Clustering with Bregman Divergences

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

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 propose and analyze parametric hard and soft clustering algorithms based on a large class of distortion functions known as Bregman divergences. The proposed algorithms unify centroid-based parametric clustering approaches, such as classical `kmeans` and information-theoretic clustering, which arise by special choices of the Bregman divergence. The algorithms maintain the simplicity and scalability of the classical `kmeans` algorithm, while generalizing the basic idea to a very large class of clustering loss functions. There are two main contributions in this paper. First, 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. Secondly, we show an explicit bijection between Bregman divergences and exponential families. The bijection enables the development of an alternative interpretation of an efficient EM scheme for learning models involving mixtures of exponential distributions. This leads to a simple soft clustering algorithm for all Bregman divergences.

1 Introduction

Data clustering is a fundamental "unsupervised" learning procedure that has been extensively studied across varied disciplines over several decades [14]. It has produced several parametric clustering methods which partition 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. For *hard* clustering, wherein the partitions are disjoint, the most well-known and widely used algorithm of this type is the iterative relocation scheme of Euclidean `kmeans` [14]. The popularity of this algorithm stems from its simplicity and scalability. The corresponding *soft* [1] clustering algorithm obtained by applying EM [9] to a mixture model of Gaussians with identical, isotropic covariances, is also popular and can be scaled to large data sets [6].

Underlying both hard and soft Euclidean `kmeans` is a Gaussian "noise" model, which corresponds to a squared-Euclidean distortion function [15]. This distortion function is also implicit in several other scalable techniques in the data mining literature. However, in many data mining applications, this distortion function is not a good match with the data, and consequently `kmeans` performs poorly as compared to other approaches [25]. In fact, in such situations `kmeans` often becomes a convenient strawman to show the superiority of a competing technique! This has also led to the search for more appropriate distance functions for specific applications [1, 25].

Is it possible to devise an algorithm which has the simplicity and scalability of `kmeans` but can cater to a much larger class of distortion functions? A hint towards an affirmative answer to this question is provided by the Linde-Buzo-Gray (LBG) algorithm [17] based on the Itakura-Saito distance, which has been used in the signal-processing community for clustering speech data. The more recent information theoretic clustering algorithm [10] for clustering probability distributions also has a flavor similar to `kmeans`. This algorithm uses the KL-divergence as the distortion function and is well-suited for various clustering tasks in the analysis of high-dimensional text data.

Our question can now be posed as: *what class of distortion functions admit an iterative relocation scheme where a global objective function based on the distortion with cluster centroids is progressively decreased?* In this paper, we give a precise answer to this question: we show that *such a scheme works for arbitrary Bregman divergences.* In fact, it can be shown [4] that such a scheme *only* works for Bregman divergences. The scope of this result is vast since Bregman divergences include a large number of useful loss functions such as square loss, KL divergence, logistic loss, Mahalanobis distance, Itakura-Saito distance, hinge loss, etc.

We pose the hard clustering problem as one of obtaining an optimal quantization in terms of minimizing the loss in *Bregman information*, a quantity motivated by rate-distortion theory. A simple analysis then yields a version of the loss function that readily suggests a natural algorithm to solve the clustering problem for arbitrary Bregman divergences. Partitional hard clustering to minimize the loss in *mutual information*, a topic of recent study [10], is seen to be a special case of our approach. Thus, this paper unifies several parametric

*Dept. of ECE, University of Texas at Austin, TX, USA.

†Dept. of CS, University of Texas at Austin, TX, USA

[1] In soft clustering, data points can have non-zero probabilities of belonging to multiple partitions.

partitional clustering approaches.

Further, we present a fundamental theoretical result by showing that there exists a *bijection between Bregman divergences and exponential families.* Since generative model-based soft clustering algorithms typically use mixtures of exponential distributions to model data, we revisit EM for mixture model estimation for this class of problems. We show that, with proper representation, the bijection gives an alternative interpretation of a well known efficient EM scheme [22] applicable in this case. The scheme simplifies the computationally intensive maximization-step of the EM algorithm, resulting in a general soft-clustering algorithm for all members of the exponential family, e.g., Poisson, Bernoulli, Binomial and Multinomial models. Both hard and soft clustering versions have essentially the same scalability as `kmeans`. Moreover they can be readily adapted to mixed data types, where different distortion functions within the family of Bregman divergences are appropriate for different subsets of features. This makes our theory and techniques suitable for a much wider class of data mining applications.

The remainder of this article is organized as follows. 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 2. In section 3, we establish a connection between exponential families and Bregman divergences and use it develop a soft Bregman clustering algorithm in section 4. In section 5, we present some experimental results that illustrate the usefulness of the Bregman clustering algorithm. In section 6, we discuss related work. Finally, in section 7, we present concluding remarks.

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. $\mathbb{R}, \mathbb{R}_{++}$ and $\mathbb{R}^d$ denote the set of reals, the set of positive reals and the d-dimensional real vector space respectively. $\|\mathbf{x}\|$ denotes 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 X is distributed as p, we denote this by $X \sim p$. Expectation of functions of a random variable $X \sim p$ are denoted by $E_p[\cdot]$ when the random variable is clear from the context. The inverse of a function f is denoted by f^{-1}.

