Review of Last Week

- Three classification models
  - Discriminant Model: learn the decision boundary directly and apply it to determine the class of each data point
  - Discriminative Model: learn P(Y|X) directly
  - Generative Model: Learn P(Y|X) through P(X|Y) and P(Y). The joint probability P(X,Y) and marginal probability P(X) can also be learned.
- SML vs. non-statistical ML
  - Objective function for SML: MLE or MAP,
  - Objective function for ML: MSE, maximizing the margin, or others.
- Unsupervised Learning
  - Clustering (group data together)
  - EM (learning given incomplete data, based-on MLE)

Revisit Bayesian: Prior vs. Smoothing Technique

Estimating the Likelihood

- Assuming we randomly observe N coin tosses to find head occurs H times, what is the frequency of head for this coin?
  - Assuming the probability is p, then according to **MLE** we want to optimize $\log(p^H(1-p)^{N-H})$
  - After performing derivatives on p, we can learn that the **MLE solution of p is H/N**
- However, the H/N model suffers a major drawback that **unseen** events will receive zero probability
  - If we toss a coin 6 times and find zero heads, H/N model tells us the probability of head is 0
  - However, unseen objects should receive a **tiny probability** (rather than zero), given the fact that we know they do exist.
- **Smoothing:** a technique to assign non-zero probability to unseen objects
  - Add-one smoothing: assuming everything occurs at least once.
  - Under this assumption, the frequency of W becomes (H+1)/(N+2), because both head and tail occurs once.
  - This is a commonly applied techniques for n-gram Langauge Model Learning

Add-one Smoothing vs. Bayesian Prior

- Last week after class, I received an email from a student in this class, Shao-Chuan Wang, saying that right after the class, he proved that add-one smoothing can be interpreted as an MAP solution for coin toss.
- Recall that the MLE solution aims at optimizing the likelihood probability $p^{H}(1-p)^{N-H}$, and that **max(posterior probability) = max(likelihood probability* Prior probability)**
- Assuming the prior probability is set to be proportional to p(1-p)
  - to prohibit assigning a very large or very small value to p
  - then the posterior probability becomes $p^{H+1}(1-p)^{N-H+1}$
  - optimizing the posterior w.r.t p will obtain p=(H+1)/(N+2).
- It can also be proved that add-lambda smoothing can be regarded as an MAP, given slightly different prior.
- Shao-Chuan's finding in fact tells us that this two basic smoothing techniques is simply a special case for Bayesian learning.

Unsupervised Learning II

Clustering

Prof. Shou-de Lin
CSIE/GINM, NTU
Sdlin@csie.ntu.edu.tw

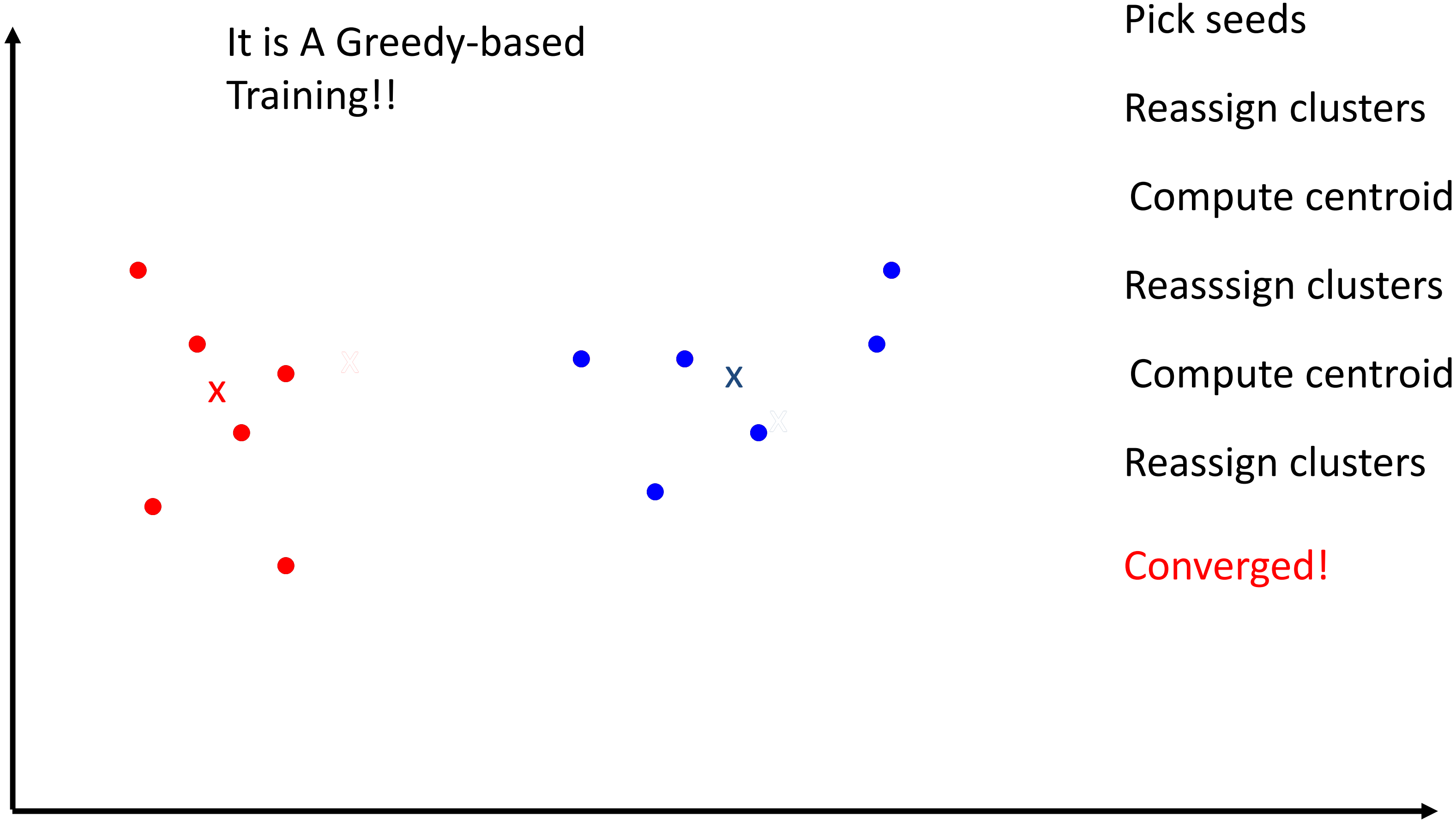
What is clustering ?

