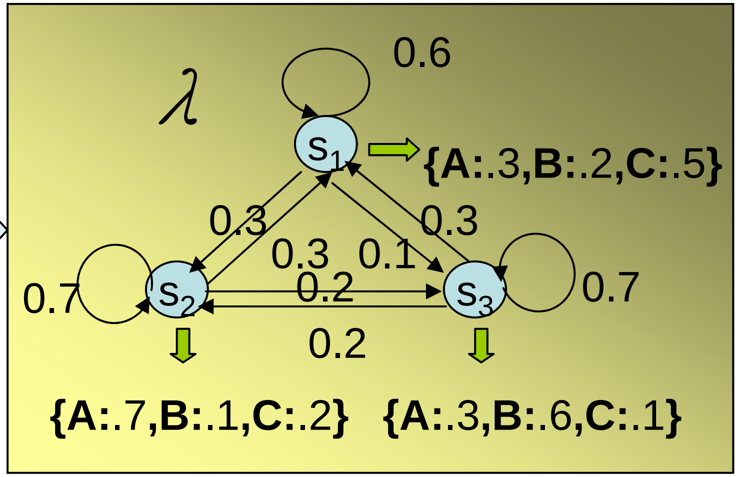
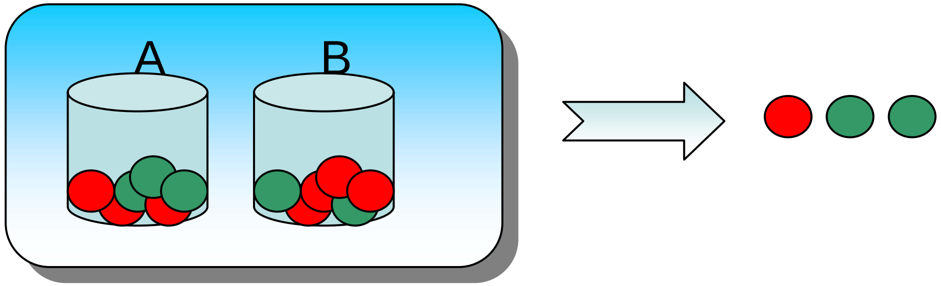
The EM Algorithm

$o_1, o_2, \ldots\ldots, o_T$



$p(\boldsymbol{O}|\lambda)$

$p(\boldsymbol{O}|\bar{\lambda}) > p(\boldsymbol{O}|\lambda)$



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

Latent data : ***S*** : "bottle sequence"

Parameters to be estimated to maximize $\log P(\boldsymbol{O}|\lambda)$

$\lambda=\{P(A),P(B),P(B|A),P(A|B),P(R|A),P(G|A),P(R|B),P(G|B)\}$

The EM Algorithm

- Introduction of EM (Expectation Maximization):
  - Why EM?
    - Simple optimization algorithms for likelihood function relies on the intermediate variables, called latent (隱藏的)data
      In our case here, ***the state sequence** is the latent data*
    - Direct access to the data necessary to estimate the parameters is impossible or difficult
      In our case here, it is almost impossible to estimate {***A***,***B***, π} without consideration of the ***state sequence***
  - Two Major Steps :
    - ***E*** : expectation with respect to the latent data using the current estimate of the parameters and conditioned on the observations $E[\bullet]_{S|\lambda,O}$
    - ***M***: provides a new estimation of the parameters according to Maximum likelihood (ML) or Maximum A Posterior (MAP) Criteria

The EM Algorithm

ML and MAP

- Estimation principle based on observations:

$$\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n) \Longleftrightarrow \boldsymbol{X} = \{X_1, X_2, \ldots, X_n\}$$

  – **The Maximum Likelihood (ML) Principle**
  find the model parameter $\boldsymbol{\Phi}$ so that the likelihood $p(\boldsymbol{x}|\boldsymbol{\Phi})$ is maximum
  for example, if $\boldsymbol{\Phi} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ *is the parameters of a multivariate normal distribution, and **X** is i.i.d.* (independent, identically distributed), *then the ML estimate of* $\boldsymbol{\Phi} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ *is*

$$\boldsymbol{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i \;, \quad \boldsymbol{\Sigma}_{ML} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu}_{ML})(\boldsymbol{x}_i - \boldsymbol{\mu}_{ML})^t$$

  – **The Maximum A Posteriori (MAP) Principle**
  find the model parameter $\boldsymbol{\Phi}$ so that the likelihood $p(\boldsymbol{\Phi}|\boldsymbol{x})$ is maximum

