Learning Continuous Latent Variable Models with Bregman Divergences

Shaojun Wang[1] and Dale Schuurmans[2]

[1] Department of Statistics, University of Toronto, Canada
[2] School of Computer Science, University of Waterloo, Canada

Abstract. We present a class of unsupervised statistical learning algorithms that are formulated in terms of minimizing Bregman divergences—a family of generalized entropy measures with convex functions. We obtain novel training algorithms that extract hidden latent structure by minimizing a Bregman divergence on training data, subject to a set of non-linear constraints which consider hidden variables. An alternating minimization procedure with nested iterative scaling is proposed to find feasible solutions for the resulting constrained optimization problem. The convergence of this algorithm along with its information geometric properties are characterized.

Index Terms — statistical machine learning, unsupervised learning, Bregman divergence, information geometry, alternating minimization, forward projection, backward projection, iterative scaling.

1 Introduction

A variety of machine learning and statistical inference problems focus on supervised learning from labeled training data. In such problems, convexity often plays a central role in formulating the loss function to be minimized during training. For example, a standard approach to formulating a training loss is to distinguish a preferred value from a set of candidate prediction values, and measure prediction error by a convex error measure. Examples of this include least squares regression, decision tree learning, boosting, on-line learning, maximum likelihood for exponential models, logistic regression, maximum entropy, support vector machines, statistical signal processing (e.g. Burg's spectral estimation for speech signal analysis and image reconstruction) and optimal portfolio selection. Such problems can often be naturally cast as convex optimization problems involving a Bregman divergence [5, 10, 23], which can lead to new algorithms, analytical tools, and insights derived from the powerful methods of convex analysis [2, 3, 7, 13]. Training algorithms that solve these problems can be cast as implementing a *minimum Bregman divergence* (MB) principle.

However, in practice, many of the natural patterns we wish to classify are the result of causal processes that have hidden hierarchical structure—yielding data that does not report the value of *latent* variables. For example, in natural language learning the observed data rarely reports the value of hidden semantic variables or syntactic structure, in speech signal analysis gender information is not explicitly marked, etc. Obtaining fully labeled data is tedious or impossible in most realistic cases. This motivates us to propose a class of *unsupervised* statistical learning algorithms that are still formulated in terms of minimizing a Bregman divergence, except that we must now change the problem formulation to respect hidden variables. In this paper we propose training algorithms for solving the *latent minimum Bregman divergence* (LMB) principle: given a set of training data and features that one would like to match in the training data, compute a model that minimizes a convex objective function (a Bregman divergence) subject to a set of non-linear constraints that take into account possible latent structure. Our treatment of the LMB principle closely parallels the results presented in [24] for the Kullback-Leibler divergence, but the extension proposed here is not trivial. For probabilistic models under the Kullback-Leibler divergence, we can show the equivalence between satisfying the constraints (i.e. achieving feasibility) and locally maximizing likelihood under log-linear assumption, thus we can resort to the technique of EM algorithm [14] to develop practical algorithm for finding feasible solutions and prove its convergence. However, the general Bregman divergence brings a more difficult technical challenge since the technique of EM algorithm breaks down for these generalized entropy measures, and we will overcome this by using alternating minimization apparoch [9] in a clever way, see Figure 1 for illustration. Thus, beyond the generalized KL divergence used for unsupervised boosting for clustering [25], the techniques of this paper can also handle a broader class of functions, such as the Itakura-Siato distortion for speech signal analysis. (In this case, gender information can be encoded implicitly and thus leads to more accurate modeling of speech signals.)

2 The LMB principle

To express a joint (probability) model, let $X \in \mathcal{X}$ denote the complete data, $Y \in \mathcal{Y}$ be the observed incomplete data and $Z \in \mathcal{Z}$ be the missing data. That is, $X = (Y, Z)$. Let $\phi(t) : \Re \to \Re$ be a strictly convex function on an interval $\mathcal{I} \subset \Re$, and differentiable in the interior of $\mathcal{I}$. Define a closed convex set $\mathcal{S} \subset \Re^X$, where $\mathcal{S}$ is typically assumed to be the set of (probability) distributions (or positive measures) p over $\mathcal{X}$. For functions p and q on $\mathcal{X}$ with values in $\mathcal{I}$, a *Bregman divergence* [3] [4, 7, 10–13, 18] is a generalized entropy measure that is associated with a convex function ϕ

$$B_\phi(p; q) = \int_{x \in \mathcal{X}} \Delta_\phi(p(x); q(x))\ \mu(dx)$$

[3] Machine learning community [8, 13, 18, 19] is familiar with the discrete case, since in supervised learning there are limited number of sampling points (training examples), so we can write the constraints as pertaining to a finite dimensional vector. However, in unsupervised learning, we are in general dealing with continuous variables. Instead of a vector, we are working with an infinite dimensional space.

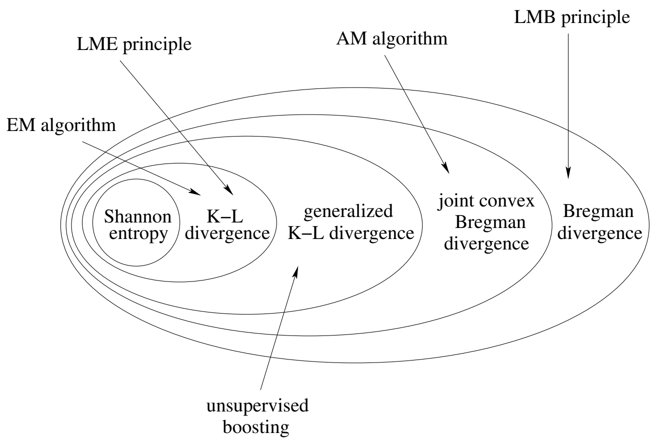

Fig. 1. *The nested ellipsoids are the functions of Bregman divergence, and they can be either discrete or continuous variables. The AM algorithm proposed in this paper is valid for the family of joint convex Bregman divergences, but the EM algorithm proposed in [24] is only valid for the K-L divergence. The unsupervised boosting case is dealing with generalized K-L divergence, thus it can be solved by using the AM algorithm to find feasible solutions under latent minimum Bregman divergence principle.*

where

$$\Delta_\phi(p(x);q(x)) \;=\; \phi(p(x)) - \phi(q(x)) \;-\; \phi'\left(q(x)\right)\left(p(x)-q(x)\right)$$

and ϕ' denotes the derivative of ϕ.[4] That is, the Bregman divergence B_ϕ measures the discrepancy between two distributions p and q by integrating the difference between ϕ evaluated at p and ϕ's first-order Taylor expansion about q, evaluated at p over $\mathcal{X}$.

