An Information Theoretic Analysis of Maximum Likelihood Mixture Estimation for Exponential Families

Arindam Banerjee ABANERJE@ECE.UTEXAS.EDU
Dept of ECE, Univ of Texas at Austin, Austin, TX 78712

Inderjit Dhillon INDERJIT@CS.UTEXAS.EDU
Dept of CS, Univ of Texas at Austin, Austin, TX 78712

Joydeep Ghosh GHOSH@ECE.UTEXAS.EDU
Srujana Merugu MERUGU@ECE.UTEXAS.EDU
Dept of ECE, Univ of Texas at Austin, Austin, TX 78712

Abstract

An important task in unsupervised learning is maximum likelihood mixture estimation (MLME) for exponential families. In this paper, we prove a mathematical equivalence between this MLME problem and the rate distortion problem for Bregman divergences. We also present new theoretical results in rate distortion theory for Bregman divergences. Further, an analysis of the problems as a trade-off between compression and preservation of information is presented that yields the information bottleneck method as an interesting special case.

1. Introduction

An important task in unsupervised learning is maximum likelihood mixture estimation (MLME) (Redner & Walker, 1984) for exponential families, which are the most widely used class of parametric distributions for statistical analysis and learning. Recent years have seen a lot of interest and research on information theoretic formulations of unsupervised learning (Tishby et al., 1999; Dhillon et al., 2003) that draw from rate distortion theory (Berger, 1971; Cover & Thomas, 1991). In this paper, we prove a mathematical equivalence between the MLME problem for exponential families and the rate distortion problem for Bregman divergences (Azoury & Warmuth, 2001), which include a large class of popular distortion functions such as squared Euclidean distortion, KL-divergence, Itakura-Saito distance, Mahalanobis distance, generalized I-divergence, etc. We achieve the equivalence by deriving new theoretical results in rate distortion theory for Bregman divergences, and using a recent bijection theorem between exponential families and Bregman divergences. It is interesting to note that special cases of this equivalence have been observed in the literature (Kearns et al., 1997; Slonim & Weiss, 2002).

Theoretical results in rate distortion theory typically involve squared Euclidean distortion. In this paper, we generalize certain results known for squared Euclidean distortion (Rose, 1994) to all Bregman divergences. In particular, we prove a new lower bound on the rate distortion function applicable for all Bregman divergences. Further, we show that at any given distortion level either (a) the rate distortion function is equal to the new lower bound, in which case it can be analytically computed, or (b) the optimal support[1] of the reproduction random variable is finite under mild assumptions, so that the rate distortion function can be algorithmically computed. The result is significant since it provides a way to solve the rate distortion problem for *all* Bregman divergences. Further, it theoretically justifies the use of finite reproduction alphabets for lossy compression and motivates the equivalence with finite mixture models.

We also analyze an alternative viewpoint of the rate distortion problem for Bregman divergences in terms of a trade-off between compression and preservation of *Bregman information*, a concept recently proposed in (Banerjee et al., 2004) that includes

Proceedings of the 21st International Conference on Machine Learning, Banff, Canada, 2004. Copyright 2004 by the authors.

[1] Support of $X \sim p(x)$ is the set $\mathcal{X}_s = \{x : p(x) > 0\}$.

variance, mutual information, etc. as special cases. We discuss an interesting special case of this setting, namely the trade-off between compression and preservation of mutual information, which has become popular in the recent past as the information bottleneck method (Tishby et al., 1999).

We begin by defining Bregman divergences. Let ϕ be a real-valued strictly convex function defined on the convex set $S = \mathrm{dom}(\phi)(\subseteq \mathbb{R}^m)$, the domain of ϕ, such that ϕ is differentiable on $\mathrm{int}(S)$, the interior of S (Rockafeller, 1970). The *Bregman divergence* $d_\phi : S \times \mathrm{int}(S) \mapsto [0, \infty)$ is defined as $d_\phi(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}, \nabla\phi(\mathbf{y})\rangle$, where $\nabla\phi$ is the gradient of ϕ. The squared Euclidean distance and KL-divergence are examples of Bregman divergences, corresponding to $\phi(\mathbf{x}) = \langle \mathbf{x}, \mathbf{x}\rangle,\ \mathbf{x} \in \mathbb{R}^m$ and $\phi(\mathbf{p}) = \sum_{j=1}^m p_j \log p_j,\ \mathbf{p} \in m$-simplex respectively.

2. Rate Distortion with Bregman Divergences

Rate distortion theory (Berger, 1971) 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 & 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

$$R(D) = \min_{p(\hat{x}|x): E_{X,\hat{X}}[d(X,\hat{X})] \leq D} I(X; \hat{X}) \ , \qquad (2.1)$$

where $I(X; \hat{X})$ is the mutual information of X and $\hat{X}$. Note that in general X and $\hat{X}$ could either be discrete or continuous random variables. As appropriate, the expectations will involve summations or integrals where p represents the corresponding probability mass function or probability density function.

The rate distortion problem (2.1) is a convex optimization problem and can be solved using the Blahut-Arimoto algorithm (Cover & Thomas, 1991). However, numerical computation of the rate distortion function through the Blahut-Arimoto algorithm is often infeasible in practice, primarily due to the lack of knowledge of the optimal support of the reproduction random variable. Analytic closed form expressions of the rate distortion function exist only for certain well-behaved source and distortion measure combinations. Therefore, exact computation of the rate distortion function has remained a difficult problem.

