HMMs for Acoustic Modeling (Part II)

Lecture 3

CS 753

Instructor: Preethi Jyothi

Recap: HMMs for Acoustic Modeling

✓ What are (first-order) HMMs?

✓ What are the simplifying assumptions governing HMMs?

✓ What are the three fundamental problems related to HMMs?

✓ 1. What is the forward algorithm? What is it used to compute?

Computing Likelihood: Given an HMM $\lambda = (A,B)$ and an observation sequence O, determine the likelihood $P(O|\lambda)$.

✓ 2. What is the Viterbi algorithm? What is it used to compute?

Decoding: Given as input an HMM $\lambda = (A,B)$ and a sequence of observations $O = o_1, o_2, ..., o_T$, find the most probable sequence of states $Q = q_1 q_2 q_3 \ldots q_T$.

Problem 3: Learning in HMMs

Problem 1 (Likelihood):	Given an HMM $\lambda = (A, B)$ and an observation sequence O, determine the likelihood $P(O\|\lambda)$.
Problem 2 (Decoding):	Given an observation sequence O and an HMM $\lambda = (A, B)$, discover the best hidden state sequence Q.
Problem 3 (Learning):	Given an observation sequence O and the set of states in the HMM, learn the HMM parameters A and B.

Learning: Given an observation sequence O and the set of possible states in the HMM, learn the HMM parameters A and B.

Standard algorithm for HMM training: Forward-backward or Baum-Welch algorithm

Forward and Backward Probabilities

Baum-Welch algorithm iteratively estimates transition & observation probabilities and uses these values to derive even better estimates.

Require two probabilities to compute estimates for the transition and observation probabilities:

1. Forward probability: Recall $\alpha_t(j) = P(o_1, o_2 \ldots o_t, q_t = j|\lambda)$

2. Backward probability: $\beta_t(i) = P(o_{t+1}, o_{t+2} \ldots o_T|q_t = i, \lambda)$

Backward probability

1. **Initialization:**

$$\beta_T(i) \;=\; 1, \quad 1 \leq i \leq N$$

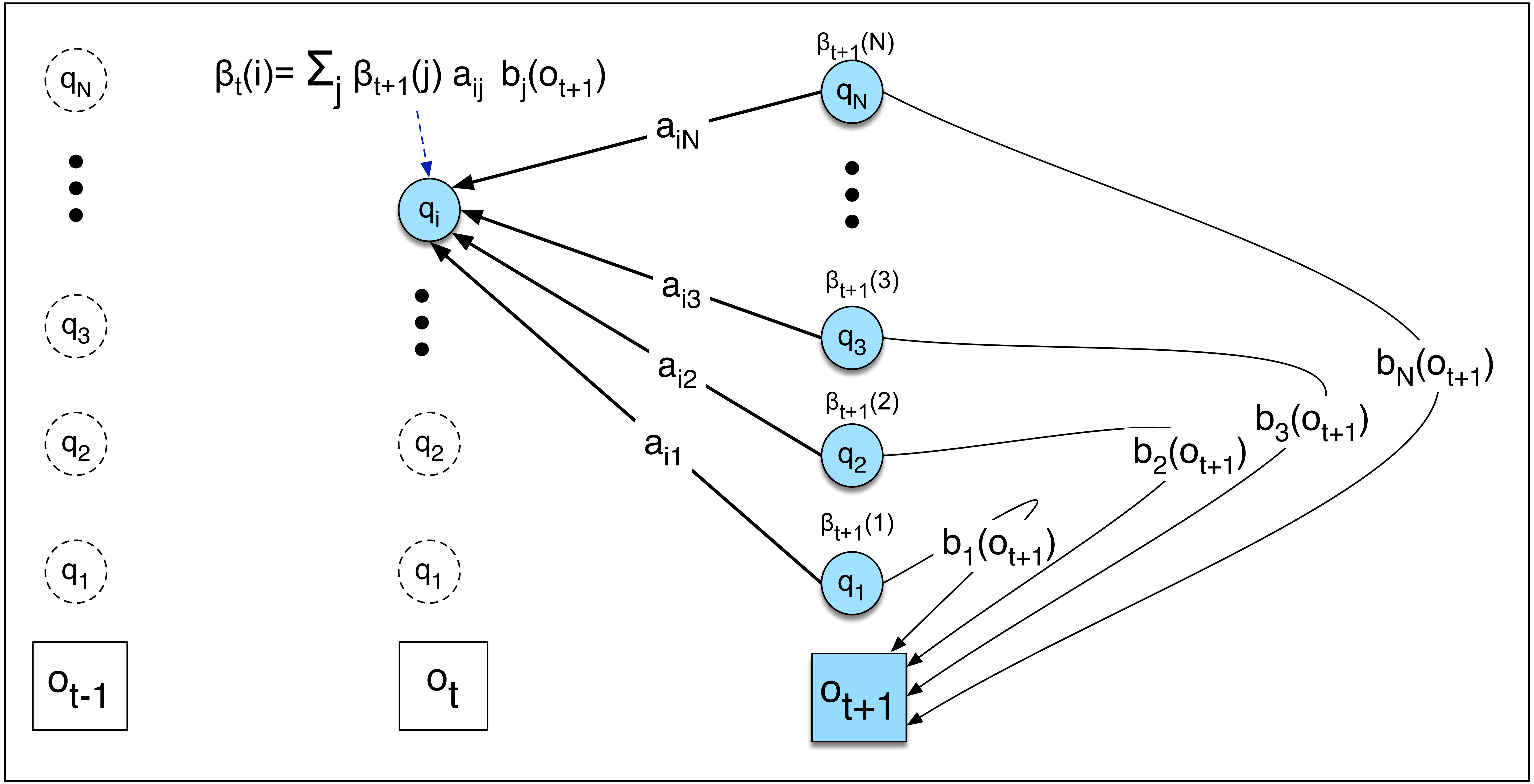
2. **Recursion**

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij}\, b_j(o_{t+1})\, \beta_{t+1}(j), \quad 1 \leq i \leq N, 1 \leq t < T$$

