16.0 An Important Fundamental Approach – EM Algorithm

References: 1. 4.3.2, 4.4.2 of Huang, or 9.1-9.3 of Jelinek

2. 6.4.3 of Rabiner and Juang

3. http://www.stanford .edu/class/cs229/materials.html

4.http://melodi.ee.washington.edu/people/bilmes/mypapers/em.pdf

5.http://www.academia.edu/2785880/A_note_on_EM_algorithm_for_probabilistic_latent_semantic_analysis

EM (Expectation and Maximization) Algorithm

- **Goal**
  estimating the parameters for some probabilistic models based on some criteria
- **Parameter Estimation Principles given some observations $X=[x_1, x_2, \ldots\ldots, x_N]$:**
  - Maximum Likelihood (ML) Principle
    find the model parameter set θ such that the likelihood function is maximized, $P(\boldsymbol{X}|\theta)$= max.
    - For example, if $\theta=\{\mu,\Sigma\}$ is the parameters of a normal distribution, and $\mathbf{X}$ is i.i.d, then the ML estimate of $\theta=\{\mu,\Sigma\}$ can be shown to be

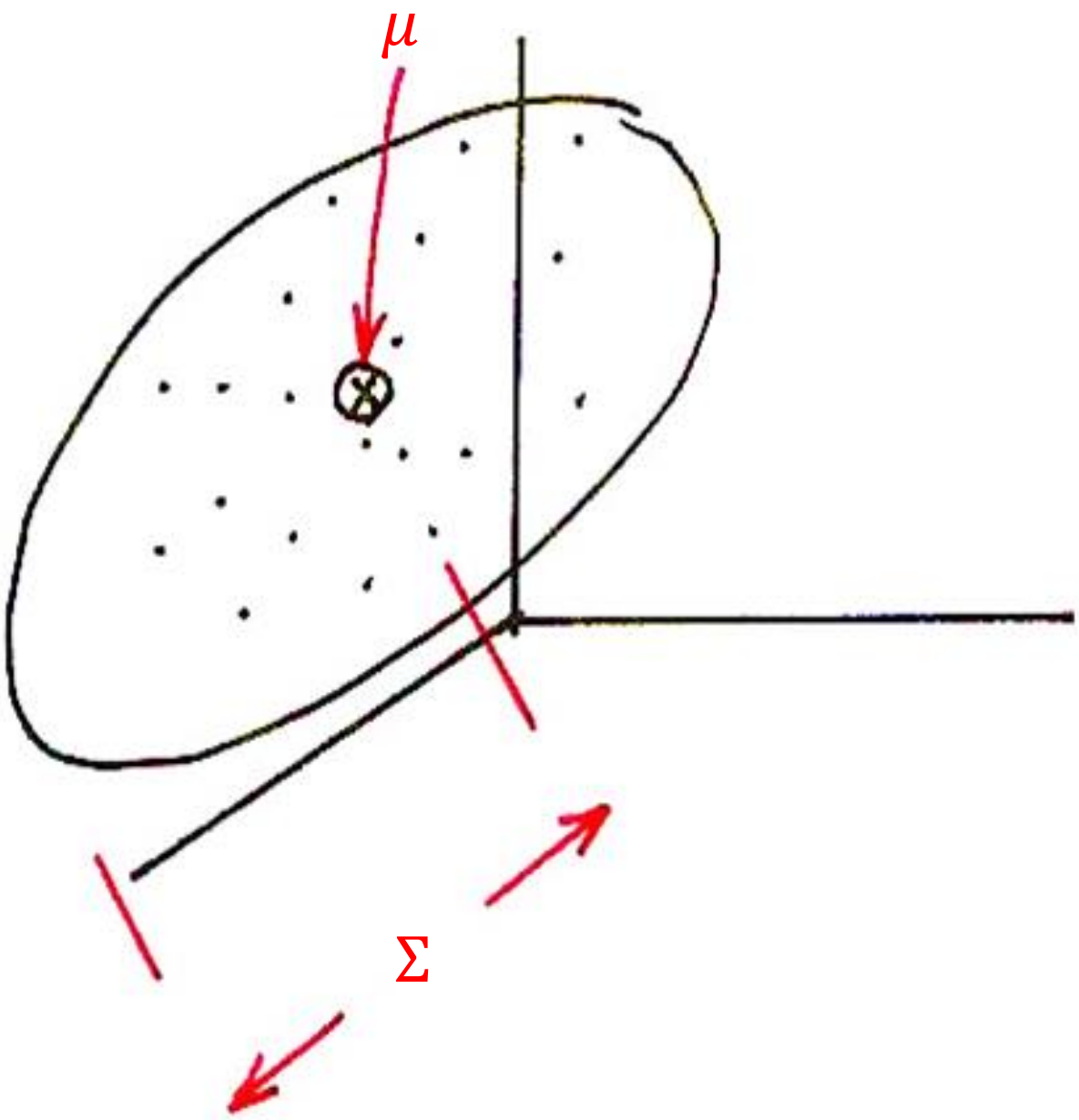
$$\mu_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i \ , \ \Sigma_{ML} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{ML})(x_i - \mu_{ML})^t$$