Rose (1994) showed that the rate distortion problem for the squared Euclidean distortion and for any source whose support is a bounded set can be solved either analytically or through a numerical computation technique called the mapping approach (Rose, 1994). In this paper, we generalize this result to all Bregman divergences. In order to define a distortion measure based on a Bregman divergence, we require a concrete representation of the source and reproduction alphabets. In general, $\mathcal{X}$ and $\hat{\mathcal{X}}$ may be abstract sets, and hence, we consider a concrete sufficient statistic representation[2] of these random variables. In particular, let T be the sufficient statistic function so that $\boldsymbol{Z} = T(X)$ and $\hat{\boldsymbol{Z}} = T(\hat{X})$ are the sufficient statistic representations of the source and reproduction random variables. The distortion measure $d : \mathcal{X} \times \hat{\mathcal{X}} \mapsto \mathbb{R}_+$ can now be defined as a Bregman divergence in the sufficient statistic space, i.e., $d(x, \hat{x}) = d_\phi(\mathbf{z}, \hat{\mathbf{z}}),\ \mathbf{z} = T(x), \hat{\mathbf{z}} = T(\hat{x})$. The rate distortion problem can now be formulated entirely in the sufficient statistic space as follows:

$$\min_{p(\hat{\mathbf{z}}|\mathbf{z}): E_{\boldsymbol{Z},\hat{\boldsymbol{Z}}}[d_\phi(\boldsymbol{Z},\hat{\boldsymbol{Z}})] \leq D} I(\boldsymbol{Z}; \hat{\boldsymbol{Z}}) \ . \qquad (2.2)$$

Now, we present a new analytic lower bound on the rate distortion function for Bregman divergences, which we call the *Shannon-Bregman lower bound.*

Theorem 1 [3] *The rate distortion function for a source $\boldsymbol{Z} \sim p(\mathbf{z})$ and a Bregman divergence d_ϕ is always lower bounded by the* **Shannon-Bregman lower bound** *$R_L(D)$ defined as*

$$R_L(D) = H(\boldsymbol{Z}) + \sup_{\gamma \geq 0} \left\{-\gamma D + E_{\boldsymbol{Z}}[\log f_{\gamma\phi}(\boldsymbol{Z})]\right\} ,$$

where $H(\boldsymbol{Z})$ denotes the (differential) entropy of $\boldsymbol{Z}$ and $f_{\gamma\phi}$ is the unique function that satisfies

$$\int_{\mathrm{dom}(\phi)} \exp(-d_{\gamma\phi}(\mathbf{t}, \mu)) f_{\gamma\phi}(\mathbf{t})\, d\mathbf{t} = 1,\ \forall \mu \in \mathrm{dom}(\phi) \ .$$

The Shannon-Bregman lower bound plays the same role for Bregman divergences as the *Shannon lower*

[2] If $X \sim p_\theta(\mathbf{x})$, parameterized by θ, then $T(X)$ is a sufficient statistic if the (conditional) distribution of $X|T(X)$ is independent of θ.

[3] We omit proofs due to lack of space.

bound for "difference" distortion measures (Berger, 1971), i.e., distortions of the form $d(x, y) = \rho(x - y)$ for any non-negative function $\rho(\cdot)$. Rose (1994) used the Shannon lower bound for squared Euclidean distortion to reduce the rate distortion problem into two mutually exclusive solvable cases. More specifically, Rose (1994) showed that for squared Euclidean distortion and any source whose support is a bounded set, either (a) the rate-distortion function equals the Shannon lower bound, or (b) the optimal support of the reproduction random variable is finite, in which case the rate distortion function can be numerically computed using the mapping approach. Our following theorem states a significantly more general result.

Theorem 2 *Consider the rate distortion problem for a source $\boldsymbol{Z} \sim p(\mathbf{z})$ and a Bregman divergence d_ϕ. Let $\hat{\mathcal{Z}}_s(D)$ be the support of the optimal reproduction random variable for an expected distortion D. If $\hat{\mathcal{Z}}_s(D)$ contains an accumulation point, then $R(D) = R_L(D)$.*

If $R(D) > R_L(D)$, then $\hat{\mathcal{Z}}_s(D)$ does not contain an accumulation point. Further, if the source alphabet is a bounded set, $\hat{\mathcal{Z}}_s(D)$ can be shown to be a finite set using the Bolzano-Weierstrass theorem. Thus, the rate distortion problem for Bregman divergences and sources with bounded support can be divided into two cases, of which the first one can be solved analytically using the Shannon-Bregman lower bound and the second one requires a numerical solution involving a finite reproduction alphabet. Therefore, in the rest of the section, we focus only on the second case. In fact, we solve a simpler problem assuming that the cardinality of the optimal support of the reproduction random variable is known. This assumption is reasonable since deterministic annealing methods (Rose, 1998) can be applied to empirically determine the appropriate cardinality at any distortion value.

2.1. Rate Distortion for Fixed Finite Cardinality Reproduction Alphabet

In this subsection, we consider the *joint* problem of finding the optimal support $\hat{\mathcal{Z}}_s$ of the reproduction random variable with $|\hat{\mathcal{Z}}_s| = k$ as well as the optimal probabilistic assignments $p(\hat{\mathbf{z}}|\mathbf{z})$ that achieve the rate-distortion function for a given source. When the distortion measure is a Bregman divergence, the problem can be formally stated as follows:

$$\min_{\substack{\hat{\mathcal{Z}}_s,\ p(\hat{\mathbf{z}}|\mathbf{z}) \\ |\hat{\mathcal{Z}}_s| = k}} \{I(\boldsymbol{Z}; \hat{\boldsymbol{Z}}) + \beta_D E_{\boldsymbol{Z},\hat{\boldsymbol{Z}}}[d_\phi(\boldsymbol{Z}, \hat{\boldsymbol{Z}})], \qquad (2.3)$$

where β_D is the optimal Lagrange multiplier that depends on the chosen tolerance level D of the expected distortion. We shall refer to problem (2.3) as rate distortion with a support of fixed finite cardinality (RDFC). It is important to note that unlike the original rate distortion problem (2.1), the RDFC problem is not a convex optimization problem, since it involves optimizing over both $\hat{\mathcal{Z}}_s$ and $p(\hat{\mathbf{z}}|\mathbf{z})$. Hence, it is difficult to obtain the globally optimal solution. However, since the minimization is over two sets of arguments, namely $p(\hat{\mathbf{z}}|\mathbf{z})$ and $\hat{\mathcal{Z}}_s$, the objective function in (2.3) can be greedily minimized by iteratively optimizing over the individual arguments yielding a solution that is locally optimal.

