1

Machine Learning

Chapter 9: Mixture Models &EM

林嘉文 (Chia-Wen Lin)
清華大學電機系
cwlin@ee.nthu.edu.tw

4/18/2022

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
- Chapter 10 provides Bayesian treatment using variational inference

K-Means Clustering: Distortion Measure

- Dataset $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$
- Partition in K clusters
- Cluster prototype: $\boldsymbol{\mu}_k$
- Binary indicator variable, 1-of-K Coding scheme

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

  if $\mathbf{x}_n$ is assigned to cluster k then $r_{nk} = 1$, and $r_{nj} = 0$ for $j \neq k$.

  (Hard assignment)
- Distortion measure

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

K-Means Clustering: Expectation Maximization

- Find values for $\{r_{nk}\}$ and $\{\boldsymbol{\mu}_k\}$ to minimize

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

- Iterative procedure:
  - Minimize J w.r.t. r_{nk}, keep $\boldsymbol{\mu}_k$ fixed (Expectation)

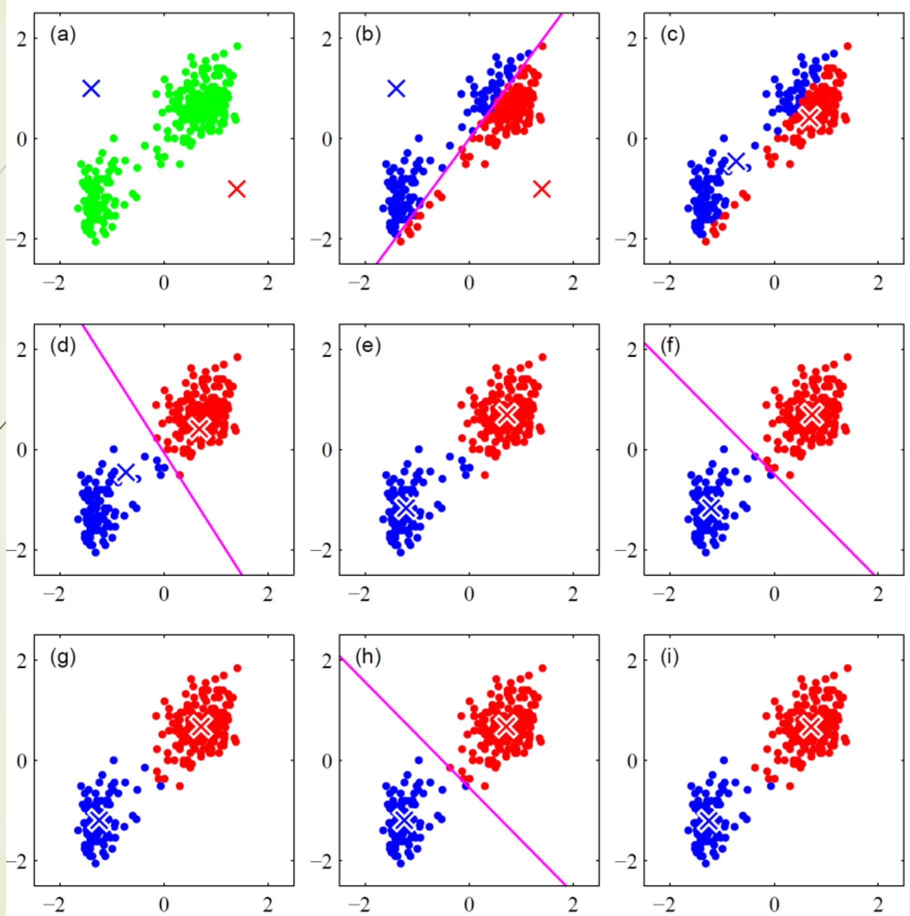
$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \left\|\mathbf{x}_n - \boldsymbol{\mu}_j\right\|^2 \\ 0 & \text{otherwise} \end{cases}$$

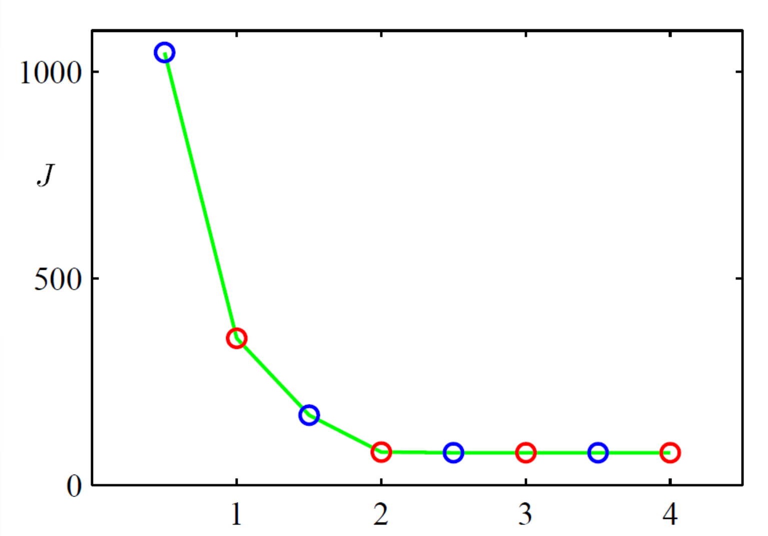
  - Minimize J w.r.t. $\boldsymbol{\mu}_k$, keep r_{nk} fixed (Maximization)

$$2 \sum_{n=1}^{N} r_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k) = 0$$

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

K-Means Clustering: Example





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

K-Means Clustering: Concluding Remarks

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

$$\boldsymbol{\mu}_k^{\mathrm{new}} = \boldsymbol{\mu}_k^{\mathrm{old}} + \eta_n\left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\mathrm{old}}\right)$$

