UVA CS 6316/4501 – Fall 2016 Machine Learning

Lecture 21: EM (Extra)

Dr. Yanjun Qi

University of Virginia

Department of
Computer Science

Where are we ? → major sections of this course

- ❑ Regression (supervised)
- ❑ Classification (supervised)
  - ❑ Feature selection
- ❑ Unsupervised models
  - ❑ Dimension Reduction (PCA)
  - ❑ Clustering (K-means, GMM/EM, Hierarchical)
- ❑ Learning theory
- ❑ Graphical models
  - ❑ (BN and HMM slides shared)

Today Outline

- Principles for Model Inference
  - Maximum Likelihood Estimation
  - ~~Bayesian Estimation~~

- Strategies for Model Inference
  - EM Algorithm – simplify difficult MLE
    - Algorithm
    - Application
    - Theory
  - ~~MCMC – samples rather than maximizing~~

Model Inference through Maximum Likelihood Estimation (MLE)

Assumption: the data is coming from a known probability distribution

The probability distribution has some parameters that are unknown to you

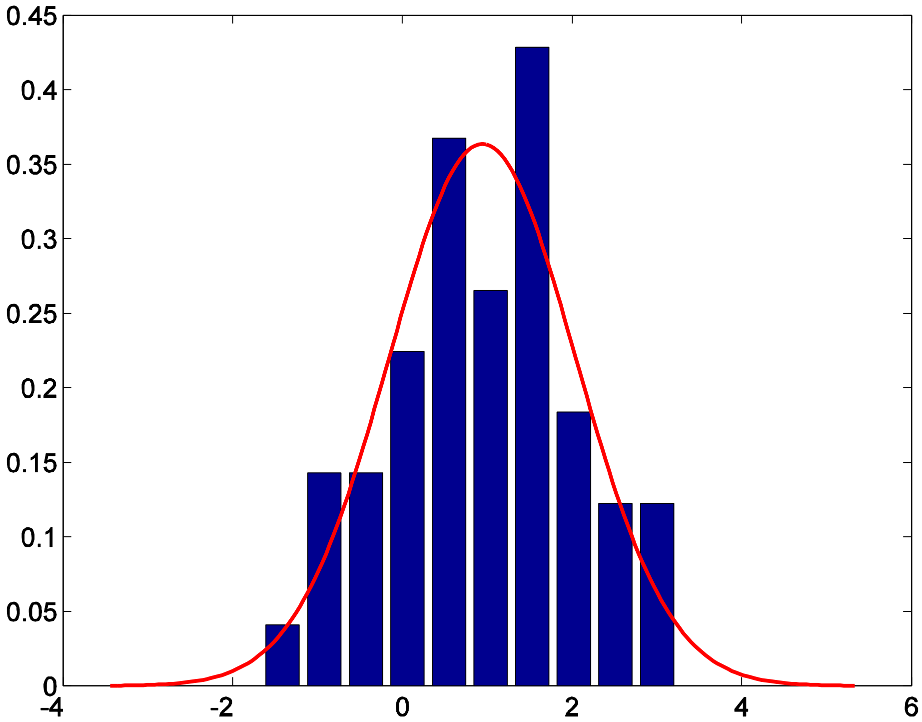
Example: data is distributed as Gaussian $y_i = N(\mu, \sigma^2)$,
so the unknown parameters here are $\theta = (\mu, \sigma^2)$

MLE is a tool that estimates the unknown parameters of the probability distribution from data

MLE: e.g. Single Gaussian Model (when p=1)

- Need to adjust the parameters (➔ model inference)

- So that the resulting distribution fits the observed data well



Maximum Likelihood revisited

$$y_i = N(\mu, \sigma^2)$$

$$Y = \{y_1, y_2, \ldots, y_N\}$$

$$l(\theta) = \log(L(\theta; Y)) = \log \prod_{i=1}^{N} p(y_i)$$

Choose θ that maximizes $l(\theta\ )$ ...

$$\frac{\partial l}{\partial \theta} = 0$$

MLE: e.g. Single Gaussian Model

- Assume observation data y_i are independent
- Form the Likelihood:

$$L(\theta;Y)=\prod_{i=1}^{N}p(y_i)=\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(y_i-\mu)^2}{2\sigma^2});$$

$$Y=\{y_1,y_2,\ldots,y_N\}$$

- Form the Log-likelihood:

$$l(\theta)=\log(\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma}}\exp(-\frac{(y_i-\mu)^2}{2\sigma^2}))=-\sum_{i=1}^{N}\frac{(y_i-\mu)^2}{2\sigma^2}-N\log(\sqrt{2\pi}\sigma)$$

MLE: e.g. Single Gaussian Model

- To find out the unknown parameter values, maximize the log-likelihood with respect to the unknown parameters:

Choose θ that maximizes $l(\theta)$ . . .

$$\frac{\partial l}{\partial \theta} = 0$$

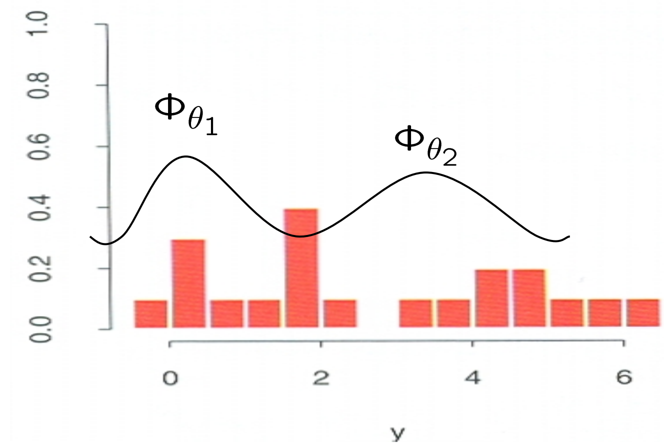
$$\frac{\partial l}{\partial \mu} = 0 \Rightarrow \mu = \frac{\sum_{i=1}^{N} y_i}{N}; \quad \frac{\partial l}{\partial \sigma^2} = 0 \Rightarrow \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2$$

MLE: A Challenging Mixture Example

$$Y_1 \sim N(\mu_1, \sigma_1^2); \quad Y_2 \sim N(\mu_2, \sigma_2^2)$$

$$Y = (1-\Delta)Y_1 + \Delta Y_2; \quad \Delta \in \{0,1\}$$

Indicator variable



histogram

marginal prob $\Rightarrow p(y \mid \overbrace{\mu_1, \sigma_1, \mu_2, \sigma_2, \pi}^{\theta})$

Mixture model: $g_Y(y) = (1-\pi)\Phi_{\theta_1}(y) + \pi\Phi_{\theta_2}(y)$ $\quad (\pi = Pr(\Delta = 1))$

$$\theta_1 = (\mu_1, \sigma_1); \quad \theta_2 = (\mu_2, \sigma_2)$$

π *is the probability with which the observation is chosen from density model 2*

(1- π) is the probability with which the observation is chosen from density 1

MLE: Gaussian Mixture Example

$p(y|\theta)$

$$g_Y(y) = (1-\pi)\Phi_{\theta_1}(y) + \pi\Phi_{\theta_2(y)} \qquad (\pi = Pr(\Delta = 1))$$

$\{y_1, y_2, \cdots, y_n\}$

Maximum likelihood fitting for parameters: $\theta = (\pi, \mu_1, \mu_2, \sigma_1, \sigma_2)$

$$l(\theta) = \sum_{i=1}^{N} log[(1-\pi)\Phi_{\theta_1}(y_i) + \pi\Phi_{\theta_2(y_i)}]$$

$$\frac{\partial l}{\partial \theta} = 0$$

Numerically (and of course analytically, too) Challenging to solve!!