Lemma 1 (Cover & Thomas, 1991) [4] *The solution to the problem*

$$\min_{p(\hat{\mathbf{z}}|\mathbf{z})} \{I(\boldsymbol{Z}; \hat{\boldsymbol{Z}}) + \beta_D E_{\boldsymbol{Z},\hat{\boldsymbol{Z}}}[d_\phi(\boldsymbol{Z}, \hat{\boldsymbol{Z}})],$$

for a fixed $\hat{\mathcal{Z}}_s$ is given by

$$p(\hat{\mathbf{z}}|\mathbf{z}) = \frac{p(\hat{\mathbf{z}})}{N(\mathbf{z}, \beta_D)} \exp(-\beta_D d_\phi(\mathbf{z}, \hat{\mathbf{z}})),$$

where $p(\hat{\mathbf{z}}) = E_{\boldsymbol{Z}|\hat{\mathbf{z}}}[p(\boldsymbol{Z})]$ and $N(\mathbf{z}, \beta_D)$ is the partition function.

Lemma 2 *The solution to the problem,*

$$\min_{\hat{\mathcal{Z}}_s} \{I(\boldsymbol{Z}; \hat{\boldsymbol{Z}}) + \beta_D E_{\boldsymbol{Z},\hat{\boldsymbol{Z}}}[d_\phi(\boldsymbol{Z}, \hat{\boldsymbol{Z}})],$$

for fixed probabilistic assignments $p(\hat{\mathbf{z}}|\mathbf{z})$ is given by

$$\hat{\mathbf{z}}^* = E_{\boldsymbol{Z}|\hat{\mathbf{z}}}[\boldsymbol{Z}].$$

Lemma 1 follows directly from the self-consistent equations for the solution of the rate distortion problem (Cover & Thomas, 1991) while Lemma 2 follows from the result that the expectation of a random variable minimizes its expected Bregman divergence to a point (Banerjee et al., 2004). Based on these results, we obtain an alternate minimization algorithm for computing the rate distortion function (Algorithm 1), guaranteed to achieve local optimality.

Theorem 3 *The alternate minimization algorithm (Algorithm 1) for the RDFC problem (2.3) converges to a solution that is locally optimal, i.e., the objective function in (2.3) cannot be decreased by either changing $p(\hat{\mathbf{z}}|\mathbf{z})$ or $\hat{\mathcal{Z}}_s$.*

[4] Lemmas 1 and 2 hold irrespective of whether $\boldsymbol{Z}$ is a continuous or discrete random variable, but practical computation (Algorithm 1) is feasible only for finite $\mathcal{Z}$.

Algorithm 1 Computation of Rate Distortion Curve for Bregman Divergences

```
Input: $\boldsymbol{Z} \sim p(\mathbf{z})$ over $\mathcal{Z} = \{\mathbf{z}_i\}_{i=1}^n \subset \mathrm{dom}(\phi) \subseteq \mathbb{R}^m$,
  Bregman divergence $d_\phi$, $k = |\hat{\mathcal{Z}}_s|$, variational parameter $\beta$ corresponding to a point on $R(D)$ curve
Output: $\hat{\mathcal{Z}}_s^* = \{\hat{\mathbf{z}}_h\}_{h=1}^k$, $P^* = \{\{p(\hat{\mathbf{z}}_h|\mathbf{z}_i)\}_{h=1}^k\}_{i=1}^n$
  that (locally) optimizes (2.3), rate-distortion trade-off $(R_\beta, D_\beta)$ at $\beta$.
Method:
  Initialize with some $\{\hat{\mathbf{z}}_h\}_{h=1}^k \subset \mathrm{dom}(\phi)$
  repeat
    {Blahut Arimoto Step ($p(\hat{\mathbf{z}}|\mathbf{z})$ using Lemma 1)}
    repeat
      for $i = 1$ to $n$ do
        for $h = 1$ to $k$ do
          $p(\hat{\mathbf{z}}_h|\mathbf{z}_i) \leftarrow \frac{p(\hat{\mathbf{z}}_h)}{N(\mathbf{z}_i,\beta)} \exp(-\beta d_\phi(\mathbf{z}_i, \hat{\mathbf{z}}_h))$,
        end for
      end for
      for $h = 1$ to $k$ do
        $p(\hat{\mathbf{z}}_h) \leftarrow \sum_{i=1}^n p(\hat{\mathbf{z}}_h|\mathbf{z}_i)p(\mathbf{z}_i)$
      end for
    until convergence
    {Support Estimation Step ($\hat{\mathcal{Z}}_s$ using Lemma 2)}
    for $h = 1$ to $k$ do
      $\hat{\mathbf{z}}_h \leftarrow \sum_{i=1}^n p(\mathbf{z}_i|\hat{\mathbf{z}}_h)\mathbf{z}_i$
    end for
  until convergence
  Compute $D_\beta = \sum_{\mathbf{z},\hat{\mathbf{z}}} p(\mathbf{z})p(\hat{\mathbf{z}}|\mathbf{z})d_\phi(\mathbf{z}, \hat{\mathbf{z}})$
  Compute $R_\beta = \sum_{\mathbf{z},\hat{\mathbf{z}}} p(\mathbf{z})p(\hat{\mathbf{z}}|\mathbf{z}) \log \frac{p(\hat{\mathbf{z}}|\mathbf{z})}{p(\hat{\mathbf{z}})}$
```

