Clustering with Bregman Divergences.
Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, Joydeep Ghosh

Xiaowei Song
Math 710

Oct 15, 2015
Instructor: Prof. Jacob Kogan
Outline, Banerjee et al. [2005]

1. Bregman Divergence
   - Definition
   - Examples
   - Properties

2. Bregman Hard Clustering
   - Bregman Information
   - Clustering formulation
   - Clustering Algorithm

3. Bijection with Exponential Families
   - Exponential Families
   - Expectation parameters and Legendre duality
   - Exponential Families and Bregman Divergences
   - Bijection with Regular Bregman Divergences
   - Examples

4. Bregman Soft clustering

References
Bregman Divergence Definition

Bregman, 1967; Censor and Zenios, 1998

Definition (Bregman Divergence)

Let \( \Phi : S \mapsto \mathbb{R} \), \( S = \text{dom}(\Phi) \) be a strictly convex function defined on a convex set \( S \subseteq \mathbb{R}^d \) such that \( \Phi \) is differentiable on \( \text{ri}(S) \), assumed to be nonempty. The Bregman divergence \( d_\Phi : S \times \text{ri}(S) \mapsto [0, \infty) \) is defined as:

\[
d_\Phi(x, y) = \Phi(x) - \Phi(y) - \langle x - y, \nabla \Phi(y) \rangle
\]

where \( \nabla \Phi(y) \) represents the gradient vector of \( \Phi \) evaluated at \( y \).
Euclidean distance

\[ \Phi(x) = \langle x, x \rangle \text{ strictly convex and differentiable on } \mathbb{R}^d \Rightarrow \]
\[ d_\Phi(x, y) = \langle x, x \rangle - \langle y, y \rangle - \langle x - y, 2y \rangle = \|x - y\|^2 \]
\[ d_\Phi(x, y) \geq 0 \text{ as long as } \Phi \text{ convex} \]

(http://mark.reid.name/blog/meet-the-bregman-divergences.html)
Experiments about underlying distributions

<table>
<thead>
<tr>
<th>Generative Model</th>
<th>$d_{\text{Gaussian}}$</th>
<th>$d_{\text{Poisson}}$</th>
<th>$d_{\text{Binomial}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian</td>
<td>0.675 ± 0.032</td>
<td>0.659 ± 0.036</td>
<td>0.668 ± 0.035</td>
</tr>
<tr>
<td>Poisson</td>
<td>0.691 ± 0.036</td>
<td>0.724 ± 0.036</td>
<td>0.716 ± 0.036</td>
</tr>
<tr>
<td>Binomial</td>
<td>0.777 ± 0.038</td>
<td>0.799 ± 0.0345</td>
<td>0.798 ± 0.034</td>
</tr>
</tbody>
</table>

Each of 3 types’ mixed density generated 300 points, were clustered 100 trials. Compared to ground-truth with NMI.

KL-divergence

\[ \sum_{j=1}^{d} p_j = 1, \text{ neg-entropy } \Phi(p) = \sum_{j=1}^{d} p_j \log_2 p_j \text{ convex} \]

\[ d_{\Phi}(p, q) = \sum_{j=1}^{d} p_j \log_2 p_j - \sum_{j=1}^{d} q_j \log_2 q_j - \langle p - q, \nabla \Phi(q) \rangle \]
\[ = \sum_{j=1}^{d} p_j \log_2 p_j - \sum_{j=1}^{d} q_j \log_2 q_j - \sum (p_j - q_j) \left( \log_2 q_j + \log_2 e \right) \]
\[ = \sum_{j=1}^{d} p_j \log_2 \left( \frac{p_j}{q_j} \right) - (\log_2 e) \cdot \sum (p_j - q_j) \]
\[ = \text{KL}(p\| q) \]

for \( f(p) = p \log_2 p, 0 \leq p \leq 1, \frac{df}{dp} = \log_2 p + \log_2 e \),
\[ \frac{d^2 f}{dp^2} = \frac{1}{p} \log_2 e > 0 \Rightarrow f(p) \text{ convex in } [0, 1], \text{ thus } \sum f(p_j) \text{ convex in } 0 \leq p_j \leq 1 \]
Itakura-Saito distance

If $F(e^{j\theta})$ is the power spectrum 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 neg-entropy rate of the signal assuming it was generated by a stationary Gaussian process.

$$d_\Phi(F, G) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[ -\log\left(F(e^{j\theta})\right) + \log\left(G(e^{j\theta})\right) ight.$$

$$\left. - \left(F(e^{j\theta}) - G(e^{j\theta})\right) \left(-\frac{1}{G(e^{j\theta})}\right) \right] \, d\theta$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} \left( -\log\left(\frac{F(e^{j\theta})}{G(e^{j\theta})}\right) + \frac{F(e^{j\theta})}{G(e^{j\theta})} - 1 \right) \, d\theta$$
Bregman divergences generated from convex functions

<table>
<thead>
<tr>
<th>Domain</th>
<th>( \Phi(x) )</th>
<th>( d_\Phi(x, y) )</th>
<th>Divergence</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathbb{R} )</td>
<td>( x^2 )</td>
<td>( (x - y)^2 )</td>
<td>Squared loss</td>
</tr>
<tr>
<td>( \mathbb{R}^d )</td>
<td>( |x|^2 )</td>
<td>( |x - y|^2 )</td>
<td>Squared Euclidean distance</td>
</tr>
<tr>
<td>( \mathbb{R}^d )</td>
<td>( x^T A x )</td>
<td>( (x - y)^T A (x - y) )</td>
<td>Mahalanobis distance</td>
</tr>
<tr>
<td>( \mathbb{R}^d_+ )</td>
<td>( \log x )</td>
<td>( x \log \frac{x}{y} - (x - y) )</td>
<td>KL-divergence</td>
</tr>
<tr>
<td>( \mathbb{R}^d_+ )</td>
<td>( \sum_{j=1}^d x_j \log_2 x_j )</td>
<td>( \sum_{j=1}^d x_j \log_2 \frac{x_j}{y_j} - \log_2 e \times \left[ \sum_{j=1}^d (x_j - y_j) \right] )</td>
<td>Generalized I-divergence</td>
</tr>
<tr>
<td>([0,1] )</td>
<td>( x \log x + (1 - x) \log(1 - x) )</td>
<td>( x \log \frac{x}{y} + (1 - x) \log \frac{1-x}{1-y} )</td>
<td>Logistic Loss</td>
</tr>
<tr>
<td>( \mathbb{R}^d_+ )</td>
<td>( \sum_{j=1}^d x_j \log x_j )</td>
<td>( \sum_{j=1}^d x_j \log \frac{x_j}{y_j} - \log e \times \left[ \sum_{j=1}^d (x_j - y_j) \right] )</td>
<td>Itakura-Satio distance</td>
</tr>
</tbody>
</table>