  - the Maximum A Posteriori (MAP) Principle
    - Find the model parameter θ so that the A Posterior probability is maximized

      i.e. $P(\theta|\boldsymbol{X})= P(\boldsymbol{X}|\theta)\, P(\theta)/\, P(\boldsymbol{X})$= max

      $\Rightarrow P(\boldsymbol{X}|\theta)\, P(\theta)$ = max

Parameter Estimation



EM (Expectation and Maximization) Algorithm

- **Why EM?**
  - In some cases the evaluation of the objective function (e.g. likelihood function) depends on some intermediate variables (latent data) which are not observable (e.g. the state sequence for HMM parameter training)
  - direct estimation of the desired parameters without such latent data is impossible or difficult
    e.g. to estimate $\{A,B, \pi\}$ for HMM without knowing the state sequence

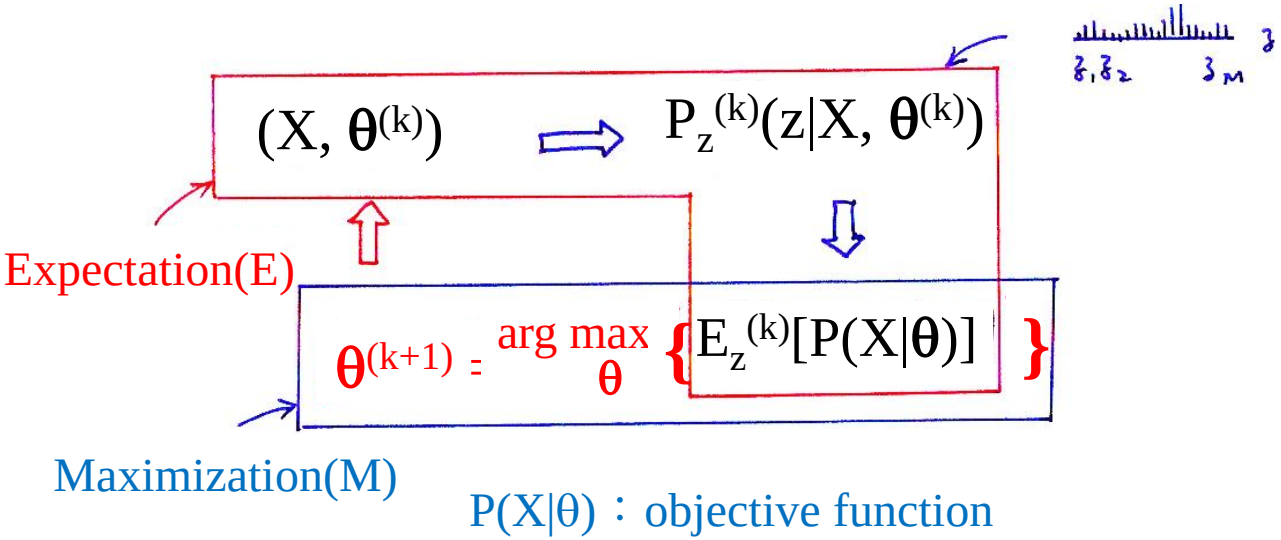
EM (Expectation and Maximization) Algorithm

- **Iterative Procedure with Two Steps in Each Iteration:**
  - ***E*** (Expectation): expectation of the objective function with respect to a distribution (values and probabilities) of the latent data based on the current estimates of the desired parameters conditioned on the given observations
  - ***M*** (Maximization): generating a new set of estimates of the desired parameters by maximizing the objective function (e.g. according to ML or MAP)
  - the objective function increased after each iteration, eventually converged