3. Equivalence with Mixture Estimation for Exponential Families

The maximum likelihood mixture estimation (MLME) problem involves finding the mixture of k distributions from a specified parametric family $\mathcal{F}$ that best fits the observed data in terms of the log-likelihood. This problem seems different from the rate distortion problem since MLME only assumes knowledge of a finite set of independent samples of the random variable and not the actual distribution. However, as we shall show, it is equivalent to the rate distortion problem when the source distribution in the rate distortion setting equals the empirical distribution over the sampled data points.

The standard way to address the mixture estimation problem is to introduce a hidden random variable associated with the choice of mixture component. Let $\mathcal{Z}_d$ be the finite set of independent samples corresponding to the observed random variable $\boldsymbol{Z}$. Let $\hat{\boldsymbol{Z}}$ be the hidden random variable corresponding to the choice of the mixture component and taking values in $\hat{\mathcal{Z}}_s$ with $|\hat{\mathcal{Z}}_s| = k$. The mixture distribution $p(\mathbf{z})$ can be viewed as the marginal induced from the joint distribution $p(\mathbf{z}, \hat{\mathbf{z}})$ such that each conditional distribution $p(\mathbf{z}|\hat{\mathbf{z}})$ belongs to the specified parametric family $\mathcal{F}$. The mixture estimation problem can be formally stated as the problem of maximizing the average incomplete log-likelihood of the data, i.e., $\frac{1}{n}\sum_{\mathbf{z}\in\mathcal{Z}_d} \log p(\mathbf{z})$, over all mixture distributions $p(\mathbf{z})$ consisting of k component distributions from $\mathcal{F}$. The MLME problem has been shown (Neal & Hinton, 1998; Rose, 1998) to be equivalent to the problem of minimizing the variational free energy of a statistical system, where the physical states correspond to the values of the unknown random variable $\hat{\boldsymbol{Z}}$ and the energy of each state is given by the negative joint log-likelihood $(-\log p(\mathbf{z}, \hat{\mathbf{z}}))$. The negative of this variational free energy can be expressed as the sum of the entropy of the conditional distribution $p(\hat{\mathbf{z}}|\mathbf{z})$ and the expected complete log-likelihood with respect to $p(\mathbf{z}, \hat{\mathbf{z}})$. Therefore, the minimum free energy problem and hence, the MLME problem can be expressed as

$$\min_{\substack{\hat{\mathcal{Z}}_s,\ p(\hat{\mathbf{z}}|\mathbf{z}) \\ |\hat{\mathcal{Z}}_s|=k}} \left\{ -E_{\boldsymbol{Z},\hat{\boldsymbol{Z}}}[\log p(\boldsymbol{Z}, \hat{\boldsymbol{Z}})] - H(\hat{\boldsymbol{Z}}|\boldsymbol{Z}) \right\}, \quad (3.4)$$

where $\boldsymbol{Z} \sim p_d(\mathbf{z})$, the empirical distribution over the sample set $\mathcal{Z}_d$, the joint distribution $p(\mathbf{z}, \hat{\mathbf{z}}) = p(\hat{\mathbf{z}})p(\mathbf{z}|\hat{\mathbf{z}})$ such that $p(\mathbf{z}|\hat{\mathbf{z}}) \in \mathcal{F}$ and the minimization is performed over $\hat{\mathcal{Z}}_s$ and $p(\hat{\mathbf{z}}|\mathbf{z})$, which uniquely determine the mixture distribution $p(\mathbf{z})$.

Now, consider the case when the specified parametric family $\mathcal{F}$ is an exponential family $\mathcal{F}_\psi$ with a log-partition function ψ (Azoury & Warmuth, 2001) so that $p_\psi(\mathbf{z}|\theta) \in \mathcal{F}_\psi$ is given by

$$p_\psi(\mathbf{z}|\theta) = \exp\left(\langle \mathbf{z}, \theta \rangle - \psi(\mathbf{z})\right) ,$$

over some measure $\nu(\mathbf{z})$ where $\theta \in \mathrm{dom}(\psi)$ is the natural parameter. Although, the MLME problem (3.4) assumes that $\mathcal{F}_\psi$ is fully specified, typically, only a meta family $\mathcal{M}_\psi$ consisting of scaled versions of $\mathcal{F}_\psi$ is specified. To make this more precise, we define the scaled versions of $\mathcal{F}_\psi$ as the parametric families $\mathcal{F}_\psi^{(\beta)}$, $\beta \geq 0$, such that $p_\psi^{(\beta)}(\mathbf{z}|\theta) \in \mathcal{F}_\psi^{(\beta)}$ is given by

$$p_\psi^{(\beta)}(\mathbf{z}|\theta) \quad \propto \quad (p_\psi(\mathbf{z}|\theta))^\beta , \qquad (3.5)$$

where $p_\psi(\mathbf{z}|\theta) \in \mathcal{F}_\psi$ and $\mathcal{M}_\psi = \{\mathcal{F}_\psi^{(\beta)}, \beta \geq 0\}$. It can be shown that each $\mathcal{F}_\psi^{(\beta)}$ is itself an exponential family with a log-partition function $\psi_\beta(\theta) = \beta\psi(\theta/\beta)$. For example, the set of all unit variance Gaussian distributions over $\mathbb{R}$ is an exponential family F_ψ with $\psi(\theta) = \theta^2/2$. All *constant* variance Gaussian families are the scaled versions of this F_ψ, and M_ψ is the set of all the scaled versions. To perform mixture modeling, we need to choose a particular member of the meta family M_ψ, i.e., a particular value for the scaling factor β. Usually, β is implicitly chosen to be 1 with F_ψ being a canonical representation of the meta family M_ψ. In practice, appropriate choice of β has led to improved results on natural datasets, e.g., see (Nigam, 2001) for scaled families on the mixture of multinomials model.