2 Bregman Hard Clustering

In this section, we introduce a new concept called the Bregman information of a random variable 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 its equivalence 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.

We begin by defining Bregman divergence [21]. 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 $\text{int}(S)$, the interior of S [23]. The **Bregman divergence** $D_\phi : S \times \text{int}(S) \mapsto [0, \infty)$ is defined as $D_\phi(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}, \nabla\phi(\mathbf{y})\rangle$, where $\nabla\phi$ is the gradient of ϕ. Table 1 contains a list of some convex functions and their corresponding Bregman divergences. Bregman divergences have several interesting and useful properties, such as non-negativity, convexity in the first argument, etc. For details see [3] and [5].

2.1 Bregman Information

The dual formulation of Shannon's celebrated rate distortion problem [13] involves finding a coding scheme with a given rate, such that the expected distortion between the source random variable and the decoded random variable is minimized. The achieved distortion is called the *distortion-rate function*, i.e., infimum distortion achievable for a given rate. Now consider a simple coding scheme for a random variable X that takes values in a finite set $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n \subset S \subseteq \mathbb{R}^d$ (S is convex), following a discrete probability measure ν and the distortion function is a Bregman divergence D_ϕ. The encoding scheme involves representing the random variable by a constant vector $\mathbf{s}$, i.e., codebook size is one, or rate is zero. The solution to the rate-distortion problem in this case is the trivial assignment. The corresponding distortion-rate function is given by $\mathbf{E}_\nu[D_\phi(X, \mathbf{s})]$ that depends on the choice of the representative $\mathbf{s}$ and can be further optimized by picking the right representative. We call this optimal distortion-rate function, i.e.,

$$\min_{\mathbf{s}\in S} \mathbf{E}_\nu[D_\phi(X, \mathbf{s})] = \min_{\mathbf{s}\in S} \sum_{i=1}^{n} \nu_i \; D_\phi(\mathbf{x}_i, \mathbf{s}), \tag{2.1}$$

the **Bregman information** of the random variable X for the Bregman divergence, D_ϕ and denote it by $I_\phi(X)$. The optimal $\mathbf{s}$ that achieves the minimal distortion will be called the *Bregman representative* or, simply the *representative* of X. The following theorem states that this representative always exists, is uniquely determined and, surprisingly, *does not depend* on the choice of the Bregman divergence.

Table 1: Bregman divergences corresponding to some convex functions.

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 [2]
$\mathbb{R}_{++}$	$-\log x$	$\frac{x}{y}-\log(\frac{x}{y})-1$	Itakura-Saito distance
$\mathbb{R}\setminus\{0\}$	$\lvert x\rvert$	$\max\{0,-2\,\mathrm{sign}(y)x\}$	Hinge loss
$\mathbb{R}^d$	$\lVert\mathbf{x}\rVert^2$	$\lVert\mathbf{x}-\mathbf{y}\rVert^2$	Squared Euclidean distance
$\mathbb{R}^d$	$\mathbf{x}^TA\mathbf{x}$	$(\mathbf{x}-\mathbf{y})^TA(\mathbf{x}-\mathbf{y})$	Mahalanobis distance [3]
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

Theorem 1 *Let X be a random variable taking values in $\mathcal{X}=\{\mathbf{x}_i\}_{i=1}^n\subset S\subseteq\mathbb{R}^d$ following ν. Given a Bregman divergence $D_\phi: S\times\mathrm{int}(S)\mapsto[0,\infty)$, the problem*

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

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

Proof. The function we are trying to minimize is $J_\phi(\mathbf{s})=E_\nu[D_\phi(X,\mathbf{s})]=\sum_{i=1}^n\nu_iD_\phi(\mathbf{x}_i,\mathbf{s})$. We prove the required result by showing that for $\forall\mathbf{s}\in S$,

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

with equality only when $\mathbf{s}=\boldsymbol{\mu}$ by the strict convexity of ϕ [3]. Hence, $\boldsymbol{\mu}$ is the unique minimizer of the function, J_ϕ. Now, we argue that $\boldsymbol{\mu}\in S$. Since $\mathcal{X}\subset S$ and S is a convex set, $\mathrm{co}(\mathcal{X})\subset S$, where $\mathrm{co}(\mathcal{X})$ is the convex hull of $\mathcal{X}$. But $\boldsymbol{\mu}=E_\nu[X]\in\mathrm{co}(\mathcal{X})$, so $\boldsymbol{\mu}\in S$. ∎

The above result shows that the representative, i.e., the minimizer of the expected Bregman divergence, is always the expectation of the set. Interestingly, the converse of theorem 1 is also true, i.e., for all random variables X, if $E_\nu[X]$ minimizes the expected distortion of the elements of the set to a fixed point for a distortion function $F(x,y)$, then, under mild conditions, it can be shown that $F(x,y)$ is a Bregman divergence [4]. Thus, Bregman divergences are exhaustive with respect to the property proved in theorem 1.

Using theorem 1, we can now give a more direct definition of the Bregman information as follows:

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

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

To start appreciating the potential of such a treatment, we note that the elements of $\mathcal{X}$ can be quite general. For instance, the elements can be probability distributions, functionals, operators or just plain vectors.