### Function Names

<table>
<thead>
<tr>
<th>Function Name</th>
<th>( \varphi(x) )</th>
<th>( \text{dom } \varphi )</th>
<th>( D_\varphi(x, y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Squared norm</td>
<td>( \frac{1}{2} x^2 )</td>
<td>((-\infty, +\infty))</td>
<td>( \frac{1}{2} (x - y)^2 )</td>
</tr>
<tr>
<td>Shannon entropy</td>
<td>( x \log x - x )</td>
<td>( [0, +\infty) )</td>
<td>( x \log \frac{x}{x+y} )</td>
</tr>
<tr>
<td>Bit entropy</td>
<td>( x \log x + (1-x) \log(1-x) )</td>
<td>([0,1])</td>
<td>( x \log \frac{x}{x+y} + (1-x) \log \frac{1-x}{1-y} )</td>
</tr>
<tr>
<td>Burg entropy</td>
<td>( - \log x )</td>
<td>((0, +\infty))</td>
<td>( - \frac{\log x}{y} - \log \frac{y}{x} )</td>
</tr>
<tr>
<td>Hellinger</td>
<td>( - \sqrt{-x^2} )</td>
<td>([-1, 1])</td>
<td>( (1-xy)(1-y^2)^{-1/2} - (1-x^2)^{1/2} )</td>
</tr>
<tr>
<td>( \ell_p ) quasi-norm</td>
<td>( - x^p )</td>
<td>((0 &lt; p &lt; 1))</td>
<td>( - x^{p+1} + p x y y^{p-1} - (p-1) y^p )</td>
</tr>
<tr>
<td>( \ell_p ) norm</td>
<td>(</td>
<td>x</td>
<td>^p )</td>
</tr>
<tr>
<td>Exponential</td>
<td>( e^x )</td>
<td>((-\infty, +\infty))</td>
<td>( e^x - (x + y + 1) e^y )</td>
</tr>
</tbody>
</table>

**Hellinger:**
\[
\varphi(x) = -\sqrt{1 - \|x\|^2}
\]
\[
D_\varphi(x, y) = \frac{1-x^T y}{\sqrt{1-\|x\|^2}} - \sqrt{1-\|y\|^2}
\]
Appendix A. Properties

1. Non-negativity. $d_\Phi(x, y) \geq 0, \forall x \in S, y \in ri(S)$, and equality holds IFF $x = y$. (Not a metric: not symmetric and triangle inequality not hold)

2. Convexity. $d_\Phi$ is always convex in the 1st argument, but not necessary convex in the 2nd argument. While, Squared Euclidean distance and KL-divergence are convex in both of their arguments.

3. Linearity. Bregman divergence is a linear operator, i.e.,

$$\forall x \in S, y \in ri(S),$$

$$d_{\Phi_1 + \Phi_2}(x, y) = d_{\Phi_1}(x, y) + d_{\Phi_2}(x, y)$$

$$d_{c\Phi}(x, y) = cd_\Phi(x, y), \ c \geq 0$$
Appendix A. Properties

4 Equivalence classes. The Bregman divergences of functions that differ only in affine terms are identical, i.e.,
if \( \Phi(x) = \Phi_0(x) + \langle b, x \rangle + c, b \in \mathbb{R}^d, c \in \mathbb{R} \), then
\( d_\Phi(x, y) = d_{\Phi_0}(x, y), \forall x \in S, y \in ri(S) \). Hence, the set of all strictly convex, differentiable functions on a convex set \( S \) can be partitioned into equivalence classes of the form

\[ [\Phi_0] = \{ \Phi | d_\Phi(x, y) = d_{\Phi_0}(x, y), \forall x \in S, y \in ri(S) \} \]

5 Linear separation.

\[ d_\Phi(x, \mu_1) = d_\Phi(x, \mu_2) \]

\[ \Rightarrow \Phi(x) - \Phi(\mu_1) - \langle x - \mu_1, \nabla \Phi(\mu_1) \rangle = \]

\[ \Phi(x) - \Phi(\mu_2) - \langle x - \mu_2, \nabla \Phi(\mu_2) \rangle \]

\[ \Rightarrow \langle x, \nabla \Phi(\mu_2) - \nabla \Phi(\mu_1) \rangle = \]

\[ (\Phi(\mu_1) - \Phi(\mu_2)) - (\langle \mu_1, \nabla \Phi(\mu_1) \rangle - \langle \mu_2, \nabla \Phi(\mu_2) \rangle) \]
Appendix A. Properties

6 Dual divergences/Conjugate duality: let \( \Psi(\theta) = \Phi^*(\theta) \) be the conjugate of \( \Phi(u) \). Then \( d_\Phi(\mu_1, \mu_2) = d_\Psi(\theta_2, \theta_1) \)

\[
\Psi(\theta) = \Phi^*(\theta) = \sup_u \{ \theta^T u - \Phi(u) \}
\]

Properties of conjugate function:
1). let \( 0 = \nabla_u g(\theta, u) = \theta - \nabla \Phi(u^*) \)
2). \( \Phi \) convex \( \Rightarrow \) \( \Psi \) convex
3). \( \Phi \) convex and closed \( \Rightarrow \) \((\Phi^*)^* = \Phi\)
Proof of Conjugate duality

\[ d_\Phi(u_1, u_2) = \Phi(u_1) - \Phi(u_2) - (u_1 - u_2)^T \nabla \Phi(u_2) \]

\[ = \Phi(u_1) - \Phi(u_2) - (u_1 - u_2)^T \theta_2 + u_1^T \theta_1 - u_1^T \theta_1 \nabla \psi(\theta_1) \]

\[ = \Phi(u_1) - \Phi(u_2) - (\theta_2 - \theta_1)^T \nabla \psi(\theta_1) + u_2^T \theta_2 - u_1^T \theta_1 \]

\[ = [\theta_2^T u_2 - \Phi(u_2)] - [\theta_1^T u_1 - \Phi(u_1)] - (\theta_2 - \theta_1)^T \nabla \psi(\theta_1) \]

\[ = \psi(\theta_2) - \psi(\theta_1) - (\theta_2 - \theta_1)^T \nabla \psi(\theta_1) \]

\[ = d_\psi(\theta_2, \theta_1) \]
Let $\mathcal{F}_\Psi$ be an exponential family with $\Psi$ as the cumulant function. 

