Theoretical Basis for EM

Sargur Srihari srihari@cedar.buffalo.edu

EM Algorithm

- Given joint distribution $p(X,Z \mid \theta)$ over observed variables X and latent variables Z governed by parameters θ goal is to find θ that maximizes likelihood function $p(X \mid \theta)$
- Step 1: Choose an initial setting for the parameters $\, heta^{o}$
- Step 2: E Step: Evaluate $p(Z \mid X, \theta^t)$ Can be done since joint $p(X,Z|\theta)$ is known
- Step 3: M Step: Evaluate θ^{t+1} given by

$$\left| heta^{t+1} = rg \max_{ heta} Q(heta, heta^t)
ight|$$

where

$$Q(\theta, \theta^t) = \sum_{Z} p(Z \mid X, \theta^t) \ln p(X, Z \mid \theta)$$

Summation due to expectation

- Check for convergence
 - of either log-likelihood or parameter values
- If not satisfied then let $\theta^t \leftarrow \theta^{t+1}$
- Return to Step 2

General EM Algorithm defined heuristically

- EM is a general technique
 - Find maximum likelihood solutions (parameters) for probabilistic models with latent variables
- EM was defined heuristically
 - It can be proved to maximize the likelihood function
 - Monotonically increases observed data log-likelihood until it reaches a local maximum
 - Proof involves obtaining lower bound on log- likelihood function
 - Dempster 1977

Likelihood Function with Latent Variables

• Given observed variables $X=\{x_1,...x_N\}$ and hidden variables $Z=\{z_1,...z_N\}$, likelihood function is

$$p(X \mid \theta) = \sum_{Z} p(X, Z \mid \theta) = \sum_{\mathbf{z}_i} \prod_{i=1}^{N} p\left(\mathbf{x}_i, \mathbf{z}_i \mid \theta\right) = \prod_{i=1}^{N} \sum_{\mathbf{z}_i} p\left(\mathbf{x}_i, \mathbf{z}_i \mid \theta\right)$$

Log-likelihood is obtained by taking log of above:

$$\ell \left(\boldsymbol{\theta} \right) = \ln p(\boldsymbol{X} \mid \boldsymbol{\theta}) = \sum_{i=1}^{N} \ln \left[\sum_{\mathbf{Z}_i} p \left(\mathbf{x}_i, \mathbf{z}_i \mid \boldsymbol{\theta} \right) \right]$$

- Goal is to maximize log-likelihood $\ell(\theta)$
 - Direct optimization of $\ell(\theta)$, i.e., $p(X|\theta)$ is difficult, whereas complete-data likelihood $p(X,Z|\theta)$ is easier (since component forms are tractable)
 - **Example of GMM:** $p(\mathbf{x}, \mathbf{z} \mid \theta) = p(\mathbf{x} \mid \mathbf{z}, \theta) p(\mathbf{z} \mid \theta) \text{ and } p(\mathbf{x} \mid \mathbf{z}, \theta) = \prod_{k=1}^{K} N(\mathbf{x} \mid \mu_{k}, \Sigma_{k})^{z_{k}}, \ p(\mathbf{z} \mid \theta) = \prod_{k=1}^{K} \pi_{k}^{z_{k}}$

Decomposition of $p(X|\theta)$

• For any distribution q(Z), this decomposition holds:

$$\ln p(X \mid \theta) = L(q, \theta) + KL(q \parallel p)$$
where we define

$$L(q,\theta) = \sum_{z} q(z) \ln \left\{ \frac{p(X,Z \mid \theta)}{q(Z)} \right\}$$

$$KL(q \parallel p) = -\sum_{z} q(z) \ln \left\{ \frac{p(z \mid X, \theta)}{q(z)} \right\} = \ln \frac{p(x, z \mid \theta) - \ln q(z) - \ln p(z \mid X, \theta) - \ln q(z)}{q(z)} = \ln \frac{p(z \mid X, \theta)}{q(z)} + \ln p(x \mid \theta)$$

L is a *functional* that takes a function as input and produces a value as output, like entropy It contains joint distribution of *X* and *Z*

KL is the Kullback-Leibler Divergence Contains conditional distribution of *Z* given *X*

