CSE 575
Statistical Machine Learning

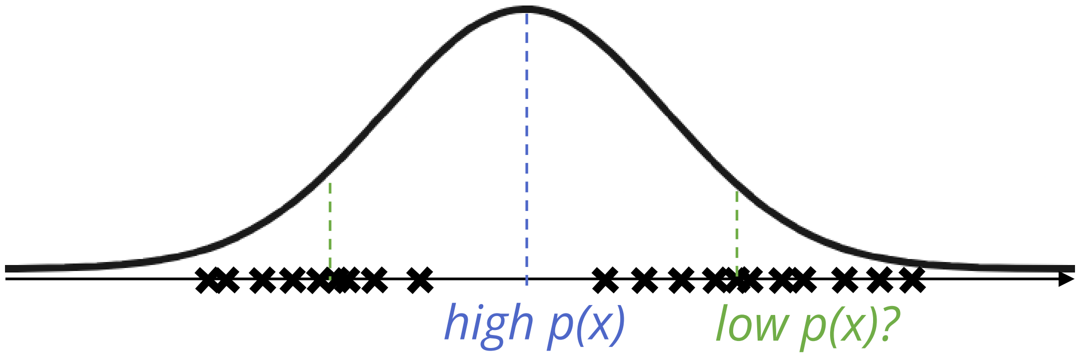
Lecture 16

YooJung Choi

Fall 2022

Revisiting Gaussians

- Consider the following 1D continuous data
- Suppose we model it using a Gaussian distribution
- Limitation: a Gaussian is *unimodal* (single "peak")



Revisiting Gaussians

- Consider the following 1D continuous data
- Suppose we model it using a Gaussian distribution
- Limitation: a Gaussian is *unimodal* (single "peak")
- Instead, let's model each "group" using a Gaussian

Then take the sum to obtain
a single density function

Gaussian mixture models (GMMs)

- Mixture of Gaussians: *weighted sum of K Gaussian distributions*

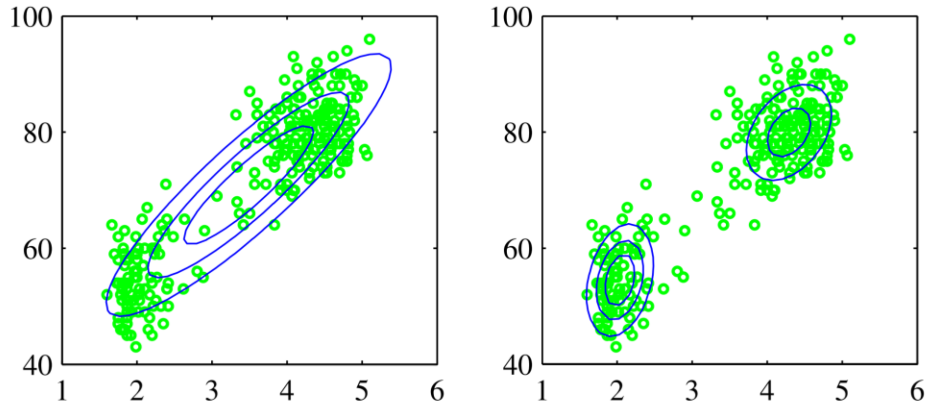
$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

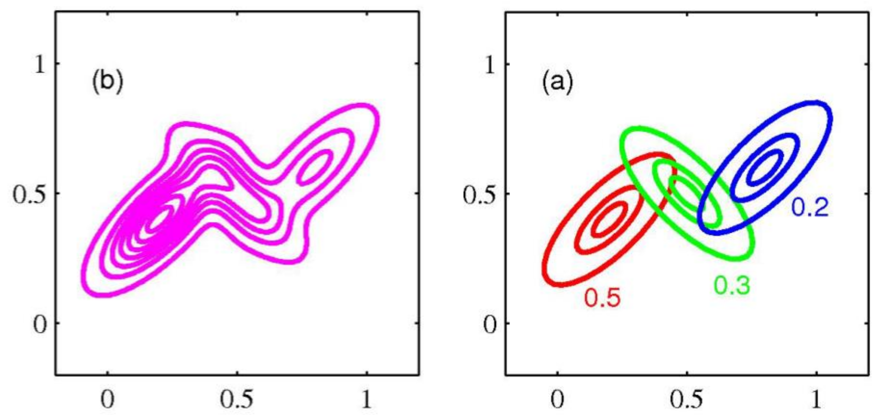
convex combination

- Mixing coefficients π_k: $\pi_k \geq 0$ and

π_k can be probabilities!