The EM Algorithm

- The EM Algorithm is important to HMMs and other learning techniques
  - Discover new model parameters to maximize the log-likelihood of incomplete data $\log P(\boldsymbol{O}|\lambda)$ by iteratively maximizing the expectation of log-likelihood from complete data $\log P(\boldsymbol{O}, \boldsymbol{S}|\lambda)$

- Using scalar random variables to introduce the EM algorithm
  - The observable training data $\boldsymbol{O}$
    - We want to maximize $P(\boldsymbol{O}|\lambda)$, λ is a parameter vector
  - The hidden (unobservable) data $\boldsymbol{S}$
    - E.g. the component densities of observable data $\boldsymbol{O}$, or the underlying state sequence in HMMs

The EM Algorithm

– Assume we have λ and estimate the probability that each $\boldsymbol{S}$ occurred in the generation of $\boldsymbol{O}$

– Pretend we had in fact observed a complete data pair $(\boldsymbol{O},\boldsymbol{S})$ with frequency proportional to the probability $P(\boldsymbol{O},\boldsymbol{S}|\lambda)$, to computed a new $\bar{\lambda}$, the maximum likelihood estimate of λ

– Does the process converge?

– **Algorithm** unknown model setting

$$P(\boldsymbol{O},\boldsymbol{S}|\bar{\lambda}) = P(\boldsymbol{S}|\boldsymbol{O},\bar{\lambda})P(\boldsymbol{O}|\bar{\lambda})$$

Bayes' rule

complete data likelihood incomplete data likelihood

- **Log-likelihood expression** and expectation taken over ***S***

$$\log P(\boldsymbol{O}|\bar{\lambda}) = \log P(\boldsymbol{O},\boldsymbol{S}|\bar{\lambda}) - \log P(\boldsymbol{S}|\boldsymbol{O},\bar{\lambda})$$

take expectation over **S**

$$\log P(\boldsymbol{O}|\bar{\lambda}) = \sum_{S}\left[P(\boldsymbol{S}|\boldsymbol{O},\lambda)\log P(\boldsymbol{O}|\bar{\lambda})\right]$$

$$= \sum_{\boldsymbol{S}}\left[P(\boldsymbol{S}|\boldsymbol{O},\lambda)\log P(\boldsymbol{O},\boldsymbol{S}|\bar{\lambda})\right] - \sum_{\boldsymbol{S}}\left[P(\boldsymbol{S}|\boldsymbol{O},\lambda)\log P(\boldsymbol{S}|\boldsymbol{O},\bar{\lambda})\right]$$

The EM Algorithm

– Algorithm (Cont.)
  - We can thus express $\log P(\boldsymbol{O}|\bar{\lambda})$ as follows

$$
\begin{aligned}
&\log P(\boldsymbol{O}|\bar{\lambda}) \\
&= \sum_{\boldsymbol{S}} \left[P(\boldsymbol{S}|\boldsymbol{O},\lambda) \log P(\boldsymbol{O},\boldsymbol{S}|\bar{\lambda}) \right] - \sum_{\boldsymbol{S}} \left[P(\boldsymbol{S}|\boldsymbol{O},\lambda) \log P(\boldsymbol{S}|\boldsymbol{O},\bar{\lambda}) \right] \\
&= Q(\lambda,\bar{\lambda}) - H(\lambda,\bar{\lambda})
\end{aligned}
$$

where

$$
Q(\lambda,\bar{\lambda}) = \sum_{\boldsymbol{S}} \left[P(\boldsymbol{S}|\boldsymbol{O},\lambda) \log P(\boldsymbol{O},\boldsymbol{S}|\bar{\lambda}) \right]
$$

$$
H(\lambda,\bar{\lambda}) = \sum_{\boldsymbol{S}} \left[P(\boldsymbol{S}|\boldsymbol{O},\lambda) \log P(\boldsymbol{S}|\boldsymbol{O},\bar{\lambda}) \right]
$$

  - We want $\log P(\boldsymbol{O}|\bar{\lambda}) \geq \log P(\boldsymbol{O}|\lambda)$

$$
\begin{aligned}
&\log P(\boldsymbol{O}|\bar{\lambda}) - \log P(\boldsymbol{O}|\lambda) \\
&= \left[Q(\lambda,\bar{\lambda}) - H(\lambda,\bar{\lambda}) \right] - \left[Q(\lambda,\lambda) - H(\lambda,\lambda) \right] \\
&= Q(\lambda,\bar{\lambda}) - Q(\lambda,\lambda) - H(\lambda,\bar{\lambda}) + H(\lambda,\lambda)
\end{aligned}
$$