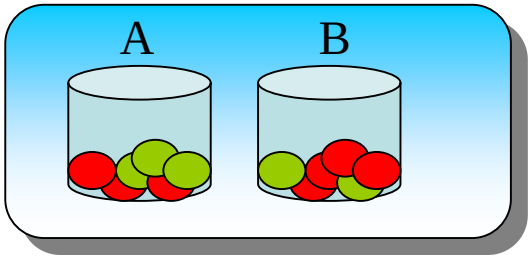
X：available data

$\theta^{(k)}$ ：k-th estimate of the parameter set θ

z：latent data



EM Algorithm: An example



output

(RGG)

Observed data : **O** : "ball sequence": RGG

Latent data : **q** : "bottle sequence": AAB

Parameter to be estimated : $\boldsymbol{\lambda}$={P(A),P(B),P(R|A),P(G|A), P(R|**B**), P(G|B)}, logP(O| λ) = max

- **First, randomly assigned** $\lambda^{(0)}$={$P^{(0)}$(A),$P^{(0)}$(B),$P^{(0)}$(R|A),$P^{(0)}$(G|A), $P^{(0)}$(R|B), $P^{(0)}$(G|B)}
  for example :
  {$P^{(0)}$(A)=0.4,$P^{(0)}$(B)=0.6,$P^{(0)}$(R|A)=0.5,$P^{(0)}$(G|A) =0.5, $P^{(0)}$(R|B) =0.5, $P^{(0)}$(G|B) =0.5}
- **Expectation Step : find the** *expectation* **of logP(O| λ)**
  8 possible state sequences q_i :{AAA},{BBB},{AAB},{BBA},{ABA},{BAB},{ABB},{BAA}

$$E_{\mathbf{q}}\left(\log P\left(O\middle|\lambda\right)\right)=\sum_{i=1}^{8}\log P\left(\mathbf{O},\mathbf{q}_i\middle|\lambda\right)P\left(\mathbf{q}_i\middle|\mathbf{O},\lambda^{(0)}\right)=\sum_{i=1}^{8}\log P\left(\mathbf{O},\mathbf{q}_i\middle|\lambda\right)\frac{P\left(\mathbf{O},\mathbf{q}_i\middle|\lambda^{(0)}\right)}{P\left(\mathbf{O}\middle|\lambda^{(0)}\right)}=\frac{1}{P\left(\mathbf{O}\middle|\lambda^{(0)}\right)}\sum_{i=1}^{8}\log P\left(\mathbf{O},\mathbf{q}_i\middle|\lambda\right)P\left(\mathbf{O},\mathbf{q}_i\middle|\lambda^{(0)}\right)$$

For example, when q_i = {AAB}

$$P\left(\mathbf{O}=RGG,\mathbf{q}_i=AAB\middle|\lambda^{(0)}\right)=P\left(\mathbf{O}=RGG\middle|\mathbf{q}_i=AAB,\lambda^{(0)}\right)P\left(\mathbf{q}_i=AAB\middle|\lambda^{(0)}\right)$$
$$=\left[P^{(0)}(R|A)P^{(0)}(G|A)P^{(0)}(G|B)\right]\left[P^{(0)}(A)P^{(0)}(A)P^{(0)}(B)\right]=0.5*0.5*0.5*0.4*0.4*0.6 \quad \text{(known values)}$$
$$\log P\left(\mathbf{O}=RGG,\mathbf{q}_i=AAB\middle|\lambda\right)=\log\left[P(R|A)P(G|A)P(G|B)\right]\left[P(A)P(A)P(B)\right] \text{(with unknown parameters)}$$

- **Maximization Step : find $\lambda^{(1)}$ to maximize the expectation function** E_q(logP(O|λ))
- **Iterations :** $\lambda^{(0)}$➔ $\lambda^{(1)}$ ➔ $\lambda^{(2)}$ ➔....

EM Algorithm

- **In Each Iteration (assuming logP(x |θ) is the objective function)**
  - E step: expressing the log-likelihood logP($\boldsymbol{x}$|θ) in terms of *the distribution of the latent data conditioned on* [x, $\theta^{(k)}$]
  - M step: find a way to maximized the above function, such that the above function increases monotonically, i.e., logP($\boldsymbol{x}$|$\theta^{(k+1)}$)≥logP($\boldsymbol{x}$|$\theta^{(k)}$)
- **The Conditions for each Iteration to Proceed based on the Criterion**
  - $\boldsymbol{x}$: observed (incomplete) data, $\boldsymbol{z}$: latent data, {$\boldsymbol{x}$, $\boldsymbol{z}$} : complete data