$$KL\left(p_{(\Psi, \theta_1)} \parallel p_{(\Psi, \theta_2)}\right) = d_{\Psi}(\theta_2, \theta_1) = d_{\Phi}(\mu_1, \mu_2)$$

where $\mu_1, \mu_2$ are the expectation parameters corresponding to $\theta_1, \theta_2$. Further, if $\Psi(0) = 0$, then $p_{(\Psi, 0)}(x) = p_0(x)$ is itself a valid probability density and $KL\left(p_{(\Psi, \theta)} \parallel p_{(\Psi, 0)}\right) = \Phi(\mu)$, where $\mu = \nabla \Psi(\theta)$.
Generalized Pythagoras theorem:
\[ \forall x \in \Omega: \ D_\varphi(x, y) \geq D_\varphi(x, P_\Omega(y)) + D_\varphi(P_\Omega(y), y) \]

Opposite to triangle inequality:
"Law of cosine"

Three point property generalizes the "law of cosine":

$$D_\varphi(x, y) = D_\varphi(x, z) + D_\varphi(z, y) - (x - z)^T (\nabla \varphi(y) - \nabla \varphi(z))$$

Euclidean special case:

$$\|x - y\|^2 = \|x - z\|^2 + \|z - y\|^2 - 2 (x - z)^T (y - z)$$
A divergence measure \( d : S \times ri(S) \rightarrow [0, \infty) \) is a Bregman divergence IFF there exists \( a \in ri(S) \) such that the function \( \Phi_a(x) = d(x, a) \) satisfies the following conditions:

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

Proof of necessity: any strictly convex, differentiable function \( \Phi \), the Bregman divergence evaluated with a fixed value for the 2nd argument differs from it only by a linear term, i.e.,

\[
\Phi_a(x) = d_\Phi(x, a) = \Phi(x) - \Phi(a) - \langle x - a, \nabla \Phi(a) \rangle = \Phi(x) - \langle x, \nabla \Phi(a) \rangle - \Phi(a) + \langle a, \nabla \Phi(a) \rangle = \Phi(x) + \langle b, x \rangle + c
\]

where \( b = -\nabla \Phi(a) \), \( c = -\Phi(a) + \langle a, \nabla \Phi(a) \rangle \).
Proved in class by Prof. Kogan.

For the data points above $a_i, 1 \leq i \leq m$, we want to find one point closest to all data points, define cost function: $f(x) = \sum_{i=1}^{m} |x - a_i|^2$, we want to get $\min_{x \in \mathbb{R}^n} f(x)$, then use the found $x$ to represent $a_i, 1 \leq i \leq m$

let $0 = \frac{df(x)}{dx} = \frac{d}{dx} \left[ \sum (x - a_i)^2 \right] = 2 \sum (x - a_i) = 2 \sum x - 2 \sum a_i$

$\Rightarrow mx = \sum_{i=1}^{m} a_i \Rightarrow x = \frac{1}{m} \sum_{i=1}^{m} a_i$ which is the mean of all data points.
Proposition 1

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

$$\min_{s \in ri(S)} E_\nu [d_\Phi(X, s)]$$

has a unique minimizer given by $s^\dagger = \mu = E_\nu[X]$. Note the minimization is with respect to 2nd argument, surprising since Bregman divergences are not necessarily convex in the 2nd argument.
Proposition 1 - Proof

The function we are trying to minimize is

\[ J_\Phi(s) = E_v [d_\Phi(X, s)] = \sum_{i=1}^{n} v_i d_\Phi(x_i, s) \]. Since

\[ \mu = E_v [X] \in ri(S) \], the objective function is well-defined at \( \mu \).

Now \( \forall s \in ri(S) \),

\[ J_\Phi(s) - J_\Phi(\mu) \]

\[ = \sum_{i=1}^{n} v_i d_\Phi(x_i, s) - \sum_{i=1}^{n} v_i d_\Phi(x_i, \mu) \]

\[ = \Phi(\mu) - \Phi(s) - \left\langle \sum_{i=1}^{n} v_i x_i - s, \nabla \Phi(s) \right\rangle + \left\langle \sum_{i=1}^{n} v_i x_i - \mu, \nabla \Phi(\mu) \right\rangle \]

\[ = \Phi(\mu) - \Phi(s) - \langle \mu - s, \nabla \Phi(s) \rangle \]

\[ = d_\Phi(\mu, s) \geq 0 \]

with equality holds only when \( s = \mu \)
Definition (Bregman Information)

Let $X$ be a random variable that takes values in $\mathcal{X} = \{x_i\}_{i=1}^n \subset S$ following a probability measure $\nu$. Let $\mu = E_\nu[X] = \sum_{i=1}^n v_i x_i \in ri(S)$ and let $d_\Phi : S \times ri(S) \mapsto [0, \infty)$ be a Bregman divergence. Then the Bregman Information of $X$ in terms of $d_\Phi$ is defined as:

$$I_\Phi(X) = E_\nu[d_\Phi(X, \mu)] = \sum_{i=1}^n v_i d_\Phi(x_i, \mu)$$

Example 5. Variance:

Let $\mathcal{X} = \{x_i\}_{i=1}^n$ be a set in $\mathbb{R}^d$, and uniform measure $v_i = \frac{1}{n}$ over $\mathcal{X}$. The Bregman Information of $X$ with squared Euclidean distance as the Bregman divergence is given by:

$$I_\Phi(X) = \sum_{i=1}^n v_i d_\Phi(x_i, \mu),$$

which is sample variance.
Example 6. Mutual information:

By definition, the mutual information $I(U; V)$ between 2 discrete random variables $U$ and $V$ with joint distribution $\{\{p(u_i, v_j)\}_{i=1}^n\}_{j=1}^m$ is given by

$$I(U; V) = \sum_{i=1}^n \sum_{j=1}^m p(u_i, v_j) \log \frac{p(u_i, v_j)}{p(u_i)p(v_j)}$$

$$= \sum_{i=1}^n p(u_i) \sum_{j=1}^m p(v_j \mid u_i) \log \frac{p(v_j \mid u_i)}{p(v_j)}$$

$$= \sum_{i=1}^n p(u_i) KL(p(V \mid u_i) \parallel p(V))$$

Consider RV $Z_u$ taking values in the set of probability distributions $Z_u = \{p(V \mid u_i)\}_{i=1}^n$ following the probability measure $\{v_i\}_{i=1}^n = \{p(u_i)\}_{i=1}^n$ over this set. The mean (distribution) of $Z_u$ is given by:

$$\mu = E_v [p(V \mid u)] = \sum_{i=1}^n p(u_i)p(V \mid u_i) = \sum_{i=1}^n p(u_i, V) = p(V)$$

hence, $I(U, V) = \sum_{i=1}^n v_i d_\phi(p(V \mid u_i), \mu) = I_\phi(Z_u)$, similarly, $I(U; V) = I_\phi(Z_V)$
Jensen’s Inequality and Bregman Information

Given any convex function \( \Phi \), for any random variable \( X \), Jensen’s inequality:

\[
E[\Phi(X)] \geq \Phi(E[X])
\]

\[
E[\Phi(X)] - \Phi(E[X]) = E[\Phi(X)] - \Phi(E[X]) - E[\langle X - E[X], \nabla \Phi(E[X]) \rangle] = E[\Phi(X) - \Phi(E[X]) - \langle X - E[X], \nabla \Phi(E[X]) \rangle] = E[d_\Phi(X, E(X))] = I_\Phi(X) \geq 0
\]
Clustering by Expected Bregman divergence

RV $X$ takes values in $\mathcal{X} = \{x_i\}_{i=1}^n$ following prob measure $v$. When $X$ has large Bregman information, it may not suffice to encode $X$ using single representative since lower quantization error may be desired.

Split the set $\mathcal{X}$ into $k$ disjoint partitions $\{\mathcal{X}_h\}_{h=1}^k$, each with its own Bregman representative, RV $M$ over the partition representatives as an appropriate quantization of $X$, which is $\mathcal{M} = \{\mu_h\}_{h=1}^k$, its probability as $\pi_h = \sum_{x_i \in \mathcal{X}_h} v_i$.

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

$$E_X [d_\Phi(X, M)] = \sum_{h=1}^k \sum_{x_i \in \mathcal{X}_h} v_i d_\Phi(x_i, \mu_h)$$

$$= \sum_{h=1}^k \pi_h \sum_{x_i \in \mathcal{X}_h} \frac{v_i}{\pi_h} d_\Phi(x_i, \mu_h)$$

$$= E_\pi [I_\Phi(X_h)]$$
In Information-theoretic clustering, the quality of partitioning is measured in terms of loss in mutual information resulting from the quantization of the original RV $X$, i.e., $I_{\Phi}(X) - I_{\Phi}(M)$.

Hard clustering problem is defined as finding a partitioning of $X$, 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.

**Theorem (Information theoretic clustering)**

Let $X$ be a RV that takes values in $\mathcal{X} = \{x_i\}_{i=1}^n \subset S \subseteq \mathbb{R}^d$ following positive probability measure $\nu$. Let $\{\mathcal{X}_h\}_{h=1}^k$ be a partitioning of $\mathcal{X}$ and let $\pi_h = \sum_{x_i \in \mathcal{X}_h} \nu_i$ be the induced measure $\pi$ on the partitions. Let $X_h$ be the RV that takes values in $\mathcal{X}_h$ following $\frac{\nu_i}{\pi_h}$ for $x_i \in \mathcal{X}_h$, $h = 1, \ldots, k$. Let $\mathcal{M} = \{\mu_h\}_{h=1}^k$ with $\mu_h \in ri(S)$ denote the set of representatives of $\{\mathcal{X}_h\}_{h=1}^k$, and $M$ be a RV that takes values in $\mathcal{M}$ following $\pi$. Then

$$L_{\Phi}(M) = I_{\Phi}(X) - I_{\Phi}(M) = E_{\pi} [I_{\Phi}(X_h)] = \sum_{h=1}^k \pi_h \sum_{x_i \in \mathcal{X}_h} \frac{\nu_i}{\pi_h} d_{\Phi}(x_i, \mu_h)$$
Information-theoretic clustering Proof

\[ I_\Phi(X) \]

\[ = \sum_{i=1}^{n} v_i d_\Phi(x_i, \mu) = \sum_{h=1}^{k} \sum_{x_i \in X_h} v_i d_\Phi(x_i, \mu) = \sum_{h=1}^{k} \sum_{x_i \in X_h} v_i \left\{ \Phi(x_i) - \Phi(\mu) - \langle x_i - \mu, \nabla \Phi(\mu) \rangle \right\} \]

\[ = \sum_{h=1}^{k} \sum_{x_i \in X_h} v_i \left\{ \Phi(x_i) - \Phi(\mu_h) - \langle x_i - \mu_h, \nabla \Phi(\mu_h) \rangle + \langle x_i - \mu_h, \nabla \Phi(\mu_h) \rangle \right\} \]

\[ + \Phi(\mu_h) - \Phi(\mu) - \langle (x_i - \mu_h) + (\mu_h - \mu), \nabla \Phi(\mu) \rangle \right\} \]

\[ = \sum_{h=1}^{k} \sum_{x_i \in X_h} v_i \left\{ d_\Phi(x_i, \mu_h) + \langle x_i - \mu_h, \nabla \Phi(\mu_h) - \nabla \Phi(\mu) \rangle + d_\Phi(\mu_h, \mu) \right\} \]

\[ = \sum_{h=1}^{k} \pi_h \sum_{x_i \in X_h} v_i \left\{ d_\Phi(x_i, \mu_h) + \sum_{h=1}^{k} \sum_{x_i \in X_h} v_i d_\Phi(\mu_h, \mu) + \sum_{h=1}^{k} \pi_h \sum_{x_i \in X_h} v_i \langle x_i - \mu_h, \nabla \Phi(\mu_h) - \nabla \Phi(\mu) \rangle \right\} \]

\[ = \sum_{h=1}^{k} \pi_h l_\Phi(X_h) + \sum_{h=1}^{k} \pi_h d_\Phi(\mu_h, \mu) + \sum_{h=1}^{k} \pi_h \left\langle \sum_{x_i \in X_h} v_i x_i - \mu_h, \nabla \Phi(\mu_h) - \nabla \Phi(\mu) \right\rangle \]

\[ = E_\pi [l_\Phi(X_h)] + I_\Phi(M) \]
Information-theoretic clustering interpretation