The EM Algorithm

- $-H(\lambda,\bar{\lambda})+H(\lambda,\lambda)$ has the following property

$$
\begin{aligned}
&-H(\lambda,\bar{\lambda})+H(\lambda,\lambda)\\
&=-\sum_{\mathbf{S}}\left[P(\mathbf{S}|\mathbf{O},\lambda)\log\frac{P(\mathbf{S}|\mathbf{O},\bar{\lambda})}{P(\mathbf{S}|\mathbf{O},\lambda)}\right]\\
&\geq\sum_{\mathbf{S}}\left[P(\mathbf{S}|\mathbf{O},\lambda)\left(1-\frac{P(\mathbf{S}|\mathbf{O},\bar{\lambda})}{P(\mathbf{S}|\mathbf{O},\lambda)}\right)\right]\quad(\because \log x\leq x-1)\\
&=\sum_{\mathbf{S}}\left[P(\mathbf{S}|\mathbf{O},\lambda)-P(\mathbf{S}|\mathbf{O},\bar{\lambda})\right]\\
&=0\\
&\therefore -H(\lambda,\bar{\lambda})+H(\lambda,\lambda)\geq 0
\end{aligned}
$$

Kullbuack-Leibler (KL) distance

Jensen's inequality

– Therefore, for maximizing $\log P(\mathbf{O}|\bar{\lambda})$, we only need to maximize the *Q*-function (auxiliary function)

$$Q(\lambda,\bar{\lambda})=\sum_{\mathbf{S}}\left[P(\mathbf{S}|\mathbf{O},\lambda)\log P(\mathbf{O},\mathbf{S}|\bar{\lambda})\right]$$

Expectation of the complete data log likelihood with respect to the latent state sequences

EM Applied to Discrete HMM Training

- Apply EM algorithm to iteratively refine the HMM parameter vector $\boldsymbol{\lambda} = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$
  - By maximizing the auxiliary function

$$Q(\lambda, \bar{\lambda}) = \sum_{\boldsymbol{S}} \left[P(\boldsymbol{S} | \boldsymbol{O}, \lambda) log\ P(\boldsymbol{O}, \boldsymbol{S} | \bar{\lambda}) \right]$$

$$= \sum_{\boldsymbol{S}} \left[\frac{P(\boldsymbol{O}, \boldsymbol{S} | \lambda)}{P(\boldsymbol{O} | \lambda)} log\ P(\boldsymbol{O}, \boldsymbol{S} | \bar{\lambda}) \right]$$

  - Where $P(\boldsymbol{O}, \boldsymbol{S} | \lambda)$ and $P(\boldsymbol{O}, \boldsymbol{S} | \bar{\lambda})$ can be expressed as

$$P(\boldsymbol{O}, \boldsymbol{S} | \lambda) = \pi_{s_1} \left[\prod_{t=1}^{T-1} a_{s_t s_{t+1}} \right] \left[\prod_{t=1}^{T} b_{s_t}(\boldsymbol{o}_t) \right]$$

$$\log P(\boldsymbol{O}, \boldsymbol{S} | \lambda) = \log \pi_{s_1} + \sum_{t=1}^{T-1} \log a_{s_t s_{t+1}} + \sum_{t=1}^{T} \log b_{s_t}(\boldsymbol{o}_t)$$

$$\log P(\boldsymbol{O}, \boldsymbol{S} | \bar{\lambda}) = \log \bar{\pi}_{s_1} + \sum_{t=1}^{T-1} \log \bar{a}_{s_t s_{t+1}} + \sum_{t=1}^{T} \log \bar{b}_{s_t}(\boldsymbol{o}_t)$$

EM Applied to Discrete HMM Training

- Rewrite the auxiliary function as

w_i y_i

$$Q(\lambda,\bar{\lambda}) = Q_{\pi}(\lambda,\bar{\pi}) + Q_{a}(\lambda,\bar{a}) + Q_{b}(\lambda,\bar{b})$$

$$Q_{\pi}(\lambda,\bar{\pi}) = \sum_{\text{all } S}\left[\frac{P(O,S|\lambda)}{P(O|\lambda)}\log \bar{\pi}_{s_1}\right] \overset{?}{=} \sum_{i=1}^{N}\left[\frac{P(O,s_1=i|\lambda)}{P(O|\lambda)}\log \bar{\pi}_i\right]$$

$$Q_{a}(\lambda,\bar{a}) = \sum_{\text{all } S}\left[\frac{P(O,S|\lambda)}{P(O|\lambda)}\sum_{t=1}^{T-1}\log \bar{a}_{s_t s_{t+1}}\right] \overset{?}{=} \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T-1}\left[\frac{P(O,s_t=i,s_{t+1}=j|\lambda)}{P(O|\lambda)}\log \bar{a}_{ij}\right]$$

$$Q_{b}(\lambda,\bar{b}) = \sum_{\text{all } S}\left[\frac{P(O,S|\lambda)}{P(O|\lambda)}\sum_{t=1}^{T}\log \bar{b}_{s_t}(k)\right] \overset{?}{=} \sum_{j=1}^{N}\sum_{k}\sum_{t\in o_t=v_k}\left[\frac{P(O,s_t=j|\lambda)}{P(O|\lambda)}\log \bar{b}_j(k)\right]$$