To strengthen the interpretation of $B_\phi(p;q)$ as a measure of distance, we make the following assumptions.

- $\Delta_\phi(u,v)$ is strictly convex in u and in v separately, but also satisfies the stronger property that it is jointly convex in u, v. Thus our choice of Bregman divergence $B_\phi(p;q)$ is strictly convex in p and in q separately, and also jointly convex. This assumption lies at the heart of the analysis below.
- $B_\phi(p;q)$ is lower-semi-continuous in p and q jointly.
- For any fixed $p \in \mathcal{S}$, the level sets $\{q : B_\phi(p;q) \le \ell\}$ are bounded.
- If $B_\phi(p^k;q^k) \to 0$ and p^k or q^k is bounded, then $p^k \to q^k$ and $q^k \to p^k$.
- If $p \in \mathcal{S}$ and $q^k \to p$, then $B_\phi(p;q^k) \to 0$.

[4] In this paper, μ denotes a given σ-finite measure on $\mathcal{X}$. If $\mathcal{X}$ is finite or countably infinite, then μ is the counting measure, and integrals reduce to sums. If $\mathcal{X}$ is a subset of a finite dimensional space, μ is the Lebesgue measure. If $\mathcal{X}$ is a combination of both cases, μ will be a combination of both measures.

Examples

1. Let $\phi(t) = t \log t$ be defined on $I = [0, \infty)$. Then $\phi'(t) = \log t + 1$, and
$$B_\phi(p;q) = D(p;q) = \int_{x \in \mathcal{X}} \left(p(x) \log \frac{p(x)}{q(x)} - p(x) + q(x) \right) \mu(dx)$$
which is the generalized KL divergence. This is the objective function of the primal problem for AdaBoost [20]. When p and q are restricted to be probability measures, it becomes the KL divergence, the objective function of the primal problem for LogitBoost [16, 20]. Furthermore, when q is chosen to be uniform, it becomes the Shannon entropy.
2. Let $\phi(t) = t^2$ be defined on $I = (-\infty, \infty)$. Then $\phi'(t) = 2t$, and
$$B_\phi(p;q) = ||p(x) - q(x)||^2_{L^2(\mu)}$$
which is the measure of energy.
3. Let $\phi(t) = -\log t$ be defined on $I = (0, \infty)$. Then $\phi'(t) = -\frac{1}{t}$, and
$$B_\phi(p;q) = \int_{x \in \mathcal{X}} \left(\log \frac{q(x)}{p(x)} + \frac{p(x)}{q(x)} - 1 \right) \mu(dx)$$
which is the Itakura-Saito distortion that arises in the spectral analysis of speech signals. When $q = 1$, it becomes the Burg entropy [17].
4. Let $\phi(t) = t \log t + (1-t)\log(1-t)$ be defined on $I = [0, 1]$. Then $\phi'(t) = \log \frac{t}{1-t}$, and
$$B_\phi(p;q) = \int_{x \in \mathcal{X}} \left(p(x) \log \frac{p(x)}{q(x)} + (1 - p(x)) \log \frac{1-p(x)}{1-q(x)} \right) \mu(dx)$$
which is the Bernoulli entropy. When $q = \frac{1}{2}$, it becomes the Fermi-Dirac entropy.

To formulate the minimum Bregman divergence principle, assume we have a finite set of *features* $f_1(x), ..., f_N(x)$ which correspond to sufficient statistics in a log-linear model, weak learners in boosting, or basis function in non-parametric estimation. Given a set of complete data points $\tilde{\mathcal{X}} = (\tilde{\mathcal{Y}}, \tilde{\mathcal{Z}})$, where $\tilde{\mathcal{Y}}$ are observed "descriptions" and $\tilde{\mathcal{Z}}$ are observed "labels", the minimum Bregman divergence principle (MB) is:

MB principle Choose a conditional distribution $p(z|y)$ to minimize
$$\min_{p(z|y) \in \mathcal{S}} \; B_\phi(\tilde{p}(y)p(z|y); q_0(x)) \tag{1}$$
subject to the constraints
$$\sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z \in \mathcal{Z}} f_i(x) \;\; p(z|y)\, \mu(dz) = \sum_{x \in \tilde{\mathcal{X}}} \tilde{p}(x) f_i(x) \qquad \text{for } i = 1, ..., N \tag{2}$$

where $x = (y, z)$, $q_0 \in \mathcal{S}$ is a default distribution chosen so that $\phi'(q_0) = 0$, and quite often we set q_0 te be uniform, and $\tilde{p}(x)$ and $\tilde{p}(y)$ denote the empirical distributions of the complete and marginal data respectively. ∎

In general, iterative scaling [11–13, 18] is used to obtain the (global) optimal solution for the MB principle.

In contrast to MB principle, if the labels $\tilde{\mathcal{Z}}$ are *unobserved*, we propose the latent minimum Bregman divergence principle (LMB) as follows.

LMB principle Choose a *joint* distribution $p(x)$ to minimize

$$\min_{p \in \mathcal{S}} \; B_\phi(p; q_0) \tag{3}$$

subject to the constraints

$$\int_{x \in \mathcal{X}} f_i(x) \;\; p(x)\, \mu(dx) = \sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z \in \mathcal{Z}} f_i(x) \;\; p(z|y)\, \mu(dz), \;\; \text{for } i = 1, ..., N \tag{4}$$

where $x = (y, z)$, $q_0 \in \mathcal{S}$ is a default distribution chosen so that $\phi'(q_0) = 0$, and quite often we set q_0 te be uniform. Here $\tilde{p}(y)$ is the empirical distribution of the observed data and $p(z|y)$ is the conditional distribution of latent variables given the observed data. ∎

Note that the conditional $p(z|y)$ implicitly encodes the latent structure and is a nonlinear mapping of $p(x)$. That is, $p(z|y) = p(y, z) / \int_{z' \in \mathcal{Z}} p(y, z')\mu(dz) = p(x) / \int_{x' = (y, z')} p(x')\mu(dx')$ where $x = (y, z)$ and $x' = (y, z')$. Clearly $p(z|y)$ is a non-linear function of $p(x)$ because of the division. This means we are faced with minimizing an objective (3) subject to a system of non-linear constraints (4). Therefore, even though the objective function (3) is convex, no unique minimum can be guaranteed to exist. In fact, maxima and saddle points may exist. Nevertheless, one can still attempt to derive an iterative training procedure that finds *feasible* solutions to the LMB problem. With such a subroutine in hand, one can then heuristically solve the LMB principle by gathering several feasible solutions (by starting with different initial points) and then choosing the feasible p that obtains the smallest Bregman divergence.