Bayesian Methods & Maximum Likelihood

- Bayesian

  Pr(model|data) i.e. posterior

  =>Pr(data|model) Pr(model)

  => Likelihood * prior

  θ as random variable

- Assume prior is uniform, equal to MLE

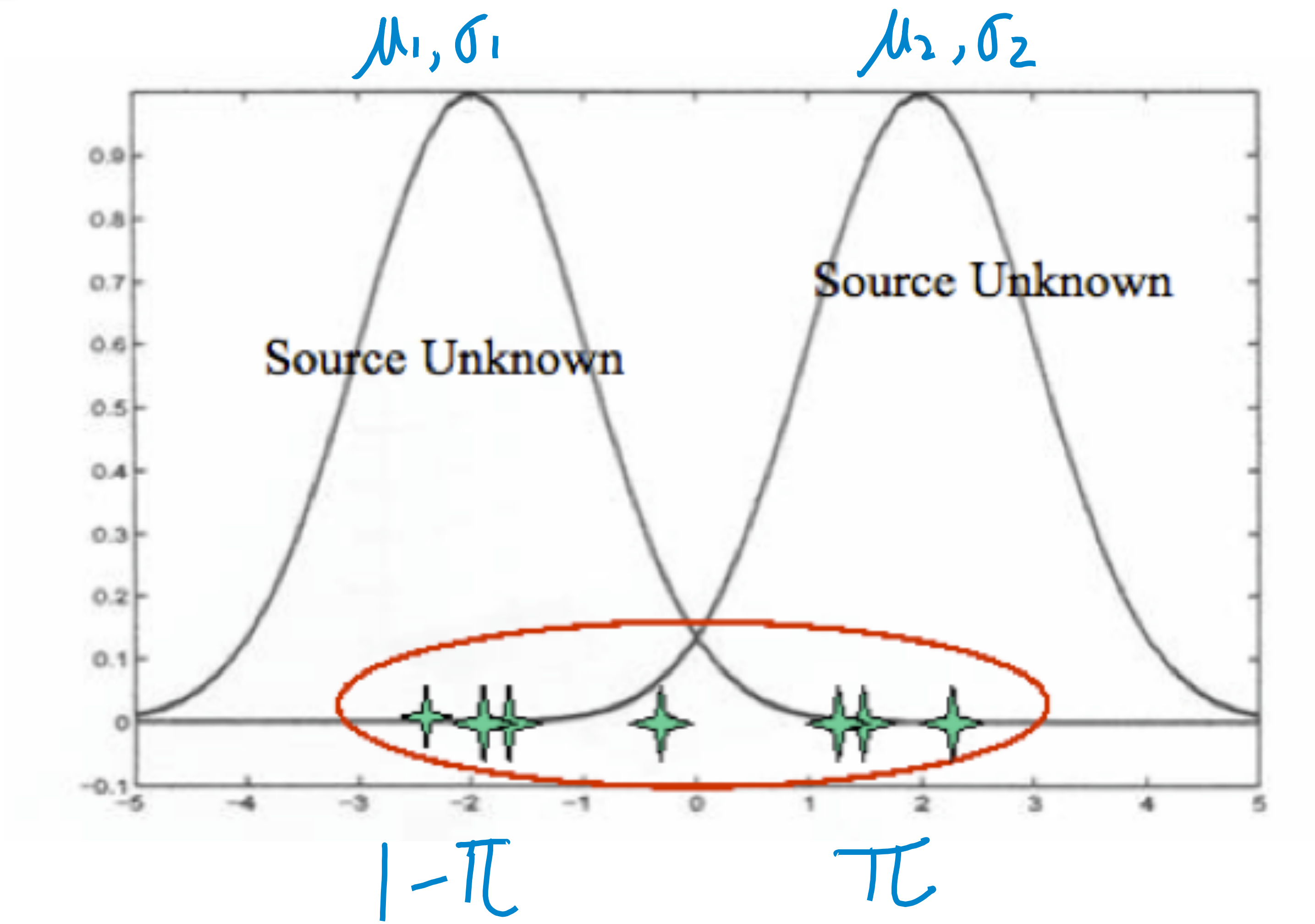
  $\arg\max_{model}$ Pr(data | model) Pr(model)

  = $\arg\max_{model}$ Pr(data | model)

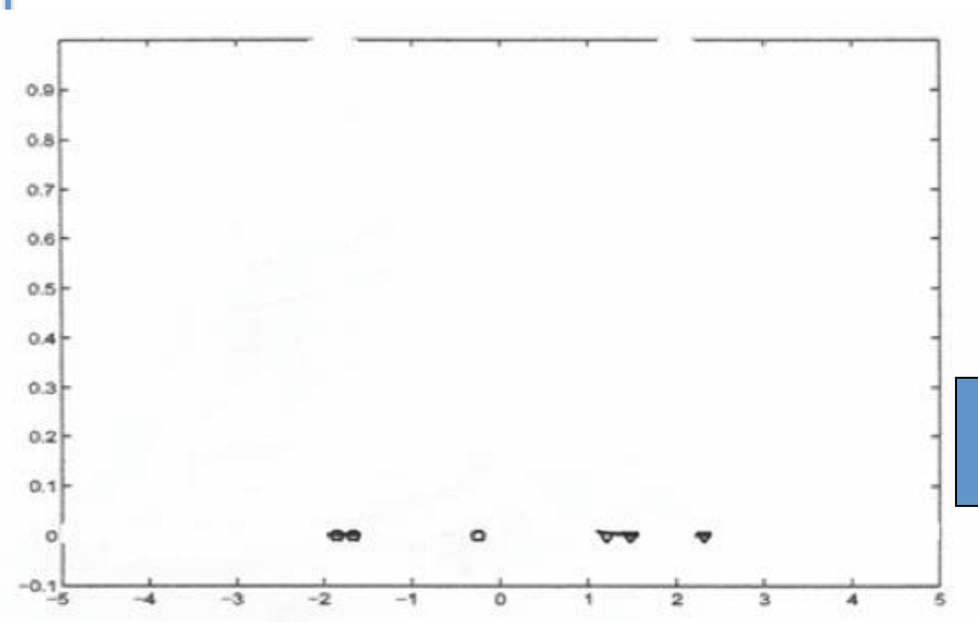
Today Outline

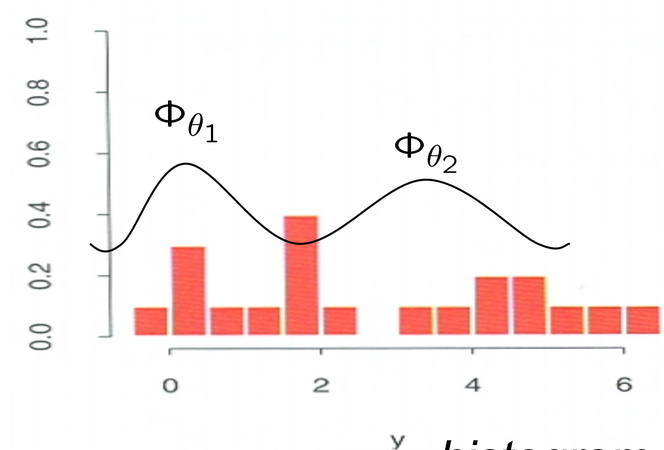
- Principles for Model Inference
  - Maximum Likelihood Estimation
  - ~~Bayesian Estimation~~
- Strategies for Model Inference
  - EM Algorithm – simplify difficult MLE
    - Algorithm
    - Application
    - Theory
  - ~~MCMC – samples rather than maximizing~~

Here is the problem



All we have is





From which we need to infer the likelihood function which generate the observations

Expectation Maximization: add latent variable Δ → latent data Δ_i

EM augments the data space– assumes with latent data

$\Delta_i \in 0, 1$ (latent data)

if($\Delta_i = 0$)

y_i was generated from first component

if($\Delta_i = 1$)

y_i was generated from second component

$\{y_1, y_2, \cdots, y_n\}$
$\{\Delta_1, \Delta_2, \cdots, \Delta_n\}$

Complete data: $t_i = (y_i, \Delta_i)$

$$p(t_i|\theta) = p(y_i, \Delta_i|\theta) = p(y_i|\Delta_i, \theta)Pr(\Delta_i)$$

$$p(t_i|\theta) = [\Phi_{\theta_1}(y_i)(1-\pi)]^{(1-\Delta_i)}[\Phi_{\theta_2}(y_i)\pi]^{\Delta_i}$$

Computing log-likelihood based on complete data

$$p(t_i|\theta) = [\Phi_{\theta_1}(y_i)(1-\pi)]^{(1-\Delta_i)}[\pi\Phi_{\theta_2}(y_i)\pi]^{\Delta_i}$$