- K-medoids, general distortion measure

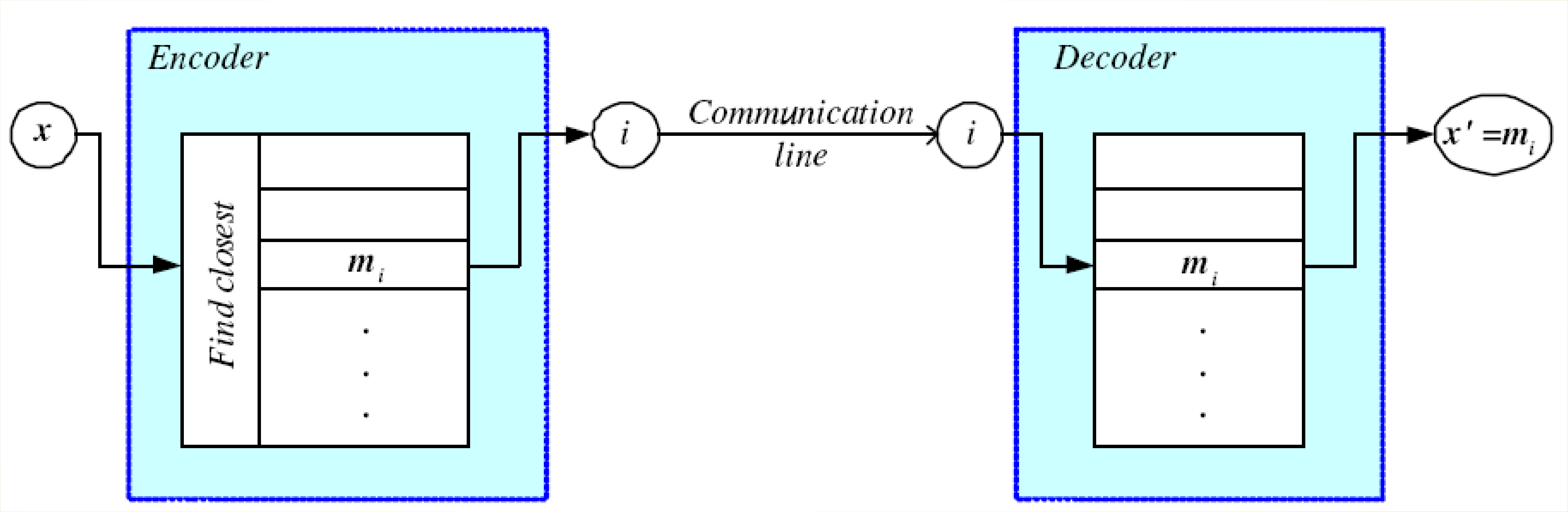
$$\tilde{J} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \mathcal{V}(\mathbf{x}_n, \boldsymbol{\mu}_k)$$

where $\mathcal{V}(\cdot,\cdot)$ is any kind of dissimilarity measure, $\boldsymbol{\mu}_k$ should be assigned with a sample value

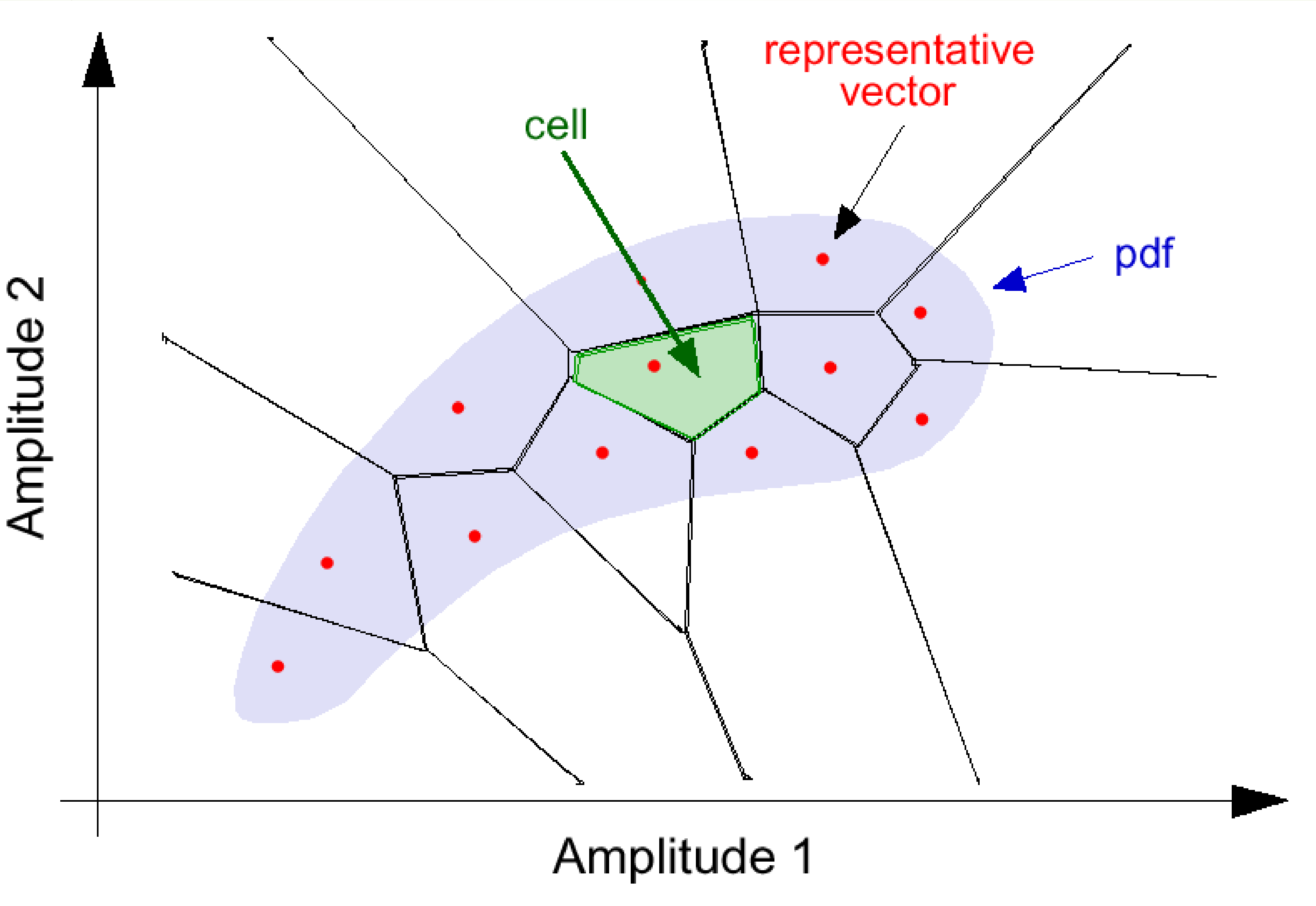
K-Means Clustering: Image Segmentation

Vector Quantization (VQ)