- **Clustering** is the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait.
- Difference between **clustering** and **classification**
  – Clustering: divide input into partitions (without label). It's unsupervised.
  – Classification: classify inputs into Y labeled classes (supervised)
- We will introduce two famous clustering algorithms
  – K-means clustering
  – EM clustering for Gaussian Mixture Model

Steps of K-means clustering

1. Randomly select k points as cluster center.
2. Assign the rest of the points to the cluster of its closest center
3. Re-calculating the mean point of each cluster.
4. Constructing a new partition by associating each point with the cluster whose centroid is the closest.
5. Go back to 3

K Means Example
(K=2)



Pick seeds

Reassign clusters

Compute centroids

Reasssign clusters

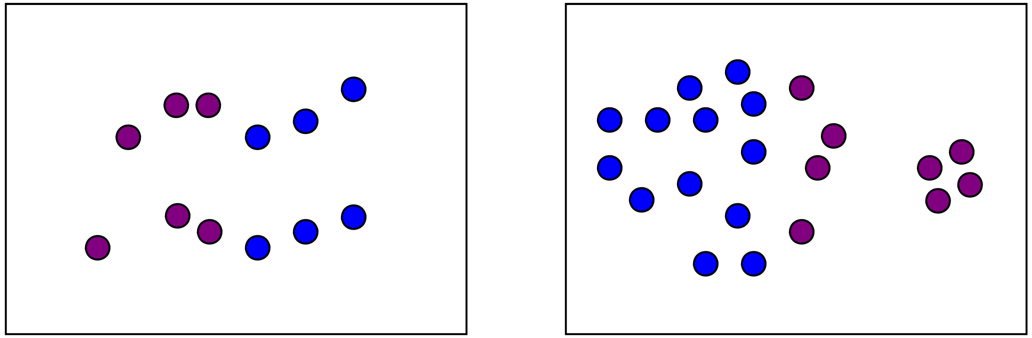
Compute centroids

Reassign clusters

Converged!

k-Means Clustering sometimes failed

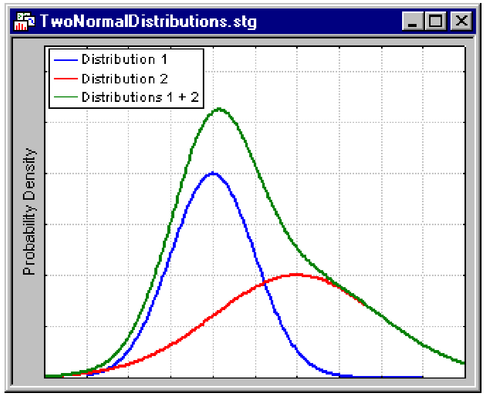
- Failure Cases:



- Viterbi (or greedy) Training got stuck to sub-optimal easily
  - If a node is equally close to several clusters, it can cause problems since we can only assign it to one class.
- Can we do better? Yes, using EM-clustering

EM clustering (for Gaussian Mixtures) Problem

- Suppose you measure a single continuous variable in a large sample of observations.
- Suppose the sample consists of several clusters of Gaussian observations with different means and variances.
- Our job is to determine the value of the 3k-1 parameters:
  - The mean and variance for cluster 1
  - The mean and variance for cluster 2
  - …
  - The mean and variance for cluster k
  - The sampling probability for cluster 1…k➔ π_k



