Instructor: Sael Lee

CS549 Spring – Computational Biology

LECTURE 7: MIXTURE MODELS

Reference:

0. T. Mensink and J. Verbeek's 2007 slides on Mixture Models and EM
1. "Pattern Recognition and Machine Learning" Chapter 9: Mixture Models and EM
2. Estimating Gaussian Mixture Densities with EM – A Tutorial
3. A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and HMMs.

- K-means clustering
  - Getting the idea with a simple example
- Mixtures of Gaussians
  - Gradient fixed-points & responsibilities
- An alternative view of EM
  - Completing the data with latent variables
- The EM algorithm in general
  - Understanding EM as coordinate ascent

MIXTURE MODELS AND EM: INTRODUCTION

- Additional **latent variables** allows to express relatively complex marginal distributions over latent variables in terms of more tractable joint distributions over the expanded space.
- Maximum-Likelihood estimator in such a space is the **Expectation-Maximization (EM)** algorithm.

K-MEANS CLUSTERING: DISTORTION MEASURE

- Dataset {x1, . . . , xN}
- **Partition in K clusters**
- Cluster prototype: μk
- Binary indicator variable, 1-of-K Coding scheme

$$r_{nk} \in \{0, 1\}$$

$r_{nk} = 1$, and $r_{nj} = 0$ for $j \neq k$. Only one is 1 and all other 0

- **Hard assignment.**
- **Distortion measure:** a measure of how much data point deviate from the center of their clusters

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

K-MEANS CLUSTERING: EXPECTATION MAXIMIZATION

- Goal: Find values for $\{r_{nk}\}$ and $\{\mu_k\}$ to minimize:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \mu_k\|^2$$

- Iterative procedure:

1. Minimize J w.r.t. r_{nk}, keep μ_k fixed (Expectation)

Calculate the membership

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

2. Minimize J w.r.t. μ_k, keep r_{nk} fixed (Maximization)

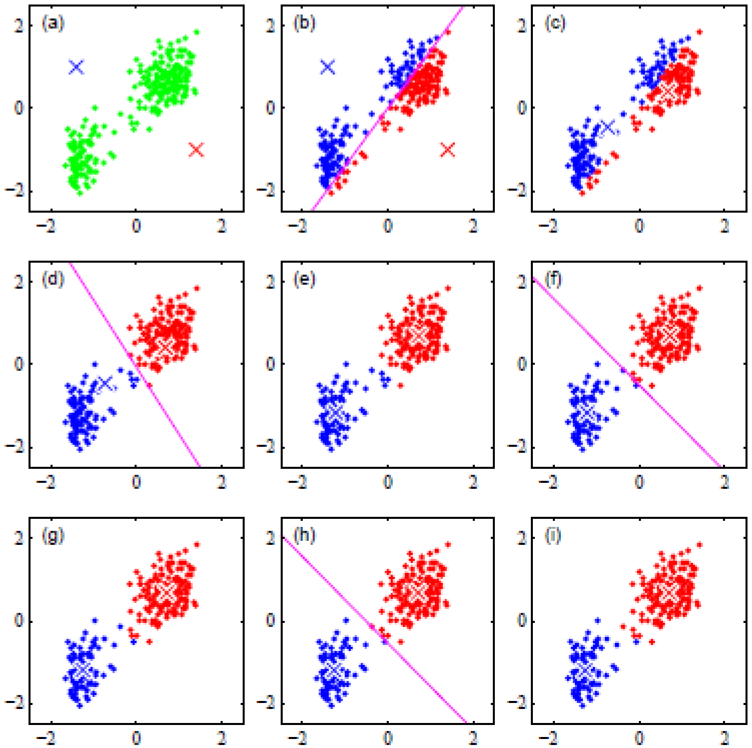
Calculate the center

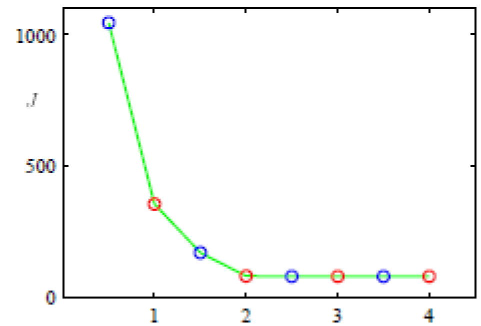
$$2 \sum_{n=1}^{N} r_{nk}(x_n - \mu_k) = 0$$

$$\mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}}$$

K-MEANS CLUSTERING: EXAMPLE

- Each E or M step reduces the value of the objective function J
- Convergence to a **local** maximum





