

EM: An alternative view  
and a general view

# An alternative view

- to find M.L. solutions for models having latent variables
- key role played by the latent variables
- $X$ : set of observed data  $:= \{x_n^T\}$
- $Z$ : set of latent data  $:= \{z_n^T\}$
- $\theta$ : set of all parameters
  - Gaussian:  $(\mu, \Sigma)$
  - Poisson:  $(\lambda)$
  - ...

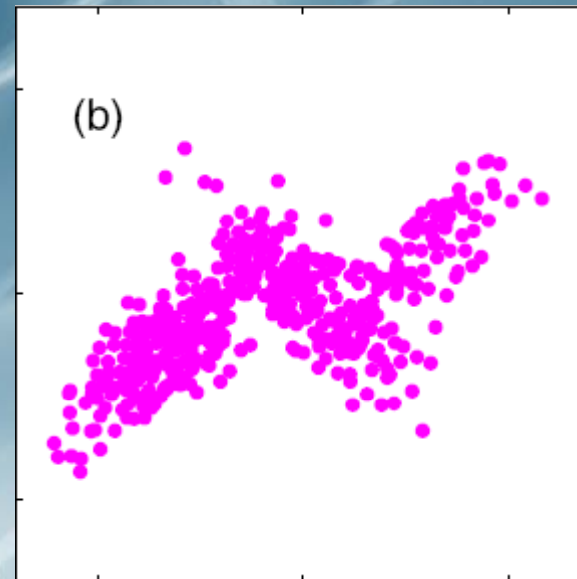
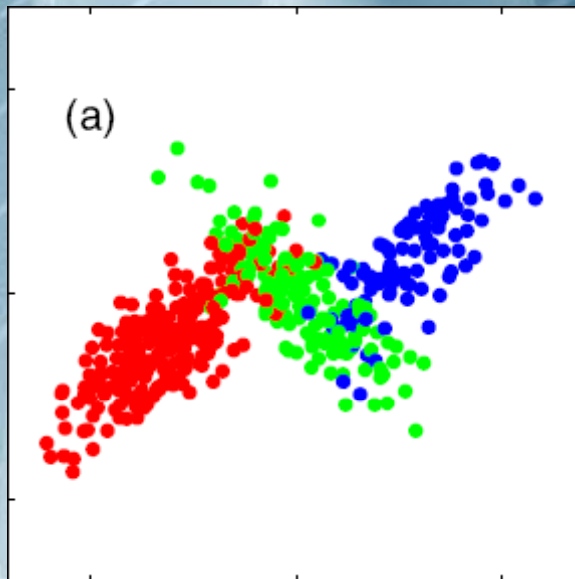
# An alternative view (cont'd)

- log likelihood function:
- $\{X, Z\}$  : complete-data
  - never given
  - instead we know values of  $Z$  only by posterior distribution  $p(Z|X, \theta^{\text{old}})$
- goal: maximize the expectation for this likelihood function

$$\ln p(\mathbf{X}|\theta) = \ln \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta) \right\}$$

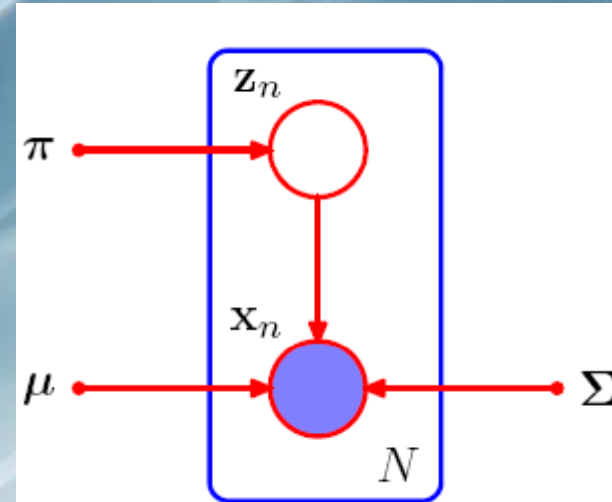
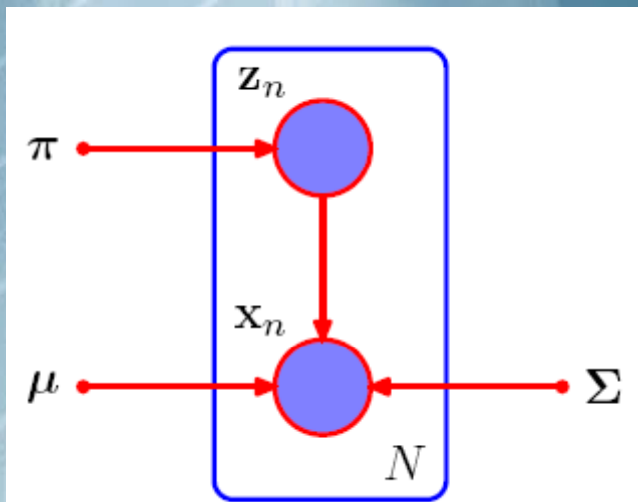
# complete-data vs. incomplete-data

