

EM (Expectation and Maximization) Algorithm

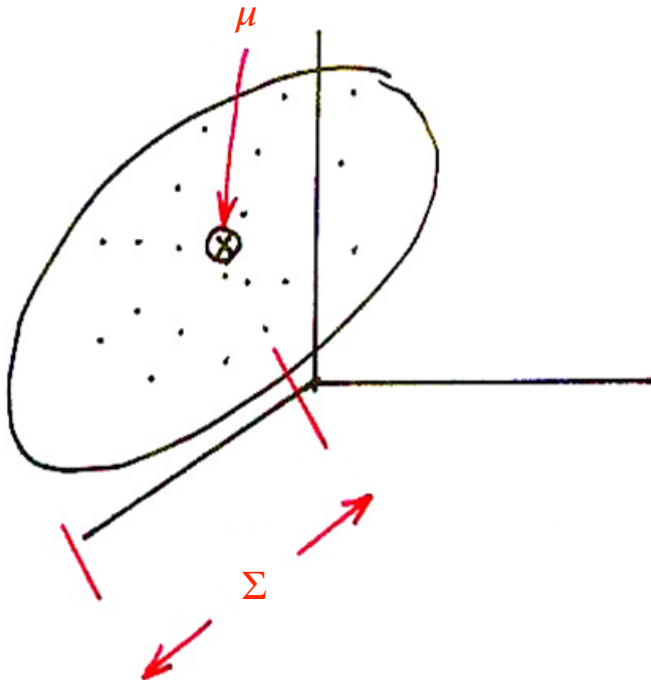
16.0 An Important Fundamental Approach – EM Algorithm

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

- **Goal**
estimating the parameters for some probabilistic models based on some criteria
- **Parameter Estimation Principles given some observations**
 $X = [x_1, x_2, \dots, x_N]$:
 - Maximum Likelihood (ML) Principle
find the model parameter set θ such that the likelihood function is maximized, $P(X|\theta) = \max$.
 - For example, if $\theta = \{\mu, \Sigma\}$ is the parameters of a normal distribution, and X is i.i.d, then the ML estimate of $\theta = \{\mu, \Sigma\}$ can be shown to be
$$\mu_{ML} = \frac{1}{N} \sum_{i=1}^N x_i, \quad \Sigma_{ML} = \frac{1}{N} \sum_{i=1}^N (x_i - \mu_{ML})(x_i - \mu_{ML})^T$$
 - the Maximum A Posteriori (MAP) Principle
 - Find the model parameter θ so that the A Posterior probability is maximized
i.e. $P(\theta|X) = P(X|\theta)P(\theta)/P(X) = \max$
 $\Rightarrow P(X|\theta)P(\theta) = \max$

2

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

3

4

EM (Expectation and Maximization) Algorithm

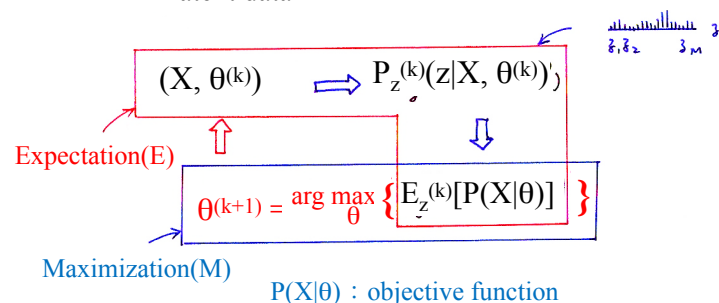
• Iterative Procedure with Two Steps in Each Iteration:

- **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

• In Each Iteration (assuming $\log P(x|\theta)$ is the objective function)

- E step: expressing the log-likelihood $\log P(x|\theta)$ in terms of *the distribution of the latent data conditioned on* $[x, \theta^{(k)}]$
- M step: find a way to maximize the above function, such that the above function increases monotonically, i.e., $\log P(x|\theta^{(k+1)}) \geq \log P(x|\theta^{(k)})$

• The Conditions for each Iteration to Proceed based on the Criterion

- x : observed (incomplete) data, z : latent data, $\{x, z\}$: complete data

$$p(x, z|\theta) = p(z|x, \theta) p(x|\theta)$$

$$\Rightarrow \log p(x|\theta) = \log p(x, z|\theta) - \log p(z|x, \theta)$$

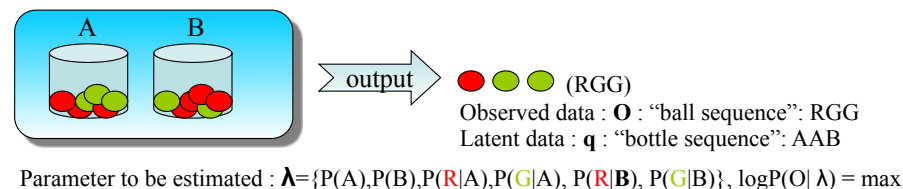
assuming z is generated based on $p(z|x, \theta^{[k]})$