K-MEANS CLUSTERING: CONCLUDING REMARKS

- Direct implementation of K-Means can be slow
- **Online version:**

$$\mu_k^{\text{new}} = \mu_k^{\text{old}} + \eta_n(\mathbf{x}_n - \mu_k^{\text{old}})$$

- **K-mediods**, general distortion measure

$$\tilde{J} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \mathcal{V}(\mathbf{x}_n, \mu_k)$$

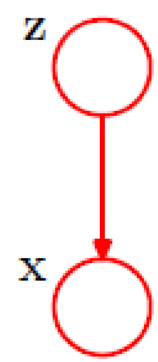
Any type of dissimilarity measure
* K-means uses Euclidean measure which is limited

MIXTURE OF GAUSSIANS: LATENT VARIABLES

- Gaussian Mixture Distribution:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

- Introduce latent variable z
  - z is binary 1-of-K coding variable
  - p(x, z) = p(z)p(x|z)



MIXTURE OF GAUSSIANS: LATENT VARIABLES (2)

The use of the joint probability p(x, z), leads to significant simplifications

- Prior probability of components

$$p(z_k = 1) = \pi_k$$
constraints: $0 \leq \pi_k \leq 1$, and $\sum_k \pi_k = 1$
$$p(\mathbf{z}) = \prod_k \pi_k^{z_k}$$

- Gaussian function of each K mixing components

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$$
$$p(\mathbf{x}|\mathbf{z}) = \prod_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)^{z_k}$$

- Redistribution of Gaussian mixture model

- $$p(\mathbf{x}) = \sum_z p(\mathbf{x}, \mathbf{z}) = \sum_z p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_k \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

MIXTURE OF GAUSSIANS: LATENT VARIABLES (3)

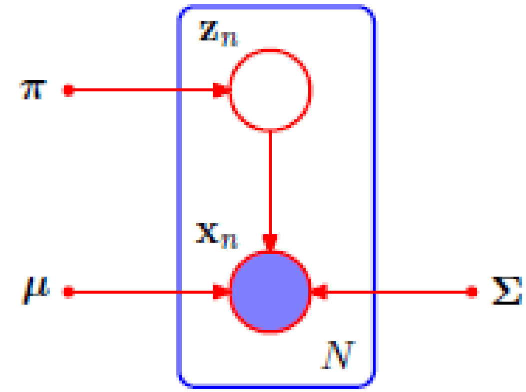
- Responsibility that component k takes for "explaining" observation x:
  - the posterior probability once we observed X.

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_k p(z_k = 1)p(\mathbf{x}|z_k = 1)}$$
$$= \frac{\pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}{\sum_k \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}$$

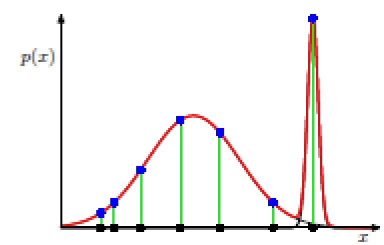
MIXTURE OF GAUSSIANS: MAXIMUM LIKELIHOOD

- Log Likelihood function of observations $X = \{x_1, \ldots, x_N\}$

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



- Problems with Log Likelihood
  - Singularity when a mixture component collapses on a data point
  - Identifiability for a ML solution in a K-component mixture there are K! equivalent solutions.
  - * We assume we can use heuristics to overcome these problems.



MIXTURE OF GAUSSIANS: EM FOR GAUSSIAN MIXTURES

- Informal introduction of expectation-maximization algorithm (Dempster et al., 1977).
- Maximum of log likelihood:
  - Derivatives of $\ln p(X|\pi, \mu, \Sigma)$ w.r.t parameters to 0.