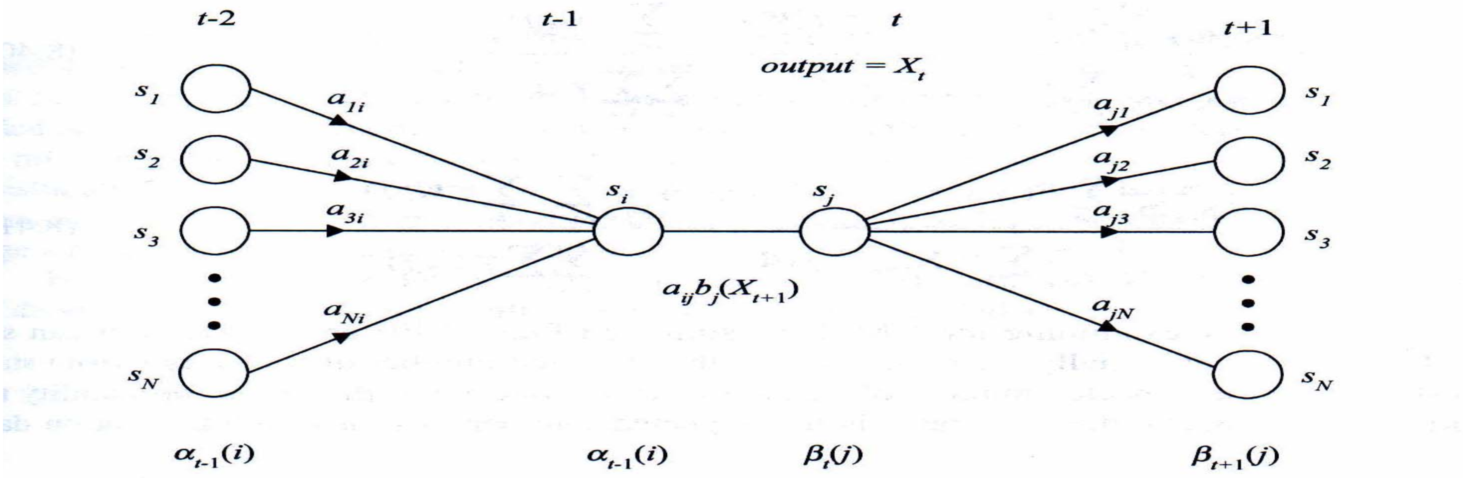


Figure 8.7 Illustration of the operations required for the computation of $\gamma_t(i,j)$, which is the probability of taking the transition from state i to state j at time t.

EM Applied to Discrete HMM Training

- The auxiliary function contains three independent terms, π_i , a_{ij} and $b_j(k)$
  - Can be maximized individually
  - All of the same form

$$F(\boldsymbol{y}) = g(y_1, y_2, \ldots, y_N) = \sum_{j=1}^{N} w_j \log y_j, \quad \text{where} \sum_{j=1}^{N} y_j = 1, \text{ and } y_j \geq 0$$

$$F(\boldsymbol{y}) \text{ has maximum value when : } y_j = \frac{w_j}{\sum_{j=1}^{N} w_j}$$

EM Applied to Discrete HMM Training

- **Proof**: Apply Lagrange Multiplier

By applying Lagrange Multiplier ℓ

Suppose that $F = \sum_{j=1}^{N} w_j \log y_j = \sum_{j=1}^{N} w_j \log y_j + \ell \left(\sum_{j=1}^{N} y_j - 1 \right)$

Constraint

$$\frac{\partial F}{\partial y_j} = \frac{w_j}{y_j} + \ell = 0 \Rightarrow \ell = -\frac{w_j}{y_j} \quad \forall\ j$$

$$\ell \sum_{j=1}^{N} y_j = -\sum_{j=1}^{N} w_j \Rightarrow \ell = -\sum_{j=1}^{N} w_j$$

$$\therefore y_j = \frac{w_j}{\sum_{j=1}^{N} w_j}$$

EM Applied to Discrete HMM Training

- The new model parameter set $\bar{\lambda} = \left(\bar{\boldsymbol{\pi}}, \bar{\boldsymbol{A}}, \bar{\boldsymbol{B}}\right)$ can be expressed as:

$$\overline{\pi}_i = \frac{P\left(\boldsymbol{O}, s_1 = i \middle| \lambda\right)}{P\left(\boldsymbol{O} \middle| \lambda\right)} = \gamma_1(i)$$

$$\overline{a}_{ij} = \frac{\sum_{t=1}^{T-1} P\left(\boldsymbol{O}, s_t = i, s_{t+1} = j \middle| \lambda\right)}{\sum_{t=1}^{T-1} P\left(\boldsymbol{O}, s_t = i \middle| \lambda\right)} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

$$\overline{b}_i(k) = \frac{\sum_{\substack{t=1 \\ \text{s.t. } o_t = v_k}}^{T} P\left(\boldsymbol{O}, s_t = i \middle| \lambda\right)}{\sum_{t=1}^{T} P\left(\boldsymbol{O}, s_t = i \middle| \lambda\right)} = \frac{\sum_{\substack{t=1 \\ \text{s.t. } o_t = v_k}}^{T} \gamma_t(i)}{\sum_{t=1}^{T} \gamma_t(i)}$$

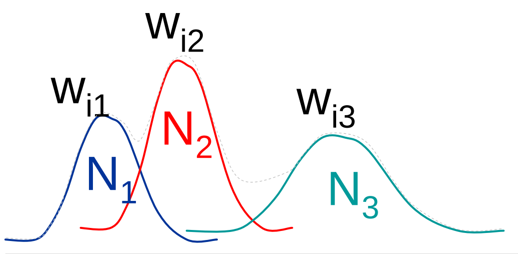
EM Applied to Continuous HMM Training

- Continuous HMM: the state observation does not come from a finite set, but from a continuous space
  - The difference between the discrete and continuous HMM lies in a different form of state output probability
  - Discrete HMM requires the quantization procedure to map observation vectors from the continuous space to the discrete space
- Continuous Mixture HMM
  - The state observation distribution of HMM is modeled by multivariate Gaussian mixture density functions (*M* mixtures)

$$b_j(\boldsymbol{o}) = \sum_{k=1}^{M} c_{jk} b_{jk}(\boldsymbol{o})$$

$$= \sum_{k=1}^{M} c_{jk} N(\boldsymbol{o}; \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}) = \sum_{k=1}^{M} c_{jk} \left(\frac{1}{\left(\sqrt{2\pi}\right)^L \left|\boldsymbol{\Sigma}_{jk}\right|^{1/2}} \exp\left(-\frac{1}{2} (\boldsymbol{o} - \boldsymbol{\mu}_{jk})^t \boldsymbol{\Sigma}_{jk}^{-1} (\boldsymbol{o} - \boldsymbol{\mu}_{jk}) \right) \right)$$

$$\sum_{k=1}^{M} c_{jk} = 1$$



Distribution for State *i*

EM Applied to Continuous HMM Training

- Express $b_j(\boldsymbol{o})$ with respect to each single mixture component $b_{jk}(\boldsymbol{o})$

Note:

$$\prod_{t=1}^{T}\left(\sum_{k_t=1}^{M} a_{tk_t}\right)$$

$$=(a_{11}+a_{12}+\ldots+a_{1M})(a_{21}+a_{22}+\ldots+a_{2M})\ldots(a_{T1}+a_{T2}+\ldots+a_{TM})$$

$$=\sum_{k_1=1}^{M}\sum_{k_2=1}^{M}\ldots\sum_{k_T=1}^{M}\prod_{t=1}^{T} a_{tk_t}$$

$$P(\boldsymbol{O},\boldsymbol{S}|\boldsymbol{\lambda})=\pi_{s_1}\left\{\prod_{t=1}^{T-1} a_{s_t s_{t+1}}\right\}\left\{\prod_{t=1}^{T} b_{s_t}(\boldsymbol{o}_t)\right\}$$

$$=\pi_{s_1}\left\{\prod_{t=1}^{T-1} a_{s_t s_{t+1}}\right\}\left\{\sum_{k_1=1}^{M}\sum_{k_2=1}^{M}\cdots\sum_{k_T=1}^{M}\prod_{t=1}^{T}\left[c_{s_t k_t} b_{s_t k_t}(\boldsymbol{o}_t)\right]\right\}$$

$$P(\boldsymbol{O},\boldsymbol{S},\boldsymbol{K}|\boldsymbol{\lambda})=\pi_{s_1}\left\{\prod_{t=1}^{T-1} a_{s_t s_{t+1}}\right\}\left\{\prod_{t=1}^{T}\left[c_{s_t k_t} b_{s_t k_t}(\boldsymbol{o}_t)\right]\right\}$$

