Hidden Markov Models for Automatic Speech Recognition Part I

R. Hegde
EE627 : Speech Signal Processing
Dept. of Electrical Engg. IIT Kanpur

Bayes Rule and ASR

BAYES RULE

$$P(W|A) = \frac{P(A|W)\,P(W)}{P(A)}$$

$P(A|W)$ → is of interest in ASR

HMM Targets → $P(A|W)$

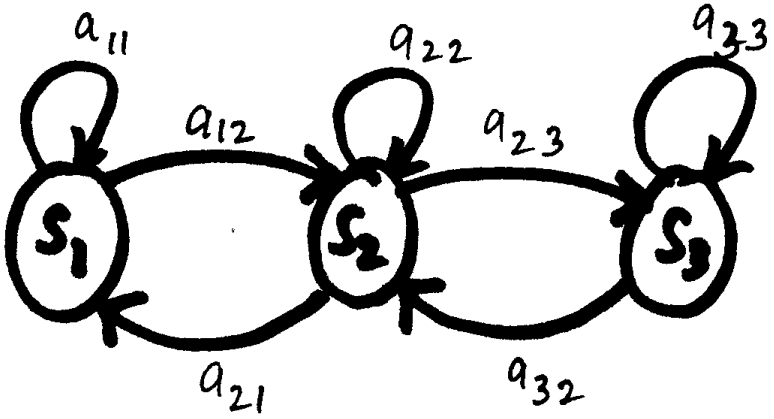
NOTATION MAPPING IN ASR

$A \rightarrow O$

$W \rightarrow \lambda$

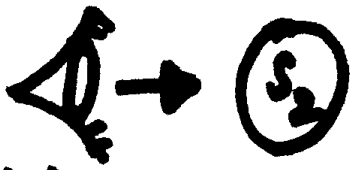
$P(A|W) \rightarrow P(O|\lambda)$

The Markovian Model



Has 'N' states
Here N=3

Let t : Current time step; q_t : is the state at 't'
t+1 : Next time step



Looking into

$P(q_{t+1} = S_1 \mid q_t = S_2) \rightarrow a_{21}$

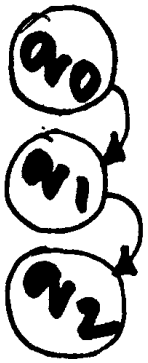
$P(q_{t+1} = S_2 \mid q_t = S_2) \rightarrow a_{22}$

$P(q_{t+1} = S_3 \mid q_t = S_2) \rightarrow a_{23}$

Similar Expressions can be generated for looking into each state!!

A Simple Bayes Network

Consider a simple Bayes Net to model the jt. distribution (q_0, q_1, q_2)



State i	1	2
1	a_{11}	a_{12}
2	a_{21}	a_{22}

Note: In practice We usually have a dummy Start and End states.

* $P(q_t = s) = \sum P(Q)$, where Q consists of all paths of length 't', that end in 's'

* Let $Q = q_1, q_2, q_3, \cdots q_t$ then

$$P(Q) = P(q_1 q_2 \cdots q_t) = P(q_1 q_2 \cdots q_{t-1}) P(q_t | q_{t-1}) = P(q_2|q_1) P(q_3|q_2) \cdots P(q_t|q_{t-1})$$

Blind Estimation of State Probability

'Blind Estimation' of $P(q_t = s)$

$$P(q_1 q_2 \cdots q_t) = P(q_1 q_2 \cdots q_{t-1}) P(q_t | q_1 q_2 \cdots q_{t-1})$$

$$= P(q_1 q_2 \cdots q_{t-1}) P(q_t | q_{t-1})$$

$$= P(q_2 | q_1) P(q_3 | q_2) \cdots P(q_t | q_{t-1})$$

$$\therefore P(q_t = s) = \sum_{Q \in \text{Paths of length } t \text{ that end in } s} P(Q)$$

$O(N^t)$ — computing cost

Estimating State Probability with DP

Estimating $P(q_t = S)$ via DP [unconditionally with no observed evid.]