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

EM FOR GAUSSIAN MIXTURES: SOLVE FOR μ_k

- Set derivative of $\ln p(X|\pi, \mu, \Sigma)$ w.r.t means μ_k of the Gaussian components to zero.

$$0 = -\sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}{\sum_k \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}}_{\gamma(z_k)} \Sigma_k^{-1}(\mathbf{x}_n - \mu_k)$$

Multiply by Σ_k

$$\mu_k = \frac{1}{\sum_n \gamma(z_k)} \sum_n \gamma(z_k)\mathbf{x}_n$$

EM FOR GAUSSIAN MIXTURES: SOLVE FOR Σ_k

- Set derivative of $\ln p(X|\pi, \mu, \Sigma)$ w.r.t Σ_k of the Gaussian components to zero.

$$\Sigma_k = \frac{1}{\sum_n \gamma(z_k)} \sum_n \gamma(z_k)(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T$$

- Each data point weighted by the corresponding posterior probability and with the denominator given by the effective number of point.

EM FOR GAUSSIAN MIXTURES: SOLVE FOR π_k

- Take into account constraint $\sum_k \pi_k = 1$
  - Can be done by introducing Lagrange multiplier

$$\ln p(\mathbf{X}|\pi, \mu, \Sigma) + \lambda(\sum_k \pi_k - 1)$$

- Set derivative of modified log likelihood w.r.t π_k of the Gaussian components to zero

$$0 = \sum_n \frac{\mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}{\sum_k \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)} + \lambda$$

$$\pi_k = \frac{\sum_n \gamma(z_k)}{N}$$

MIXTURE OF GAUSSIANS: EM FOR GAUSSIAN MIXTURES SUMMARY

1. Initialize $\{\mu_k, \Sigma_k, \pi_k\}$ and evaluate log-likeihood
2. E-Step: Evaluate responsibilities $\gamma(z_k)$
3. M-Step: Re-estimate paramters θ, using current responsibilities $\gamma(z_k)$

$$\mu_k^{\text{new}} = \frac{1}{\sum_n \gamma(z_k)} \sum_n \gamma(z_k) \mathbf{x}_n$$

$$\Sigma_k^{\text{new}} = \frac{1}{\sum_n \gamma(z_k)} \sum_n \gamma(z_k) (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T$$

$$\pi_k^{\text{new}} = \frac{\sum_n \gamma(z_k)}{N}$$

4. Evaluate log-likelihood $\ln p(X|\pi, \mu, \Sigma)$ and check for convergence of either the parameters or the log likelihood.

   If convergence criterion is not satisfied return to step 2.

AN ALTERNATIVE VIEW OF EM: LATENT VARIABLES

- Let X observed data, Z latent variables, parameters.
- Goal: maximize marginal log-likelihood of observed data

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

- Optimization problematic due to log-sum.
- Assume straightforward maximization for complete data

$$\ln p(X, Z \,|\theta)$$

- Latent Z is known only through $p(X, Z \,|\theta)$.

AN ALTERNATIVE VIEW OF EM: GENERAL EM ALGORITHM

Consider expectation of complete data log-likelihood.

1. Initialization: Choose initial set of parameters θ^{old}
2. E-step: use current parameters θ^{old} to compute $p(X, Z\ |\theta^{old})$.

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

3. M-step: determine θ^{new} by maximizing $Q(\theta, \theta^{old})$

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

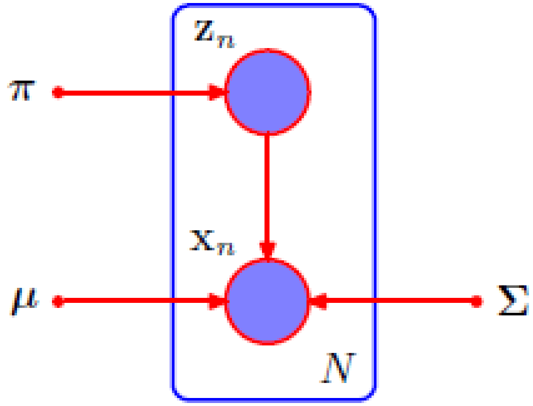
4. Check convergence either the log likelihood or the parameter values : stop, or $\theta^{old} \longleftarrow\ \theta^{new}$ and go to step 2.