$l_0(\theta; \mathbf{T})$ $\qquad T = \{t_i = (y_i, \Delta_i), i = 1 \ldots N\}$

$$= \sum_{i=1}^{N} (1-\Delta_i) log[(1-\pi)\Phi_{\theta_1}(y_i)] + \Delta_i log[\pi\Phi_{\theta_2}(y_i)]$$

$$= \sum_{i=1}^{N} (1-\Delta_i) log\Phi_{\theta_1}(y_i) + \Delta_i log\Phi_{\theta_2}(y_i)]$$
$$+ \sum_{i=1}^{N} [(1-\Delta_i) log(1-\pi) + \Delta_i log\pi \qquad (8.40)$$

only about π

Maximizing this form of log-likelihood is now tractable

Note that we cannot analytically maximize the previous log-likelihood with only observed Y={y_1, y_2, …, y_n}

EM: The Complete Data Likelihood

By simple differentiations we have:

$$\frac{\partial l_0}{\partial \mu_1} = 0 \Rightarrow \mu_1 = \frac{\sum_{i=1}^{N} (1-\Delta_i) y_i}{\sum_{i=1}^{N} (1-\Delta_i)};$$

$$\frac{\partial l_0}{\partial \sigma_1^2} = 0 \Rightarrow \sigma_1^2 = \frac{\sum_{i=1}^{N} (1-\Delta_i)(y_i - \mu_1)^2}{\sum_{i=1}^{N} (1-\Delta_i)};$$

So, maximization of the complete data likelihood is much easier!

How do we get the latent variables?

EM: The Complete Data Likelihood

By simple differentiations we have:

$$\frac{\partial l_0}{\partial \mu_2} = 0 \Rightarrow \mu_2 = \frac{\sum_{i=1}^{N} \Delta_i y_i}{\sum_{i=1}^{N} \Delta_i};$$

$$\frac{\partial l_0}{\partial \sigma_2^2} = 0 \Rightarrow \sigma_2^2 = \frac{\sum_{i=1}^{N} \Delta_i (y_i - \mu_2)^2}{\sum_{i=1}^{N} \Delta_i};$$

So, maximization of the complete data likelihood is much easier!

$$\frac{\partial l_0}{\partial \pi} = 0 \Rightarrow \pi = \frac{\sum_{i=1}^{N} \Delta_i}{N};$$

How do we get the latent variables?

Obtaining Latent Variables

The latent variables are computed as expected values given the data and parameters:

$\Delta_i \rightarrow$ $$\gamma_i(\theta) = E(\Delta_i \mid \theta, y_i) = \Pr(\Delta_i = 1 \mid \theta, y_i)$$

Apply Bayes' rule:

$$\gamma_i(\theta) = \Pr(\Delta_i = 1 \mid \theta, y_i) = \frac{\Pr(y_i \mid \Delta_i = 1, \theta)\Pr(\Delta_i = 1 \mid \theta)}{\Pr(y_i \mid \Delta_i = 1, \theta)\Pr(\Delta_i = 1 \mid \theta) + \Pr(y_i \mid \Delta_i = 0, \theta)\Pr(\Delta_i = 0 \mid \theta)}$$

$$= \frac{\Phi_{\theta_2}(y_i)\pi}{\Phi_{\theta_1}(y_i)(1-\pi) + \Phi_{\theta_2}(y_i)\pi}$$

$(y_i, \theta^{(t)}) \rightarrow E(\Delta_i)^{(t)}$

Dilemma Situation

- We need to know latent variable / data to maximize the complete log-likelihood to get the parameters
- We need to know the parameters to calculate the expected values of latent variable / data
- ➔ Solve through iterations

So we iterate → EM for Gaussian Mixtures…

1. Initialize parameters $\hat{\mu_1}, \hat{\sigma_1^2}, \hat{\mu_2}, \hat{\sigma_2^2}, \hat{\pi}$
2. Expectation Step:

$\{\theta^{(t)}, Y\} \Rightarrow E(\Delta_i)^{(t)}$

$$\gamma_i(\theta) = E(\Delta_i|\theta, Y) = Pr(\Delta_i = 1|\theta, Y)$$

By Bayes' theroem:

$$Pr(\Delta_i = 1|\theta, y_i) = \frac{p(y_i|\Delta_i{=}1,\theta).P(\Delta_i{=}1|\theta)}{p(y_i|\theta)}$$

$$= \frac{\Phi_{\hat{\theta_2}}(y_i).\hat{\pi}}{(1-\hat{\pi})\Phi_{\hat{\theta_1}}(y_i)+\hat{\pi}\Phi_{\hat{\theta_2}}(y_i)}$$

$$E[l_0(\theta;\mathbf{T}|Y,\hat{\theta}^{(j)})] = \sum_{i=1}^{N}[(1-\hat{\gamma}_i)log\Phi_{\theta_1}(y_i) + \hat{\gamma}_i log\Phi_{\theta_2}(y_i)] + \sum_{i=1}^{N}[(1-\hat{\gamma}_i)log(1-\pi) + \hat{\gamma}_i log\pi]$$

EM for Gaussian Mixtures…

3. Maximization Step:

$\{Y, E(\Delta_i)^{(t)}\} \Rightarrow \theta^{(t+1)}$

$$Q(\theta', \hat{\theta}^{(j)}) = E[l_0(\theta'; \mathbf{T}|Y, \hat{\theta}^{(j)})]$$

$$= \sum_{i=1}^{N} [(1-\hat{\gamma}_i) log \Phi_{\theta_1}(y_i) + \hat{\gamma}_i log \Phi_{\theta_2}(y_i)]$$
$$+ \sum_{i=1}^{N} [(1-\hat{\gamma}_i) log(1-\pi) + \hat{\gamma}_i log \pi]$$

Find θ' that maximizes $Q(\theta', \hat{\theta}^{(j)})$ ...

Set $\frac{\partial Q}{\partial \hat{\mu_1}}, \frac{\partial Q}{\partial \hat{\mu_2}}, \frac{\partial Q}{\partial \hat{\sigma_1}}, \frac{\partial Q}{\partial \hat{\sigma_2}}, \frac{\partial Q}{\partial \hat{\pi}} = 0$

to get $\hat{\theta}^{(j+1)}$

4. Use this $\hat{\theta}^{j+1}$ to compute the expected values $\hat{\gamma}_i$ and repeat…until convergence

EM for Two-component Gaussian Mixture

- Initialize $\mu_1, \sigma_1, \mu_2, \sigma_2, \pi$
- Iterate until convergence
  - Expectation of latent variables Δ

$$\gamma_i(\theta) = \frac{\Phi_{\theta_2}(y_i)\pi}{\Phi_{\theta_1}(y_i)(1-\pi) + \Phi_{\theta_2}(y_i)\pi} = \frac{1}{1 + \frac{1-\pi}{\pi}\frac{\sigma_2}{\sigma_1}\exp(-\frac{(y_i-\mu_1)^2}{2\sigma_1^2} + \frac{(y_i-\mu_2)^2}{2\sigma_2^2})}$$

  - Maximization for finding parameters θ

$$\mu_1 = \frac{\sum_{i=1}^{N}(1-\gamma_i)y_i}{\sum_{i=1}^{N}(1-\gamma_i)}; \quad \mu_2 = \frac{\sum_{i=1}^{N}\gamma_i y_i}{\sum_{i=1}^{N}\gamma_i}; \quad \sigma_1^2 = \frac{\sum_{i=1}^{N}(1-\gamma_i)(y_i-\mu_1)^2}{\sum_{i=1}^{N}(1-\gamma_i)}; \quad \sigma_2^2 = \frac{\sum_{i=1}^{N}\gamma_i(y_i-\mu_2)^2}{\sum_{i=1}^{N}\gamma_i}; \quad \pi = \frac{\sum_{i=1}^{N}\gamma_i}{N};$$

