Journal of Machine Learning Research 6 (2005) 1705–1749 Submitted 10/03; Revised 2/05; Published 10/05

Clustering with Bregman Divergences

Arindam Banerjee ABANERJE@ECE.UTEXAS.EDU
Srujana Merugu MERUGU@ECE.UTEXAS.EDU
Department of Electrical and Computer Engineering
University of Texas at Austin
Austin, TX 78712, USA.

Inderjit S. Dhillon INDERJIT@CS.UTEXAS.EDU
Department of Computer Sciences
University of Texas at Austin
Austin, TX 78712, USA

Joydeep Ghosh GHOSH@ECE.UTEXAS.EDU
Department of Electrical and Computer Engineering
University of Texas at Austin
Austin, TX 78712, USA.

Editor: John Lafferty

Abstract

A wide variety of distortion functions, such as squared Euclidean distance, Mahalanobis distance, Itakura-Saito distance and relative entropy, have been used for clustering. 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`, the Linde-Buzo-Gray (LBG) algorithm 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 method to a large class of clustering loss functions. This is achieved by first posing the hard clustering problem in terms of minimizing the loss in Bregman information, a quantity motivated by rate distortion theory, and then deriving an iterative algorithm that monotonically decreases this loss. In addition, we show that there is a bijection between regular exponential families and a large class of Bregman divergences, that we call regular Bregman divergences. This result enables the development of an alternative interpretation of an efficient EM scheme for learning mixtures of exponential family distributions, and leads to a simple soft clustering algorithm for regular Bregman divergences. Finally, we discuss the connection between rate distortion theory and Bregman clustering and present an information theoretic analysis of Bregman clustering algorithms in terms of a trade-off between compression and loss in Bregman information.

Keywords: clustering, Bregman divergences, Bregman information, exponential families, expectation maximization, information theory

1. Introduction

Data clustering is a fundamental "unsupervised" learning procedure that has been extensively studied across varied disciplines over several decades (Jain and Dubes, 1988). Most of the existing parametric clustering methods 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. Typically, these clustering methods 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. In soft clustering, each data point has a certain probability of belonging to each of the partitions. One can think of hard clustering as a special case of soft clustering where these probabilities only take values 0 or 1. The popularity of parametric clustering algorithms stems from their simplicity and scalability.

Several algorithms for solving particular versions of parametric clustering problems have been developed over the years. Among the hard clustering algorithms, the most well-known is the iterative relocation scheme for the Euclidean `kmeans` algorithm (MacQueen, 1967; Jain and Dubes, 1988; Duda et al., 2001). Another widely used clustering algorithm with a similar scheme is the Linde-Buzo-Gray (LBG) algorithm (Linde et al., 1980; Buzo et al., 1980) based on the Itakura-Saito distance, which has been used in the signal-processing community for clustering speech data. The recently proposed information theoretic clustering algorithm (Dhillon et al., 2003) for clustering probability distributions also has a similar flavor.

The observation that for certain distortion functions, e.g., squared Euclidean distance, KL-divergence (Dhillon et al., 2003), Itakura-Saito distance (Buzo et al., 1980) etc., the clustering problem can be solved using appropriate `kmeans` type iterative relocation schemes leads to a natural question: *what class of distortion functions admit such an iterative relocation scheme where a global objective function based on the distortion with respect to cluster centroids*[1] *is progressively decreased?* In this paper, we provide an answer to this question: we show that *such a scheme works for arbitrary Bregman divergences*. In fact, it can be shown (Banerjee et al., 2005) that such a simple scheme works *only* when the distortion is a Bregman divergence. The scope of this result is vast since Bregman divergences include a large number of useful loss functions such as squared loss, KL-divergence, logistic loss, Mahalanobis distance, Itakura-Saito distance, I-divergence, 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*, otherwise known as information theoretic clustering (Dhillon et al., 2003), is seen to be a special case of our approach. Thus, this paper unifies several parametric partitional clustering approaches.

Several researchers have observed relationships between Bregman divergences and exponential families (Azoury and Warmuth, 2001; Collins et al., 2001). In this paper, we formally prove an observation made by Forster and Warmuth (2000) that *there exists a unique Bregman divergence corresponding to every regular exponential family*. In fact, we show that there is a bijection between regular exponential families and a class of Bregman divergences, that we call regular Bregman divergences. We show that, with proper representation, the bijection provides an alternative interpretation of a well known efficient EM scheme (Redner and Walker, 1984) for learning mixture models of exponential family distributions. This scheme simplifies the computationally intensive maximization step of the EM algorithm, resulting in a general soft-clustering algorithm for all regular Bregman divergences. We also present an information theoretic analysis of Bregman clustering algorithms in terms of a trade-off between compression and loss in Bregman information.

1. We use the term "cluster centroid" to denote the expectation of the data points in that cluster.

1.1 Contributions

We briefly summarize the main contributions of this paper:

1. In the context of hard clustering, we introduce the concept of *Bregman Information* (Section 3) that measures the minimum expected loss incurred by encoding a set of data points using a constant, where loss is measured in terms of a Bregman divergence. Variance and mutual information are shown to be special cases of Bregman information. Further, we show a close connection between Bregman information and Jensen's inequality.

2. Hard clustering with Bregman divergences is posed as a quantization problem that involves minimizing loss of Bregman information. We show (Theorem 1 in Section 3) that for any given clustering, the loss in Bregman information is equal to the expected Bregman divergence of data points to their respective cluster centroids. Hence, minimizing either of these quantities yields the same optimal clustering.

3. Based on our analysis of the Bregman clustering problem, we present a meta hard clustering algorithm that is applicable to *all* Bregman divergences (Section 3). The meta clustering algorithm retains the simplicity and scalability of `kmeans` and is a direct generalization of all previously known centroid-based parametric hard clustering algorithms.

4. To obtain a similar generalization for the soft clustering case, we show (Theorem 4, Section 4) that there is a uniquely determined Bregman divergence corresponding to every regular exponential family. This result formally proves an observation made by Forster and Warmuth (2000). In particular, in Section 4.3, we show that the log-likelihood of any parametric exponential family is equal to the negative of the corresponding Bregman divergence to the expectation parameter, up to a fixed additive non-parametric function. Further, in Section 4.4, we define regular Bregman divergences using exponentially convex functions and show that there is a bijection between regular exponential families and regular Bregman divergences.

5. Using the correspondence between exponential families and Bregman divergences, we show that the mixture estimation problem based on regular exponential families is identical to a Bregman soft clustering problem (Section 5). Further, we describe an EM scheme to efficiently solve the mixture estimation problem. Although this particular scheme for learning mixtures of exponential families was previously known (Redner and Walker, 1984), the Bregman divergence viewpoint explaining the efficiency is new. In particular, we give a correctness proof of the efficient M-step updates using properties of Bregman divergences.

6. Finally, we study the relationship between Bregman clustering and rate distortion theory (Section 6). Based on the results in Banerjee et al. (2004a), we observe that the Bregman hard and soft clustering formulations correspond to the "scalar" and asymptotic rate distortion problems respectively, where distortion is measured using a regular Bregman divergence. Further, we show how each of these problems can be interpreted as a trade-off between compression and loss in Bregman information. The information-bottleneck method (Tishby et al., 1999) can be readily derived as a special case of this trade-off.

A word about the notation: bold faced variables, e.g., $\mathbf{x}, \boldsymbol{\mu}$, are used to represent vectors. Sets are represented by calligraphic upper-case alphabets, e.g., $\mathcal{X}, \mathcal{Y}$. Random variables are represented

by upper-case alphabets, e.g., X, Y. The symbols $\mathbb{R}, \mathbb{N}, \mathbb{Z}$ and $\mathbb{R}^d$ denote the set of reals, the set of natural numbers, the set of integers and the d-dimensional real vector space respectively. Further, $\mathbb{R}_+$ and $\mathbb{R}_{++}$ denote the set of non-negative and positive real numbers. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, $\|\mathbf{x}\|$ denotes the L_2 norm, and $\langle \mathbf{x}, \mathbf{y} \rangle$ denotes the inner product. Unless otherwise mentioned, log will represent the natural logarithm. Probability density functions (with respect to the Lebesgue or the counting measure) are denoted by lower case alphabets such as p, q. If a random variable X is distributed according to ν, expectation of functions of X are denoted by $E_X[\cdot]$, or by $E_\nu[\cdot]$ when the random variable is clear from the context. The interior, relative interior, boundary, closure and closed convex hull of a set $\mathcal{X}$ are denoted by $\text{int}(\mathcal{X})$, $\text{ri}(\mathcal{X})$, $\text{bd}(\mathcal{X})$, $\text{cl}(\mathcal{X})$ and $\text{co}(\mathcal{X})$ respectively. The effective domain of a function f, i.e., set of all x such that $f(x) < +\infty$ is denoted by $\text{dom}(f)$ while the range is denoted by $\text{range}(f)$. The inverse of a function f, when well-defined, is denoted by f^{-1}.

2. Preliminaries

In this section, we define the Bregman divergence corresponding to a strictly convex function and present some examples.

Definition 1 (Bregman, 1967; Censor and Zenios, 1998) Let $\phi : \mathcal{S} \mapsto \mathbb{R}, \mathcal{S} = \text{dom}(\phi)$ be a strictly convex function defined on a convex set $\mathcal{S} \subseteq \mathbb{R}^d$ such that ϕ is differentiable on $\text{ri}(\mathcal{S})$, assumed to be nonempty. The *Bregman divergence* $d_\phi : \mathcal{S} \times \text{ri}(\mathcal{S}) \mapsto [0, \infty)$ 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 $\nabla\phi(\mathbf{y})$ represents the gradient vector of ϕ evaluated at $\mathbf{y}$.

Example 1 Squared Euclidean distance is perhaps the simplest and most widely used Bregman divergence. The underlying function $\phi(\mathbf{x}) = \langle \mathbf{x}, \mathbf{x} \rangle$ is strictly convex, differentiable on $\mathbb{R}^d$ and

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

Example 2 Another widely used Bregman divergence is the KL-divergence. If $\mathbf{p}$ is a discrete probability distribution so that $\sum_{j=1}^d p_j = 1$, the negative entropy $\phi(\mathbf{p}) = \sum_{j=1}^d p_j \log_2 p_j$ is a convex function. The corresponding Bregman divergence is

$$\begin{aligned} d_\phi(\mathbf{p}, \mathbf{q}) &= \sum_{j=1}^d p_j \log_2 p_j - \sum_{j=1}^d q_j \log_2 q_j - \langle \mathbf{p} - \mathbf{q}, \nabla\phi(\mathbf{q}) \rangle \\ &= \sum_{j=1}^d p_j \log_2 p_j - \sum_{j=1}^d q_j \log_2 q_j - \sum_{j=1}^d (p_j - q_j)(\log_2 q_j + \log_2 e) \\ &= \sum_{j=1}^d p_j \log_2 \left(\frac{p_j}{q_j} \right) - \log_2 e \sum_{j=1}^d (p_j - q_j) \\ &= KL(\mathbf{p} \| \mathbf{q}) \ , \end{aligned}$$

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

Table 1: Bregman divergences generated from some convex functions.

Domain	$\phi(\mathbf{x})$	$d_\phi(\mathbf{x},\mathbf{y})$	**Divergence**
$\mathbb{R}$	x^2	$(x-y)^2$	Squared 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 [3]
$\mathbb{R}_{++}$	$-\log x$	$\frac{x}{y}-\log(\frac{x}{y})-1$	Itakura-Saito distance
$\mathbb{R}$	e^x	$e^x-e^y-(x-y)e^y$	
$\mathbb{R}^d$	$\|\mathbf{x}\|^2$	$\|\mathbf{x}-\mathbf{y}\|^2$	Squared Euclidean distance
$\mathbb{R}^d$	$\mathbf{x}^TA\mathbf{x}$	$(\mathbf{x}-\mathbf{y})^TA(\mathbf{x}-\mathbf{y})$	Mahalanobis distance [4]
d-Simplex	$\sum_{j=1}^d x_j\log_2 x_j$	$\sum_{j=1}^d x_j\log_2(\frac{x_j}{y_j})$	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

Example 3 Itakura-Saito distance is another Bregman divergence that is widely used in signal processing. If $F(e^{j\theta})$ is the power spectrum[2] of a signal $f(t)$, then the functional $\phi(F)=-\frac{1}{2\pi}\int_{-\pi}^{\pi}\log(F(e^{j\theta}))d\theta$ is convex in F and corresponds to the negative entropy rate of the signal assuming it was generated by a stationary Gaussian process (Palus, 1997; Cover and Thomas, 1991). The Bregman divergence between $F(e^{j\theta})$ and $G(e^{j\theta})$ (the power spectrum of another signal $g(t)$) is given by

$$\begin{aligned} d_\phi(F,G) &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(-\log(F(e^{j\theta}))+\log(G(e^{j\theta}))-(F(e^{j\theta})-G(e^{j\theta}))\left(-\frac{1}{G(e^{j\theta})}\right)\right)d\theta \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(-\log\left(\frac{F(e^{j\theta})}{G(e^{j\theta})}\right)+\frac{F(e^{j\theta})}{G(e^{j\theta})}-1\right)d\theta, \end{aligned}$$

which is exactly the Itakura-Saito distance between the power spectra $F(e^{j\theta})$ and $G(e^{j\theta})$ and can also be interpreted as the I-divergence (Csiszár, 1991) between the generating processes under the assumption that they are equal mean, stationary Gaussian processes (Kazakos and Kazakos, 1980).

Table 1 contains a list of some common 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 Appendix A.

3. 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 data

2. Note that $F(\cdot)$ is a function and it is possible to extend the notion of Bregman divergences to the space of functions (Csiszár, 1995; Grünwald and Dawid, 2004).
3. For $x\in\{0,1\}$ (Bernoulli) and $y\in(0,1)$ (posterior probability for $x=1$), we have $x\log(\frac{x}{y})+(1-x)\log(\frac{1-x}{1-y})=\log(1+\exp(-f(x)g(y)))$, i.e., the logistic loss with $f(x)=2x-1$ and $g(y)=\log(\frac{y}{1-y})$.
4. The matrix A is assumed to be positive definite; $(\mathbf{x}-\mathbf{y})^TA(\mathbf{x}-\mathbf{y})$ is called the Mahalanobis distance when A is the inverse of the covariance matrix.

points from their representatives is minimized. We then present a clustering algorithm that generalizes the iterative relocation scheme of `kmeans` to monotonically decrease the loss in Bregman information.

3.1 Bregman Information

The dual formulation of Shannon's celebrated rate distortion problem (Cover and Thomas, 1991; Grünwald and Vitányi, 2003) involves finding a coding scheme with a given rate, i.e., average number of bits per symbol, 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., the infimum distortion achievable for a given rate. Now consider a random variable X that takes values in a finite set $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{S} \subseteq \mathbb{R}^d$ ($\mathcal{S}$ is convex) following a discrete probability measure ν. Let the distortion be measured by a Bregman divergence d_ϕ. Consider a simple encoding scheme that represents 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 $E_\nu[d_\phi(X,\mathbf{s})]$ that depends on the choice of the representative $\mathbf{s}$ and can be optimized by picking the right representative. We call this optimal distortion-rate function the *Bregman information* of the random variable X for the Bregman divergence d_ϕ and denote it by $I_\phi(X)$, i.e.,

$$I_\phi(X) = \min_{\mathbf{s}\in \mathrm{ri}(\mathcal{S})} E_\nu[d_\phi(X,\mathbf{s})] = \min_{\mathbf{s}\in \mathrm{ri}(\mathcal{S})} \sum_{i=1}^n \nu_i \, d_\phi(\mathbf{x}_i,\mathbf{s}) \, . \tag{1}$$

The optimal vector $\mathbf{s}$ 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 Bregman divergence. In fact, the minimizer is just the expectation of the random variable X.

Proposition 1 *Let X be a random variable that take values in $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{S} \subseteq \mathbb{R}^d$ following a positive probability measure ν such that $E_\nu[X] \in \mathrm{ri}(\mathcal{S})$.*[5] *Given a Bregman divergence $d_\phi : \mathcal{S} \times \mathrm{ri}(\mathcal{S}) \mapsto [0,\infty)$, the problem*

$$\min_{\mathbf{s}\in \mathrm{ri}(\mathcal{S})} E_\nu[d_\phi(X,\mathbf{s})] \tag{2}$$

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

Proof The function we are trying to minimize is $J_\phi(\mathbf{s}) = E_\nu[d_\phi(X,\mathbf{s})] = \sum_{i=1}^n \nu_i d_\phi(\mathbf{x}_i,\mathbf{s})$. Since $\boldsymbol{\mu} = E_\nu[X] \in \mathrm{ri}(S)$, the objective function is well-defined at $\boldsymbol{\mu}$. Now, $\forall \mathbf{s} \in \mathrm{ri}(\mathcal{S})$,

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

5. The assumption that $E_\nu[X] \in \mathrm{ri}(\mathcal{S})$ is not restrictive since a violation can occur only when $\mathrm{co}(\mathcal{X}) \subset \mathrm{bd}(S)$, i.e., the entire convex hull of $\mathcal{X}$ is on the boundary of S.

with equality only when $\mathbf{s} = \boldsymbol{\mu}$ by the strict convexity of ϕ (Appendix A, Property 1). Hence, $\boldsymbol{\mu}$ is the unique minimizer of J_ϕ. ■

Note that the minimization in (2) is with respect to the second argument of d_ϕ. Proposition 1 is somewhat surprising since Bregman divergences are not necessarily convex in the second argument as the following example demonstrates.