Within/Between cluster interpretation

- **Total Bregman Information** = $I_{\Phi}(X) = L_{\Phi}(M) + I_{\Phi}(M)$
- **Within-cluster Bregman Information**

$$L_{\Phi}(M) = I_{\Phi}(X) - I_{\Phi}(M) = E_{\pi} [I_{\Phi}(X_h)] = \sum_{h=1}^{k} \sum_{x_i \in X_h} v_i d_{\Phi}(x_i, \mu_h)$$

- **Between-cluster Bregman Information** = $I_{\Phi}(M)$

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

$$\min_{M} \left( I_{\Phi}(X) - I_{\Phi}(M) \right) = \min_{M} \left( \sum_{h=1}^{k} \sum_{x_i \in X_h} v_i d_{\Phi}(x_i, \mu_h) \right)$$
Bregman Hard Clustering Algorithm

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

Output: $\mathcal{M}^\dagger$, local minimizer of $L_\Phi(\mathcal{M}) = \sum_{h=1}^k \sum_{x_i \in \mathcal{X}_h} \nu_i d_\Phi(x_i, \mu_h)$ where $\mathcal{M} = \{\mu_h\}_{h=1}^k$, hard partitioning $\{\mathcal{X}_h\}_{h=1}^k$ of $\mathcal{X}$.

Method: Initialize $\{\mu_h\}_{h=1}^k$ with $\mu_h \in ri(S)$ (one possible initialization is to choose $\mu_h \in ri(S)$ at random)

repeat
  * The assignment Step
  Set $\mathcal{X}_h \leftarrow \emptyset$, $1 \leq h \leq k$
  for $i=1$ to $n$ do
    $\mathcal{X}_h \leftarrow \mathcal{X}_h \cup \{x_i\}$
  where $h = h^\dagger(x_i) = \arg\min_{h'} d_\Phi(x_i, \mu_h')$
  endfor
  * The Re-estimation Step
  for $h = 1$ to $k$ do
    $\pi_h \leftarrow \sum_{x_i \in \mathcal{X}_h} \nu_i$
    $\mu_h \leftarrow \frac{1}{\pi_h} \sum_{x_i \in \mathcal{X}_h} \nu_i x_i$
  endfor
  until convergence
return $\mathcal{M}^\dagger \leftarrow \{\mu_h\}_{h=1}^k$
Proof: Convergence and terminates in a finite steps at local optimal partition

The Bregman hard clustering algorithm monotonically decreases the loss function
\[ \min_M (I_\Phi(X) - I_\Phi(M)) = \min_M \left( \sum_{h=1}^{k} \sum_{x_i \in \mathcal{X}_h} v_id_\Phi(x_i, \mu_h) \right). \]

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

\[
L_\Phi(M^{(t)}) = \sum_{h=1}^{k} \sum_{x_i \in \mathcal{X}_h^{(t)}} v_i d_\Phi(x_i, \mu_h^{(t)})
\]

\[
\geq \sum_{h=1}^{k} \sum_{x_i \in \mathcal{X}_h^{(t)}} v_i d_\Phi(x_i, \mu_h^{(t)}(X_i))
\]

\[
\geq \sum_{h=1}^{k} \sum_{x_i \in \mathcal{X}_h^{(t+1)}} v_i d_\Phi(x_i, \mu_h^{(t+1)}) = L_\Phi(M^{(t)})
\]
Properties of hard clustering

- **Exhaustiveness**: the algorithm works for all Bregman divergences and only for Bregman divergences since the arithmetic mean is the best predictor only for Bregman divergences.

- **Linear Separators**: The locus of points that are equidistant to 2 fixed points $\mu_1, \mu_2$ in terms of a Bregman divergence is given by $\mathcal{X} = \{ x | d_\Phi (x, \mu_1) = d_\Phi (x, \mu_2) \}$, i.e., the set of points, $\{x | \langle x, \nabla \Phi (\mu_2) - \nabla \Phi (\mu_1) \rangle = (\Phi (\mu_1) - \Phi (\mu_2)) - (\langle \mu_1, \nabla \Phi (\mu_1) \rangle - \langle \mu_2, \nabla \Phi (\mu_2) \rangle)\}$

- **Scalability**: computational complexity of each iteration is linear in number of data points and number of desired cluster for all Bregman divergences.

- **Applicability to mixed data types**: One can choose different convex functions appropriate and meaningful for different subsets of the features. We can build a convex combination corresponding to Bregman divergence.
Consider a family $\mathcal{F}$ of probability densities on a measurable space $(\Omega, \mathcal{B})$ where $\mathcal{B}$ is a $\sigma$-algebra on the set $\Omega$. Suppose every probability density, $p_\theta \in \mathcal{F}$, is parameterized by $d$ real-valued variables $\theta = \{\theta_j\}_{j=1}^d$ so that

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

Let $H : \mathcal{B} \mapsto \mathcal{G}$ transforms any RV $U : \mathcal{B} \mapsto \mathbb{R}$ to a RV $V : \mathcal{G} \mapsto \mathbb{R}$ with $V = H(U)$. Then given the probability density $p_\theta$ of $U$, $H$ uniquely determines the probability density $q_\theta$ governing the RV $V$.

**Definition (sufficient statistic)**

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

Definition (exponential family, natural parameter)

If $d$-dimensional model $\mathcal{F} = \{p_\theta | \theta \in \Gamma\}$ can be expressed in terms of $(d + 1)$-real-valued linearly independent functions $\{C, H_1, \ldots, H_d\}$ on $\mathcal{B}$ and a function $\Psi$ on $\Gamma$ as $f(\omega; \theta) = \exp \left\{ \sum_{j=1}^{d} \theta_j H_j(\omega) - \psi(\theta) + C(\omega) \right\}$, then $\mathcal{F}$ is called an exponential family, and $\theta$ is called its natural parameter.