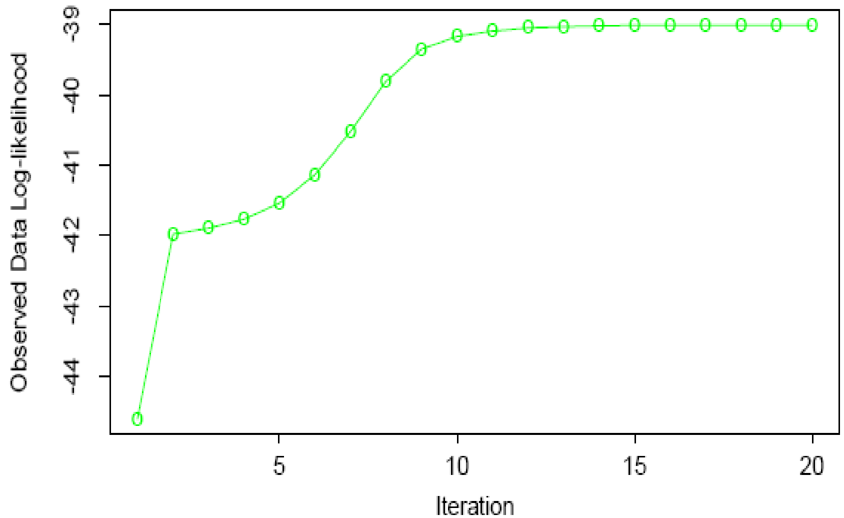
EM in....simple words

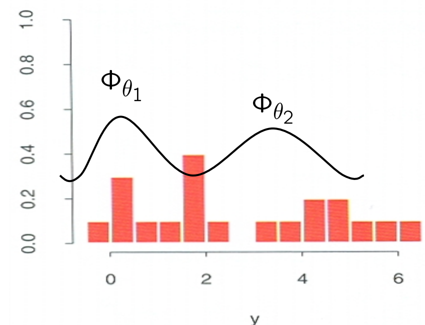
- Given observed data, you need to come up with a generative model
- You choose a model that comprises of some hidden variables Δ_i (this is your belief!)
- Problem: To estimate the parameters of model
  - Assume some initial values parameters
  - Replace values of hidden variable with their expectation (given the old parameters)
  - Recompute new values of parameters (given Δ_i)
  - Check for convergence using log-likelihood

1 stationary

2 until parameters stabalize ★★

EM – Example (cont'd)





histogram

Figure 8.6: *EM algorithm: observed data log-likelihood as a function of the iteration number.*

Selected iterations of the EM algorithm
For mixture example

Iteration	π
1	0.485
5	0.493
10	0.523
15	0.544
20	0.546

EM Summary

- An iterative approach for MLE
- Good idea when you have missing or latent data
- Has a nice property of convergence
- Can get stuck in local minima (try different starting points)
- Generally hard to calculate expectation over all possible values of hidden variables
- Still not much known about the rate of convergence

Today Outline

- Principles for Model Inference
  - Maximum Likelihood Estimation
  - ~~Bayesian Estimation~~

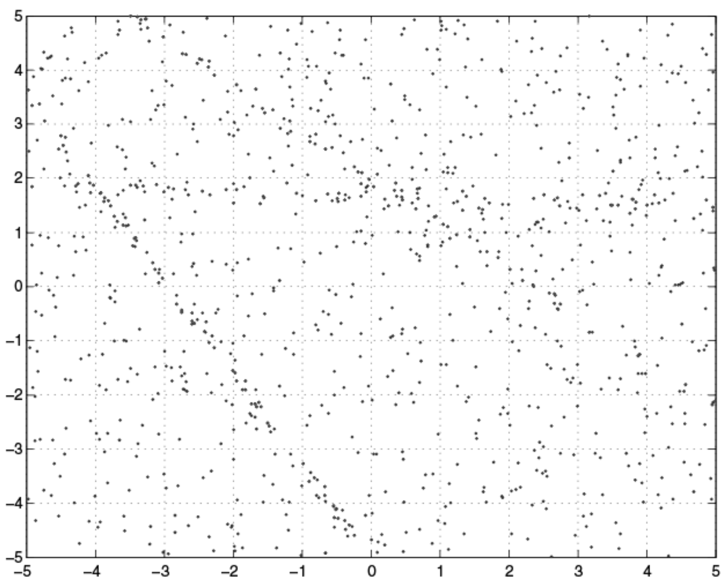
- Strategies for Model Inference
  - EM Algorithm – simplify difficult MLE
    - Algorithm
    - Application
    - Theory
  - ~~MCMC – samples rather than maximizing~~

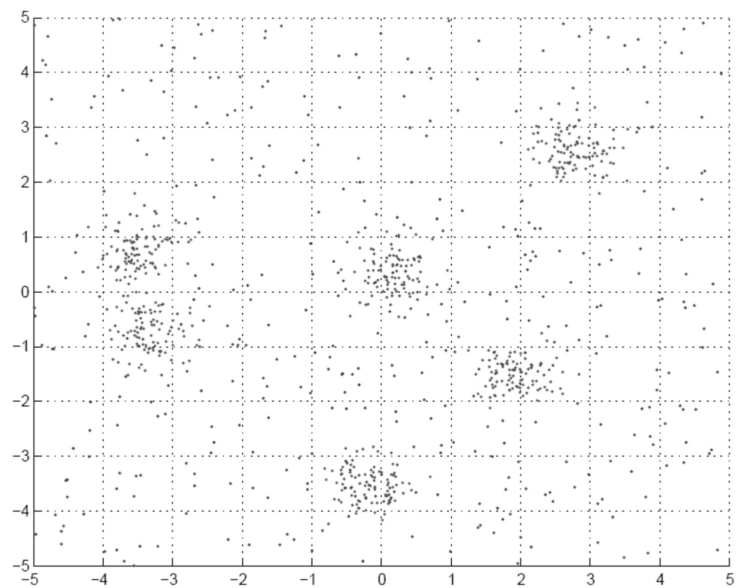
Applications of EM

- Mixture models
- HMMs
- Latent variable models
- Missing data problems
- …

Applications of EM (1)