Algebraic Derivation of expression:

$$\begin{aligned} p(X,Z\mid\theta) &= p(Z\mid X,\theta)p(X\mid\theta) \text{ from product rule} \\ \ln p(X,Z\mid\theta) &= \ln p(Z\mid X,\theta) + \ln p(X\mid\theta) \\ \ln p(X,Z\mid\theta) &= \ln p(Z\mid X,\theta) - \ln q(Z) + \ln p(X\mid\theta) \\ \ln \frac{p(X,Z\mid\theta)}{q(Z)} &= \ln \frac{p(Z\mid X,\theta)}{q(Z)} + \ln p(X\mid\theta) \\ q(Z) \ln \frac{p(X,Z\mid\theta)}{q(Z)} &= q(Z) \ln \frac{p(Z\mid X,\theta)}{q(Z)} + q(Z) \ln p(X\mid\theta) \\ \sum_{Z} q(Z) \ln \frac{p(X,Z\mid\theta)}{q(Z)} &= \sum_{Z} q(Z) \ln \frac{p(Z\mid X,\theta)}{q(Z)} + \sum_{Z} q(Z) \ln p(X\mid\theta) \\ \sum_{Z} q(Z) \ln \frac{p(X,Z\mid\theta)}{q(Z)} &= \sum_{Z} q(Z) \ln \frac{p(Z\mid X,\theta)}{q(Z)} + \ln p(X\mid\theta) \\ \ln p(X\mid\theta) &= \sum_{Z} q(Z) \ln \frac{p(X,Z\mid\theta)}{q(Z)} - \sum_{Z} q(Z) \ln \frac{p(Z\mid X,\theta)}{q(Z)} \end{aligned}$$

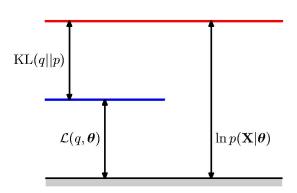
Lower bound on log-likelihood

$$\ell(\theta) = \ln p(X \mid \theta) = L(q, \theta) + KL(q \mid\mid p)$$

L is the lower bound

- Since $KL(q||p) \ge 0$ It follows that $\ln p(X|\theta) \ge L(q,\theta)$
 - Or $L(q, \theta)$ is a lower bound on $\ln p(X|\theta)$

$$\ell(\theta) \ge L(q, \theta)$$



E-step maximizes Lower Bound

• For any distribution $q(z_i)$ the following holds

$$\boxed{\ell\left(\theta\right) = \ln p(X \mid \theta) = \sum_{i=1}^{N} \ln \left[\sum_{\mathbf{Z}_{i}} p\left(\mathbf{x}_{i}, \mathbf{z}_{i} \mid \theta\right) \right] = \sum_{i=1}^{N} \ln \left[\sum_{\mathbf{Z}_{i}} q\left(\mathbf{z}_{i}\right) \frac{p\left(\mathbf{x}_{i}, \mathbf{z}_{i} \mid \theta\right)}{q\left(\mathbf{z}_{i}\right)} \right]} \qquad \text{Multiplying numerator and denominator by } q(\mathbf{z}_{i})$$

- Using Jensen's inequality $\left| f\left(\sum_{i=1}^{N} \lambda_i \mathbf{x}_i\right) \le \sum_{i=1}^{N} \lambda_i f\left(\mathbf{x}_i\right) \right|$ we get a lower bound

$$\ell\left(\boldsymbol{\theta}\right) \geq \sum_{i=1}^{N} \sum_{\mathbf{Z}_{i}} q\left(\mathbf{z}_{i}\right) \frac{p\left(\mathbf{x}_{i}, \mathbf{z}_{i} \mid \boldsymbol{\theta}\right)}{q\left(\mathbf{z}_{i}\right)}$$

$$- \ \, \text{The lower bound is} \quad \boxed{Q\!\left(\theta,q\right) = \sum_{i} E_{q_i}\!\left[\ln p\!\left(x_i,z_i\mid\theta\right)\right] + H\!\left(q_i\right)} \quad \text{H is the entropy of q_i}$$

- What form of q yields the tightest lower bound?
 - Lower bound is also sum of i terms of the form

$$L\left(\boldsymbol{\theta}, \boldsymbol{q}_{i}\right) = \sum_{\mathbf{z}_{i}} \boldsymbol{q}_{i}\left(\mathbf{z}_{i}\right) \ln \frac{\boldsymbol{p}\left(\mathbf{x}_{i}, \mathbf{z}_{i} \mid \boldsymbol{\theta}\right)}{\boldsymbol{q}_{i}\left(\mathbf{z}_{i}\right)} = -KL\left(\boldsymbol{q}_{i}\left(\mathbf{z}_{i}\right) \mid\mid \boldsymbol{p}\left(\mathbf{z}_{i} \mid \mathbf{x}_{i}, \boldsymbol{\theta}\right)\right) + \ln \boldsymbol{p}\left(\mathbf{x}_{i} \mid \boldsymbol{\theta}\right)$$