Using (3.5) in (3.4), we note that the scaling factor β determines the relative importance of expected complete log-likelihood and the assignment entropy terms in the maximum likelihood problem (3.4), and consequently, the degree of "softness" in the assignments $p(\hat{\mathbf{z}}|\mathbf{z})$. In particular, the assignment entropy term $H(\hat{\boldsymbol{Z}}|\boldsymbol{Z})$ is significant for low β leading to an almost uniform assignment, whereas for high β, the entropy term becomes insignificant resulting in hard assignments between $\boldsymbol{Z}$ and $\hat{\boldsymbol{Z}}$. It is, therefore, important to choose β appropriately based on the desired accuracy and softness constraints. We present an information theoretic analysis for making this choice by demonstrating an equivalence between the RDFC problem for a specified distortion constraint and the MLME problem based on a particular member of a meta exponential family with scaling factor β that depends on D.

3.1. Equivalence Theorem

We start by reviewing a bijection result involving Bregman divergences and exponential families. Since the log-partition function ψ of an exponential family is a convex function (Azoury & Warmuth, 2001), Legendre duality (Rockafeller, 1970) can be invoked to establish the following bijection theorem.

Theorem 4 (Banerjee et al., 2004) *Let $P_{(\psi,\boldsymbol{\theta})}$ be an exponential probability distribution function with log-partition function $\psi(\boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \Gamma$ is the natural parameter. Let $\boldsymbol{\mu}$ be the corresponding expectation parameter. Let (ψ, Γ) and (ϕ, S) be Legendre conjugates, and let d_ϕ be the Bregman divergence derived from ϕ. Then,*

$$dP_{(\psi,\boldsymbol{\theta})}(\mathbf{z}) = \exp(-d_\phi(\mathbf{z},\boldsymbol{\mu}))d\nu_\phi(\mathbf{z}), \tag{3.6}$$

where ν_ϕ is a uniquely determined measure on S. Hence, there is a bijection between exponential distributions $P_{(\psi,\boldsymbol{\theta})}$ and Bregman divergences $d_\phi(\cdot,\boldsymbol{\mu})$.

Note that the distribution $P_{(\psi,\theta)}$ corresponds to the exponential density $p_\psi(\mathbf{z}|\theta) \in \mathcal{F}_\psi$. Based on the above theorem, the conditional distribution $p_\psi(\mathbf{z}|\hat{\mathbf{z}}) \in \mathcal{F}_\psi$ is given by $p_\psi(\mathbf{z}|\hat{\mathbf{z}}) = \exp(-d_\phi(\mathbf{z},\hat{\mathbf{z}}))$, where $\hat{\mathbf{z}}$ is the expectation parameter, ϕ is the Legendre conjugate of ψ and d_ϕ is the Bregman divergence derived from ϕ. Hence, the Bregman divergence $d_\phi(\mathbf{z},\hat{\mathbf{z}})$ is related to the negative log-likelihood $(-\log p_\psi(\mathbf{z}|\hat{\mathbf{z}}))$ of the corresponding exponential distribution. We use this observation to prove the following equivalence.

Theorem 5 *Consider a source $\boldsymbol{Z} \sim p_d(\mathbf{z})$. Then, the RDFC problem (2.3) for $\boldsymbol{Z}$ with Bregman distortion d_ϕ, tolerable expected distortion D with $|\hat{\mathcal{Z}}_s| = k$ is equivalent to the MLME problem (3.4) for a mixture model with k distributions from the scaled exponential family $\mathcal{F}_\psi^{(\beta_D)}$, where β_D is the optimal Lagrange multiplier for the RDFC problem and ψ is the Legendre conjugate of ϕ.*

Proof: It is sufficient to compare the objective functions of the problems (2.3) and (3.4) as both are minimization problems with identical arguments and constraints. For the RDFC problem (2.3) based on Bregman divergence d_ϕ and tolerable level of distortion D, the objective function is given by

$$\begin{aligned}
&J_{RDFC}(\hat{\mathcal{Z}}_s, p(\hat{\mathbf{z}}|\mathbf{z})) \\
&= I(\boldsymbol{Z};\hat{\boldsymbol{Z}}) + \beta_D E_{\boldsymbol{Z},\hat{\boldsymbol{Z}}}[d_\phi(\boldsymbol{Z},\hat{\boldsymbol{Z}})] \\
&= E_{\boldsymbol{Z},\hat{\boldsymbol{Z}}}[\log p(\hat{\boldsymbol{Z}}|\boldsymbol{Z}) - \log p(\hat{\boldsymbol{Z}}) + \beta_D d_\phi(\boldsymbol{Z},\hat{\boldsymbol{Z}})]
\end{aligned}$$