Multivariate Gaussian Distribution

- Single value Gaussian Distribution:

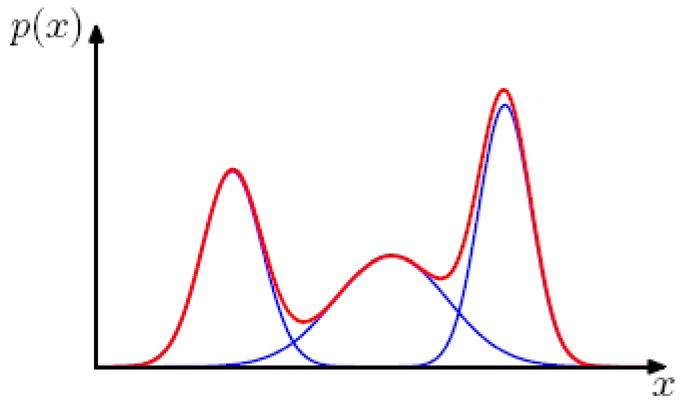
$$N(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$$

- Multivariate Gaussian Distribution:

***N*(x | μ,Σ)=** $\frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) \right\}$

Gaussian Mixture Distribution (GMD)

- A linear combination of several Gaussian Distributions
- It can model almost any continuous density given sufficient number of Gaussians.



- $$p(x) = \sum_{k=1}^{K} \pi_k \mathrm{N}(x \mid \mu_k, \Sigma_k),$$

prior likelihood

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

MLE for GMD

- Suppose we have a dataset D of observations $\{x_1, x_2, \ldots, x_N\}$, and we wish to model this data using a mixture of Gaussians.
- Then the log likelihood function is given by

$$\ln p(\mathrm{X} \mid \pi, \mu, \Sigma) = \ln\left\{ \prod_{n=1}^{N} \left(\sum_{k=1}^{K} \pi_k \mathrm{N}(x_n \mid \mu_k, \Sigma_k) \right) \right\}$$

$$= \sum_{n=1}^{N} \ln\left\{ \sum_{k=1}^{K} \pi_k \mathrm{N}(x_n \mid \mu_k, \Sigma_k) \right\}$$

MLE solution for μ

$$\ln p(\mathrm{X} \mid \pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathrm{N}(x_n \mid \mu_k, \Sigma_k) \right\}$$

$$\frac{\partial \ln p(\mathrm{X} \mid \pi, \mu, \Sigma)}{\partial \mu_k} = -\sum_{n=1}^{N} \boxed{\frac{\pi_k \mathrm{N}(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathrm{N}(x_n \mid \mu_j, \Sigma_j)}} \Sigma_k (x_n - \mu_k)$$

$P(x \in c_k \mid x) \equiv p(z_{nk})$

- Solve the above equation:

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} p(z_{nk}) x_n, \text{ and } N_k = \sum_{n=1}^{N} p(z_{nk})$$

The weighted mean of all the points in the dataset, in which the weight for data point x_n is given by the posterior probability that x belongs to a cluster c_k

Can be interpreted as the effective number of points assigned to cluster k

MLE solution for Σ

- If we find the zero partial derivatives of Σ, we will learn the MLE solution for Σ is

$$\Sigma_k = \frac{1}{N_k}\sum_{n=1}^{N} p(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T$$

MLE solution for π

- Since there is a constraint that the sum of π_k is 1, we apply Lagrange multiplier to maximize

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

- The derivatives of the above (w.r.t. π_k) equal 0 gives

$$\sum_{n=1}^{N} \frac{N(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n \mid \mu_j, \Sigma_j)} + \lambda = 0 \Rightarrow \sum_k \pi_k \sum_{n=1}^{N} \frac{N(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n \mid \mu_j, \Sigma_j)} + \sum_k \pi_k \lambda = 0, \lambda = -N$$

$$\sum_{n=1}^{N} \frac{\pi_k N(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n \mid \mu_j, \Sigma_j)} + \pi_k \lambda = 0 \Rightarrow N_k + \pi_k(-N) = 0 \Rightarrow \pi_k = \frac{N_k}{N}$$

How to Produce the MLE solutions?

- MLE solution for μ, Σ, and π cannot be easily obtained (i.e. no close-form solution) since $p(z_{nk})$ contains μ, Σ, and π
- Since $p(x_{nk})$ can be generated by μ, Σ, and π; and μ, Σ, and π can be generated by $p(x_{nk})$. We can treat $p(x_{nk})$ as a hidden variable and apply EM algorithm to iteratively learn the parameters.