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





GMM: Latent variable interpretation

- Introduce a K-valued discrete random variable z such that:

$$p(z = k) = \pi_k$$

- Now the model represents the joint distribution $p(\mathbf{x}, z)$

- We can interpret the probability of $\mathbf{x}$ given by a GMM as

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

- Using sum rule and product rule: $p(\mathbf{x}) = \sum_{k=1}^{K} p(z = k) p(\mathbf{x} \mid z = k)$

- Each Gaussian component is a conditional distribution: $p(\mathbf{x} \mid z = k) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

Maximum likelihood estimation

- Gaussian mixture model: K mixture components, each associated with
  - A mixture coefficient π_k (representing $p(z = k)$)
  - A Gaussian distribution $\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ (representing $p(\mathbf{x} \mid z = k)$)
- How to learn the maximum-likelihood estimates for π_k, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$?
- Recall: Gaussian discriminant analysis (general case, no shared covariance)

x_1	x_2	$\cdots$	x_D	z (class)
1.5	3	$\cdots$	-4	1
0	4	$\cdots$	3.5	1
-1	1	$\cdots$	-6.3	2
0.7	1	$\cdots$	1.4	1
-0.2	2.5	$\cdots$	1.0	2
$\vdots$	$\vdots$		$\vdots$	$\vdots$

$$\pi_k = \frac{\#\{z = k\}}{N} \qquad \boldsymbol{\mu}_k = \frac{\sum_{n:z_n=k} \mathbf{x}_n}{\#\{z = k\}}$$

$$\boldsymbol{\Sigma}_k = \frac{\sum_{n:z_n=k} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T}{\#\{z = k\}}$$

Maximum likelihood estimation

- Gaussian mixture model: K mixture components, each associated with
  - A mixture coefficient π_k (representing $p(z = k)$)
  - A Gaussian distribution $\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ (representing $p(\mathbf{x} \mid z = k)$)
- How to learn the maximum-likelihood estimates for π_k, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$?
- For a GMM, z is a latent (hidden) variable!

x_1	x_2	$\cdots$	x_D	z
1.5	3	$\cdots$	-4	?
0	4	$\cdots$	3.5	?
-1	1	$\cdots$	-6.3	?
0.7	1	$\cdots$	1.4	?
-0.2	2.5	$\cdots$	1.0	?
$\vdots$	$\vdots$		$\vdots$	?

Maximum likelihood estimation

- Gaussian mixture model: K mixture components, each associated with
  - A mixture coefficient π_k (representing $p(z=k)$)
  - A Gaussian distribution $\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ (representing $p(\mathbf{x} \mid z=k)$)
- How to learn the maximum-likelihood estimates for π_k, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$?
- Log-likelihood of a GMM given N data examples $\mathbf{x}_1, \dots, \mathbf{x}_N$:

$$ll(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \sum_{n=1}^{N} \log p(\mathbf{x}_n) = \sum_{n=1}^{N} \log \left\{ \sum_{k=1}^{K} p(z=k) p(\mathbf{x}_n | z=k) \right\}$$

Note: *marginal log-likelihood*

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

Log cannot "reach" exp due to the summation

$$= \sum_{n=1}^{N} \log \left\{ \sum_{k=1}^{K} \pi_k \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) \right\} \right\}$$

Expectation maximization

- Iterative approach
- Algorithm (informally):
  - Starting with some initial parameters π_k, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, repeat until convergence:
  - E step: "guess" the values of z_n for $n = 1, \ldots, N$, informed by the current parameters
  - M step: update the parameters based on the guess

x_1	x_2	$\cdots$	x_D	z
1.5	3	$\cdots$	-4	?
0	4	$\cdots$	3.5	?
-1	1	$\cdots$	-6.3	?
0.7	1	$\cdots$	1.4	?
-0.2	2.5	$\cdots$	1.0	?
$\vdots$	$\vdots$		$\vdots$	?