To illustrate how the LMB principle is related to unsupervised learning, assume given a collection of samples without labels, and we want to incrementally build linear combinations of weak "decision stumps" to generate a "strong" predicative model for clustering. In this case, we can formulate the problem as minimizing the generalized K-L divergence subject to the nonlinear constraints where the features are the decision stumps and we are looking for a feasible unnormalized exponential model which has the smallest Bregman divergence.

Below we focus on developing an iterative algorithm for finding feasible solutions. In general, solving (3) subject to (4) is quite complex. Since the original problem does not yield a simple closed form solution for p, we instead look for an approximate solution. First, we restrict the model to have an *additive* form.

Definition 1. *[19] Let $S \subset \Re^X$ be a set of measures. An additive model for S is defined by an operation $\mathcal{L} : \Re^X \times \mathcal{S} \to \mathcal{S}$ satisfying the homomorphism property $\mathcal{L}(r_1 + r_2, s) = \mathcal{L}(r_1, \mathcal{L}(r_2, s))$ for all $r_1, r_2 \in \Re^X$ and $s \in \mathcal{S}$.*

Lemma 1. *[19] Given a convex function $\phi : \Re \to \Re$, let B_ϕ be the Bregman divergence defined on measures $p \in \mathcal{S}$. Define the Legendre transform $v \circ_\phi q_0$ by*

$$v \circ_\phi q_0 = \arg\min_{p \in \mathcal{S}} \quad B_\phi(p; q_0) - \langle v \cdot p \rangle$$

Then we have $v \circ_\phi q_0 = \mathcal{L}_\phi(q_0, v)$ such that $(\mathcal{L}_\phi(q_0, v))(x) = (\phi')^{-1}(\phi'(q_0(x)) - v(x))$ for all x. Also the map $(v, q_0) \mapsto v \circ_\phi q_0$ is an additive model for $\mathcal{S}$.

By adopting an additive model restriction, we can make valuable progress toward formulating a practical algorithm for approximately satisfying the LMB principle.

In the following, we use a doubly iterative projection algorithm to obtain feasible additive solutions, and also provide a characterization of its convergence properties and information geometry.

3 Preliminaries: Convergence of alternating projections

We present a generalization of the alternating projection method of Csiszar and Tusnady [9] for Bregman divergences, and show how this technique can be used to find feasible solutions for the LMB principle. In developing our method we need to derive a slightly more general convergence result than [9], which is due to [15]. **These results are originally shown in [6, 15] for discrete case, here we extend them for continuous variables.**

Since projections onto closed convex sets may be thought of as solutions of minimum divergence problems, we begin by introducing suitable definitions for the Bregman divergence.

Definition 2. (Forward projection) *Suppose $\mathcal{Q} \subset \mathcal{S}$ is a nonempty closed convex set, and let $p \in \mathcal{S}$. We define the forward projection of p onto $\mathcal{Q}$ as the unique element $q^* \in \mathcal{Q}$ such that $B_\phi(p; q^*) = \min_{q \in \mathcal{Q}}\{B_\phi(p; q)\}$.*

Definition 3. (Backward projection) *Suppose $\mathcal{P} \subset \mathcal{S}$ is a nonempty closed convex set, and let $q \in \mathcal{S}$. We define the backward projection of q onto $\mathcal{P}$ as the unique element $p^* \in \mathcal{P}$ such that $B_\phi(p^*; q) = \min_{p \in \mathcal{P}}\{B_\phi(p; q)\}$.*

We can then define the alternating projection algorithm associated with the Bregman divergence.

Alternating minimization (AM) algorithm Consider two nonempty closed convex sets $\mathcal{P}, \mathcal{Q} \subset \mathcal{S}$.

Initialization: Let $q^0 \in \mathcal{Q}$ be an arbitrary distribution such that there exists $p \in \mathcal{P}$ with $B_\phi(p; q^0) < \infty$.

Iterative step: Given q^k, find p^k by backward projection onto $\mathcal{P}$:

$$p^k = \arg\min_{p \in \mathcal{P}} B_\phi(p; q^k)$$

Then calculate q^{k+1} by forward projection onto $\mathcal{Q}$:

$$q^{k+1} = \arg\min_{q \in \mathcal{Q}} B_\phi(p^k; q)$$

Repeat until convergence. ∎

To prove that this procedure converges, we first demonstrate the "three points" and "four points" properties for the Bregman divergence.

Lemma 2. (Three points property) *Consider a Bregman divergence B_ϕ on two nonempty closed convex sets $\mathcal{P}, \mathcal{Q} \subset \mathcal{S}$. Let $q \in \mathcal{Q}$ be such that $B_\phi(p; q) < \infty$ for all $p \in \mathcal{P}$, and let $p^* = \arg\min_{p \in \mathcal{P}} B_\phi(p; q)$. Then for all $p \in \mathcal{P}$ we have*

$$B_\phi(p; q) - B_\phi(p^*; q) \geq B_\phi(p; p^*)$$

Proof. By definition of Bregman divergence, we have

$$\begin{aligned}
&B_\phi(p; q) - B_\phi(p^*; q) \\
&= \int_{x \in \mathcal{X}} \phi(p(x)) - \phi(p^*(x)) - \phi'(q(x)) \, (p(x) - p^*(x)) \, \mu(dx) \\
&= \int_{x \in \mathcal{X}} \phi(p(x)) - \phi(p^*(x)) - \phi'(p^*(x))(p(x) - p^*(x)) \\
&\qquad\quad + (\phi'(p^*(x)) - \phi'(q(x))) \, (p(x) - p^*(x)) \, \mu(dx) \\
&= B_\phi(p; p^*) + \int_{x \in \mathcal{X}} (\phi'(p^*(x)) - \phi'(q(x))) \, (p(x) - p^*(x)) \, \mu(dx)
\end{aligned}$$

Denote the partial gradient of B_ϕ with respect to its first argument as $\nabla_p B_\phi(p; q)$, and note that $\partial B_\phi(p; q)/\partial p(x) = \phi'(p(x)) - \phi'(q(x))$. Therefore we have

$$(\nabla_p B_\phi(p^*; q))\,(x) = \phi'(p^*(x)) - \phi'(q(x))$$

for all x. Since p^* minimizes $B_\phi(p; q)$ over convex set $\mathcal{P}$, we must have

$$\langle \nabla_p B_\phi(p^*; q) \cdot (p - p^*) \rangle \geq 0$$

The result then follows since

$$\int_{x \in \mathcal{X}} (\phi'(p^*(x)) - \phi'(q(x))) \, (p(x) - p^*(x)) \, \mu(dx) = \langle \nabla_p B_\phi(p^*; q) \cdot (p - p^*) \rangle \quad ∎$$