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} p(z_{nk}) x_n,$$

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} p(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T$$

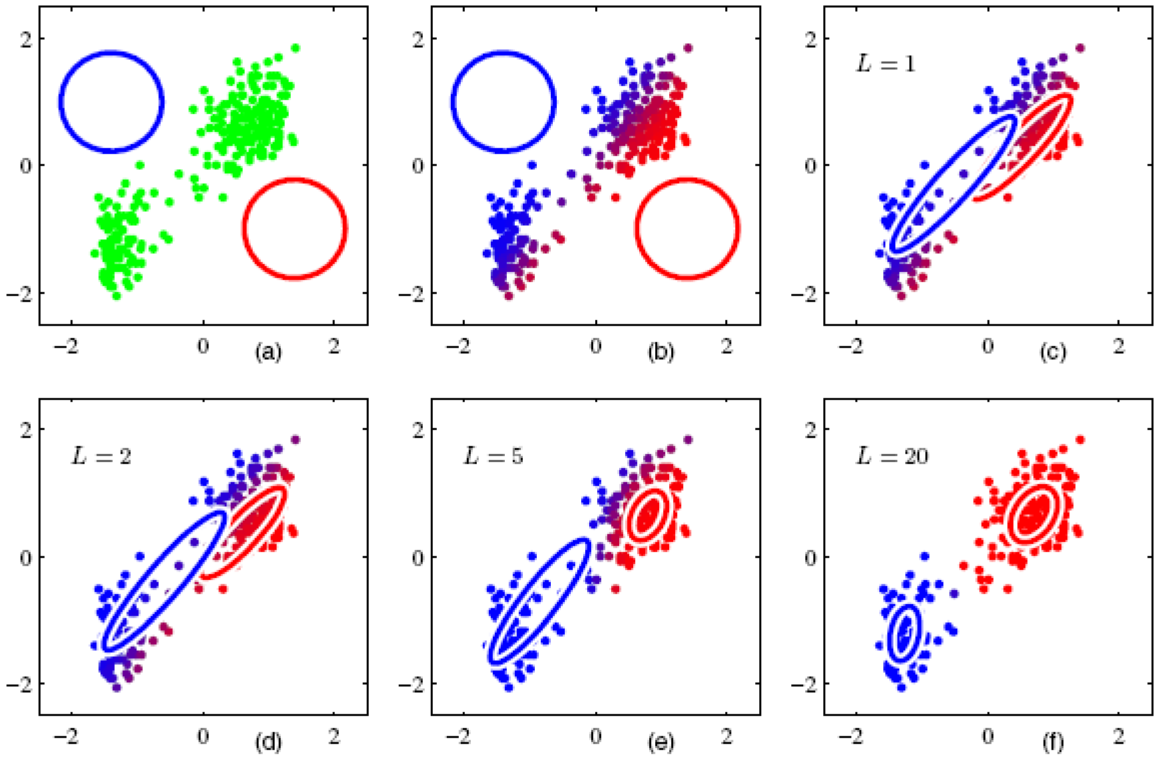
$$\pi_k = \frac{N_k}{N}, N_k = \sum_{n=1}^{N} p(z_{nk})$$

$$p(z_{nk}) = \frac{\pi_k \mathrm{N}(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathrm{N}(x_n \mid \mu_j, \Sigma_j)}$$

EM for Gaussian Mixtures Clustering

- Goal: Maximize the Likelihood function w.r.t. the parameters μ, Σ, and π
- Steps
  - Initialize the parameters μ, Σ, and π, and evaluate the initial value of the log likelihood $\ln p(X|\ \mu, \Sigma, \pi)$.
  - E step: generate the posterior probabilities using current parameters $$p(z_{nk}) = \frac{\pi_k \mathrm{N}(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathrm{N}(x_n \mid \mu_j, \Sigma_j)}$$
  - M-step. Re-estimate the parameters using the current posterior probabilities (see previous page).
  - Check the log likelihood given current parameters to see if they converge. If not, return to E-step.

Graphic Example of EM Clustering



EM Framework and EM Theory

Goal

- The goal of EM algorithm is to find maximum likelihood solutions for models that contain latent (or missing) variables. We represent the input data as X, and the latent variables as Z. The parameters of the model is represented as θ.
- {X,Z} is called the complete data and X is called the incomplete data.
- Assuming the posterior distribution $P(Z|X,\theta)$ can be generated given X and θ is known.
- The log likelihood function to maximize is $\ln p(X|\theta)$. Since Z is unknown, we can represent $\ln p(X|\theta)$ as $\ln \sum_{Z} p(X, Z \mid \theta)$

Problem: the summation over the latent variables appears inside the logarithm. Therefore even $P(X,Z|\theta)$ belongs to Gaussian, $P(X|\theta)$ will still be hard to compute.