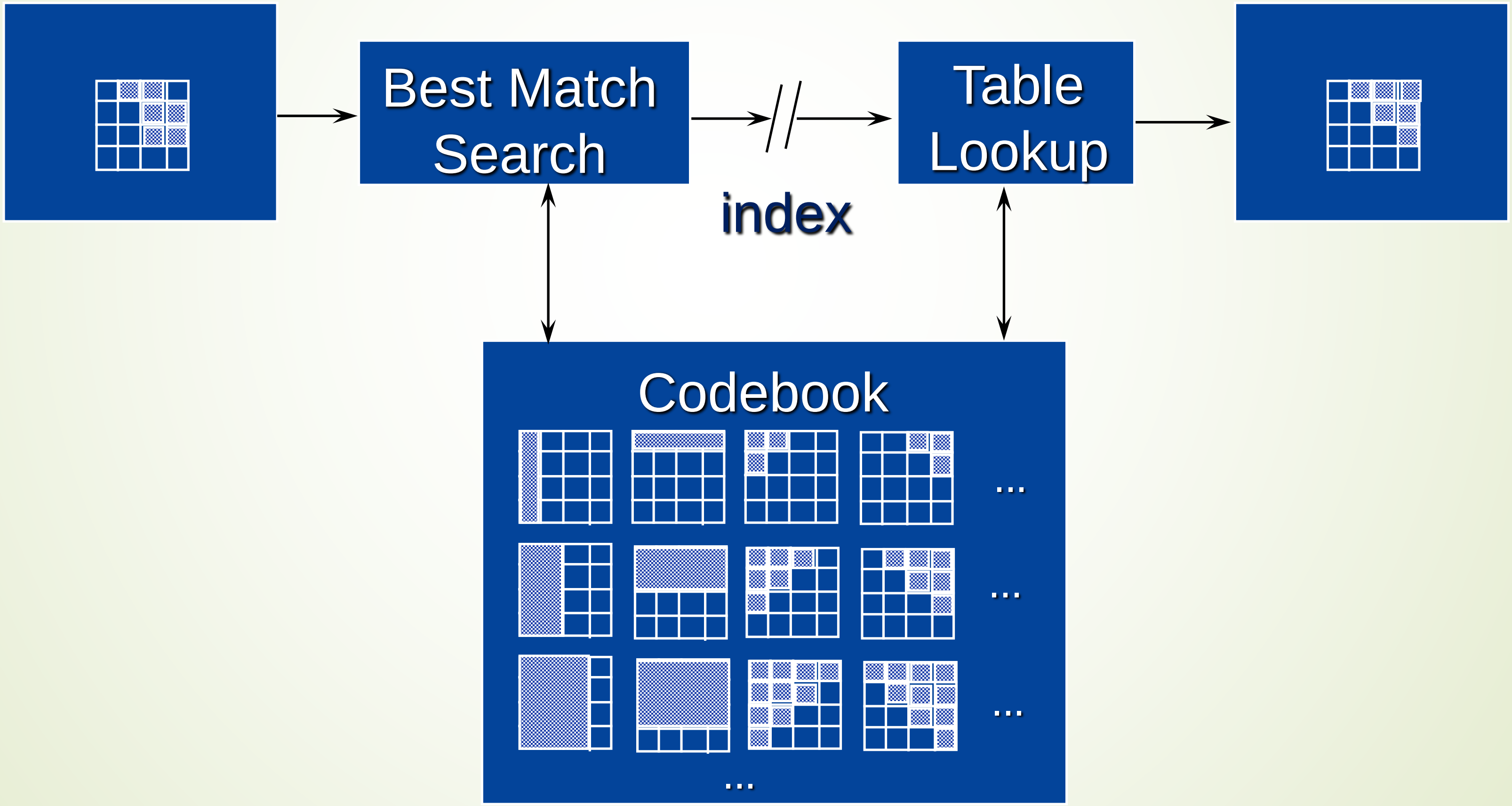
K-Means Clustering: Vector Quantization



K-Means Clustering: Vector Quantization



K-Means Clustering: Vector Quantization

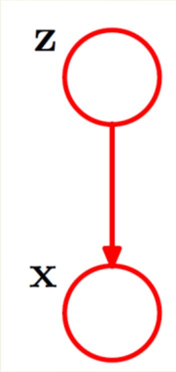


Mixture of Gaussians: Latent variables

- Gaussian Mixture Distribution:

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

- Introduce latent variable $\mathbf{z}$
  - $\mathbf{z}$ is a binary 1-of-K coding variable
  - $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$



Mixture of Gaussians: Latent variables

- $p(z_k = 1) = \pi_k$

  Constraints: $0 \leq \pi_k \leq 1$, and $\sum_k \pi_k = 1$
- Because $\mathbf{z}$ uses a 1-of-K representation, we can rewrite $p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k{}^{z_k}$
- $p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

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

- $p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}) = \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{x}|\mathbf{z}) = \sum_k \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- The use of the joint probability $p(\mathbf{x}, \mathbf{z})$, leads to significant simplifications

Mixture of Gaussians: Latent variables

- Responsibility of component k to generate observation $\mathbf{x}$

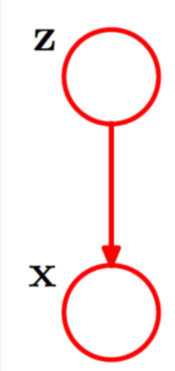
$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_k p(z_k = 1)p(\mathbf{x}|z_k = 1)} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_k \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$$