Lemma 3. (Four points property) *Consider a* **jointly convex** *Bregman divergence B_ϕ on two nonempty closed convex sets $\mathcal{P}, \mathcal{Q} \subset \mathcal{S}$. Let $p \in \mathcal{P}$ such that $B_\phi(p; q) < \infty$ for all $q \in \mathcal{Q}$, and let $q^* = \arg\min_{q \in \mathcal{Q}} B_\phi(p; q)$. Then for all $u \in \mathcal{P}, v \in \mathcal{Q}$ we have*

$$B_\phi(u; q^*) \leq B_\phi(u; p) + B_\phi(u; v)$$

Proof. By the joint convexity assumption of $\Delta_\phi(p(x); q(x))$, we have

$$\begin{aligned}\Delta_\phi(u(x); v(x)) \;\geq\; \Delta_\phi(p(x); q^*(x)) + &\tfrac{\partial}{\partial p(x)}\Delta_\phi(p(x); q^*(x))\,(u(x)-p(x)) \\ + &\tfrac{\partial}{\partial q^*(x)}\Delta_\phi(p(x); q^*(x))\,(v(x)-q^*(x))\end{aligned}$$

for all x. Therefore

$$B_\phi(u; v) \geq B_\phi(p; q^*) + \langle \nabla_p B_\phi(p; q^*) \cdot (u-p)\rangle + \langle \nabla_{q^*} B_\phi(p; q^*) \cdot (v-q^*)\rangle$$

Since q^* minimizes $B_\phi(p; q)$ over the convex set $\mathcal{Q}$, we have

$$\langle \nabla_{q^*} B_\phi(p; q^*) \cdot (v-q^*)\rangle \geq 0$$

Thus

$$B_\phi(u; v) - B_\phi(p; q^*) - \langle \nabla_p B_\phi(p; q^*) \cdot (u-p)\rangle \geq 0$$

On the other hand, by the definition of Bregman divergence, we have

$$\begin{aligned}B_\phi(u; p) - B_\phi(u; q^*) &= \textstyle\int_{x\in\mathcal{X}} \phi(q^*(x)) - \phi(p(x)) + \phi'(q^*(x))(u(x)-q^*(x)) \\ &\qquad\qquad - \phi'(p(x))(u(x)-p(x)) \quad \mu(dx) \\ &= -B_\phi(p; q^*) - \textstyle\int_{x\in\mathcal{X}} (\phi'(p(x)) - \phi'(q^*(x)))(u(x)-p(x))\,\mu(dx) \\ &= -B_\phi(p; q^*) - \langle \nabla_p B_\phi(p; q^*) \cdot (u-p)\rangle\end{aligned}$$

Thus we obtain

$$\begin{aligned}B_\phi(u; p) + B_\phi(u; v) - B_\phi(u; q^*) &= B_\phi(u; v) - B_\phi(p; q^*) - \langle \nabla_p B_\phi(p; q^*) \cdot (u-p)\rangle \\ &\geq 0 \quad \blacksquare\end{aligned}$$

Given these two lemmas, following [15] we obtain the following convergence result.

Theorem 1. *The alternating minimization algorithm (AM) converges. That is, $p^1, p^2, \ldots$ converges to some $p^\infty \in \mathcal{P}$, and $q^1, q^2, \ldots$ converges to some $q^\infty \in \mathcal{Q}$, such that*

$$B_\phi(p^\infty; q^\infty) = \min_{p\in\mathcal{P}, q\in\mathcal{Q}} B_\phi(p; q)$$

Proof. We first show that the sequence $B_\phi(p^k; q^k)$ is non-increasing. First note that since $q^{k+1} = \arg\min_{q\in\mathcal{Q}} B_\phi(p^k; q)$ we have $B_\phi(p^k; q^k) - B_\phi(p^k; q^{k+1}) \geq 0$. Next, by the three points lemma, we have $B_\phi(p^k; q^{k+1}) - B_\phi(p^{k+1}; q^{k+1}) \geq B_\phi(p^k; p^{k+1})$. Combining these two results yields

$$\begin{aligned}&B_\phi(p^k; q^k) - B_\phi(p^{k+1}; q^{k+1}) \\ &= B_\phi(p^k; q^k) - B_\phi(p^k; q^{k+1}) + B_\phi(p^k; q^{k+1}) - B_\phi(p^{k+1}; q^{k+1}) \\ &\geq 0 + B_\phi(p^k; p^{k+1}) \;\geq\; 0\end{aligned}$$

Therefore, for all $k \geq 1$ the sequence $B_\phi(p^k; q^k)$ is non-increasing and non-negative.

Let $p^\infty, q^\infty = \arg\min_{p\in\mathcal{P}, q\in\mathcal{Q}} B_\phi(p;q)$. We next show that the sequence $B_\phi(p^\infty; p^k)$ must be non-increasing. By setting $p = p^k, q^* = q^{k+1}, u = p^\infty$ and $v = q^\infty$ in four points lemma, we obtain

$$B_\phi(p^\infty; q^{k+1}) \leq B_\phi(p^\infty; p^k) + B_\phi(p^\infty; q^\infty)$$

Next, by setting $p^* = p^{k+1}, q = q^{k+1}, p = p^\infty$ and $q = q^\infty$ in three points lemma

$$B_\phi(p^\infty; q^{k+1}) - B_\phi(p^{k+1}; q^{k+1}) \geq B_\phi(p^\infty; p^{k+1})$$

Combining these two results yields the inequality

$$B_\phi(p^\infty; p^k) - B_\phi(p^\infty; p^{k+1}) \geq B_\phi(p^{k+1}; q^{k+1}) - B_\phi(p^\infty; q^\infty) \ \geq \ 0 \qquad (5)$$

Therefore, for all $k \geq 1$ the sequence $B_\phi(p^\infty; p^k)$ is also non-increasing and non-negative.

Since the sequence $B_\phi(p^\infty; p^k)$ is bounded and non-increasing it must have a limit. In particular, as $k \to \infty$, the left hand side of (5) must go to 0. Moreover (5) implies further that $\lim_{k\to\infty} B_\phi(p^k; q^k) = B_\phi(p^\infty; q^\infty)$.