AN ALTERNATIVE VIEW OF EM: GAUSSIAN MIXTURES REVISITED

- For mixture assign each **x** latent assignment variables z_{nk}. (the kth component of z_n)
- Complete-data (log-)likelihood,

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

$$\ln p(\mathbf{x}, \mathbf{z}|\theta) = \sum_{k=1}^{K} z_k \{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}$$



- If we know z_n mixing coefficients is simply

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

- PROBLEM: We don't know z_n

- Consider the expectation, with respect to the posterior distribution of the latent variables, of the complete-data log likelihood
- Posterior distribution : since $p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$. $p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$

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

- Expected value of the indicator variable z_{nk} under this posterior distribution

$$\mathbb{E}[z_{nk}] = \frac{\sum_{\mathbf{z}_n} z_{nk} \prod_{k'} [\pi_{k'} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})]^{z_{nk'}}}{\sum_{\mathbf{z}_n} \prod_{j} [\pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)]^{z_{nj}}}$$

$$= \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} = \gamma(z_{nk})$$

responsibility of component k for data point $\mathbf{x}_n$

- Expected value of the complete-data log likelihood function is therefore given by

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

- Use the derivatives to find for parameter μ_k, Σ_k, π_k

RELATION TO *K*-MEANS

- *K*-means algorithm with the EM algorithm for Gaussian mixtures shows that there is a close similarity
  - *K*-means algorithm performs a *hard* assignment of data points to clusters, in which each data point is associated uniquely with one cluster,
  - the EM algorithm makes a *soft* assignment based on the posterior probabilities.

THE EM ALGORITHM IN GENERAL

- Let X observed data, Z latent variables, θ parameters
- Goal: maximize marginal log-likelihood of observed data

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

- Maximization of $\mathrm{p}(\mathrm{X}, \mathrm{Z}|\theta)$ simple, but difficult for $\mathrm{p}(\mathrm{X}|\theta)$.
- Given any q(Z), we decompose the data log-likelihood

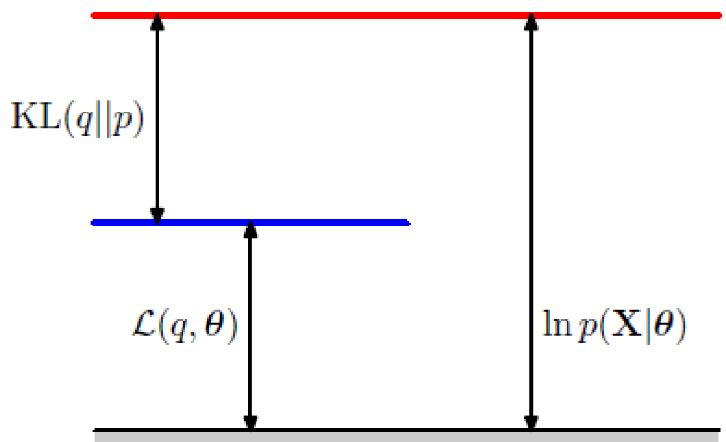
$$\begin{aligned}
\ln p(\mathbf{X}|\theta) &= \mathcal{L}(q, \theta) + \mathrm{KL}(q(\mathbf{Z}) \| p(\mathbf{Z}|\mathbf{X}, \theta)), \\
\mathcal{L}(q, \theta) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})}, \\
\mathrm{KL}(q(\mathbf{Z}) \| p(\mathbf{Z}|\mathbf{X}, \theta)) &= - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{Z}|\mathbf{X}, \theta)}{q(\mathbf{Z})} \geq 0.
\end{aligned}$$

THE EM ALGORITHM IN GENERAL: THE EM BOUND

- $\mathrm{L}(\mathrm{q}|\theta)$ is a lower bound on the data log-likelihood
  - $-\mathrm{L}(\mathrm{q}|\theta)$ known as variational free-energy

$$\mathcal{L}(q, \theta) \quad = \quad \ln p(\mathbf{X}|\theta) - \mathrm{KL}(q(\mathbf{Z}) \| p(\mathbf{Z}|\mathbf{X}, \theta)) \leq \ln p(\mathbf{X}|\theta)$$

- The EM algorithm performs coordinate ascent on L
  - E-step maximizes L w.r.t. q for fixed θ
  - M-step maximizes L w.r.t. for fixed q

