Clustering with Bregman Divergences

Arindam Banerjee* Srujana Merugu† Inderjit Dhillon‡ Joydeep Ghosh§

UT CS Technical Report #TR-03-19

Abstract

A wide variety of distortion functions are used for clustering, e.g., squared Euclidean distance, Mahalanobis distance and relative entropy. In this paper we consider the general case where the distortion is a Bregman divergence. We pose the hard clustering problem in terms of minimizing the loss in Bregman information, a quantity motivated by rate-distortion theory, and present an algorithm to minimize this loss. The proposed algorithm unifies several well-known partitional methods, such as classical `kmeans` and information-theoretic clustering, which arise by special choices of the Bregman divergence. Further, we show an explicit bijection between Bregman divergences and exponential families. The bijection enables the development of an efficient viewpoint of EM for learning models involving mixtures of exponential distributions. This leads to a simple soft clustering algorithm involving Bregman divergences.

1 Introduction

Data clustering is a fundamental "unsupervised" learning procedure that has been extensively studied across varied disciplines over several years [JD88]. Parametric clustering of data involves finding a partitioning of the data into a pre-specified number of partitions with a *cluster representative* corresponding to every cluster such that a well-defined cost function involving the data and the representatives is minimized. The cost function is normally the expected value of a well-motivated distortion measure between the data-points and their cluster representatives. Usually, the distribution over the data is assumed to be uniform and hence it is sufficient to work with the sum of the distortions, since this is equal to the expectation, in this case, with a multiplicative constant. The `kmeans` [Mac67] problem is perhaps the most well-studied and widely used member of this class of problems.

Typically, parametric clustering problems come in two flavors: *hard* and *soft*. In hard clustering, one obtains a disjoint partitioning of the data such that each data-point belongs to exactly one of the partitions. Moreover, the cluster representative of every partition depends only on the data-points in that partition. In soft clustering, each data-point has a certain probability of belonging to each of the partitions. The cluster representatives are computed using all the data-points with contributions appropriately weighted according to their probability of being in that cluster. In some sense, one can think of hard clustering as a special case of soft clustering where the probabilities of a data-point belonging to a cluster can either be 1 or 0.

Algorithms for solving particular versions of parametric clustering problems have been developed over the years. As far as hard clustering algorithms are concerned, the most well-known

*Department of ECE, University of Texas at Austin
†Department of ECE, University of Texas at Austin
‡Department of CS, University of Texas at Austin
§Department of ECE, University of Texas at Austin

algorithm is the iterative relocation scheme for the `kmeans` problem [JD88, DHS00]. The recently proposed information theoretic clustering algorithm [DMK03] for clustering probability distributions has a similar flavor. On the other hand, the domain of soft clustering algorithms is more well developed. Since most of the practical soft clustering problems can be posed as a problem of finding the parameters of a mixture density under the assumption that the observed data has been sampled from the mixture distribution, the clustering problem boils down to a problem of maximum likelihood parameter estimation. However, for a given data-point, since one does not know the exact component following which it was sampled, the problem is one of maximum-likelihood estimation with incomplete information. The well-known EM algorithm [DLR77] is used to solve the estimation and hence, the soft clustering problem.

Although the `kmeans` and the information theoretic clustering algorithms seem to have a similar flavor in that both of them employ an iterative relocation scheme using a certain distortion function — square of the Euclidean distance in the `kmeans` case, and the KL-divergence in the case of information theoretic clustering — it is not clear for exactly what type of distortion functions such simple schemes will give a clustering of the data. In this article, we answer this question. We introduce a concept called Bregman information of a set and pose an optimal quantization problem such that the loss in Bregman information due to quantization is minimized for a quantization of a given size. We show that the optimal quantization problem is exactly equivalent to a Bregman clustering problem, where the distortion function used for clustering is a Bregman divergence. Then, we show that if the distortion function is a Bregman divergence, there is always an iterative relocation scheme that clusters the data minimizing the given distortion function. In other words, like many other problems in machine learning, the clustering algorithm comes along with the choice of the loss function.

The practical generative models used for soft clustering typically use a mixture density involving an appropriate member of the exponential family. Except for an excellent analysis by Kearns et al [KMN97] involving hard and soft assignments, there does not appear to be much literature on the connection between hard and soft clustering algorithms involving exponential families. In this article, we prove that the density of any exponential distribution can be written as the product of an exponential function of the negative of a uniquely determined Bregman divergence and a function independent of the parameters, thereby exhibiting a bijection between Bregman divergences and exponential families. Using this result, and the Bregman clustering results developed for the hard clustering case, we revisit the EM algorithm for mixture density learning. We demonstrate that the M-step of EM, where most of the computation is involved, simplifies to finding a simple expectation if one is working with an appropriate representation of the sufficient statistic. This also demonstrates the exact connection between the hard and soft Bregman clustering algorithms.

The remainder of this article is organized as follows. We first define Bregman divergences and provide a few illustrative examples in section 2. Then, we introduce the concept of Bregman information to motivate the Bregman hard clustering problem and propose an algorithm to solve this clustering problem in section 3. In section 4, we establish a connection between exponential families and Bregman divergences and use it develop a soft Bregman clustering algorithm in section 5. In section 6, we present some experimental results that illustrate the usefulness of the Bregman clustering algorithm. In section 7, we briefly discuss related work.

A word about the notation: bold faced variables, e.g., $\mathbf{x}, \boldsymbol{\mu}$, etc., represent vectors, sets are represented by calligraphic upper-case alphabets, e.g., $\mathcal{X}, \mathcal{Y}$, etc. and enumerated as $\{\mathbf{x}_i\}_{i=1}^{n}$ where $\mathbf{x}_i$ are the elements of the set and vectors $\mathbf{x} = (x_1, \cdots, x_d)$ are represented as $[x_j]_{j=1}^{d}$. φ denotes the null set and $\mathbb{R}, \mathbb{R}_{++}$ and $\mathbb{R}^d$ denote the set of reals, the set of positive reals and the d-dimensional real vector space respectively. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, $\langle \mathbf{x}, \mathbf{y} \rangle$ is the natural inner product in $\mathbb{R}^d$ and $\|x\|$

is the L_2 norm. Probability density functions are denoted by lower case alphabets, e.g., p, q, etc. Probability measure on a set is denoted by ν. If a random variable $\mathbf{x}$ is distributed as p, we denote this by $\mathbf{x} \sim p$. Expectation of functions of a random variable $\mathbf{x} \sim p$ are denoted by $E_{\mathbf{x}\sim p}[\cdot]$, or, simply $E_p[\cdot]$ when it is clear which random variable is being specified. The inverse of a function f is denoted by f^{-1}.

2 Preliminaries

First, we revisit some basic concepts in analysis [KF75] that we need in order to define Bregman divergences. Then, we define Bregman divergence for a well-behaved class of convex functions and look at a few examples.

Definition 1 *The* **interior** *of a set $S \subseteq \mathbb{R}^d$ is defined as*

$$\text{int}(S) = \{\mathbf{x} \in \mathbb{R}^d | \exists \epsilon > 0, B_\epsilon(\mathbf{x}) \subset S\},$$

where $B_\epsilon(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^d | \rho(\mathbf{y}, \mathbf{x}) < \epsilon\}$ denotes the open metric ϵ-ball centered at $\mathbf{x}$ for a metric ρ defined on $\mathbb{R}^d$.

Definition 2 *Let $\phi : S \mapsto \mathbb{R}$ be a strictly convex function defined on a convex set $S \subseteq \mathbb{R}^d$ such that ϕ is differentiable on $\text{int}(S) \neq \varphi$. The* **Bregman divergence** *$D_\phi : S \times \text{int}(S) \mapsto [0, \infty)$ is defined as*

$$D_\phi(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}, \nabla\phi(\mathbf{y}) \rangle.$$

Example 1: The well-known squared Euclidean distance is perhaps the simplest and most widely used Bregman divergence. In this case, $\phi(\mathbf{x}) = \langle \mathbf{x}, \mathbf{x} \rangle$ is a strictly convex, differentiable function on $\mathbb{R}^d$ and

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

which is exactly the squared Euclidean distance. ∎

Example 2: Another widely used Bregman divergence is the KL-divergence. If $\mathbf{p}$ is a discrete probability distribution so that $\sum_{j=1}^d p_j = 1$, then $\phi(\mathbf{p}) = \sum_{j=1}^d p_j \log p_j$, which is the negative entropy is a convex function in this domain (the d-simplex). Then, the corresponding Bregman divergence

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

Table 1: Bregman divergences corresponding to some convex functions. Note that logarithm is to the natural base and A is positive definite.

Domain	$\phi(\mathbf{x})$	$D_\phi(\mathbf{x}, \mathbf{y})$	**Divergence**
$\mathbb{R}$	x^2	$(x-y)^2$	Square loss
$\mathbb{R}_{++}$	$x \log x$	$x \log(\frac{x}{y}) - (x-y)$	
$(0,1)$	$x \log x + (1-x)\log(1-x)$	$x \log(\frac{x}{y}) + (1-x)\log(\frac{1-x}{1-y})$	Logistic loss [1]
$\mathbb{R}_{++}$	$-\log x$	$\frac{x}{y} - \log(\frac{x}{y}) - 1$	Itakura-Saito distance
$\mathbb{R}$	e^x	$e^x - e^y - (x-y)e^y$	
$\mathbb{R} \setminus \{0\}$	$\lvert x \rvert$	$\max\{0, -2\,\mathrm{sign}(y)x\}$	Hinge loss
$\mathbb{R}^d$	$\lVert \mathbf{x} \rVert^2$	$\lVert \mathbf{x} - \mathbf{y} \rVert^2$	Squared Euclidean distance
$\mathbb{R}^d$	$\mathbf{x}^T A \mathbf{x}$	$(\mathbf{x}-\mathbf{y})^T A (\mathbf{x}-\mathbf{y})$	Mahalanobis distance [2]
d-Simplex	$\sum_{j=1}^d x_j \log x_j$	$\sum_{j=1}^d x_j \log(\frac{x_j}{y_j})$	KL-divergence
$\mathbb{R}_+^d$	$\sum_{j=1}^d x_j \log x_j$	$\sum_{j=1}^d x_j \log(\frac{x_j}{y_j}) - \sum_{j=1}^d (x_j - y_j)$	Generalized I-divergence

the KL-divergence between the two distributions as $\sum_{j=1}^d q_j = \sum_{j=1}^d p_j = 1$. ■

Table 1 contains a list of some common convex functions and their corresponding Bregman divergences. Some basic properties of Bregman divergences are listed in the Appendix.