If $\exists x \in \mathbb{R}^d$ such that $x_j = H_j(\omega)$, then density function $g(x; \theta) = \exp \left\{ \sum_{j=1}^{d} \theta_j x_j - \psi(\theta) - \lambda(x) \right\}$ for a uniquely determined function $\lambda(x)$, is such that $f(\omega; \theta)/g(x; \theta)$ does not depend on $\theta$. Thus $x$ is sufficient statistic for the family.

Definition (exponential family, log-partition/cumulant function)

A multivariate parametric family $\mathcal{F}_\Psi$ of distribution $\{p(\psi, \theta) | \theta \in \Gamma \subseteq \mathbb{R}^d\}$ is called an exponential family if the probability density is of the form: $p(\psi, \theta) = \exp (\langle x, \theta \rangle - \psi(\theta) - \lambda(x))$. The function $\psi(\theta)$ is known as log partition function or the cumulant function and it uniquely determines the exponential family $\mathcal{F}_\Psi$. Further, given $\mathcal{F}_\Psi, \psi$ is uniquely determined up to a constant additive term. Amari [1995] showed $\Gamma$ is a convex set in $\mathbb{R}^d$ and $\psi$ is a strictly convex and differentiable function on $\text{int}(\Gamma)$. 
Consider a d-dimensional real RV \( X \) following an exponential family density \( p(\psi, \theta) \) specified by natural parameter \( \theta \in \Gamma \). The expectation of \( X \) with respect to \( p(\psi, \theta) \), also called the expectation parameter, is given by:

\[
\mu = \mu(\theta) = E_{p(\psi, \theta)}[X] = \int_{\mathbb{R}^d} x p(\psi, \theta)(x) dx.
\]

Amari [1995] showed that expectation and natural parameters have a one-one correspondence with each other and span spaces that exhibit a dual relationship.

**Theorem (Rockafellar, 1970)**

Let \( \Psi \) be a real-valued proper closed convex function with conjugate function \( \Psi^* \). Let \( \Theta = \text{int}(\text{dom}(\Psi)) \) and \( \Theta^* = \text{int}(\text{dom}(\Psi^*)) \). If \((\Theta, \Psi)\) is a convex function of Legendre type, then

1. \((\Theta^*, \Psi^*)\) is a convex function of Legendre type.
2. \((\Theta, \Psi)\) and \((\Theta^*, \Psi^*)\) are Legendre duals of each other,
3. The gradient function \( \nabla \Psi : \Theta \mapsto \Theta^* \) is a one-to-one function from the open convex set \( \Theta \) onto the open convex set \( \Theta^* \),
4. The gradient functions \( \nabla \Psi, \nabla \Psi^* \) are continuous, and \( \nabla \Psi^* = (\nabla \Psi)^{-1} \).
Differentiating $1 = \int p(\psi, \theta)(x)dx$ with respect to $\theta$

$0 = \frac{\partial}{\partial \theta} \int \exp (\langle x, \theta \rangle - \psi(\theta) - \lambda(x)) \, dx = \int (x - \nabla \psi(\theta)) \, p(\psi, \theta)(x) \, dx$

$\Leftrightarrow \nabla \psi(\theta) \int p(\psi, \theta)(x) \, dx = \int xp(\psi, \theta)(x) \, dx$

$\Leftrightarrow \nabla \Psi(\theta) = \mu(\theta) = \mu$

Let $\Phi$ be defined as the conjugate of $\Psi$, i.e.,

$\Phi(\mu) = \sup_{\theta \in \Theta} \{ \langle \mu, \theta \rangle - \Psi(\theta) \}$.

Then $\Phi = \Psi^*$ and $\text{int}(\text{dom}(\Phi)) = \Theta^*$, thus by Legendre transformation:

$\mu(\theta) = \nabla \Psi(\theta)$ and $\theta(\mu) = \nabla \Phi(\mu)$,

$\Phi(\mu) = \langle \theta(\mu), \theta \rangle - \Psi(\theta(\mu))$, $\forall \mu \in \text{int}(\text{dom}(\Phi))$
**Exponential Families and Bregman Divergences**

\[
\log \left( p_{(\psi, \theta)}(x) \right) = \langle x, \theta \rangle - \psi(\theta) - \lambda(x) \\
= [\langle \mu, \theta \rangle - \psi(\theta)] - \lambda(x) + \langle x - \mu, \theta \rangle \\
= \Phi(\mu) + \langle x - \mu, \nabla \Phi(\mu) \rangle - \lambda(x) \\
= [\Phi(\mu) + \langle x - \mu, \nabla \Phi(\mu) \rangle - \Phi(x)] + \Phi(x) - \lambda(x) \\
= -d_\Phi(x, \mu(\theta)) + \Phi(x) - \lambda(x)
\]

**Theorem (4. pdf expressed by Bregman Divergence)**

Let \( p_{(\psi, \theta)} \) be the pdf of a regular exponential family distribution. Let \( \Phi \) be the conjugate function of \( \Psi \) so that \((\text{int} \ (\text{dom} (\Phi)), \Phi)\) is the Legendre dual of \((\Theta, \Psi)\). Let \( \theta \in \Theta \) be the natural parameter and \( \mu \in \text{int} \ (\text{dom} (\Phi)) \) be the corresponding expectation parameter. Let \( d_\Phi \) be the Bregman divergence derived from \( \Phi \). Then \( p_{(\psi, \theta)} \) can be uniquely expressed as \( p_{(\psi, \theta)}(x) = \exp \left( -d_\Phi(x, \mu) \right) b_\Phi(x), \forall x \in \text{dom} (\Phi) \), where \( b_\Phi: \text{dom} (\Phi) \mapsto \mathbb{R}_+ \) is a uniquely determined function.
Theorem (Devinatz, 1955)

Let $\Theta \in \mathbb{R}^d$ be an open convex set. A necessary and sufficient condition that there exists a unique, bounded, non-negative measure $\nu$ such that $f : \Theta \mapsto \mathbb{R}^{++}$ can be represented as $f(\theta) = \int_{x \in \mathbb{R}^d} \exp(\langle x, \theta \rangle) \, d\nu(x)$ is that $f$ is continuous and exponentially convex.