$\xrightarrow{\text{E step}}$

x_1	x_2	$\cdots$	x_D	z
1.5	3	$\cdots$	-4	z_1
0	4	$\cdots$	3.5	z_2
-1	1	$\cdots$	-6.3	z_3
0.7	1	$\cdots$	1.4	z_4
-0.2	2.5	$\cdots$	1.0	z_5
$\vdots$	$\vdots$		$\vdots$	$\vdots$

$\xrightarrow{\text{M step}}$

Maximum likelihood estimates for π_k, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ a la GDA

E step

- How to "guess" the values of z_n?
- Use the GMM to compute $p(z_n = k \mid \mathbf{x}_n)$

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

Called the responsibility of component k on $\mathbf{x}_n$. Denoted as $\gamma(z_n = k)$

- *Complete* the data using $\gamma(z_n = k)$

x_1	x_2	$\cdots$	x_D	z
1.5	3	$\cdots$	-4	?
0	4	$\cdots$	3.5	?
-1	1	$\cdots$	-6.3	?
$\vdots$	$\vdots$		$\vdots$	?

weighted dataset

x_1	x_2	$\cdots$	x_D	z	
1.5	3	$\cdots$	-4	1	$\gamma(z_1 = 1) = 0.24$
1.5	3	$\cdots$	-4	2	$\gamma(z_1 = 2) = 0.76$
0	4	$\cdots$	3.5	1	$\gamma(z_2 = 1) = 0.99$
0	4	$\cdots$	3.5	2	$\gamma(z_2 = 1) = 0.01$
-1	1	$\cdots$	-6.3	$\vdots$	
$\vdots$	$\vdots$		$\vdots$	$\vdots$	

M step

- How to update the parameters given a weighted dataset?
- In the case of complete dataset (i.e. all weights are 1):

$$\pi_k = \frac{\#\{z = k\}}{N} \qquad \boldsymbol{\mu}_k = \frac{\sum_{n:z_n=k} \mathbf{x}_n}{\#\{z = k\}} \qquad \boldsymbol{\Sigma}_k = \frac{\sum_{n:z_n=k}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T}{\#\{z = k\}}$$

- Using $\gamma(z_n = k)$:

$$\pi_k = \frac{\sum_{n=1}^{N} \gamma(z_n = k)}{N} \qquad \boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma(z_n = k)\mathbf{x}_n}{\sum_{n=1}^{N} \gamma(z_n = k)} \qquad \boldsymbol{\Sigma}_k = \frac{\sum_{n=1}^{N} \gamma(z_n = k)(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T}{\sum_{n=1}^{N} \gamma(z_n = k)}$$

In practice, updates would be made without explicitly constructing the weighted dataset