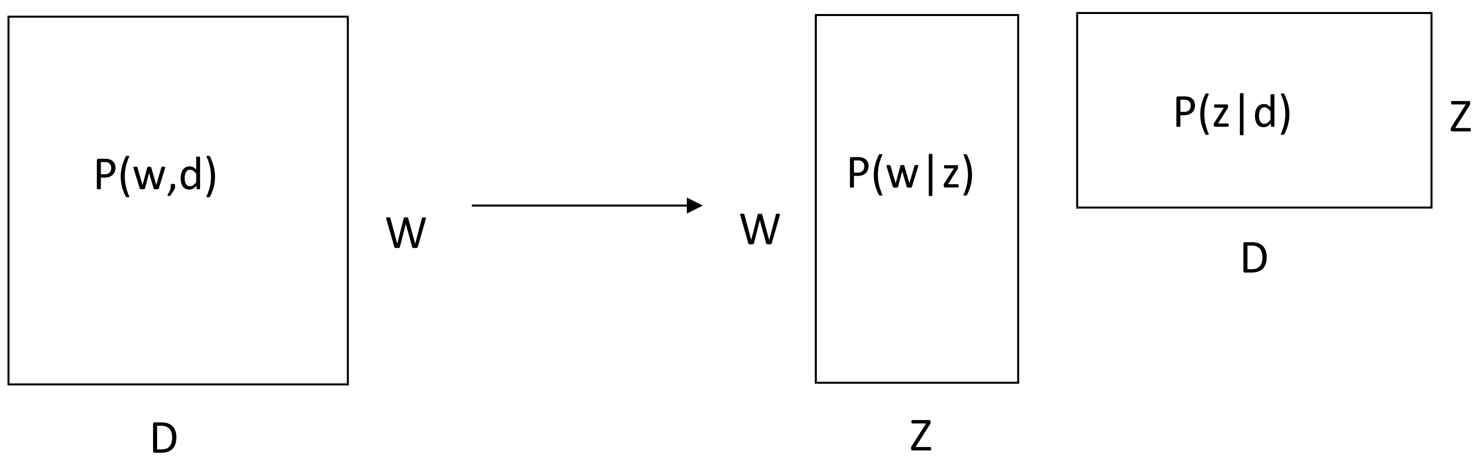
- Fitting mixture models





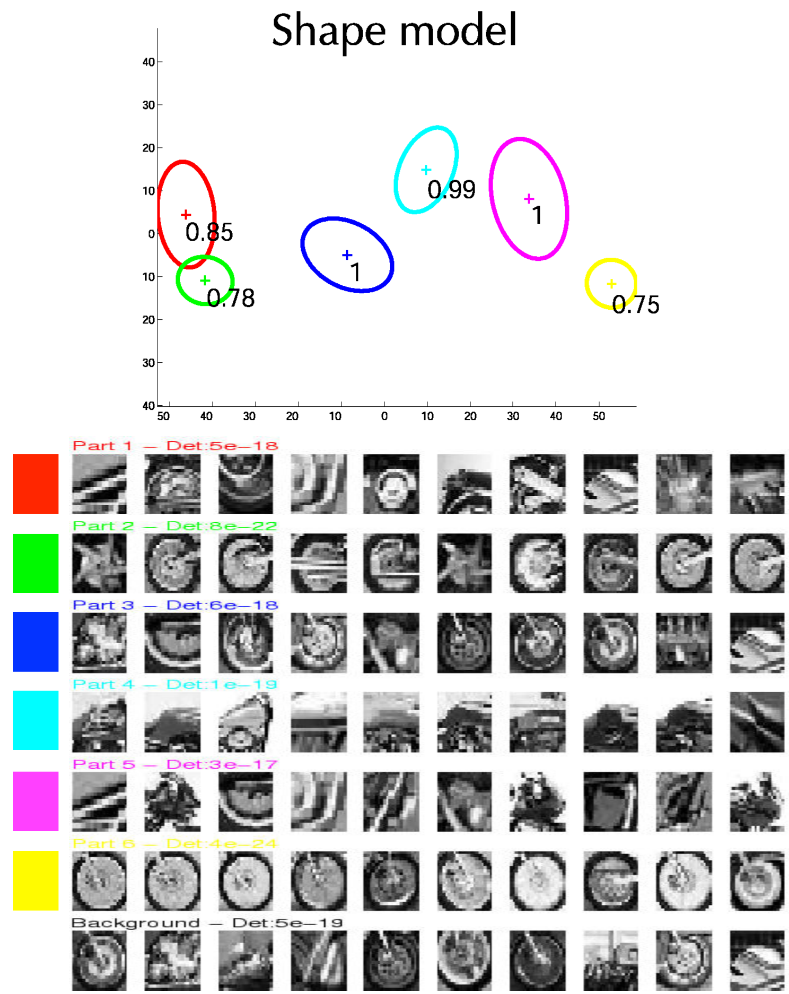
Applications of EM (2)

- Probabilistic Latent Semantic Analysis (pLSA)
  - Technique from text for topic modeling



Applications of EM (3)

- Learning parts and structure models



Applications of EM (4)

- Automatic segmentation of layers in video

http://www.psi.toronto.edu/images/figures/cutouts_vid.gif

Expectation Maximization (EM)

- Old idea (late 50' s) but formalized by Dempster, Laird and Rubin in 1977

- Subject of much investigation. See McLachlan & Krishnan book 1997.

single-
variable

+

two-
cluster
case

page 10 $\quad \pi = P(\Delta = 1)$

Joint Prob. Model :

① $p(y_i, \Delta_i | \theta) = p(y_i | \Delta_i, \theta)\, \underline{p(\Delta_i)} \quad \begin{cases} \Delta_i = 1 \\ \Delta_i = 0 \end{cases}$

$= [N(y_i | \mu_1, \sigma_1)(1-\pi)]^{1-\Delta_i} [N(y_i | \mu_2, \sigma_2)\, \pi]^{\Delta_i}$

② $\boxed{\text{Marginal}}$ Prob.

$p(y_i | \theta) = \sum_{\Delta_i} p(y_i | \Delta_i, \theta)\, p(\Delta_i)$

$= N(y_i | \mu_1, \sigma_1)(1-\pi) + N(y_i | \mu_2, \sigma_2)\, \pi$

③ $\boxed{\text{conditional}}$

$\Rightarrow p(y_i | \Delta_i, \theta) = \begin{cases} \Delta_i = 1 & N(y_i | \mu_1, \sigma_1) \\ \Delta_i = 0 & N(y_2 | \mu_2, \sigma_2) \end{cases}$

Estep

$\Rightarrow p(\Delta_i = 1 | y_i, \theta) = \dfrac{Pr(y_i | \Delta_i = 1)\, Pr(\Delta_i = 1 | \theta)}{p(y_i | \theta)}$

$=$

multi-variable

$+$

multi-cluster case

multi-variate $\Rightarrow$ Given $(x_1, x_2, \cdots, x_n)$

multi-cluster $\Rightarrow$ complete $(\vec{z}_1, \vec{z}_2, \cdots, \vec{z}_n)$

with

each vector $\vec{z}_i = (0, 0, 0, \cdots, \underbrace{1}_{j\text{th position}}, 0, 0, 0)_K$ $\longrightarrow z_i^{(j)} = 1 \Longrightarrow z_i^{(j)} = 1$ BasisVector

$\Rightarrow$ parameters $\vec{\theta}$ includes

$\begin{cases} \{\mu_j, \Sigma_j\}, & j = 1, 2, \cdots, K \\ \vec{\pi} \text{ vector}, & \pi_j = p(z^{(j)} = 1) \end{cases}$

$\boxed{\text{s.t. } \sum_{j=1}^{K} \pi_j = 1}$

① Joint Prob.

$$p(x_i, \vec{z}_i | \theta) = \prod_{j=1}^{K} \left[\pi_j N(x_i | \mu_j, \Sigma_j)\right]^{z_i^{(j)}}$$

$$p(x_i, z_i^{(j)} = 1 | \theta) = \pi_j N(x_i | \mu_j, \Sigma_j)$$

② Marginal

$$p(x_i | \theta) = \sum_{j=1}^{K} \pi_j N(x_i | \mu_j, \Sigma_j)$$

③ Conditional

$$p(z_i^{(j)} = 1 | x_i, \mu_j, \Sigma_j) \overset{\text{Bayes Rule}}{=\!=\!=} \frac{\pi_j N(x_i | \mu_j, \Sigma_j)}{\sum_{k=1}^{K} \pi_k N(x_i | \mu_k, \Sigma_k)}$$

Today Outline

- Principles for Model Inference
  - Maximum Likelihood Estimation
  - ~~Bayesian Estimation~~
- Strategies for Model Inference
  - EM Algorithm – simplify difficult MLE
    - Algorithm
    - Application
    - Theory
  - ~~MCMC – samples rather than maximizing~~

Why is Learning Harder?

- In fully observed iid settings, the complete log likelihood decomposes into a sum of local terms.

θ_z θ_x

$$\ell_c(\theta;D)=\log p(\underbrace{x,z}|\theta)=\log \underbrace{p(z|\theta_z)}+\log \underbrace{p(x|z,\theta_x)}$$

- When with latent variables, all the parameters become coupled together via *marginalization*

$$l(\theta;D)=\log p(x|\theta)=\log\sum_z p(z|\underbrace{\theta_z}_{\pi})p(x|z,\underbrace{\theta_x}_{\mu_i,\ \sigma_i})$$

Gradient Learning for mixture models

- We can learn mixture densities using gradient descent on the observed log likelihood. The gradients are quite interesting:

$$\ell(\theta) = \log p(\mathrm{x} \mid \theta) = \underbrace{\log \sum_k \pi_k p_k(\mathrm{x}|\theta_k)}$$

$$\begin{aligned}
\frac{\partial \ell}{\partial \theta} &= \frac{1}{p(\mathrm{x} \mid \theta)} \sum_k \pi_k \frac{\partial p_k(\mathrm{x}|\theta_k)}{\partial \theta} \\
&= \sum_k \frac{\pi_k}{p(\mathrm{x} \mid \theta)} p_k(\mathrm{x}|\theta_k) \frac{\partial \log p_k(\mathrm{x}|\theta_k)}{\partial \theta} \\
&= \sum_k \underbrace{\pi_k \frac{p_k(\mathrm{x}|\theta_k)}{p(\mathrm{x} \mid \theta)}} \frac{\partial \log p_k(\mathrm{x}|\theta_k)}{\partial \theta_k} = \boxed{\sum_k r_k \frac{\partial \ell_k}{\partial \theta_k}}
\end{aligned}$$

- In other words, the gradient is the responsibility weighted sum of the individual log likelihood gradients.
- Can pass this to a conjugate gradient routine.