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

$$I_\phi(X)=\sum_{i=1}^n\nu_iD_\phi(\mathbf{x}_i,\boldsymbol{\mu})=\frac{1}{n}\sum_{i=1}^n\lVert\mathbf{x}_i-\boldsymbol{\mu}\rVert^2$$

which is just the sample variance. ∎

Example 2: 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 a random variable taking values in the

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

[3] It is the Mahalanobis distance when A is the inverse of the covariance matrix. In general, A is positive definite.

set of conditional distributions $\{p(V|\mathbf{u}_i)\}_{i=1}^n$ following $\{p(\mathbf{u}_i)\}_{i=1}^n$, 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 a random variable $Z_{\mathbf{u}}$ that takes values in the set of probability distributions $\mathcal{Z}_{\mathbf{u}} = \{p(V|\mathbf{u}_i)\}_{i=1}^n$ following the probability measure $\{\nu_i\}_{i=1}^n = \{p(\mathbf{u}_i)\}_{i=1}^n$ over this set. For $Z_{\mathbf{u}}$, the mean distribution is given by

$$\begin{aligned} \boldsymbol{\mu} = E_\nu[p(V|\mathbf{u})] &= \sum_{i=1}^{n} p(\mathbf{u}_i)p(V|\mathbf{u}_i) = p(V)\ . \\ \therefore \qquad I(U;V) &= \sum_{i=1}^{n} p(\mathbf{u}_i) KL(\ p(V|\mathbf{u}_i)\ \|\ p(V)\ ) \\ &= \sum_{i=1}^{n} \nu_i D_\phi(p(V|\mathbf{u}_i),\boldsymbol{\mu}) = I_\phi(Z_{\mathbf{u}}), \end{aligned}$$

i.e., *mutual information is a special case of Bregman information.* Further, for a random variable $Z_{\mathbf{v}}$ taking values in the set of probability distributions $\mathcal{Z}_{\mathbf{v}} = \{p(U|\mathbf{v}_j)\}_{j=1}^m$ following the probability measure $\nu_j = p(\mathbf{v}_j)$ over this set, one can similarly show that $I(U;V) = I_\phi(Z_{\mathbf{v}})$. The Bregman information of $Z_{\mathbf{v}}$ and $Z_{\mathbf{u}}$ can also be interpreted as the Jensen-Shannon divergence of the sets $\mathcal{Z}_{\mathbf{u}}$ and $\mathcal{Z}_{\mathbf{v}}$ [10]. ■

2.2 Clustering Formulation If X is a random variable with large Bregman information, it may not suffice to have a single representative for X if a low quantization error is desired. In such a situation, partitioning the set $\mathcal{X}$ into k relatively homogeneous groups, each with its own Bregman representative, such that the set $\mathcal{M}$ of these representatives along with the induced measure on $\mathcal{M}$ preserve Bregman information of X, seems a natural goal. The **Bregman hard clustering problem** is thus to find a partitioning of $\mathcal{X}$, or, equivalently, the set of representatives $\mathcal{M}$, such that if M is a random variable taking values in $\mathcal{M}$ following the induced measure for the corresponding partitions of $\mathcal{X}$, *the loss in Bregman information* due to quantization, $L_\phi(M) = I_\phi(X) - I_\phi(M)$, *is minimized.* This loss function can be re-written in a form that suggests a natural solution to this problem.

Theorem 2 *Let X be a random variable taking values in $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n \subset S \subseteq \mathbb{R}^d$ following ν. Let $\{\mathcal{X}_h\}_{h=1}^k$ be a partitioning of $\mathcal{X}$ and let $\pi_h = \sum_{\mathbf{x}_i \in \mathcal{X}_h} \nu_i$ be the induced measure π on the partitions. Let $\{X_h\}_{h=1}^k$ be random variables taking values in $\{\mathcal{X}_h\}_{h=1}^k$ following $\{\nu|\pi_h\}_{h=1}^k$ respectively. If $\mathcal{M} = \{\boldsymbol{\mu}_h\}_{h=1}^k$ denotes the set of representatives, and M be a random variable taking values in $\mathcal{M}$ following π, then*

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

Proof. Let $\boldsymbol{\mu} = \mathbf{E}_\nu[X]$. After some algebra [5], it can be shown that

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

Rearranging terms completes the proof. ■

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

$$\text{(2.2)} \qquad \min_{\mathcal{M}} L_\phi(M) = \min_{\mathcal{M}} \sum_{h=1}^{k}\sum_{\mathbf{x}_i \in \mathcal{X}_h} \nu_i D_\phi(\mathbf{x}_i,\boldsymbol{\mu}_h).$$

Thus, 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.