3. **Termination:**

$$P(O|\lambda) = \sum_{j=1}^{N} \pi_j\, b_j(o_1)\, \beta_1(j)$$

Visualising backward probability computation



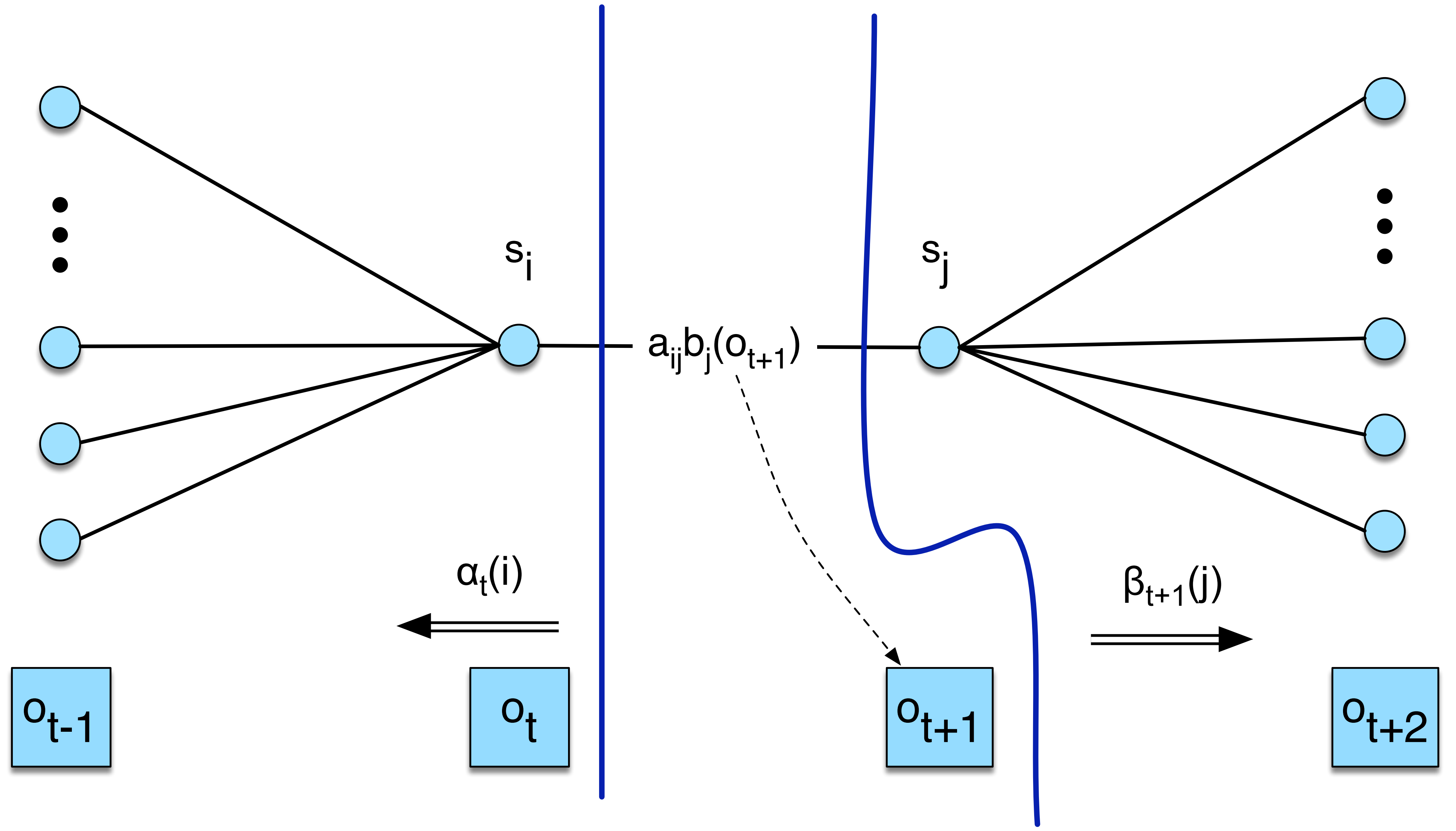
1. Baum-Welch: Estimating a_{ij}

We need to define $\xi_t(i,j)$ to estimate a_{ij}

where $\xi_t(i,j) = P(q_t = i, q_{t+1} = j | O, \lambda)$

which works out to be $\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{j=1}^{N} \alpha_t(j) \beta_t(j)}$

Then, $\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \sum_{k=1}^{N} \xi_t(i,k)}$