Parameter Constraints

- Often we have constraints on the parameters, e.g. Σ_k being symmetric positive definite.
- We can use constrained optimization, or we can re-parameterize in terms of unconstrained values.
  - For normalized weights, softmax to e.g. $\sum_{j=1}^{K} \pi_j = 1$
  - For covariance matrices, use the Cholesky decomposition:

$$\Sigma^{-1} = \mathbf{A}^T \mathbf{A}$$

where A is upper diagonal with positive diagonal:

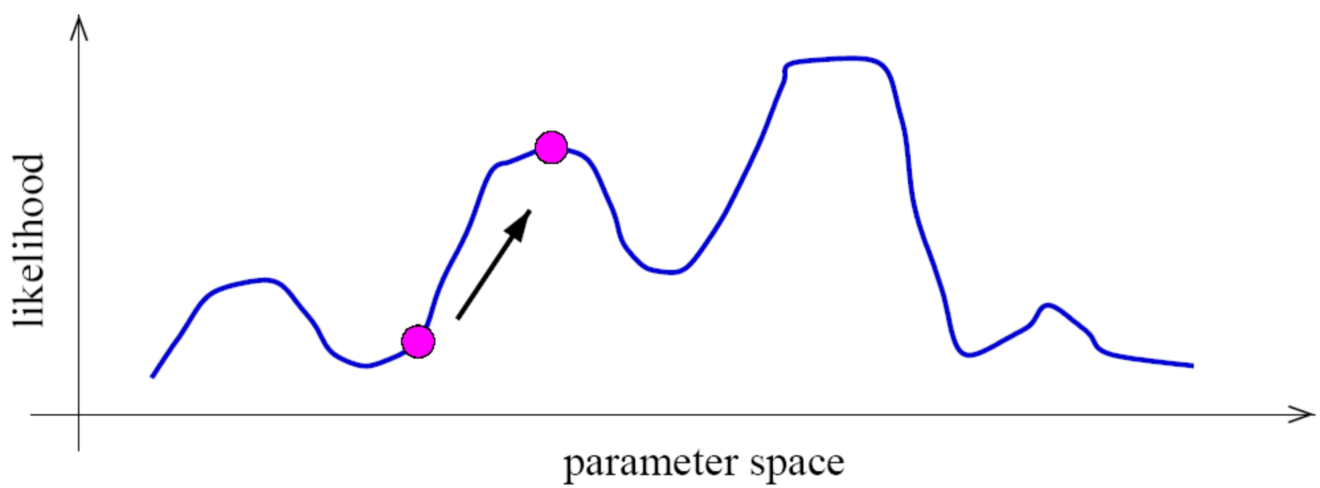
$$\mathbf{A}_{ii} = \exp(\lambda_i) > 0 \quad \mathbf{A}_{ij} = \eta_{ij} \ (j > i) \quad \mathbf{A}_{ij} = 0 \ (j < i)$$

  - Use chain rule to compute

$$\frac{\partial \ell}{\partial \pi}, \frac{\partial \ell}{\partial \mathbf{A}}.$$

Identifiability

- A mixture model induces a multi-modal likelihood.
- Hence gradient ascent can only find a local maximum.
- Mixture models are unidentifiable, since we can always switch the hidden labels without affecting the likelihood.
- Hence we should be careful in trying to interpret the "meaning" of latent variables.



Expectation-Maximization (EM) Algorithm

- EM is an Iterative algorithm with two linked steps:
  - E-step: fill-in hidden values using inference: $p(z|x, \theta^t)$.
  - M-step: update parameters (t+1) rounds using standard MLE/MAP method applied to completed data
- We will prove that this procedure monotonically improves (or leaves it unchanged). Thus it always converges to a local optimum of the likelihood.

Theory underlying EM

- What are we doing?

- Recall that according to MLE, we intend to learn the model parameter that would have maximize the likelihood of the data.

- But we do not observe z, so computing

$$\ell_c(\theta; D) = \log \sum_z p(x, z \mid \theta) = \log \sum_z p(z \mid \theta_z) p(x \mid z, \theta_x)$$

is difficult!

- What shall we do?

(1) Incomplete Log Likelihoods

- Incomplete log likelihood

With z unobserved, our objective becomes the log of a marginal probability:

– **This objective won't decouple**

$$l(\theta;x) = \log \underbrace{p(x|\theta)}_{\text{marginal given observed } x} = \log \sum_z p(x,z|\theta)$$

[One sample]

(2) Complete Log Likelihoods

[a random quantity]

- Complete log likelihood

  Let X denote the observable variable(s), and Z denote the latent variable(s).
  If Z could be observed, then

  Joint Prob.

$$l_c(\theta; x, z) \stackrel{\text{def}}{=} \underbrace{\log p(x, z \mid \theta)} = \log p(z \mid \theta_z) p(x \mid z, \theta_x)$$

  - Usually, optimizing $l_c()$ given both z and x is straightforward (c.f. MLE for fully observed models).
  - Recalled that in this case the objective for, e.g., MLE, decomposes into a sum of factors, the parameter for each factor can be estimated separately.
  - **But given that Z is not observed, $l_c()$ is a random quantity, cannot be maximized directly.**

Three types of log-likelihood

over multiple observed samples (x_1, x_2, …, x_N)

Observed data	$x = (x_1, x_2, \ldots, x_N)$
Latent variables	$z = (z_1, z_2, \ldots, z_N)$
Iteration index	t

$E_q[f(z)] = \sum_z q(z) f(z)$

Log-likelihood [Incomplete log-likelihood (ILL)]

$$l(\theta; x) = log\, p(x|\theta) = log\, \prod_x p(x|\theta)$$
$$= \sum_x log\, \sum_z p(x, z|\theta)$$

Complete log-likelihood (CLL)

$$l_c(\theta; x, z) \triangleq \sum_x \log p(x, z \mid \theta)$$

$z \sim q(z|x,\theta)$

Expected complete log-likelihood (ECLL)

$$E_q[f(z)] = \langle l_c(\theta; x, z)\rangle_q \triangleq \sum_{\substack{x \\ x_1, x_2, \cdots, x_n}} \sum_z q(z \mid x, \theta) \log p(x, z \mid \theta)$$

(3) Expected Complete Log Likelihood

- For ***any*** distribution $q(z)$, define *expected complete log likelihood (ECLL)*:
  - CLL is random variable ➔ ECLL is a deterministic function of *q*
  - Linear in CLL() --- inherit its factorizabiility
  - Does maximizing this surrogate yield a maximizer of the likelihood?

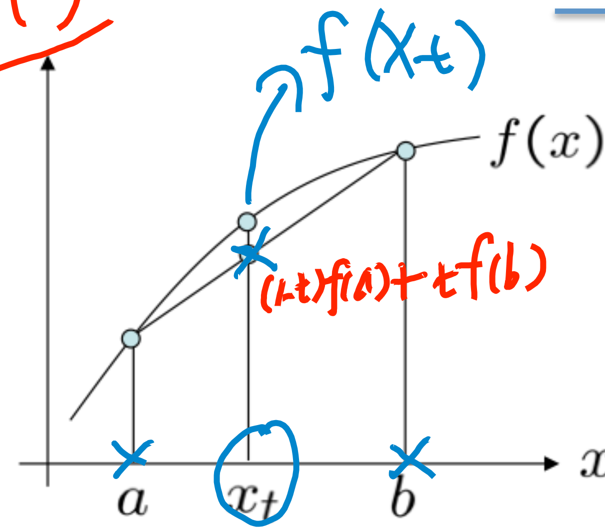
$$ECLL = \langle l_c(\theta; x, z) \rangle_q \stackrel{\text{def}}{=} \sum_z q(z \mid x, \theta) \log p(x, z \mid \theta)$$

