The EM Algorithm in General

The EM Algorithm is a general algorithm for finding maximum likelihood solutions for probabilistic models having latent variables.

- Proof that heuristic algorithm does maximize likelihood function
- Basis for variational inference

Consider prob model with

$\underline{X}$ observed variables
Z hidden variables
θ parameters

joint distribution $P(X, Z \mid \theta)$

Goal: Maximize likelihood function

$$P(X \mid \theta) = \sum_Z p(X, Z \mid \theta)$$

Z can be missing data, marginalized params, etc.

Assumption 1) Z discrete, all args hold for continuous var's
2) Direct optimization of $p(X \mid \theta)$ w.r.t. θ hard

usually true because of sum

3) Optimization of complete likelihood function easier
$P(X, Z \mid \theta)$

no summation, log passes through

$$= \sum_z q(z) \ln \frac{p(z|x,\theta)\, p(x|\theta)}{q(z)} - \underbrace{\sum_z q(z) \ln \frac{p(z|x,\theta)}{q(z)}}$$

$$= \cancel{\sum_z q(z) \ln \frac{p(z|x,\theta)}{q(z)}} + \sum_z q(z) \ln \frac{p(x|\theta)}{\cancel{q(z)}} - \cancel{\sum_z q(z) \ln q(z)}$$

$$- \left[\cancel{\sum_z q(z) \ln p(z|x,\theta)} - \cancel{\sum_z q(z) \ln q(z)} \right]$$

$$= \sum_z q(z) \ln p(x|\theta)$$

$$= \ln p(x|\theta) \cdot 1 \quad \blacksquare$$

$KL(q \| p) = 0$ iff $\frac{q(x)}{p(x)} = 1 \;\; \forall x$

$\log(1) = 0 \ldots$

General EM

Introduce $q(Z)$, function over latent vars.

Note the following decomposition holds for all q

$$\ln p(X|\theta) = \mathcal{L}(q,\theta) + KL(q\|p)$$

where

9.71 $$\mathcal{L}(q,\theta) = \sum_Z q(Z) \ln \left\{ \frac{p(X,Z|\theta)}{q(Z)} \right\}$$

and

9.72 $$KL(q\|p) = -\sum_Z q(Z) \ln \left\{ \frac{p(Z|X,\theta)}{q(Z)} \right\}$$

Note

a) signs not equal
b) $\mathcal{L}(q,\theta)$ contains complete-data likelihood
c) $KL(q\|p)$ is the KL divergence between $q(Z)$ and $p(Z|X,\theta)$ (and contains $p(Z|X,\theta)$.
   - Recall $KL(q\|p) \geq 0$
     $KL(q\|p) \neq KL(p\|q)$ in general
     and $KL(q\|p) = 0$ if $q = p$

Because $KL(q\|p) \geq 0$

$$\mathcal{L}(q,\theta) \leq \ln p(X|\theta)$$

i.e. $\mathcal{L}(q,\theta)$ is a lower bound on $\ln p(X|\theta)$.

EM is two stage procedure
Suppose current param vector is θ^{old}.

E step: maximize $\mathcal{L}(q, \theta^{old})$ w.r.t. $q(Z)$
- i.e. find $q(Z)$ that maxes $\mathcal{L}(q, \theta^{old})$, this will be by making KL as small as possible
- ideal, set $q(Z) = p(Z|X, \theta^{old})$, KL vanishes

M step: $q(Z)$ held fixed and $\mathcal{L}(q, \theta)$ is maxed w.r.t. to θ yielding θ^{new}

- this causes the lower bound to increase and thereby $\ln p(X|\theta)$ to increase as well.

- since q is learned using θ^{old}, in general $p(Z|X, \theta^{new})$ will be different from q and KL div will be non-zero

Substituting $q(Z) = p(Z|X, \theta^{old})$ into

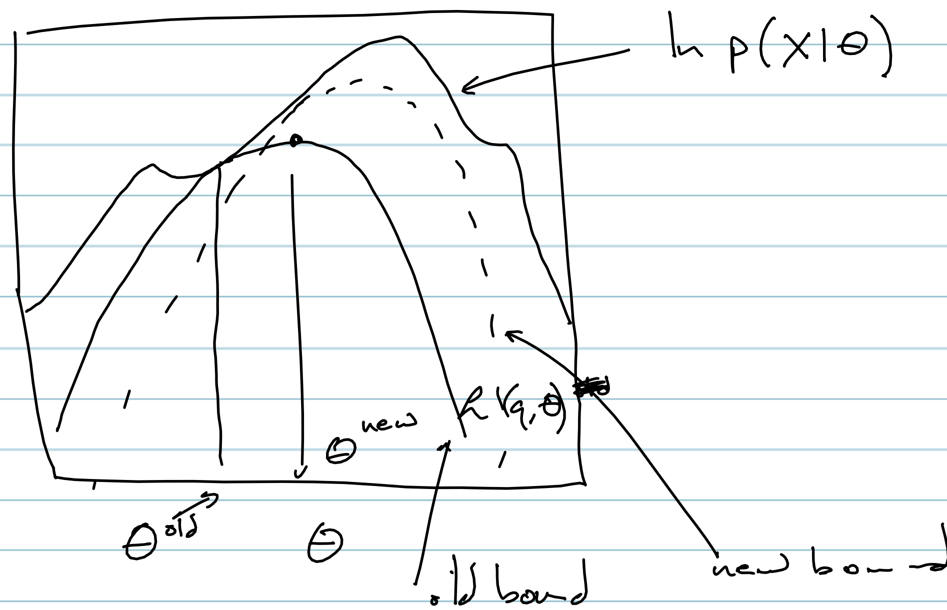
$$\mathcal{L}(q, \theta) = \sum_Z q(Z) \ln \left\{ \frac{p(X, Z|\theta)}{q(Z)} \right\}$$

we have

$$\mathcal{L}(q, \theta) = \underbrace{\sum_Z p(Z|X, \theta^{old}) \ln p(X, Z|\theta)}_{Q(\theta, \theta^{old})} + \text{const.}$$

where $Q(\theta, \theta^{old})$ involves the log of the complete-data likelihood, a quantity assumed to be easy to work with.

Graphically



Note that in the case of i.i.d. data, i.e.

$$p(X,Z) = \prod_n p(x_n, z_n)$$

the posterior over Z has a nice form

$$p(Z|X,\theta) = \frac{p(X,Z|\theta)}{\sum_Z p(X,Z|\theta)}$$

$$= \frac{\prod_{n=1}^{N} p(x_n, z_n|\theta)}{\sum_Z \prod_{n=1}^{N} p(x_n, z_n|\theta)}$$

$$= \prod_{n=1}^{N} p(z_n|x_n,\theta)$$

$\therefore$ So in E step, each data point's posterior "responsibility" can be computed independently from the others.

Note:

Any moves that increase $\mathcal{L}(q, \theta)$ ~~to~~ will increase $\ln p(X|\theta)$.

Possibilities include:

- sampling z's
- numerical gradient ascent on θ's
- one data point at a time

etc.