Define $p_t(i)$ = Probability at time 't' is s_i

$$p_t(i) = P(q_t = s_i)$$

By Induction

$$\forall i \quad p_0(i) = \begin{cases} 1 ; \text{if } s_i \text{ is the start state} \\ 0 ; \text{otherwise} \end{cases}$$

$$\forall j \quad p_{t+1}(j) = P(q_{t+1} = s_j) = \sum_{i=1}^{N} P(q_{t+1} = s_j \,\&\, q_t = s_i)$$

$$= \sum_{i=1}^{N} P(q_{t+1} = s_j \mid q_t = s_i) P(q_t = s_i) = \sum_{i=1}^{N} a_{ij} p_t(i)$$

$$\therefore P(q_t = S) = \sum_{i=1}^{N} a_{ij} p_t(i)$$

? Associate Observations with States

Lets look at a model where we can 'observe some stuff' which are effected by each state S_i

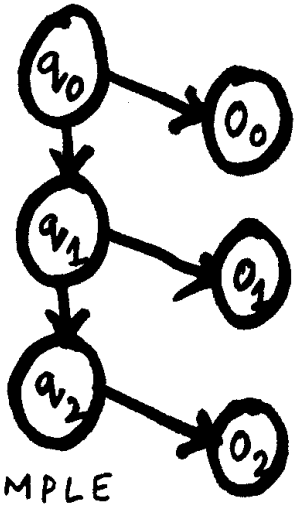
→ Observation symbols O_t → unreliable but significant

Assume O_t is conditionally independent of $\{ q_{t-1}\, q_{t-2} \cdots q_1, q_0, O_{t-1}\, O_{t-2} \cdots O_1, O_0 \}$ given q_t

Associate Observations with states

Given the conditional independence of O_t wrt $\{a_{t-1}\ a_{t-2} \cdots a_1 a_0, O_{t-1}\ O_{t-2} \cdots O_1\ O_0\}$ given a_t

$$P(O_t = X \mid a_t = s_i) = P(O_t = X \mid a_t = s_i, \text{any earlier history})$$



SIMPLE
Bayes Net to represent
$(a_0, a_1, a_2, O_0, O_1, O_2)$

i \ P	$P(O_t=1 \mid a_t=s_i)$	$\cdots$	$P(O_t=k \mid a_t=s_i)$	$\cdots$	$P(O_t=M \mid a_t=s_i)$
1	$b_1(1)$	$\cdots$	$b_1(k)$	$\cdots$	$b_1(M)$
2	$b_2(1)$	$\cdots$	$b_2(k)$	$\cdots$	$b_2(M)$
3	$b_3(1)$	$\cdots$	$b_3(k)$	$\cdots$	$b_3(M)$
$\vdots$					
i	$b_i(1)$	$\cdots$	$b_i(k)$	$\cdots$	$b_i(M)$
N	$b_N(1)$	$\cdots$	$b_N(k)$	$\cdots$	$b_N(M)$

$$b_i(k) = P(O_t = k \mid a_t = s_i)$$

HMM Basic Questions

Hidden Markov Models : Questions?

1: State Estimation

$$P(q_T = S_i \mid O_1 O_2 \cdots O_T) = ?$$

Use DP

2: Most probable path

Given $O_1 O_2 \cdots O_T$, what is the most probable path?

Viterbi Algorithm

3: Learning HMM

Given $O_1 O_2 \cdots O_T$ what is the max likelihood HMM that could have produced this

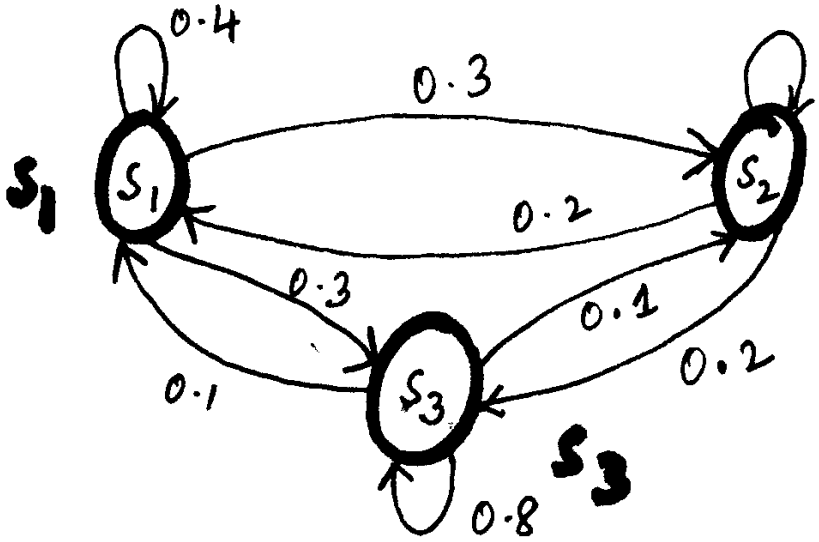
EM

Observable Markov Model

Observable Markov Model : Eg.