Since ψ is the Legendre conjugate of ϕ, the exponential family $\mathcal{F}_\psi$ corresponding to the Bregman divergence d_ϕ is given by (3.6) and the scaled version $\mathcal{F}_\psi^{(\beta_D)}$ is obtained using (3.5). Hence, the objective function of the MLME problem (3.4) based on the exponential family $\mathcal{F}_\psi^{(\beta_D)}$ is given by

$$\begin{aligned}
&J_{MLME}(\hat{\mathcal{Z}}_s, p(\hat{\mathbf{z}}|\mathbf{z})) \\
&= -E_{\boldsymbol{Z},\hat{\boldsymbol{Z}}}[\log p(\boldsymbol{Z},\hat{\boldsymbol{Z}})] - H(\hat{\boldsymbol{Z}}|\boldsymbol{Z}) \\
&= E_{\boldsymbol{Z},\hat{\boldsymbol{Z}}}[\log p(\hat{\boldsymbol{Z}}|\boldsymbol{Z}) - \log p(\hat{\boldsymbol{Z}}) + \beta_D d_\phi(\boldsymbol{Z},\hat{\boldsymbol{Z}})]
\end{aligned}$$

The objective functions J_{MLME} and J_{RDFC} are exactly same and hence, the equivalence follows. ■

The equivalence theorem gives an information theoretic recipe for choosing the appropriate scaled exponential family for a mixture modeling based on the desired model accuracy constraints. In particular, the Bregman distortion constraint $E_{\boldsymbol{Z},\hat{\boldsymbol{Z}}}[d_\phi(\boldsymbol{Z},\hat{\boldsymbol{Z}})] \leq D$, which is equivalent to a conditional entropy constraint $H(\boldsymbol{Z}|\hat{\boldsymbol{Z}}) \leq D$, specifies the desired level of model accuracy. Then, the appropriate exponential family for mixture modeling is $\mathcal{F}_{\psi}^{(\beta_D)}$ where β_D is the optimal Lagrange multiplier of the RDFC problem. This follows since the optimal solution $(p(\hat{\mathbf{z}}|\mathbf{z}), \hat{\mathcal{Z}}_s)$ of the RDFC problem exactly satisfies the condition $p(\mathbf{z}|\hat{\mathbf{z}}) \in \mathcal{F}_{\psi}^{(\beta_D)}$.

The objective function of the MLME problem corresponding to $F_{\psi}^{(\beta)}$ can be written as $I(\boldsymbol{Z}, \hat{\boldsymbol{Z}}) + \beta E_{\boldsymbol{Z},\hat{\boldsymbol{Z}}}[d_\phi(\boldsymbol{Z}, \hat{\boldsymbol{Z}})]$. Therefore, solving the MLME problem based on any exponential family $\mathcal{F}_{\psi}^{(\beta)}$, $\beta \leq \beta_D$ such that $E_{\boldsymbol{Z},\hat{\boldsymbol{Z}}}[d_\phi(\boldsymbol{Z}, \hat{\boldsymbol{Z}})] \leq D$ yields a solution identical to that of the unconstrained MLME problem based on $\mathcal{F}_{\psi}^{(\beta_D)}$, and the equivalent RDFC problem. In particular, the constrained MLME problem based on $\mathcal{F}_\psi \equiv \mathcal{F}_{\psi}^{(1)}$ such that $E_{\boldsymbol{Z},\hat{\boldsymbol{Z}}}[d_\phi(\boldsymbol{Z}, \hat{\boldsymbol{Z}})] \leq D$ is equivalent to RDFC problem for all D such that $\beta_D \geq 1$. Further, the constrained MLME problem based on $F_{\psi}^{(0^+)}$ is equivalent to the RDFC problem for all D.

The equivalence also suggests that that the RDFC problem can also be solved by the expectation maximization (EM) algorithm (Redner & Walker, 1984). The update equations in both the algorithms are identical, the only difference being the order in which they are executed, i.e., the two algorithms correspond to two different ways of cyclic minimization. Both the algorithms are guaranteed to converge to a locally optimal solution, but the actual solutions could be different. In fact, any algorithm that alternates between the three updates, viz, $p(\hat{\mathbf{z}}|\mathbf{z}), p(\hat{\mathbf{z}})$ and $\hat{\mathbf{z}}$, will have similar guarantees. However, this class of algorithms have two drawbacks. First, the algorithms assume that the optimal cardinality $k(D)$ of the reproduction alphabet or the mixture model for a given tolerable distortion D is known though it is not the case in practice. Secondly, the algorithms are guaranteed to provide only locally optimal solutions. A practical technique that addresses these deficiencies is the deterministic annealing approach that starts with a high value of D (i.e., a low positive β_D), where the optimal support of the reproduction random variable and the mixture model have cardinality one, and slowly decreases the tolerable distortion level D (i.e., increases β_D) while detecting the phase transitions corresponding to changes in the cardinality. For details, see Rose (1998).

4. Compression vs. Bregman Information Trade-off

In this section, we provide an alternate view of the RDFC problem as a lossy compression problem where the objective is to balance the trade-off between compression and loss in *Bregman information* (Banerjee et al., 2004). Further, we show that the information bottleneck method (Tishby et al., 1999) is an interesting special case of this setting for a particular choice of sufficient statistic vector and Bregman divergence.