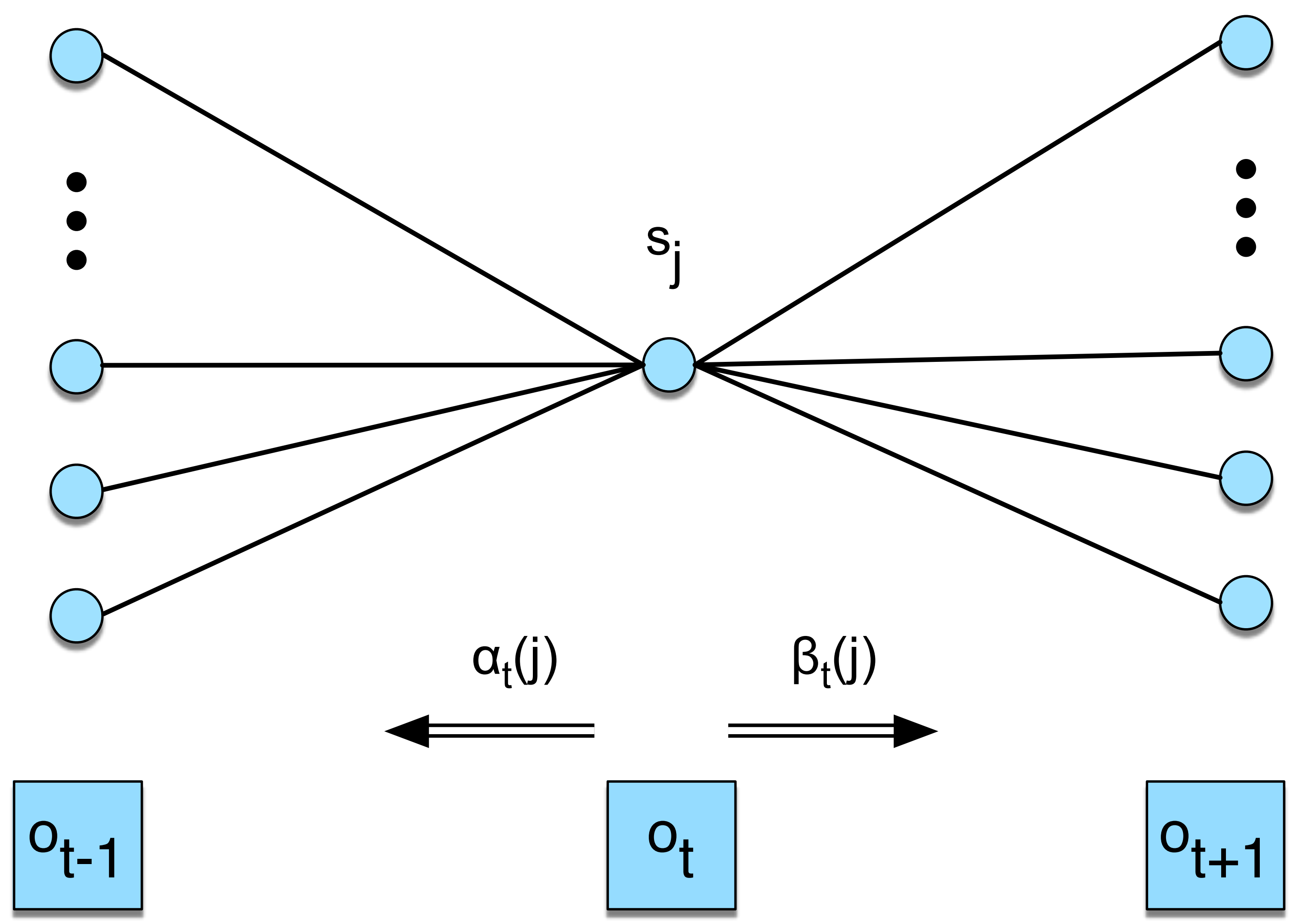
2. Baum-Welch: Estimating $b_j(v_k)$

We need to define $\gamma_t(j)$ to estimate $b_j(v_k)$

where $\gamma_t(j) = P(q_t = j|O,\lambda)$

which works out to be $\gamma_t(j) = \dfrac{\alpha_t(j)\beta_t(j)}{P(O|\lambda)}$ State occupancy probability

Then, $\hat{b}_j(v_k) = \dfrac{\sum_{t=1 s.t. O_t=v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}$ for discrete outputs



Bringing it all together: Baum-Welch

Estimating HMM parameters iteratively using the EM algorithm.
For each iteration, do:

E step: For all time-state pairs, compute the state occupation probabilities $\gamma_t(j)$ and $\xi_t(i, j)$

M step: Reestimate HMM parameters, i.e. transition probabilities, observation probabilities, based on the estimates derived in the E step

Baum-Welch algorithm (pseudocode)

function FORWARD-BACKWARD(*observations* of len T, *output vocabulary V*, *hidden state set Q*) **returns** *HMM=(A,B)*

initialize A and B
iterate until convergence
E-step

$$\gamma_t(j) = \frac{\alpha_t(j)\beta_t(j)}{\alpha_T(q_F)} \quad \forall\, t \text{ and } j$$

$$\xi_t(i,j) = \frac{\alpha_t(i)\,a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\alpha_T(q_F)} \quad \forall\, t,\ i,\ \text{and } j$$

M-step

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \sum_{k=1}^{N} \xi_t(i,k)}$$

$$\hat{b}_j(v_k) = \frac{\sum_{t=1 s.t.\ O_t=v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}$$

return A, B

Discrete to continuous outputs

We derived Baum-Welch updates for discrete outputs.

