EM Algorithm and Online Extensions

Ehsan Amid and Manfred Warmuth

University of California, Santa Cruz

eamid@ucsc.edu, manfred@ucsc.edu
Overview

1. Preliminaries
   - Exponential Family and Bregman Divergences

2. Previous Work
   - Expectation Maximization (EM)

3. Our Online EM
For a real-valued convex function \( G : \mathbb{R}^d \to \mathbb{R} \), the Bregman divergence is defined as

\[
\Delta_G(\tilde{\Theta}, \Theta) = G(\tilde{\Theta}) - G(\Theta) - g(\Theta) \cdot (\tilde{\Theta} - \Theta)
\]  

(1)

The gradient with respect to the first argument takes the following simple form

\[
\nabla_{\tilde{\theta}} \Delta_G(\tilde{\Theta}, \Theta) = g(\tilde{\Theta}) - g(\Theta)
\]  

(2)
Exponential Family

Exponential family is defined as the collection of probability distributions

\[ P_G(x|\Theta) = P_\circ(x) \exp(\Theta \cdot \phi(x) - G(\Theta)) \] \hspace{1cm} (3)

The quantity \( \phi(x) = (\phi_\alpha(x))_{\alpha \in \mathcal{I}} \) denotes the set of all sufficient statistics \( \phi_\alpha(x) \) with the associate set of canonical parameters \( \Theta_\alpha \). \( G(\Theta) \) is called the log partition function

\[ G(\Theta) = \log \int P_\circ(x) \exp(\Theta \cdot \phi(x)) dx \] \hspace{1cm} (4)

which ensures that \( P_G(x|\Theta) \) sums up to one (assuming that the integral exists).
There exists a close connection between the exponential family and the Bregman divergence

\[
\int P_G(x|\Theta) \log \frac{P_G(x|\Theta)}{P_G(x|\Theta)} \, dx = \Delta_G(\tilde{\Theta}, \Theta)
\]  

(5)
Let $\mu = (\mu_\alpha)_{\alpha \in \mathcal{I}}$ denote the mean vector associated with the sufficient statistics $\phi_\alpha(x)$

$$\mu = \mathbb{E}_\Theta[\phi(x)] = \int \phi(x) P_G(x|\Theta) \, dx \quad (6)$$

It can be shown that the parameter $\Theta$ satisfies the moment matching relation,

$$\mathbb{E}_\Theta[\phi(x)] = g(\Theta) = \mu \quad (7)$$
Given a set of iid observations $V = \{v_n\}_{n=1}^N$, the negative log-likelihood of a mixture model is defined as

$$L(\Theta; V) = -\sum_{n=1}^N \log \sum_{h_n} P(h_n|\Theta) P(v_n|h_n, \Theta)$$  \hspace{1cm} (8)$$

$h_n$ is the hidden r.v. corresponding to $v_n$, $\Theta$ is the mixture parameters.

The maximum-likelihood parameter estimation

$$\tilde{\Theta} = \arg \min_{\Theta} L(V; \Theta)$$  \hspace{1cm} (9)$$
EM Algorithm

\[ \mathcal{L}(\mathcal{V}|\Theta) = - \sum_n \log P(v_n|\Theta) \]

\[ \leq \sum_n \sum_h \gamma_n^h \log \frac{\gamma_n^h}{P(h_n|v_n, \Theta)} - \sum_n \log P(v_n|\Theta) \quad (10) \]

\[ = \sum_n \sum_h \gamma_n^h \log \frac{\gamma_n^h}{P(h_n, v_n|\Theta)} - \sum_n \log P(v_n|\Theta) \]

\[ = - \sum_n \sum_h \gamma_n^h \log P(h_n, v_n|\Theta) + \sum_h \gamma_n^h \log \gamma_n^h \quad (11) \]
Start with

$$\sum_n \sum_h \gamma_n^h \log \frac{\gamma_n^h}{P(h_n|v_n, \Theta)} - \sum_n \log P(v_n|\Theta)$$  \hspace{1cm} (10)$$

Set

$$\gamma_n^h = \frac{P(h_n|\Theta)P(v_n|h_n, \Theta)}{\sum_{h_n} P(h_n|\Theta)P(v_n|h_n, \Theta)}$$

(10) equals $- \sum_n \log P(v_n|\Theta)$ \textbf{(E step)}.