The Bregman information $I_\phi(\boldsymbol{Z})$ of any random variable $\boldsymbol{Z}$ for a specified Bregman divergence d_ϕ is defined as the expected Bregman divergence between the random variable $\boldsymbol{Z}$ and its expectation, i.e., $I_\phi(\boldsymbol{Z}) = E_{\boldsymbol{Z}}[d_\phi(\boldsymbol{Z}, E_{\boldsymbol{Z}}[\boldsymbol{Z}])]$, where the expectations are with respect to the distribution of $\boldsymbol{Z}$. Examples of Bregman information include variance for squared Euclidean distortion and mutual information for KL-divergence. Since $E_{\boldsymbol{Z}}[\boldsymbol{Z}]$ minimizes the expected Bregman divergence of a random variable to a point (Banerjee et al., 2004), i.e., $E_{\boldsymbol{Z}}[\boldsymbol{Z}] = \operatorname{argmin}_{\mathbf{a}}\ E_{\boldsymbol{Z}}[d_\phi(\boldsymbol{Z}, \mathbf{a})]$, the Bregman information also corresponds to the minimum achievable expected distortion at zero rate, which indicates the uncertainty or the "information" contained in the random variable. Similarly, the Bregman information of the reproduction random variable $\hat{\boldsymbol{Z}}$ is given by $I_\phi(\hat{\boldsymbol{Z}}) = E_{\hat{\boldsymbol{Z}}}[d_\phi(\hat{\boldsymbol{Z}}, E_{\hat{\boldsymbol{Z}}}[\hat{\boldsymbol{Z}}])]$, where the expectations are with respect to the distribution of $\hat{\boldsymbol{Z}}$ given by $p(\hat{\boldsymbol{Z}}) = E_{\boldsymbol{Z}|\hat{\boldsymbol{Z}}}[p(\boldsymbol{Z})]$. Further, choosing $\hat{\boldsymbol{Z}} = E_{\boldsymbol{Z}|\hat{\boldsymbol{Z}}}[\boldsymbol{Z}]$ implies that

$$E_{\hat{\boldsymbol{Z}}}[\hat{\boldsymbol{Z}}] = E_{\hat{\boldsymbol{Z}}}[E_{\boldsymbol{Z}|\hat{\boldsymbol{Z}}}[\boldsymbol{Z}]] = E_{\boldsymbol{Z},\hat{\boldsymbol{Z}}}[\boldsymbol{Z}] = E_{\boldsymbol{Z}}[\boldsymbol{Z}] = \mu\ .$$

Hence, we have $I_\phi(\boldsymbol{Z}) = E_{\boldsymbol{Z}}[d_\phi(\boldsymbol{Z}, \mu)]$ and $I_\phi(\hat{\boldsymbol{Z}}) = E_{\hat{\boldsymbol{Z}}}[d_\phi(\hat{\boldsymbol{Z}}, \mu)]$.

The alternate view of the RDFC problem is based on the observation that the reproduction random variable $\hat{\boldsymbol{Z}}$ is a coarser representation of the source random variable $\boldsymbol{Z}$ with less "information" than $\boldsymbol{Z}$. In rate distortion theory, the loss in "information" is quantified by the expected Bregman distortion between $\boldsymbol{Z}$ and $\hat{\boldsymbol{Z}}$. Intuitively, if the expected Bregman distortion is low, then $\boldsymbol{Z}$ and $\hat{\boldsymbol{Z}}$ are "close" to each other and it is reasonable to expect that $\hat{\boldsymbol{Z}}$ contains most of the "information" in $\boldsymbol{Z}$. The following theorem provides a direct way of quantifying the intuitive loss in "information".

Theorem 6 *The expected Bregman distortion between the source and the reproduction random variables is exactly equal to the loss in the Bregman information due to compression, i.e.,*

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

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

The RDFC problem (2.3) can, therefore, be viewed as an optimization problem involving a trade-off between the mutual information $I(\boldsymbol{Z}; \hat{\boldsymbol{Z}})$ that measures the compression, and the loss in Bregman information $I_\phi(\boldsymbol{Z}) - I_\phi(\hat{\boldsymbol{Z}})$. Since the source random variable $\boldsymbol{Z}$ is known, the Bregman information $I_\phi(\boldsymbol{Z})$ is a constant and minimizing the expected distortion is equivalent to maximizing the Bregman information of the compressed random variable $\hat{\boldsymbol{Z}}$. Hence, this constrained form of the RDFC problem (2.3) can be written as:

$$\min_{p(\hat{\mathbf{z}}|\mathbf{z})} \{I(\boldsymbol{Z}; \hat{\boldsymbol{Z}}) - \beta I_\phi(\hat{\boldsymbol{Z}})\}, \tag{4.7}$$

where β is the variational parameter corresponding to the desired point in the rate distortion curve and $\hat{\boldsymbol{Z}} = E_{\boldsymbol{Z}|\hat{\boldsymbol{Z}}}[\boldsymbol{Z}]$. The variational parameter β also determines the trade-off between the achieved compression and the preserved Bregman information. Further, corresponding to each rate distortion curve, one can obtain a compression vs. Bregman information curve where the achieved compression is quantified by the rate and the preserved Bregman information is negatively related to the expected distortion.

4.1. Information Bottleneck Revisited

Let $Y \sim p(y),\ y \in \mathcal{Y}$ be a random variable and let the sufficient statistic random vector $\boldsymbol{Z}$ corresponding to a source X be the conditional distribution of Y given X, i.e., $\boldsymbol{Z} = p(Y|X)$. $\boldsymbol{Z}$ is a concrete representation of the source X. Similarly, the random variable $\hat{\boldsymbol{Z}} = p(Y|\hat{X})$ 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, e.g., random variables taking values over documents and words. For the above choice of sufficient statistic mapping, an additional constraint that $\hat{\boldsymbol{Z}}$ is the conditional expectation of $\boldsymbol{Z}$ leads to the lossy compression problem (4.7) where we need to find the optimal assignments that balance the trade-off between compression and the loss in Bregman information. Now, the Bregman information $I_\phi(\hat{\boldsymbol{Z}})$ of the random variable $\hat{\boldsymbol{Z}}$ taking values over the set of conditional distributions $\{p(Y|\hat{x})\}$ with probability $p(\hat{x})$ is same as the mutual information $I(\hat{X}; Y)$ of $\hat{X}$ and Y when the Bregman divergence is the KL-divergence (Banerjee et al., 2004). Hence, the original problem (4.7) reduces to