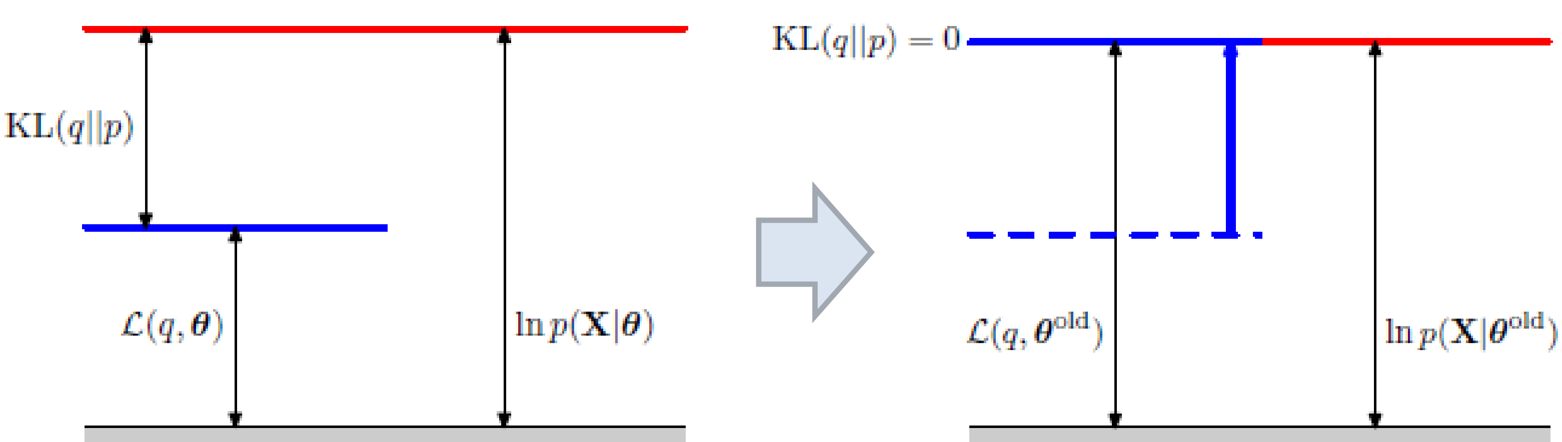


THE EM ALGORITHM IN GENERAL: THE E-STEP

- E-step maximizes L(q|θ) w.r.t. q for fixed θ

$$\mathcal{L}(q,\theta) \quad = \quad \ln p(\mathbf{X}|\theta) - \mathrm{KL}(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X},\theta))$$

- L maximized for $q(\mathbf{Z}) \leftarrow p(\boldsymbol{Z}|\boldsymbol{X},\theta)$