Example 4 Consider $\phi(\mathbf{x}) = \sum_{j=1}^{3} x_j^3$ defined on $\mathbb{R}_+^3$ so that $d_\phi(\mathbf{x},\mathbf{s}) = \sum_{j=1}^{3}(x_j^3 - s_j^3 - 3(x_j - s_j)s_j^2)$. For the random variable X distributed uniformly over the set $\mathcal{X} = \{(1,1,1),(2,2,2),(3,3,3),(4,4,4),(5,5,5)\}$,

$$E[d_\phi(X,\mathbf{s})] = 135 + 2\sum_{j=1}^{3} s_j^3 - 9\sum_{j=1}^{3} s_j^2 \ ,$$

which is clearly not convex in $\mathbf{s}$ since the Hessian $\nabla^2 J_\phi(\mathbf{s}) = \mathrm{diag}(12\mathbf{s} - 18)$ is not positive definite. However, $J_\phi(\mathbf{s})$ is uniquely minimized by $\mathbf{s} = (3,3,3)$, i.e., the expectation of the random variable X.

Interestingly, the converse of Proposition 1 is also true, i.e., for all random variables X, if $E[X]$ minimizes the expected distortion of X to a fixed point for a smooth distortion function $F(x,y)$ (see Appendix B for details), then $F(x,y)$ has to be a Bregman divergence (Banerjee et al., 2005). Thus, Bregman divergences are *exhaustive* with respect to the property proved in Proposition 1.

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

Definition 2 Let X be a random variable that takes values in $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{S}$ following a probability measure ν. Let $\boldsymbol{\mu} = E_\nu[X] = \sum_{i=1}^n \nu_i \mathbf{x}_i \in \mathrm{ri}(\mathcal{S})$ and let $d_\phi : \mathcal{S} \times \mathrm{ri}(\mathcal{S}) \mapsto [0,\infty)$ be a Bregman divergence. Then the *Bregman Information* of X in terms of d_ϕ is defined as

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

Example 5 (**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

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

which is just the sample variance.

Example 6 (**Mutual Information**) By definition, the mutual information $I(U;V)$ between two discrete random variables U and V with joint distribution $\{\{p(u_i,v_j)\}_{i=1}^n\}_{j=1}^m$ is given by

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

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

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

Hence,

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

i.e., mutual information is the Bregman information of Z_u when d_ϕ is the KL-divergence. Similarly, for a random variable Z_v that takes values in the set of probability distributions $\mathcal{Z}_v = \{p(U|v_j)\}_{j=1}^m$ following the probability measure $\{\nu_j\}_{j=1}^m = \{p(v_j)\}_{j=1}^m$ over this set, one can show that $I(U;V) = I_\phi(Z_v)$. The Bregman information of Z_u and Z_v can also be interpreted as the Jensen-Shannon divergence of the sets $\mathcal{Z}_u$ and $\mathcal{Z}_v$ (Dhillon et al., 2003).

Example 7 The Bregman information corresponding to Itakura-Saito distance also has a useful interpretation. Let $\mathcal{F} = \{F_i\}_{i=1}^n$ be a set of power spectra corresponding to n different signals, and let ν be a probability measure on $\mathcal{F}$. Then, the Bregman information of a random variable F that takes values in $\mathcal{F}$ following ν, with Itakura-Saito distance as the Bregman divergence, is given by

$$\begin{aligned} I_\phi(F) &= \sum_{i=1}^n \nu_i d_\phi(F_i,\bar{F}) = \sum_{i=1}^n \frac{\nu_i}{2\pi}\int_{-\pi}^{\pi}\left(-\log\left(\frac{F_i(e^{j\theta})}{F(e^{j\theta})}\right)+\frac{F_i(e^{j\theta})}{F(e^{j\theta})}-1\right)d\theta \\ &= -\frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{i=1}^n \nu_i \log\left(\frac{F_i(e^{j\theta})}{F(e^{j\theta})}\right)d\theta, \end{aligned}$$

where $\bar{F}$ is the marginal average power spectrum. Based on the connection between the corresponding convex function ϕ and the negative entropy of Gaussian processes (Cover and Thomas, 1991; Palus, 1997), it can be shown that the Bregman information $I_\phi(F)$ is the Jensen-Shannon divergence of the generating processes under the assumption that they are equal mean, stationary Gaussian processes. Further, consider a n-class signal classification problem where each class of signals is assumed to be generated by a certain Gaussian process. Now, if $P_e(t)$ is the optimal Bayes error for this classification problem averaged upto time t, then $P_e(t)$ is bounded above and below by functions of the Chernoff coefficient $B(t)$ (Kazakos and Kazakos, 1980) of the generating Gaussian processes. The asymptotic value of this Chernoff coefficient as t tends to ∞ is a function of the Bregman information of F, i.e.,

$$\lim_{t\to\infty} B(t) = \exp(-\frac{1}{2}I_\phi(F)).$$

and is directly proportional to the optimal Bayes error.

3.1.1 JENSEN'S INEQUALITY AND BREGMAN INFORMATION

An alternative interpretation of Bregman information can also be made in terms of Jensen's inequality (Cover and Thomas, 1991). Given any convex function ϕ, for any random variable X, Jensen's inequality states that

$$E[\phi(X)] \geq \phi(E[X]) \ .$$

A direct calculation using the definition of Bregman information shows that (Banerjee et al., 2004b)

$$\begin{aligned} E[\phi(X)] - \phi(E[X]) &\overset{(a)}{=} E[\phi(X)] - \phi(E[X]) - E[\langle X - E[X], \nabla\phi(E[X])\rangle] \\ &\overset{(b)}{=} E[\phi(X) - \phi(E[X]) - \langle X - E[X], \nabla\phi(E[X])\rangle] \\ &= E[d_\phi(X, E[X])] \;=\; I_\phi(X) \;\geq\; 0\,, \end{aligned}$$

where (a) follows since the last term is 0, and (b) follows from the linearity of expectation. Thus the difference between the two sides of Jensen's inequality is exactly equal to the Bregman information.

3.2 Clustering Formulation

Let X be a random variable that takes values in $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$ following the probability measure ν. When X has a large Bregman information, it may not suffice to encode X using a single representative since a lower quantization error may be desired. In such a situation, a natural goal is to split the set $\mathcal{X}$ into k disjoint partitions $\{\mathcal{X}_h\}_{h=1}^k$, each with its own Bregman representative, such that a random variable M over the partition representatives serves as an appropriate quantization of X. Let $\mathcal{M} = \{\boldsymbol{\mu}_h\}_{h=1}^k$ denote the set of representatives, and $\pi = \{\pi_h\}_{h=1}^k$ with $\pi_h = \sum_{\mathbf{x}_i \in \mathcal{X}_h} \nu_i$ denote the induced probability measure on $\mathcal{M}$. Then the induced random variable M takes values in $\mathcal{M}$ following π.

The quality of the quantization M can be measured by the expected Bregman divergence between X and M, i.e., $E_{X,M}[d_\phi(X,M)]$. Since M is a deterministic function of X, the expectation is actually over the distribution of X, so that

$$E_X[d_\phi(X,M)] \;=\; \sum_{h=1}^{k} \sum_{\mathbf{x}_i \in \mathcal{X}_h} \nu_i d_\phi(\mathbf{x}_i, \boldsymbol{\mu}_h) \;=\; \sum_{h=1}^{k} \pi_h \sum_{\mathbf{x}_i \in \mathcal{X}_h} \frac{\nu_i}{\pi_h} d_\phi(\mathbf{x}_i, \boldsymbol{\mu}_h) \;=\; E_\pi[I_\phi(X_h)]\,,$$

where X_h is the random variable that takes values in the partition $\mathcal{X}_h$ following a probability distribution $\frac{\nu_i}{\pi_h}$, and $I_\phi(X_h)$ is the Bregman information of X_h. Thus, the quality of the quantization is equal to the expected Bregman information of the partitions.

An alternative way of measuring the quality of the quantization M can be formulated from an information theoretic viewpoint. In information-theoretic clustering (Dhillon et al., 2003), the quality of the partitioning is measured in terms of the loss in mutual information resulting from the quantization of the original random variable X. Extending this formulation, we can measure the quality of the quantization M by the loss in Bregman information due to the quantization, i.e., by $I_\phi(X) - I_\phi(M)$. For $k = n$, the best choice is of course $M = X$ with no loss in Bregman information. For $k = 1$, the best quantization is to pick $E_\nu[X]$ with probability 1, incurring a loss of $I_\phi(X)$. For intermediate values of k, the solution is less obvious.

Interestingly the two possible formulations outlined above turn out to be identical (see Theorem 1 below). We choose the information theoretic viewpoint to pose the problem, since we will study the connections of both the hard and soft clustering problems to rate distortion theory in Section 6. Thus we define the *Bregman hard clustering problem* as that of finding a partitioning of $\mathcal{X}$, or, equivalently, finding the random variable M, such that *the loss in Bregman information* due to quantization, $L_\phi(M) = I_\phi(X) - I_\phi(M)$, *is minimized*. Typically, clustering algorithms assume a uniform measure, i.e., $\nu_i = \frac{1}{n}, \forall i$, over the data, which is clearly a special case of our formulation.

The following theorem shows that the loss in Bregman information and the expected Bregman information of the partitions are equal.

Theorem 1 *Let X be a random variable that takes values in $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{S} \subseteq \mathbb{R}^d$ following the positive probability measure 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 be the random variable that takes values in $\mathcal{X}_h$ following $\frac{\nu_i}{\pi_h}$ for $\mathbf{x}_i \in \mathcal{X}_h$, for $h = 1, \ldots, k$. Let $\mathcal{M} = \{\boldsymbol{\mu}_h\}_{h=1}^k$ with $\boldsymbol{\mu}_h \in \mathrm{ri}(\mathcal{S})$ denote the set of representatives of $\{X_h\}_{h=1}^k$, and M be a random variable that takes values in $\mathcal{M}$ following π. Then,*

$$L_\phi(M) \;=\; I_\phi(X) - I_\phi(M) \;=\; E_\pi[I_\phi(X_h)] \;=\; \sum_{h=1}^{k} \pi_h \sum_{\mathbf{x}_i \in \mathcal{X}_h} \frac{\nu_i}{\pi_h}\, d_\phi(\mathbf{x}_i, \boldsymbol{\mu}_h)\ .$$

Proof By definition,

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

since $\sum_{\mathbf{x}_i \in \mathcal{X}_h} \frac{\nu_i}{\pi_h} \mathbf{x}_i = \boldsymbol{\mu}_h$. ■

Note that $I_\phi(X)$ can be interpreted as the "total Bregman information", and $I_\phi(M)$ can be interpreted as the "between-cluster Bregman information" since it is a measure of divergence between the cluster representatives, while $L_\phi(M)$ can be interpreted as the "within-cluster Bregman information". Thus Theorem 1 states that the total Bregman information equals the sum of the within-cluster Bregman information and between-cluster Bregman information. This is a generalization of the corresponding result for squared Euclidean distances (Duda et al., 2001).

Using Theorem 1, the Bregman clustering problem of minimizing the loss in Bregman information can be written as

$$\min_M \; \big(I_\phi(X) - I_\phi(M)\big) = \min_M \; \left(\sum_{h=1}^{k} \sum_{\mathbf{x}_i \in \mathcal{X}_h} \nu_i d_\phi(\mathbf{x}_i, \boldsymbol{\mu}_h) \right) . \tag{3}$$

Algorithm 1 Bregman Hard Clustering

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

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

Method:

Initialize $\{\boldsymbol{\mu}_h\}_{h=1}^k$ with $\boldsymbol{\mu}_h \in \mathrm{ri}(\mathcal{S})$ (one possible initialization is to choose $\boldsymbol{\mu}_h \in \mathrm{ri}(\mathcal{S})$ at random)
repeat
{The Assignment Step}
Set $\mathcal{X}_h \leftarrow \emptyset$, $1 \leq h \leq k$
for $i = 1$ to n **do**
$\mathcal{X}_h \leftarrow \mathcal{X}_h \cup \{\mathbf{x}_i\}$
where $h = h^\dagger(\mathbf{x}_i) = \operatorname{argmin}_{h'} d_\phi(\mathbf{x}_i, \boldsymbol{\mu}_{h'})$
end for
{The Re-estimation Step}
for $h = 1$ to k **do**
$\pi_h \leftarrow \sum_{\mathbf{x}_i \in \mathcal{X}_h} \nu_i$
$\boldsymbol{\mu}_h \leftarrow \frac{1}{\pi_h} \sum_{\mathbf{x}_i \in \mathcal{X}_h} \nu_i \mathbf{x}_i$
end for
until *convergence*
return $M^\dagger \leftarrow \{\mu_h\}_{h=1}^k$

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. We shall investigate the relationship of this formulation to rate distortion theory in detail in Section 6.

3.3 Clustering Algorithm

The objective function given in (3) suggests a natural iterative relocation algorithm for solving the Bregman hard clustering problem (see Algorithm 1). It is easy to see that classical `kmeans`, the LBG algorithm (Buzo et al., 1980) and the information theoretic clustering algorithm (Dhillon et al., 2003) are special cases of Bregman hard clustering for squared Euclidean distance, Itakura-Saito distance and KL-divergence respectively. The following propositions prove the convergence of the Bregman hard clustering algorithm.

Proposition 2 *The Bregman hard clustering algorithm (Algorithm 1) monotonically decreases the loss function in (3).*

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,

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

where (a) follows from the assignment step, and (b) follows from the re-estimation step and Proposition 1. Note that if equality holds, i.e., if the loss function value is equal at consecutive iterations, then the algorithm will terminate. ■

Proposition 3 *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) the assignment step 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. ■

In addition to local optimality, the Bregman hard clustering algorithm has the following interesting properties.

Exhaustiveness: The Bregman hard clustering algorithm with cluster centroids as optimal representatives works for *all* Bregman divergences and *only* for Bregman divergences since the arithmetic mean is the best predictor *only* for Bregman divergences (Banerjee et al., 2005). However, it is possible to have a similar alternate minimization based clustering algorithm for distance functions that are not Bregman divergences, the primary difference being that the optimal cluster representative, when it exists, will no longer be the arithmetic mean or the expectation. The `convex-kmeans` clustering algorithm (Modha and Spangler, 2003) and the generalizations of the LBG algorithm (Linde et al., 1980) are examples of such alternate minimization schemes where a (unique) representative exists because of convexity.

Linear Separators: For all Bregman divergences, the partitions induced by the Bregman hard clustering algorithm are separated by hyperplanes. In particular, the locus of points that are equidistant to two fixed points $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ in terms of a Bregman divergence is given by $\mathcal{X} = \{\mathbf{x} \mid d_\phi(\mathbf{x}, \boldsymbol{\mu}_1) = d_\phi(\mathbf{x}, \boldsymbol{\mu}_2)\}$, i.e., the set of points,

$$\{\mathbf{x} \mid \langle \mathbf{x}, \nabla\phi(\boldsymbol{\mu}_2) - \nabla\phi(\boldsymbol{\mu}_1)\rangle = (\phi(\boldsymbol{\mu}_1) - \langle \boldsymbol{\mu}_1, \nabla\phi(\boldsymbol{\mu}_1)\rangle) - (\phi(\boldsymbol{\mu}_2) - \langle \boldsymbol{\mu}_2, \nabla\phi(\boldsymbol{\mu}_2)\rangle)\} \ ,$$

which corresponds to a hyperplane.

Scalability: The computational complexity of each iteration of the Bregman hard clustering algorithm is linear in the number of data points and the number of desired clusters for *all* Bregman divergences, which makes the algorithm scalable and appropriate for large clustering problems.

Applicability to mixed data types: The Bregman hard clustering algorithm is applicable to mixed data types that are commonly encountered in machine learning. One can choose different convex functions that are appropriate and meaningful for different subsets of the features. The Bregman divergence corresponding to a convex combination of the component convex functions can then be used to cluster the data.

4. Relationship with Exponential Families

We now turn our attention to *soft* clustering with Bregman divergences. To accomplish our goal, we first establish that there is a unique Bregman divergence corresponding to every regular exponential

family distribution. Later, we make this relation more precise by establishing a bijection between regular exponential families and *regular Bregman divergences*. The correspondence will be used to develop the Bregman soft clustering algorithm in Section 5. To present our results, we first review some background information on exponential families and Legendre duality in Sections 4.1 and 4.2 respectively.