2.3 Clustering Algorithm Eq. 2.2 suggests a natural algorithm (Algorithm 1) to solve the Bregman hard clustering problem. It is easy to see that classical `kmeans`, the LBG algorithm [17] and the information theoretic clustering algorithm [10] are special cases of Bregman hard clustering for squared Euclidean distance, Itakura-Saito distance and KL-divergence respectively. For all these cases, the induced partitions are known to have linear separators. It is easy to 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 (2.2).*

Algorithm 1 Bregman hard-clustering

Input: Set $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n \subset S \subseteq \mathbb{R}^d$, probability measure ν over $\mathcal{X}$, Bregman divergence $D_\phi : S \times \text{int}(S) \mapsto \mathbb{R}$, num. of clusters k.
Output: $\mathcal{M}^* = \underset{\mathcal{M}}{\text{argmin}} \sum_{h=1}^k \sum_{\mathbf{x}_i \in \mathcal{X}_h} \nu_i D_\phi(\mathbf{x}_i, \boldsymbol{\mu}_h)$ where $\mathcal{M} = \{\boldsymbol{\mu}_h\}_{h=1}^k$, corresponding partitioning $\{\mathcal{X}_h\}_{h=1}^k$ of $\mathcal{X}$.
Method:
Initialize $\{\boldsymbol{\mu}_h\}_{h=1}^k$ at random with $\boldsymbol{\mu}_h \in S$
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 $h = h^*(\mathbf{x}_i) = \underset{h'}{\text{argmin}}\ D_\phi(\mathbf{x}_i, \boldsymbol{\mu}_{h'})$
  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*

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

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

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. ∎

3 Bijection with Exponential Families

We now turn our attention to soft clustering with Bregman divergences. To this end, we establish a bijection between Bregman divergences and exponential families in this section. We also list examples of Bregman divergences obtained from some popular exponential families. The bijection will be used to develop the Bregman soft clustering algorithm in section 4.

It has been observed in the literature [3, 11] that exponential families and Bregman divergences have certain relationships that can be exploited for several learning problems. We provide 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. To present the bijection result, we need to review the following background material.

3.1 Exponential families Consider a family $\mathcal{F}$ of probability densities on a measurable space $(\Omega, \mathcal{B})$ where $\mathcal{B}$ is a σ-algebra on the set Ω [12]. 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 2 *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 $\mathcal{F}$.*

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 3 *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 [2] that Γ is a convex set in $\mathbb{R}^d$ and ψ is a strictly convex and differentiable function on $\mathrm{int}(\Gamma)$.

3.2 Expectation parameters and Legendre duality Consider a d-dimensional real random variable X following an exponential family density $p_{(\psi,\boldsymbol{\theta})}$ specified by the natural parameter $\boldsymbol{\theta} \in \Gamma$. The expectation of X with respect to $p_{(\psi,\boldsymbol{\theta})}$, also called the **expectation parameter**, is given by

$$\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta}) = E_{p_{(\psi,\boldsymbol{\theta})}}[X] = \int_{\mathbb{R}^d} \mathbf{x} p_{(\psi,\boldsymbol{\theta})}(\mathbf{x}) d\mathbf{x}. \tag{3.3}$$

It can be shown [2] 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 [23]. 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 [23] 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 4 *[23] 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 (3.3). 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

$$\phi(\boldsymbol{\mu}) = \langle \boldsymbol{\theta}(\boldsymbol{\mu}), \boldsymbol{\mu}\rangle - \psi(\boldsymbol{\theta}(\boldsymbol{\mu})). \tag{3.4}$$

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,

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \nabla\psi(\boldsymbol{\theta}) \text{ and } \boldsymbol{\theta}(\boldsymbol{\mu}) = \nabla\phi(\boldsymbol{\mu}). \tag{3.5}$$

3.3 Bijection Theorem We are now ready to state the connection between exponential families of distributions and Bregman divergences. As mentioned earlier, connections between Bregman divergences and exponential families have been observed earlier in the literature. In particular, [11] showed that the negative log-likelihood of an exponential distribution can be written as the sum of a Bregman divergence and a function that does not depend on the parameters, i.e.,

$$-\log p(\mathbf{x}|\boldsymbol{\theta}) = D_\phi(\mathbf{x}, \boldsymbol{\mu}) + f(\mathbf{x}). \tag{3.6}$$

This result was used later on by [7] to extend PCA to exponential families. We make the relationship between Bregman divergences and exponential families exact by showing that there is actually a *bijection* between Bregman divergences and exponential families. More precisely, we show that for a given exponential family, D_ϕ and f in (3.6) are *uniquely determined* (one-one); further, there is an exponential family corresponding to every Bregman divergence (onto). Note that the duality between expectation and natural parameters mentioned in [11], [3], and [7] is basically the Legendre duality and the bijection result follows using properties of Legendre conjugates.

Theorem 3 *Let (ϕ, S) and (ψ, Γ) be Legendre conjugates of each other. Let $D_\phi : S \times \mathrm{int}(S) \mapsto \mathbb{R}$ 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,*

$$p_{(\psi,\boldsymbol{\theta})}(\mathbf{x}) = \exp(-D_\phi(\mathbf{x}, \boldsymbol{\mu})) f_\phi(\mathbf{x}), \tag{3.7}$$