The spirit of EM

1. Create an initial model, θ_0.
   – The initialization can be arbitrarily, randomly, or with a small set of training examples.
2. Use the existing model θ^{old} to obtain another model θ^{new} such that

$$\ln p(X \mid \theta^{new}) > \ln p(X \mid \theta^{old})$$

3. Repeat the above step until reaching a local maximum.
4. Challenge: How can we guaranteed to find a better model after each iteration given the hidden variable exists?

$$Ans: \theta^{new} = \arg\max_{\theta} \sum_{Z} p(Z \mid X, \theta^{old}) \ln p(X, Z \mid \theta)$$

EM Theorem

- If we can find a θ^{new} that guarantee

$$\sum_Z p(Z \mid X, \theta^{old}) \ln p(X, Z \mid \theta^{new}) > \sum_Z p(Z \mid X, \theta^{old}) \ln p(X, Z \mid \theta^{old})$$

then the same θ^{new} will also satisfy the condition

$$\ln p(X \mid \theta^{new}) > \ln p(X \mid \theta^{old})$$

- If EM theorem is true, then we can try to find

$$\theta^{new} = \arg\max_{\theta} \sum_Z p(Z \mid X, \theta^{old}) \ln p(X, Z \mid \theta)$$

Then such θ^{new} will lead to better $P(X|\theta)$

- How can we prove EM theorem?
  - If we can prove the equation below, then we are done

$$\ln p(X \mid \theta^{new}) - \ln p(X \mid \theta^{old}) \geq$$

$$\sum_Z p(Z \mid X, \theta^{old}) \ln p(X, Z \mid \theta^{new}) - \sum_Z p(Z \mid X, \theta^{old}) \ln p(X, Z \mid \theta^{old})$$

Proof of EM Theorem (1/2)

$$\ln p(X\mid\theta^{new}) - \ln p(X\mid\theta^{old}) \ge$$

$$\boxed{\sum_Z p(Z\mid X,\theta^{old})\ln p(X,Z\mid\theta^{new}) - \sum_Z p(Z\mid X,\theta^{old})\ln p(X,Z\mid\theta^{old})}$$

- Since P(X,Z|θ) =P(X|θ)* P(Z|X,θ) ➔
  ln P(X|θ)=ln P(X,Z|θ)- ln P(Z|X,θ)
- ln P(X|θ^{new})- ln P(X|θ^{old})= {ln P(X,Z|θ^{new})- ln P(Z|X,θ^{new})}-{ln P(X,Z|θ^{old})- ln P(Z|X,θ^{old})}
- Apply $\sum_Z p(Z\mid X,\theta^{old})$ on both ends:

$$\sum_Z p(Z\mid X,\theta^{old})[\ln p(X\mid\theta^{new}) - \ln p(X\mid\theta^{old})] = \ln p(X\mid\theta^{new}) - \ln p(X\mid\theta^{old}) =$$

$$\boxed{\sum_Z p(Z\mid X,\theta^{old})\{\ln p(X,Z\mid\theta^{new}) - \ln p(X,Z\mid\theta^{old})\}} -$$

$$\sum_Z p(Z\mid X,\theta^{old})\{\ln p(Z\mid X,\theta^{new}) - \ln p(Z\mid X,\theta^{old})\}$$

- If we can prove <0
  then we are done.

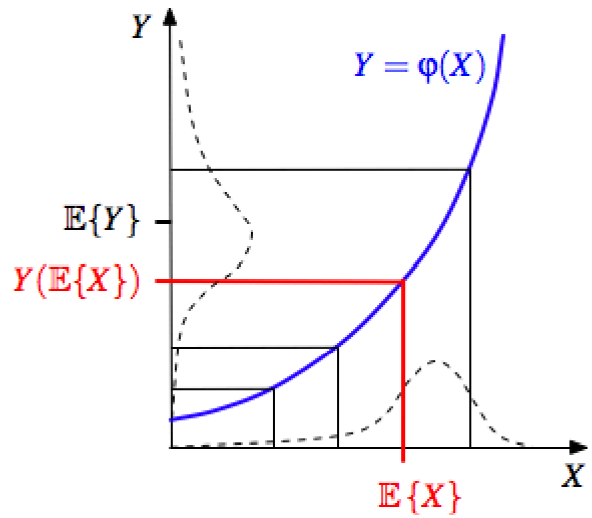
Proof of EM Theorem (2/2)

$$\sum_{Z} p(Z \mid X, \theta^{old})\{\ln \frac{p(Z \mid X, \theta^{new})}{p(Z \mid X, \theta^{old})}\} < 0$$

- The proof used Jensen's Inequality

$$-\sum_{t}\left\{P_{\theta'}(t \mid y_i)\log\frac{P_{\theta}(t \mid y_i)}{P_{\theta'}(t \mid y_i)}\right\} \geq 0$$

- More generally, if p and q are probability distributions $-\sum_{x} p(x)\log\frac{q(x)}{p(x)} \geq 0$



Optimization Process of EM

$$\theta^{new} = \arg\max_{\theta} \sum_{Z} p(Z \mid X, \theta^{old}) \ln p(X, Z \mid \theta)$$

- Initial step: randomly choose $\theta^{old} = \theta_0$
- E-step: using an existing parameter θ^{old} to estimate $P(Z|X, \theta^{old})$
- M-step: find $\theta=\theta^{new}$ that maximize $Q(\theta)= \sum_{Z} p(Z \mid X, \theta^{old}) \ln p(X, Z \mid \theta)$
- Check the convergence of $Q(\theta)$ or θ, if not satisfied, then set $\theta^{old} \leftarrow \theta^{new}$ and go back to E step.

