)$$

the maximization step for the density parameters in Algorithm 2 reduces to

$$\mu_k = \frac{\sum_{i=1}^n p(k|x_i) x_i}{\sum_{i=1}^n p(k|x_i)}$$

for $1 \leq k \leq K$. This is the M-step in Algorithm 3.

Proof. The maximization step in Algorithm 2 for $\theta_k, 1 \leq k \leq K$ is given by

$$\theta_k \leftarrow \arg\max_{\theta} \sum_{i=1}^{n} \log(p_{\psi,\theta}(x_i))p(k|x_i)$$

The component densities are

$$p_{\psi,\theta_k}(x) = e^{-d_\phi(x,\mu_k)} b_\phi(x)$$

for $1 \leq k \leq K$. Substituting these into the maximization step, we get update equations for the expectations μ_k. For $1 \leq k \leq K$,

$$\begin{aligned}
\mu_k &= \arg\max_{\mu} \sum_{i=1}^{n} \Big(\log(b_\phi(x_i)) - d_\phi(x_i, \mu) \Big) p(k|x_i) \\
&= \arg\min_{\mu} d_\phi(x_i, \mu) p(k|x_i) \\
&= \arg\min_{\mu} d_\phi(x_i, \mu) p(k|x_i) \left(\frac{1}{\sum_{j=1}^{n} p(k|x_j)} \right) \\
&= \arg\min_{\mu} d_\phi(x_i, \mu) \nu(x_i)
\end{aligned}$$

where $\nu(x_i) = \frac{p(k|x_i)}{\sum_{j=1}^{n} p(k|x_j)}$. By Theorem 3.1, we must have

$$\mu_k = \frac{\sum_{i=1}^{n} x_i p(k|x_i)}{\sum_{i=1}^{n} p(k|x_i)}$$

This is the desired update equation for the expectations $\{\mu_k\}_{k=1}^{K}$. □

We can also consider the Bregman soft clustering problem where the samples are weighted, that is, associate each x_i with a weight ν_i such that $\sum_{i=1}^{n} \nu_i = 1$. We then maximize the weighted likelihood function

$$\log L_\phi(\Gamma) = \sum_{i=1}^{n} \nu_i \log \left(\sum_{k=1}^{K} \pi_k e^{-d_\phi(x_i,\mu_k)} b_\phi(x_i) \right)$$

Then the E-step will remain the same and the M-step will have update equations

$$\begin{aligned}
\pi_k &= \sum_{i=1}^{n} \mu_i p(k|x_i) \\
\mu_k &= \frac{\sum_{i=1}^{n} \nu_i x_i p(k|x_i)}{\sum_{i=1}^{n} \nu_i p(k|x_i)}
\end{aligned}$$

Notice that the original case is identical to $\nu_i = 1/n$ for all i, since the samples are weighted equally.

Another interesting viewpoint is to consider the Bregman hard clustering as a limit of Bregman soft clustering. For a convex function ϕ and constant $c > 0$, $c\phi$ is also convex with Bregman divergence $d_{c\phi} = cd_\phi$. Hence if we take $c \to \infty$, the probabilities $p(k|x_i) \to \{0, 1\}$ in the E-step. This means the EM algorithm of the soft clustering problem reduces to the hard clustering algorithm.

5.3 Clustering in the Second Argument

So far, we have considered the case where the Bregman divergence took the data point as the first argument and the cluster representative as the second argument. In this case, we get unique cluster representatives. We can consider an alternative clustering problem where the data point is the second argument and the cluster representative in the Bregman divergence. Formally, we consider an alternate version of the hard clustering problem. Given a set $\mathcal{X} = \{x_i\}$ and a positive probability measure ν, we find a partition $\{\mathcal{X}_h\}_{h=1}^k$ and corresponding cluster representatives $\{\mu_h\}_{h=1}^k$ that solve

$$\min_{\mu_h, h=1,\ldots,k} \sum_{h=1}^{k} \sum_{x_i \in \mathcal{X}_h} \nu_i d_\phi(\mu_h, x_i)$$

Because Bregman divergences are not necessarily convex in the second argument, the cluster representatives are not necessarily the expectation. However, we can show that this problem is equivalent to the original Bregman hard clustering problem using a different Bregman divergence and representation. We first state and prove a proposition about the duality of the Bregman divergence.

Proposition 5.3. Let $\phi : S \to \mathbb{R}$ be a strictly convex function and d_ϕ its corresponding Bregman divergence. Let ψ be its conjugate. For all $\mu_1, \mu_2 \in \text{ri}(S)$, we have a duality between the Bregman divergences:

$$d_\phi(\mu_1, \mu_2) = d_\psi(\theta_2, \theta_1)$$

where $\theta_i = \nabla\phi(\mu_i), \mu_i = \nabla\psi(\theta_i), i = 1, 2$. We call d_ψ the dual Bregman divergence to d_ϕ.