- sample point representation



# complete-data vs. incomplete-data

- graphical representation



# An alternative view (cont'd)

- E step:
  - posterior =  $p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$
  - expectation of the complete-data log likelihood evaluated for some general parameter value  $\theta$

$$Q(\theta, \theta^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\theta).$$

# An alternative view (cont'd)

- M step:
  - determine the revised parameter by maximizing the above expectation function

$$\theta^{\text{new}} = \arg \max_{\theta} Q(\theta, \theta^{\text{old}})$$

# General EM algorithm

- Choose an initial setting for the parameters:  $\theta^{\text{old}}$
- E step: evaluate  $p(Z|X, \theta^{\text{old}})$
- M step: evaluate  $\theta^{\text{new}}$
- Check for convergence; if not satisfied:  
$$\theta^{\text{old}} \leftarrow \theta^{\text{new}}$$

# General EM algorithm (cont'd)

- Each EM cycle increases the incomplete-data log likelihood, unless @ a local minimum
- EM for finding maximum posterior solutions for models with  $p(\theta)$  over parameters:
  - new expectation function using Lagrange method:

$$Q(\theta, \theta^{\text{old}}) + \ln p(\theta).$$

# EM for GMM

- Again the comparison between complete and incomplete data
- Recall:

$$p(\mathbf{z}) = \prod_{k=1}^K \pi_k^{z_k}$$

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^K \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$$

# EM for GMM (cont'd)

- incomplete-data:

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

- complete-data:

$$\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^N \sum_{k=1}^K z_{nk} \{ \ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \}$$

- 
- log acts directly on GD
  - much simpler solution

# EM for GMM (cont'd)

- Maximization with respect to  $\mu$  and  $\Sigma$ :
  - Is exactly as for a single G.D., except that it involves only the subset of data points that are assigned to that component  $z_n$
  - All  $z_{nk}$  are 0 except for one  $k$
  - Log likelihood function is simply summation on  $k=1..K$  of  $K$  independent contributions for any of mixture components

# EM for GMM (cont'd)

- Maximization with respect to the mixing coefficients  $\Pi$ :

- Recall:

$$\sum_{k=1}^K \pi_k = 1$$

- A Lagrange multiplier is added to:

$$p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{n=1}^N \prod_{k=1}^K \pi_k^{z_{nk}} \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}$$

- The result:

$$\pi_k = \frac{1}{N} \sum_{n=1}^N z_{nk}$$

# EM for GMM (cont'd)

- Maximization with respect to the mixing coefficients  $\Pi$ :
- The result:
  - “Mixing coefficients are equal to the fractions of data points assigned to the corresponding components”

$$\pi_k = \frac{1}{N} \sum_{n=1}^N z_{nk}$$

# EM for GMM (cont'd)

- Back to incomplete-data, we consider the expectation with respect to the posterior distribution of the latent variables:

$$p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) \propto \prod_{n=1}^N \prod_{k=1}^K [\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]^{z_{nk}}$$

# EM for GMM (cont'd)

- Factorizing over  $n$  so that under the posterior distribution the  $\{z_n\}$  are independent

$$\begin{aligned}\mathbb{E}[z_{nk}] &= \frac{\sum_{z_{nk}} z_{nk} [\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]^{z_{nk}}}{\sum_{z_{nj}} [\pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)]^{z_{nj}}} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} = \gamma(z_{nk})\end{aligned}$$


responsibility of  
component  $k$   
for data point  $n$

# EM for GMM (cont'd)

- For complete-data log likelihood function:

$$\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] = \sum_{n=1}^N \sum_{k=1}^K \gamma(z_{nk}) \{ \ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \}$$