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

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,

$$\begin{aligned}
p_{(\psi,\boldsymbol{\theta})}(\mathbf{x}) &= \exp(\langle \mathbf{x}, \boldsymbol{\theta} \rangle - \psi(\boldsymbol{\theta}) - \lambda(\mathbf{x})) \\
&= \exp(\langle \mathbf{x}, \nabla\phi(\boldsymbol{\mu}) \rangle + (\phi(\boldsymbol{\mu}) - \langle \boldsymbol{\mu}, \nabla\phi(\boldsymbol{\mu}) \rangle) \\
&\qquad - \lambda(\mathbf{x})) \quad \text{(using (3.4) and (3.5))} \\
&= \exp(-\{\phi(\mathbf{x}) - \phi(\boldsymbol{\mu}) - \langle (\mathbf{x} - \boldsymbol{\mu}), \nabla\phi(\boldsymbol{\mu}) \rangle\} \\
&\qquad + \{\phi(\mathbf{x}) - \lambda(\mathbf{x})\}) \\
&= \exp(-D_\phi(\mathbf{x}, \boldsymbol{\mu}))\ f_\phi(\mathbf{x}).
\end{aligned}$$

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, since linear additive terms to a convex function do not change the corresponding Bregman divergence [5]. This also implies that f_ϕ is a uniquely determined function on S.

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 $D_\phi(\cdot, \boldsymbol{\mu})$ by (3.7), i.e., the mapping is onto. That completes the proof. ■

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.

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

4.1 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 the bijection theorem (theorem 3), minimizing the negative log-likelihood is the same as minimizing the corresponding expected Bregman divergence. Using Theorem 1, we conclude that the optimal distribution is the one with $\boldsymbol{\mu} = \mathbf{E}[X]$ as the expectation parameter where the expectation is over the empirical distribution. Further, note that the minimum negative log-likelihood of X under a particular exponential model with log-partition function ψ is the Bregman information of X, i.e., $I_\phi(X)$, up to additive constants, where ϕ is the Legendre dual of ψ.

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

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[4]	$\frac{1}{\sqrt{(2\pi\sigma^2)}}\exp(-\frac{(x-a)^2}{2\sigma^2})$	$\frac{a}{\sigma^2}$	$\frac{\sigma^2}{2}\theta^2$
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[4]	$\frac{N!}{(x)!(N-x)!}q^x(1-q)^{N-x}$	$\log(\frac{q}{1-q})$	$N\log(1+e^\theta)$
1-D Exponential	$\lambda\exp(-\lambda x)$	$-\lambda$	$-\log(-\theta)$
d-D Sph. Gaussian[4]	$\frac{1}{\sqrt{(2\pi\sigma^2)^d}}\exp(-\frac{\lVert\mathbf{x}-\mathbf{a}\rVert^2}{2\sigma^2})$	$\frac{\mathbf{a}}{\sigma^2}$	$\frac{\sigma^2}{2}\lVert\boldsymbol{\theta}\rVert^2$
d-D Multinomial[4]	$\frac{N!}{\prod_{j=1}^d(x_j)!}\prod_{j=1}^d(q_j)^{x_j}$	$[\log(\frac{q_j}{q_d})]_{j=1}^{d-1}$	$N\log(1+\sum_{j=1}^{d-1}e^{\theta_j})$

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

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 Exponential	$\lambda\exp(-\lambda x)$	$1/\lambda$	$-\ln\mu-1$	$\frac{x}{\mu}-\ln\left(\frac{x}{\mu}\right)-1$
d-D Sph. Gaussian	$\frac{1}{\sqrt{(2\pi\sigma^2)^d}}\exp(-\frac{\lVert\mathbf{x}-\mathbf{a}\rVert^2}{2\sigma^2})$	$\mathbf{a}$	$\frac{1}{2\sigma^2}\lVert\boldsymbol{\mu}\rVert^2$	$\frac{1}{2\sigma^2}\lVert\mathbf{x}-\boldsymbol{\mu}\rVert^2$
d-D Multinomial	$\frac{N!}{\prod_{j=1}^d x_j!}\prod_{j=1}^d q_j^{x_j}$	$[Nq_j]_{j=1}^{d-1}$	$\sum_{j=1}^d \mu_j\log(\frac{\mu_j}{N})$	$\sum_{j=1}^d x_j\log(\frac{x_j}{\mu_j})$

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

$$p(\mathbf{x}|\Theta) = \sum_{h=1}^{k}\pi_h f_\phi(\mathbf{x})\exp(-D_\phi(\mathbf{x},\boldsymbol{\mu}_h)), \tag{4.8}$$

where the last equality follows from the bijection theorem (theorem 3). The above problem is a special case of the general maximum likelihood parameter estimation problem for mixture models and can be solved by applying the EM algorithm. Note that, by the bijection between Bregman divergences and exponential families, (4.8) encompasses the soft clustering problem for *all* exponential families.

4.2 EM for Mixture Models based on Bregman Divergences Algorithm 2 describes the well known application of EM for mixture density estimation. This algorithm 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 [18].

As stated earlier, the Bregman soft clustering problem is to estimate the maximum likelihood parameters for a mixture model of the form given in (4.8). 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 gives an alternative interpretation of a well known efficient EM scheme applicable to mixture of exponential distributions [22]. This significantly simplifies the algorithm, especially the computationally intensive M-step. The resulting update equations look very similar to those for learning mixture models of unit variance Gaussians. However, note that these equations are applicable to mixtures of any exponential distributions, and, as mentioned earlier, $\mathbf{x}$ denotes the sufficient statistic vector in the general case.