Now minimize the equivalent expression (11) w.r.t. $\Theta$ \textbf{(M step)}.

$$\tilde{\Theta} = \arg \max_{\Theta} - \sum_n \sum_h \gamma_n^h \log P(h_n, v_n|\Theta) + \sum_h \gamma_n^h \log \gamma_n^h$$  \hspace{1cm} (11)$$

Switch back from (11) to (10) with $\Theta \leftarrow \tilde{\Theta}$ and repeat.
Log-sum inequality on the integral

\[ \Delta(\tilde{\Theta}, \Theta) = \int \sum_h \omega_h P(x|\Theta_h) \log \frac{\sum_h \omega_h P(x|\Theta_h)}{\sum_h \tilde{\omega}_h P(x|\tilde{\Theta}_h)} \, dx \]

\[ \leq \int \sum_h \omega_h P(x|\Theta_h) \log \frac{\omega_h P(x|\Theta_h)}{\tilde{\omega}_h P(x|\tilde{\Theta}_h)} \, dx \]

\[ = \sum_h \omega_h \log \frac{\omega_h}{\tilde{\omega}_h} + \sum_h \omega_h P(x|\Theta_h) \log \frac{P(x|\Theta_h)}{P(x|\tilde{\Theta}_h)} \]

\[ = \sum_h \omega_h \log \frac{\omega_h}{\tilde{\omega}_h} + \sum_h \omega_h \Delta(\tilde{\Theta}_h, \Theta_h) \]

\[ = \hat{\Delta}(\tilde{\Theta}, \Theta) \] (12)
Our Approach

Relative entropy as *regularizer* + negative log-likelihood as *loss*

\[
\mathcal{L}_{\text{our}}(\tilde{\Theta}; \Theta, X) = \sum_{h=1}^{K} \omega_h \log \frac{\omega_h}{\tilde{\omega}_h} + \sum_{h=1}^{K} \omega_h \Delta(\tilde{\Theta}_h, \Theta_h) - \eta \sum_{n=1}^{N} \sum_{h=1}^{K} \gamma_n^h \log \tilde{\omega}_h P(x_n|\tilde{\Theta}_h) + \eta \sum_{n=1}^{N} \sum_{h=1}^{K} \gamma_n^h \log \gamma_n^h
\]

(13)

*Reds* are current (old) parameters.

Fix *greens* as before (*E step*).

Update *blues* (*M step*).

Replace *reds* with *blues* and repeat.
Again, taking the derivatives

\[
\frac{\partial}{\partial \tilde{\Theta}_h} : \quad \omega_h \left( g(\tilde{\Theta}_h) - g(\Theta_h) \right) - \eta \sum_n \gamma_h^n \left( \phi(x_n) - g(\tilde{\Theta}_h) \right) = 0 \quad (14)
\]

\[
\frac{\partial}{\partial \tilde{\omega}_h} : \quad - \frac{\omega_h}{\tilde{\omega}_h} - \eta \sum_n \gamma_h^n + \lambda = 0 \quad (15)
\]
Our Approach: Updates

Solving (14) and (15) yields the following updates

\[
\tilde{\omega}_h = \frac{\omega_h + \eta \sum_n \gamma_n}{1 + \eta N} \tag{16}
\]

\[
\tilde{\mu}_h = \frac{\mu_h + \eta \sum_n \beta_n \phi(x_n)}{1 + \eta \sum_n \beta_n} \tag{17}
\]

where we use the notation \( \beta_n = \gamma_n / \omega_h \).

Note that updates are additive this time, thus, more stable (compared to joint-entropy algorithm [SW99] and online EM [CE07]).
References