3 Bregman Hard Clustering

In this section, we first introduce a new concept called the Bregman information of a set based on ideas from Shannon's rate-distortion theory. Then, we motivate the Bregman hard clustering problem as a quantization problem that involves minimizing the loss in Bregman information and show that it is equivalent to a more direct formulation, i.e., the problem of finding a partitioning and a representative for each of the partitions such that the expected Bregman divergence of the points from their representatives is minimized. We also propose a clustering algorithm that is a generalization of the `kmeans` algorithm and is guaranteed to converge to a local minimum of the Bregman hard clustering problem.

3.1 Bregman Information

Before we go on to define Bregman Information, we briefly relate the relevant concepts in Shannon's rate distortion theory to the Bregman clustering problem. In the general rate-distortion setting [CT91], a random variable is coded using a scheme that consists of an encoding and a decoding function. The **rate** of the coding scheme is the number of bits used for encoding and can be considered a measure of the size of codebook (2^R where R is rate). The performance of the coding scheme is determined in terms of the **expected distortion** between the source random variable and the decoded random variable, for an appropriate application dependent distortion function. The rate distortion problem [GV03], can be stated as the problem of finding a coding scheme with a given rate, R, such that the expected distortion between the source random variable and the decoded random variable, is minimized. The achieved distortion is called the **distortion-**

[1] $x \log(\frac{x}{y}) + (1-x)\log(\frac{1-x}{1-y}) = \log(1 + exp(-f(x)g(y)))$, i.e. logistic loss where $f(x) = 2x - 1$ and $g(y) = \log(\frac{y}{1-y})$

[2] $(\mathbf{x}-\mathbf{y})^T A (\mathbf{x}-\mathbf{y})$ is the Mahalanobis distance when A is the inverse of the covariance matrix

rate function, i.e., infimum distortion achievable for a given rate, or, in other words, for a given codebook size.

For the current analysis, let us consider a simple coding scheme for a random variable X that takes values in a finite set $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n \subset S \subset \mathbb{R}^d$ following a probability measure ν. The encoding scheme involves a quantization of the random variable and decoding is just the identity map. The size of the codebook, i.e., the set of quantized values, determines the rate of the coding scheme or rate of quantization. Assuming the distortion function to be a Bregman divergence D_ϕ, the rate-distortion problem for this coding scheme can be stated as

$$\min_F \{\mathbf{E}_\nu[D_\phi(X, \hat{X})]\}, \tag{1}$$

where F is the quantization mapping with a fixed rate, R and $\hat{X} = F(X)$ is the encoded version of X. First consider the case where the rate of quantization, $R = 0$ i.e., the codebook is a singleton set. From a stochastic viewpoint, the encoded version $\hat{X}$ of the random variable X is a constant, say $\mathbf{s} \in S$ and the joint distribution of $(X, \hat{X})$ is equivalent to the marginal distribution of X. The distortion-rate function for rate $R = 0$ and distortion function, D_ϕ is given by

$$\min_{\mathbf{s}\in S} \mathbf{E}_\nu[D_\phi(X, \mathbf{s})] = \min_{\mathbf{s}\in S} \sum_{i=1}^n \nu_i \; D_\phi(\mathbf{x}_i, \mathbf{s}). \tag{2}$$

We call this distortion-rate function the **Bregman information** of the set $\mathcal{X}$ for the Bregman divergence, D_ϕ and denote it by $I_\phi(\mathcal{X})$. The optimal value of $\mathbf{s}$ that achieves the minimal distortion will be called the *Bregman representative* or, simply the *representative* of the set $\mathcal{X}$. In the subsequent analysis, we shall show that this representative always exists, is uniquely determined and, surprisingly, does not depend on the choice of the Bregman divergence. Note that a lower value of Bregman information indicates that the elements of the set are, in a Bregman divergence sense, closer to the representative. On the other hand, a higher value of Bregman information indicates that the single representative cannot capture the diversity in the set, and it may be a good idea to partition the set into more homogeneous subsets and have a representative for each subset. This directly leads to the Bregman clustering problem.
Before proceeding further, we show the existence and uniqueness of the Bregman representative.

Theorem 1 *Given a set $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n \subset S \subseteq \mathbb{R}^d$, a probability measure ν over $\mathcal{X}$ and a Bregman divergence $D_\phi : S \times \mathrm{int}(S) \mapsto [0, \infty)$, the problem*

$$\min_{\mathbf{s}\in S} E_\nu[D_\phi(\mathbf{x}, \mathbf{s})]$$

has a unique minimizer given by $\mathbf{s}^* = \boldsymbol{\mu} = \mathbf{E}_\nu[\mathbf{x}]$.

Proof: The function we are trying to minimize is

$$J_\phi(\mathbf{s}) = E_\nu[D_\phi(\mathbf{x}, \mathbf{s})] = \sum_{i=1}^n \nu_i D_\phi(\mathbf{x}_i, \mathbf{s}).$$

We prove the required result by showing that for $\forall \mathbf{s} \in S$, $J_\phi(\mathbf{s}) \geq J_\phi(\boldsymbol{\mu})$ where $\boldsymbol{\mu} = \mathbf{E}_\nu[\mathbf{x}]$ and

equality holds only when $\mathbf{s} = \boldsymbol{\mu}$. To this end, we note that

$$\begin{aligned} J_\phi(\mathbf{s}) - J_\phi(\boldsymbol{\mu}) &= \sum_{i=1}^{n} \nu_i D_\phi(\mathbf{x}_i, \mathbf{s}) - \sum_{i=1}^{n} \nu_i D_\phi(\mathbf{x}_i, \boldsymbol{\mu}) \\ &= \sum_{i=1}^{n} \nu_i (\phi(\mathbf{x}_i) - \phi(\mathbf{s}) - \langle \mathbf{x}_i - \mathbf{s}, \nabla\phi(\mathbf{s})\rangle - \phi(\mathbf{x}_i) + \phi(\boldsymbol{\mu}) + \langle \mathbf{x}_i - \boldsymbol{\mu}, \nabla\phi(\boldsymbol{\mu})\rangle) \\ &= \phi(\boldsymbol{\mu}) - \phi(\mathbf{s}) - \langle (\sum_{i=1}^{n} \nu_i \mathbf{x}_i) - \mathbf{s}, \nabla\phi(\mathbf{s})\rangle + \langle (\sum_{i=1}^{n} \nu_i \mathbf{x}_i) - \boldsymbol{\mu}, \nabla\phi(\boldsymbol{\mu})\rangle \\ &= \phi(\boldsymbol{\mu}) - \phi(\mathbf{s}) - \langle \boldsymbol{\mu} - \mathbf{s}, \nabla\phi(\mathbf{s})\rangle \\ &= D_\phi(\boldsymbol{\mu}, \mathbf{s}) \geq 0, \end{aligned}$$

with equality only when $\mathbf{s} = \boldsymbol{\mu}$. Hence, $\boldsymbol{\mu}$ is the unique minimizer of the function, J_ϕ.

Finally, we argue that $\boldsymbol{\mu} \in S$. To this end, since $\mathcal{X} \subset S$ and S is a convex set, $\text{co}(\mathcal{X}) \subset S$, where $\text{co}(\mathcal{X})$ is the convex hull of $\mathcal{X}$. But $\boldsymbol{\mu} = E_\nu[x] \in \text{co}(\mathcal{X})$. Hence $\boldsymbol{\mu} \in S$. That completes the proof. ∎

The above result shows that the representative, i.e., the minimizer of the expected Bregman divergence, is always the expectation of the set even when the Bregman divergence is not convex in the second argument[1].

Using the above theorem, we can now give a more direct definition of the Bregman information as follows:

Definition 3 *Let $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n \subset S$ be a finite subset of S and ν be a probability measure on $\mathcal{X}$ and let $\boldsymbol{\mu} = \mathbf{E}_\nu[\mathbf{x}] = \sum_{i=1}^n \nu_i \mathbf{x}_i$. Let $D_\phi : S \times \text{int}(S) \mapsto [0, \infty)$ be a Bregman divergence. Then,* **Bregman Information** *of $\mathcal{X}$ in terms of D_ϕ is defined as*

$$I_\phi(\mathcal{X}) = E_\nu[D_\phi(\mathbf{x}, \boldsymbol{\mu})] = \sum_{i=1}^{n} \nu_i D_\phi(\mathbf{x}_i, \boldsymbol{\mu}).$$

To start appreciating the potential of such a treatment, we note that the elements of the set $\mathcal{X}$ can be quite general. For instance, the elements can be probability distributions, functionals, operators or just plain vectors. In the following examples, we look at sets of vectors and probability distributions on a fixed discrete output space.