The following proposition shows how theorems 1 and 3 can be used to simplify the M-step of the clustering algorithm. Note that proposition 3 has appeared in various forms in the literature (see, for example, [22, 18]). We give an alternative proof using results involving Bregman divergences developed in the earlier sections.

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

Algorithm 2 EM for Mixture Density Estimation [18]

Input: Set $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n \subset S \subseteq \mathbb{R}^d$, num. 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**
$\pi_h \leftarrow \frac{1}{n}\sum_{i=1}^n p(h|\mathbf{x}_i)$
$\boldsymbol{\theta}_h \leftarrow \underset{\boldsymbol{\theta}}{\text{argmax}} \sum_{i=1}^n \log(p_h(\mathbf{x}_i|\boldsymbol{\theta}))p(h|\mathbf{x}_i)$
end for
until *convergence*
return $\Theta^* = \{\boldsymbol{\theta}_h, \pi_h\}_{h=1}^k$

Algorithm 3 Bregman Soft Clustering

Input: Set $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n \subset S \subseteq \mathbb{R}^d$, Bregman divergence D_ϕ, num. 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 $\Theta^* = \{\boldsymbol{\mu}_h, \pi_h\}_{h=1}^k$

Proposition 3 *For a mixture model with density given by (4.8), the maximization step for the density parameters in the EM algorithm (Algorithm 2), $\forall h, 1 \leq h \leq k$, reduces to:*

$$\boldsymbol{\mu}_h = \frac{\sum_{i=1}^n p(h|\mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^n p(h|\mathbf{x}_i)}. \tag{4.9}$$

Proof. The maximization step for the density parameters in the EM algorithm, $\forall h, 1 \leq h \leq k$, is given by

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

For the given mixture density, the component densities, $\forall h, 1 \leq h \leq k$, are given by

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

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

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

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 $\mathbf{x}$,

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

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

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

The update equations for the posterior probabilities (E-step) $\forall \mathbf{x} \in \mathcal{X}$, $\forall h, 1 \leq h \leq k$, are given by

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

as the $f_\phi(\mathbf{x})$ factor cancels out. The prior update equations are independent of the parametric form of the densities and remain unaltered. Hence, for a mixture model with density given by (4.8), the EM algorithm (Algorithm 2) reduces to the Bregman soft clustering algorithm (Algorithm 3).

So far, we considered the Bregman soft clustering problem for a set $\mathcal{X}$ where all the elements are equally important and assumed to have been independently sampled from some particular exponential distribution. In practice, it might be desirable to associate weights with the individual samples and optimize a weighted log-likelihood function. A slight modification to the M-step of the Bregman soft clustering algorithm is sufficient to address this new optimization problem. The E-step remains same and the new update equations for the M-step $\forall h, 1 \leq h \leq k$, are given by

$$\pi_h = \sum_{i=1}^{n} \nu_i p(h|\mathbf{x}_i), \tag{4.10}$$

$$\boldsymbol{\mu}_h = \frac{\sum_{i=1}^{n} \nu_i p(h|\mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^{n} \nu_i p(h|\mathbf{x}_i)}. \tag{4.11}$$

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 [5]). 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.

5 Experiments

There are a number of experimental results in existing literature [17, 10, 20, 16] that illustrate the usefulness of Bregman divergences and the Bregman clustering algorithms in specific domains. The classical `kmeans` algorithm, which is a special case of the Bregman hard clustering algorithm for the squared Euclidean distance has been successfully applied to a large number of domains where a Gaussian distribution assumption is valid. Besides this, there are at least two other domains where special cases of Bregman clustering methods have been shown to provide good results.

The first is the text-clustering domain where the information-theoretic clustering algorithm [10] and the EM algorithm based on the Naive-Bayes model [20], which are, respectively, special cases of the Bregman hard and soft clustering algorithms for KL-divergence have been shown to provide superior results to other existing algorithms on large real datasets such as the 20-Newsgroups, Reuters and Dmoz datasets. This is to be expected as text documents can be effectively modeled using multinomial distributions whose corresponding Bregman divergence is just the KL-divergence between the word distributions. Recently, [16] showed that a convex combination of KL-divergence and squared Euclidean distance seems to give even better performance on this domain. We note that [16] essentially uses a Bregman divergence derived from a convex function that is convex combination of two convex functions [23], and hence, is a special case of the proposed Bregman clustering framework.

Speech coding is another domain where a special case of the Bregman clustering algorithm based on the Itakura-Saito distance, namely the Linde-Buzo-Gray (LBG) algorithm [17], has been widely and successfully applied. Again, this is to be expected as speech power spectra tend to follow exponential family density of the form $p(x) = \lambda e^{-\lambda x}$, whose corresponding Bregman divergence is the Itakura-Saito distance.

Since special cases of Bregman clustering algorithms have already been known to provide substantial improvements over other existing methods in certain domains, we do not experimentally re-evaluate the Bregman clustering algorithms against other methods. Instead, we only focus on showing that the quality of the clustering depends on the match between the data characteristics and the choice of Bregman divergence used for clustering.