Each State → deterministically observable event

Weather Model at t = Noon (once a day)



S_1

S_2

S_3

S1: Precipitation
S2: Cloudy
S3: Sunny

$$\{a_{ij}\} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

$\{a_{ij}\}$ ← A

Matrix of state trans. probabilities.

Example Problem

What is the probability (according to the model) that the weather for eight consecutive days is "sun-sun-sun-rain-rain-sun-cloudy-sun"?

Solution

We define the observation sequence, **O**, as

O	=	(sunny,	sunny,	sunny,	rain,	rain,	sunny,	cloudy,	sunny)
	=	(3,	3,	3,	1,	1,	3,	2,	3)
day		1	2	3	4	5	6	7	8

corresponding to the postulated set of weather conditions over the eight-day period and we want to calculate $P(\mathbf{O}|\text{Model})$, the probability of the observation sequence **O**, given the model of Figure 6.2. We can directly determine $P(\mathbf{O}|\text{Model})$ as:

$$\begin{aligned} P(\mathbf{O}|\text{Model}) &= P[3, 3, 3, 1, 1, 3, 2, 3|\text{Model}] \\ &= P[3]\,P[3|3]^2 P[1|3]\,P[1|1] \\ &\quad P[3|1]\,P[2|3]\,P[3|2] \\ &= \pi_3 \cdot (a_{33})^2\, a_{31}\, a_{11}\, a_{13}\, a_{32}\, a_{23} \\ &= (1.0)(0.8)^2(0.1)(0.4)(0.3)(0.1)(0.2) \\ &= 1.536 \times 10^{-4} \end{aligned}$$

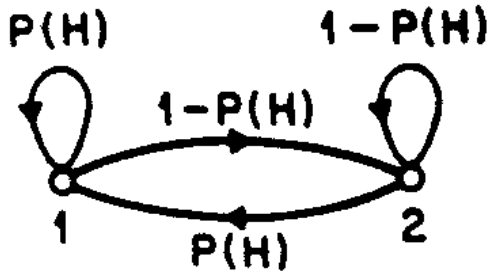
where we use the notation:

$$\pi_i = P[q_1 = i], \qquad 1 \leq i \leq N \tag{6.4}$$

to denote the initial state probabilities.

1 and 2 coin model

(a)

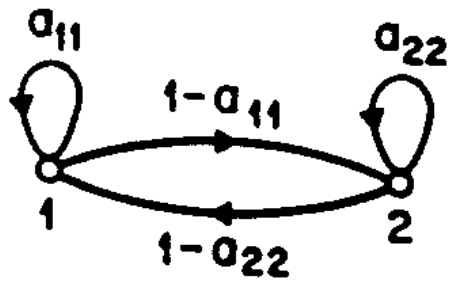


HEADS TAILS

1-COIN MODEL
(OBSERVABLE MARKOV MODEL)

$O = H\ H\ T\ T\ H\ T\ H\ H\ T\ T\ H\ \ldots$
$S = 1\ 1\ 2\ 2\ 1\ 2\ 1\ 1\ 2\ 2\ 1\ \ldots$

(b)



$P(H) = P_1$ $P(H) = P_2$

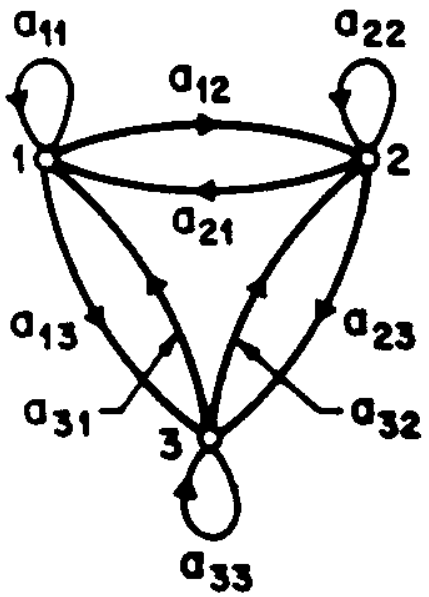
$P(T) = 1-P_1$ $P(T) = 1-P_2$

2-COINS MODEL
(HIDDEN MARKOV MODEL)

$O = H\ H\ T\ T\ H\ T\ H\ H\ T\ T\ H\ \ldots$
$S = 2\ 1\ 1\ 2\ 2\ 2\ 1\ 2\ 2\ 1\ 2\ \ldots$