Finally, we must show that the distributions p^k and q^k themselves converge. From the boundedness of $B_\phi(p^\infty; p^k)$ we have that p^k is bounded, so it has a convergent subsequence, denoted $\{p^{k_i}\}$. Denote the limit of this subsequence as p_0. Similarly, one can show that the corresponding subsequence $\{q^{k_i}\}$ is bounded, and therefore has a convergent subsequence. Without loss of generality, we may assume that $\{q^{k_i}\}$ is convergent with limit q_0. By the lower semi-continuity assumption of the Bregman divergence, we have

$$B_\phi(p_0; q_0) \leq \liminf_i B_\phi(p^{k_i}; q^{k_i}) = B_\phi(p^\infty; q^\infty)$$

Denote $\mathcal{P}^\infty = \{p : \min_{p\in\mathcal{P}, q\in\mathcal{Q}} B_\phi(p;q)\}$ and $\mathcal{Q}^\infty = \{q : \min_{p\in\mathcal{P}, q\in\mathcal{Q}} B_\phi(p;q)\}$. Then we have $p_0 \in \mathcal{P}^\infty$ and $q_0 \in \mathcal{Q}^\infty$.

To prove the convergence of the entire sequence, apply the above with p^∞ replaced by p_0. Then the sequence $\{B_\phi(p_0; p^k)\}$ is bounded and non-increasing, and it has a convergent subsequence $\{p^{k_i}\} \to p_0$ such that $\{B_\phi(p_0; p^{k_i})\} \to 0$. This implies that $\{B_\phi(p_0; p^k)\} \to 0$, which establishes $\{p^k\} \to p_0$. Now since $\{q^k\}$ is bounded, every subsequence has itself a convergent subsequence. Denote the limit as $q_{(0)}$. By the lower semi-continuity assumption of the Bregman divergence, as above we have $B_\phi(p_0; q_{(0)}) \leq B_\phi(p^\infty; q^\infty)$. So $q_{(0)} = \arg\min_{q\in\mathcal{Q}} B_\phi(p_0; q)$ and $\{q^k\} \to q_{(0)}$.

Finally since $p^\infty \in \mathcal{P}^\infty$ and $p^\infty = \arg\min_{p\in\mathcal{P}} B_\phi(p; q^\infty)$, we have $B_\phi(p^\infty; q^\infty) = \min_{p\in\mathcal{P}, q\in\mathcal{Q}} B_\phi(p;q)$. ∎

4 The AM-IS iterative algorithm for learning latent structure

We now extend this alternating minimization algorithm to finding feasible solutions to the LMB principle. To understand the algorithm and its information geometry, we first define some useful sub-manifolds in $\mathcal{S}$.

$$\mathcal{C} = \left\{ p \in \mathcal{S} : \int_{x\in\mathcal{X}} f_i(x)\, p(x)\mu(dx) \;=\; \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z\in\mathcal{Z}} f_i(x)\, p(z|y)\mu(dz),\; i = 1...N \right\}$$

$$\mathcal{M} = \left\{ p \in \mathcal{S} : \int_{z\in\mathcal{Z}} p(y,z)\,\mu(dz) \;=\; \tilde{p}(y), \quad \text{for each}\;\; y \in \tilde{\mathcal{Y}} \right\}$$

$$\mathcal{G}_a = \left\{ p \in \mathcal{S} : \int_{x\in\mathcal{X}} p(x) f_i(x)\,\mu(dx) \;=\; a_i, \quad i = 1...N \right\}$$

$$\mathcal{E} = \left\{ p_\lambda \in \mathcal{S} : p_\lambda(x) = \mathcal{L}_\phi\left(q_0, \sum_{i=1}^N \lambda_i f_i(x) \right),\; \lambda \in \Omega \right\}$$

where $\mathcal{C}$ denotes the set of nonlinear constraints the model should satisfies, $\mathcal{M}$ denotes the set of distributions whose observed marginal distribution matches the observed empirical distribution, $\mathcal{G}_a$ denotes the set of distributions whose features' expectations are constant, $\mathcal{E}$ denotes the set of additive models, and

$$\Omega = \left\{ \lambda \in \Re^N : \mathcal{L}_\phi\left(q_0, \sum_{i=1}^N \lambda_i f_i(x) \right) \;<\; \infty \right\}$$

Now by choosing the closed convex set $\bar{\mathcal{M}}$ to play the role of $\mathcal{P}$ in the previous discussion, and choosing the closed convex set $\bar{\mathcal{E}}$ to play the role of $\mathcal{Q}$, we can define the corresponding forward projection and backward projection operators, and then use these to iterate toward feasible LMB solutions.

First, to derive a backward projection operator, take a current $p_{\lambda^{(k)}}$ playing the role of q in the previous discussion, and use this to determine a distribution $p^* \in \bar{\mathcal{M}}$ that minimizes

$$B_\phi(p^*; p_{\lambda^{(k)}}) = \min_{p\in\bar{\mathcal{M}}} B_\phi(p; p_{\lambda^{(k)}})$$

That is, p^* is the backward projection of $p_{\lambda^{(k)}}$ onto $\bar{\mathcal{M}}$. To solve for p^*, one can formulate the Lagrangian $\Lambda(p,\alpha)$

$$\Lambda(p,\alpha) = B_\phi(p; p_{\lambda^{(k)}}) + \sum_{y\in\tilde{\mathcal{Y}}} \alpha_y \left(\int_{z\in\mathcal{Z}} p(y,z)\mu(dz) - \tilde{p}(y) \right)$$

Now since

$$\frac{\partial}{\partial p(x)} \Lambda(p,\alpha) = \phi'(p(x)) - \phi'(p_{\lambda^{(k)}}(x)) + \alpha_y$$

it is not hard to see that the solution p^* must satisfy

$$p^*(x) = (\phi')^{-1}\left(\phi'(p_{\lambda^{(k)}}(x)) - \alpha_y\right)$$

for all $y \in \tilde{\mathcal{Y}}$, where α_y is chosen so that

$$\int_{z\in\mathcal{Z}} p^*(y,z)\,\mu(dz) = \int_{z\in\mathcal{Z}} (\phi')^{-1}\left(\phi'(p_{\lambda^{(k)}}(y,z)) - \alpha_y\right)\,\mu(dz) \;=\; \tilde{p}(y) \quad (6)$$

Lemma 4. *For ϕ corresponding to Examples 1 and 2 above, the backward projection of $p_{\lambda^{(k)}}$ onto $\bar{\mathcal{M}}$ is given by the closed form solution $p^*(x) = \tilde{p}(y)p_{\lambda^{(k)}}(z|y)$ for all $x = (y,z)$.*

Proof. Note that for $\phi(t) = t\log t$ (Example 1) or $\phi(t) = t^2$ (Example 2) we have

$$\phi'(p_{\lambda^{(k)}}(x)) - \phi'(p_{\lambda^{(k)}}(y)) + \phi'(\tilde{p}(y)) = \phi'\left(p_{\lambda^{(k)}}(z|y)\tilde{p}(y)\right)$$

Therefore, if we let $\alpha_y = \phi'(p_{\lambda^{(k)}}(y)) - \phi'(\tilde{p}(y))$ we obtain

$$\begin{aligned} p^*(x) &= (\phi')^{-1}\left(\phi'(p_{\lambda^{(k)}}(x)) - \alpha_y\right) \\ &= (\phi')^{-1}\left(\phi'(p_{\lambda^{(k)}}(z|y)\tilde{p}(y))\right) \\ &= \tilde{p}(y)p_{\lambda^{(k)}}(z|y) \quad \blacksquare \end{aligned}$$

Thus in many cases we can implement the backward projection step for AM merely by calculating the conditional distribution $p_{\lambda^{(k)}}(z|y)$ of the current model. In general, one has to solve for the Lagrange multipliers that satistfy (6) to yield a general form of the backward projection $p_{\alpha,\lambda^{(k)}}(z|y)$.