EM: Why P(Z|X, θ^{old}) in the E step?

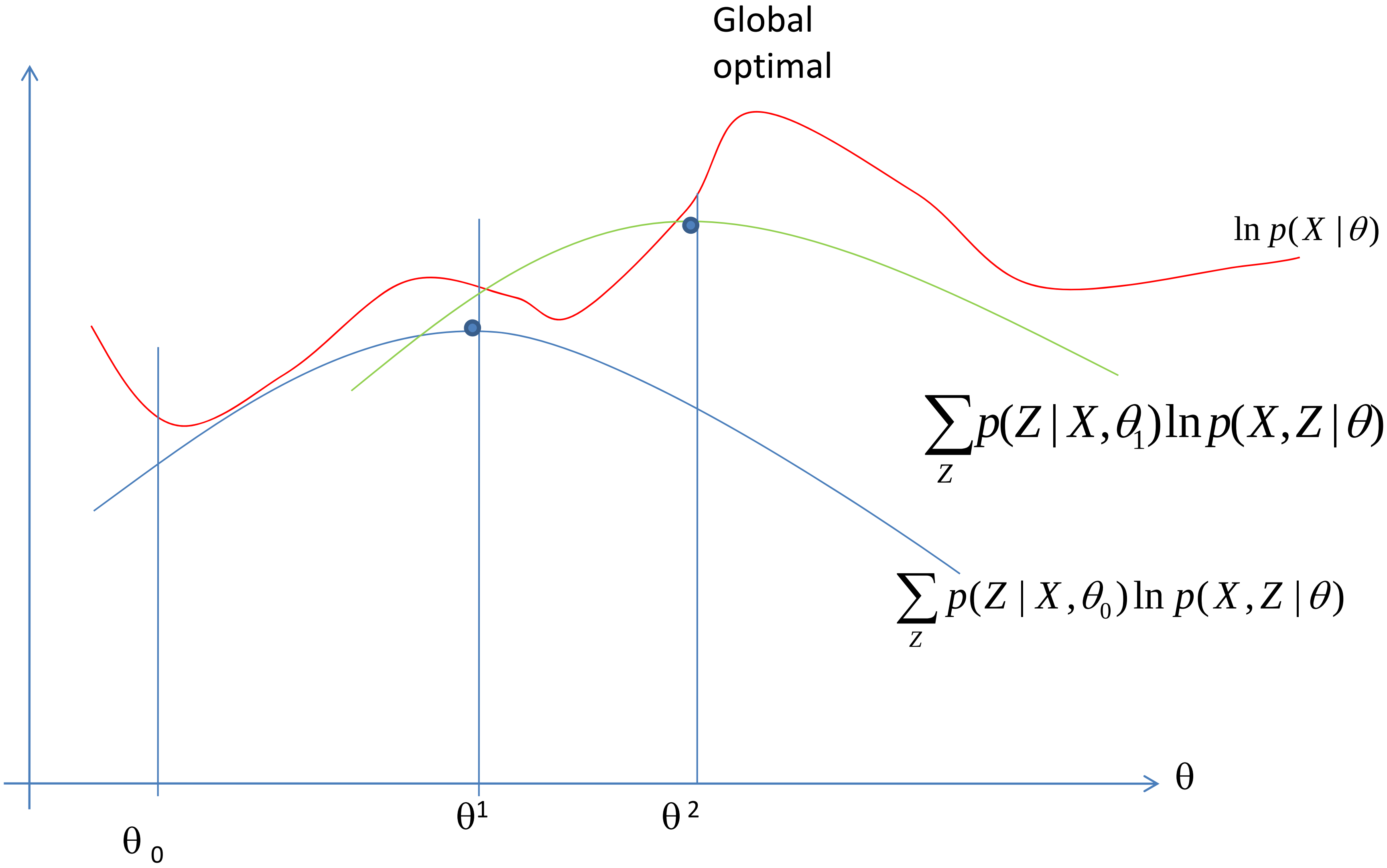
- Let's go back to lnP(X|θ)=lnP(X,Z|θ)- lnP(Z|X,θ)
- $$\sum_Z q(z)\ln p(X\mid\theta) = \sum_Z q(z)\ln p(X,Z\mid\theta) - \sum_Z q(z)\ln p(Z\mid X,\theta)$$

$$\ln p(X\mid\theta) = \sum_Z q(z)\ln\frac{p(X,Z\mid\theta)}{q(z)} \boxed{-\sum_Z q(z)\ln\frac{p(Z\mid X,\theta)}{q(z)}}$$