4.1 Exponential Families

Consider a measurable space $(\Omega, \mathcal{B})$ where $\mathcal{B}$ is a σ-algebra on the set Ω. Let $\mathbf{t}$ be a measurable mapping from Ω to a set $\mathcal{T} \subseteq \mathbb{R}^d$, where $\mathcal{T}$ may be discrete (e.g., $\mathcal{T} \subset \mathbb{N}$). Let $p_0 : \mathcal{T} \mapsto \mathbb{R}_+$ be any function such that if $(\Omega, \mathcal{B})$ is endowed with a measure $dP_0(\omega) = p_0(\mathbf{t}(\omega))d\mathbf{t}(\omega)$, then $\int_{\omega\in\Omega} dP_0(\omega) < \infty$. The measure P_0 is absolutely continuous with respect to the Lebesgue measure $d\mathbf{t}(\omega)$. When $\mathcal{T}$ is a discrete set, $d\mathbf{t}(\omega)$ is the counting measure and P_0 is absolutely continuous with respect to the counting measure.[6]

Now, $\mathbf{t}(\omega)$ is a random variable from $(\Omega, \mathcal{B}, P_0)$ to $(\mathcal{T}, \sigma(\mathcal{T}))$, where $\sigma(\mathcal{T})$ denotes the σ-algebra generated by $\mathcal{T}$. Let Θ be defined as the set of all parameters $\boldsymbol{\theta} \in \mathbb{R}^d$ for which

$$\int_{\omega\in\Omega} \exp\left(\langle \boldsymbol{\theta}, \mathbf{t}(\omega)\rangle\right) dP_0(\omega) < \infty \, .$$

Based on the definition of Θ, it is possible to define a function $\psi : \Theta \mapsto \mathbb{R}$ such that

$$\psi(\boldsymbol{\theta}) = \log\left(\int_{\omega\in\Omega} \exp(\langle \boldsymbol{\theta}, \mathbf{t}(\omega)\rangle) dP_0(\omega)\right) \, . \tag{4}$$

A family of probability distributions $\mathcal{F}_\psi$ parameterized by a d-dimensional vector $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$ such that the probability density functions with respect to the measure $d\mathbf{t}(\omega)$ can be expressed in the form

$$f(\omega; \boldsymbol{\theta}) = \exp\left(\langle \boldsymbol{\theta}, \mathbf{t}(\omega)\rangle - \psi(\boldsymbol{\theta})\right) p_0(\mathbf{t}(\omega)) \tag{5}$$

is called an *exponential family* with *natural statistic* $\mathbf{t}(\omega)$, *natural parameter* $\boldsymbol{\theta}$ and *natural parameter space* Θ. In particular, if the components of $\mathbf{t}(\omega)$ are affinely independent, i.e., $\nexists$ non-zero $\mathbf{a} \in \mathbb{R}^d$ such that $\langle \mathbf{a}, \mathbf{t}(\omega)\rangle = c$ (a constant) $\forall \omega \in \Omega$, then this representation is said to be *minimal*.[7] For a minimal representation, there exists a unique probability density $f(\omega; \boldsymbol{\theta})$ for every choice of $\boldsymbol{\theta} \in \Theta$ (Wainwright and Jordan, 2003). $\mathcal{F}_\psi$ is called a *full exponential family* of *order d* in such a case. In addition, if the parameter space Θ is open, i.e., $\Theta = \text{int}(\Theta)$, then $\mathcal{F}_\psi$ is called a *regular exponential family*.

It can be easily seen that if $\mathbf{x} \in \mathbb{R}^d$ denotes the natural statistic $\mathbf{t}(\omega)$, then the probability density function $g(\mathbf{x}; \boldsymbol{\theta})$ (with respect to the appropriate measure $d\mathbf{x}$) given by

$$g(\mathbf{x}; \boldsymbol{\theta}) = \exp(\langle \boldsymbol{\theta}, \mathbf{x}\rangle - \psi(\boldsymbol{\theta})) p_0(\mathbf{x}) \tag{6}$$

is such that $f(\omega; \boldsymbol{\theta})/g(\mathbf{x}; \boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}$. Thus, $\mathbf{x}$ is a sufficient statistic (Amari and Nagaoka, 2001) for the family, and in fact, can be shown (Barndorff-Nielsen, 1978) to be minimally

6. For conciseness, we abuse notation and continue to use the Lebesgue integral sign even for counting measures. The integral in this case actually denotes a sum over $\mathcal{T}$. Further, the use of absolute continuity in the context of counting measure is non-standard. We say the measure P_0 is absolutely continuous with respect to the counting measure μ_c if $P_0(E) = 0$ for every set with $\mu_c(E) = 0$, where E is a discrete set.
7. Strictly speaking, $\nexists$ non-zero $\mathbf{a}$ such that $P_0(\{\omega : \langle \mathbf{t}(\omega), \mathbf{a}\rangle = c\}) = 1$.

sufficient. For instance, the natural statistic for the one-dimensional Gaussian distributions denoted by $f(\omega;\sigma,\mu) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(\omega-\mu)^2}{2\sigma^2})$ is given by $\mathbf{x} = [\omega, \omega^2]$ and the corresponding natural parameter turns out to be $\boldsymbol{\theta} = [\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}]$, which can be easily verified to be minimally sufficient. For our analysis, it is convenient to work with the minimal natural sufficient statistic $\mathbf{x}$ and hence, we redefine regular exponential families in terms of the probability density of $\mathbf{x} \in \mathbb{R}^d$, noting that the original probability space can actually be quite general.

Definition 3 A multivariate parametric family $\mathcal{F}_\psi$ of distributions $\{p_{(\psi,\boldsymbol{\theta})} | \boldsymbol{\theta} \in \Theta = \text{int}(\Theta) = \text{dom}(\psi) \subseteq \mathbb{R}^d\}$ is called a regular exponential family if each probability density is of the form

$$p_{(\psi,\boldsymbol{\theta})}(\mathbf{x}) = \exp(\langle \mathbf{x}, \boldsymbol{\theta} \rangle - \psi(\boldsymbol{\theta})) p_0(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^d,$$

where $\mathbf{x}$ is a minimal sufficient statistic for the family.

The function $\psi(\boldsymbol{\theta})$ is known as the *log partition function* or the *cumulant function* corresponding to the exponential family. Given a regular exponential family $\mathcal{F}_\psi$, the log-partition function ψ is uniquely determined up to a constant additive term. It can be shown (Barndorff-Nielsen, 1978) that Θ is a non-empty convex set in $\mathbb{R}^d$ and ψ is a convex function. In fact, it is possible to prove a stronger result that characterizes ψ in terms of a special class of convex functions called Legendre functions, which are defined below.

Definition 4 (Rockafellar (1970)) Let ψ be a proper, closed[8] convex function with $\Theta = \text{int}(\text{dom}(\psi))$. The pair (Θ, ψ) is called a convex function of Legendre type or a Legendre function if the following properties are satisfied:

(L1) Θ is non-empty,

(L2) ψ is strictly convex and differentiable on Θ,

(L3) $\forall \boldsymbol{\theta}_b \in \text{bd}(\Theta)$, $\lim_{\boldsymbol{\theta} \to \boldsymbol{\theta}_b} \|\nabla \psi(\boldsymbol{\theta})\| \to \infty$ where $\boldsymbol{\theta} \in \Theta$.

Based on this definition, we now state a critical property of the cumulant function of any regular exponential family.

Lemma 1 *Let ψ be the cumulant function of a regular exponential family with natural parameter space $\Theta = \text{dom}(\psi)$. Then ψ is a proper, closed convex function with $\text{int}(\Theta) = \Theta$ and (Θ, ψ) is a convex function of Legendre type.*

The above result directly follows from Theorems 8.2, 9.1 and 9.3 of Barndorff-Nielsen (1978).

4.2 Expectation Parameters and Legendre Duality

Consider a d-dimensional real random vector X distributed according to a regular exponential family density $p_{(\psi,\boldsymbol{\theta})}$ specified by the natural parameter $\boldsymbol{\theta} \in \Theta$. The expectation of X with respect to $p_{(\psi,\boldsymbol{\theta})}$, also called the *expectation parameter*, is given by

$$\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta}) = E_{p_{(\psi,\boldsymbol{\theta})}}[X] = \int_{\mathbb{R}^d} \mathbf{x} p_{(\psi,\boldsymbol{\theta})}(\mathbf{x}) d\mathbf{x}. \tag{7}$$

8. A convex function ψ is proper if $\text{dom}(\psi)$ is non-empty and $\forall \mathbf{x} \in \text{dom}(\psi), \psi(\mathbf{x}) > -\infty$. A convex function is closed if it is lower semi-continuous.

It can be shown (Barndorff-Nielsen, 1978; Amari, 1995) 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 conjugate functions.

Definition 5 (Rockafellar (1970)) Let ψ be a real-valued function on $\mathbb{R}^d$. Then its *conjugate function* ψ^* is given by

$$\psi^*(\mathbf{t}) = \sup_{\boldsymbol{\theta} \in \mathrm{dom}(\psi)} \{\langle \mathbf{t}, \boldsymbol{\theta} \rangle - \psi(\boldsymbol{\theta})\}. \tag{8}$$

Further, if ψ is a proper closed convex function, ψ^* is also a proper closed convex function and $\psi^{**} = \psi$.

When ψ is strictly convex and differentiable over $\Theta = \mathrm{int}(\mathrm{dom}(\psi))$, we can obtain the unique $\boldsymbol{\theta}^\dagger$ that corresponds to the supremum in (8) by setting the gradient of $\langle \mathbf{t}, \boldsymbol{\theta} \rangle - \psi(\boldsymbol{\theta})$ to zero, i.e.,

$$\nabla(\langle \mathbf{t}, \boldsymbol{\theta} \rangle - \psi(\boldsymbol{\theta}))\,|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\dagger} = 0 \qquad \Rightarrow \qquad \mathbf{t} = \nabla\psi(\boldsymbol{\theta}^\dagger)\ . \tag{9}$$

The strict convexity of ψ implies that $\nabla\psi$ is monotonic and it is possible to define the inverse function $(\nabla\psi)^{-1} : \Theta^* \mapsto \Theta$, where $\Theta^* = \mathrm{int}(\mathrm{dom}(\psi^*))$. If the pair (Θ, ψ) is of Legendre type, then it can be shown (Rockafellar, 1970) that (Θ^*, ψ^*) is also of Legendre type, and (Θ, ψ) and (Θ^*, ψ^*) are called Legendre duals of each other. Further, the gradient mappings are continuous and form a bijection between the two open sets Θ and Θ^*. The relation between (Θ, ψ) and (Θ^*, ψ^*) result is formally stated below.

Theorem 2 (Rockafellar (1970)) *Let ψ be a real-valued proper closed convex function with conjugate function ψ^*. Let $\Theta = \mathrm{int}(\mathrm{dom}(\psi))$ and $\Theta^* = \mathrm{int}(\mathrm{dom}(\psi^*))$. If (Θ, ψ) is a convex function of Legendre type, then*

(i) *(Θ^*, ψ^*) is a convex function of Legendre type,*

(ii) *(Θ, ψ) and (Θ^*, ψ^*) are Legendre duals of each other,*

(iii) *The gradient function $\nabla\psi : \Theta \mapsto \Theta^*$ is a one-to-one function from the open convex set Θ onto the open convex set Θ^*,*

(iv) *The gradient functions $\nabla\psi, \nabla\psi^*$ are continuous, and $\nabla\psi^* = (\nabla\psi)^{-1}$.*

Let us now look at the relationship between the natural parameter $\boldsymbol{\theta}$ and the expectation parameter $\boldsymbol{\mu}$ defined in (7). 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 ϕ be defined as the conjugate of ψ, i.e.,

$$\phi(\boldsymbol{\mu}) = \sup_{\boldsymbol{\theta} \in \Theta} \{\langle \boldsymbol{\mu}, \boldsymbol{\theta} \rangle - \psi(\boldsymbol{\theta})\}. \tag{10}$$

Since (Θ, ψ) is a convex function of Legendre type (Lemma 1), the pairs (Θ, ψ) and $(\mathrm{int}(\mathrm{dom}(\phi)), \phi)$ are Legendre duals of each other from Theorem 2, i.e., $\phi = \psi^*$ and $\mathrm{int}(\mathrm{dom}(\phi)) = \Theta^*$. Thus, the mappings between the dual spaces $\mathrm{int}(\mathrm{dom}(\phi))$ and Θ are given by the Legendre transformation

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \nabla\psi(\boldsymbol{\theta}) \qquad \text{and} \qquad \boldsymbol{\theta}(\boldsymbol{\mu}) = \nabla\phi(\boldsymbol{\mu})\ . \tag{11}$$

Further, the conjugate function ϕ can be expressed as

$$\phi(\boldsymbol{\mu}) = \langle \boldsymbol{\theta}(\boldsymbol{\mu}), \boldsymbol{\mu} \rangle - \psi(\boldsymbol{\theta}(\boldsymbol{\mu})), \ \ \forall \boldsymbol{\mu} \in \mathrm{int}(\mathrm{dom}(\phi))\ . \tag{12}$$

4.3 Exponential Families and Bregman Divergences

We are now ready to explicitly state the formal connection between exponential families of distributions and Bregman divergences. It has been observed in the literature that exponential families and Bregman divergences have a close relationship that can be exploited for several learning problems. In particular, Forster and Warmuth (2000)[Section 5.1] remarked that the log-likelihood of the density of an exponential family distribution $p_{(\psi,\boldsymbol{\theta})}$ can be written as the sum of the negative of a uniquely determined Bregman divergence $d_\phi(\mathbf{x},\boldsymbol{\mu})$ and a function that does not depend on the distribution parameters. In our notation, this can be written as

$$\log(p_{(\psi,\boldsymbol{\theta})}(\mathbf{x})) = -d_\phi(\mathbf{x},\boldsymbol{\mu}(\boldsymbol{\theta})) + \log(b_\phi(\mathbf{x}))\ , \tag{13}$$

where ϕ is the conjugate function of ψ and $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta}) = \nabla\psi(\boldsymbol{\theta})$ is the expectation parameter corresponding to $\boldsymbol{\theta}$. The result was later used by Collins et al. (2001) to extend PCA to exponential families. However, as we explain below, a formal proof is required to show that (13) holds for all instances $\mathbf{x}$ of interest. We focus on the case when $p_{(\psi,\boldsymbol{\theta})}$ is a *regular* exponential family.

To get an intuition of the main result, observe that the log-likelihood of any exponential family, considering only the parametric terms, can be written as

$$\begin{aligned}\langle \mathbf{x},\boldsymbol{\theta}\rangle - \psi(\boldsymbol{\theta}) &= (\langle \boldsymbol{\mu},\boldsymbol{\theta}\rangle - \psi(\boldsymbol{\theta})) + \langle \mathbf{x}-\boldsymbol{\mu},\boldsymbol{\theta}\rangle \\ &= \phi(\boldsymbol{\mu}) + \langle \mathbf{x}-\boldsymbol{\mu},\nabla\phi(\boldsymbol{\mu})\rangle\ ,\end{aligned}$$

from (11) and (12), where $\boldsymbol{\mu} \in \text{int}(\text{dom}(\phi))$. Therefore, for any $\mathbf{x} \in \text{dom}(\phi)$, $\boldsymbol{\theta} \in \Theta$, and $\boldsymbol{\mu} \in \text{int}(\text{dom}(\phi))$, one can write

$$\langle \mathbf{x},\boldsymbol{\theta}\rangle - \psi(\boldsymbol{\theta}) \ = \ -d_\phi(\mathbf{x},\boldsymbol{\mu}) + \phi(\mathbf{x})\ .$$

Then considering the density of an exponential family with respect to the appropriate measure $d\mathbf{x}$, we have

$$\log(p_{(\psi,\boldsymbol{\theta})}(\mathbf{x})) \ = \ \langle \mathbf{x},\boldsymbol{\theta}\rangle - \psi(\boldsymbol{\theta}) + \log p_0(\mathbf{x}) \ = \ -d_\phi(\mathbf{x},\boldsymbol{\mu}) + \log(b_\phi(\mathbf{x}))\ ,$$

where $b_\phi(\mathbf{x}) = \exp(\phi(\mathbf{x}))p_0(\mathbf{x})$.

Thus (13) follows directly from Legendre duality for $\mathbf{x} \in \text{dom}(\phi)$. However, for (13) to be useful, one would like to ensure that it is true for all individual "instances" $\mathbf{x}$ that can be drawn following the exponential distribution $p_{(\psi,\boldsymbol{\theta})}$. Let I_ψ denote the set of such instances. Establishing (13) can be tricky for all $\mathbf{x} \in I_\psi$ since the relationship between I_ψ and $\text{dom}(\phi)$ is not apparent. Further, there are distributions for which the instances space I_ψ and the expectation parameter space $\text{int}(\text{dom}(\phi))$ are disjoint, as the following example shows.

Example 8 A Bernoulli random variable X takes values in $\{0,1\}$ such that $p(X=1) = q$ and $p(X=0) = 1-q$, for some $q \in [0,1]$. The instance space for X is just $I_\psi = \{0,1\}$. The cumulant function for X is $\psi(\theta) = \log(1+\exp(\theta))$ with $\Theta = \mathbb{R}$ (see Table 2). A simple calculation shows that the conjugate function $\phi(\mu) = \mu\log\mu + (1-\mu)\log(1-\mu)$, $\forall \mu \in (0,1)$. Since ϕ is a closed function, we obtain $\phi(\mu) = 0$ for $\mu \in \{0,1\}$ by taking limits. Thus, the effective domain of ϕ is $[0,1]$ and $\mu = q$, whereas the expectation parameter space is given by $\text{int}(\text{dom}(\phi)) = (0,1)$. Hence the instance space I_ψ and the expectation parameter space $\text{int}(\text{dom}(\phi))$ are disjoint; however $I_\psi \subset \text{dom}(\phi)$.