$$p(x,z|\theta) = p(z|x,\theta)p(x|\theta)$$
$$\Rightarrow \log p(x|\theta) = \log p(x,z|\theta) - \log p(z|x,\theta)$$
$$\text{assuming } z \text{ is generated based on } p\left(z|x,\theta^{[k]}\right),$$
$$E_z\left[\log p(x|\theta)\right] = E_z\left[\log p(x,z|\theta)\right] - E_z\left[\log p(z|x,\theta)\right]$$
$$= \int \log p(x,z|\theta)\, p\left(z|x,\theta^{[k]}\right)dz - \int \log p\left(z|x,\theta\right) p\left(z|x,\theta^{[k]}\right)dz$$
$$= Q\left(\theta,\theta^{[k]}\right) - H\left(\theta,\theta^{[k]}\right)$$

EM Algorithm

- **For the EM Iterations to Proceed based on the Criterion:**

$$E_z\left[\log p\left(x|\theta\right)\right] = E_z\left[\log p\left(x,z|\theta\right)\right] - E_z\left[\log p\left(z|x,\theta\right)\right]$$
$$= \int \log p\left(x,z|\theta\right) p\left(z|x,\theta^{[k]}\right) dz - \int \log p\left(z|x,\theta\right) p\left(z|x,\theta^{[k]}\right) dz$$
$$= Q\left(\theta,\theta^{[k]}\right) - H\left(\theta,\theta^{[k]}\right) \text{ (estimate of the objective function given } \theta^{[k]})$$

  - to make sure $\log P(\boldsymbol{x}|\boldsymbol{\theta}^{[k+1]}) \geq \log P(\boldsymbol{x}|\boldsymbol{\theta}^{[k]})$

$$\Rightarrow Q\left(\theta^{[k+1]},\theta^{[k]}\right) - Q\left(\theta^{[k]},\theta^{[k]}\right) - H\left(\theta^{[k+1]},\theta^{[k]}\right) + H\left(\theta^{[k]},\theta^{[k]}\right) \geq 0$$

  - $H(\boldsymbol{\theta}^{[k+1]},\boldsymbol{\theta}^{[k]}) \leq H(\boldsymbol{\theta}^{[k]},\boldsymbol{\theta}^{[k]})$ due to Jenson's Inequality

$$\sum_i p_i \log p_i \geq \sum_i p_i \log q_i \text{ , or } \sum_i p_i \log p_i - \sum_i p_i \log q_i \geq 0$$
$$= \text{when } p_i = q_i$$

  - the only requirement is to have $\boldsymbol{\theta}^{[k+1]}$ such that

$$Q(\boldsymbol{\theta}^{[k+1]},\boldsymbol{\theta}^{[k]}) - Q(\boldsymbol{\theta}^{[k]},\boldsymbol{\theta}^{[k]}) \geq 0$$

  - E-step: to estimate $Q(\boldsymbol{\theta},\boldsymbol{\theta}^{[k]})$: auxiliary function (increase in this function means increase in objective function, maximizing this function may be easier), the expectation of the objective function in terms of the distribution of the latent data conditioned on $(x,\boldsymbol{\theta}^{[k]})$
  - M-step: $\boldsymbol{\theta}^{[k+1]} = \underset{\boldsymbol{\theta}}{\arg\max}\ Q(\boldsymbol{\theta},\boldsymbol{\theta}^{[k]})$

Example: Use of EM Algorithm in Solving Problem 3 of HMM

- **Observed data : *observations* O, latent data : *state sequence* q**
- **The probability of the complete data is**

$$P(\mathbf{O},\mathbf{q}|\lambda) = P(\mathbf{O}|\mathbf{q},\lambda)P(\mathbf{q}|\lambda)$$

- **E-Step :**

$$Q(\lambda, \lambda^{[k]}) = E[\log P(O,q|\lambda)|O, \lambda^{[k]}] = \Sigma_q P(q|O,\lambda^{[k]}) \log[P(O,q|\lambda)]$$

  - $\lambda^{[k]}$: k-th estimate of λ (known), λ: unknown parameter to be estimated
- **M-Step :**
  - Find $\lambda^{[k+1]}$ such that $\lambda^{[k+1]} = \arg\max_{\lambda} Q(\lambda, \lambda^{[k]})$
- **Given the Various Constraints** (e.g. $\sum_i \pi_i = 1, \sum_j a_{ij} = 1,$ *etc.*), **It can be shown**
  - the above maximization leads to the formulas obtained previously
  - $P(O|\lambda^{[k+1]}) \geq P(O|\lambda^{[k]})$