# K-means vs. EM

- K-means: hard assignment of data points to clusters, each point is associated uniquely with one cluster
  - EM: soft assignment based on the posterior probability
- 

K-means as a particular  
limit

of EM for GM

# K-means vs. EM (cont'd)

- Consider a GMM with the covariance matrix as  $\epsilon I$ , where  $\epsilon$  is a variance parameter that is shared by all of the components

$$p(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi\epsilon)^{1/2}} \exp \left\{ -\frac{1}{2\epsilon} \|\mathbf{x} - \boldsymbol{\mu}_k\|^2 \right\}$$

- We treat our EM with K such Gaussians with a constant  $\epsilon$ , with no need to be re-estimated

## K-means vs. EM (cont'd)

- So we have:

$$\gamma(z_{nk}) = \frac{\pi_k \exp \left\{ -\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2 / 2\epsilon \right\}}{\sum_j \pi_j \exp \left\{ -\|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 / 2\epsilon \right\}}$$

- Considering  $\epsilon \rightarrow 0$ , the term for which  $\|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2$  is smallest will go to zero most slowly  
Responsibilities will go to zero except for that specific  $j$  which goes to unity

## K-means vs. EM (cont'd)

- So we have:

$$\gamma(z_{nk}) = \frac{\pi_k \exp \left\{ -\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2 / 2\epsilon \right\}}{\sum_j \pi_j \exp \left\{ -\|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 / 2\epsilon \right\}}$$

- Considering  $\epsilon \rightarrow 0$ , the term for which  $\|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2$  is smallest will go to zero most slowly  
Responsibilities will go to zero except for that specific  $j$  which goes to unity

# K-means vs. EM (cont'd)

- We have a hard assignment of data points to clusters, with responsibilities turned into:

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg \min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}_n$$

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk} \mathbf{x}_n}{\sum_n r_{nk}}$$

$$\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] = \sum_{n=1}^N \sum_{k=1}^K \gamma(z_{nk}) \{ \ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \}$$

$$\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] \rightarrow -\frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2 + \text{const}$$

# EM in General

- Recall: EM to find M.L. solutions for models having latent variables
- a general treatment
- a proof that EM does indeed maximize the likelihood function

$$p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

direct optimization is difficult

- optimization of the complete-data is easier

# EM in General (cont'd)

- Recall: EM for finding M.L. solutions for prob. models having latent variables

$$p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

Difficult

Much  
Easier

# EM in General (cont'd)

- Introducing  $q(\mathbf{Z})$  over latent variables, for any  $q$ :

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

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

$$\text{KL}(q||p) = - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right\}$$

# EM in General (cont'd)

- Using the product rule of prob.:

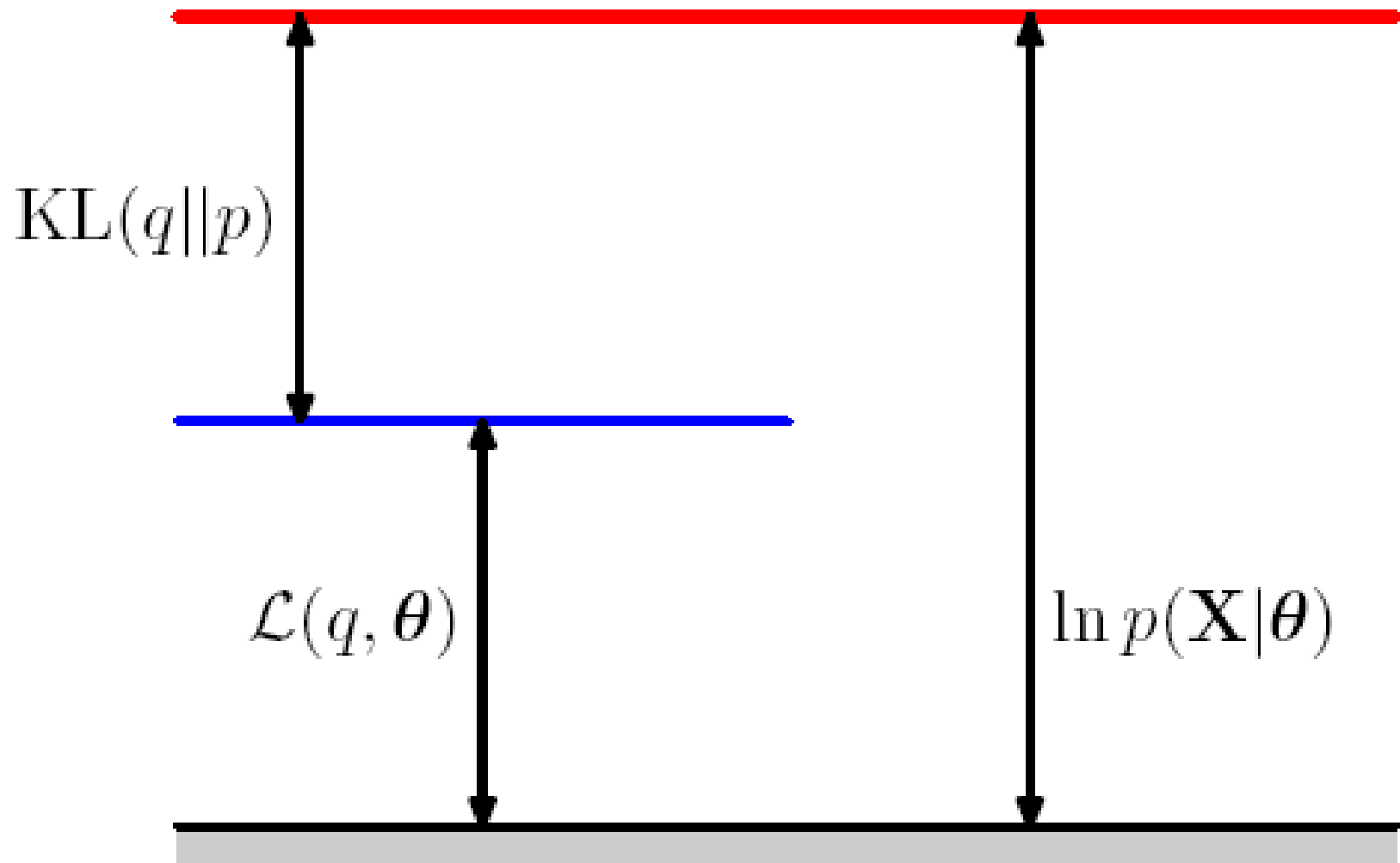
$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \ln p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) + \ln p(\mathbf{X}|\boldsymbol{\theta})$$

- , the term appear to be similar to above

- So :  $\mathcal{L}(q, \boldsymbol{\theta}) \leq \ln p(\mathbf{X}|\boldsymbol{\theta})$