EM for GMMs

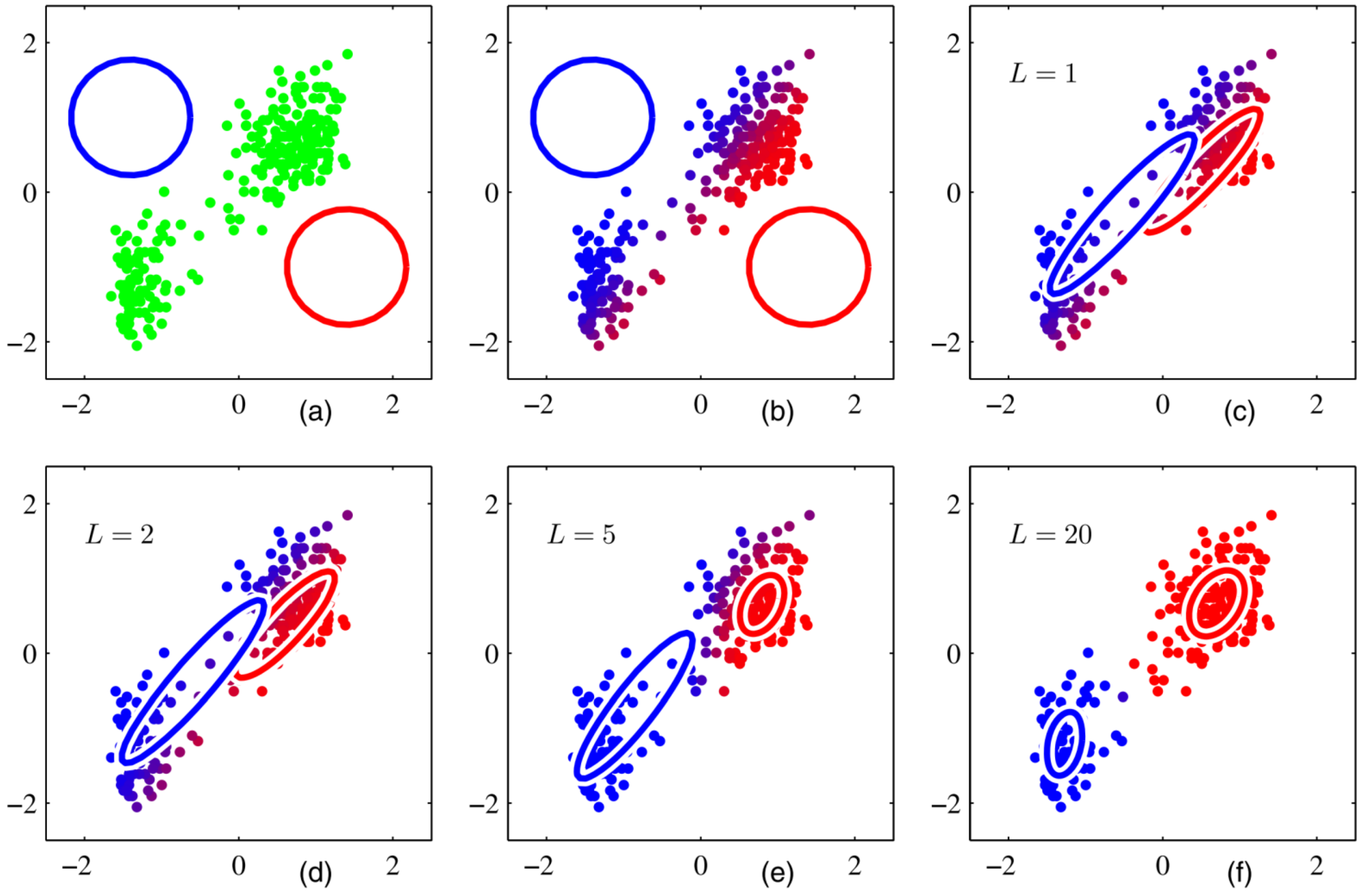
Putting everything together:

1. Initialize π_k, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$
2. Until convergence, repeat:
    1. E-step: for all n and k, compute $\gamma(z_n = k) = \dfrac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$
    2. M-step: for all k, let $N_k = \sum_{n=1}^{N} \gamma(z_n = k)$ and compute

$$\pi_k^{(new)} = \frac{N_k}{N}, \qquad \boldsymbol{\mu}_k^{(new)} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_n = k) \mathbf{x}_n ,$$

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

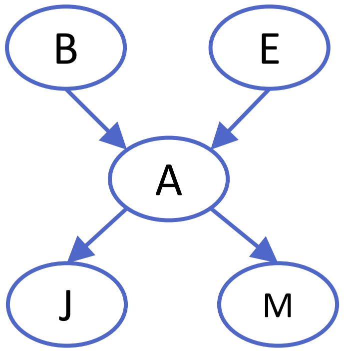
EM for GMMs



2
0
−2
−2
0
(a)
2
2
0
−2
−2
0
(b)
2
2
$L = 1$
0
−2
−2
0
(c)
2
2
$L = 2$
0
−2
−2
0
(d)
2
2
$L = 5$
0
−2
−2
0
(e)
2
2
$L = 20$
0
−2
−2
0
(f)
2