Example 3: One simple example of Bregman information is the variance. Let $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$ be a set in $\mathbb{R}^d$. The Bregman information of the set $\mathcal{X}$ with the squared Euclidean distance as the Bregman divergence is given by

$$\begin{aligned} I_\phi(\mathcal{X}) &= \sum_{i=1}^{n} \nu_i D_\phi(\mathbf{x}_i, \boldsymbol{\mu}) \\ &= \frac{1}{n} \sum_{i=1}^{n} \|\mathbf{x}_i - \boldsymbol{\mu}\|^2 \ (\text{ Assuming uniform measure, } \nu_i = \frac{1}{n}), \end{aligned}$$

[1] For example, consider $\phi(\mathbf{x}) = \sum_{j=1}^3 x_j^3$ defined on $\mathbb{R}^3_{++}$ so that $D_\phi(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^3 (x_j^3 - y_j^3 - 3(x_j - y_j)y_j^2)$ is not convex in $\mathbf{y}$. Now for the set $\mathcal{X} = \{(1,1,1), (2,2,2), (3,3,3), (4,4,4), (5,5,5)\}$, the expected Bregman divergence with respect to a point $\mathbf{y}$ is given by $(135 + 2\sum_{j=1}^3 y_j^3 - 9\sum_{j=1}^3 y_j^2)$, which is minimized when $\mathbf{y} = (3,3,3)$, i.e., the expectation of the original set $\mathcal{X}$.

which is just the sample variance of the set $\mathcal{X}$. ■

Example 4: Another example involves a set of probability distributions, which can also be interpreted as conditional distributions given a random variable. In particular, we show that if random variables (U, V) are jointly distributed according to $\{\{p(\mathbf{u}_i, \mathbf{v}_j)\}_{i=1}^n\}_{j=1}^m$, then the mutual information $I(U;V)$ is the Bregman information of the set of conditional distributions $\{p(V|\mathbf{u}_i)\}_{i=1}^n$ with KL-divergence as the Bregman divergence. By definition,

$$\begin{aligned} I(U;V) &= \sum_{i=1}^{n}\sum_{j=1}^{m} p(\mathbf{u}_i, \mathbf{v}_j)\left(\log \frac{p(\mathbf{u}_i, \mathbf{v}_j)}{p(\mathbf{u}_i)p(\mathbf{v}_j)}\right) \\ &= \sum_{i=1}^{n} p(\mathbf{u}_i)\sum_{j=1}^{m} p(\mathbf{v}_j|\mathbf{u}_i)\left(\log \frac{p(\mathbf{v}_j|\mathbf{u}_i)}{p(\mathbf{v}_j)}\right) \\ &= \sum_{i=1}^{n} p(\mathbf{u}_i) KL(\ p(V|\mathbf{u}_i)\ \|\ p(V)\ ). \end{aligned}$$

Consider the set of probability distributions $\mathcal{Z}_\mathbf{u} = \{p(V|\mathbf{u}_i)\}_{i=1}^n$ and the probability measure $\{\nu_i\}_{i=1}^n = \{p(\mathbf{u}_i)\}_{i=1}^n$ over this set. For this set, the mean distribution is given by

$$\boldsymbol{\mu} = E_\nu[p(V|\mathbf{u})] = \sum_{i=1}^{n} p(\mathbf{u}_i)p(V|\mathbf{u}_i) = \sum_{i=1}^{n} p(\mathbf{u}_i, V) = p(V).$$

Hence,

$$\begin{aligned} I(U;V) &= \sum_{i=1}^{n} p(\mathbf{u}_i) KL(\ p(V|\mathbf{u}_i)\ \|\ p(V)\ ) \\ &= \sum_{i=1}^{n} \nu_i D_\phi(p(V|\mathbf{u}_i), \boldsymbol{\mu}) \\ &= I_\phi(\mathcal{Z}_\mathbf{u}), \end{aligned}$$

i.e., mutual information is a special case of Bregman information. Further, for the set of probability distributions $\mathcal{Z}_\mathbf{v} = \{p(U|\mathbf{v}_j)\}_{j=1}^m$ and the probability measure $\nu_j = p(\mathbf{v}_j)$ over this set, one can similarly show that $I(U;V) = I_\phi(\mathcal{Z}_\mathbf{v})$. The Bregman information of the two sets of probability distributions, $\mathcal{Z}_\mathbf{v}$ and $\mathcal{Z}_\mathbf{u}$ can also be interpreted as the Jensen-Shannon divergence [DMK03] of those sets. ■

3.2 Problem Formulation

As mentioned earlier, if $\mathcal{X}$ is a set with high Bregman information, it may not be justifiable to have a single Bregman representative for the entire set from a quantization error point of view. In such a situation, partitioning the set into relatively homogeneous groups and having a representative for each group seems like a natural idea. Note that this implies a higher rate of quantization, i.e., a codebook of larger size. Now, in addition to achieving minimal expected Bregman divergence within their corresponding partitions, the set of representatives should also preserve the Bregman information content of the original set as much as possible. If the codebook size is greater than or equal to the cardinality of $\mathcal{X}$, a trivial solution is to use $|\mathcal{X}|$ representatives, being one-to-one with the actual elements in the set. A more interesting case arises when the allowed codebook size

is strictly less than $|\mathcal{X}|$. If $\mathcal{M}$ denotes the set of Bregman representatives with $|\mathcal{M}| = k < |\mathcal{X}|$, the problem is to choose $\mathcal{M}$, or, equivalently, the k-partitioning of $\mathcal{X}$ such that $\mathcal{M}$ preserves as much Bregman information as possible compared to the original set. This is precisely the Bregman clustering problem.

On more concrete terms, the set of representatives, $\mathcal{M}$, will have its own Bregman information. If $I_\phi(\mathcal{M})$ denotes the Bregman information of $\mathcal{M}$, the loss in Bregman information due to quantization is given by

$$L_\phi(\mathcal{M}) = I_\phi(\mathcal{X}) - I_\phi(\mathcal{M}). \tag{3}$$

The **Bregman hard clustering problem** is to find a partitioning of $\mathcal{X}$, or, equivalently, the set of representatives $\mathcal{M}$, such that *the loss in Bregman information, $L_\phi(\mathcal{M})$, is minimized.* Note that the classical vector quantization problem is a special case of this formulation.

In the rest of this section, we show that this loss function can be written in a different form that suggests a natural solution to this problem. This result will be used to develop an algorithm for solving arbitrary Bregman clustering problems efficiently.

Theorem 2 *Let $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n \subset S \subseteq \mathbb{R}^d$ and let ν be a probability measure on $\mathcal{X}$. Let $\{\mathcal{X}_h\}_{h=1}^k$ be a partitioning of $\mathcal{X}$ and let $\pi_h = \sum_{\mathbf{x}_i \in \mathcal{X}_h} \nu_i$. If $\mathcal{M} = \{\boldsymbol{\mu}_h\}_{h=1}^k$ denotes the set of representatives, then*

$$L_\phi(\mathcal{M}) = I_\phi(\mathcal{X}) - I_\phi(\mathcal{M}) = \mathbf{E}_\pi[I_\phi(\mathcal{X}_h)], \tag{4}$$

where

$$\mathbf{E}_\pi[I_\phi(\mathcal{X}_h)] = \sum_{h=1}^k \pi_h I_\phi(\mathcal{X}_h) = \sum_{h=1}^k \pi_h \sum_{\mathbf{x}_i \in \mathcal{X}_h} \frac{\nu_i}{\pi_h} D_\phi(\mathbf{x}_i, \boldsymbol{\mu}_h). \tag{5}$$

Proof: By definition,

$$\begin{aligned}
I_\phi(\mathcal{X}) &= \sum_{i=1}^n \nu_i D_\phi(\mathbf{x}_i, \boldsymbol{\mu}) = \sum_{h=1}^k \sum_{\mathbf{x}_i \in \mathcal{X}_h} \nu_i D_\phi(\mathbf{x}_i, \boldsymbol{\mu}) \\
&= \sum_{h=1}^k \sum_{\mathbf{x}_i \in \mathcal{X}_h} \nu_i \left\{\phi(\mathbf{x}_i) - \phi(\boldsymbol{\mu}) - \langle \mathbf{x}_i - \boldsymbol{\mu}, \nabla\phi(\boldsymbol{\mu})\rangle\right\} \\
&= \sum_{h=1}^k \sum_{\mathbf{x}_i \in \mathcal{X}_h} \nu_i \left\{\phi(\mathbf{x}_i) - \phi(\boldsymbol{\mu}_h) - \langle (\mathbf{x}_i - \boldsymbol{\mu}_h), \nabla\phi(\boldsymbol{\mu}_h)\rangle + \langle (\mathbf{x}_i - \boldsymbol{\mu}_h), \nabla\phi(\boldsymbol{\mu}_h)\rangle \right. \\
&\qquad \left. + \phi(\boldsymbol{\mu}_h) - \phi(\boldsymbol{\mu}) - \langle \mathbf{x}_i - \boldsymbol{\mu}_h + \boldsymbol{\mu}_h - \boldsymbol{\mu}, \nabla\phi(\boldsymbol{\mu})\rangle\right\} \\
&= \sum_{h=1}^k \sum_{\mathbf{x}_i \in \mathcal{X}_h} \nu_i \left\{ D_\phi(\mathbf{x}_i, \boldsymbol{\mu}_h) + D_\phi(\boldsymbol{\mu}_h, \boldsymbol{\mu}) + \langle (\mathbf{x}_i - \boldsymbol{\mu}_h), (\nabla\phi(\boldsymbol{\mu}_h) - \nabla\phi(\boldsymbol{\mu}))\rangle \right\} \\
&= \sum_{h=1}^k \pi_h \sum_{\mathbf{x}_i \in \mathcal{X}_h} \frac{\nu_i}{\pi_h} D_\phi(\mathbf{x}_i, \boldsymbol{\mu}_h) + \sum_{h=1}^k \sum_{\mathbf{x}_i \in \mathcal{X}_h} \nu_i D_\phi(\boldsymbol{\mu}_h, \boldsymbol{\mu}) \\
&\qquad + \sum_{h=1}^k \pi_h \sum_{x_i \in \mathcal{X}_h} \frac{\nu_i}{\pi_h} \langle \mathbf{x}_i - \boldsymbol{\mu}_h, (\nabla\phi(\boldsymbol{\mu}_h) - \nabla\phi(\boldsymbol{\mu}))\rangle
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{h=1}^{k} \pi_h I_\phi(\mathcal{X}_h) + \sum_{h=1}^{k} \pi_h D_\phi(\boldsymbol{\mu}_h, \boldsymbol{\mu}) + \sum_{h=1}^{k} \pi_h \langle (\sum_{\mathbf{x}_i \in \mathcal{X}_h} \frac{\nu_i}{\pi_h} \mathbf{x}_i - \boldsymbol{\mu}_h), \nabla\phi(\boldsymbol{\mu}_h) - \nabla\phi(\boldsymbol{\mu}) \rangle \\
&= \mathbf{E}_\pi[I_\phi(\mathcal{X}_h)] + I_\phi(\mathcal{M}),
\end{aligned}
$$

since $\sum_{\mathbf{x}_i \in \mathcal{X}_h} \frac{\nu_i}{\pi_h} \mathbf{x}_i = \boldsymbol{\mu}_h$ by definition. By rearranging terms, we get the desired result. ∎