3 coin model

(c)



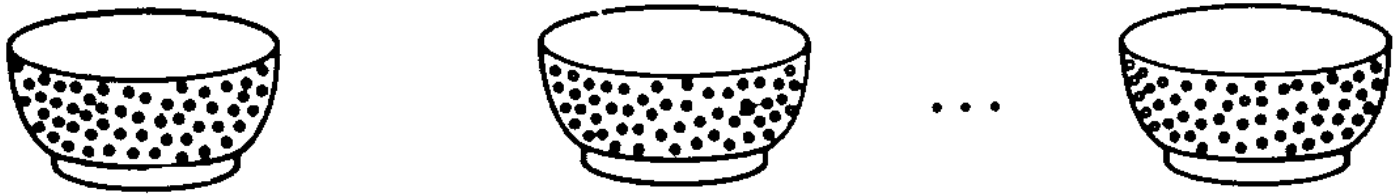
3-COINS MODEL
(HIDDEN MARKOV MODEL)

$O = H\ H\ T\ T\ H\ T\ H\ H\ T\ T\ H\ \ldots$
$S = 3\ 1\ 2\ 3\ 3\ 1\ 1\ 2\ 3\ 1\ 3\ \ldots$

STATE

	1	2	3
$P(H)$	P_1	P_2	P_3
$P(T)$	$1-P_1$	$1-P_2$	$1-P_3$

Urn Ball Model



URN 1

$P(RED) = b_1(1)$
$P(BLUE) = b_1(2)$
$P(GREEN) = b_1(3)$
$P(YELLOW) = b_1(4)$
$\vdots$
$P(ORANGE) = b_1(M)$

URN 2

$P(RED) = b_2(1)$
$P(BLUE) = b_2(2)$
$P(GREEN) = b_2(3)$
$P(YELLOW) = b_2(4)$
$\vdots$
$P(ORANGE) = b_2(M)$

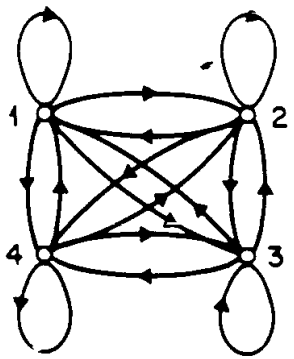
URN N

$P(RED) = b_N(1)$
$P(BLUE) = b_N(2)$
$P(GREEN) = b_N(3)$
$P(YELLOW) = b_N(4)$
$\vdots$
$P(ORANGE) = b_N(M)$

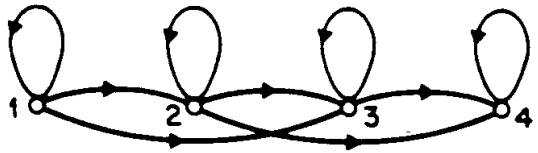
$O = \{GREEN, GREEN, BLUE, RED, YELLOW, RED, \ldots\ldots, BLUE\}$

Figure 6.4 An N-state urn-and-ball model illustrating the general case of a discrete symbol HMM.

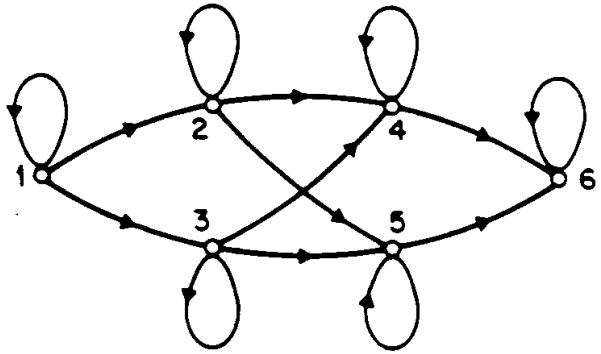
HMM types



(a)



(b)



(c)