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

Theorem (Bijection)

There is a bijection between regular exponential families and regular Bregman divergences.
## Examples

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$p(x; \theta)$</th>
<th>$\mu$</th>
<th>$\Phi(\mu)$</th>
<th>$d_\Phi(x, \mu)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-D Gaussian</td>
<td>$\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(x-a)^2}{2\sigma^2} \right]$</td>
<td>$a$</td>
<td>$\frac{1}{2\sigma^2} \mu^2$</td>
<td>$\frac{1}{2\sigma^2} (x - \mu)^2$</td>
</tr>
<tr>
<td>1-D Poisson</td>
<td>$\frac{1}{\sqrt{2\pi\sigma^2}} \lambda^x \exp (-\lambda)$</td>
<td>$\lambda$</td>
<td>$\mu \log \mu - \mu$</td>
<td>$x \log \frac{x}{\mu} - (x - \mu)$</td>
</tr>
<tr>
<td>1-D Bernoulli</td>
<td>$q^x (1 - q)^{1-x}$</td>
<td>$q$</td>
<td>$\mu \log \mu + (1 - \mu) \log (1 - \mu)$</td>
<td>$x \log \frac{x}{\mu} + (1 - x) \log \frac{1-x}{1-\mu}$</td>
</tr>
<tr>
<td>1-D Binomial</td>
<td>$\binom{N}{x} q^x (1 - q)^{N-x}$</td>
<td>$Nq$</td>
<td>$\mu \log \frac{\mu}{N} + (N - \mu) \log \frac{N-\mu}{N}$</td>
<td>$x \log \frac{x}{\mu} + (N - x) \log \frac{N-x}{N-\mu}$</td>
</tr>
<tr>
<td>1-D Exponential</td>
<td>$\lambda \exp (-\lambda x)$</td>
<td>$\frac{1}{\lambda}$</td>
<td>$-\log \mu - 1$</td>
<td>$\frac{x}{\mu} - \log \frac{x}{\mu} - 1$</td>
</tr>
<tr>
<td>d-D Sph. Gaussian</td>
<td>$\frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp \left[ -\frac{|x-a|^2}{2\sigma^2} \right]$</td>
<td>$a$</td>
<td>$\frac{1}{2\sigma^2} |\mu|^2$</td>
<td>$\frac{1}{2\sigma^2} |x - \mu|^2$</td>
</tr>
<tr>
<td>d-D multinomial</td>
<td>$\frac{1}{\prod_{j=1}^k x_j!} \prod_{j=1}^d q_{x_j}^{x_j} \left[ \binom{N}{x_j} \right]_{j=1}^{d-1}$</td>
<td>$\sum_{j=1}^d \mu_j \log \frac{\mu_j}{N}$</td>
<td>$\sum_{j=1}^d x_j \log \frac{x_j}{\mu_j}$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$\theta$</th>
<th>$\Psi(\theta)$</th>
<th>$\text{dom}(\Psi)$</th>
<th>$\text{dom}(\Phi)$</th>
<th>$h_\Psi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-D Gaussian</td>
<td>$\frac{a}{\sigma^2}$</td>
<td>$\frac{\sigma^2}{2} \theta^2$</td>
<td>$\mathbb{R}$</td>
<td>$\mathbb{R}$</td>
<td>$\mathbb{R}$</td>
</tr>
<tr>
<td>1-D Poisson</td>
<td>$\log \lambda$</td>
<td>$\exp \theta$</td>
<td>$\mathbb{R}$</td>
<td>$\mathbb{R}_+$</td>
<td>$[0, 1]$</td>
</tr>
<tr>
<td>1-D Bernoulli</td>
<td>$\log \frac{q}{1-q}$</td>
<td>$\log (1 + \exp \theta)$</td>
<td>$\mathbb{R}$</td>
<td>$\mathbb{R}_+$</td>
<td>${0, 1}$</td>
</tr>
<tr>
<td>1-D Binomial</td>
<td>$\log \frac{q}{1-q}$</td>
<td>$N \log (1 + \exp \theta)$</td>
<td>$\mathbb{R}$</td>
<td>$\mathbb{R}_+$</td>
<td>${0, 1, \ldots, N}$</td>
</tr>
<tr>
<td>1-D Exponential</td>
<td>$-\lambda$</td>
<td>$-\log (-\theta)$</td>
<td>$\mathbb{R}_{--}$</td>
<td>$\mathbb{R}_{++}$</td>
<td>$\mathbb{R}_+$</td>
</tr>
<tr>
<td>d-D Sph. Gaussian</td>
<td>$\frac{a}{\sigma^2}$</td>
<td>$\frac{\sigma^2}{2} |\theta|^2$</td>
<td>$\mathbb{R}^d$</td>
<td>$\mathbb{R}^d$</td>
<td>$\mathbb{R}^d$</td>
</tr>
<tr>
<td>d-D multinomial</td>
<td>$\left[ \log \frac{q_j}{a_j} \right]_{j=1}^{d-1}$</td>
<td>$N \log \left( 1 + \sum_{j=1}^{d-1} \exp \theta_j \right)$</td>
<td>$\mathbb{R}^{d-1}$</td>
<td>${ \mu \in \mathbb{R}_{++}^{d-1},</td>
<td>\mu</td>
</tr>
</tbody>
</table>
Example 9, spherical Gaussian distributions

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

\[ = \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp \left( \langle x, \frac{a}{\sigma^2} \rangle - \frac{\|a\|^2}{2\sigma^2} - \frac{x^2}{2\sigma^2} \right) \]

\[ = \exp \left( \langle x, \theta \rangle - \frac{\sigma^2}{2} \|\theta\|^2 \right) \exp \left( -\frac{1}{2\sigma^2} \|x\|^2 \right) \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \]

\[ = \exp (\langle x, \theta \rangle - \Psi(\theta)) p_0(x) \]

\[ \therefore \mu = \nabla \Psi(\theta) = \nabla \left( \frac{\sigma^2}{2} \|\theta\|^2 \right) = \theta \sigma^2 = a \]

\[ \therefore \Phi(\mu) = \langle \mu, \theta \rangle - \Psi(\theta) = \langle \mu, \frac{\mu}{\sigma^2} \rangle - \frac{\sigma^2}{2} \|\theta\|^2 = \frac{\|\mu\|^2}{2\sigma^2} \]

\[ \therefore d_\Phi(x, \mu) = \Phi(x) - \Phi(\mu) - \langle x - \mu, \nabla \Phi(\mu) \rangle = \frac{\|x - \mu\|^2}{2\sigma^2} \]

\[ b_\Phi(x) = \exp(\Phi(x)) p_0(x) = \exp \left( \frac{\|x\|^2}{2\sigma^2} \right) \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp \left( -\frac{\|x\|^2}{2\sigma^2} \right) = \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \]

\[ p(\psi, \theta)(x) = \exp(-d_\Phi(x, \mu)) b_\Phi(x) \]
Definition (Bregman Soft clustering problem)

as that of of learning the maximum likelihood parameters 
\[ \Gamma = \{ \theta_h, \pi_h \}_{h=1}^k \equiv \{ \mu_h, \pi_h \}_{h=1}^k \] of a mixture model of the form

\[ p(x|\Gamma) = \sum_{h=1}^{k} \pi_h p_{(\psi, \theta_h)}(x) = \sum_{h=1}^{k} \pi_h \exp(-d_\Phi(x, \mu_h)) b_\Phi(x) \]