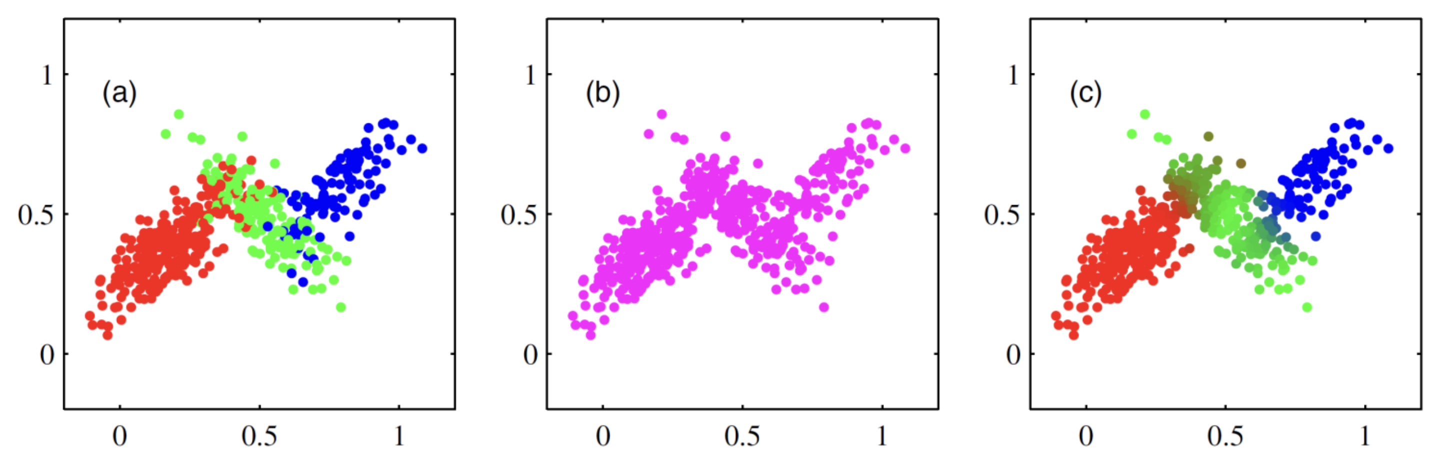
  is the posterior probability

- Generate random samples with ancestral sampling:

  First: generate $\hat{\mathbf{z}}$ from $p(\mathbf{z})$

  Second: generate a value for $\mathbf{x}$ from $p(\mathbf{x}|\hat{\mathbf{z}})$

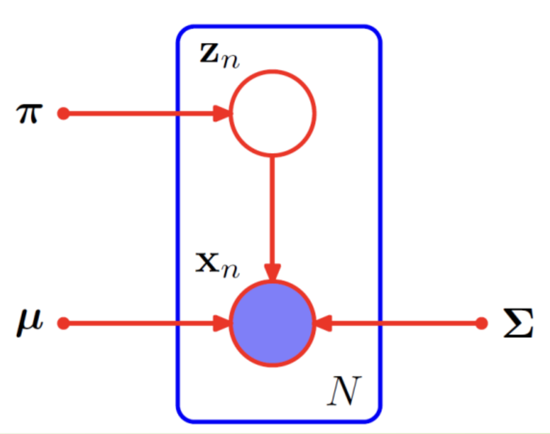




Mixture of Gaussians: Maximum Likelihood

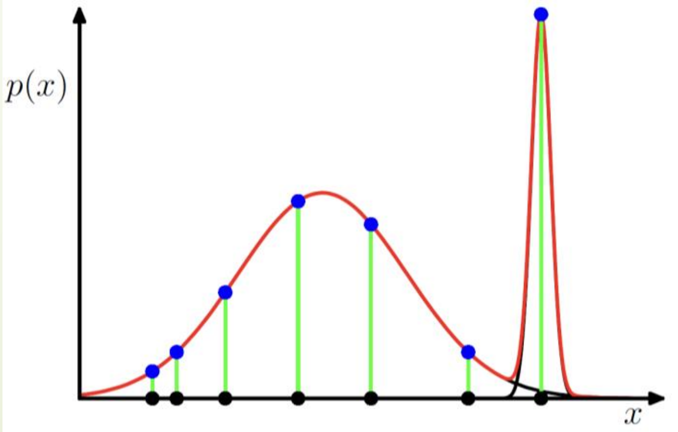
- Log Likelihood

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



- Singularity when a mixture component collapses on a datapoint
- Identifiability for a ML solution in a K-component mixture there are $K!$ equivalent solutions

Singularity



EM for Gaussian Mixtures

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

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

- For $\boldsymbol{\mu}_k$

$$0 = \sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)}}_{\gamma(z_{nk})} \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)$$

$$\gamma(z_{nk}) \equiv p(z_{nk} = 1|\mathbf{x}_n)$$

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

EM for Gaussian Mixtures

- For $\boldsymbol{\Sigma}_k$

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

- For π_k
  - Take into account constraint $\sum_k \pi_k = 1$
  - Lagrange multiplier

$$\ln p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right)$$

$$0 = \sum_{n=1}^{N}\frac{\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_j \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)} + \lambda \qquad (\lambda = -N)$$

$$\pi_k = \frac{N_k}{N}$$

EM for Gaussian Mixtures Example

- No closed form solutions: $\gamma(z_{nk})$ depends on parameters
- But these equations suggest simple iterative scheme for finding maximum likelihood:

  Alternate between estimating the current $\gamma(z_{nk})$ and updating the parameters $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k\}$
- More iterations needed to converge than K-means algorithm, and each cycle requires more computation
- It's common to use K-means to initialize parameters