Hence, the Bregman clustering problem of minimizing the loss in Bregman information can be written as

$$
\min_{\mathcal{M}} L_\phi(\mathcal{M}) = \min_{\mathcal{M}} \sum_{h=1}^{k} \pi_h \sum_{\mathbf{x}_i \in \mathcal{X}_h} \frac{\nu_i}{\pi_h} D_\phi(\mathbf{x}_i, \boldsymbol{\mu}_h), \tag{6}
$$

where $\mathcal{M}$ is the set of representatives. The loss in Bregman information is minimized if the set of representatives $\mathcal{M}$ is such that the expected Bregman divergence of points in the original set $\mathcal{X}$ to their corresponding representatives is minimized. Since, the original set is partitioned such that every $\mathbf{x} \in \mathcal{X}$ belongs to exactly one of the final partitions, we call this version of the problem the Bregman hard clustering problem. A soft version, where the points can have non-zero probabilities of belonging to multiple partitions, will be developed and discussed in section 5.

3.3 Clustering Algorithm

Given a set $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$ and a probability measure ν over $\mathcal{X}$, the Bregman hard clustering problem is to find, for a given k, a k-partitioning $\{\mathcal{X}_h\}_{h=1}^k$ of $\mathcal{X}$ and a set of representatives, $\mathcal{M} = \{\boldsymbol{\mu}_h\}_{h=1}^k$, such that the loss function

$$
L_\phi(\mathcal{M}) = \sum_{h=1}^{k} \pi_h I_\phi(\mathcal{X}_h) = \sum_{h=1}^{k} \pi_h \sum_{\mathbf{x}_i \in \mathcal{X}_h} \frac{\nu_i}{\pi_h} D_\phi(\mathbf{x}_i, \boldsymbol{\mu}_h) \tag{7}
$$

is minimized. The above formulation suggests a natural iterative relocation algorithm to solve the Bregman hard clustering problem. Details of the proposed method are given in Algorithm 1. We call this the Bregman hard clustering algorithm. It is easy to see that classical `kmeans` and information theoretic clustering are special cases of the Bregman hard clustering algorithm for squared Euclidean distance and KL-divergence respectively. For both these cases, the induced partitions are known to have linear separators. We now see that this is true for all Bregman divergences since the locus of points that are equidistant to two fixed points in terms of a Bregman divergence is always a hyperplane. The following theorems prove the convergence of the Bregman hard clustering algorithm.

Proposition 1 *The Bregman hard clustering algorithm (Algorithm 1) monotonically decreases the loss function in (7).*

Proof: Let $\{\mathcal{X}_h^{(t)}\}_{h=1}^k$ be the partitioning of $\mathcal{X}$ after the t^{th} iteration and let $\mathcal{M}^{(t)} = \{\boldsymbol{\mu}_h^{(t)}\}_{h=1}^k$ be the corresponding set of cluster representatives. Then,

$$
\begin{aligned}
L_\phi(\mathcal{M}^{(t)}) &= \sum_{h=1}^{k} \sum_{\mathbf{x}_i \in \mathcal{X}_h^{(t)}} \nu_i D_\phi(\mathbf{x}_i, \boldsymbol{\mu}_h^{(t)}) \qquad \geq \sum_{h=1}^{k} \sum_{\mathbf{x}_i \in \mathcal{X}_h^{(t)}} \nu_i D_\phi(\mathbf{x}_i, \boldsymbol{\mu}_{h^*(\mathbf{x}_i)}^{(t)}) \\
&\geq \sum_{h=1}^{k} \sum_{\mathbf{x}_i \in \mathcal{X}_h^{(t+1)}} \nu_i D_\phi(\mathbf{x}_i, \boldsymbol{\mu}_h^{(t+1)}) = L_\phi(\mathcal{M}^{(t+1)}),
\end{aligned}
$$

Algorithm 1 Bregman Hard-Clustering

Input: Set $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n \subset S \subseteq \mathbb{R}^d$, probability measure ν[1] over $\mathcal{X}$, Bregman divergence $D_\phi : \mathcal{S} \times \text{int}(S) \mapsto \mathbb{R}$, number of clusters k.

Output: $\mathcal{M}^*$, a local minimizer of $\sum_{h=1}^k \sum_{\mathbf{x}_i \in \mathcal{X}_h} \nu_i D_\phi(\mathbf{x}_i, \boldsymbol{\mu}_h)$ where $\mathcal{M} = \{\boldsymbol{\mu}_h\}_{h=1}^k$, corresponding partitioning $\{\mathcal{X}_h\}_{h=1}^k$ of $\mathcal{X}$.

Method:

Initialize $\{\boldsymbol{\mu}_h\}_{h=1}^k$ with some $\boldsymbol{\mu}_h \in S$
repeat
{The Assignment Step}
Set $\mathcal{X}_h \leftarrow \varphi,\ h = 1, \cdots, k$
for $i = 1$ to n **do**
$\mathcal{X}_h \leftarrow \mathcal{X}_h \cup \{\mathbf{x}_i\}$ where $h = h^*(\mathbf{x}_i) = \operatorname{argmin}_{h'} D_\phi(\mathbf{x}_i, \boldsymbol{\mu}_{h'})$
end for
{The Re-estimation Step}
for $h = 1$ to k **do**
$\pi_h \leftarrow \sum_{\mathbf{x}_i \in \mathcal{X}_h} \nu_i$
$\boldsymbol{\mu}_h \leftarrow \sum_{\mathbf{x}_i \in \mathcal{X}_h} \frac{\nu_i}{\pi_h} \mathbf{x}_i$
end for
until *convergence*
return $\mathcal{M}^* = \{\boldsymbol{\mu}_h\}_{h=1}^k$

where the first inequality follows trivially from the criteria used for the assignment of each of the points in the assignment step, and the second inequality follows from the re-estimation procedure using Theorem 1. Note that if equality holds, i.e., if the loss function value is equal at consecutive iterations, then the algorithm terminates. ∎

Proposition 2 *The Bregman hard clustering algorithm (Algorithm 1) terminates in a finite number of steps at a partition that is locally optimal, i.e., the total loss cannot be decreased by either (a) reassignment of points to different clusters or by (b) changing the means of any existing clusters.*

Proof: The result follows since the algorithm monotonically decreases the objective function value, and the number of distinct clusterings is finite. ∎

4 Bijection with Exponential Families

In this section, we establish a bijection between Bregman divergences and exponential families. We also provide examples of Bregman divergences obtained from some popular exponential families. The bijection will be used to develop Bregman soft clustering algorithm in section 5. Sections 4.1 and 4.2, which provide the background concepts needed to state and prove the bijection result, may be skipped in a first reading by those who are not familiar with this subject. It has been observed in the literature [Ama95] that exponential families and Bregman divergences have certain relationships that can be exploited for several learning problems. In particular, [FW00] observes

[1] We could, in general, have any non-negative weights and normalize them so as to sum to 1.

that Bregman divergences are a generalization of the negative log-likelihood of any member of the exponential family. We state this connection more precisely by providing a constructive proof of an explicit bijection between Bregman divergences and exponential families. This result is useful as it enables us to obtain the appropriate divergence for any given exponential family.

4.1 Exponential families

Consider a family $\mathcal{F}$ of probability densities[1] on a measurable space $(\Omega, \mathcal{B})$ where $\mathcal{B}$ is a σ-algebra on the set Ω [FG97]. Suppose every probability density, $p_{\boldsymbol{\theta}} \in \mathcal{F}$, is parameterized by d real-valued variables $\boldsymbol{\theta} = \{\theta_j\}_{j=1}^d$ so that

$$\mathcal{F} = \{p_{\boldsymbol{\theta}} = f(\omega; \boldsymbol{\theta}) | \omega \in \mathcal{B}, \boldsymbol{\theta} \in \Gamma \subseteq \mathbb{R}^d\}.$$

Then, $\mathcal{F}$ is called a d-dimensional parametric model on $(\Omega, \mathcal{B})$. Let $H : \mathcal{B} \mapsto \mathcal{G}$ be a ($\mathcal{B}$-$\mathcal{G}$ measurable) function that transforms any random variable $U : \mathcal{B} \mapsto \mathbb{R}$ to a random variable $V : \mathcal{G} \mapsto \mathbb{R}$ with $V = H(U)$. Then, given the probability density $p_{\boldsymbol{\theta}}$ of U, this function uniquely determines the probability density $q_{\boldsymbol{\theta}}$ governing the random variable V.

Definition 4 *If $\forall \omega \in \mathcal{B}$, $p_{\boldsymbol{\theta}}(\omega)/q_{\boldsymbol{\theta}}(\omega)$ exists and does not depend on $\boldsymbol{\theta}$, then H is called a sufficient statistic for the model $\mathcal{F}$.*

The identity map $H(U) = U$ is a trivial example of a sufficient statistic.