Basic Elements of A DHMM

- number of states in the model : N
  individual state : $S = \{s_1, s_2, \cdots, s_N\}$
- number of distinct obser. symbols per state : M
  symbol set : $V = \{v_1, v_2, \cdots, v_M\}$
- initial state distribution : $\pi = \{\pi_i\}$

  $$\pi_i = P(s_i \text{@time } 1)$$

- state transition probability distribution : $A = \{a_{ij}\}$

  $$a_{ij} = P(s_j \text{@time } t+1 \mid s_i \text{@time } t)$$

- observation symbol probability distribution in s_j : $B = \{b_j(k)\}$

  $$b_j(k) = P(v_k \text{@time } t \mid s_j \text{@time } t)$$

Basic Elements in CHMM

- number of states in the model : N
  individual state : $S = \{s_1, s_2, \cdots, s_N\}$
- number of mixtures per state : M
- initial state distribution : $\pi = \{\pi_i\}$

  $$\pi_i = P(s_i \text{@time } 1)$$

- state transition probability distribution : $A = \{a_{ij}\}$

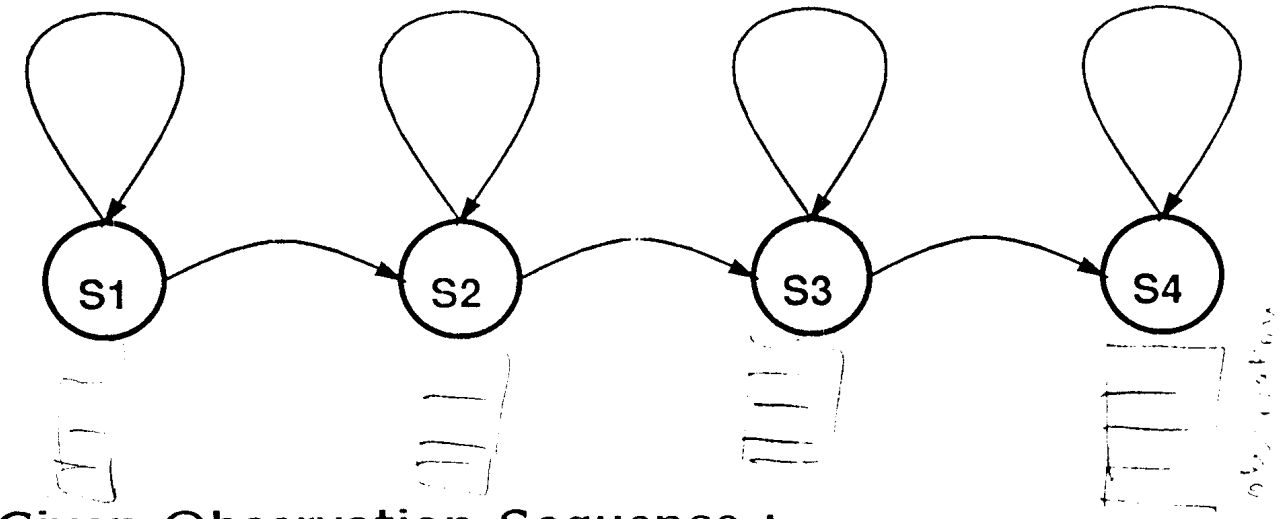
  $$a_{ij} = P(s_j \text{@time } t+1 \mid s_i \text{@time } t)$$

- continuous observation density in s_j : $B = \{b_j(\mathbf{x})\}$

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

  note : mixture coefficients

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



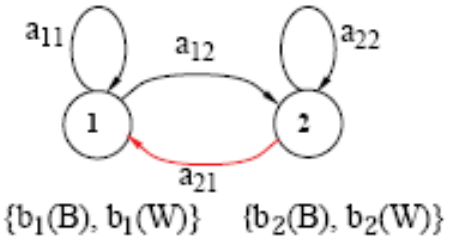
Given Observation Sequence :

- number of observations in the sequence : T
- discrete observation sequence :
  $\mathcal{O} = o_1, o_2, \cdots, o_T, \qquad o_t \in V$ (i.e. code symbol)
- continuous Observation Sequence :
  $\mathcal{O} = \mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_T, \qquad \mathbf{o}_t$: feature vector

Consider Corresponding State Sequence :

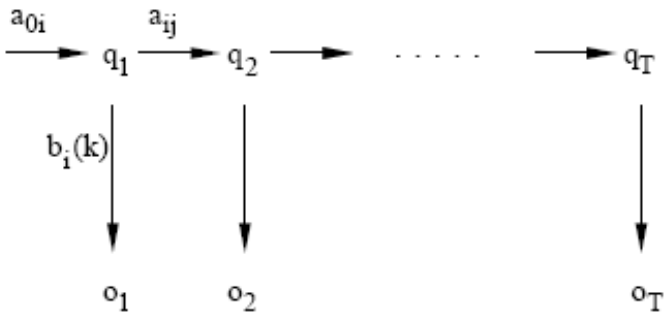
- $\mathcal{Q} = q_1, q_2, \cdots, q_T, \qquad q_t \in S$