EM for BNs

- Expectation maximization is a general algorithm for learning with latent variables or incomplete data
- E.g. Bayesian network parameter learning from incomplete data



B	E	A	J	M
1	?	0	1	0
0	1	?	1	1
1	1	1	0	0
?	0	?	1	0

B	E	A	J	M	
1	0	0	1	0	$P(E=0 \mid B=1, A=0, J=1, M=0)$
1	1	0	1	0	$P(E=1 \mid B=1, A=0, J=1, M=0)$
0	1	0	1	1	$P(A=0 \mid B=0, E=1, J=1, M=1)$
0	1	1	1	1	$P(A=1 \mid B=0, E=1, J=1, M=1)$
1	1	1	0	0	1
0	0	0	1	0	$P(B=0, A=0 \mid E=0, J=1, M=0)$
0	0	1	1	0	$P(B=0, A=1 \mid E=0, J=1, M=0)$
1	0	0	1	0	$P(B=1, A=0 \mid E=0, J=1, M=0)$
1	0	1	1	0	$P(B=1, A=1 \mid E=0, J=1, M=0)$

EM for BNs

- Expectation maximization is a general algorithm for learning with latent variables or incomplete data
- E.g. Bayesian network parameter learning from incomplete data

B	E	A	J	M	
1	0	0	1	0	0.2
1	1	0	1	0	0.8
0	1	0	1	1	0.9
0	1	1	1	1	0.1
1	1	1	0	0	1
0	0	0	1	0	0.15
0	0	1	1	0	0.50
1	0	0	1	0	0.25
1	0	1	1	0	0.1

$N = 4$

$$\theta_E^{(new)} = \frac{0.8 + 0.9 + 0.1 + 1}{4}$$

$$\theta_{A|1,0}^{(new)} = \frac{0.1}{0.2 + 0.25 + 0.1}$$

$\vdots$

...and repeat until convergence

EM for BNs

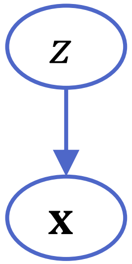
- Expectation … variables or incomplet…
- E.g. Bayes…

B	E				
1	0				
1	1				
0	1				
0	1				
1	1				
0	0				
0	0				
1	0	0			
1	0	1	1	0	0.1

$N = 4$

…ntil convergence

- Note: Bayesian networks can also have continuous variables
- Instead of conditional probability tables (CPTs), conditional probability densities (CPDs)
- Gaussian mixture model as a Bayesian network:



$$p(z = k) = \pi_k$$

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

EM in general

- Want to maximize the likelihood: $p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \boldsymbol{Z}|\boldsymbol{\theta})$
- Assume $p(\mathbf{X}|\boldsymbol{\theta})$ is difficult to optimize, while $p(\mathbf{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ is significantly easier to optimize
- Introduce a distribution $q(\mathbf{Z})$ and write

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

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

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

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

$$= \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})}}_{\mathcal{L}(q, \boldsymbol{\theta})} \underbrace{- \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})}}_{\mathrm{KL}(p \parallel q)}$$

KL-divergence between $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$ and $q(\mathbf{Z})$

$\mathrm{KL}(p \parallel q) \geq 0$

$\mathrm{KL}(p \parallel q) = 0$ iff $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) = q(\mathbf{Z})$

EM in general

- Want to maximize:

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \underbrace{\sum\nolimits_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})}}_{\mathcal{L}(q,\boldsymbol{\theta})} \underbrace{- \sum\nolimits_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{q(\mathbf{Z})}}_{\mathrm{KL}(p \parallel q)}$$

- Expectation step: maximize $\mathcal{L}(q, \boldsymbol{\theta}^{(\text{old})})$ w.r.t. q

  Equivalently, maximize $\log p\left(\mathbf{X}\middle|\boldsymbol{\theta}^{(\text{old})}\right) - \mathrm{KL}(p \parallel q)$ w.r.t. q