EM for Gaussian Mixtures Example

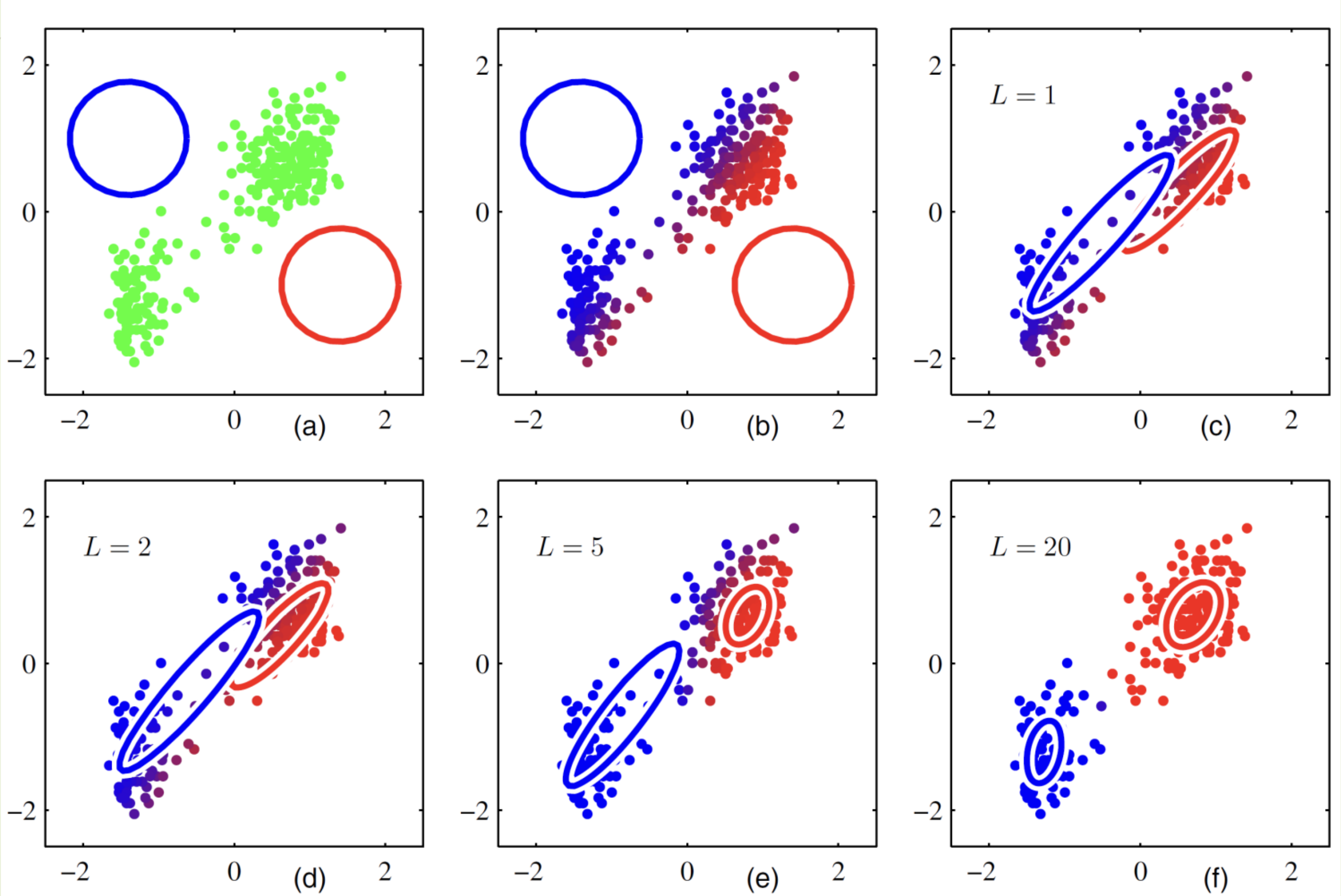


Illustration of the EM algorithm using the Old Faithful set as used for the illustration of the K-means algorithm

EM for Gaussian Mixtures Summary

1. Initialize $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k\}$ and evaluate log-likelihood
2. E-Step: Evaluate responsibilities

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_k \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)}$$

3. M-Step: Re-estimate parameters, using current responsibilities

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

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

$$\pi_k = \frac{N_k}{N} = \frac{\sum_{n=1}^N \gamma(z_{nk})}{N} \qquad N_k = \sum_{n=1}^N \gamma(z_{nk})$$

4. Evaluate log-likelihood $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and check for convergence (go to step 2)

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

An Alternative View of EM: Latent Variables

- Let $\mathbf{X}$ observed data, $\mathbf{Z}$ latent variables, $\boldsymbol{\theta}$ parameters
- Goal: maximize marginal log-likelihood of observed data

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

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

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

$$\gamma(z_{nk}) \equiv p(z_{nk}=1|\mathbf{x}_n) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_k \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}$$

- Latent $\mathbf{Z}$ is known only through $p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})$
- We will consider expectation of complete data log-likelihood

An Alternative View of EM: Algorithm

1. Initialization: Choose initial set of parameters $\boldsymbol{\theta}^{\text{old}}$
2. E-step: use current parameters $\boldsymbol{\theta}^{\text{old}}$ to compute $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$ to find the expected complete-data log-likelihood for general $\boldsymbol{\theta}$

   Evaluate $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$

   $$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})]$$

3. M-step: determine $\boldsymbol{\theta}^{\text{new}}$ by maximizing

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

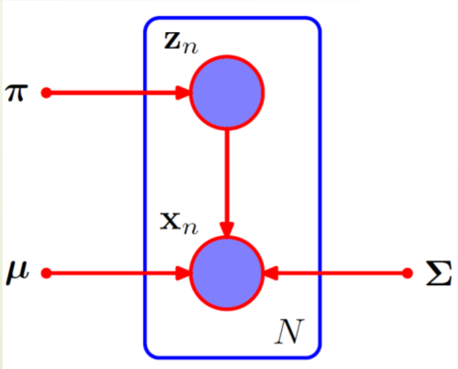
4. Check convergence: stop, or $\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}}$ and go to E-step

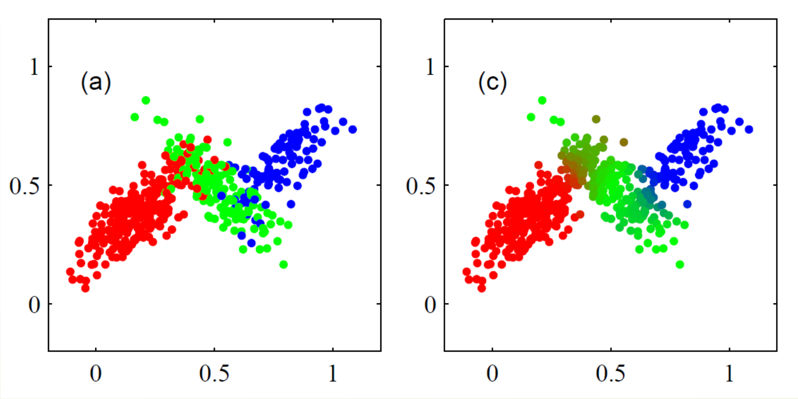
Gaussian Mixtures Revisited

- For mixture assign each $\mathbf{x}$ latent assignment variables z_{nk}
- Complete-data (log-)likelihood, and expectation:

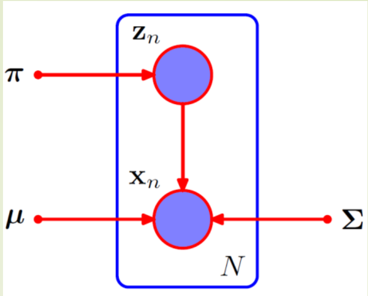
$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}$$

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





Gaussian Mixtures Revisited



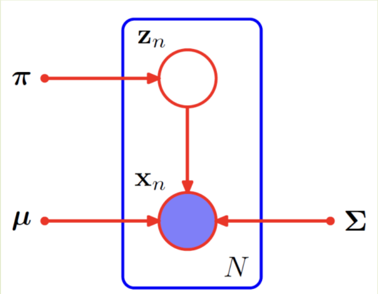
- But these equations suggest a simple iterative scheme for finding maximum likelihood:

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

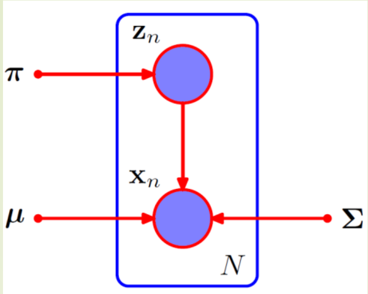
$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} (\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))$$

- Cross-reference: the log-likelihood of incomplete data

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



Gaussian Mixtures Revisited



- But these equations suggest a simple iterative scheme for finding maximum likelihood:

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

- The expected value of z_{nk} under this posterior distribution is

$$\mathbb{E}[z_{nk}] = \frac{\sum_{\mathbf{z}_n} z_{nk} \prod_{k'}\left[\pi_{k'}\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_{k'},\boldsymbol{\Sigma}_{k'}\right)\right]^{z_{nk'}}}{\sum_{\mathbf{z}_n}\prod_i[\pi_i\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)]^{z_{ni}}} = \frac{\pi_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K}\pi_j\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j\right)} = \gamma(z_{nk})$$

- The expected value of the complete-data log likelihood function:

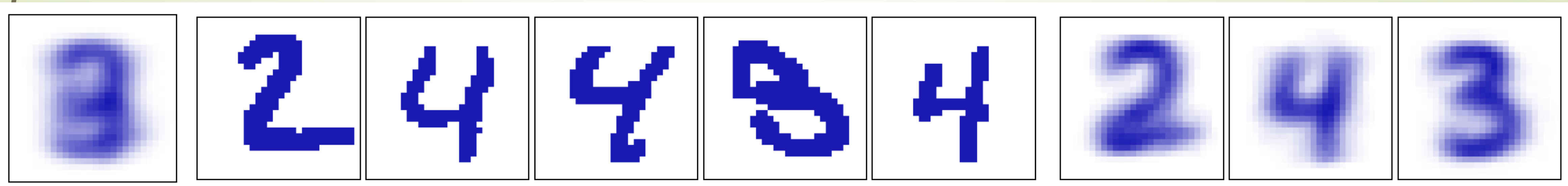
$$\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi})] = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})(\ln\pi_k + \ln\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k))$$

EM Example: Bernoulli Mixtures

- Bernoulli distributions over binary data vectors

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{i=1}^{D} \mu_i^{x_i}(1-\mu_i)^{1-x_i}$$

- Mixture of Bernoullis can model variable correlations
- Same as the Gaussian, Bernoulli is member of exponential family
  - Model log-linear, mixture not, complete-data log-likelihood is
- Simple EM algorithm to find ML parameters
  - E-step: compute responsibilities $\gamma(z_{nk}) \propto \pi_k \, p(\mathbf{x}_n|\boldsymbol{\mu}_k)$
  - M-step: update parameters $\pi_k = N^{-1} \sum_n \gamma(z_{nk})$ and $\boldsymbol{\mu}_k = N_{\pi_k}^{-1} \sum_n \gamma(z_{nk})\mathbf{x}_n$



EM Example: Bayesian Linear Regression

- Recall Bayesian linear regression: it's a latent variable model

$$p(\mathbf{t}|\mathbf{w},\beta,\mathbf{X}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{t}_n|\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_N),\beta^{-1})$$

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0},\alpha^{-1}\,\mathbf{I})$$

$$p(\mathbf{t}|\alpha,\beta,\mathbf{X}) = \int p(\mathbf{t}|\mathbf{w},\beta)p(\mathbf{w}|\alpha)d\mathbf{w}$$

- Simple EM algorithm to find ML parameters (α,β)
  - E-step: compute responsibilities over latent variable $\mathbf{w}$
  - $p(\mathbf{w}|\mathbf{t},\beta,\mathbf{X}) = \mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{S})$, $\mathbf{m} = \beta\mathbf{S}\boldsymbol{\Phi}^\top\mathbf{t}$, $\mathbf{S}^{-1} = \alpha\mathbf{I} + \beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi}$
  - M-step: update parameters using complete-data log-likelihood

$$\alpha = \frac{M}{\mathbf{m}_N^\top\mathbf{m}_N + \mathrm{Tr}(\mathbf{S}_N)}$$

$$(\beta^{\text{new}})^{-1} = \frac{1}{N}\left(\|\mathbf{t}-\boldsymbol{\Phi}\mathbf{m}_N\|^2 + \beta^{-1}\textstyle\sum_i \gamma^i\right)$$

The EM Algorithm in General

- EM is a general technique for finding maximum likelihood solutions for probabilistic models having latent variables
- Let $\mathbf{X}$ denote observed data, $\mathbf{Z}$ latent variables, $\boldsymbol{\theta}$ parameters
- Goal: maximize the marginal log-likelihood of observed data

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

- Maximization of $p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})$ is simple, but it's difficult for $p(\mathbf{X}|\boldsymbol{\theta})$
- Given any $q(\mathbf{Z})$, we decompose the data log-likelihood

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

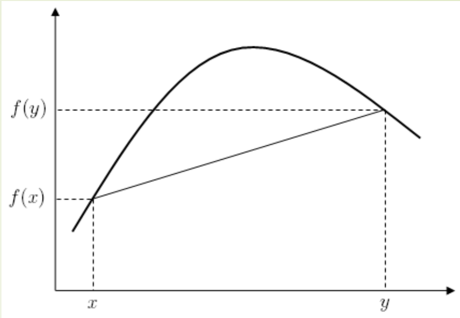
$$\mathcal{L}(q,\boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})}$$

$$\mathrm{KL}(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{q(\mathbf{Z})} \geq \mathbf{0}$$