If a d-dimensional model $\mathcal{F} = \{p_{\boldsymbol{\theta}} | \boldsymbol{\theta} \in \Gamma\}$ can be expressed in terms of $(d+1)$ real-valued linearly independent functions $\{C, H_1, \cdots, H_d\}$ on $\mathcal{B}$ and a function ψ on Γ as

$$f(\omega; \boldsymbol{\theta}) = \exp\left\{\sum_{j=1}^{d} \theta_j H_j(\omega) - \psi(\boldsymbol{\theta}) + C(\omega)\right\},$$

then $\mathcal{F}$ is called an **exponential family**, and $\boldsymbol{\theta}$ is called its **natural parameter**. It can be easily seen that if $\mathbf{x} \in \mathbb{R}^d$ is such that $x_j = H_j(\omega)$, then the density function $g(\mathbf{x}; \boldsymbol{\theta})$ given by

$$g(\mathbf{x}; \boldsymbol{\theta}) = \exp\left\{\sum_{j=1}^{d} \theta_j x_j - \psi(\boldsymbol{\theta}) - \lambda(\mathbf{x})\right\},$$

for a uniquely determined function $\lambda(\mathbf{x})$, is such that $f(w; \boldsymbol{\theta})/g(\mathbf{x}; \boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}$. Thus, $\mathbf{x}$ is a sufficient statistic for the family. For our analysis, it is convenient to work with the sufficient statistic $\mathbf{x}$ and hence, we redefine exponential families in terms of the probability density of the sufficient statistic variable in $\mathbb{R}^d$, noting that the original σ-algebra $\mathcal{B}$ can actually be quite general.

Definition 5 *A multivariate parametric family $\mathcal{F}_\psi$ of distributions $\{p_{(\psi,\boldsymbol{\theta})} | \boldsymbol{\theta} \in \Gamma \subseteq \mathbb{R}^d\}$ is called an exponential family if the probability density is of the form*

$$p_{(\psi,\boldsymbol{\theta})}(\mathbf{x}) = \exp(\langle \mathbf{x}, \boldsymbol{\theta} \rangle - \psi(\boldsymbol{\theta}) - \lambda(\mathbf{x})).$$

The function $\psi(\boldsymbol{\theta})$ is known as the **log partition function** or the **cumulant function** and it uniquely determines the exponential family $\mathcal{F}_\psi$. Further, given an exponential family $\mathcal{F}_\psi$, the log-partition function, ψ is uniquely determined up to a constant additive term. It can be shown [Ama95] that Γ is a convex set in $\mathbb{R}^d$ and ψ is a strictly convex and differentiable function on $\text{int}(\Gamma)$.

[1] It is possible to have probability distributions without a corresponding well-defined density function, i.e., Radon Nikodym derivative with respect to the Lebesgue measure, but all exponential distributions have well-defined densities.

4.2 Expectation parameters and Legendre duality

Consider a d-dimensional real random variable X following an exponential density[1] $p_{(\psi,\boldsymbol{\theta})}$. specified by the natural parameter $\boldsymbol{\theta} \in \Gamma$. The expectation of X with respect to $p_{(\psi,\boldsymbol{\theta})}$, also called as the **expectation parameter** is given by

$$\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta}) = E_{p_{(\psi,\boldsymbol{\theta})}}[X] = \int_{\mathbb{R}^d} \mathbf{x} p_{(\psi,\boldsymbol{\theta})}(\mathbf{x}) d\mathbf{x}. \tag{8}$$

It can be shown [Ama95] that the expectation and natural parameters have a one-one correspondence with each other and span spaces that exhibit a dual relationship. To specify the duality more precisely, we first define Legendre conjugates. The Legendre conjugate ψ^c of the function ψ is given by

$$\psi^c(\mathbf{s}) = \sup_{\boldsymbol{\theta}} \{\langle \mathbf{s}, \boldsymbol{\theta} \rangle - \psi(\boldsymbol{\theta})\}.$$

As ψ is a strictly convex and differentiable function over its domain Γ, we can obtain the $\boldsymbol{\theta}$ corresponding to the supremum by setting the gradient of the corresponding function to zero, i.e.,

$$\nabla(\langle \mathbf{s}, \boldsymbol{\theta} \rangle - \psi(\boldsymbol{\theta}))\,|_{\boldsymbol{\theta}=\boldsymbol{\theta}*} = 0 \Rightarrow \mathbf{s} = \nabla\psi(\boldsymbol{\theta}^*)$$

From the above equation, we can see that the conjugate function is well defined on the gradient space of the function ψ, say Γ^c. Further, the strict convexity of ψ implies that $\nabla\psi$ is monotonic and hence, is a bijection from Γ to Γ^c. Hence, for every $\mathbf{s} \in \Gamma^c$, there exists a $\boldsymbol{\theta} = \boldsymbol{\theta}(\mathbf{s}) \in \Gamma$ and for every $\boldsymbol{\theta} \in \Gamma$, there exists a $\mathbf{s} = \mathbf{s}(\boldsymbol{\theta}) \in \Gamma^c$ such that $\mathbf{s} = \nabla\psi(\boldsymbol{\theta})$. It is, therefore, possible to define the inverse function $(\nabla\psi)^{-1} : \Gamma^c \mapsto \Gamma$ and write the conjugate function ψ^c in a closed form as

$$\psi^c(\mathbf{s}) = \langle (\nabla\psi)^{-1}(\mathbf{s}), \mathbf{s} \rangle - \psi((\nabla\psi)^{-1}(\mathbf{s})).$$

It can be shown [Roc70] that the function ψ^c is also a strictly convex and differentiable function on its domain and that the pairs (ψ, Γ) and (ψ^c, Γ^c) are Legendre conjugates of each other. This is stated more formally below.

Definition 6 *[Roc70] Let $\psi : \Gamma \mapsto \mathbb{R}$ be a strictly convex, differentiable function, then the Legendre conjugate of (ψ, Γ) is given by (ψ^c, Γ^c) where Γ^c is the image of Γ under the gradient mapping $\nabla\psi$ and $\psi^c : \Gamma^c \mapsto \mathbb{R}$ is a strictly convex, differentiable function given by*

$$\psi^c(\mathbf{s}) = \langle (\nabla\psi)^{-1}(\mathbf{s}), \mathbf{s} \rangle - \psi((\nabla\psi)^{-1}(\mathbf{s})).$$

Further, (ψ, Γ) is the Legendre conjugate of (ψ^c, Γ^c). The gradient functions $\nabla\psi : \Gamma \mapsto \Gamma^c$ and $\nabla\psi^c : \Gamma^c \mapsto \Gamma$ are both continuous, one-one functions and also form inverses of each other.

Let us now look at the relationship between $\boldsymbol{\theta}$ and the expectation parameter $\boldsymbol{\mu}$ defined in (8). Differentiating the identity $\int p_{(\psi,\boldsymbol{\theta})}(\mathbf{x})d\mathbf{x} = 1$ with respect to $\boldsymbol{\theta}$ gives us $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta}) = \nabla\psi(\boldsymbol{\theta})$, i.e., the expectation parameter $\boldsymbol{\mu}$ is the image of the natural parameter $\boldsymbol{\theta}$ under the gradient mapping $\nabla\psi$. Let S be the expectation parameter space, $\boldsymbol{\theta}(\boldsymbol{\mu}) = (\nabla\psi)^{-1}(\boldsymbol{\mu})$ be the natural parameter corresponding to $\boldsymbol{\mu}$ and the function $\phi : S \mapsto \mathbb{R}$ be defined as

$$\phi(\boldsymbol{\mu}) = \langle \boldsymbol{\theta}(\boldsymbol{\mu}), \boldsymbol{\mu} \rangle - \psi(\boldsymbol{\theta}(\boldsymbol{\mu})). \tag{9}$$

Then, the pairs (ψ, Γ) and (ϕ, S) form Legendre conjugates of each other, i.e., $\phi = \psi^c$ and $S = \Gamma^c$ and the mappings between the dual spaces are given by the Legendre transformation,

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \nabla\psi(\boldsymbol{\theta}) \text{ and } \boldsymbol{\theta}(\boldsymbol{\mu}) = \nabla\phi(\boldsymbol{\mu}). \tag{10}$$

[1] Exponential densities, in the present context, refer to the probability densities corresponding to members of an exponential family and are not to be confused with density functions of the form $p(x) = \lambda e^{-\lambda x}$

4.3 Bijection Theorem

We are now ready to state the connection between exponential families of distributions and Bregman divergences.

Theorem 3 *Let (ϕ, S) and (ψ, Γ) be Legendre conjugates of each other. Let $D_\phi : S \times \text{int}(S) \mapsto \mathbb{R}$ be the Bregman divergence derived from ϕ. For $\boldsymbol{\theta} \in \Gamma$, let $p_{(\psi,\boldsymbol{\theta})}$ be the exponential probability density derived using $\psi(\boldsymbol{\theta})$ as the log-partition function with $\boldsymbol{\theta}$ as the natural parameter. Let $\boldsymbol{\mu}$ be the corresponding expectation parameter. Then,*

$$p_{(\psi,\boldsymbol{\theta})}(\mathbf{x}) = \exp(-D_\phi(\mathbf{x}, \boldsymbol{\mu}))f_\phi(\mathbf{x}), \tag{11}$$

where $f_\phi : S \mapsto \mathbb{R}$ is a uniquely determined function. Hence, there is a bijection between exponential densities $p_{(\psi,\boldsymbol{\theta})}$ and Bregman divergences $D_\phi(\cdot, \boldsymbol{\mu})$.