By assuming the mixture components from same family, it can be solved by EM algorithm.
EM example of coin flipping

Do and Batzoglou [2008]
Algorithm 2 EM for Mixture Density Estimation [18]

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

Output: \( \Theta^* \), local maximizer of

\[ L_\mathcal{X}(\Theta) = \prod_{i=1}^n \left( \sum_{h=1}^k \pi_h p_h(x_i|\theta_h) \right) \]

where \( \Theta = \{\theta_h, \pi_h\}_{h=1}^k \), soft partitioning \( \{p(h|x_i)\}_{h=1}^k \).

Method:

Initialize \( \{\theta_h, \pi_h\}_{h=1}^k \) with some \( \theta_h \in S \),
\( \pi_h \geq 0 \), \( \sum_{h=1}^k \pi_h = 1 \)

repeat

\{The Expectation Step\}

for \( i = 1 \) to \( n \) do

for \( h = 1 \) to \( k \) do

\[ p(h|x_i) \leftarrow \frac{\pi_h p_h(x_i|\theta_h)}{\sum_{h'=1}^k \pi_{h'} p_{h'}(x_i|\theta_{h'})} \]

end for

end for

\{The Maximization Step\}

for \( h = 1 \) to \( k \) do

\[ \pi_h \leftarrow \frac{1}{n} \sum_{i=1}^n p(h|x_i) \]

\[ \theta_h \leftarrow \arg \max_{\theta} \sum_{i=1}^n \log(p_h(x_i|\theta))p(h|x_i) \]

end for

until convergence

return \( \Theta^* = \{\theta_h, \pi_h\}_{h=1}^k \)

Algorithm 3 Bregman Soft Clustering

Input: Set \( \mathcal{X} = \{x_i\}_{i=1}^n \subset S \subseteq \mathbb{R}^d \), Bregman divergence \( D_\phi \), num. of clusters \( k \).

Output: \( \Theta^* \), local maximizer of

\[ \prod_{i=1}^n \left( \sum_{h=1}^k \pi_h f_\phi(x_i) \exp(-D_\phi(x_i, \mu_h)) \right) \]

where \( \Theta = \{\mu_h, \pi_h\}_{h=1}^k \), soft partitioning \( \{p(h|x_i)\}_{h=1}^k \).

Method:

Initialize \( \{\mu_h, \pi_h\}_{h=1}^k \) with some \( \mu_h \in S \), \( \pi_h \geq 0 \), and \( \sum_{h=1}^k \pi_h = 1 \)

repeat

\{The Expectation Step\}

for \( i = 1 \) to \( n \) do

for \( h = 1 \) to \( k \) do

\[ p(h|x_i) \leftarrow \frac{\pi_h \exp(-D_\phi(x_i, \mu_h))}{\sum_{h'=1}^k \pi_{h'} \exp(-D_\phi(x_i, \mu_{h'}))} \]

end for

end for

\{The Maximization Step\}

for \( h = 1 \) to \( k \) do

\[ \pi_h \leftarrow \frac{1}{n} \sum_{i=1}^n p(h|x_i) \]

\[ \mu_h \leftarrow \frac{\sum_{i=1}^n p(h|x_i) x_i}{\sum_{i=1}^n p(h|x_i)} \]

end for

until convergence

return \( \Theta^* = \{\mu_h, \pi_h\}_{h=1}^k \)
Geography faculty at the University of North Carolina like to point out that in 1986, those who graduated with a major in Geography had the highest average starting salaries in the class — $250,000. The punchline to this joke is that basketball legend, Michael Jordan, graduated from UNC with a major in Geography in 1986. In that particular dataset, Michael Jordan is clearly an outlier whose astronomical earnings skew the results and obscure the real market for geography majors. (Ref: http://www.forest2market.com/about/methodology/stumpage-price-database)

**Definition (Robustness Check, Liu [2011])**

Let \( \bar{x} \) be the true centroid of set \( X = \{ x_1, \ldots, x_n \} \). When \( \epsilon \% \) (\( \epsilon \) small) of outlier \( y \) is mixed into the set \( X \), then the estimation of the centroid would be influenced by the outliers, and denote the estimation as \( \hat{x} = \bar{x} + \epsilon z(y) \), where the \( z(y) \) is called the influence function. For ordinary Bregman divergence, \( z = y \), thus the breakdown point is 0\%.
Definition (Total Bregman divergence (TBD))

TBD\(\delta\) associated with a real valued strictly convex and differentiable function \(f\) defined on a convex set \(X\) between points \(x, y \in X\) is defined as,

\[
\delta_f (x, y) = \frac{f(x) - f(y) - \langle x - y, \nabla f(y) \rangle}{\sqrt{1 + \|\nabla f(y)\|^2}}
\]
Extensions: Symmetry

Definition (Symmetry Extension, Leonenko et al. [2008])

\[ D_q(f, g) = \int_{\mathbb{R}^m} \left[ g^q(x) + \frac{f^q(x)}{q - 1} - \frac{q}{q - 1} f(x) g^{q-1}(x) \right] dx \]

\[ K_q(f, g) = \frac{1}{q} \left[ D_q(f, g) + D_q(g, f) \right] \]

\[ = \frac{1}{q - 1} \int_{\mathbb{R}^m} \left[ f(x) - g(x) \right] \left[ f^{q-1}(x) - g^{q-1}(x) \right] dx \]


Acknowledgements

Thanks to:
- Prof. Jacob Kogan, Math 710