Independent of q(Z)

Larger than 0, equal holds when q(z)=p(Z|X,θ)

- When θ= θ^{old} , setting q(z)=ln p(Z|X, θ^{old}) can cause $\sum_Z q(z)\ln\frac{p(X,Z\mid\theta^{old})}{q(z)} = \ln p(X\mid\theta^{old})$
- Then we find θ= θ^{new} that maximize $\sum_Z \ln p(Z\mid X,\theta^{old})\ln p(X,Z\mid\theta)$, this θ^{new} will also make >0


Global optimal
$\ln p(X \mid \theta)$
$\sum_Z p(Z \mid X, \theta_1) \ln p(X, Z \mid \theta)$
$\sum_Z p(Z \mid X, \theta_0) \ln p(X, Z \mid \theta)$
θ
θ_0
θ^1
θ^2

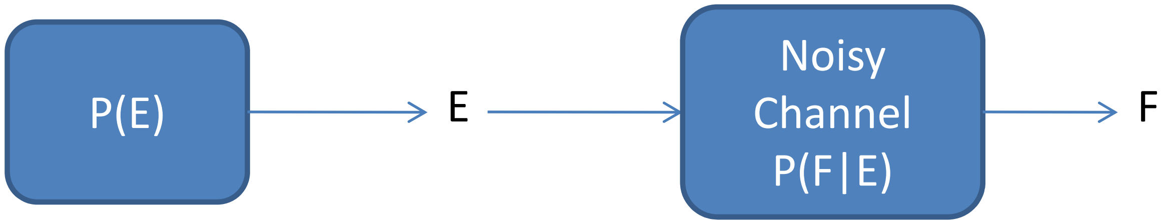
Generalized EM (GEM)

$$\ln p(X \mid \theta^{new}) - \ln p(X \mid \theta^{old}) \geq$$
$$\sum_Z p(Z \mid X, \theta^{old}) \ln p(X, Z \mid \theta^{new}) - \sum_Z p(Z \mid X, \theta^{old}) \ln p(X, Z \mid \theta^{old})$$

- Since the above is always true. In the M-step we don't really need to find a θ^{new} that optimizes $\sum_Z p(Z \mid X, \theta^{old}) \ln p(X, Z \mid \theta)$
- If we can guarantee that θ^{new} always does a better job than θ^{old} in $\sum_Z p(Z \mid X, \theta^{old}) \ln p(X, Z \mid \theta)$ then we are guarantee to reach a local optimal

Ideal vs. Available Data – Alignment Problem for Machine Translation

- MT:



- Ideal: $e_1\ e_2\ e_3$ ….. (solvable by SL)

  $f_1\ f_2\ f_3$ ….
- Available: $e_1\ e_2\ e_3$ ….. (need EM)

  $f_1\ f_2\ f_3$ ….

Ex: English-French Alignment

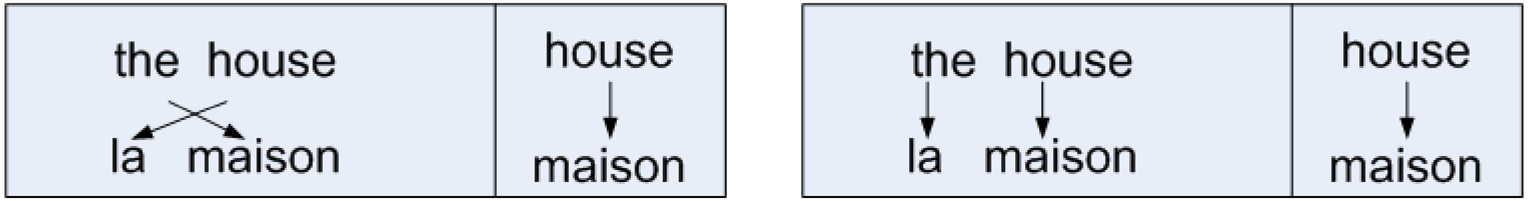
- Data: the house → la maison,

  house → maison
- Alignments are missing!!
- Theory: English words are translated first, then permuted.
- Parameters: P(la|the), p(maison|the), p(la|house), p(maison|house)

Ex: EMTraining on MT

Model to learn:
P(la|the)=?
P(maison|the)=?
P(la|house)=?
P(maison|house)=?