Proof: We prove the bijection between the exponential densities $p_{(\psi,\boldsymbol{\theta})}$ and the Bregman divergences $D_\phi(\cdot, \boldsymbol{\mu})$ by first showing that each exponential density $p_{(\psi,\boldsymbol{\theta})}$ corresponds to a unique Bregman divergence $D_\phi(\cdot, \boldsymbol{\mu})$ (one-one) and then arguing that there exists an exponential density corresponding to every Bregman divergence (onto). By definition,

$$\begin{aligned}
p_{(\psi,\boldsymbol{\theta})}(\mathbf{x}) &= \exp(\langle \mathbf{x}, \boldsymbol{\theta} \rangle - \psi(\boldsymbol{\theta}) - \lambda(\mathbf{x})) \\
&= \exp(\langle \mathbf{x}, \nabla\phi(\boldsymbol{\mu}) \rangle + (\phi(\boldsymbol{\mu}) - \langle \boldsymbol{\mu}, \nabla\phi(\boldsymbol{\mu}) \rangle) - \lambda(\mathbf{x})) \quad \text{(using (9) and (10))} \\
&= \exp(-\{\phi(\mathbf{x}) - \phi(\boldsymbol{\mu}) - \langle (\mathbf{x} - \boldsymbol{\mu}), \nabla\phi(\boldsymbol{\mu}) \rangle\} + \{\phi(\mathbf{x}) - \lambda(\mathbf{x})\}) \\
&= \exp(-D_\phi(\mathbf{x}, \boldsymbol{\mu}))\ f_\phi(\mathbf{x}).
\end{aligned}$$

We observe that $p_{(\psi,\boldsymbol{\theta})}$ uniquely determines the log-partition function ψ to a constant additive term so that the gradient space of all the possible functions ψ is the same and the corresponding conjugate functions, ϕ differ only by a constant additive term. Hence the Bregman divergence $D_\phi(\mathbf{x}, \boldsymbol{\mu})$ derived from any of these conjugate functions will be identical, i.e., the mapping is one-one. This also implies that f_ϕ is a uniquely determined function on S.

$$\begin{aligned}
\psi_2(\boldsymbol{\theta}) = \psi_1(\boldsymbol{\theta}) + c \quad &\Rightarrow \quad \nabla\psi_2(\boldsymbol{\theta}) = \nabla\psi_1(\boldsymbol{\theta}) = \boldsymbol{\mu} \\
&\Rightarrow \quad \phi_2(\boldsymbol{\mu}) = \langle \boldsymbol{\mu}, \boldsymbol{\theta} \rangle - \psi_2(\boldsymbol{\theta}) = \{\langle \boldsymbol{\mu}, \boldsymbol{\theta} \rangle - \psi_1(\boldsymbol{\theta})\} - c = \phi_1(\boldsymbol{\mu}) - c \\
&\Rightarrow \quad D_{\phi_2}(\mathbf{x}, \boldsymbol{\mu}) = D_{\phi_1}(\mathbf{x}, \boldsymbol{\mu}) \text{ (as linear terms do not change Bregman divergences)}
\end{aligned}$$

Now, consider any Bregman divergence $D_\phi(\cdot, \boldsymbol{\mu})$ on S. There exists at least one strictly convex, differentiable function ϕ on S that generates this divergence. The Legendre conjugates of (ϕ, S), i.e. (ψ, Γ) are well-defined. Hence, there exists an exponential density $p_{(\psi,\boldsymbol{\theta})}$ that is related to $D_\phi(\mathbf{x}.\boldsymbol{\mu})$ by (11), i.e., the mapping is onto. That completes the proof. ∎

4.4 Examples

We now look at two common exponential families and obtain the corresponding Bregman divergences using the bijection theorem stated above.

Example 5: The most well-known exponential family is that of Gaussian distributions, in particular uniform variance, spherical Gaussian distributions with densities of the form

$$p(\mathbf{x}; \mathbf{a}) = \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp(-\frac{1}{2\sigma^2}\|\mathbf{x} - \mathbf{a}\|^2).$$

where $\mathbf{x}, \mathbf{a} \in \mathbb{R}^d$ and $\sigma \in \mathbb{R}$ is a constant. It is easy to see that the density can be expressed in the canonical form for exponential families with natural parameter, $\boldsymbol{\theta} = \frac{\mathbf{a}}{\sigma^2}$ and cumulant function, $\psi(\boldsymbol{\theta}) = \frac{\sigma^2}{2}\|\boldsymbol{\theta}\|^2$

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

where $e^{-\lambda(\mathbf{x})}$ is independent of $\boldsymbol{\theta}$. The expectation parameter for this distribution is given by

$$\boldsymbol{\mu} = \nabla\psi(\boldsymbol{\theta}) = \nabla(\frac{\sigma^2}{2}\|\boldsymbol{\theta}\|^2) = \sigma^2\boldsymbol{\theta} \text{ (using (10))}$$

and the Legendre conjugate function, ϕ is obtained as

$$\phi(\boldsymbol{\mu}) = \langle \boldsymbol{\mu}, \boldsymbol{\theta}\rangle - \psi(\boldsymbol{\theta}) = \langle \boldsymbol{\mu}, \frac{\boldsymbol{\mu}}{\sigma^2}\rangle - \frac{\sigma^2}{2}\|\boldsymbol{\theta}\|^2 = \frac{\|\boldsymbol{\mu}\|^2}{2\sigma^2}, \text{ (using (9))}$$

a constant multiple of the squared Euclidean norm. From Example 1, we know that the corresponding Bregman divergence, D_ϕ will be given by a similar multiple of the squared Euclidean distance.

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

The function $f_\phi(\mathbf{x})$, mentioned in the bijection theorem, turns out to be constant and is given by

$$\begin{aligned} f_\phi(\mathbf{x}) &= \exp(\phi(\mathbf{x}) - \lambda(\mathbf{x})) \\ &= \exp(\frac{\|\mathbf{x}\|^2}{2\sigma^2} - \frac{\|\mathbf{x}\|^2}{2\sigma^2})\frac{1}{\sqrt{(2\pi\sigma^2)^d}} \\ &= \frac{1}{\sqrt{(2\pi\sigma^2)^d}}. \end{aligned}$$

■

Example 6: Another exponential family that has been widely used to model text data is the family of multinomial distributions with densities of the form,

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

where frequencies of events, $x_j \in \mathbb{Z}_+$, $\sum_{j=1}^d x_j = N$ and probabilities of events, $q_j \geq 0, \sum_{j=1}^d q_j = 1$. This can be expressed as the density of an exponential distribution in $\mathbf{x} = \{x_j\}_{j=1}^{d-1}$ with natural parameter, $\boldsymbol{\theta} = \{\log(\frac{q_j}{q_d})\}_{j=1}^{d-1}$ and cumulant function, $\psi(\boldsymbol{\theta}) = -N \log q_d = N \log(1 + \sum_{j=1}^{d-1} e^{\theta_j})$.

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

where N is a constant and $\vartheta(\mathbf{x}) = e^{-\lambda(x)}$ is independent of $\boldsymbol{\theta}$. The expectation parameter $\boldsymbol{\mu}$ is given by

$$\boldsymbol{\mu} = \nabla \psi(\boldsymbol{\theta}) = \nabla(N \log(1 + \sum_{j=1}^{d-1} e^{\theta_j})) = \left[\frac{N e^{\theta_j}}{(1 + \sum_{j=1}^{d-1} e^{\theta_j})} \right]_{j=1}^{d-1} = [N q_j]_{j=1}^{d-1}$$

and the Legendre conjugate function, ϕ is obtained as

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

where $\mu_d = N q_d$ so that $\sum_{i=1}^d \mu_j = N$. This is a constant multiple of negative entropy for the discrete probability distribution given by $\{\frac{\mu_j}{N}\}_{j=1}^d$. From Example 2, we know that the corresponding Bregman divergence will be a similar multiple of KL-divergence.

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

Table 2: Various functions of interest for some popular exponential distributions

Distribution	$p(\mathbf{x};\boldsymbol{\theta})$	$\boldsymbol{\theta}$	$\psi(\boldsymbol{\theta})$
1-D Gaussian[1]	$\frac{1}{\sqrt{(2\pi\sigma^2)}}\exp(-\frac{(x-a)^2}{2\sigma^2})$	$\frac{a}{\sigma^2}$	$\frac{\sigma^2}{2}\theta^2$
1-D Poisson	$\frac{\lambda^x e^{-\lambda}}{x!}$	$log\lambda$	e^θ
1-D Bernoulli	$q^x(1-q)^{1-x}$	$\log(\frac{q}{1-q})$	$\log(1+e^\theta)$
1-D Binomial[1]	$\frac{N!}{(x)!(N-x)!}q^x(1-q)^{N-x}$	$\log(\frac{q}{1-q})$	$N\log(1+e^\theta)$
1-D Geometric	$\lambda\exp(-\lambda x)$	$-\lambda$	$-\log(-\theta)$
d-D Sph. Gaussian[1]	$\frac{1}{\sqrt{(2\pi\sigma^2)^d}}\exp(-\frac{\|\mathbf{x}-\mathbf{a}\|^2}{2\sigma^2})$	$\frac{\mathbf{a}}{\sigma^2}$	$\frac{\sigma^2}{2}\|\boldsymbol{\theta}\|^2$
d-D Multinomial[1]	$\frac{N!}{\prod_{j=1}^d (x_j)!}\prod_{j=1}^d (q_j)^{x_j}$	$[\log(\frac{q_j}{q_d})]_{j=1}^{d-1}$	$N\log(1+\sum_{j=1}^{d-1} e^{\theta_j})$

Table 3: Various functions of interest for some popular exponential distributions (cont.)