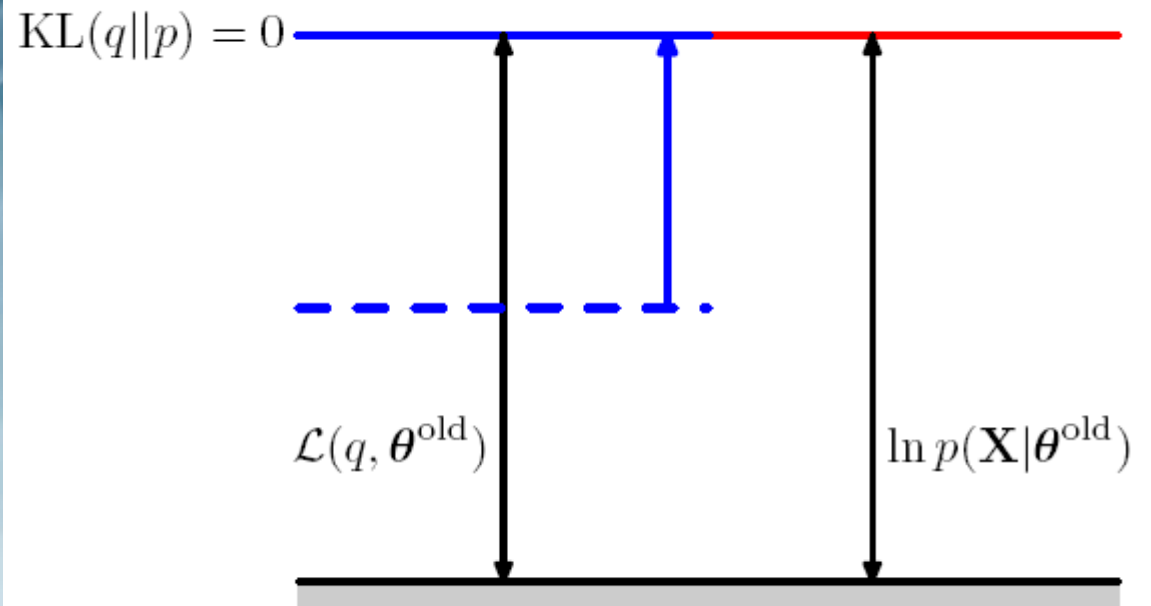
- while :  $\text{KL}(q||p) \geq 0$

# EM in General (cont'd)



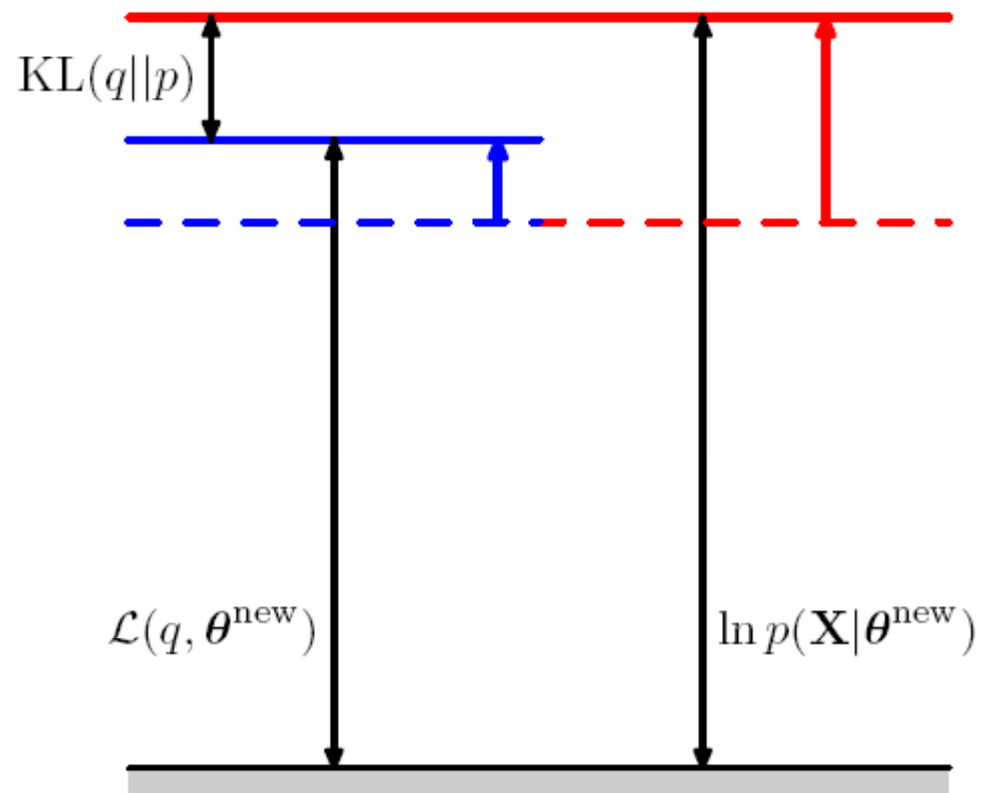
# EM in General (cont'd)

- Back to two steps of EM:



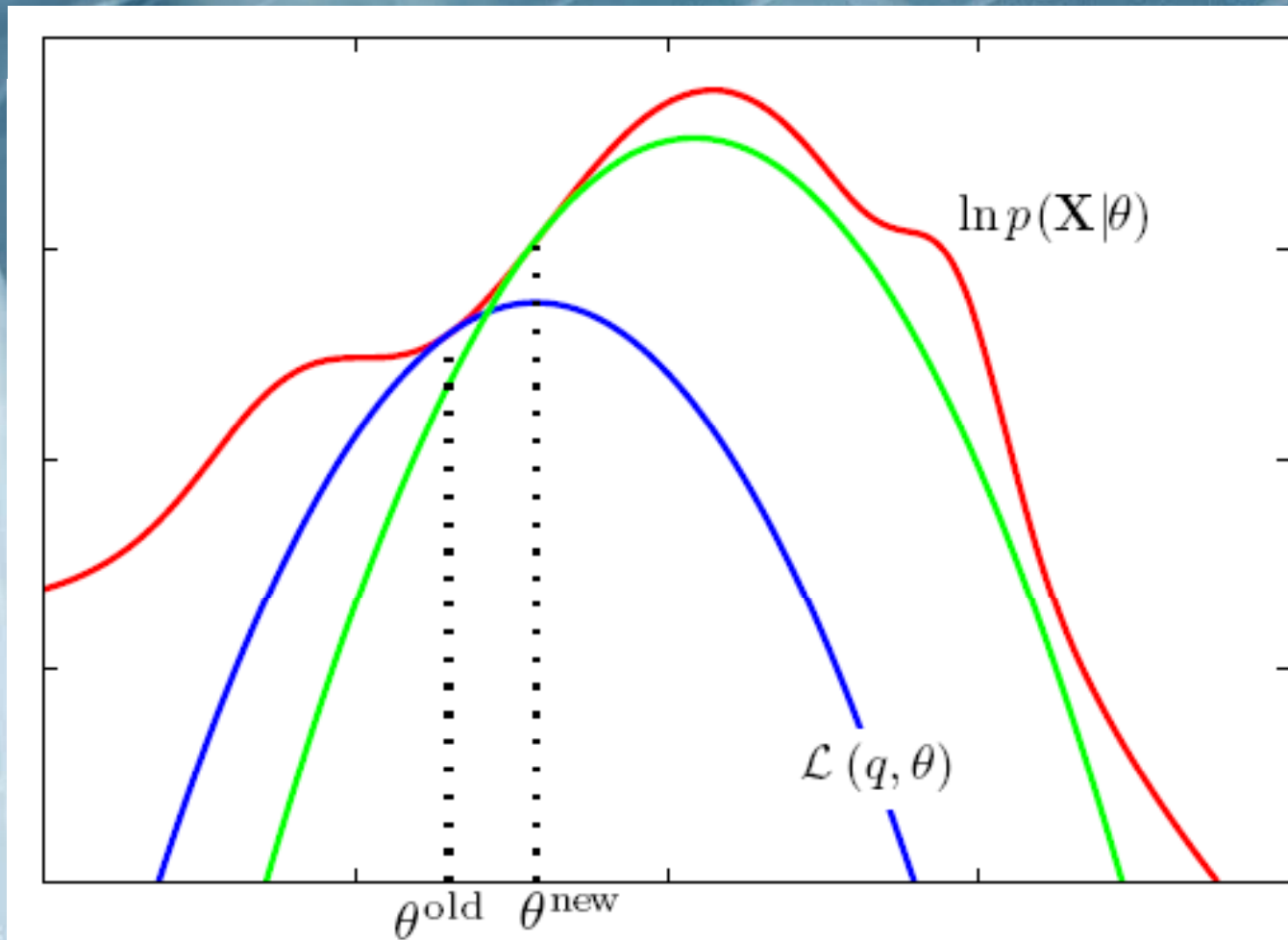
# EM in General (cont'd)

- Back to two steps of EM:



# EM in General (cont'd)

- The lower bound for log likelihood



incomplete-data  
1<sup>st</sup> E step  
2<sup>nd</sup> E step

# Intractable E/M Steps

- GEM
  - conjugate gradient algorithm for M step
- ECM
  - partitioned parameters, multiple M steps
- Data points' independence

$$\mu_k^{\text{new}} = \mu_k^{\text{old}} + \left( \frac{\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N_k^{\text{new}}} \right) (\mathbf{x}_m - \mu_k^{\text{old}})$$

$$N_k^{\text{new}} = N_k^{\text{old}} + \gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})$$