Next, to formulate the forward projection step, we exploit the following lemma.

Lemma 5. *For any $p^{(k)} \in \bar{\mathcal{M}}$, the forward projection of $p^{(k)}$ onto $\bar{\mathcal{E}}$ is equivalent to solving the minimization*

$$\min_{p\in\mathcal{G}_a} \quad B_\phi(p;q_0) \qquad where \qquad a_i = \int_{x\in\mathcal{X}} p^{(k)}(x)f_i(x)\mu(dx) \quad (7)$$

Proof. To find the solution of (7), form the Lagrangian $\Psi(p,\lambda)$

$$\Psi(p,\lambda) = B_\phi(p;q_0) + \sum_{i=1}^N \lambda_i\left(\int_{x\in\mathcal{X}} p(x)f_i(x)\mu(dx) - \int_{x\in\mathcal{X}} p^{(k)}(x)f_i(x)\mu(dx)\right)$$

Now since

$$\frac{\partial}{\partial p(x)}\Psi(p,\lambda) = \phi'(p(x)) - \phi'(q_0(x)) + \sum_{i=1}^N \lambda_i f_i(x)$$

any solution must satisfy

$$p_\lambda(x) = (\phi')^{-1}\left(\phi'(q_0(x)) - \sum_{i=1}^N \lambda_i f_i(x)\right)$$

Plugging into Ψ, we are left with the problem of maximizing

$$\begin{aligned}\Psi(p_\lambda,\lambda) &= \int_{x\in\mathcal{X}} \phi(p_\lambda(x)) - \phi(q_0(x)) - \phi'(q_0(x))(p_\lambda(x) - q_0(x))\ \mu(dx) \\ &\quad + \sum_{i=1}^N \lambda_i \left(\int_{x\in\mathcal{X}} p_\lambda(x) f_i(x)\mu(dx) - \int_{x\in\mathcal{X}} p^{(k)}(x) f_i(x)\mu(dx) \right) \\ &= B_\phi(p^{(k)}; q_0) - B_\phi(p^{(k)}; p_\lambda)\end{aligned}$$

which is equivalent to minimizing $B_\phi(p^{(k)}; p_\lambda)$ over $p_\lambda \in \bar{\mathcal{E}}$, the forward projection of $p^{(k)}$ onto $\bar{\mathcal{E}}$. ∎

To solve the minimization problem specified in (7) one can use *iterative scaling*. By using an auxiliary function to bound the change in Bregman divergence from below, the iterative scaling algorithm can be derived. Following [13], define an auxiliary function as the following:

Definition 4. *We call $\mathcal{A} : \mathcal{S} \times \Re^N \to \Re$ an auxiliary function for p^k and $\underline{f}$ if it satisfies the following conditions:*

1. *$\mathcal{A}(q,\lambda)$ is continuous in q and $\mathcal{A}(q,0) = 0$.*
2. *$B_\phi(p^k; q) - B_\phi(p^k; \mathcal{L}_\phi(q, \sum_{i=1}^N f_i(x)\lambda_i)) \geq \mathcal{A}(q,\lambda)$.*
3. *If $\lambda = 0$ is a maximum of $\mathcal{A}(q,\lambda)$, then*

$$\int_{x\in\mathcal{X}} f_i(x)\ \ q(x)\, \mu(dx) = \int_{x\in\mathcal{X}} f_i(x)\ \ p^k(x)\, \mu(dx), \qquad i = 1, ..., N$$

Maximizing this auxiliary function we obtain new parameters $\lambda' = \lambda + \triangle\lambda$ and a new model given by

$$\begin{aligned} q_{\lambda+\triangle\lambda} &= \mathcal{L}\left(q_\lambda, \sum_{i=1}^N \triangle\lambda_i f_i(x) \right) \\ &= \mathcal{L}\left(\mathcal{L}\left(q_0, \sum_{i=1}^N \lambda_i f_i(x) \right), \sum_{i=1}^N \triangle\lambda_i f_i(x) \right) \\ &= \mathcal{L}\left(q_\lambda, \sum_{i=1}^N (\lambda_i + \triangle\lambda_i) f_i(x) \right)\end{aligned}$$

When $\triangle\lambda = 0$, we have that $q_\lambda = \min_{p\in\mathcal{G}_a} B_\phi(p; q_0)$.

Lemma 6. *Define $f(x) = \sum_{i=1}^N |f_i(x)|$, $\sigma_i(x) = \mathrm{sign}(f_i(x))$ and $l_\phi(q,v) = \sup_{p\in\mathcal{S}} \langle v \cdot p\rangle - B_\phi(p;q)$. Then*

$$\begin{aligned}\mathcal{A}(q,\lambda) &\stackrel{\text{def}}{=} \sum_{i=1}^N \lambda_i \int_{x\in\mathcal{X}} f_i(x)\ \ p^k(x)\, \mu(dx) \\ &\quad - \int_{x\in\mathcal{X}} \sum_{i=1}^N \frac{|f_i(x)|}{f(x)}\ \ l_\phi\left(q, \sum_{i=1}^N \sigma_i(x) f(x) \lambda_i \right) \mu(dx)\end{aligned} \tag{8}$$

is an auxiliary function for p^k and $\underline{f}$, and the corresponding iterative scaling update scheme is given by

$$q_{t+1}^k = \mathcal{L}_\phi\left(q_t^k, \sum_{i=1}^{N} \lambda_i f_i(x)\right) \tag{9}$$

where $\lambda_i, i = 1, ..., N$ satisfies

$$\int_{x\in\mathcal{X}} f_i(x)\mathcal{L}_\phi\left(q_t^k, \sigma_i(x) f(x)\lambda_i\right)\mu(dx) = \int_{x\in\mathcal{X}} f_i(x)\ \ p^k(x)\,\mu(dx) \tag{10}$$

and

$$\lim_{t\to\infty} q_t^k = q_\infty^k = \arg\min_{q\in\bar{\mathcal{E}}} B_\phi(p^k; q) \tag{11}$$

Proof. Following [13], which considers discrete state distribution, we consider the continuous case. The proof is essentially identical.