Distribution	$p(\mathbf{x};\boldsymbol{\theta})$	$\boldsymbol{\mu}$	$\phi(\boldsymbol{\mu})$	$D_\phi(\mathbf{x},\boldsymbol{\mu})$
1-D Gaussian	$\frac{1}{\sqrt{(2\pi\sigma^2)}}\exp(-\frac{(x-a)^2}{2\sigma^2})$	a	$\frac{1}{2\sigma^2}\mu^2$	$\frac{1}{2\sigma^2}(x-\mu)^2$
1-D Poisson	$\frac{\lambda^x e^{-\lambda}}{x!}$	λ	$\mu\log\mu-\mu$	$x\log(\frac{x}{\mu})-(x-\mu)$
1-D Bernoulli	$q^x(1-q)^{1-x}$	q	$\mu\log\mu+(1-\mu)log(1-\mu)$	$x\log(\frac{x}{\mu})+(1-x)\log(\frac{1-x}{1-\mu})$
1-D Binomial	$\frac{N!}{(x)!(N-x)!}q^x(1-q)^{N-x}$	Nq	$\mu\log(\frac{\mu}{N})+(N-\mu)log(\frac{N-\mu}{N})$	$x\log(\frac{x}{\mu})+(N-x)\log(\frac{N-x}{N-\mu})$
1-D Geometric	$\lambda\exp(-\lambda x)$	λ	$-\mu^2+\log\mu$	$log(\frac{x}{\mu})-\frac{x}{\mu}+1-(x-\mu)^2$
d-D Sph. Gaussian	$\frac{1}{\sqrt{(2\pi\sigma^2)^d}}\exp(-\frac{\|\mathbf{x}-\mathbf{a}\|^2}{2\sigma^2})$	$\mathbf{a}$	$\frac{1}{2\sigma^2}\|\boldsymbol{\mu}\|^2$	$\frac{1}{2\sigma^2}\|\mathbf{x}-\boldsymbol{\mu}\|^2$
d-D Multinomial	$\frac{N!}{\prod_{j=1}^d x_j!}\prod_{j=1}^d q_j^{x_j}$	$[Nq_j]_{j=1}^{d-1}$	$\sum_{j=1}^d \mu_j\log(\frac{\mu_j}{N})$	$\sum_{j=1}^d x_j\log(\frac{x_j}{\mu_j})$

The function $f_\phi(\mathbf{x})$ for this case is given by

$$\begin{aligned} f_\phi(\mathbf{x}) &= \exp(\phi(\mathbf{x})-\lambda(\mathbf{x})) \\ &= \exp(\sum_{j=1}^d x_j\log(\frac{x_j}{N}))\frac{N!}{\prod_{j=1}^d x_j!} \\ &= \frac{\prod_{j=1}^d x_j^{x_j}}{N^N}\frac{N!}{\prod_{j=1}^d x_j!}. \end{aligned}$$

■

Tables 2 and 3 shows the various functions of interest for some popular exponential distribution families. For all the cases shown in the table, $\mathbf{x}$ is itself the sufficient statistic.

5 Bregman Soft Clustering

Using the bijection between exponential families and Bregman divergences, we first pose the Bregman soft clustering problem as a parameter estimation problem for mixture models based on exponential distributions . Then, we revisit the Expectation-Maximization (EM) framework for estimating mixture densities and develop the Bregman soft clustering algorithm (Algorithm 3). We also present the Bregman soft clustering algorithm for a set with a probability measure and

[1]The variance σ and the number of trials N are assumed to be constant for the distributions.

show how the hard clustering algorithm can be interpreted as a special case of the soft clustering algorithm.

5.1 Soft Clustering as Mixture Density Estimation

Given a set $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$ drawn independently from a stochastic source, consider the problem of modeling the source using a single parametric exponential distribution. This is the problem of maximum likelihood estimation, or, equivalently, minimum negative log-likelihood estimation of the parameter(s) of the parametric density belonging to a given exponential family. Now, from Theorem 3, minimizing the negative log-likelihood is the same as minimizing the expected Bregman divergence. Using Theorem 1, we conclude that the optimal distribution is the one with $\boldsymbol{\mu} = \mathbf{E}[\mathbf{x}]$ as the expectation parameter where the expectation is over the empirical distribution. Further, note that the minimum negative log-likelihood is the Bregman information of the set, $I_\phi(\mathcal{X})$, up to additive constants.

Now, consider the problem of modeling the stochastic source with a mixture of k densities of the same exponential family. This also yields a soft clustering where clusters correspond to the components of the mixture model, and the soft membership of a data point in each cluster is proportional to the probability of the data point being generated by the corresponding density function. Thus the **Bregman soft clustering problem** can be stated to be that of learning the maximum likelihood parameters $\Theta = \{\boldsymbol{\mu}_h, \pi_h\}_{h=1}^k$ of a mixture model of the form

$$p(\mathbf{x}|\Theta) = \sum_{h=1}^{k} \pi_h p_h(\mathbf{x}|\boldsymbol{\theta}_h) = \sum_{h=1}^{k} \pi_h f_\phi(\mathbf{x}) \exp(-D_\phi(\mathbf{x}, \boldsymbol{\mu}_h)). \tag{12}$$

The above problem is a special case of the general maximum likelihood parameter estimation problem for mixture models. So we first revisit the general problem and its solution using the EM framework. Later, we use this to develop the Bregman soft clustering algorithm for the special case in which we are interested. Note that, by the bijection between Bregman divergences and exponential families, (12) encompasses the soft clustering problem for *all* exponential families.

5.2 EM for Mixture Models based on Bregman Divergences

The maximum likelihood parameter estimation problem for a mixture model can be stated formally as follows. Let $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$ and $p(\mathbf{x}|\Theta)$ be a mixture density given by

$$p(\mathbf{x}|\Theta) = \sum_{h=1}^{k} \pi_h p_h(\mathbf{x}|\boldsymbol{\theta}_h),$$

where $p_h(\mathbf{x}|\boldsymbol{\theta}_h)$ are the individual component densities, π_h their priors and $\Theta = \{\boldsymbol{\theta}_h, \pi_h\}_{h=1}^k$. Then, the likelihood of observing $\mathcal{X}$ given the mixture model is obtained as

$$L_{\mathcal{X}}(\Theta) = \prod_{i=1}^{n} \left(\sum_{h=1}^{k} \pi_h p_h(\mathbf{x}_i|\boldsymbol{\theta}_h) \right).$$

Estimating the mixture densities for the dataset is equivalent to solving the optimization problem,

$$\max_{\Theta} L_{\mathcal{X}}(\Theta),$$

where $\Theta = \{\boldsymbol{\theta}_h, \pi_h\}_{h=1}^k$. The Expectation-Maximization (EM) framework provides a nice solution

Algorithm 2 EM for Mixture Density Estimation

Input: Set $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n \subset S \subseteq \mathbb{R}^d$,number of clusters k.

Output: Θ^*, local maximizer of $L_{\mathcal{X}}(\Theta) = \prod_{i=1}^n (\sum_{h=1}^k \pi_h p_h(\mathbf{x}_i|\boldsymbol{\theta}_h))$ where $\Theta = \{\boldsymbol{\theta}_h, \pi_h\}_{h=1}^k$, soft partitioning $\{\{p(h|\mathbf{x}_i)\}_{h=1}^k\}_{i=1}^n$.

Method:

Initialize $\{\boldsymbol{\theta}_h, \pi_h\}_{h=1}^k$ with some $\boldsymbol{\theta}_h \in S,\ \pi_h \geq 0,\ \sum_{h=1}^k \pi_h = 1$

repeat

{The Expectation Step}

for $i = 1$ to n **do**

for $h = 1$ to k **do**

$p(h|\mathbf{x}_i) \leftarrow \frac{\pi_h p_h(\mathbf{x}_i|\boldsymbol{\theta}_h)}{\sum_{h'=1}^k \pi_{h'} p_{h'}(\mathbf{x}_i|\boldsymbol{\theta}_{h'})}$

end for

end for

{The Maximization Step}

for $h = 1$ to k **do**

$\pi_h \leftarrow \frac{1}{n} \sum_{i=1}^n p(h|\mathbf{x}_i)$

$\boldsymbol{\theta}_h \leftarrow \operatorname{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^n \log(p_h(\mathbf{x}_i|\boldsymbol{\theta}))p(h|\mathbf{x}_i)$

end for

until *convergence*

return $\Theta^* = \{\boldsymbol{\theta}_h, \pi_h\}_{h=1}^k$

to the parameter estimation problem stated above. The resulting algorithm is an iterative procedure for obtaining the maximum likelihood estimator of the parameters using the conditional expectation for the missing data, which in this case is the posterior probabilities of the clusters for each data point. The algorithm is known to have the following convergence property.

Proposition 3 *The EM algorithm (Algorithm 2) has the property that the likelihood of the data, $L_{\mathcal{X}}(\Theta)$ is non-decreasing at each iteration. Further, if there exists at least one local maximum for the likelihood function, then the algorithm will converge to a local maximum of the likelihood.*

For a detailed proof and other related results, please see [Col97] and [Bil97].

As stated earlier, the Bregman soft clustering problem is to estimate the maximum likelihood parameters for a mixture model of the form,

$$p(\mathbf{x}|\Theta) = \sum_{h=1}^{k} \pi_h f_\phi(\mathbf{x}) \exp(-D_\phi(\mathbf{x}, \boldsymbol{\mu}_h)).$$

Applying the EM algorithm to this problem gives us locally optimal parameters Θ^* for this mixture model. The resulting mixture model also provides a soft clustering of the dataset based on the Bregman divergence D_ϕ. Hence, we call this application of the EM algorithm, the Bregman soft clustering algorithm. The Bregman divergence viewpoint also helps in significantly simplifying the algorithm, especially the computationally intensive M-step. The resulting update equations are very similar to those for learning mixture models of uniform variance, spherical Gaussians. The following propositions prove the correctness of the Bregman soft clustering algorithm.