In this particular case, since the "instances" lie within $\mathrm{dom}(\phi)$, the relation (13) does hold for all $\mathbf{x} \in I_\psi$. However, it remains to be shown that $I_\psi \subseteq \mathrm{dom}(\phi)$ for all regular exponential family distributions.

In order to establish such a result for all regular exponential family distributions, we need to formally define the set of instances I_ψ. If the measure P_0 is absolutely continuous with respect to the counting measure, then $\mathbf{x} \in I_\psi$ if $p_{(\psi,\boldsymbol{\theta})}(\mathbf{x}) > 0$. On the other hand, if P_0 is absolutely continuous with respect to the Lebesgue measure, then $\mathbf{x} \in I_\psi$ if all sets with positive Lebesgue measure that contain $\mathbf{x}$ have positive probability mass. A closer look reveals that the set of instances I_ψ is independent of the choice of $\boldsymbol{\theta}$. In fact, I_ψ is just the support of P_0 and can be formally defined as follows.

Definition 6 Let I_ψ denote the set of instances that can be drawn following $p_{(\psi,\boldsymbol{\theta})}(\mathbf{x})$. Then, $\mathbf{x}_0 \in I_\psi$ if $\forall I$ such that $\mathbf{x}_0 \in I$ and $\int_I d\mathbf{x} > 0$, we have $\int_I dP_0(\mathbf{x}) > 0$, where P_0 is as defined in Section 4.1; also see footnote 6.

The following theorem establishes the crucial result that the set of instances I_ψ is always a subset of $\mathrm{dom}(\phi)$.

Theorem 3 *Let I_ψ be the set of instances as in Definition 6. Then, $I_\psi \subseteq \mathrm{dom}(\phi)$ where ϕ is the conjugate function of ψ.*

The above result follows from Theorem 9.1 and related results in Barndorff-Nielsen (1978). We have included the proof in Appendix C.

We are now ready to formally show that there is a unique Bregman divergence corresponding to every regular exponential family distribution. Note that, by Theorem 3, it is sufficient to establish the relationship for all $\mathbf{x} \in \mathrm{dom}(\phi)$.

Theorem 4 *Let $p_{(\psi,\boldsymbol{\theta})}$ be the probability density function of a regular exponential family distribution. Let ϕ be the conjugate function of ψ so that $(\mathrm{int}(\mathrm{dom}(\phi)), \phi)$ is the Legendre dual of (Θ, ψ). Let $\boldsymbol{\theta} \in \Theta$ be the natural parameter and $\boldsymbol{\mu} \in \mathrm{int}(\mathrm{dom}(\phi))$ be the corresponding expectation parameter. Let d_ϕ be the Bregman divergence derived from ϕ. Then $p_{(\psi,\boldsymbol{\theta})}$ can be uniquely expressed as*

$$p_{(\psi,\boldsymbol{\theta})}(\mathbf{x}) = \exp(-d_\phi(\mathbf{x},\boldsymbol{\mu}))b_\phi(\mathbf{x}), \qquad \forall \mathbf{x} \in \mathrm{dom}(\phi) \tag{14}$$

where $b_\phi : \mathrm{dom}(\phi) \mapsto \mathbb{R}_+$ is a uniquely determined function.

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

$$\begin{aligned} p_{(\psi,\boldsymbol{\theta})}(\mathbf{x}) &= \exp(\langle \mathbf{x}, \boldsymbol{\theta} \rangle - \psi(\boldsymbol{\theta}))p_0(\mathbf{x}) \\ &= \exp(\phi(\boldsymbol{\mu}) + \langle \mathbf{x} - \boldsymbol{\mu}, \nabla\phi(\boldsymbol{\mu}) \rangle)p_0(\mathbf{x}) \qquad \text{(using (11) and (12))} \\ &= \exp(-\{\phi(\mathbf{x}) - \phi(\boldsymbol{\mu}) - \langle \mathbf{x} - \boldsymbol{\mu}, \nabla\phi(\boldsymbol{\mu}) \rangle\} + \phi(\mathbf{x}))p_0(\mathbf{x}) \\ &= \exp(-d_\phi(\mathbf{x},\boldsymbol{\mu}))b_\phi(\mathbf{x}) \ , \end{aligned}$$

where $b_\phi(\mathbf{x}) = \exp(\phi(\mathbf{x}))p_0(\mathbf{x}))$.

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, i.e., the expectation parameter $\boldsymbol{\mu} = \nabla\psi(\boldsymbol{\theta})$ corresponding to $\boldsymbol{\theta}$ is uniquely determined 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 since constant additive terms do not change the corresponding Bregman divergence (Appendix A, Property 4). The Legendre duality between ϕ and ψ also ensures that no two exponential families correspond to the same Bregman divergence, i.e., the mapping is one-to-one. Further, since $p_{(\psi,\boldsymbol{\theta})}(\mathbf{x})$ is well-defined on $\mathrm{dom}(\phi)$, and the corresponding $d_\phi(\mathbf{x},\boldsymbol{\mu})$ is unique, the function $b_\phi(\mathbf{x}) = \exp(d_\phi(\mathbf{x},\boldsymbol{\mu}))p_{(\psi,\boldsymbol{\theta})}(\mathbf{x})$ is uniquely determined. ■

4.4 Bijection with Regular Bregman Divergences

From Theorem 4 we note that every regular exponential family corresponds to a unique and distinct Bregman divergence (one-to-one mapping). Now, we investigate whether there is a regular exponential family corresponding to every choice of Bregman divergence (onto mapping).

For regular exponential families, the cumulant function ψ as well as its conjugate ϕ are convex functions of Legendre type. Hence, for a Bregman divergence generated from a convex function ϕ to correspond to a regular exponential family, it is necessary that ϕ be of Legendre type. Further, it is necessary that the Legendre conjugate ψ of ϕ to be C^∞, since cumulant functions of regular exponential families are C^∞. However, it is not clear if these conditions are sufficient. Instead, we provide a sufficiency condition using exponentially convex functions (Akhizer, 1965; Ehm et al., 2003), which are defined below.

Definition 7 A function $f : \Theta \mapsto \mathbb{R}_{++}$, $\Theta \subseteq \mathbb{R}^d$ is called exponentially convex if the kernel $K_f(\alpha,\beta) = f(\alpha+\beta)$, with $\alpha+\beta \in \Theta$, satisfies

$$\sum_{i=1}^{n}\sum_{j=1}^{n} K_f(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j)u_i\bar{u}_j \geq 0$$

for any set $\{\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_n\} \subseteq \Theta$ with $\boldsymbol{\theta}_i+\boldsymbol{\theta}_j \in \Theta$, $\forall i, j$, and $\{u_1,\cdots,u_n\} \subset \mathbb{C}$ ($\bar{u}_j$ denotes the complex conjugate of u_j), i.e, the kernel K_f is positive semi-definite.

Although it is well known that the logarithm of an exponentially convex function is a convex function (Akhizer, 1965), we are interested in the case where the logarithm is strictly convex with an open domain. Using this class of exponentially convex functions, we now define a class of Bregman divergences called *regular Bregman divergences*.

Definition 8 Let $f : \Theta \mapsto \mathbb{R}_{++}$ be a continuous exponentially convex function such that Θ is open and $\psi(\boldsymbol{\theta}) = \log(f(\boldsymbol{\theta}))$ is strictly convex. Let ϕ be the conjugate function of ψ. Then we say that the Bregman divergence d_ϕ derived from ϕ is a *regular Bregman divergence*.

We will now prove that there is a bijection between regular exponential families and regular Bregman divergences. The crux of the argument relies on results in harmonic analysis connecting positive definiteness to integral transforms (Berg et al., 1984). In particular, we use a result due to Devinatz (1955) that relates exponentially convex functions to Laplace transforms of bounded non-negative measures.

Theorem 5 (Devinatz (1955)) *Let $\Theta \subseteq \mathbb{R}^d$ be an open convex set. A necessary and sufficient condition that there exists a unique, bounded, non-negative measure ν such that $f : \Theta \mapsto \mathbb{R}_{++}$ can be*

represented as

$$f(\boldsymbol{\theta}) = \int_{\mathbf{x}\in\mathbb{R}^d} \exp(\langle \mathbf{x}, \boldsymbol{\theta}\rangle) d\nu(\mathbf{x}) \tag{15}$$

is that f is continuous and exponentially convex.

We also need the following result to establish the bijection.

Lemma 2 *Let ψ be the cumulant of an exponential family with base measure P_0 and natural parameter space $\Theta \subseteq \mathbb{R}^d$. Then, if P_0 is concentrated on an affine subspace of $\mathbb{R}^d$ then ψ is not strictly convex.*

Proof Let $P_0(\mathbf{x})$ be concentrated on an affine subspace $S = \{\mathbf{x} \in \mathbb{R}^d | \langle \mathbf{x}, \mathbf{b}\rangle = c\}$ for some $\mathbf{b} \in \mathbb{R}^d$ and $c \in \mathbb{R}$. Let $I = \{\boldsymbol{\theta} | \boldsymbol{\theta} = \alpha\mathbf{b},\ \alpha \in \mathbb{R}\}$. Then, for any $\boldsymbol{\theta} = \alpha\mathbf{b} \in I$, we have $\langle \mathbf{x}, \boldsymbol{\theta}\rangle = \alpha c\ \forall \mathbf{x} \in S$ and the cumulant is given by

$$\begin{aligned}
\psi(\boldsymbol{\theta}) &= \log\left(\int_{\mathbf{x}\in\mathbb{R}^d} \exp(\langle \mathbf{x}, \boldsymbol{\theta}\rangle) dP_0(\mathbf{x})\right) = \log\left(\int_{\mathbf{x}\in S} \exp(\langle \mathbf{x}, \boldsymbol{\theta}\rangle) dP_0(\mathbf{x})\right) \\
&= \log\left(\int_{\mathbf{x}\in S} \exp(\alpha c) dP_0(\mathbf{x})\right) = \log(\exp(\alpha c) P_0(S)) = \alpha c + \log(P_0(S)) \\
&= \langle \mathbf{x}_0, \boldsymbol{\theta}\rangle + \log(P_0(S)) ,
\end{aligned}$$

for any $\mathbf{x}_0 \in S$, implying that ψ is not strictly convex. ∎

There are two parts to the proof leading to the bijection result. Note that we have already established in Theorem 4 that there is a unique Bregman divergence corresponding to every exponential family distribution. In the first part of the proof, we show that these Bregman divergences are regular (one-to-one). Then we show that there exists a unique regular exponential family determined by every regular Bregman divergence (onto).

Theorem 6 *There is a bijection between regular exponential families and regular Bregman divergences.*

Proof First we prove the 'one-to-one' part, i.e., there is a regular Bregman divergence corresponding to every regular exponential family F_ψ with cumulant function ψ and natural parameter space Θ. Since $\mathcal{F}_\psi$ is a regular exponential family, there exists a non-negative bounded measure ν such that for all $\boldsymbol{\theta} \in \Theta$,

$$\begin{aligned}
1 &= \int_{\mathbf{x}\in\mathbb{R}^d} \exp(\langle \mathbf{x}, \boldsymbol{\theta}\rangle - \psi(\boldsymbol{\theta})) d\nu(\mathbf{x}) \\
\Rightarrow \qquad \exp(\psi(\boldsymbol{\theta})) &= \int_{\mathbf{x}\in\mathbb{R}^d} \exp(\langle \mathbf{x}, \boldsymbol{\theta}\rangle) d\nu(\mathbf{x}).
\end{aligned}$$

Thus, from Theorem 5, $\exp(\psi(\boldsymbol{\theta}))$ is a continuous exponentially convex function with the open set Θ as its domain. Further, being the cumulant of a regular exponential family, ψ is strictly convex. Therefore, the Bregman divergence d_ϕ derived from the conjugate function ϕ of ψ is a regular Bregman divergence.

Next we prove the 'onto' part, i.e., every regular Bregman divergence corresponds to a unique regular exponential family. Let the regular Bregman divergence d_ϕ be generated by ϕ and let ψ be

the conjugate of ϕ. Since d_ϕ is a regular Bregman divergence, by Definition 8, ψ is strictly convex with $\mathrm{dom}(\psi) = \Theta$ being an open set. Further, the function $\exp(\psi(\boldsymbol{\theta}))$ is a continuous, exponentially convex function. From Theorem 5, there exists a unique non-negative bounded measure ν that satisfies (15). Since Θ is non-empty, we can choose some fixed $\mathbf{b} \in \Theta$ so that

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

and so $dP_0(\mathbf{x}) = \exp(\langle \mathbf{x}, \mathbf{b}\rangle - \psi(\mathbf{b}))d\nu(\mathbf{x})$ is a probability density function. The set of all $\boldsymbol{\theta} \in \mathbb{R}^d$ for which

$$\int_{\mathbf{x}\in\mathbb{R}^d} \exp(\langle \mathbf{x}, \boldsymbol{\theta}\rangle) dP_0(\mathbf{x}) < \infty$$

is same as the set $\{\boldsymbol{\theta} \in \mathbb{R}^d | \exp(\psi(\boldsymbol{\theta}+\mathbf{b}) - \psi(\mathbf{b})) < \infty\} = \{\boldsymbol{\theta} \in \mathbb{R}^d | \boldsymbol{\theta}+\mathbf{b} \in \Theta\}$ which is just a translated version of Θ itself. For any $\boldsymbol{\theta}$ such that $\boldsymbol{\theta}+\mathbf{b} \in \Theta$, we have

$$\int_{\mathbf{x}\in\mathbb{R}^d} \exp(\langle \mathbf{x}, \boldsymbol{\theta}+\mathbf{b}\rangle - \psi(\boldsymbol{\theta}+\mathbf{b})) d\nu(\mathbf{x}) = 1 \ .$$

Hence, the exponential family $\mathcal{F}_\psi$ consisting of densities of the form

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

with respect to the measure ν has Θ as its natural parameter space and $\psi(\boldsymbol{\theta})$ as the cumulant function.

Since ψ is strictly convex on Θ, it follows from Lemma 2 that the measure P_0 is not concentrated in an affine subspace of $\mathbb{R}^d$, i.e., $\mathbf{x}$ is a minimal statistic for $\mathcal{F}_\psi$. Therefore, the exponential family generated by P_0 and $\mathbf{x}$ is full. Since Θ is also open, it follows that $\mathcal{F}_\psi$ is a regular exponential family.

Finally we show that the family is unique. Since only d_ϕ is given, the generating convex function could be $\bar{\phi}(\mathbf{x}) = \phi(\mathbf{x}) + \langle \mathbf{x}, \mathbf{a}\rangle + c$ for $\mathbf{a} \in \mathbb{R}^d$ and a constant $c \in \mathbb{R}$. The corresponding conjugate function $\bar{\psi}(\boldsymbol{\theta}) = \psi(\boldsymbol{\theta} - \mathbf{a}) - c$ differs from ψ only by a constant. Hence, the exponential family is exactly $\mathcal{F}_\psi$. That completes the proof. ■

4.5 Examples

Table 2 shows the various functions of interest for some popular exponential families. We now look at two of these distributions in detail and obtain the corresponding Bregman divergences.

Example 9 The most well-known exponential family is that of Gaussian distributions, in particular uniform variance, spherical Gaussian distributions with densities of the form

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

Table 2: Various functions of interest for some popular exponential families. For all the cases shown in the table, $\mathbf{x}$ is the sufficient statistic. Note that for the Gaussian examples the variance σ is assumed to be constant. The number of trials, N, for the binomial and multinomial examples is also assumed to be constant.