However, HMMs in acoustic models emit real-valued vectors as observations.

Before we understand how Baum-Welch works for acoustic modelling using HMMs, let's look at an overview of the Expectation Maximization (**EM**) algorithm and establish some notation.

EM Algorithm: Fitting Parameters to Data

Observed data: i.i.d samples x_i, i=1, …, N

Goal: Find $\arg\max_\theta \mathcal{L}(\theta)$ where $\mathcal{L}(\theta) = \sum_{i=1}^{N} \log \Pr(x_i; \theta)$

Initial parameters: θ^0 (x is observed and z is hidden)

Iteratively compute θ^ℓ as follows:

$$Q(\theta, \theta^{\ell-1}) = \sum_{i=1}^{N} \sum_{z} \Pr(z | x_i; \theta^{\ell-1}) \log \Pr(x_i, z; \theta)$$

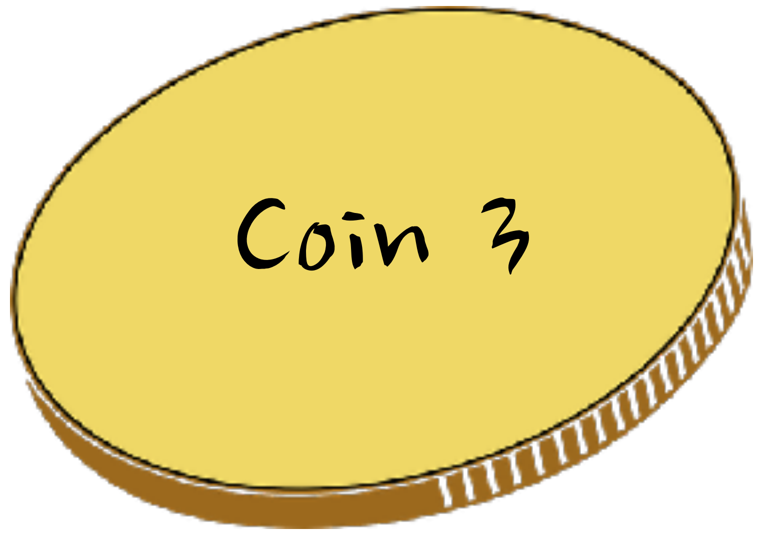
$$\theta^\ell = \arg\max_\theta Q(\theta, \theta^{\ell-1})$$

Estimate θ^ℓ cannot get worse over iterations because for all θ:

$$\mathcal{L}(\theta) - \mathcal{L}(\theta^{\ell-1}) \geq Q(\theta, \theta^{\ell-1}) - Q(\theta^{\ell-1}, \theta^{\ell-1})$$

EM is guaranteed to converge to a local optimum or saddle points [Wu83]

Coin example to illustrate EM



$\rho_1 = \Pr(\mathrm{H})$ $\rho_2 = \Pr(\mathrm{H})$ $\rho_3 = \Pr(\mathrm{H})$

Repeat:

Toss Coin 1 privately

if it shows H:

Toss Coin 2 twice

else

Toss Coin 3 twice

The following sequence is observed: "HH, TT, HH, TT, HH"

How do you estimate ρ_1, ρ_2 and ρ_3?

Coin example to illustrate EM

Recall, for partially observed data, the log likelihood is given by:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log \Pr(x_i; \theta) = \sum_{i=1}^{N} \log \sum_{z} \Pr(x_i, z; \theta)$$

where, for the coin example:

- each observation $x_i \in \mathcal{X} = \{\text{HH,HT,TH,TT}\}$
- the hidden variable $z \in \mathcal{Z} = \{\text{H,T}\}$

Coin example to illustrate EM

Recall, for partially observed data, the log likelihood is given by:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log \Pr(x_i; \theta) = \sum_{i=1}^{N} \log \sum_{z} \Pr(x_i, z; \theta)$$

$$\Pr(x, z; \theta) = \Pr(x|z; \theta) \Pr(z; \theta)$$

ρ_1 = Pr(H)

ρ_2 =Pr(H)