Lower Bound of $\ln p(\mathbf{X}|\boldsymbol{\theta})$

- Jensen's inequality

$$\begin{cases} \varphi(\mathbb{E}[\mathbf{Y}]) \geq \mathbb{E}[\varphi(\mathbf{Y})] & \text{if } \varphi(\mathbf{Y}) \text{ is concave} \\ \varphi(\mathbb{E}[\mathbf{Y}]) \leq \mathbb{E}[\varphi(\mathbf{Y})] & \text{if } \varphi(\mathbf{Y}) \text{ is convex} \end{cases}$$



- Since $\ln f(\mathbf{X})$ is concave, $\ln \mathbb{E}[f(\mathbf{X})] \geq \mathbb{E}[\ln f(\mathbf{X})]$

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) = \ln \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} = \ln \mathbb{E}_{\mathbf{Z}\sim q(\mathbf{Z})}\left[\frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})}\right]$$

$$\geq \mathbb{E}_{\mathbf{Z}\sim q(\mathbf{Z})}\left[\ln \frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})}\right] = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} = \mathcal{L}(q,\boldsymbol{\theta})$$

- Given any $q(\mathbf{Z})$, we decompose the data log-likelihood

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

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

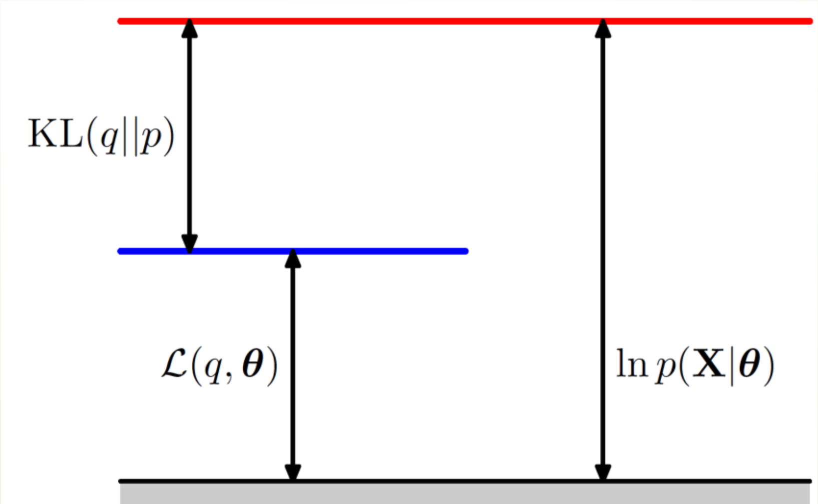
$$\mathrm{KL}(q(\mathbf{Z})\|p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{q(\mathbf{Z})} \geq \mathbf{0}$$

The EM Algorithm in General: The EM Bound

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

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

- The EM algorithm performs coordinate ascent on $\mathcal{L}$
  - E-step maximizes $\mathcal{L}$ w.r.t. q for fixed $\boldsymbol{\theta}$
  - M-step maximizes $\mathcal{L}$ w.r.t. $\boldsymbol{\theta}$ for fixed q

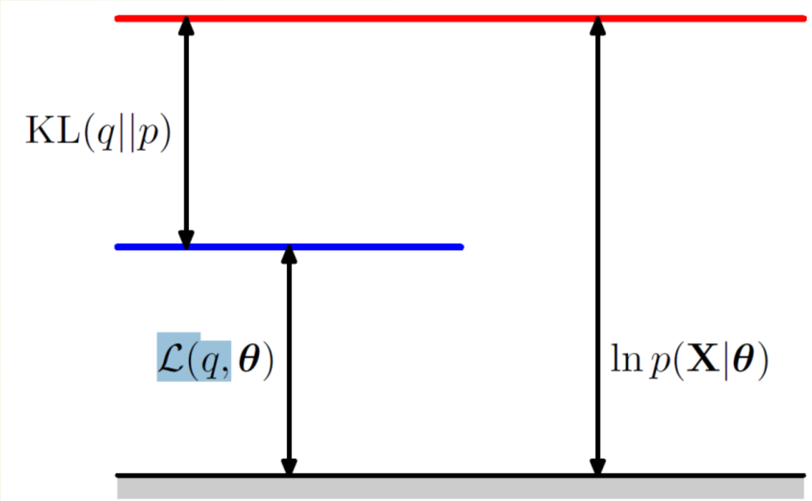


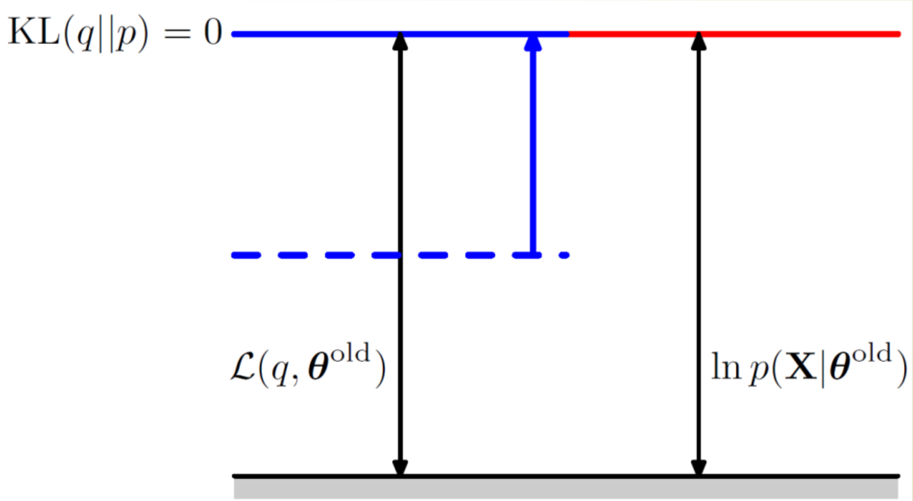
The EM Algorithm in General: The E-step

- E-step maximizes $\mathcal{L}$ w.r.t. q for fixed $\boldsymbol{\theta}^{\text{old}}$