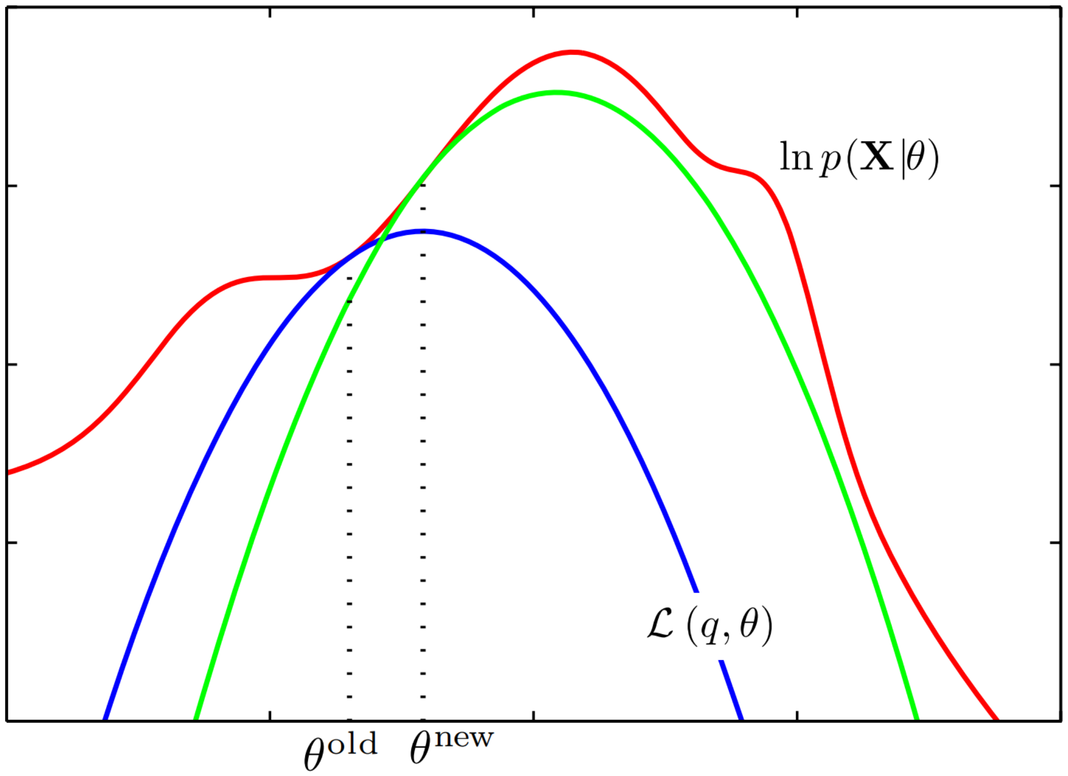
  Equivalently, minimize $\mathrm{KL}(p \parallel q)$

  I.e., set $q(\mathbf{Z}) = p\left(\mathbf{Z}\middle|\mathbf{X}, \boldsymbol{\theta}^{(\text{old})}\right)$

- Maximization step: maximize $\mathcal{L}(q, \boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$

  I.e., set $\boldsymbol{\theta}^{(\text{new})} = \operatorname{argmax}_{\boldsymbol{\theta}} \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$

EM in general



Probabilistic models: big picture

- Exploit constraints or structures (e.g. conditional independence for BNs) to more concisely represent distributions
- Use latent variables to learn expressive models composed of simple ones (e.g. GMMs)
- What can we do other than compute probability mass/density?
  - Generating new samples

    BN: sample parents first then children
  - Probabilistic reasoning: computing marginal & conditional probabilities; probability of having only one of symptoms A and B given a positive test?; most likely route given partial information about traffic conditions?

    In general, at least NP-hard other than special cases (e.g. Naïve Bayes, tree BN)
- Many more models with variety of expressiveness, inference efficiency, learnability
  - E.g. tractable probabilistic models, probabilistic programming, … (CSE 598 Spring 2023)