$\boldsymbol{K}$: one of the possible mixture component sequence along with the state sequence $\boldsymbol{S}$

$$P(\boldsymbol{O}|\boldsymbol{\lambda})=\sum_{\boldsymbol{S}}\sum_{\boldsymbol{K}} P(\boldsymbol{O},\boldsymbol{S},\boldsymbol{K}|\boldsymbol{\lambda})$$

EM Applied to Continuous HMM Training

- Therefore, an auxiliary function for the EM algorithm can be written as:

$$Q(\lambda, \bar{\lambda}) = \sum_{\boldsymbol{S}} \sum_{\boldsymbol{K}} \left[P(\boldsymbol{S}, \boldsymbol{K} | \boldsymbol{O}, \lambda) \log P(\boldsymbol{O}, \boldsymbol{S}, \boldsymbol{K} | \bar{\lambda}) \right]$$

$$= \sum_{\boldsymbol{S}} \sum_{\boldsymbol{K}} \left[\frac{P(\boldsymbol{O}, \boldsymbol{S}, \boldsymbol{K} | \lambda)}{P(\boldsymbol{O} | \lambda)} \log P(\boldsymbol{O}, \boldsymbol{S}, \boldsymbol{K} | \bar{\lambda}) \right]$$

$$\log P(\boldsymbol{O}, \boldsymbol{S}, \boldsymbol{K} | \bar{\lambda}) = \log \bar{\pi}_{s_1} + \sum_{t=1}^{T-1} \log \bar{a}_{s_t s_{t+1}} + \sum_{t=1}^{T} \log \bar{b}_{s_t k_t}(\boldsymbol{o}_t) + \sum_{t=1}^{T} \log \bar{c}_{s_t k_t}$$

$$\Downarrow$$

$$Q(\lambda, \bar{\lambda}) = Q_{\boldsymbol{\pi}}(\lambda, \bar{\pi}) + Q_{\boldsymbol{a}}(\lambda, \bar{\boldsymbol{a}}) + Q_{\boldsymbol{b}}(\lambda, \bar{\boldsymbol{b}}) + Q_{c}(\lambda, \bar{c})$$

initial probabilities | state transition probabilities | Gaussian density functions | mixture components

EM Applied to Continuous HMM Training

- The only difference we have when compared with Discrete HMM training

$$\gamma_t(j,k)$$

$$Q_{\boldsymbol{b}}\left(\boldsymbol{\lambda},\overline{\boldsymbol{b}}\right)=\sum_{t=1}^{T}\left\{\left[\sum_{j=1}^{N}\sum_{k=1}^{M}P\left(s_t=j,k_t=k\middle|\boldsymbol{O},\boldsymbol{\lambda}\right)\right]\log\overline{b}_{jk}\left(\boldsymbol{o}_t\right)\right\}$$

$$Q_c\left(\boldsymbol{\lambda},\overline{\boldsymbol{c}}\right)=\sum_{t=1}^{T}\left\{\left[\sum_{j=1}^{N}\sum_{k=1}^{M}P\left(s_t=j,k_t=k\middle|\boldsymbol{O},\boldsymbol{\lambda}\right)\right]\log\overline{c}_{jk}\left(\boldsymbol{o}_t\right)\right\}$$

EM Applied to Continuous HMM Training

Let $\gamma_t(j,k) = \sum_{k=1}^{M} P(s_t = j, k_t = k | \boldsymbol{O}, \boldsymbol{\lambda})$

$$\overline{b}_{jk}(\boldsymbol{o}_t) = N(\boldsymbol{o}_t ; \overline{\boldsymbol{\mu}}_{jk}, \overline{\boldsymbol{\Sigma}}_{jk}) = \frac{1}{(2\pi)^{L/2} |\overline{\boldsymbol{\Sigma}}_{jk}|^{1/2}} \exp\left(-\frac{1}{2} (\boldsymbol{o}_t - \overline{\boldsymbol{\mu}}_{jk})^t \overline{\boldsymbol{\Sigma}}_{jk}^{-1} (\boldsymbol{o}_t - \overline{\boldsymbol{\mu}}_{jk}) \right)$$

$$\log \overline{b}_{jk}(\boldsymbol{o}_t) = -L/2 \cdot \log(2\pi) + 1/2 \cdot \log |\overline{\boldsymbol{\Sigma}}_{jk}^{-1}| - \left(\frac{1}{2} (\boldsymbol{o}_t - \overline{\boldsymbol{\mu}}_{jk})^t \overline{\boldsymbol{\Sigma}}_{jk}^{-1} (\boldsymbol{o}_t - \overline{\boldsymbol{\mu}}_{jk}) \right)$$