Proof. We know by Section 4.2 that $\phi(\mu) = \langle\theta(\mu), \mu\rangle - \psi(\theta(\mu))$. From the definition of Bregman divergence, we have

$$\begin{aligned}
d_\phi(\mu_1, \mu_2) &= \phi(\mu_1) - \phi(\mu_2) - \langle \mu_1 - \mu_2, \nabla\phi(\mu_2)\rangle \\
&= \langle\theta_1, \mu_1\rangle - \psi(\theta_1) - \langle\theta_2, \mu_2\rangle + \psi(\theta_2) - \langle\mu_1 - \mu_2, \theta_2\rangle \\
&= \psi(\theta_2) - \psi(\theta_1) - \langle\theta_1, \mu_1\rangle + \langle\mu_1, \theta_2\rangle \\
&= \psi(\theta_2) - \psi(\theta_1) - \langle\theta_2 - \theta_2, \nabla\psi(\theta_1)\rangle \\
&= d_\psi(\theta_2, \theta_1)
\end{aligned}$$

The change in argument in the duality is what will give our result. □

Now, we are ready to state the alternate Bregman hard clustering problem in terms of the original Bregman hard clustering problem. Let ϕ be a strictly convex function and d_ϕ its corresponding Bregman divergence such that $(\text{int}(\text{dom}(\phi)), \phi)$ is a convex function of Legendre type. Let $(\text{int}(\text{dom}(\psi)), \psi)$ be the corresponding Legendre dual. Let $\mathcal{X} = \{x_i\}_{i=1}^n$ be our original set. Let $\mathcal{X}^\theta = \{\theta_{x_i}\}_{i=1}^n = \{\nabla\phi(x_i)\}_{i=1}^n$ be the dual space. Let $\theta_h = \nabla\phi(\mu_h)$ denote the cluster representatives in the dual space. For a probability measure ν over $\mathcal{X}$, the alternative Bregman hard clustering problem can be expressed as

$$\min_{\mu_h, h=1,\ldots,k} \sum_{h=1}^{k} \sum_{x_i \in \mathcal{X}_h} \nu_i d_\phi(\mu_h, x_i) = \min_{\theta_h, h=1,\ldots,k} \sum_{h=1}^{k} \sum_{\theta_{x_i} \in \mathcal{X}_h^\theta} \nu_i d_\psi(\theta_{x_i}, \theta_h)$$

where $\mathcal{X}_h^\theta$ are the clusters in the dual space. This implies the alternative Bregman hard clustering problem is equivalent to the original Bregman hard clustering problem with the dual divergence d_ψ. We can easily get the original cluster representatives $\mu_h = \nabla\psi(\theta_h)$ from θ_h given by the expectation. Of course, this method will work efficiently only if d_ψ can be found easily, just like in the EM scheme. This alternative formulation also extends to the Bregman soft clustering case by replacing $d_\phi(x_i, \mu_h)$ by $d_\psi(\theta_h, \theta_{x_i})$.

6 Conclusion

In this paper, we presented clustering algorithms based on minimizing functions of Bregman divergence. In the hard clustering case, we showed it was a direct generalization of the K-Means algorithm, and that Bregman divergences are the only such possible function for which cluster centers are the mean. In the soft clustering case, we developed the theory behind regular exponential families and their bijection with regular Bregman divergences. This bijection allowed us simplify the EM algorithm for mixture density estimation for the soft clustering problem. Bregman divergences are truly an important theoretical tool that allow us to solve many applied problems, not just in unsupervised learning, but also in problems tied to rate distortion theory and information theory.

References

[1] N. I. Akhiezer, *The Classical Moment Problem and some related questions in analysis*, Oliver & Boyd, 1965.

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

[3] A. Banerjee, X. Guo, and H.Wang, *On the optimality of conditional expectation as a Bregman predictor*, IEEE Transactions on Information Theory, 51 (July 2005), pp. 2664–2669.

[4] O. Barndorff-Nielsen, *Information and Exponential Families in Statistical Theory*, John Wiley & Sons, 1978.

[5] 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 (1967), pp. 200–217.

[6] A. P. Dempster, N. M. Laird, and D. Rubin, *Maximum Likelihood from Incomplete Data via the EM Algorithm*, Journal of the Royal Statistical Society. Series B (Methodological), 39 (1977), pp. 1–38.

[7] A. Devinatz, *The representation of functions as Laplace-Stieltjes integrals*, Duke Mathematical Journal, 24 (1955), pp. 281–298.

[8] G. J. McLachlan and T. Krishnan, *The EM Algorithms and Extensions*, Wiley-Interscience, 2 ed., 2008.

[9] C. Niculescu and L. E. Persson, *Convex Functions and Their Applications: A Comtemporary Approach*, Springer Science+Business Media Inc., 2006.

[10] M. Perlman, *Class notes for stat 513.* Department of Statistics, University of Washington, 2020. Course Notes.

[11] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.

[12] P. Smyth, *Mixture Models and the EM Algorithm.* Department of Computer Science, University of California, Irvine, 2017. Course Notes.