Jensen's inequality

Concave func $f(x)$ e.g. $\log(\ )$

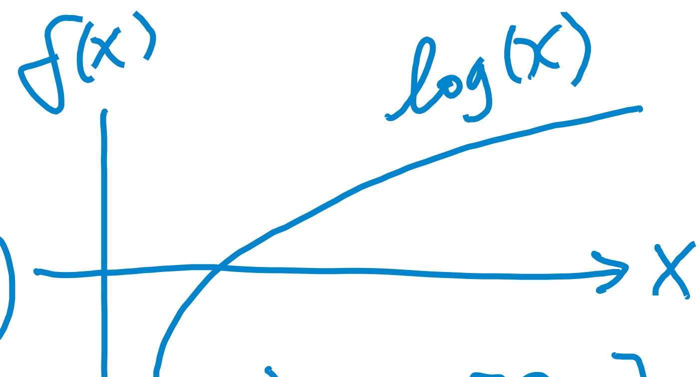
$x_t = (1-t)a + tb$

$$\Rightarrow f(x_t) \geq (1-t) f(a) + t f(b)$$



$$\Rightarrow f\left(\sum_{j=1}^{M} \lambda_j x_j\right) \geq \sum_{j=1}^{M} \lambda_j f(x_j)$$

$\sum \lambda_j = 1$



$$f(E[X]) \geq E[f(X)]$$

Jensen's inequality

- Jensen's inequality

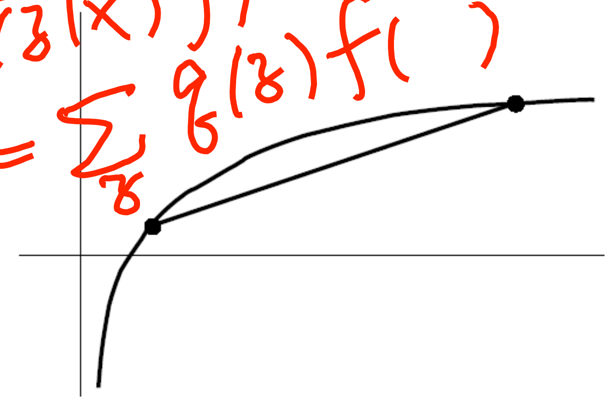
$$ECLL = \langle l_c(\theta; x, z) \rangle_q \stackrel{\text{def}}{=} \sum_z q(z|x,\theta) \log p(x,z|\theta)$$

$f = \log(\;)$

$z \sim q(z|x)$

$$\begin{aligned}
ILL = l(\theta;x) &= \log p(x|\theta) \\
&= \log \sum_z p(x,z|\theta) \\
&= \log \sum_z q(z|x) \frac{p(x,z|\theta)}{q(z|x)} \Rightarrow f\left(E_q\left[\frac{p(x,z|\theta)}{q(z|x)}\right]\right) \\
\text{Jensen's} \quad &\geq \sum_z q(z|x) \log \frac{p(x,z|\theta)}{q(z|x)} \Rightarrow E_q[f(\;)] = \sum_z q(z) f(\;) \\
&= \sum_z q(z|x) \log p(x,z|\theta) - \sum_z q(z|x) \log q(z|x) \\
&= ECLL + H_q
\end{aligned}$$

Entropy term



$$\Rightarrow \quad \ell(\theta;x) \geq \langle \ell_c(\theta;x,z) \rangle_q + H_q$$

$ILL \geq ECLL + H_q$

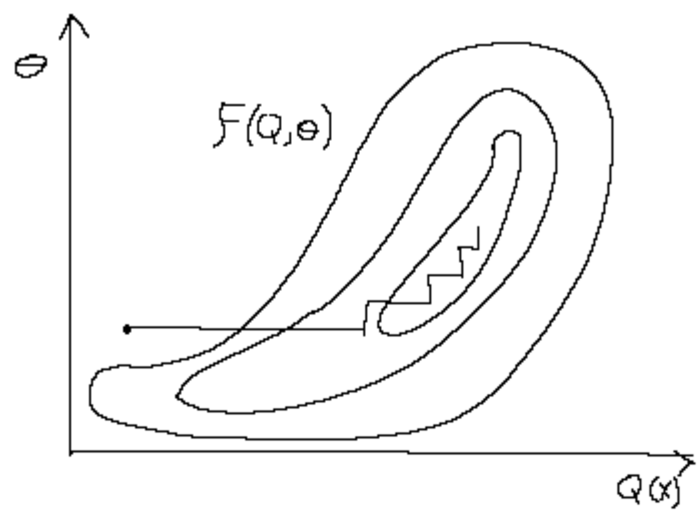
Lower Bounds and Free Energy

- For fixed data x, define a functional called the free energy:

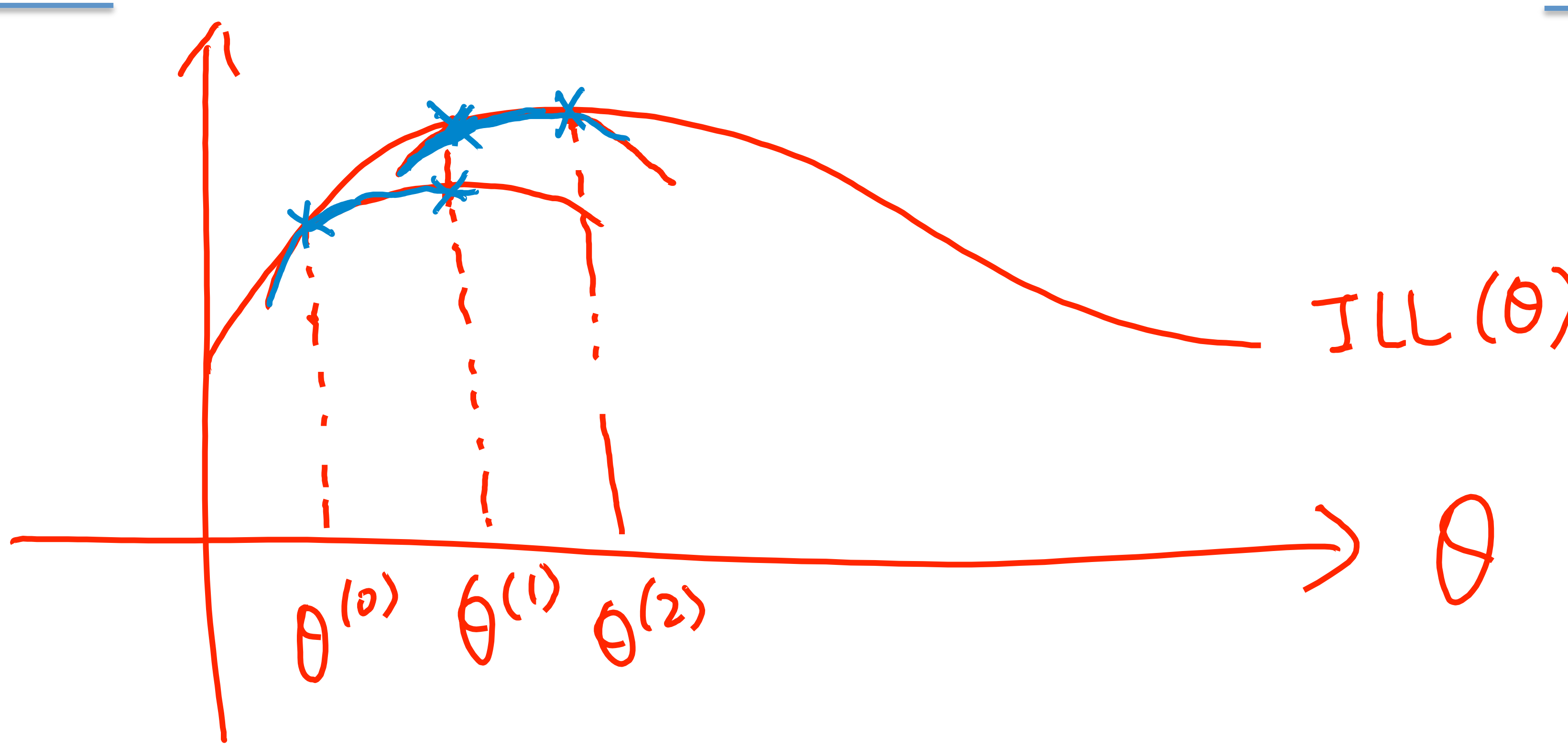
$E_{q(z)} f(\;(z))$

$$F(q,\theta) \overset{\text{def}}{=} \sum_z q(z \mid x) \log \frac{p(x,z \mid \theta)}{q(z \mid x)} \leq \ell(\theta; x)$$

- The EM algorithm is coordinate-ascent on F :
  - **E-step:** $q^{t+1} = \arg\max_q F(q, \theta^t)$
  - **M-step:** $\theta^{t+1} = \arg\max_\theta F(q^{t+1}, \theta^t)$



How EM optimize ILL ?