$$\frac{\partial \log \overline{b}_{jk}(\boldsymbol{o}_t)}{\partial \overline{\boldsymbol{\mu}}_{jk}} = \overline{\boldsymbol{\Sigma}}_{jk}^{-1} (\boldsymbol{o}_t - \overline{\boldsymbol{\mu}}_{jk})$$

$$\frac{d(\boldsymbol{x}^T \boldsymbol{C} \boldsymbol{x})}{d\boldsymbol{x}} = (\boldsymbol{C} + \boldsymbol{C}^T) \boldsymbol{x}$$

and Σ_{jk}^{-1} is symmetric here

$$\frac{\partial Q_{\boldsymbol{b}}(\boldsymbol{\lambda}, \overline{\boldsymbol{b}})}{\partial \overline{\boldsymbol{\mu}}_{jk}} = \frac{\partial \sum_{t=1}^{T} \left\{ \left[\sum_{j=1}^{N} \sum_{k=1}^{M} \gamma_t(j,k) \log \overline{b}_{jk}(\boldsymbol{o}_t) \right] \right\}}{\partial \overline{\boldsymbol{\mu}}_{jk}}$$

$$\Rightarrow \sum_{t=1}^{T} \left\{ \gamma_t(j,k) \overline{\boldsymbol{\Sigma}}_{jk}^{-1} (\boldsymbol{o}_t - \overline{\boldsymbol{\mu}}_{jk}) \right\} = 0$$

$$\Rightarrow \overline{\boldsymbol{\mu}}_{jk} = \frac{\sum_{t=1}^{T} [\gamma_t(j,k) \cdot \boldsymbol{o}_t]}{\sum_{t=1}^{T} \gamma_t(j,k)}$$

EM Applied to Continuous HMM Training

$$\log \bar{b}_{jk}(\boldsymbol{o}_t) = -\frac{L}{2}\cdot\log(2\pi) - \frac{1}{2}\cdot\log\left|\bar{\boldsymbol{\Sigma}}_{jk}\right| - \left(\frac{1}{2}(\boldsymbol{o}_t - \bar{\boldsymbol{\mu}}_{jk})^t\,\bar{\boldsymbol{\Sigma}}_{jk}^{-1}(\boldsymbol{o}_t - \bar{\boldsymbol{\mu}}_{jk})\right)$$

$$\frac{\partial \log \bar{b}_{jk}(\boldsymbol{o}_t)}{\partial(\bar{\boldsymbol{\Sigma}}_{jk})} = -\left[\frac{1}{2}\cdot\left|\bar{\boldsymbol{\Sigma}}_{jk}\right|^{-1}\cdot\left|\bar{\boldsymbol{\Sigma}}_{jk}\right|\cdot\bar{\boldsymbol{\Sigma}}_{jk}^{-1} - \left(\bar{\boldsymbol{\Sigma}}_{jk}^{-1}\frac{1}{2}(\boldsymbol{o}_t - \bar{\boldsymbol{\mu}}_{jk})(\boldsymbol{o}_t - \bar{\boldsymbol{\mu}}_{jk})^t\,\bar{\boldsymbol{\Sigma}}_{jk}^{-1}\right)\right]$$

$$= -\frac{1}{2}\cdot\left[\bar{\boldsymbol{\Sigma}}_{jk}^{-1} - \bar{\boldsymbol{\Sigma}}_{jk}^{-1}(\boldsymbol{o}_t - \bar{\boldsymbol{\mu}}_{jk})(\boldsymbol{o}_t - \bar{\boldsymbol{\mu}}_{jk})^t\,\bar{\boldsymbol{\Sigma}}_{jk}^{-1}\right]$$

$$\frac{d\,(\boldsymbol{a}^T\boldsymbol{X}^{-1}\boldsymbol{b})}{d\boldsymbol{X}} = -\boldsymbol{X}^T\boldsymbol{a}\boldsymbol{b}^T\boldsymbol{X}^T$$

$$\frac{d[\det(\boldsymbol{X})]}{d\boldsymbol{X}} = \det(\boldsymbol{X})\cdot\boldsymbol{X}^{-T}$$

and Σ_{jk} is symmetric here

$$\frac{\partial Q_{\boldsymbol{b}}(\lambda,\bar{\boldsymbol{b}})}{\partial(\bar{\boldsymbol{\Sigma}}_{jk})} = \frac{\partial \sum_{t=1}^{T}\left\{\left[\sum_{j=1}^{N}\sum_{k=1}^{M}\gamma_t(j,k)\log \bar{b}_{jk}(\boldsymbol{o}_t)\right]\right\}}{\partial(\bar{\boldsymbol{\Sigma}}_{jk}^{-1})}$$