- Possible assignments:



(a) (b)

initialize uniformly:

P(la|the)=1/2
P(maison|the)=1/2
P(la|house)=1/2
P(maison|house)=1/2

E-step

p(a)= 1/8
p(b)= 1/8

M-step
(MLE)

C(la|the)=0*1/8+1*1/8=1/8
C(maison|the)=1 *1/8+0*1/8=1/8
C(la|house)=1*1/8+0*1/8=1/8
C(maison|house)=1*1/8+2*1/8=3/8

normalize

p(la|the)=1/2
p(maison|the)=1/2
p(la|house)=1/4
p(maison|house)=3/4

E-step

P(a)=3/32
P(b)=9/32

M-step

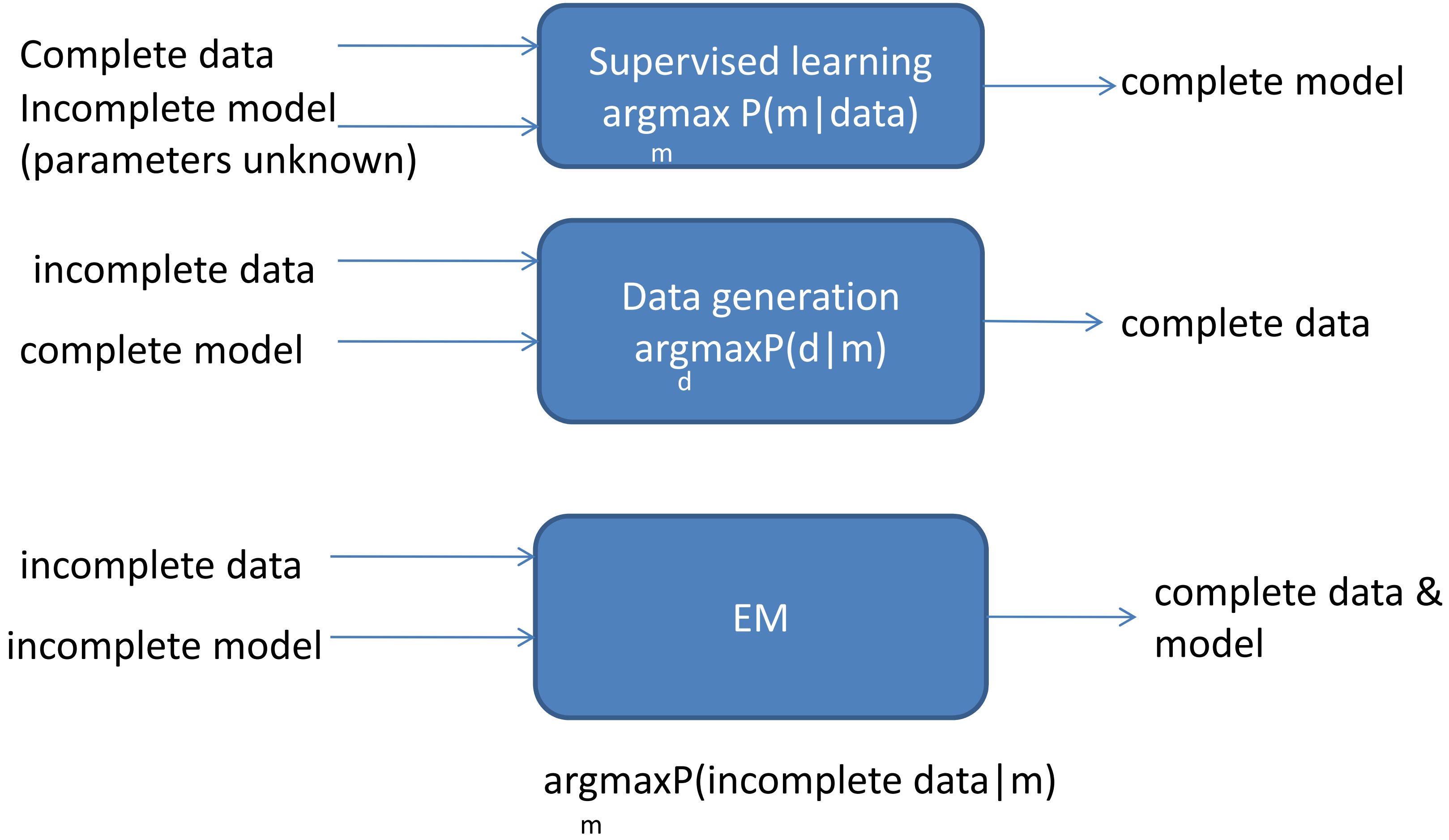
C(la|the)=9/32
C(maison|the)=3/32
C(la|house)=3/32
C(maison|house)=21/32

normalize

p(la|the)=3/4
p(maison|the)=1/4
p(la|house)=1/8
p(maison|house)=7/8

P(a)=7/256
P(b)=147/256

Data and Model



Recommend Reading

- Pattern Recognition and Machine Learning (Bishop Chapter 2, Chapter 9)
- "Bayesian Inference with Tears“ Kevin Knight (Sep 2009)