ρ_3 = Pr(H)

where $\Pr(z; \theta) = \begin{cases} \rho_1 & \text{if } z = \text{H} \\ 1 - \rho_1 & \text{if } z = \text{T} \end{cases}$

$$\Pr(x|z; \theta) = \begin{cases} \rho_2^h (1 - \rho_2)^t & \text{if } z = \text{H} \\ \rho_3^h (1 - \rho_3)^t & \text{if } z = \text{T} \end{cases}$$

h : number of heads, t : number of tails

Coin example to illustrate EM

Our observed data is: {HH, TT, HH, TT, HH}

Let's use EM to estimate $\theta = (\rho_1, \rho_2, \rho_3)$

[EM Iteration, E-step]

Compute quantities involved in

$$Q(\theta, \theta^{\ell-1}) = \sum_{i=1}^{N} \sum_{z} \gamma(z, x_i) \log \Pr(x_i, z; \theta)$$

where $\gamma(z, x) = \Pr(z \mid x ; \theta^{\ell-1})$

i.e., compute $\gamma(z, x_i)$ for all z and all i

Suppose $\theta^{\ell-1}$ is $\rho_1 = 0.3$, $\rho_2 = 0.4$, $\rho_3 = 0.6$:

What is $\gamma(\mathrm{H}, \mathrm{HH})$? $= 0.16$

What is $\gamma(\mathrm{H}, \mathrm{TT})$? $= 0.49$

Coin example to illustrate EM

Our observed data is: {HH, TT, HH, TT, HH}

Let's use EM to estimate θ = (ρ_1, ρ_2, ρ_3)

[EM Iteration, M-step]

Find θ which maximises

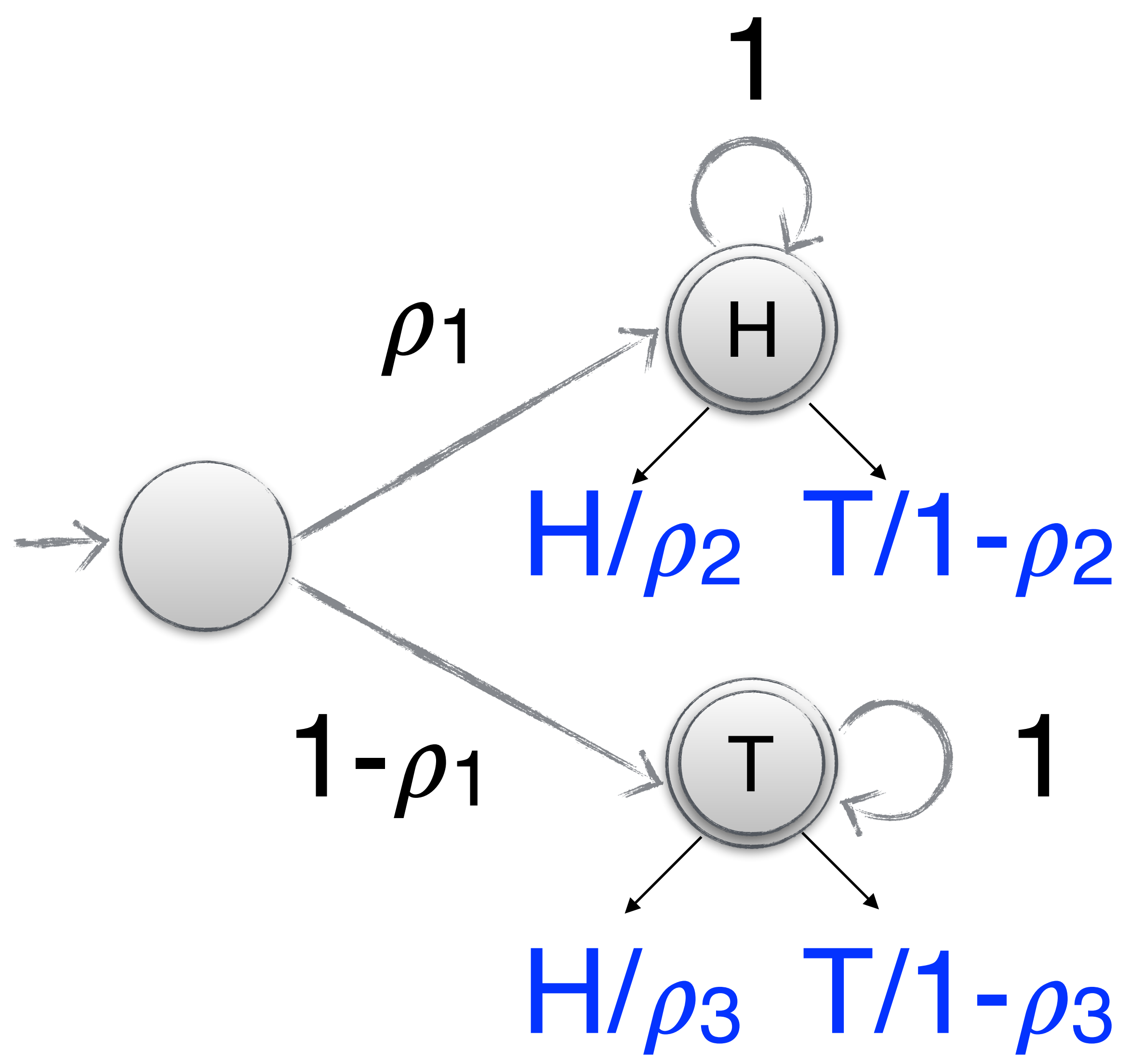
$$Q(\theta, \theta^{\ell-1}) = \sum_{i=1}^{N} \sum_{z} \gamma(z, x_i) \log \Pr(x_i, z; \theta)$$