$$\Rightarrow \sum_{t=1}^{T}\left\{\gamma_t(j,k)\left(-\frac{1}{2}\right)\cdot\left[\bar{\boldsymbol{\Sigma}}_{jk}^{-1} - \bar{\boldsymbol{\Sigma}}_{jk}^{-1}(\boldsymbol{o}_t - \bar{\boldsymbol{\mu}}_{jk})(\boldsymbol{o}_t - \bar{\boldsymbol{\mu}}_{jk})^t\,\bar{\boldsymbol{\Sigma}}_{jk}^{-1}\right]\right\} = 0$$

$$\Rightarrow \sum_{t=1}^{T}\gamma_t(j,k)\bar{\boldsymbol{\Sigma}}_{jk}^{-1} = \sum_{t=1}^{T}\gamma_t(j,k)\bar{\boldsymbol{\Sigma}}_{jk}^{-1}(\boldsymbol{o}_t - \bar{\boldsymbol{\mu}}_{jk})(\boldsymbol{o}_t - \bar{\boldsymbol{\mu}}_{jk})^t\,\bar{\boldsymbol{\Sigma}}_{jk}^{-1}$$

$$\Rightarrow \sum_{t=1}^{T}\gamma_t(j,k)\bar{\boldsymbol{\Sigma}}_{jk}\bar{\boldsymbol{\Sigma}}_{jk}^{-1}\bar{\boldsymbol{\Sigma}}_{jk} = \sum_{t=1}^{T}\gamma_t(j,k)\bar{\boldsymbol{\Sigma}}_{jk}\bar{\boldsymbol{\Sigma}}_{jk}^{-1}(\boldsymbol{o}_t - \bar{\boldsymbol{\mu}}_{jk})(\boldsymbol{o}_t - \bar{\boldsymbol{\mu}}_{jk})^t\,\bar{\boldsymbol{\Sigma}}_{jk}^{-1}\bar{\boldsymbol{\Sigma}}_{jk}$$

$$\Rightarrow \bar{\boldsymbol{\Sigma}}_{jk} = \frac{\sum_{t=1}^{T}\left[\gamma_t(j,k)\cdot(\boldsymbol{o}_t - \bar{\boldsymbol{\mu}}_{jk})(\boldsymbol{o}_t - \bar{\boldsymbol{\mu}}_{jk})^t\right]}{\sum_{t=1}^{T}\gamma_t(j,k)}$$

EM Applied to Continuous HMM Training

- The new model parameter set for each mixture component and mixture weight can be expressed as:

$$\overline{\boldsymbol{\mu}}_{jk} = \frac{\sum_{t=1}^{T}\left[\frac{P\left(\boldsymbol{O}, s_t = j, k_t = k \middle| \boldsymbol{\lambda}\right)}{P\left(\boldsymbol{O} \middle| \boldsymbol{\lambda}\right)}\boldsymbol{o}_t\right]}{\sum_{t=1}^{T}\frac{P\left(\boldsymbol{O}, s_t = j, k_t = k \middle| \boldsymbol{\lambda}\right)}{P\left(\boldsymbol{O} \middle| \boldsymbol{\lambda}\right)}} = \frac{\sum_{t=1}^{T}\left[\gamma_t(j,k)\boldsymbol{o}_t\right]}{\sum_{t=1}^{T}\gamma_t(j,k)}$$

$$\overline{\boldsymbol{\Sigma}}_{jk} = \frac{\sum_{t=1}^{T}\left[\frac{P\left(\boldsymbol{O}, s_t = j, k_t = k \middle| \boldsymbol{\lambda}\right)}{P\left(\boldsymbol{O} \middle| \boldsymbol{\lambda}\right)}\left(\boldsymbol{o}_t - \overline{\boldsymbol{\mu}}_{jk}\right)\left(\boldsymbol{o}_t - \overline{\boldsymbol{\mu}}_{jk}\right)^t\right]}{\sum_{t=1}^{T}\frac{P\left(\boldsymbol{O}, s_t = j, k_t = k \middle| \boldsymbol{\lambda}\right)}{P\left(\boldsymbol{O} \middle| \boldsymbol{\lambda}\right)}} = \frac{\sum_{t=1}^{T}\left[\gamma_t(j,k)\left(\boldsymbol{o}_t - \overline{\boldsymbol{\mu}}_{jk}\right)\left(\boldsymbol{o}_t - \overline{\boldsymbol{\mu}}_{jk}\right)^t\right]}{\sum_{t=1}^{T}\gamma_t(j,k)}$$

$$\overline{c}_{jk} = \frac{\sum_{t=1}^{T}\gamma_t(j,k)}{\sum_{t=1}^{T}\sum_{k=1}^{M}\gamma_t(j,k)}$$