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

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





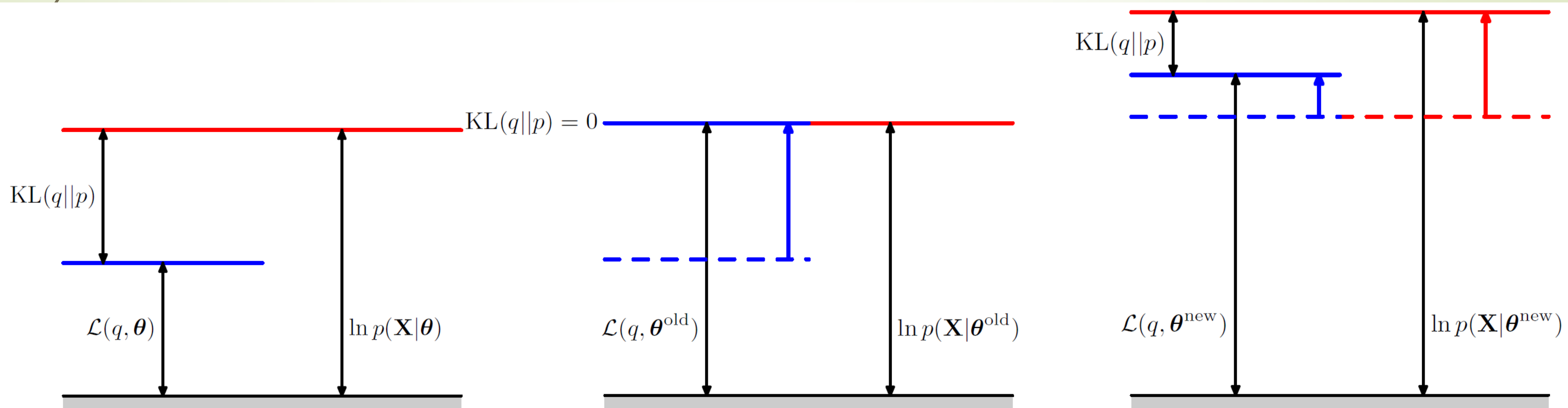
The EM Algorithm in General: The M-step

- M-step maximizes $\mathcal{L}$ w.r.t. $\boldsymbol{\theta}$ for fixed q

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

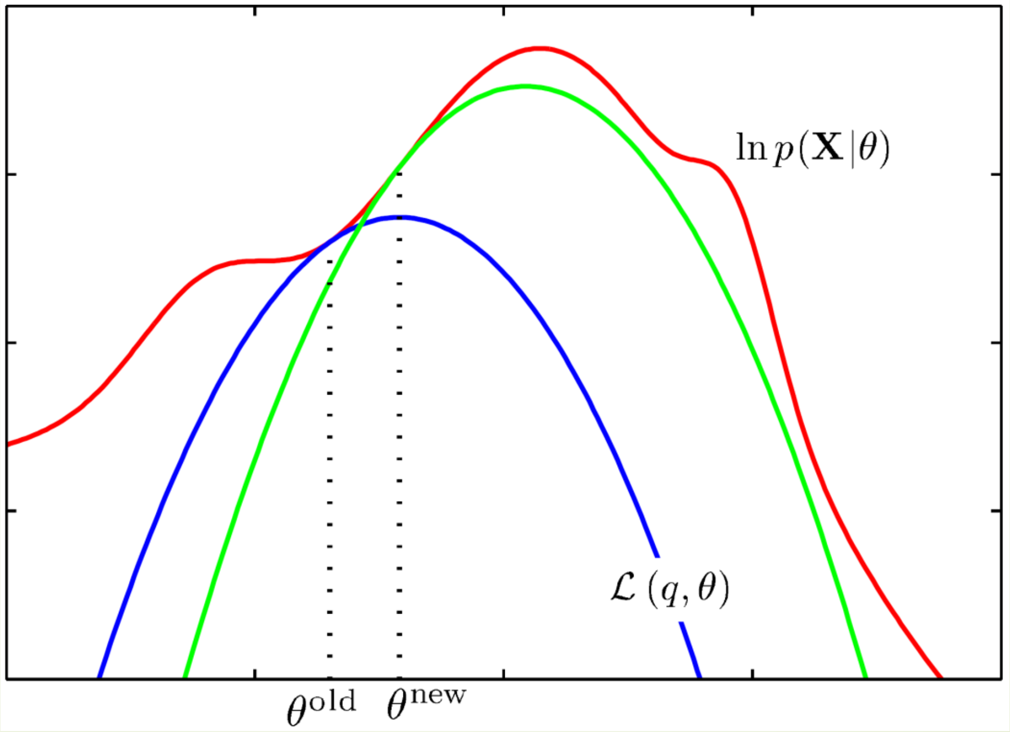
$$= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln q(\mathbf{Z})$$

- $\mathcal{L}$ maximized for $\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$



Picture in Parameter Space

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



The EM Algorithm in General: Final Thoughts

- (local) maxima of $\mathcal{L}(q, \boldsymbol{\theta})$ correspond to those of $\ln p(\mathbf{X}|\boldsymbol{\theta})$
- EM converges to (local) maximum of likelihood
  - Coordinate ascent on $\mathcal{L}(q, \boldsymbol{\theta})$, and $\mathcal{L} = \ln p(\mathbf{X}|\boldsymbol{\theta})$ after E-step
- Alternative schemes to optimize the bound
  - Generalized EM: relax M-step from maximizing to increasing $\mathcal{L}$
  - 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 $\mathcal{L}$
    - no longer $\mathcal{L} = \ln p(\mathbf{X}|\boldsymbol{\theta})$ after E-step
- Same applies for MAP estimation $p(\boldsymbol{\theta}|\mathbf{X}) = p(\boldsymbol{\theta})\, p(\mathbf{X}|\boldsymbol{\theta})/p(\mathbf{X})$
  - bound second term: $\ln p(\boldsymbol{\theta}|\mathbf{X}) = \ln p(\boldsymbol{\theta}) + \mathcal{L}(q, \boldsymbol{\theta}) - \ln p(\mathbf{X})$