$$\min_{p(\hat{x}|x)} \{I(X; \hat{X}) - \beta I(\hat{X}; Y)\}, \tag{4.8}$$

since $p(\hat{x}|x) = p(\hat{\mathbf{z}}|\mathbf{z})$ and $I(X; \hat{X}) = I(\boldsymbol{Z}; \hat{\boldsymbol{Z}})$, where β is the variational parameter. This problem is exactly the same as the information bottleneck (IB) problem (Tishby et al., 1999). 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 $\hat{\boldsymbol{Z}}$ is the conditional expectation of $\boldsymbol{Z}$, i.e., $\hat{\mathbf{z}} = p(Y|\hat{x}) = E_{X|\hat{x}}[p(Y|X)] = E_{\boldsymbol{Z}|\hat{\mathbf{z}}}[\boldsymbol{Z}]$.

From the above discussion, it follows that the information bottleneck problem is a special case of the rate distortion problem (2.3). Hence, from Theorem 5, it 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 as has been previously shown in (Slonim & Weiss, 2002).

5. Related Work

Over the years, significant theoretical progress has been made in rate distortion theory (Berger, 1971; Cover & Thomas, 1991), with various results involving optimality (Rose, 1994) and algorithmic computations (Finamore & Pearlman, 1980) using finite reproduction alphabets. In this paper, we have significantly extended the work of Rose (1994) by presenting new results that provide a way to solve the rate distortion problem for *all* Bregman divergences.

Maximum likelihood mixture estimation using the expectation maximization (EM) algorithm has been widely applied to a number of unsupervised learning problems. It was observed (Redner & Walker, 1984) that the EM algorithm can be significantly simplified when applied for mixture estimation using exponential families. The variational free energy interpretation (Neal & Hinton, 1998) broadened the context of the problem.

The equivalence in Theorem 5 is based on a bijection between exponential families and Bregman divergences (Forster & Warmuth, 2000; Banerjee et al., 2004). A few special cases of this equivalence have been observed in the literature. For example, several researchers (Kearns et al., 1997; Rose, 1998) have observed the connection between Euclidean vector quantization, which corresponds to a special case of the RDFC problem, and mixture estimation problem for Gaussian distributions. Further, Slonim and Weiss (2002) established the connection between the information bottleneck method (Tishby et al., 1999), that implicitly uses KL-divergence (Gilad-Bachrach et al., 2003) in the rate distortion setting, and the maximum likelihood mixture estimation based on multinomial distributions.

6. Discussion

The results of theorems 1 and 2 give a way to compute the rate distortion function for all Bregman divergences. In fact, it maybe possible to get analytic closed form solutions of the rate distortion function for sources belonging to the exponential family with the corresponding Bregman divergence. Analytic solutions for Gaussians with squared Euclidean distance exist to encourage the exploration of this possibility. Further, the equivalence result of theorem 5 suggest that analytic results in rate distortion theory such as bounds on the rate distortion function, bounds on the appropriate output alphabet size, etc., can possibly be directly translated to useful results for mixture estimation based on the corresponding exponential family. Finally, since the bijection of theorem 4 is the key result used to establish the equivalence, further investigation of the bijection theorem may potentially lead to more connections between lossy compression and learning.

Acknowledgements: This research was supported in part by NSF grant IIS-0307792, NSF CAREER Award No. ACI-0093404, and by an IBM PhD Fellowship to Arindam Banerjee. The authors are grateful to the anonymous reviewers for their constructive comments.

References

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

Banerjee, A., Merugu, S., Dhillon, I., & Ghosh, J. (2004). Clustering with Bregman divergences. *(To appear) Proc. SIAM Intl. Conf. on Data Mining.*

Berger, T. (1971). *Rate distortion theory: A mathematical basis for data compression.* Prentice-Hall.

Collins, M., Dasgupta, S., & Schapire, R. (2001). A generalization of principal component analysis to the exponential family. *15th NIPS.*

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory.* Wiley-Interscience.

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

Finamore, W. A., & Pearlman, W. A. (1980). Optimal encoding of discrete-time continuous-amplitude memoryless sources with finite output alphabet. *IEEE Transactions on Information Theory*, *26*, 144–155.

Forster, J., & Warmuth, M. K. (2000). Relative expected instantaneous loss bounds. *13th COLT.*

Gilad-Bachrach, R., Navot, A., & Tishby, N. (2003). An information theoretic tradeoff between complexity and accuracy. *16th COLT.*

Kearns, M., Mansour, Y., & Ng, A. (1997). An information-theoretic analysis of hard and soft assignment methods for clustering. *13th UAI.*

Neal, R., & Hinton, G. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. *Learning in Graphical Models* (pp. 355–368). MIT Press.

Nigam, K. (2001). *Using unlabeled data to improve text classification* (Technical Report CMU-CS-01-126). Carnegie Mellon University.

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

Rockafeller, R. T. (1970). *Convex analysis.* Princeton University Press.

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

Rose, K. (1998). Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proceedings of IEEE*, *86*, 2210–2239.

Slonim, N., & Weiss, Y. (2002). Maximum likelihood and the information bottleneck. *16th NIPS.*

Tishby, N., Pereira, F. C., & Bialek, W. (1999). The information bottleneck method. *Proc. 37th Annual Allerton Conf. on Communication, Control and Computing* (pp. 368–377).