$$E_z[\log p(x|\theta)] = E_z[\log p(x, z|\theta)] - E_z[\log p(z|x, \theta)]$$

$$= \int \log p(x, z|\theta) p(z|x, \theta^{[k]}) dz - \int \log p(z|x, \theta) p(z|x, \theta^{[k]}) dz$$

$$= Q(\theta, \theta^{[k]}) - H(\theta, \theta^{[k]})$$

EM Algorithm: An example



- **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 : $\{P^{(0)}(A)=0.5, P^{(0)}(G|A)=0.5, P^{(0)}(R|B)=0.5, P^{(0)}(G|B)=0.5\}$

- **Expectation Step** : find the expectation of $\log P(O|\lambda)$

8 possible state sequences $q_i : \{AAA\}, \{BBB\}, \{AAB\}, \{BBA\}, \{ABA\}, \{BAB\}, \{ABB\}, \{BAA\}$

$$E_q(\log P(O|\lambda)) = \sum_{i=1}^8 \log P(O, q_i|\lambda) P(q_i|O, \lambda^{(0)}) = \sum_{i=1}^8 \log P(O, q_i|\lambda) \frac{P(O, q_i|\lambda^{(0)})}{P(O|\lambda^{(0)})} = \frac{1}{P(O|\lambda^{(0)})} \sum_{i=1}^8 \log P(O, q_i|\lambda) P(O, q_i|\lambda^{(0)})$$

For example, when $q_i = \{AAB\}$

$$P(O = RGG, q_i = AAB|\lambda^{(0)}) = \frac{P(O = RGG|q_i = AAB, \lambda^{(0)}) P(q_i = AAB|\lambda^{(0)})}{P(O|\lambda^{(0)})} = \frac{P(O = RGG|q_i = AAB, \lambda^{(0)}) P(q_i = AAB|\lambda^{(0)})}{P(O|\lambda^{(0)})}$$

$$= \frac{P^{(0)}(R|A) P^{(0)}(G|A) P^{(0)}(G|B)}{P^{(0)}(A) P^{(0)}(A) P^{(0)}(B)} = 0.5 * 0.5 * 0.5 * 0.4 * 0.4 * 0.6 \quad (\text{known values})$$

$$\log P(O = RGG, q_i = AAB|\lambda) = \log [P(R|A) P(G|A) P(G|B)] + \log [P(A) P(A) P(B)] \quad (\text{with unknown parameters})$$

- **Maximization Step** : find $\lambda^{(1)}$ to maximize the expectation function $E_q(\log P(O|\lambda))$
- **Iterations** : $\lambda^{(0)} \rightarrow \lambda^{(1)} \rightarrow \lambda^{(2)} \rightarrow \dots$

5

EM Algorithm

• For the EM Iterations to Proceed based on the Criterion:

$$E_z[\log p(x|\theta)] = E_z[\log p(x, z|\theta)] - E_z[\log p(z|x, \theta)]$$

$$= \int \log p(x, z|\theta) p(z|x, \theta^{[k]}) dz - \int \log p(z|x, \theta) p(z|x, \theta^{[k]}) dz$$

$$= Q(\theta, \theta^{[k]}) - H(\theta, \theta^{[k]}) \quad (\text{estimate of the objective function given } \theta^{[k]})$$

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

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

- $H(\theta^{[k+1]}, \theta^{[k]}) \leq H(\theta^{[k]}, \theta^{[k]})$ due to Jensen'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 $\theta^{[k+1]}$ such that

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

- E-step: to estimate $Q(\theta, \theta^{[k]})$: 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 $(x, \theta^{[k]})$
- M-step: $\theta^{[k+1]} = \arg \max_{\theta} Q(\theta, \theta^{[k]})$

7

6

8

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

- Observed data : *observations* \mathbf{O} , latent data : *state sequence* \mathbf{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 :
$$Q(\lambda, \lambda^{[k]}) = E[\log P(\mathbf{O}, \mathbf{q} | \lambda) | \mathbf{O}, \lambda^{[k]}] = \sum_{\mathbf{q}} P(\mathbf{q} | \mathbf{O}, \lambda^{[k]}) \log[P(\mathbf{O}, \mathbf{q} | \lambda)]$$
 - $\lambda^{[k]}$: 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
 - $P(\mathbf{O} | \lambda^{[k+1]}) \geq P(\mathbf{O} | \lambda^{[k]})$