In order to do this, we performed two experiments using datasets of increasing level of difficulty. 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 mixture models had three components with equal priors centered at 10, 20 and 40 respectively. The standard deviation, σ of the Gaussian densities was set to 5 and the number of trials N of the Binomial distribution was set to 100 so as to make the three models somewhat similar to each other. The datasets were then each clustered using three versions of the Bregman hard clustering algorithm corresponding to the Bregman divergences obtained from the Gaussian (`kmeans`), Poisson and Binomial distributions respectively. The quality of the clustering was measured in terms of the normalized mutual information[6] [24] between the predicted clusters and the original clusters and the results were averaged over 10 trials. Table 4 shows the normalized mutual information values for the different divergences and datasets. From the table, we can see that clustering quality depends on the choice of Bregman

[6] It is meaningless to compare the clustering objective function values as they are different for the three versions of the Bregman clustering algorithm.

divergence and is significantly better when the appropriate Bregman divergence is used.

Table 4: Clustering results for the first set of datasets

Model	$\mathbf{D_{Gaussian}}$	$\mathbf{D_{Poisson}}$	$\mathbf{D_{Binomial}}$
Gaussian	$\mathbf{0.70 \pm 0.033}$	0.63 ± 0.043	0.64 ± 0.035
Poisson	0.69 ± 0.063	$\mathbf{0.73 \pm 0.057}$	0.69 ± 0.059
Binomial	0.77 ± 0.061	0.75 ± 0.048	$\mathbf{0.83 \pm 0.046}$

The second experiment involved a similar kind of comparison of clustering algorithms for multi-dimensional datasets drawn from multivariate Gaussian, Binomial and Poisson distributions respectively. The datasets were sampled from mixture models with 15 overlapping components and had 2000 10-dimensional samples each. The results of the Bregman clustering algorithms shown in table 5 lead to the same conclusion as before, i.e., the choice of the Bregman divergence used for clustering is crucial for obtaining good quality.

Table 5: Clustering results for the second set of datasets

Model	$\mathbf{D_{Gaussian}}$	$\mathbf{D_{Poisson}}$	$\mathbf{D_{Binomial}}$
Gaussian	$\mathbf{0.73 \pm 0.005}$	0.66 ± 0.007	0.67 ± 0.005
Poisson	0.79 ± 0.013	$\mathbf{0.82 \pm 0.014}$	0.80 ± 0.013
Binomial	0.82 ± 0.006	0.83 ± 0.011	$\mathbf{0.85 \pm 0.012}$

6 Related Work

This work is largely inspired by three broad and overlapping ideas. First, considering the clustering problem from an information theoretic viewpoint is very insightful. Such considerations occur in several techniques, from classical vector quantization to information theoretic clustering [10] and the information bottleneck method [26]. In particular, the information theoretic clustering [10] approach solved the problem of distributional clustering with a formulation involving loss in Shannon's mutual information. In this paper, we have significantly generalized that work by proposing techniques for obtaining optimal quantizations by minimizing loss in Bregman information corresponding to arbitrary Bregman divergences.

Second, our soft clustering approach is based on the relationship between Bregman divergences and exponential distributions and the suitability of Bregman divergences as distortion or loss functions for data drawn from exponential distributions. It has been previously shown [3] that the KL-divergence, which is the most natural distance measure for this parameter space, between two members $p_{\boldsymbol{\theta}}$ and $p_{\tilde{\boldsymbol{\theta}}}$ of an exponential family, is always a Bregman divergence. In particular, it is the Bregman divergence $D_\psi(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ corresponding to the cumulant function ψ of the exponential family. In our work, we extend this concept to say that the Bregman divergence of the Legendre conjugate of the cumulant function is, in some sense, a natural distance function for the data drawn according to that exponential family.

The third broad idea is that many learning algorithms can be viewed as solutions for minimizing loss functions based on Bregman divergences. Elegant techniques for the design of algorithms and the analysis of relative loss bounds in the online learning setting extensively use this framework [3]. In the unsupervised learning setting, use of this framework typically involves development of alternate minimization procedures [8]. For example, [21, 27] analyze and develop iterative alternate projection procedures for solving unsupervised optimization problems involving objective functions based on Bregman divergences under various kinds of constraints. Further, [7] develops a generalization of PCA for exponential families using loss functions based on the corresponding Bregman divergences and proposes alternate minimization schemes for solving the problem.

There has also been work on learning algorithms that involve minimizing loss functions based on distortion measures that are somewhat different from Bregman divergences. For example, [19] presents the convex-`kmeans` clustering for distortion measures that are always non-negative and convex in the second argument, using the notion of a generalized centroid. Bregman divergences, on the other hand, are not necessarily convex in the second argument and also, the minimizer of the Bregman clustering loss function always happens to be the actual centroid, i.e., the expectation of the set.

7 Conclusion

In this paper, we presented hard and soft clustering algorithms minimizing loss functions involving Bregman divergences. This analysis presents a unified view of an entire class of parametric clustering algorithms. First, in the hard-clustering framework, we show that a `kmeans` type iterative relocation scheme solves the Bregman hard-clustering problem for all Bregman divergences. Further, from a related result, we see that Bregman divergences are the only distortion functions for which such a centroid-based clustering scheme is possible. Secondly, using insights from existing literature, we show that there is a bijection between Bregman divergences and exponential families. This result is useful in developing an alternative interpretation of the EM algorithm for learning mixtures of exponential distributions, eventually resulting in a set of Bregman soft-clustering algorithms.