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	$\frac{\lambda^x\exp(-\lambda)}{x!}$	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	$\lambda\exp(-\lambda x)$	$1/\lambda$	$-\log\mu-1$	$\frac{x}{\mu}-\log\left(\frac{x}{\mu}\right)-1$
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	$\frac{N!}{\prod_{j=1}^d x_j!}\prod_{j=1}^d q_j^{x_j}$	$[Nq_j]_{j=1}^{d-1}$	$\sum_{j=1}^d \mu_j\log(\frac{\mu_j}{N})$	$\sum_{j=1}^d x_j\log(\frac{x_j}{\mu_j})$

Distribution	$\boldsymbol{\theta}$	$\psi(\boldsymbol{\theta})$	dom(ψ)	dom(ϕ)	I_ψ
1-D Gaussian	$\frac{a}{\sigma^2}$	$\frac{\sigma^2}{2}\theta^2$	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$
1-D Poisson	$\log\lambda$	$\exp(\theta)$	$\mathbb{R}$	$\mathbb{R}_+$	$\mathbb{N}$
1-D Bernoulli	$\log(\frac{q}{1-q})$	$\log(1+\exp(\theta))$	$\mathbb{R}$	$[0,1]$	$\{0,1\}$
1-D Binomial	$\log(\frac{q}{1-q})$	$N\log(1+\exp(\theta))$	$\mathbb{R}$	$[0,N]$	$\{0,1,\ldots,N\}$
1-D Exponential	$-\lambda$	$-\log(-\theta)$	$\mathbb{R}_{--}$	$\mathbb{R}_{++}$	$\mathbb{R}_{++}$
d-D Sph. Gaussian	$\frac{\mathbf{a}}{\sigma^2}$	$\frac{\sigma^2}{2}\|\boldsymbol{\theta}\|^2$	$\mathbb{R}^d$	$\mathbb{R}^d$	$\mathbb{R}^d$
d-D Multinomial	$[\log(\frac{q_j}{q_d})]_{j=1}^{d-1}$	$N\log(1+\sum_{j=1}^{d-1}\exp(\theta_j))$	$\mathbb{R}^{d-1}$	$\{\boldsymbol{\mu}\in\mathbb{R}_+^{d-1},\|\boldsymbol{\mu}\|\le N\}$	$\{\mathbf{x}\in\mathbb{Z}_+^{d-1},\|\mathbf{x}\|\le N\}$

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

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

where $p_0(\mathbf{x})$ is independent of $\boldsymbol{\theta}$. By (11), the expectation parameter for this distribution is given by

$$\boldsymbol{\mu}=\nabla\psi(\boldsymbol{\theta})=\nabla\left(\frac{\sigma^2}{2}\|\boldsymbol{\theta}\|^2\right)=\sigma^2\boldsymbol{\theta}=\mathbf{a}\ .$$

By using (12), the Legendre dual ϕ of ψ is

$$\phi(\boldsymbol{\mu})=\langle\boldsymbol{\mu},\boldsymbol{\theta}\rangle-\psi(\boldsymbol{\theta})=\left\langle\boldsymbol{\mu},\frac{\boldsymbol{\mu}}{\sigma^2}\right\rangle-\frac{\sigma^2}{2}\|\boldsymbol{\theta}\|^2=\frac{\|\boldsymbol{\mu}\|^2}{2\sigma^2}\ .$$

The corresponding Bregman divergence equals

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

The function $b_\phi(\mathbf{x})$ in Theorem 4 is given by

$$b_\phi(\mathbf{x}) = \exp(\phi(\mathbf{x}))p_0(\mathbf{x}) = \exp\left(\frac{\|\mathbf{x}\|^2}{2\sigma^2} - \frac{\|\mathbf{x}\|^2}{2\sigma^2}\right)\frac{1}{\sqrt{(2\pi\sigma^2)^d}} = \frac{1}{\sqrt{(2\pi\sigma^2)^d}}\,,$$

and turns out to be a constant. Thus, $p_{(\psi,\boldsymbol{\theta})}(\mathbf{x}) = \exp(-d_\phi(\mathbf{x},\boldsymbol{\mu}))b_\phi(\mathbf{x})$.

Example 10 Another exponential family that is widely used is the family of multinomial distributions:

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

where $x_j \in \mathbb{Z}_+$ are frequencies of events, $\sum_{j=1}^d x_j = N$ and $q_j \geq 0$ are probabilities of events, $\sum_{j=1}^d q_j = 1$. As shown below, $p(\mathbf{x};\mathbf{q})$ 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 $\psi(\boldsymbol{\theta}) = -N\log q_d = N\log(1+\sum_{j=1}^{d-1} e^{\theta_j})$.

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

where $p_0(\mathbf{x})$ is independent of $\boldsymbol{\theta}$. The expectation parameter $\boldsymbol{\mu}$ is given by

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

and the Legendre dual ϕ of ψ is

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

where $\mu_d = N q_d$ so that $\sum_{i=1}^{d} \mu_j = N$. Note that $\phi(\boldsymbol{\mu})$ 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, i.e.,

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

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

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

and $p_{(\psi,\boldsymbol{\theta})}(\mathbf{x}) = \exp(-d_\phi(\mathbf{x}, \boldsymbol{\mu})) b_\phi(\mathbf{x})$.

5. Bregman Soft Clustering

Using the correspondence between regular exponential families and regular Bregman divergences, we now pose the Bregman soft clustering problem as a parameter estimation problem for mixture models based on regular exponential family distributions. We revisit the expectation maximization (EM) framework for estimating mixture densities and develop a Bregman soft clustering algorithm (Algorithm 3) for regular Bregman divergences. We also present the Bregman soft clustering algorithm for a set of data points with non-uniform non-negative weights (or measure). Finally, we show how the hard clustering algorithm can be interpreted as a special case of the soft clustering algorithm and also discuss an alternative formulation of hard clustering in terms of a dual divergence derived from the conjugate function.

5.1 Soft Clustering as Mixture Density Estimation

Given a set $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{n} \subset \mathbb{R}^d$ drawn independently from a stochastic source, consider the problem of modeling the source using a single parametric exponential family distribution. This is the problem of maximum likelihood estimation, or, equivalently, minimum negative log-likelihood estimation of the parameter(s) of a given exponential family distribution. From Theorem 4, minimizing the negative log-likelihood is the same as minimizing the corresponding expected Bregman divergence. Using Proposition 1, we conclude that the optimal distribution is the one with $\boldsymbol{\mu} = \mathbf{E}[X]$ as the expectation parameter, where X is a random variable that takes values in $\mathcal{X}$ following (by the

independence assumption) the empirical distribution over $\mathcal{X}$. Further, note that the minimum negative log-likelihood of $\mathcal{X}$ 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 conjugate of ψ.

Now, consider the problem of modeling the stochastic source with a mixture of k densities of the same exponential family. The model 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. For regular Bregman divergences, we define the *Bregman soft clustering problem* as that of learning the maximum likelihood parameters $\Gamma = \{\boldsymbol{\theta}_h, \pi_h\}_{h=1}^k \equiv \{\boldsymbol{\mu}_h, \pi_h\}_{h=1}^k$ of a mixture model of the form

$$p(\mathbf{x}|\Gamma) = \sum_{h=1}^{k} \pi_h p_{(\psi,\boldsymbol{\theta}_h)}(\mathbf{x}) = \sum_{h=1}^{k} \pi_h \exp(-d_\phi(\mathbf{x}, \boldsymbol{\mu}_h)) b_\phi(\mathbf{x}), \tag{16}$$

where the last equality follows from Theorem 4. Since the mixture components are all assumed to be from the same family, 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.

5.2 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, $L_\mathcal{X}(\Gamma)$ 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 more details, the reader is referred to Collins (1997); McLachlan and Krishnan (1996) and Bilmes (1997).

The Bregman soft clustering problem is to estimate the maximum likelihood parameters for the mixture model given in (16). Using the Bregman divergence viewpoint, we get a simplified version of the above EM algorithm that we call the Bregman soft clustering algorithm (Algorithm 3). Using Proposition 1, the computationally intensive M-step turns out to be straightforward to solve. In fact, the Bregman divergence viewpoint gives an alternative interpretation of a well known efficient EM scheme applicable to learning a mixture of exponential distributions (Redner and Walker, 1984). The resulting update equations are similar to those for learning mixture models of identity covariance Gaussians. Note that these equations are applicable to mixtures of any regular exponential distributions, as long as $\mathbf{x}$ is the (minimal) sufficient statistic vector.

It is important to note that the simplification of the M-step is applicable only when the parameterization is with respect to the expectation parameter space, i.e., when d_ϕ corresponding to an exponential family is known. Otherwise, if the parameterization is with respect to the natural parameter space, i.e., the functional form for a family is known in terms of its cumulant ψ and natural parameters $\boldsymbol{\theta}$, the problem

$$\phi(\mathbf{x}) = \sup_{\boldsymbol{\theta} \in \mathbb{R}^d} \left(\langle \boldsymbol{\theta}, \mathbf{x} \rangle - \psi(\boldsymbol{\theta}) \right), \tag{17}$$

needs to be solved to obtain $\phi(\mathbf{x})$. Since the function to be maximized in (17) is precisely the log-likelihood of the exponential family density (with respect to an appropriate measure), the transformation is equivalent to solving a maximum likelihood estimation problem (with a single sample), which is computationally expensive for several exponential family distributions. In such a situation, transforming the problem to the expectation space need not lead to any tangible computational bene-

Algorithm 2 Standard EM for Mixture Density Estimation

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

Output: $\Gamma^\dagger$: local maximizer of $L_\mathcal{X}(\Gamma) = \prod_{i=1}^n (\sum_{h=1}^k \pi_h p_{\psi,\boldsymbol{\theta}_h}(\mathbf{x}_i))$ where $\Gamma = \{\boldsymbol{\theta}_h, \pi_h\}_{h=1}^k$, 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 $\boldsymbol{\theta}_h \in \Theta$, and $\pi_h \geq 0$, $\sum_{h=1}^k \pi_h = 1$

repeat

{The Expectation Step (E-step)}

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

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

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

end for

end for

{The Maximization Step (M-step)}

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

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

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

end for

until *convergence*

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

Algorithm 3 Bregman Soft Clustering

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

Output: $\Gamma^\dagger$, local maximizer of $\prod_{i=1}^n (\sum_{h=1}^k \pi_h b_\phi(\mathbf{x}_i) \exp(-d_\phi(\mathbf{x}_i, \boldsymbol{\mu}_h)))$ where $\Gamma = \{\boldsymbol{\mu}_h, \pi_h\}_{h=1}^k$, 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 $\boldsymbol{\mu}_h \in \mathrm{ri}(\mathcal{S}), \pi_h \geq 0$, and $\sum_{h=1}^k \pi_h = 1$

repeat

{The Expectation Step (E-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 (M-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 $\Gamma^\dagger = \{\boldsymbol{\mu}_h, \pi_h\}_{h=1}^k$

fits. However, if the Bregman divergence d_ϕ corresponding to an exponential family is either known or easy to compute from the natural parameterization, then Algorithm 3 is computationally much more efficient. In fact, in some situations it may be easier to design regular Bregman divergences for mixture modeling of data than to come up with an appropriate exponential family. Such situations can take full advantage of the computationally efficient Bregman soft clustering algorithm.

The following result shows how Proposition 1 and Theorem 4 can be used to simplify the M-step of Algorithm 2. Using this result, we then show that Algorithms 2 and 3 are exactly equivalent for regular Bregman divergences and exponential families. Note that Proposition 4 has appeared in various forms in the literature (see, for example, Redner and Walker (1984); McLachlan and Krishnan (1996)). We give an alternative proof using Bregman divergences.

Proposition 4 *For a mixture model with density given by (16), the maximization step for the density parameters in the EM algorithm (Algorithm 2), $\forall h, 1 \leq h \leq k$, reduces to:*

$$\boldsymbol{\mu}_h = \frac{\sum_{i=1}^n p(h|\mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^n p(h|\mathbf{x}_i)}. \tag{18}$$

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

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

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

$$p_{(\psi,\boldsymbol{\theta}_h)}(\mathbf{x}) = b_\phi(\mathbf{x})\exp(-d_\phi(\mathbf{x},\boldsymbol{\mu}_h)).$$

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

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

From Proposition 1, we know that the expected Bregman divergence is minimized by the expectation of $\mathbf{x}$, i.e.,

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

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

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

■

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 $b_\phi(\mathbf{x})$ 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 (16), the EM algorithm (Algorithm 2) reduces to the Bregman soft clustering algorithm (Algorithm 3).

So far we have 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 ν_i with the individual samples such that $\sum_i \nu_i = 1$ 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 identical and the new update equations for the M-step, $\forall h, 1 \leq h \leq k$, are given by

$$\begin{aligned} \pi_h &= \sum_{i=1}^{n} \nu_i p(h|\mathbf{x}_i), \\ \boldsymbol{\mu}_h &= \frac{\sum_{i=1}^{n} \nu_i p(h|\mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^{n} \nu_i p(h|\mathbf{x}_i)}. \end{aligned}$$

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 $d_{\beta\phi} = \beta d_\phi$. In the limit, when $\beta \to \infty$, the posterior probabilities in the E-step take values in $\{0, 1\}$ and hence, the E and M steps of the soft clustering algorithm reduce to the assignment and re-estimation steps of the hard clustering algorithm.

5.3 An Alternative Formulation for Bregman Clustering

In earlier sections, the Bregman divergence was measured with the data points as the first argument and the cluster representative as the second argument. Since Bregman divergences are not symmetric (with the exception of squared Euclidean distance), we now consider an alternative formulation of Bregman clustering where cluster representatives are the first argument of the Bregman divergence. Using Legendre duality, we show that this alternate formulation is equivalent to our original Bregman clustering problem in a dual space using a different, but uniquely determined Bregman divergence.

We focus on the hard clustering case. Let X be a random variable that takes values in $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{n}$ following a positive probability measure ν. Then the alternative Bregman hard clustering problem is to find clusters $\{\mathcal{X}_h\}_{h=1}^{k}$ and corresponding representatives $\{\boldsymbol{\mu}_h\}_{h=1}^{k}$ that solve

$$\min_{\{\mu_h\}_{h=1}^{k}} \sum_{h=1}^{k} \sum_{\mathbf{x}_i \in \mathcal{X}_h} \nu_i d_\phi(\boldsymbol{\mu}_h, \mathbf{x}_i). \tag{19}$$

As mentioned earlier, Bregman divergences are convex in the first argument and hence, the resulting optimization problem for each cluster is convex so there is a unique optimal representative for each

cluster. However, unlike in the original formulation, the optimal cluster representative is not always the expectation and depends on the Bregman divergence d_ϕ.

It is interesting to note that this alternative formulation, though seemingly different, reduces to the original formulation with an appropriate representation. Let ϕ be the generating convex function of d_ϕ such that $(\text{int}(\text{dom}(\phi)), \phi)$ is a convex function of Legendre type and let $(\text{int}(\text{dom}(\psi)), \psi)$ be the corresponding Legendre dual. Then for any $\mathbf{x}, \mathbf{y} \in \text{int}(\text{dom}(\phi))$, the Bregman divergence $d_\phi(\mathbf{x}, \mathbf{y}) = d_\psi(\boldsymbol{\theta}_\mathbf{y}, \boldsymbol{\theta}_\mathbf{x})$ where d_ψ is the Bregman divergence derived from ψ and $\boldsymbol{\theta}_\mathbf{x} = \nabla\phi(\mathbf{x}), \boldsymbol{\theta}_\mathbf{y} = \nabla\phi(\mathbf{y})$ (Appendix A, Property 6). Using the above property, we can restate the alternative Bregman clustering problem in the dual space. More specifically, let $\mathcal{X}^{\boldsymbol{\theta}} = \{\boldsymbol{\theta}_{\mathbf{x}_i}\}_{i=1}^n$ where $\boldsymbol{\theta}_{\mathbf{x}_i} = \nabla\phi(\mathbf{x}_i), \forall \mathbf{x}_i, 1 \leq i \leq n$, and let $\boldsymbol{\theta}_h = \nabla\phi(\boldsymbol{\mu}_h), \forall \boldsymbol{\mu}_h, 1 \leq h \leq k$. Then the hard clustering problem (19) can be expressed as

$$\min_{\{\boldsymbol{\theta}_h\}_{h=1}^k} \sum_{h=1}^{k} \sum_{\boldsymbol{\theta}_{\mathbf{x}_i} \in \mathcal{X}_h^{\boldsymbol{\theta}}} \nu_i d_\psi(\boldsymbol{\theta}_{\mathbf{x}_i}, \boldsymbol{\theta}_h). \tag{20}$$

where $\mathcal{X}_h^{\boldsymbol{\theta}}$ correspond to cluster h in the dual space. It is now straightforward to see that this is our original Bregman hard clustering problem for the set $\mathcal{X}^{\boldsymbol{\theta}}$ consisting of the dual data points with the same measure ν and the dual Bregman divergence d_ψ. The optimal cluster representative in this dual space is given by the expectation, which is easy to compute. The efficiency of this approach is based on the same premise as the efficient EM scheme for exponential families, i.e, the M-step can be simplified if there is an easy transition to the dual space.

6. Lossy Compression and Generalized Loss Functions

In this section, we study the connection between Bregman clustering algorithms and lossy compression schemes. In particular, we focus on the relationship of our work with Shannon's rate distortion theory, showing connections between learning mixtures of exponential distributions, the Bregman soft clustering problem and the rate distortion problem where distortion is measured using a regular Bregman divergence (Banerjee et al., 2004a). Then we show that all these problems involve a trade-off between compression and loss in Bregman information. The information bottleneck method (Tishby et al., 1999) emerges as a special case of this viewpoint. We restrict our attention to regular exponential families and regular Bregman divergences in this section.

6.1 Rate Distortion Theory for Bregman Divergences

Rate distortion theory (Berger, 1971; Berger and Gibson, 1998) deals with the fundamental limits of quantizing a stochastic source $X \sim p(x),\ x \in \mathcal{X}$, using a random variable $\hat{X}$ over a reproduction alphabet $\hat{\mathcal{X}}$ typically assumed to embed the source alphabet $\mathcal{X}$, i.e., $\mathcal{X} \subseteq \hat{\mathcal{X}}$. In the rate distortion setting, the performance of a quantization scheme is determined in terms of the rate, i.e., the average number of bits for encoding a symbol, and the expected distortion between the source and the reproduction random variables based on an appropriate distortion function $d : \mathcal{X} \times \hat{\mathcal{X}} \mapsto \mathbb{R}_+$. The central problem in rate distortion theory (Cover and Thomas, 1991) is to compute the rate distortion function $R(D)$, which is defined as the minimum achievable rate for a specified level of expected distortion D, and can be mathematically expressed as

$$R(D) = \min_{p(\hat{x}|x): E_{X,\hat{X}}[d(X,\hat{X})] \leq D} I(X;\hat{X}) \ , \tag{21}$$

where $I(X;\hat{X})$ is the mutual information of X and $\hat{X}$.