We verify that the function defined in (9) satisfies the three properties of the definition. Property (1) holds since $l_\phi(q, 0) = 0$. Property (2) follows from the convexity of l_ϕ.

$$l_\phi\left(q, \sum_{i=1}^{N} \lambda_i f_i(x)\right) = l_\phi\left(q, \sum_{i=1}^{N} \sigma_i(x)|f_i(x)|\lambda_i\right) \tag{12}$$

$$\leq \int_{x\in\mathcal{X}} \sum_{i=1}^{N} \frac{|f_i(x)|}{f(x)}\ \ l_\phi\left(q, \sum_{i=1}^{N} \sigma_i(x) f(x)\lambda_i\right)\mu(dx) \tag{13}$$

The rest proof follows exactly the proof of Proposition 4.4 of [13]. ∎

We are then able to find feasible solutions for the LMB principle by using an algorithm that combines the previous AM algorithm with a nested IS loop to calculate the forward projection.

AM-IS algorithm:

Backward projection: Compute $p^k(x) = \tilde{p}(y)\ p_{\alpha,\lambda}^k(z|y)$, which yields $a_i = \int_{x\in\mathcal{X}} p^{(k)}(x) f_i(x)\mu(dx), i = 1, ..., N$ for the forward projection step.

Forward projection: Perform iterations of full parallel update IS as in (9) and (10) to obtain the parameter values $\lambda^\infty = (\lambda_1^\infty, ..., \lambda_N^\infty)$ and set $p_\lambda^k(x) = q_\infty^k(x)$. ∎

This alternating procedure will halt at a point where the three manifolds $\mathcal{C}$, $\mathcal{E}$ and $\mathcal{G}_a$ have a common intersection, since we reach a stationary point in that case. Due to the nonlinearity of the manifold $\mathcal{C}$, the intersection is not unique, and multiple feasible solutions may exist.

We are now ready to prove the main result of this section that AM-IS can be shown to converge and hence is guaranteed to yield feasible solutions to the LMB principle.

Theorem 2. *The AM-IS algorithm asymptotically yields feasible solutions to the LMB principle for additive models.*

Proof. By lemmas 5 and 6 and choose the closed convex set $\bar{\mathcal{M}}$ to play the role of $\mathcal{P}$ in Theorem 2, and choose the closed convex set $\bar{\mathcal{E}}$ to play the role of $\mathcal{Q}$ in Theorem 2. The conclusion immediately follows. ∎

Unlike under the K-L divergence in which the EM-IS algorithm monotonically increases the likelihood value during each iteration [24], this property will not maintain under the Bregman divergences.

5 Information geometry of AM-IS

We gain further insight by considering the well known Pythagorean theorem [13] for additive models, which in the complete data case states that if there exists $p_{\lambda^*} \in \bar{\mathcal{G}}_a \cap \bar{\mathcal{E}}$ then

$$B_\phi(p; p_\lambda) = B_\phi(p; p_{\lambda^*}) \ + \ B_\phi(p_{\lambda^*}; p_\lambda) \quad \text{for all } p \in \bar{\mathcal{G}}_a \text{ and } p_\lambda \in \bar{\mathcal{E}}$$

In the incomplete data case, this theorem needs to be modified to incorporate the effect of latent variables. Unlike the case in [24], in general, $\mathcal{M}$ is not a sub-manifold of $\bar{\mathcal{C}}$, thus the interpretation of the information geometry of Pythagorean theorem need to be slightly modified.

Theorem 3. *Pythagorean Property: For all $p_\lambda \in \bar{\mathcal{E}}$ and $p_{\lambda^*} \in \bar{\mathcal{C}} \cap \bar{\mathcal{E}}$, there exists a $p(x) \in \bar{\mathcal{M}}$ such that*

$$B_\phi(p; p_\lambda) = B_\phi(p; p_{\lambda^*}) \ + \ B_\phi(p_{\lambda^*}; p_\lambda) \tag{14}$$

Proof. For all $p_{\lambda^*} \in \bar{\mathcal{C}} \cap \bar{\mathcal{E}}$, pick $p(x) = \tilde{p}(y) p_{\alpha,\lambda^*}(z|y)$. Obviously $p \in \bar{\mathcal{M}}$. Now we show that for all $p_\lambda \in \bar{\mathcal{E}}$

$$B_\phi(\tilde{p}(y) p_{\alpha,\lambda^*}(z|y); p_\lambda(x)) = B_\phi(\tilde{p}(y) p_{\alpha,\lambda^*}(z|y); p_{\lambda^*}(x)) + B_\phi(p_{\lambda^*}(x); p_\lambda(x)) \tag{15}$$

Establishing (15) is equivalent to showing

$$\begin{aligned}
&\int_{x \in \mathcal{X}} \phi(\tilde{p}(y) p_{\alpha,\lambda^*}(z|y)) - \phi(p_\lambda(x)) - \phi'(p_\lambda(x)) \ (\tilde{p}(y) p_{\alpha,\lambda^*}(z|y) - p_\lambda(x)) \ \mu(dx) \\
&= \int_{x \in \mathcal{X}} \phi(\tilde{p}(y) p_{\alpha,\lambda^*}(z|y)) - \phi(p_{\lambda^*}(x)) - \phi'(p_{\lambda^*}(x)) \ (\tilde{p}(y) p_{\alpha,\lambda^*}(z|y) - p_{\lambda^*}(x)) \ \mu(dx) \\
&\quad + \int_{x \in \mathcal{X}} \phi(p_{\lambda^*}(x)) - \phi(p_\lambda(x)) - \phi'(p_\lambda(x)) \ (p_{\lambda^*}(x) - p_\lambda(x)) \ \mu(dx)
\end{aligned}$$

Cancelling common terms leaves

$$\begin{aligned}
&\int_{x \in \mathcal{X}} \phi'(p_\lambda(x)) \tilde{p}(y) p_{\alpha,\lambda^*}(z|y) \ \mu(dx) \\
&= \int_{x \in \mathcal{X}} \phi'(p_{\lambda^*}(x))(\tilde{p}(y) p_{\alpha,\lambda^*}(z|y) - p_{\lambda^*}(x)) \ \mu(dx) \\
&\quad + \int_{x \in \mathcal{X}} \phi'(p_\lambda(x)) p_{\lambda^*}(x) \ \mu(dx)
\end{aligned}$$