$$\rho_1 = \frac{\sum_{i=1}^{N} \gamma(\mathrm{H}, x_i)}{N}$$

$$\rho_2 = \frac{\sum_{i=1}^{N} \gamma(\mathrm{H}, x_i) h_i}{\sum_{i=1}^{N} \gamma(\mathrm{H}, x_i)(h_i + t_i)}$$

$$\rho_3 = \frac{\sum_{i=1}^{N} \gamma(\mathrm{T}, x_i) h_i}{\sum_{i=1}^{N} \gamma(\mathrm{T}, x_i)(h_i + t_i)}$$

Coin example to illustrate EM



This was a very simple HMM
(with observations from 2 states)

State remains the same after the first transition

γ estimated the distribution of this state

More generally, will need the distribution of the state at each time step

EM for general HMMs: Baum-Welch algorithm (1972)

(predates the general formulation of EM (1977))

Baum-Welch Algorithm as EM

Observed data: N sequences, x_i, $i=1\ldots N$ where $x_i \in \mathrm{V}$

Parameters θ : transition matrix A, observation probabilities B

[EM Iteration, E-step]

Compute quantities involved in $Q(\theta, \theta^{\ell\text{-}1})$

$\gamma_{\mathrm{i,t}}(j) = \Pr(z_t = j \mid x_i ; \theta^{\ell\text{-}1})$

$\xi_{i,t}(j,k) = \Pr(z_t = j,\, z_{t+1} = k \mid x_i ; \theta^{\ell\text{-}1})$

Baum-Welch Algorithm as EM

Observed data: N sequences, x_i, i=1…N where $x_i \in$ V

Parameters θ : transition matrix A, observation probabilities B

[EM Iteration, M-step]

Find θ which maximises $Q(\theta,\theta^{\ell\text{-}1})$

$$A_{j,k} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T_i-1} \xi_{i,t}(j,k)}{\sum_{i=1}^{N} \sum_{t=1}^{T_i-1} \sum_{k'} \xi_{i,t}(j,k')}$$

$$B_{j,v} = \frac{\sum_{i=1}^{N} \sum_{t:x_{it}=v} \gamma_{i,t}(j)}{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \gamma_{i,t}(j)}$$

Discrete to continuous outputs

We derived Baum-Welch updates for discrete outputs.

However, HMMs in acoustic models emit real-valued vectors as observations.