The rate distortion problem is a convex problem that involves optimizing over the probabilistic assignments $p(\hat{x}|x)$ and can be theoretically solved using the Blahut-Arimoto algorithm (Arimoto, 1972; Blahut, 1972; Csiszár, 1974; Cover and Thomas, 1991). However, numerical computation of the rate distortion function through the Blahut-Arimoto algorithm is often infeasible in practice, primarily due to lack of knowledge of the optimal support of the reproduction random variable. An efficient solution for addressing this problem is the mapping approach (Banerjee et al., 2004a; Rose, 1994), where one solves a related problem that assumes cardinality k for the support of the reproduction random variable. In this setting, the optimization is over the assignments as well as the support set, i.e.,

$$\min_{\substack{\hat{\mathcal{X}}_s,\ p(\hat{x}|x) \\ |\hat{\mathcal{X}}_s|=k}} I(X;\hat{X}) + \beta_D E_{X,\hat{X}}[d(X,\hat{X})] \ , \tag{22}$$

where β_D is the optimal Lagrange multiplier that depends on the chosen tolerance level D of the expected distortion and $\hat{\mathcal{X}}_s$ is the optimal support of the reproduction random variable with cardinality k. We shall refer to the above problem (22) as the rate distortion problem with a support set of finite cardinality (RDFC). It can be shown (Berger, 1971) that the RDFC problem and the original rate distortion problem have identical solutions when the cardinality of the optimal support set is less than or equal to k, which is known to be true for cases without an analytical solution (Banerjee et al., 2004a).

Our analysis connects the Bregman soft clustering problem to the RDFC problem following results from Banerjee et al. (2004a), which extend previous work (Rose, 1994; Gray and Neuhoff, 1998) that related `kmeans` clustering to vector quantization and rate-distortion based on squared Euclidean distortion. Let $Z,\hat{Z}$ denote suitable sufficient statistic representations of $X,\hat{X}$ so that the distortion can be measured by a Bregman divergence d_ϕ in the sufficient statistic space. The RDFC problem can now be stated directly in terms of Z and $\hat{Z}$ as

$$\min_{\substack{\hat{\mathcal{Z}}_s,\ p(\hat{\mathrm{z}}|\mathrm{z}) \\ |\hat{\mathcal{Z}}_s|=k}} I(Z;\hat{Z}) + \beta_D E_{Z,\hat{Z}}[d_\phi(Z,\hat{Z})] \ , \tag{23}$$

where $\hat{\mathcal{Z}}_s$ is the optimal support of the reproduction random variable with cardinality k.

Unlike the basic rate distortion problem (21), the RDFC problem (23) is no longer a convex problem since it involves optimization over both $\hat{\mathcal{Z}}_s$ and $p(\hat{\mathrm{z}}|\mathrm{z})$. However, when either of the arguments is fixed, the resulting sub-problem can be solved exactly. In particular, when $\hat{\mathcal{Z}}_s$ is known, then the RDFC problem reduces to that of optimizing over $p(\hat{\mathrm{z}}|\mathrm{z})$, which is a feasible convex problem and can be exactly solved by the Blahut-Arimoto algorithm (Csiszár, 1974). Similarly, when the assignments $p(\hat{\mathrm{z}}|\mathrm{z})$ are known, the RDFC problem only involves minimizing the expected distortion measured in terms of a Bregman divergence and can be exactly solved using Proposition 1. Thus the objective function in (23) can be greedily minimized by alternately optimizing over the individual arguments, yielding a solution that is locally optimal. The details of this analysis and resulting algorithm can be found in Banerjee et al. (2004a).

Interestingly, it can be shown (Banerjee et al., 2004a) that the RDFC problem based on a regular Bregman divergence is exactly equivalent to the the maximum likelihood mixture estimation problem based on a uniquely determined exponential family when the source distribution in the rate distortion setting equals the empirical distribution over the sampled data points.

Theorem 7 (Banerjee et al. (2004a)) *Consider a source $Z \sim p(\mathrm{z})$, where $p(\mathrm{z})$ is the empirical distribution over the samples. Then the RDFC problem (23) for the source Z with regular Bregman divergence d_ϕ, variational parameter β_D, and reproduction random variable $\hat{Z}$ with $|\hat{Z}| = k$ is equivalent to the maximum likelihood mixture estimation problem based on the regular exponential family $\mathcal{F}_{\beta_D \psi}$ with number of mixture components set to k (ψ is the conjugate of ϕ).*

From Section 5, we know that the maximum likelihood mixture estimation problem for any regular exponential family is equivalent to the Bregman soft clustering problem for the corresponding regular Bregman divergence. Using this in conjunction with Theorem 7, we obtain the following equivalence relation between the RDFC problem and the Bregman soft clustering problem.

Theorem 8 *Consider a source $Z \sim p(\mathrm{z})$, where $p(\mathrm{z})$ is the empirical distribution over the samples. Then the RDFC problem (23) for the source Z with regular Bregman divergence d_ϕ, variational parameter β_D, and reproduction random variable $\hat{Z}$ with $|\hat{Z}| = k$ is equivalent to the Bregman soft clustering problem (16) based on the Bregman divergence $d_{\beta_D \phi}$ with number of clusters set to k.*

From the above theorem, it follows that Algorithm 3 can be used to solve the RDFC problem. Note that the update steps for $p(h|\mathbf{x})$ and π_h in Algorithm 3 exactly correspond to the updates of $p(\hat{\mathrm{z}}|\mathrm{z})$ and $p(\hat{\mathrm{z}})$ in the Blahut-Arimoto step in Algorithm 1 of Banerjee et al. (2004a) for solving the RDFC problem. The update of $\boldsymbol{\mu}_h$ in Algorithm 3 is equivalent to the update of $\hat{\mathrm{z}}$ in the support estimation step in Algorithm 1 of Banerjee et al. (2004a). From the viewpoint of alternate minimization, the order of the three updates $p(\hat{\mathrm{z}}|\mathrm{z}), p(\hat{\mathrm{z}})$ and $\hat{\mathrm{z}}$ is interchangeable and does not affect the local optimality guarantees, although different orderings may lead to different solutions.

The Bregman soft clustering problem corresponds to the RDFC problem and not to the basic rate distortion problem (21). However, as mentioned earlier, both the problems yield the same solution for the rate distortion function when the optimal support set $|\hat{\mathcal{Z}}_\mathrm{s}|$ is finite and k is sufficiently large. The solution is the rate distortion function and refers to the asymptotic rate (Cover and Thomas, 1991) that can be achieved for a given distortion, when we are allowed to code the source symbols in blocks of size m with $m \to \infty$.

It is also possible to consider a related rate distortion problem where the source symbols are coded using blocks of size 1. The resultant rate distortion function is referred to as the "scalar" or "order 1" rate distortion function $R_1(D)$ (Gray and Neuhoff, 1998). The problem is solved by performing hard assignments of the source symbols to the closest codebook members, which is similar to the assignment step in the Bregman hard clustering problem. In fact, the "order 1" or "1-shot" rate distortion problem, assuming a known finite cardinality of the optimal reproduction support set, turns out to be exactly equivalent to the Bregman hard clustering problem.

6.2 Compression vs. Bregman Information Trade-off

We now provide yet another view of the RDFC problem (and hence, Bregman soft clustering) as a lossy compression problem where the objective is to balance the trade-off between compression and preservation of Bregman information. Intuitively, the reproduction random variable $\hat{Z}$ is a coarser representation of the source random variable Z with less "information" than Z. In rate distortion theory, the loss in "information" is quantified by the expected Bregman distortion between Z and $\hat{Z}$. The following theorem, which is along the same lines as Theorem 1, provides a direct way of quantifying the intuitive loss in "information" in terms of Bregman information.

Theorem 9 (Banerjee et al. (2004a)) *The expected Bregman distortion between the source and the reproduction random variables is exactly equal to the loss in Bregman information due to compression, i.e.,*

$$E_{Z,\hat{Z}}[d_\phi(Z,\hat{Z})] = I_\phi(Z) - I_\phi(\hat{Z}) \ ,$$

where $\hat{Z} = E_{Z|\hat{Z}}[Z]$.

The RDFC problem (23) can, therefore, be viewed as an optimization problem involving a trade-off between the mutual information $I(Z;\hat{Z})$ that measures the compression, and the loss in Bregman information $I_\phi(Z) - I_\phi(\hat{Z})$. Since the source random variable Z is known, the Bregman information $I_\phi(Z)$ is fixed and minimizing the expected distortion is equivalent to maximizing the Bregman information of the compressed random variable $\hat{Z}$. Hence, this constrained form of the RDFC problem (23) can be written as:

$$\min_{p(\hat{z}|z)} \{I(Z;\hat{Z}) - \beta I_\phi(\hat{Z})\}, \tag{24}$$

where β is the variational parameter corresponding to the desired point in the rate distortion curve and $\hat{Z} = E_{Z|\hat{Z}}[Z]$. The variational parameter β determines the trade-off between the achieved compression and the preserved Bregman information.

6.2.1 INFORMATION BOTTLENECK REVISITED

We now demonstrate how the information bottleneck (IB) method of Tishby et al. (1999) can be derived from the RDFC problem (24) for a suitable choice of Bregman divergence.

Let $Y \sim p(y),\ y \in \mathcal{Y}$ be a random variable. Let the sufficient statistic random vector Z corresponding to a source X be the conditional distribution of Y given X, i.e., $Z = p(Y|X)$. Z is just a concrete representation of the possibly abstract source X. Similarly, the random variable $\hat{Z} = p(Y|\hat{X})$ represents the reproduction random variable $\hat{X}$. This choice of sufficient statistic mapping is appropriate when the joint distribution of the random variables X and Y contains all the relevant information about X. For the above choice of sufficient statistic mapping, an additional constraint that $\hat{Z}$ is the conditional expectation of Z leads to the lossy compression problem (24) where we need to find the optimal assignments that balance the trade-off between compression and the loss in Bregman information. Now, from Example 6 in Section 3.1, the Bregman information $I_\phi(\hat{Z})$ of the random variable $\hat{Z}$ that takes values over the set of conditional distributions $\{p(Y|\hat{x})\}$ with probability $p(\hat{x})$ is the same as the mutual information $I(\hat{X};Y)$ of $\hat{X}$ and Y (when the Bregman divergence is the KL-divergence). Hence, the problem (24) reduces to

$$\min_{p(\hat{x}|x)} \{I(X;\hat{X}) - \beta I(\hat{X};Y)\}, \tag{25}$$

since $p(\hat{x}|x) = p(\hat{z}|z)$ and $I(X;\hat{X}) = I(Z;\hat{Z})$, where β is the variational parameter. This is identical to the IB formulation (Tishby et al., 1999). Our framework reveals that the IB assumption that the mutual information with respect to another random variable Y holds all the relevant information for comparing the different source entities is equivalent to assuming that (a) $p(Y|X)$ is the appropriate sufficient statistic representation, and (b) the KL-divergence between the conditional distributions of Y is the appropriate distortion measure. Further, the assumption about the conditional independence of Y and $\hat{X}$ given X, i.e., the Markov chain condition $Y \leftrightarrow X \leftrightarrow \hat{X}$, is equivalent to the constraint that $\hat{Z}$ is the conditional expectation of Z, i.e., $\hat{Z} = p(Y|\hat{X}) = E_{X|\hat{X}}[p(Y|X)] = E_{Z|\hat{Z}}[Z]$.

Thus the information bottleneck problem is seen to be a special case of the RDFC problem (23), and hence also of the Bregman soft clustering problem and mixture estimation problem for exponential families. In particular, IB is exactly equivalent to the mixture estimation problem based on the exponential family corresponding to KL-divergence, i.e., the multinomial family (Collins et al., 2001). Further, the iterative IB algorithm is the same as the EM algorithm for multinomial distributions (Slonim and Weiss, 2002), and also the Bregman soft clustering algorithm using KL-divergence.

7. Experiments

There are a number of experimental results in existing literature (MacQueen, 1967; Linde et al., 1980; Buzo et al., 1980; Dhillon et al., 2003; Nigam et al., 2000) that illustrate the usefulness of specific Bregman divergences and the corresponding Bregman clustering algorithms in important application 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 (Dhillon et al., 2003) and the EM algorithm on a mixture of multinomials based on the naive Bayes assumption (Nigam et al., 2000) have been applied. These algorithms are, respectively, special cases of the Bregman hard and soft clustering algorithms for KL-divergence, and have been shown to provide high quality results on large real datasets such as the 20-Newsgroups, Reuters and Dmoz text datasets. This success is not unexpected as text documents can be effectively modeled using multinomial distributions where the corresponding Bregman divergence is just the KL-divergence between word distributions.

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 (Linde et al., 1980; Buzo et al., 1980), has been successfully applied. Speech power spectra tend to follow exponential family densities of the form $p(x) = \lambda e^{-\lambda x}$ whose corresponding Bregman divergence is the Itakura-Saito distance (see Table 2).

Since special cases of Bregman clustering algorithms have already been shown to be effective in various 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 appropriateness of the Bregman divergence. In particular we study Bregman clustering of data generated from mixture of exponential family distributions using the corresponding Bregman divergence as well as non-matching divergences. The results indicate that the cluster quality is best when the Bregman divergence corresponding to the generative model is employed.

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, in the sense that the variance is approximately the same for all the models. Figure 1 shows the density functions of the generative

models. The datasets were then each clustered using three versions of the Bregman hard clustering

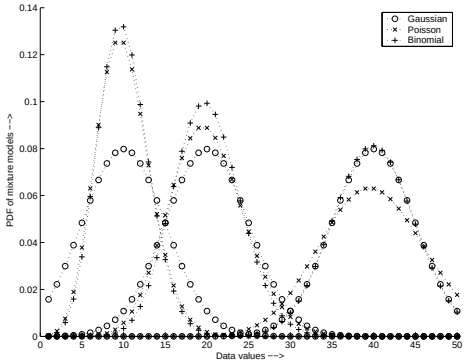

Figure 1: Generative models for data sets used in experiment 1

Table 3: Clustering results for the first data set. Columns 2-4 correspond to the normalized mutual information between original and predicted clusters obtained by applying the Bregman clustering algorithm corresponding to the Bregman divergences $d_{Gaussian}$, $d_{Poisson}$ and $d_{Binomial}$ respectively

Generative Model	$\mathbf{d_{Gaussian}}$	$\mathbf{d_{Poisson}}$	$\mathbf{d_{Binomial}}$
Gaussian	$\mathbf{0.701 \pm 0.033}$	0.633 ± 0.043	0.641 ± 0.035
Poisson	0.689 ± 0.063	$\mathbf{0.734 \pm 0.057}$	0.694 ± 0.059
Binomial	0.769 ± 0.061	0.746 ± 0.048	$\mathbf{0.825 \pm 0.046}$

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[11] (Strehl and Ghosh, 2002) between the predicted clusters and original clusters (based on the actual generating mixture component), and the results were averaged over 10 trials. Table 3 shows the normalized mutual information values for the different divergences and datasets. From Table 3, we can see that clustering quality is significantly better when the Bregman divergence used in the clustering algorithm matches that of the generative model.

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 4 lead to the same conclusion as before, i.e., the choice of the Bregman divergence used for clustering is important for obtaining good quality clusters.

In practice, an important issue that needs to be addressed is: what is the appropriate Bregman divergence for a given application? In certain situations, it may be possible to realistically characterize the data generative process using a mixture of exponential family distributions. In such a scenario, especially in the absence of a better methodology, using the divergence corresponding to

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

Table 4: Clustering results for the second set of data sets

Generative Model	$\mathbf{d_{Gaussian}}$	$\mathbf{d_{Poisson}}$	$\mathbf{d_{Binomial}}$
Gaussian	$\mathbf{0.728 \pm 0.005}$	0.661 ± 0.007	0.669 ± 0.005
Poisson	0.792 ± 0.013	$\mathbf{0.815 \pm 0.014}$	0.802 ± 0.013
Binomial	0.823 ± 0.006	0.833 ± 0.011	$\mathbf{0.849 \pm 0.012}$

the exponential family seems appropriate. In general, however, the divergence used for clustering need not necessarily have to be the one corresponding to the generative model. The final choice should depend on the relevant application, i.e., the divergence should capture the similarity properties desirable in the application, and need not necessarily depend on how the data was actually generated.