Example HMM and Generation of Observation Symbols

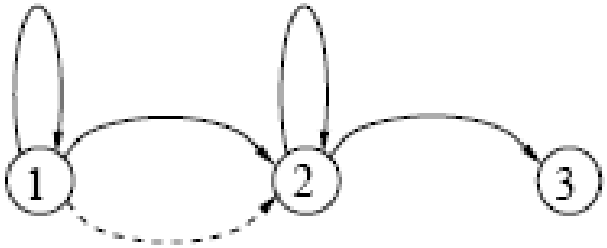


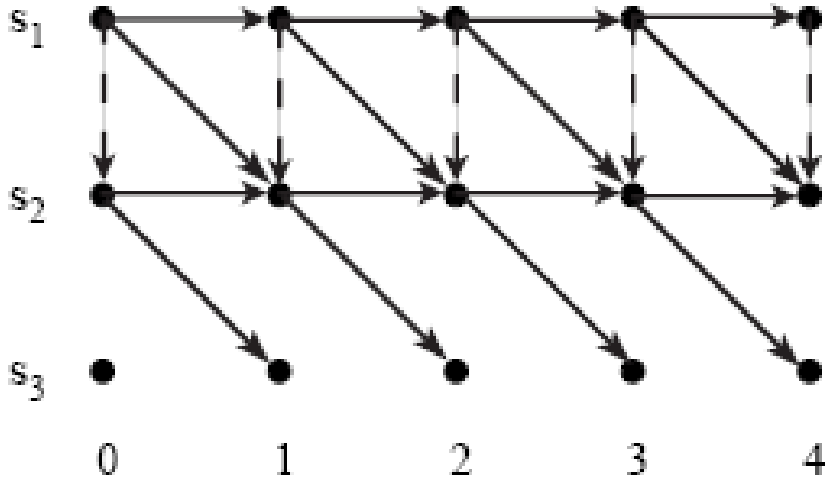
$$\pi = \{a_{01}, a_{02}\}, \quad \boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \text{ and } \quad \boldsymbol{B} = \begin{bmatrix} b_1(B) & b_1(W) \\ b_2(B) & b_2(W) \end{bmatrix}$$

1. Choose an initial state, $q_1 = s_i$, based on the initial state distribution, π
2. For $t = 1$ to T:
   - Choose $o_t = v_k$ according to the symbol probability distribution in state s_i, $b_i(k)$
   - Transition to a new state $q_{t+1} = s_j$ according to the state transition probability distribution for state s_i, a_{ij}
3. Increment t by 1, return to step 2 if $t \leq T$; else, terminate



State Diagram and Trellis





Three Problems of HMM :

- Problem (1)

  Given the observation sequence $\mathcal{O}$ and the model λ, how do we compute $P(\mathcal{O}|\lambda)$ efficiently?

- Problem (2)

  How do we adjust model parameters λ to maximize $P(\mathcal{O}|\lambda)$?

- Problem (3)

  Given the observation sequence $\mathcal{O}$ and the model λ, how do we choose a state sequence $\mathcal{Q}$, which is optimal in some meaningful sense (that is, best "explains" the observations)?

Problem 1 (Training)

Problem (1) – Evaluation Problem

Given the observation sequence $\mathcal{O}$ and the model λ, how do we compute $P(\mathcal{O}|\lambda)$?

Solution :

$$P(\mathcal{O}|\lambda) = \sum_{\mathcal{Q}} P(\mathcal{O}, \mathcal{Q}|\lambda) = \sum_{\mathcal{Q}} P(\mathcal{O}|\mathcal{Q}, \lambda) P(\mathcal{Q}|\lambda)$$

Probability of state sequence $\mathcal{Q}$:

$$P(\mathcal{Q}|\lambda) = \pi_{q_1} \cdot a_{q_1 q_2} \cdot a_{q_2 q_3} \cdots a_{q_{T-1} q_T}$$

Problem 1 : Contd.