As discussed in the paper, special cases of this analysis have been discovered and widely used by researchers in applications ranging from speech coding to text clustering. There are four salient features of this framework that make these results particularly useful for data-mining applications. First, the computational complexity of the entire class of Bregman clustering algorithms is linear in the data-points. Hence, the algorithms are extremely scalable and appropriate for large-scale data-mining tasks. Secondly, the modularity of the proposed class of algorithms is evident from the fact that only one component in the proposed schemes, viz the Bregman divergence used in the assignment step, needs to be changed to get an algorithm for a new loss function. This simplifies the implementation and application of this class of algorithms to various data types. Thirdly, the algorithms discussed are applicable to mixed data types that are commonly encountered in data-mining. Since the linear combination of convex functions with non-negative coefficients is always convex [23], one can have different convex functions appropriately chosen for different sets of features. The Bregman divergence corresponding to any linear combination of the component convex functions can now be used to cluster the data. This vastly increases the scope of the proposed techniques. Finally, because of the similarity of Bregman hard clustering to `kmeans`, existing techniques that employ `kmeans` in various settings such as data stream clustering, privacy preserving clustering, etc., can be easily extended to the general framework of Bregman clustering.

References

[1] C. C. Aggarwal. Towards systematic design of distance functions for data mining applications. In *Proc. 9th Intl. Conf. on KDD*, pages 9–19, 2003.

[2] S. Amari. Information geometry of the EM and em algorithms for neural networks. *Neural Networks*, 8(9):1379–1408, 1995.

[3] K. S. Azoury and M. K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, 2001.

[4] A. Banerjee, X. Guo, and H. Wang. On the optimality of conditional expectation as a Bregman predictor. Submitted for publication, 2003.

[5] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh. Clustering with Bregman divergences. Technical Report TR-03-19, Department of Computer Science, University of Texas at Austin, 2003.

[6] P. S. Bradley, U. M. Fayyad, and C. Reina. Scaling clustering algorithms to large databases. In *Proc. 4th Intl. Conf. on KDD*, pages 9–15, 1998.

[7] M. Collins, S. Dasgupta, and R. Schapire. A generalization of principal component analysis to the exponential family. In *Proc. 14th Ann. Conf. on NIPS*, 2001.

[8] I. Csiszár and G. Tusnády. Information geometry and alternating minimization procedures. *Statistics and Decisions, Supplement Issue*, 1(1):205–237, 1984.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[10] I. Dhillon, S. Mallela, and R. Kumar. A divisive information-theoretic feature clustering algorithm for text classification. *Journal of Machine Learning Research*, 3(4):1265–1287, 2003.

[11] J. Forster and M. K. Warmuth. Relative expected instantaneous loss bounds. In *Proc. 13th Ann. Conf. on COLT*, pages 90–99, 2000.

[12] B. Fristedt and L. Gray. *A Modern Approach to Probability Theory*. Birkhauser Verlag, 1997.

[13] P. D. Grunwald and P. Vitányi. Kolmogorov complexity and information theory. with an interpretation in terms of questions and answers. *Journal of Logic, Language and Information*, 12(4):497–529, 2003.

[14] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, New Jersey, 1988.

[15] M. Kearns, Y. Mansour, and A. Ng. An information-theoretic analysis of hard and soft assignment methods for clustering. In *Proc. 13th UAI*, pages 282–293, 1997.

[16] J. Kogan, M. Teboulle, and C. Nicholas. Optimization approach to generating families of k–means like algorithms. In *Proc. Workshop on Clustering High Dimensional Data and its Applications: 3rd SDM*, 2003.

[17] Y. Linde, A. Buzo, and R. M. Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, 28(1):84–95, 1980.

[18] G. J. McLachlan and T. Krishnan. *The EM algorithm and Extensions*. Wiley-Interscience, 1996.

[19] D. Modha and S. Spangler. Feature weighting in k-means clustering. *Machine Learning*, 52(3):217–237, 2003.

[20] K. Nigam, A. K. McCallum, S. Thrun, and T. M. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.

[21] S. D. Pietra, V. D. Pietra, and J. Lafferty. Duality and auxiliary functions for Bregman distances. Technical Report CMU-CS-01-109, School of Computer Science, Carnegie Mellon University, 2001.

[22] R. Redner and H. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, 1984.

[23] R. T. Rockafeller. *Convex Analysis*. Princeton University Press, 1970.

[24] A. Strehl and J. Ghosh. Cluster ensembles – a knowledge reuse framework for combining partitionings. *Journal of Machine Learning Research*, 3(3):583–617, 2002.

[25] A. Strehl, J. Ghosh, and R. Mooney. Impact of similarity measures on web-page clustering. In *Proc. Workshop on AI for Web Search : 17th AAAI*, pages 58–64. AAAI, 2000.

[26] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Proc. 37th Ann. Allerton Conf. on Communication, Control and Computing*, pages 368–377, 1999.

[27] S. Wang and D. Schuurmans. Learning latent variable models with Bregman divergences. In *Proc. IEEE International Symposium on Information Theory*, 2003.