8. Related Work

This work is largely inspired by three broad and overlapping ideas. First, an information theoretic viewpoint of the clustering problem is invaluable. Such considerations occur in several techniques, from classical vector quantization (Gersho and Gray, 1992) to information theoretic clustering (Dhillon et al., 2003) and the information bottleneck method (Tishby et al., 1999). In particular, the information theoretic hard clustering (Dhillon et al., 2003) 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 family distributions and the suitability of Bregman divergences as distortion or loss functions for data drawn from exponential distributions. It has been previously shown (Amari and Nagaoka, 2001; Azoury and Warmuth, 2001) that the KL-divergence, which is the most natural distance measure for comparing two members $p_{(\psi,\boldsymbol{\theta})}$ and $p_{(\psi,\tilde{\boldsymbol{\theta}})}$ 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 a natural distance function for the data drawn according to a mixture model based on 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 (Censor and Zenios, 1998). Elegant techniques for the design of algorithms and the analysis of relative loss bounds in the online learning setting extensively use this framework (Azoury and Warmuth, 2001). In the unsupervised learning setting, use of this framework typically involves development of alternate minimization procedures (Csiszár and Tusnády, 1984). For example, Pietra et al. (2001); Wang and Schuurmans (2003) analyze and develop iterative alternate projection procedures for solving unsupervised optimization problems that involve objective functions based on Bregman divergences under various kinds of constraints. Further, Collins et al. (2001) develop a generalization of PCA for exponential families using loss functions based on the corresponding Bregman divergences and propose alternate minimization schemes for solving the problem.

On a larger context, there has been research in various fields that has focussed on generalized notions of distances and on extending known methodologies to the general setting (Rao, 1982). Grünwald and Dawid (2004) recently extended the 'redundancy-capacity theorem' of information theory to arbitrary discrepancy measures. As an extension of Shannon's entropy (Cover and Thomas, 1991), they introduced generalized entropy measures that are (not necessarily differentiable) concave functions of probability distributions. Just as Shannon's entropy is the minimum number of bits (on an average) required to encode a stochastic source, the generalized entropy measures correspond to the infimum of a general class of loss functions in a game theoretic setting. Restricting their results to our setting, the generalized entropy is equivalent to the concave function $-\phi$, where ϕ determines the Bregman divergence d_ϕ. However, our framework is applicable to arbitrary vectors (or functions), whereas Grünwald and Dawid (2004) focus only on probability distributions.

As we discussed in Section 6, our treatment of clustering is very closely tied to rate distortion theory (Berger, 1971; Berger and Gibson, 1998; Gray and Neuhoff, 1998). The results presented in the paper extend vector quantization methods (Gersho and Gray, 1992) to a large class of distortion measures. Further, building on the work of Rose (1994), our results provide practical ways of computing the rate-distortion function when distortion is measured by a Bregman divergence. In addition, the results also establish a connection between the rate distortion problem with Bregman divergences and the mixture model estimation problem for exponential families (Banerjee et al., 2004a).

In the literature, there are clustering algorithms that involve minimizing loss functions based on distortion measures that are somewhat different from Bregman divergences. For example, Modha and Spangler (2003) present the `convex-kmeans` clustering algorithm 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. Linde et al. (1980) consider distortion measures of the form $d(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T A(\mathbf{x})(\mathbf{x}-\mathbf{y})$ where $\mathbf{x},\mathbf{y} \in \mathbb{R}^d$ and $A(\mathbf{x})$ is a $d \times d$ positive definite matrix, as loss functions for vector quantization. Although such distortions are Bregman divergences in some cases, e.g., when $A(\mathbf{x})$ is a constant matrix, in general one has to solve a convex optimization problem to compute the optimal representative when using the above $d(\mathbf{x},\mathbf{y})$.

9. Concluding Remarks

In this paper, we have presented hard and soft clustering algorithms to minimize loss functions involving Bregman divergences. Our analysis presents a unified view of an entire class of centroid based 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, using a related result, we see that Bregman divergences are the only distortion functions for which such a centroid-based clustering scheme is possible. Second, we formally show that there is a one-to-one correspondence between regular exponential families and regular Bregman divergences. This result is useful in developing an alternative interpretation of the EM algorithm for learning mixtures of exponential distributions, eventually resulting in a class of Bregman soft-clustering algorithms. Our formulation also turns out to be closely tied to the rate distortion theory for Bregman divergences.

As discussed in the paper, special cases of our analysis have been discovered and widely used by researchers in applications ranging from speech coding to text clustering. There are three salient features of this framework that make these results particularly useful for real-life applications. First, the computational complexity of each iteration of the entire class of Bregman clustering algorithms is linear in the number of data-points. Hence the algorithms are scalable and appropriate for large-scale machine learning tasks. Second, the modularity of the proposed class of algorithms is evident from the fact that only one component in the proposed schemes, i.e., the Bregman divergence used in the assignment step, needs to be changed to obtain an algorithm for a new loss function. This simplifies the implementation and application of this class of algorithms to various data types. Third, the algorithms discussed are also applicable to mixed data types that are commonly encountered in real applications. Since a convex combination of convex functions is always convex, one can have different convex functions appropriately chosen for different subsets of features. The Bregman divergence corresponding to a convex combination of the component functions can now be used to cluster the data, thus vastly increasing the scope of the proposed techniques.

Acknowledgments

We would like to thank Peter Grünwald and an anonymous referee for their constructive comments that greatly improved the presentation of the paper. We thank Manfred Warmuth for pointing out that a rigorous proof of (13) was required, and Alex Smola and S. V. N. Vishwanathan for their valuable comments on Section 4. Part of this research was supported by an IBM PhD fellowship, NSF grants IIS-0307792, IIS-0325116, CCF-0431257, NSF CAREER Award No. ACI-0093404 and Texas Advanced Research Program grant 003658-0431-2001.

Appendix A. Properties of Bregman Divergences

In this section, we list some well-known useful properties of Bregman divergences.

Properties of Bregman Divergences

Let $\phi : \mathcal{S} \mapsto \mathbb{R}$ be a strictly convex, differentiable function defined on a convex set $\mathcal{S} = \text{dom}(\phi) \subseteq \mathbb{R}^d$ and let $d_\phi : \mathcal{S} \times \text{ri}(\mathcal{S}) \mapsto [0, \infty)$ be its Bregman divergence, i.e, $d_\phi(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}, \nabla\phi(\mathbf{y}) \rangle$. Then, the following properties are true.

1. **Non-negativity.** $d_\phi(\mathbf{x}, \mathbf{y}) \geq 0, \ \forall \mathbf{x} \in \mathcal{S}, \mathbf{y} \in \text{ri}(\mathcal{S})$, and equality holds if and only if $\mathbf{x} = \mathbf{y}$.

2. **Convexity.** d_ϕ is always convex in the first argument, but not necessarily convex in the second argument. Squared Euclidean distance and KL-divergence are examples of Bregman divergences that are convex in both their arguments, but the Bregman divergence corresponding to the strictly convex function $\phi(x) = x^3$, defined on $\mathbb{R}_+$, given by $d_\phi(x, y) = x^3 - y^3 - 3(x - y)y^2$ an example divergence that is not convex in y.

3. **Linearity.** Bregman divergence is a linear operator i.e., $\forall \mathbf{x} \in \mathcal{S}, \mathbf{y} \in \text{ri}(\mathcal{S})$,

$$\begin{aligned} d_{\phi_1 + \phi_2}(\mathbf{x}, \mathbf{y}) &= d_{\phi_1}(\mathbf{x}, \mathbf{y}) + d_{\phi_2}(\mathbf{x}, \mathbf{y}) \ , \\ d_{c\phi}(\mathbf{x}, \mathbf{y}) &= c d_\phi(\mathbf{x}, \mathbf{y}) \quad (\text{for } c \geq 0) \ . \end{aligned}$$

4. **Equivalence classes.** The Bregman divergences of functions that differ only in affine terms are identical i.e., if $\phi(\mathbf{x}) = \phi_0(\mathbf{x}) + \langle \mathbf{b}, \mathbf{x} \rangle + c$ where $\mathbf{b} \in \mathbb{R}^d$ and $c \in \mathbb{R}$, then $d_\phi(\mathbf{x}, \mathbf{y}) = d_{\phi_0}(\mathbf{x}, \mathbf{y}), \forall \mathbf{x} \in \mathcal{S}, \mathbf{y} \in \mathrm{ri}(\mathcal{S})$. Hence, the set of all strictly convex, differentiable functions on a convex set $\mathcal{S}$ can be partitioned into equivalence classes of the form

$$[\phi_0] = \{\phi \mid d_\phi(\mathbf{x}, \mathbf{y}) = d_{\phi_0}(\mathbf{x}, \mathbf{y}) \; \forall \; \mathbf{x} \in \mathcal{S}, \; \mathbf{y} \in \mathrm{ri}(\mathcal{S})\}.$$

5. **Linear separation.** The locus of all the points $\mathbf{x} \in \mathcal{S}$ that are equidistant from two fixed points $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathrm{ri}(\mathcal{S})$ in terms of a Bregman divergence is a hyperplane, i.e., the partitions induced by Bregman divergences have linear separators given by

$$\begin{aligned} & d_\phi(\mathbf{x}, \boldsymbol{\mu}_1) = d_\phi(\mathbf{x}, \boldsymbol{\mu}_2) \\ \Rightarrow \quad & \phi(\mathbf{x}) - \phi(\boldsymbol{\mu}_1) - \langle \mathbf{x} - \boldsymbol{\mu}_1, \nabla\phi(\boldsymbol{\mu}_1) \rangle = \phi(\mathbf{x}) - \phi(\boldsymbol{\mu}_2) - \langle \mathbf{x} - \boldsymbol{\mu}_2, \nabla\phi(\boldsymbol{\mu}_2) \rangle \\ \Rightarrow \quad & \langle \mathbf{x}, \nabla\phi(\boldsymbol{\mu}_2) - \nabla\phi(\boldsymbol{\mu}_1) \rangle = (\phi(\boldsymbol{\mu}_1) - \phi(\boldsymbol{\mu}_2)) - (\langle \boldsymbol{\mu}_1, \nabla\phi(\boldsymbol{\mu}_1) \rangle - \langle \boldsymbol{\mu}_2, \nabla\phi(\boldsymbol{\mu}_2) \rangle) \end{aligned}$$

6. **Dual Divergences.** Bregman divergences obtained from a Legendre function ϕ and its conjugate ψ satisfy the duality property:

$$d_\phi(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = \phi(\boldsymbol{\mu}_1) + \psi(\boldsymbol{\theta}_2) - \langle \boldsymbol{\mu}_1, \boldsymbol{\theta}_2 \rangle = d_\psi(\boldsymbol{\theta}_2, \boldsymbol{\theta}_1),$$

where $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathrm{ri}(\mathcal{S})$ are related to $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathrm{ri}(\Theta)$ by the Legendre transformation.

7. **Relation to KL-divergence.** Let $\mathcal{F}_\psi$ be an exponential family with ψ as the cumulant function. Then the KL divergence between two members $p_{(\psi,\boldsymbol{\theta}_1)}$ and $p_{(\psi,\boldsymbol{\theta}_2)}$ in F_ψ corresponding to natural parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ can be expressed as a Bregman divergence in two possible ways. In particular,

$$KL(p_{(\psi,\theta_1)} || p_{(\psi,\theta_2)}) = d_\phi(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = d_\psi(\boldsymbol{\theta}_2, \boldsymbol{\theta}_1)$$

where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the expectation parameters corresponding to $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Further, if $\psi(\mathbf{0}) = 0$, then $p_{(\psi,\mathbf{0})}(\mathbf{x}) = p_0(\mathbf{x})$ is itself a valid probability density and $KL(p_{(\psi,\boldsymbol{\theta})} \| \; p_{(\psi,\mathbf{0})}) = \phi(\boldsymbol{\mu})$, where $\boldsymbol{\mu} = \nabla\psi(\boldsymbol{\theta})$.

8. **Generalized Pythagoras theorem.** For any $\mathbf{x}_1 \in \mathcal{S}$ and $\mathbf{x}_2, \mathbf{x}_3 \in \mathrm{ri}(\mathcal{S})$, the following three-point property holds:

$$d_\phi(\mathbf{x}_1, \mathbf{x}_3) = d_\phi(\mathbf{x}_1, \mathbf{x}_2) + d_\phi(\mathbf{x}_2, \mathbf{x}_3) - \langle \mathbf{x}_1 - \mathbf{x}_2, \nabla\phi(\mathbf{x}_3) - \nabla\phi(\mathbf{x}_2) \rangle. \tag{26}$$

When $\mathbf{x}_1, \mathbf{x}_2$ and $\mathbf{x}_3$ are such that $\mathbf{x}_1 \in \mathcal{S}'$ where $\mathcal{S}'$ is a convex subset of $\mathcal{S}$ and $\mathbf{x}_2$ is given by

$$\mathbf{x}_2 = \operatorname*{argmin}_{\mathbf{x} \in \mathcal{S}'} d_\phi(\mathbf{x}, \mathbf{x}_3),$$

then the inner product term in (26) becomes negative and we have,

$$d_\phi(\mathbf{x}_1, \mathbf{x}_2) + d_\phi(\mathbf{x}_2, \mathbf{x}_3) \leq d_\phi(\mathbf{x}_1, \mathbf{x}_3).$$

When the convex subset $\mathcal{S}'$ is an affine set, then the inner product term is zero giving rise to an equality.

Necessary and Sufficient Conditions for a Bregman Divergence

A divergence measure $d : \mathcal{S} \times \mathrm{ri}(\mathcal{S}) \mapsto [0,\infty)$ is a Bregman divergence if and only if there exists $\mathbf{a} \in \mathrm{ri}(\mathcal{S})$ such that the function $\phi_{\mathbf{a}}(\mathbf{x}) = d(\mathbf{x},\mathbf{a})$ satisfies the following conditions:

1. $\phi_{\mathbf{a}}$ is strictly convex on $\mathcal{S}$ and differentiable on $\mathrm{ri}(\mathcal{S})$.
2. $d(\mathbf{x},\mathbf{y}) = d_{\phi_{\mathbf{a}}}(\mathbf{x},\mathbf{y}),\ \forall \mathbf{x} \in \mathcal{S}, \mathbf{y} \in \mathrm{ri}(\mathcal{S})$ where $d_{\phi_{\mathbf{a}}}$ is the Bregman divergence associated with $\phi_{\mathbf{a}}$.

It is easy to see the sufficiency property from the second condition. To prove that the conditions are necessary as well, we note that for any strictly convex, differentiable function ϕ, the Bregman divergence evaluated with a fixed value for the second argument differs from it only by a linear term, i.e.,

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

where $\mathbf{b} = -\nabla\phi(\mathbf{a})$ and $c = \langle \mathbf{a}, \nabla\phi(\mathbf{a})\rangle - \phi(\mathbf{a})$. Hence, $\phi_{\mathbf{a}}$ is also strictly convex and differentiable and the Bregman divergences associated with ϕ and $\phi_{\mathbf{a}}$ are identical.

Appendix B. Proof of Exhaustiveness Result

This appendix is based on results reported in Banerjee et al. (2005) and is included in this paper for the sake of completeness. The results discussed here show the exhaustiveness of Bregman divergences with respect to the property proved in Proposition 1.

Theorem 10 (Banerjee et al. (2005)) *Let $F : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}_+$ be a continuous and differentiable function $F(x,y)$ with continuous partial derivatives $\frac{\partial F}{\partial x}$ and $\frac{\partial F}{\partial y}$ such that $F(x,x) = 0, \forall x \in \mathbb{R}$. For all sets $\mathcal{X} \subseteq \mathbb{R}$ and all probability measures ν over $\mathcal{X}$, if the random variable X takes values in $\mathcal{X}$ following ν such that $y^* = E_\nu[X]$ is the unique minimizer of $E_\nu[F(X,y)]$ over all $y \in \mathbb{R}$, i.e., if*

$$\operatorname*{argmin}_{y \in \mathbb{R}} E_\nu[F(X,y)] = E_\nu[X] \tag{27}$$

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

Proof Since the optimality property in (27) is true for all $\mathcal{X}$ and ν, we give a constructive argument with a particular choice of $\mathcal{X}$ and ν. Let $\mathcal{X} = \{a,b\} \subset \mathbb{R}$ where $a \neq b$, and let ν be $\{p,q\}$, with $p,q \in (0,1)$ and $p+q=1$ so that $E_\nu[X] = pa+qb$. Then from (27),

$$pF(a,y) + qF(b,y) = E_\nu[F(X,y)] \geq E_\nu[F(X,E_\nu[X])] = pF(a,pa+qb) + qF(b,pa+qb)$$

$\forall y \in \mathbb{R}$. If we consider the left-hand-side as a function of y, it equals the right-hand-side at $y = y^* \doteq E_\nu[X] = pa+qb$. Therefore, we must have

$$p\frac{\partial F(a,y^*)}{\partial y} + q\frac{\partial F(b,y^*)}{\partial y} = 0. \tag{28}$$

Substituting $p = (y^* - b)/(a-b)$ and rearranging terms yields

$$\frac{1}{(y^*-a)}\frac{\partial F(a,y^*)}{\partial y} = \frac{1}{(y^*-b)}\frac{\partial F(b,y^*)}{\partial s}.$$

Since a, b and p are arbitrary, the above equality implies that the function

$$\frac{1}{(y-x)}\frac{\partial F(x,y)}{\partial y}$$

is independent of x. Thus we can write, for some function H,

$$\frac{\partial F(x,y)}{\partial y} = (y-x)H(y), \tag{29}$$

for some continuous function H.