Probability of observation sequence $\mathcal{O}$ for $\mathcal{Q}$:

$$\begin{aligned} P(\mathcal{O}|\mathcal{Q},\lambda) &= \prod_{t=1}^{T} P(o_t|q_t,\lambda) \\ &= b_{q_1}(o_1)\cdot b_{q_2}(o_2)\cdots b_{q_T}(o_T) \end{aligned}$$

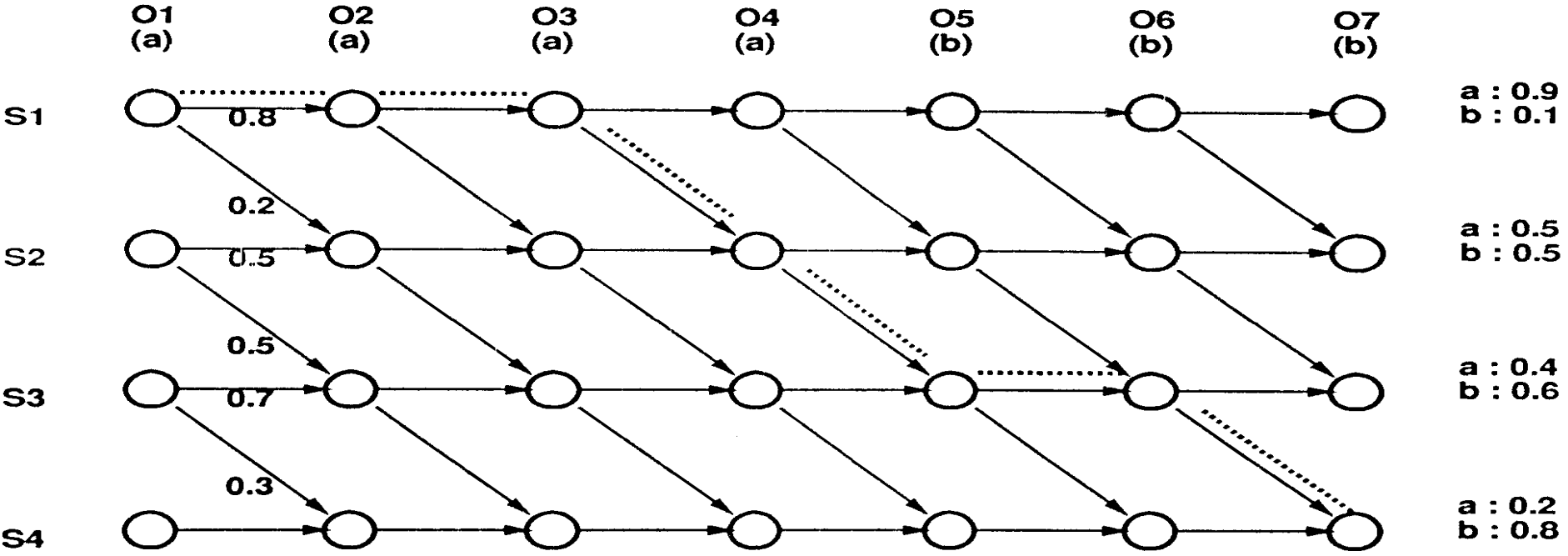
(assumption : statistical independence of observations)

Probability of $\mathcal{O}$ given the model λ :

$$P(\mathcal{O}|\lambda) = \sum_{q_1,q_2,\cdots,q_T} \pi_{q_1}\cdot b_{q_1}(o_1)\cdot a_{q_1q_2}\cdot b_{q_2}(o_2)\cdots \cdots a_{q_{T-1}q_T}\cdot b_{q_T}(o_T)$$

(not very efficient solution)

Computation Example for Problem (1) :



$$P(\mathcal{O}, \mathcal{Q}|\lambda)$$

$$= \pi_1 \cdot b_1(o_1) \cdot a_{11} b_1(o_2) \cdot a_{11} b_1(o_3) \cdot a_{12} b_2(o_4) \cdot$$

$$a_{23} b_3(o_5) \cdot a_{33} b_3(o_6) \cdot a_{34} b_4(o_7)$$

$$= (1 \cdot 0.9) \times (0.8 \cdot 0.9) \times (0.8 \cdot 0.9) \times (0.2 \cdot 0.5) \times$$

$$(0.5 \cdot 0.6) \times (0.7 \cdot 0.6) \times (0.3 \cdot 0.8)$$

$$= 0.00141 \cdots$$

Efficient Solution – Forward-Backward Procedure

Forward variables (alpha terms) :

$$\alpha(t,i) = P(o_1, o_2, \cdots, o_t,\ s_i\text{@time } t \mid \lambda)$$

Forward Procedure :

Initialization :

$$\alpha(1,i) = \pi_i b_i(o_1), \qquad i = 1, 2, \cdots, N$$

Induction :