Plugging

$$p_\lambda(x) = \mathcal{L}\left(q_0, \sum_{i=1}^N \lambda_i f_i(x)\right) = (\phi')^{-1}\left(\phi'(q_0) + \sum_{i=1}^N \lambda_i f_i(x)\right)$$
$$p_{\lambda^*}(x) = \mathcal{L}\left(q_0, \sum_{i=1}^N \lambda_i^* f_i(x)\right) = (\phi')^{-1}\left(\phi'(q_0) + \sum_{i=1}^N \lambda_i^* f_i(x)\right)$$

into the above equation, we then have

$$\begin{aligned}
&\int_{x\in\mathcal{X}} \left(\phi'(q_0) + \sum_{i=1}^N \lambda_i f_i(x)\right) \tilde{p}(y) p_{\alpha,\lambda^*}(z|y)\ \mu(dx) \\
&= \int_{x\in\mathcal{X}} \left(\phi(q_0) + \sum_{i=1}^N \lambda_i^* f_i(x)\right) \left(\tilde{p}(y) p_{\alpha,\lambda^*}(z|y) - p_{\lambda^*}(x)\right)\ \mu(dx) \\
&\quad + \int_{x\in\mathcal{X}} \left(\phi(q_0) + \sum_{i=1}^N \lambda_i f_i(x)\right) p_{\lambda^*}(x)\ \mu(dx)
\end{aligned}$$

Cancelling the common terms, we then are left with

$$\begin{aligned}
&\sum_{i=1}^N (\lambda_i - \lambda_i^*) \left(\sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z\in\mathcal{Z}} p_{\alpha,\lambda^*}(z|y) f_i(x) \mu(dz) - \int_{x\in\mathcal{X}} p_{\lambda^*}(x) f_i(x) \mu(dx)\right) \\
&= \ 0
\end{aligned}$$

The term inside the brackets is 0 since $p_{\lambda^*}(x) \in \bar{\mathcal{C}} \cap \bar{\mathcal{E}}$. ∎

6 Summary

There are a number of iterative methods for performing Bregman divergence projections onto convex sets that can be used to illustrate existing supervised machine learning techniques. In this paper, we have presented a class of *unsupervised* statistical learning algorithms formulated in terms of minimizing Bregman divergences subject to a set of non-linear constraints that consider hidden variables. We have proposed a new alternating minimization algorithm with nested iterative scaling that asymptotically finds feasible solutions to this constrained optimization problem, and provided its convergence and information geometry properties.

We are developing this framework to provide analytical tools to transform current supervised machine learning techniques to unsupervised counterparts. Prelimilary experimental results on unsupervised boosting for clustering and gender independent speech signal analysis are encouraging.

References

1. H. Bauschke and J. Borwein, "Joint and Separate Convexity of the Bregman Distance," in: *Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications*, D. Butnariu, Y. Censor and S. Reich (Editors), Elsevier Science Publishers, Amsterdam, The Netherlands, 2001, pp. 23-36
2. J. Borwein and A. Lewis, "Duality relationships for entropy-like minimization problems," *SIAM Journal on Control and Optimization*, Vol. 29, No. 2, pp. 325-338, 1991
3. J. Borwein and A. Lewis, *Convex Analysis and Nolinear Optimization*, Springer 2000
4. L. Bregman, "The relaxation method of finding the common point of convex sets and its applications to the solution of problems in convex programming," *USSR Computational Mathematics and Mathematical Physics*, Vol. 7, pp. 200-217, 1967
5. A. Buja and W. Stuetzle, "Degrees of Boosting," manuscript, 2002
6. C. Byrne and Y. Censor, "Proximity function minimization using multiple Bregman projections with applications to split feasibility and Kullback-Leibler distance minimization," *Annals of Operations Research*, Vol. 105, pp. 77-98, 2001
7. Y. Censor and S. Zenios, *Parallel Optimization: Theory, Algorithms, and Applications*, Oxford University Press, 1997
8. M. Collins, R. Schapire and Y. Singer, "Logistic regression, AdaBoost and Bregman distances," *Machine Learning*, Vol. 48, No. 1-3, pp. 253-285, 2002
9. I. Csiszar and G. Tusnady, "Information geometry and alternating minimization procedures," *Statistics and Decisions*, Supplement Issue 1, pp. 205-237, 1984
10. I. Csiszar, "Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems," *The Annals of Statisics*, Vol. 19, No. 4, pp. 2032-2066, 1991
11. I. Csiszar, "Generalized projections for non-negative functions," *Acta Mathematica Hungarica*, Vol. 68, No. 1-2, pp. 161-185, 1995
12. I. Csiszar, "Maxent, mathematics, and information theory," *Maximum Entropy and Bayesian Methods*, Edited by K. Hanson and R. Silver, pp. 35-50, Kluwer Academic Publishers, 1996
13. S. Della Pietra, V. Della Pietra and J. Lafferty, "Duality and auxiliary functions for Bregman distances," Technical Report CMU-CS-01-109, School of Computer Science, CMU, 2001
14. A. Dempster, N. Laird and D. Rubin, "Maximum likelihood estimation from incomplete data via the EM algorithm," *Journal of Royal Statistical Society, Series B*, Vol. 39, pp 1-38, 1977
15. P. Eggermont and V. LaRiccia, "On EM-like algorithms for minimum distance estimation," Technical Report, Department of Mathematical Sciences, University of Delaware, 1998
16. J. Friedman, T. Hastie and R. Tibshirani, "Additive logistic regression: A statistical view of boosting," *Annals of Statistics*, Vol. 28, No. 2, pp. 337-407, 2000
17. R. Johnson and J. Shore, "Which is the better entropy expression for speech processing: -S log S or log S?," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 32, No. 1, pp. 129 -137, 1984
18. J. Lafferty, S. Della Pietra and V. Della Pietra, "Statistical learning algorithms based on Bregman distances," *Proceedings of Canadian Workshop on Information Theory*, pp. 77-80, 1997

19. J. Lafferty, "Additive models, boosting, and inference for generalized divergences," *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pp. 125-133, 1999
20. G. Lebanon and J. Lafferty, "Boosting and maximum likelihood for exponential models," In *Advances in Neural Information Processing Systems* (NIPS), 14, 2002
21. D. Luenberger, *Optimization by Vector Space Methods*, John Wiley & Sons, 1969
22. V. Vapnik, *The Natural of Statistical Learning Theory*, Springer, 2000
23. T. Zhang, "Statistical behavior and consistency of classification methods based on convex risk minimization," to appear in *Annals of Statistics*, 2003
24. S. Wang, D. Schuurmans and Y. Zhao, "The latent maximum entropy principle," manuscript, 2002
25. S. Wang, D. Schuurmans, A. Ghodsi and J. Rosenthal, "Unsupervised Boosting with the Latent Maximum Entropy Principle," manuscript, 2003