Now define function ϕ by

$$\phi(y) = \int_0^y \int_0^{y'} H(t)dtdy'.$$

Then ϕ is differentiable with $\phi(0) = \phi'(0) = 0$, $\phi''(y) = H(y)$. Integration by parts for (29) leads to

$$F(x,y) - F(x,x) = \int_x^y (y'-x)H(y')\,dy' = \phi(x) - \phi(y) - \phi'(y)(x-y).$$

Since $F(x,x) = 0$, the non-negativity of F implies that ϕ is a convex function.

It remains to show that ϕ is strictly convex. Suppose ϕ is not strictly convex. Then there exists an interval $I = [\ell_1, \ell_2]$ such that $\ell_1 < \ell_2$ and $\phi'(y) = (\phi(\ell_1) - \phi(\ell_2))/(\ell_1 - \ell_2)$ for all $y \in I$. Consider the set $\mathcal{X} = \{\ell_1, \ell_2\}$ with $\nu = \{\frac{1}{2}, \frac{1}{2}\}$. It is easy to check that any $y \in I$ is a minimizer of $E_\nu[F(X,y)]$. This is a contradiction, and so ϕ must be strictly convex. ■

It is possible to get rid of the condition that $\frac{\partial F}{\partial y}$ has to be continuous by proper mollification arguments (Banerjee et al., 2005). Further, it is possible to generalize the result to functions in more than one dimension, i.e., $F : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}_+$.

Theorem 11 (Banerjee et al. (2005)) *Let $F : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}_+$ be a continuous function such that $F(\mathbf{x},\mathbf{x}) = 0, \forall \mathbf{x} \in \mathbb{R}^d$, and the second order partial derivatives $\frac{\partial^2 F}{\partial \mathbf{x}_i \partial \mathbf{x}_j}, 1 \leq i, j, \leq d$, are all continuous. For all sets $\mathcal{X} \subseteq \mathbb{R}^d$ and all probability measures ν over $\mathcal{X}$, if the random variable X takes values in $\mathcal{X}$ following ν such that $\mathbf{y} = E_\nu[X]$ is the unique minimizer of $E_\nu[F(X,\mathbf{y})]$ over all $\mathbf{y} \in \mathbb{R}^d$, i.e., if*

$$\operatorname*{argmin}_{\mathbf{y}\in\mathbb{R}^d} E_\nu[F(X,\mathbf{y})] = E_\nu[X]\ ,$$

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

The proof of Theorem 11 builds on the intuition of the proof of Theorem 10, but is more involved and hence skipped; the interested reader is referred to Banerjee et al. (2005).

Appendix C. Proof of Theorem 3

This appendix provides a proof of Theorem 3 in Section 4.3 and related results. Most of the ideas used in our analysis are from Section 9.1 of Barndorff-Nielsen (1978). For the sake of completeness, we give detailed proofs of the results. We begin with definitions. Let P_0 be any non-negative bounded measure on $\mathbb{R}^d$ and $\mathcal{F}_\psi = \{p_{(\psi,\boldsymbol{\theta})},\ \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d\}$ be a regular exponential family with cumulant function ψ and base measure P_0, as discussed in Section 4.1. Without loss of generality, let P_0 be a probability measure.[12] Let I_ψ be the support of P_0 (Definition 6) and hence, of all the probability distributions in $\mathcal{F}_\psi$. Let ϕ be the conjugate of ψ so that $(\text{int}(\text{dom}(\phi)), \phi)$ and (Θ, ψ) are Legendre duals of each other.

Lemma 3 *For any $\boldsymbol{\theta} \in \Theta$ and $\mathbf{x} \in \mathbb{R}^d$,*

$$\langle \boldsymbol{\theta}, \mathbf{x} \rangle - \psi(\boldsymbol{\theta}) \leq -\log \left(\inf_{\mathbf{u} \in \mathbb{R}^d, \|\mathbf{u}\|_2 = 1} P_0 \left[\langle \mathbf{u}, X \rangle \geq \langle \mathbf{u}, \mathbf{x} \rangle \right] \right) \tag{30}$$

where $X \sim P_0$. Hence

$$\inf_{\mathbf{u} \in \mathbb{R}^d, \|\mathbf{u}\|_2 = 1} P_0[\langle \mathbf{u}, X \rangle \geq \langle \mathbf{u}, \mathbf{x} \rangle] > 0 \quad \textit{implies that} \quad \mathbf{x} \in \text{dom}(\phi)\ .$$

Proof Let $\mathbf{u}_{\boldsymbol{\theta}}$ be the unit vector in the direction of $\boldsymbol{\theta}$. Given any $\mathbf{x} \in \mathbb{R}^d$, it is possible to divide $\mathbb{R}^d$ into two half spaces $\mathcal{G}_1 = \{\mathbf{x}' \in \mathbb{R}^d |\ \langle \mathbf{u}_{\boldsymbol{\theta}}, \mathbf{x}' \rangle < \langle \mathbf{u}_{\boldsymbol{\theta}}, \mathbf{x} \rangle\}$ and $\mathcal{G}_2 = \{\mathbf{x}' \in \mathbb{R}^d |\ \langle \mathbf{u}_{\boldsymbol{\theta}}, \mathbf{x}' \rangle \geq \langle \mathbf{u}_{\boldsymbol{\theta}}, \mathbf{x} \rangle\}$. For any $\boldsymbol{\theta}$, we have

$$\begin{aligned} 1 &= \int_{\mathbf{x}' \in \mathbb{R}^d} \exp(\langle \boldsymbol{\theta}, \mathbf{x}' \rangle - \psi(\boldsymbol{\theta})) dP_0(\mathbf{x}') \\ \Rightarrow \quad \exp(\psi(\boldsymbol{\theta})) &= \int_{\mathbf{x}' \in \mathbb{R}^d} \exp(\langle \boldsymbol{\theta}, \mathbf{x}' \rangle) dP_0(\mathbf{x}')\ . \end{aligned}$$

Partitioning the integral over $\mathbb{R}^d$ into $\mathcal{G}_1$ and $\mathcal{G}_2$, we obtain

$$\begin{aligned} \exp(\psi(\boldsymbol{\theta})) &= \int_{\mathbf{x}' \in \mathcal{G}_1} \exp(\langle \boldsymbol{\theta}, \mathbf{x}' \rangle) dP_0(\mathbf{x}') + \int_{\mathbf{x}' \in \mathcal{G}_2} \exp(\langle \boldsymbol{\theta}, \mathbf{x}' \rangle) dP_0(\mathbf{x}') \\ &\geq \int_{\mathbf{x}' \in \mathcal{G}_2} \exp(\langle \boldsymbol{\theta}, \mathbf{x}' \rangle) dP_0(\mathbf{x}') \\ &\geq \exp(\langle \boldsymbol{\theta}, \mathbf{x} \rangle) \int_{\mathbf{x}' \in \mathcal{G}_2} dP_0(\mathbf{x}') \quad (\text{since } \langle \mathbf{u}_{\boldsymbol{\theta}}, \mathbf{x}' \rangle \geq \langle \mathbf{u}_{\boldsymbol{\theta}}, \mathbf{x} \rangle \text{ for } \mathbf{x}' \in \mathcal{G}_2\ ) \\ &= \exp(\langle \boldsymbol{\theta}, \mathbf{x} \rangle) P_0[\langle \mathbf{u}_{\boldsymbol{\theta}}, X \rangle \geq \langle \mathbf{u}_{\boldsymbol{\theta}}, \mathbf{x} \rangle] \\ &\geq \exp(\langle \boldsymbol{\theta}, \mathbf{x} \rangle) \inf_{\mathbf{u}} P_0[\langle \mathbf{u}, X \rangle \geq \langle \mathbf{u}, \mathbf{x} \rangle]\ . \end{aligned}$$

On taking logarithms and re-arranging terms, we obtain (30).

From (30), $\inf_{\mathbf{u}} P_0[\langle \mathbf{u}, X \rangle \geq \langle \mathbf{u}, \mathbf{x} \rangle] > 0$ implies that $\forall \boldsymbol{\theta}, \langle \boldsymbol{\theta}, \mathbf{x} \rangle - \psi(\boldsymbol{\theta}) < \infty$, so that

$$\phi(\mathbf{x}) = \sup_{\boldsymbol{\theta}} (\langle \boldsymbol{\theta}, \mathbf{x} \rangle - \psi(\boldsymbol{\theta})) < \infty,$$

12. Since any non-negative bounded measure can be simply converted to a probability measure by a multiplicative constant, our analysis remains practically unchanged in the general case, except for an additive constant to the cumulant function.

i.e., $\mathbf{x} \in \text{dom}(\phi)$. ∎

We now prove the claim of Theorem 3 that $I_\psi \subseteq \text{dom}(\phi)$.

Proof of Theorem 3 Let $\mathbf{x}_0 \in I_\psi$ and let $\mathbf{u}$ be any unit vector. Let $H(\mathbf{u}, \mathbf{x}_0)$ be the hyperplane through $\mathbf{x}_0$ with unit normal $\mathbf{u}$. Let $\mathcal{H}(\mathbf{u}, \mathbf{x}_0)$ be the closed half-space determined by the hyperplane $H(\mathbf{u}, \mathbf{x}_0)$, i.e., $\mathcal{H}(\mathbf{u}, \mathbf{x}_0) = \{\mathbf{x} \in \mathbb{R}^d | \langle \mathbf{u}, \mathbf{x} \rangle \geq \langle \mathbf{u}, \mathbf{x}_0 \rangle\}$. Using this notation, we give separate proofs for the cases when P_0 is absolutely continuous with respect to the counting measure and with respect to the Lebesgue measure.

Let P_0 be absolutely continuous with respect to the counting measure. By definition, $\mathbf{x}_0 \in \mathcal{H}(\mathbf{u}, \mathbf{x}_0)$. Since $\mathbf{x}_0 \in I_\psi$, applying Definition 6 to the set $I = \{\mathbf{x}_0\}$ we have $p_{(\psi,\boldsymbol{\theta})}(\mathbf{x}_0) > 0$. Hence $p_0(\mathbf{x}_0) > 0$ as the exponential family distribution is absolutely continuous with respect to P_0. Therefore, the closed half-space $\mathcal{H}(\mathbf{u}, \mathbf{x}_0)$ has a positive measure of at least $p_0(\mathbf{x}_0)$ for any unit vector $\mathbf{u}$, i.e.,

$$\begin{aligned} & P_0\left[\langle \mathbf{u}, X \rangle \geq \langle \mathbf{u}, \mathbf{x}_0 \rangle\right] \;\geq\; p_0(\mathbf{x}_0) \;>\; 0 \quad \forall \mathbf{u} \\ \text{so that} \qquad & \inf_{\mathbf{u}} P_0\left[\langle \mathbf{u}, X \rangle \geq \langle \mathbf{u}, \mathbf{x}_0 \rangle\right] \;\geq\; p_0(\mathbf{x}_0) \;>\; 0\,. \end{aligned}$$

From Lemma 3, it follows that $\mathbf{x}_0 \in \text{dom}(\phi)$. Therefore, $I_\psi \subseteq \text{dom}(\phi)$.

Now we consider the case when P_0 is absolutely continuous with respect to the Lebesgue measure. If $\mathbf{x}_0 \in I_\psi$, then $\forall I \subseteq \mathbb{R}^d$ with $\mathbf{x}_0 \in I$ and $\int_I d\mathbf{x} > 0$, we have

$$\int_I dP_0(\mathbf{x}) > 0\,.$$

Note that since $\mathbf{x}_0 \in \mathcal{H}(\mathbf{u}, \mathbf{x}_0)$ and $\int_{\mathcal{H}(\mathbf{u},\mathbf{x}_0)} d\mathbf{x} > 0$, we must have

$$\int_{\mathcal{H}(\mathbf{u},\mathbf{x}_0)} dP_0(\mathbf{x}) > 0 \quad \forall \mathbf{u}.$$

Hence, $P_0(\langle \mathbf{u}, X \rangle \geq \langle \mathbf{u}, \mathbf{x}_0 \rangle) > 0, \forall \mathbf{u}$. Since the set of unit vectors is a compact set, $\inf_{\mathbf{u}} P_0(\langle \mathbf{u}, X \rangle \geq \langle \mathbf{u}, \mathbf{x}_0 \rangle)$ is achieved at some unit vector $\mathbf{u}^*$, so that

$$\inf_{\mathbf{u}} P_0(\langle \mathbf{u}, X \rangle \geq \langle \mathbf{u}, \mathbf{x}_0 \rangle) = P_0(\langle \mathbf{u}^*, X \rangle \geq \langle \mathbf{u}^*, \mathbf{x}_0 \rangle) > 0\,.$$

Again, Lemma 3 implies that $\mathbf{x}_0 \in \text{dom}(\phi)$ so that $I_\psi \subseteq \text{dom}(\phi)$. ∎

Finally, we present a related result from Barndorff-Nielsen (1978) involving the closed convex hull of I_ψ and $\text{dom}(\phi)$. The result is not essential to the paper, but is relevant, and interesting in its own right.

Theorem 12 (Barndorff-Nielsen (1978)) *Let I_ψ be as in Definition 6. Let C_ψ be the closure of the convex hull of I_ψ, i.e., $C_\psi = \text{co}(I_\psi)$. Then,*

$$\text{int}(C_\psi) \subseteq \text{dom}(\phi) \subseteq C_\psi$$

where ϕ is the conjugate of ψ.

Note that Theorem 12 does not imply Theorem 3.

References

N. I. Akhizer. *The Classical Moment Problem and some related questions in analysis*. Hafner Publishing Company, 1965.

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

S. Amari and H. Nagaoka. *Methods of Information Geometry*. American Mathematical Society, 2001.

S. Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18:14–20, 1972.

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.

A. Banerjee, I. Dhillon, J. Ghosh, and S. Merugu. An information theoretic analysis of maximum likelihood mixture estimation for exponential families. In *Proc. 21st International Conference on Machine Learning (ICML)*, 2004a.

A. Banerjee, X. Guo, and H. Wang. Optimal Bregman prediction and Jensen's equality. In *Proc. International Symposium on Information Theory (ISIT)*, 2004b.

A. Banerjee, X. Guo, and H. Wang. On the optimality of conditional expectation as a Bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669, July 2005.

O. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. Wiley Publishers, 1978.

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

T. Berger. *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall, 1971.

T. Berger and J. D. Gibson. Lossy source coding. *IEEE Transactions on Information Theory*, 44 (6):2691–2723, 1998.

J. Bilmes. A gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden markov models. Technical Report ICSI-TR-97-02, University of Berkeley, 1997.

R. E. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, 18:460–473, 1972.

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

A. Buzo, A. H. Gray, R. M. Gray, and J. D. Markel. Speech coding based upon vector quantization. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(5):562–574, 1980.

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

M. Collins. The EM algorithm. Technical report, Department of Computer and Information Science, University of Pennsylvania, 1997.

M. Collins, S. Dasgupta, and R. Schapire. A generalization of principal component analysis to the exponential family. In *Proc. of the 14th Annual Conference on Neural Information Processing Systems (NIPS)*, 2001.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

I. Csiszár. On the computation of rate distortion functions. *IEEE Transactions of Information Theory*, IT-20:122:124, 1974.

I. Csiszár. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *The Annals of Statistics*, 19(4):2032–2066, 1991.

I. Csiszár. Generalized projections for non-negative functions. *Acta Mathematica Hungarica*, 68 (1-2):161–185, 1995.

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

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

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.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley & Sons, 2001.

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

J. Forster and M. K. Warmuth. Relative expected instantaneous loss bounds. In *Proc. of the 13th Annual Conference on Computational Learing Theory (COLT)*, pages 90–99, 2000.

A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1992.

R. M. Gray and D. L. Neuhoff. Quantization. *IEEE Transactions on Information Theory*, 44(6): 2325–2383, 1998.

P. D. Grünwald and A. Dawid. Game theory, maximum entropy, minimum discrepancy, and robust bayesian decision theory. *Annals of Statistics*, 32(4), 2004.

P. D. Grünwald 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.

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

D. Kazakos and P.P. Kazakos. Spectral distance measures between Gaussian processes. *IEEE Transactions on Automatic Control*, 25(5):950–959, 1980.

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

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

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

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

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.

M. Palus. On entropy rates of dynamical systems and Gaussian processes. *Physics Letters A*, 227 (5-6):301–308, 1997.

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.

C. R. Rao. Diversity and dissimilarity coefficients: A unified approach. *Journal of Theoretical Population Biology*, 21:24–43, 1982.

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

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

K. Rose. A mapping approach to rate-distortion computation and analysis. *IEEE Transactions on Information Theory*, 40(6):1939–1952, 1994.

N. Slonim and Y. Weiss. Maximum likelihood and the information bottleneck. In *Proc. 16th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 335–342, 2002.

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

N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Proc. of the 37th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. Technical Report TR 649, Dept. of Statistics, University of California at Berkeley, 2003.

S. Wang and D. Schuurmans. Learning continuous latent variable models with Bregman divergences. In *Proc. IEEE International Conference on Algorithmic Learning Theory*, 2003.