- We can maximize the lower bound by setting $q_i(z_i) = p(z_i|x_i,\theta_i)$
- Since θ is unknown, instead we use $q_i^t(\mathbf{z}_i) = p(\mathbf{z}_i \mid \mathbf{x}_i, \theta^t)$
 - Where θ^t is our estimate of parameters at iteration t
- This is the output of the E step, i.e., we determine $q_i^t(\mathbf{z}_i)$

M step computes $\ell(\theta^t)$

After the E step the lower bound is

$$Q\!\left(\theta,q^{t}\right) = \sum_{i} E_{q_{i}^{t}}\!\left[\ln p\!\left(x_{i},z_{i}\mid\theta\right)\right] + H\!\left(q_{i}^{t}\right)$$

- Second term is constant wrt θ
- So the M step becomes

$$\theta^{t+1} = \arg\max_{\theta} Q\Big(\theta, q^t\Big) = \arg\max_{\theta} \sum_{i} E_{q_i^t} \Big[\ln p\Big(\mathbf{x}_i, \mathbf{z}_i \mid \theta\Big)\Big]$$

• Since we choose $q_i(\mathbf{z}_i) = p(\mathbf{z}_i | \mathbf{x}_i, \theta_i)$ KL divergence is zero $L(\theta^t, q_i) = \ln p(\mathbf{x}_i | \theta^t)$ and hence lower bound reaches $\ell(\theta^t)$

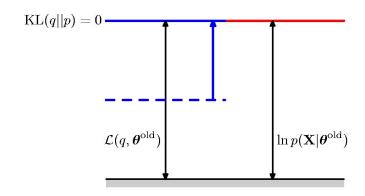
$$Q\!\left(\boldsymbol{\theta}^{t}, \boldsymbol{q}^{t}\right) = \sum_{i} \ln p\!\left(\mathbf{x}_{i} \mid \boldsymbol{\theta}^{t}\right) = \ell\!\left(\boldsymbol{\theta}^{t}\right)$$

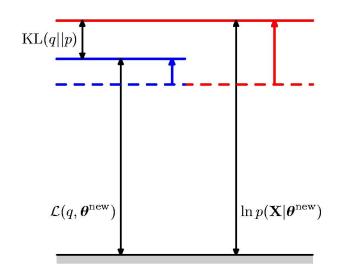
Illustration of E and M steps

- E step (Lower bound maximized keeping θ^{old} fixed)
 - θ distribution is set to posterior distribution for current parameter values θ^{old}
 - Causing lower bound to move to same value as log-likelihood with KL vanishing

M step

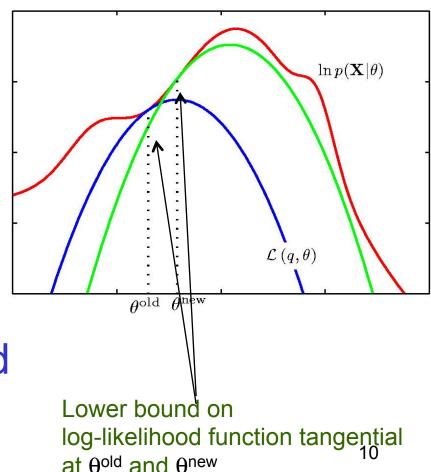
- Lower bound is maximized wrt qto give revised value θ^{new}
- Causes log-likelihood to increase by as much as lower bound





View of EM in parameter space

- EM involves alternately computing lower bound on log-likelihood for the current parameter values
- And maximizing this bound to obtain new parameter values
- Note that the lower bound is a convex function with a unique maximum



Generalized EM (GEM)

- EM breaks own potentially difficult problem of maximizing the likelihood function into two stages, the E step and the M step
- One or both may remain intractable
- GEM addresses the problem of the intractable M step
 - Instead of maximizing $L(q,\theta)$ wrt θ it changes parameters so as to increase its value
- Similar generalization of the E step can be made