Use probability density functions to define observation probabilities

If x were 1D values, HMM observation probabilities: $b_j(x) = \mathcal{N}(x \mid \mu_j, \sigma_j^2)$
where μ_j is the mean associated with state j and σ_j^2 is its variance

If $\mathbf{x} \in \mathbb{R}^d$, then we use multivariate Gaussians, $b_j(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_j, \Sigma_j)$
where Σ_j is the covariance matrix associated with state j

BW for Gaussian Observation Model

Observed data: N sequences, $x_i = (x_{i1}, \ldots, x_{iT_i})$, $i=1\ldots N$ where $x_{it} \in \mathbb{R}^d$

Parameters θ : transition matrix A, observation prob. B = {(μ$_j$,Σ$_j$)} for all j

[EM Iteration, M-step]

Find θ which maximises $Q(\theta, \theta^{\ell\text{-}1})$

A same as with discrete outputs

$$\mu_j = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \gamma_{i,t}(j) x_{it}}{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \gamma_{i,t}(j)}$$

$$\Sigma_j = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \gamma_{i,t}(j)(x_{it} - \mu_j)(x_{it} - \mu_j)^T}{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \gamma_{i,t}(j)}$$

Gaussian Mixture Model

- Assuming that observations associated with a state follow a Gaussian distribution is too simplistic.
- More generally, we use a "mixture of Gaussians" to allow for acoustic vectors associated with a state to be non-Gaussian.
- Instead of $b_j(\mathbf{x}) = \mathcal{N}(\mathbf{x} \,|\, \boldsymbol{\mu}_j, \Sigma_j)$ in the single Gaussian case, $b_j(\mathbf{x})$ can be an M-component mixture model:

$$b_j(\mathbf{x}) = \sum_{m=1}^{M} c_{jm} \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_{jm}, \Sigma_{jm})$$

where c_{jm} is the mixing probability for Gaussian component m of state j

$$\sum_{m=1}^{M} c_{jm} = 1, \;\; c_{jm} \geq 0$$

BW for Gaussian Mixture Model

Observed data: N sequences, $x_i = (x_{i1}, \ldots, x_{iT_i})$, $i=1 \ldots N$ where $x_{it} \in \mathbb{R}^d$

Parameters θ : transition matrix A, observation prob. B = {(μ_{jm},Σ_{jm},c_{jm})} for all j,m

[EM Iteration, M-step]

Find θ which maximises $Q(\theta, \theta^{\ell-1})$

$$\mu_{jm} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \gamma_{i,t}(j,m) x_{it}}{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \gamma_{i,t}(j,m)}$$

$$\Sigma_{jm} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \gamma_{i,t}(j,m)(x_{it} - \mu_{jm})(x_{it} - \mu_{jm})^T}{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \gamma_{i,t}(j,m)}$$

$$c_{jm} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \gamma_{i,t}(j,m)}{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \sum_{m'=1}^{M} \gamma_{i,t}(j,m')}$$

Prob. of component m of state j at time t

Baum Welch: In summary

[Every EM Iteration]

Compute θ = { A_{jk}, (μ_{jm},Σ_{jm},c_{jm}) } for all j,k,m

$$A_{j,k} = \frac{\sum_{i=1}^{N}\sum_{t=2}^{T_i} \xi_{i,t}(j,k)}{\sum_{i=1}^{N}\sum_{t=2}^{T_i}\sum_{k'} \xi_{i,t}(j,k')}$$

$$\mu_{jm} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T_i} \gamma_{i,t}(j,m)x_{it}}{\sum_{i=1}^{N}\sum_{t=1}^{T_i} \gamma_{i,t}(j,m)}$$

$$\Sigma_{jm} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T_i} \gamma_{i,t}(j,m)(x_{it}-\mu_{jm})(x_{it}-\mu_{jm})^T}{\sum_{i=1}^{N}\sum_{t=1}^{T_i} \gamma_{i,t}(j,m)}$$

$$c_{jm} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T_i} \gamma_{i,t}(j,m)}{\sum_{i=1}^{N}\sum_{t=1}^{T_i}\sum_{m'=1}^{M} \gamma_{i,t}(j,m')}$$