E step of the EM

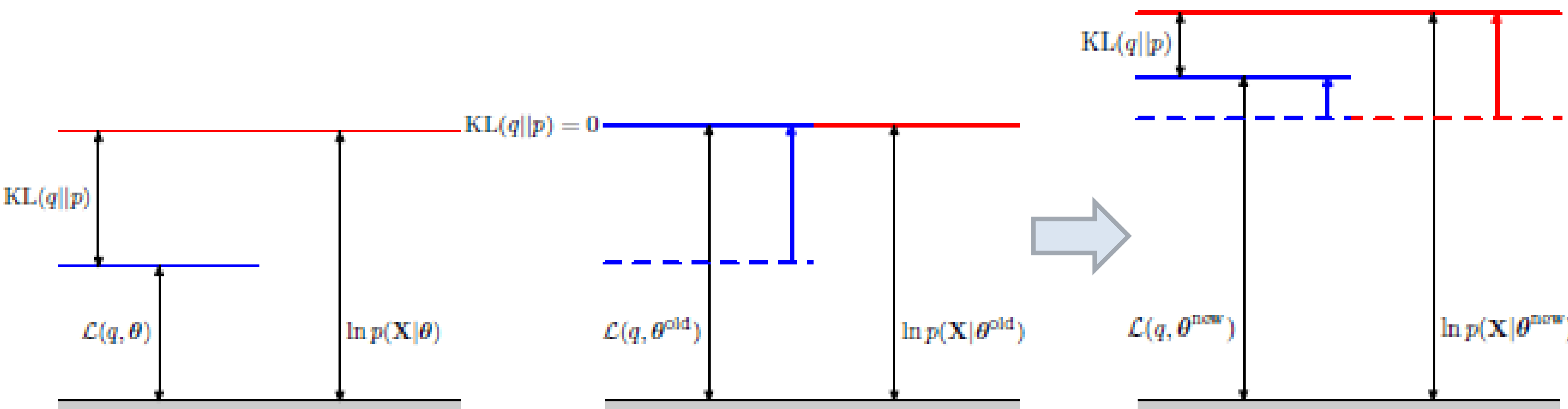
THE EM ALGORITHM IN GENERAL: THE M-STEP

- M-step maximizes L(q|θ) w.r.t. for fixed q

$$\mathcal{L}(q,\theta) \quad = \quad \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln p(\mathbf{X},\mathbf{Z}|\theta) - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln q(\mathbf{Z})$$

- L maximized for

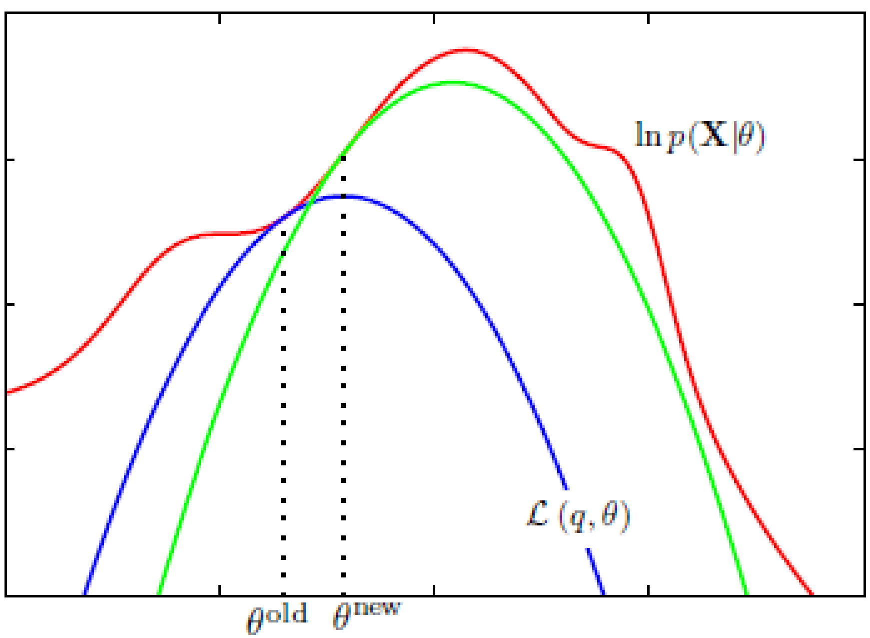
$$\theta = \arg\max_\theta \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln p(\mathbf{X},\mathbf{Z}|\theta)$$



M step of the EM

THE EM ALGORITHM IN GENERAL: PICTURE IN PARAMETER SPACE

- E-step resets bound $\mathrm{L}(\mathrm{q}|\theta)$ on $\ln p(\boldsymbol{X}|\theta)$ at $\theta = \theta^{old}$, it is
  - tight at $\theta = \theta^{old}$,
  - tangential at $\theta = \theta^{old}$,
  - convex (easy) in θ for exponential family mixture components



The EM algorithm involves alternately computing a lower bound on the log likelihood for the current parameter values and then maximizing this bound to obtain the new parameter values.

THE EM ALGORITHM IN GENERAL: FINAL THOUGHTS

- Local maxima of $\mathrm{L}(\mathrm{q}|\theta)$ correspond to those of $\ln p(\boldsymbol{X}|\theta)$
- EM converges to local maximum of likelihood
  - Coordinate ascent on $\mathrm{L}(\mathrm{q}|\theta)$ and $\mathrm{L}(\mathrm{q}|\theta) = \ln p(\boldsymbol{X}|\theta)$ after E-step
- Alternative schemes to optimize the bound
  - Generalized EM: relax M-step from maximizing to increasing L
  - Expectation Conditional Maximization: M-step maximizes w.r.t. groups of parameters in turn
  - Incremental EM: E-step per data point, incremental M-step
  - Variational EM: relax E-step from maximizing to increasing L
    - no longer L = $\ln p(\boldsymbol{X}|\theta)$ after E-step