E-step: maximization of w.r.t. q

- Claim:

$$q^{t+1} = \arg\max_q F(q, \theta^t) = p(z \mid x, \theta^t)$$

  - This is the posterior distribution over the latent variables given the data and the parameters. Often we need this at test time anyway (e.g. to perform clustering).

- Proof (easy): this setting attains the bound of ILL

$$\begin{aligned} F(p(z|x,\theta^t),\theta^t) &= \sum_z p(z|x,\theta^t) \log \frac{p(x,z \mid \theta^t)}{p(z|x,\theta^t)} \\ &= \sum_z p(z|x,\theta^t) \log p(x \mid \theta^t) \\ &= \log p(x \mid \theta^t) = \ell(\theta^t; x) \end{aligned}$$

ILL

- Can also show this result using variational calculus or the fact that

$$\ell(\theta; x) - F(q, \theta) = \mathrm{KL}\big(q \,\|\, p(z \mid x, \theta)\big)$$

E-step: Alternative derivation

$\Rightarrow \ell(\theta;x) \geq F(q,\theta)$

$$\ell(\theta;x) - F(q,\theta) = \mathrm{KL}\big(q \,\|\, p(z \mid x,\theta)\big)$$

$$= l(\theta;x) - \sum_z q(z \mid x) \log \frac{p(x,z \mid \theta)}{q(z \mid x)}$$

$$= \sum_z q(z \mid x) \log p(x \mid \theta) - \sum_z q(z \mid x) \log \frac{p(x,z \mid \theta)}{q(z \mid x)}$$

$\Rightarrow p(z \mid x,\theta)\,p(x,\theta)$

$$= \sum_z q(z \mid x) \log \frac{q(z \mid x)}{p(z \mid x,\theta)}$$

$$= D_{KL}(q(z \mid x) \,\|\, p(z \mid x,\theta)).$$

$\Rightarrow$ [$D_{KL} = 0$ iff $q = p$ almost everywhere]

M-step: maximization w.r.t. \theta

- Note that the free energy breaks into two terms:

$$F(q,\theta) = \sum_z q(z|x)\log\frac{p(x,z|\theta)}{q(z|x)}$$
$$= \sum_z q(z|x)\log p(x,z|\theta) - \sum_z q(z|x)\log q(z|x)$$
$$\boxed{= \langle \ell_c(\theta;x,z)\rangle_q + H_q}$$

ECLL + entropy$_q$

  – The first term is the expected complete log likelihood (energy) and the second term, which does not depend on *q*, is the entropy.

M-step: maximization w.r.t. \theta

- Thus, in the M-step, maximizing with respect to *q* for fixed *q* we only need to consider the first term:

ECLL

$$\theta^{t+1} = \arg\max_{\theta} \langle \ell_c(\theta; x, z) \rangle_{q^{t+1}} = \arg\max_{\theta} \sum_z q(z \mid x) \log p(x, z \mid \theta)$$

  - Under optimal q^{t+1}, this is equivalent to solving a standard MLE of fully observed model $p(x,z|q)$, with the sufficient statistics involving *z* replaced by their expectations w.r.t. $p(z|x,q)$.

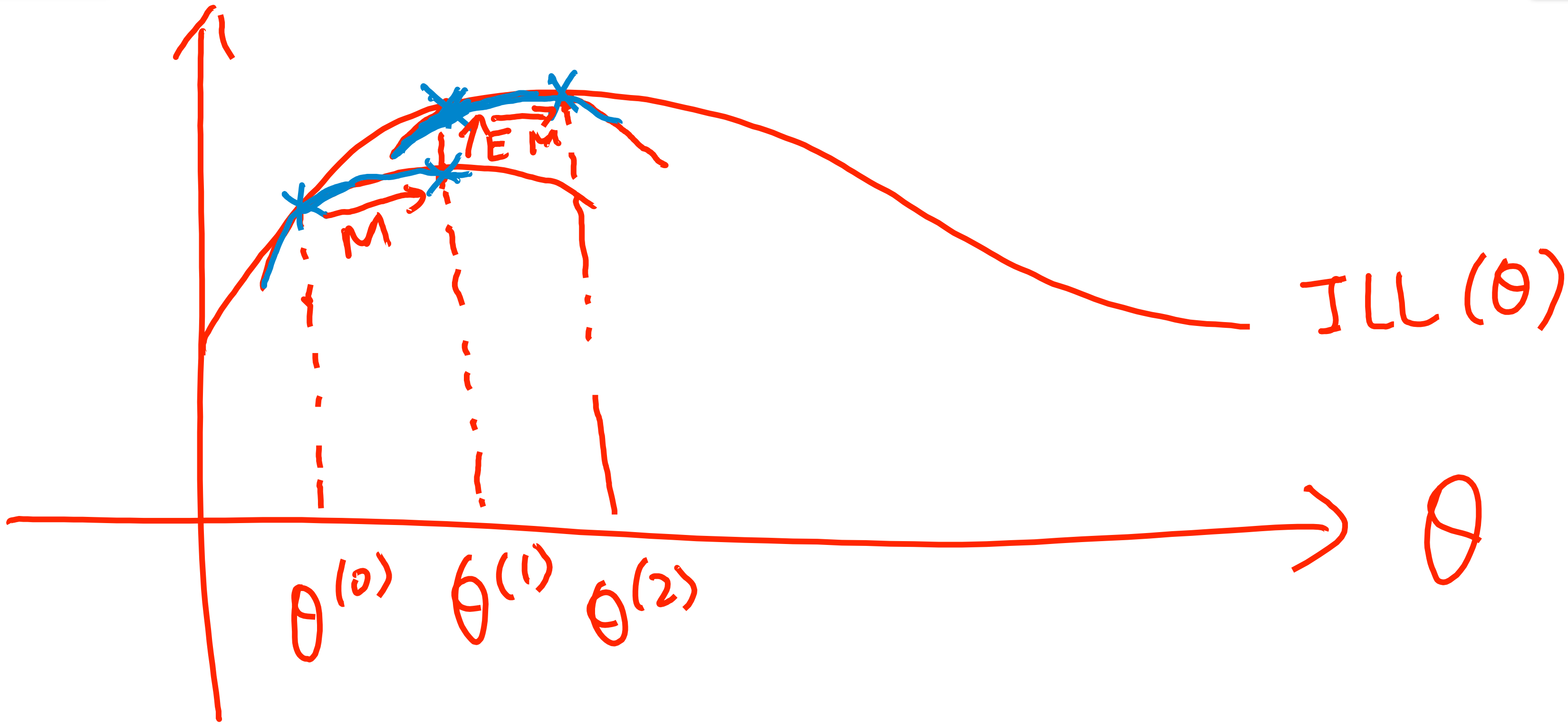
Summary: EM Algorithm

- A way of maximizing likelihood function for latent variable models. Finds MLE of parameters when the original (hard) problem can be broken up into two (easy) pieces:
  1. Estimate some "missing" or "unobserved" data from observed data and current parameters.
  2. Using this "complete" data, find the maximum likelihood parameter estimates.

- Alternate between filling in the latent variables using the best guess (posterior) and updating the parameters based on this guess:
  – E-step: $q^{t+1} = \arg\max_q F(q, \theta^t)$
  – M-step: $\theta^{t+1} = \arg\max_\theta F(q^{t+1}, \theta^t)$

- In the M-step we optimize a lower bound on the likelihood. In the E-step we close the gap, making bound=likelihood.

How EM optimize ILL ?



A Report Card for EM

- Some good things about EM:
  - no learning rate (step-size) parameter
  - automatically enforces parameter constraints
  - very fast for low dimensions
  - each iteration guaranteed to improve likelihood
  - Calls inference and fully observed learning as subroutines.

- Some bad things about EM:
  - can get stuck in local minima
  - can be slower than conjugate gradient (especially near convergence)
  - requires expensive inference step $\Rightarrow p(z|x,\theta)$
  - is a maximum likelihood/MAP method

References

- Big thanks to Prof. Eric Xing @ CMU for allowing me to reuse some of his slides
- **The EM Algorithm and Extensions** by Geoffrey J. MacLauchlan, Thriyambakam Krishnan