Proposition 4 *For a mixture model with density given by*

$$p(x|\Theta) = \sum_{h=1}^{k} \pi_h f_\phi(\mathbf{x}) \exp(-D_\phi(\mathbf{x}, \boldsymbol{\mu}_h)),$$

the maximization step for the density parameters in the EM algorithm (Algorithm 2) reduces to a simple expectation step:

$$\forall h, 1 \leq h \leq k, \quad \boldsymbol{\mu}_h = \frac{\sum_{i=1}^{n} p(h|\mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^{n} p(h|\mathbf{x}_i)}. \tag{13}$$

Proof: The maximization step for the density parameters in the EM algorithm is given by

$$\forall h, 1 \leq h \leq k, \ \boldsymbol{\theta}_h = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log(p_h(\mathbf{x}_i|\boldsymbol{\theta}))p(h|\mathbf{x}_i).$$

For the given mixture density, the component densities are given by

$$\forall h, 1 \leq h \leq k, \ p_h(\mathbf{x}|\boldsymbol{\theta}_h) = f_\phi(\mathbf{x}) \exp(-D_\phi(\mathbf{x}, \boldsymbol{\mu}_h)).$$

Substituting the above into the maximization step, we obtain the update equations for the expectation parameters $\boldsymbol{\mu}_h$: $\forall h, 1 \leq h \leq k$,

$$\begin{aligned}
\boldsymbol{\mu}_h &= \operatorname*{argmax}_{\boldsymbol{\mu}} \sum_{i=1}^{n} \log(f_\phi(\mathbf{x}_i) \exp(-D_\phi(\mathbf{x}_i, \boldsymbol{\mu})))p(h|\mathbf{x}_i) \\
&= \operatorname*{argmax}_{\boldsymbol{\mu}} \sum_{i=1}^{n} (\log(f_\phi(\mathbf{x}_i)) - D_\phi(\mathbf{x}_i, \boldsymbol{\mu}))p(h|\mathbf{x}_i) \\
&= \operatorname*{argmin}_{\boldsymbol{\mu}} \sum_{i=1}^{n} D_\phi(\mathbf{x}_i, \boldsymbol{\mu})p(h|\mathbf{x}_i) \qquad (\text{as } f_\phi(\mathbf{x}) \text{ is independent of } \boldsymbol{\mu}_h) \\
&= \operatorname*{argmin}_{\boldsymbol{\mu}} \sum_{i=1}^{n} D_\phi(\mathbf{x}_i, \boldsymbol{\mu}) \frac{p(h|\mathbf{x}_i)}{\sum_{i'=1}^{n} p(h|\mathbf{x}_{i'})},
\end{aligned}$$

so that the weights on the divergences form a valid probability measure (i.e. sum to 1). From Theorem 1, we know that the expected Bregman divergence is minimized by the expectation of $\mathbf{x}$,

$$\operatorname*{argmin}_{\boldsymbol{\mu}} \sum_{i=1}^{n} D_\phi(\mathbf{x}, \boldsymbol{\mu}))p(h|\mathbf{x}_i) = \frac{\sum_{i=1}^{n} p(h|\mathbf{x}_i) \ \mathbf{x}_i}{\sum_{i=1}^{n} p(h|\mathbf{x}_i)}.$$

Therefore, the update equation for the parameters is a weighted averaging step,

$$\forall h, 1 \leq h \leq k, \ \boldsymbol{\mu}_h = \frac{\sum_{i=1}^{n} p(h|\mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^{n} p(h|\mathbf{x}_i)}.$$

■

Proposition 5 *For a mixture model with density given by*

$$p(x|\Theta) = \sum_{h=1}^{k} \pi_h f_\phi(\mathbf{x}) \exp(-D_\phi(\mathbf{x}, \boldsymbol{\mu}_h)),$$

the EM algorithm (Algorithm 2) reduces to the Bregman soft clustering algorithm (Algorithm 3).

Algorithm 3 Bregman Soft Clustering

Input: Set $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n \subset S \subseteq \mathbb{R}^d$, Bregman divergence $D_\phi : \mathcal{S} \times \text{int}(S) \mapsto \mathbb{R}$, number of clusters k.

Output: Θ^*, local maximizer of $\prod_{i=1}^n (\sum_{h=1}^k \pi_h f_\phi(\mathbf{x}_i) \exp(-D_\phi(\mathbf{x}_i, \boldsymbol{\mu}_h)))$ where $\Theta = \{\boldsymbol{\mu}_h, \pi_h\}_{h=1}^k$, soft partitioning $\{\{p(h|\mathbf{x}_i)\}_{h=1}^k\}_{i=1}^n$

Method:

Initialize $\{\boldsymbol{\mu}_h, \pi_h\}_{h=1}^k$ with some $\boldsymbol{\mu}_h \in S, \pi_h \geq 0$, and $\sum_{h=}^k \pi_h = 1$

repeat

{The Expectation Step}

for $i = 1$ to n **do**

for $h = 1$ to k **do**

$p(h|\mathbf{x}_i) \leftarrow \frac{\pi_h \exp(-D_\phi(\mathbf{x}_i, \boldsymbol{\mu}_h))}{\sum_{h'=1}^k \pi_{h'} \exp(-D_\phi(\mathbf{x}_i, \boldsymbol{\mu}_{h'}))}$

end for

end for

{The Maximization Step}

for $h = 1$ to k **do**

$\pi_h \leftarrow \frac{1}{n} \sum_{i=1}^n p(h|\mathbf{x}_i)$

$\boldsymbol{\mu}_h \leftarrow \frac{\sum_{i=1}^n p(h|\mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^n p(h|\mathbf{x}_i)}$

end for

until *convergence*

return $\Theta^* = \{\boldsymbol{\mu}_h, \pi_h\}_{h=1}^k$

Proof: For the given mixture model, the component densities are given by

$$\forall h, 1 \leq h \leq k, \ p_h(\mathbf{x}|\boldsymbol{\theta}_h) = f_\phi(\mathbf{x}) \exp(-D_\phi(\mathbf{x}, \boldsymbol{\mu}_h)).$$

The update equations for the posterior probabilities in the EM algorithm are given by

$$\forall \mathbf{x} \in \mathcal{X}, \ \forall h, 1 \leq h \leq k, \ p(h|\mathbf{x}) = \frac{\pi_h p_h(\mathbf{x}|\boldsymbol{\theta}_h)}{\sum_{h'=1}^k \pi_{h'} p_{h'}(\mathbf{x}|\boldsymbol{\theta}_{h'})} = \frac{\pi_h \exp(-D_\phi(\mathbf{x}, \boldsymbol{\mu}_h))}{\sum_{h'=1}^k \pi_{h'} \exp(-D_\phi(\mathbf{x}, \boldsymbol{\mu}_{h'}))}$$

as the $f_\phi(\mathbf{x})$ factor cancels out. Further from Proposition 4, the parameter estimation step in the EM algorithm reduces to a simple expectation step,

$$\forall h, 1 \leq h \leq k, \ \boldsymbol{\mu}_h = \frac{\sum_{i=1}^n p(h|\mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^n p(h|\mathbf{x}_i)}.$$

The prior update equations are independent of the parametric form of the densities and remain unaltered:

$$\forall h, 1 \leq h \leq k, \ \pi_h = \frac{1}{n} \sum_{i=1}^n p(h|\mathbf{x}_i).$$

■

5.3 Bregman Soft Clustering for a Set with Probability measure

In section 5.2, we addressed the Bregman soft clustering problem for a set $\mathcal{X}$ with a uniform measure, i.e., all elements of $\mathcal{X}$ have the same weight. Now we look at the soft clustering problem

for a set with non-uniform probability measure. As in the case of the hard clustering problem, the objective is to assign the elements of the set $\mathcal{X}$ to different clusters, the only difference being that now a single element can have a non-zero probability of belonging to multiple clusters. To take the non-uniform measure into account, we consider a new set $\mathcal{X}_N$ of large size N such that every element $\mathbf{x}_i \in \mathcal{X}$ occurs $\nu_i N$ times in the set $\mathcal{X}_N$ and pose the Bregman soft clustering problem for the set $\mathcal{X}$ with a non-uniform measure ν as that for the set $\mathcal{X}_N$ with uniform measure. From the previous subsection, this is equivalent to the maximum likelihood estimation problem,

$$\max_{\Theta} L_{\mathcal{X}_N}(\Theta),$$

where $L_{\mathcal{X}_N}(\Theta)$ is the likelihood of observing the set $\mathcal{X}_N$ given a mixture density $p(\mathbf{x}|\Theta)$ based on exponential distributions with parameters Θ, and can be easily solved using the EM framework. Note that the set $\mathcal{X}_N$ has multiple instances of elements of $\mathcal{X}$ and could have an extremely low probability of being generated from a mixture model based on exponential distributions, but we are only interested in learning the parameters of the mixture model that has the highest probability of generating $\mathcal{X}_N$. The resulting algorithm is similar to the EM algorithm applied directly to $\mathcal{X}$ with the M-step modified to include the probability measure ν. The new M-step update equations are given by

$$\forall h, 1 \leq h \leq k, \quad \pi_h = \sum_{i=1}^{n} \nu_i p(h|\mathbf{x}_i), \tag{14}$$

$$\text{and} \quad \forall h, 1 \leq h \leq k, \quad \boldsymbol{\theta}_h = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log(p_h(\mathbf{x}_i|\boldsymbol{\theta}))\nu_i p(h|\mathbf{x}_i). \tag{15}$$

$$\tag{16}$$

When the mixture models are based on exponential distributions, as in the case of the Bregman soft clustering problem, (16) remains identical while (16) reduces to

$$\forall h, 1 \leq h \leq k, \quad \boldsymbol{\mu}_h = \frac{\sum_{i=1}^{n} \nu_i p(h|\mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^{n} \nu_i p(h|\mathbf{x}_i)}.$$

Hence, the Bregman soft clustering algorithm for a set $\mathcal{X}$ with probability measure ν is given by Algorithm 3 with the maximization steps replaced by the above update equations. The expectation step remains unchanged.

Finally, we note that the Bregman hard clustering algorithm is a limiting case of the above soft clustering algorithm. For every convex function ϕ and positive constant β, $\beta\phi$ is also a convex function with the corresponding Bregman divergence $D_{\beta\phi} = \beta D_\phi$ (see Property 3 in the Appendix). In the limit, when $\beta \to \infty$, both the E and M steps of the soft clustering algorithm reduce to the assignment and re-estimation step of the hard clustering algorithm. Further, this view suggests the possibility of designing annealing schemes for Bregman soft clustering interpreting $1/\beta$ as the temperature parameter.

6 Experiments

In this section, we present the results of applying Bregman clustering to datasets based on different exponential distributions and show that the clustering quality depends on the choice of the Bregman divergence. For our first experiment, we created three 1-dimensional datasets of 100 samples each, based on mixture models of Gaussian, Poisson and Binomial distributions respectively. All the