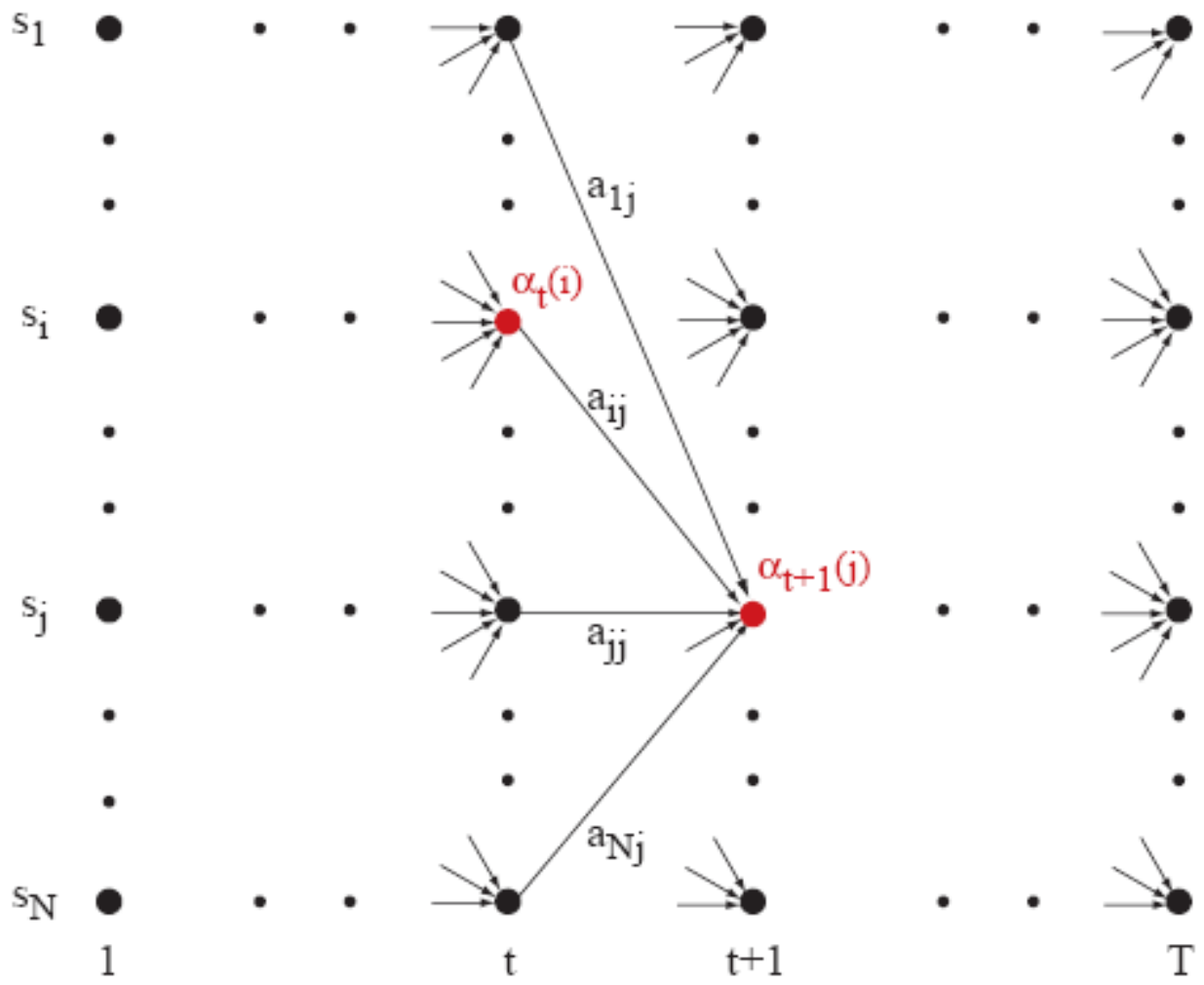
$$\alpha(t+1,j) = \left[\sum_{i=1}^{N} \alpha(t,i) a_{ij}\right] b_j(o_{t+1})$$

$$j = 1, 2, \cdots, N \qquad t = 1, 2, \cdots, T-1$$

Termination :

$$P(\mathcal{O}|\lambda) = \sum_{i=1}^{N} \alpha(T,i)$$

Forward Procedure - Illustration



Efficient Solution – Forward-Backward Procedure

Backward variables (beta terms) :

$$\beta(t,i) = P(o_{t+1}, o_{t+2}, \cdots, o_T \mid s_i \text{@time } t,\ \lambda)$$

Backward Procedure :

Initialization :

$$\beta(T,i) = \begin{cases} 1 & i = N \\ 0 & i \neq N \end{cases}$$

Induction :

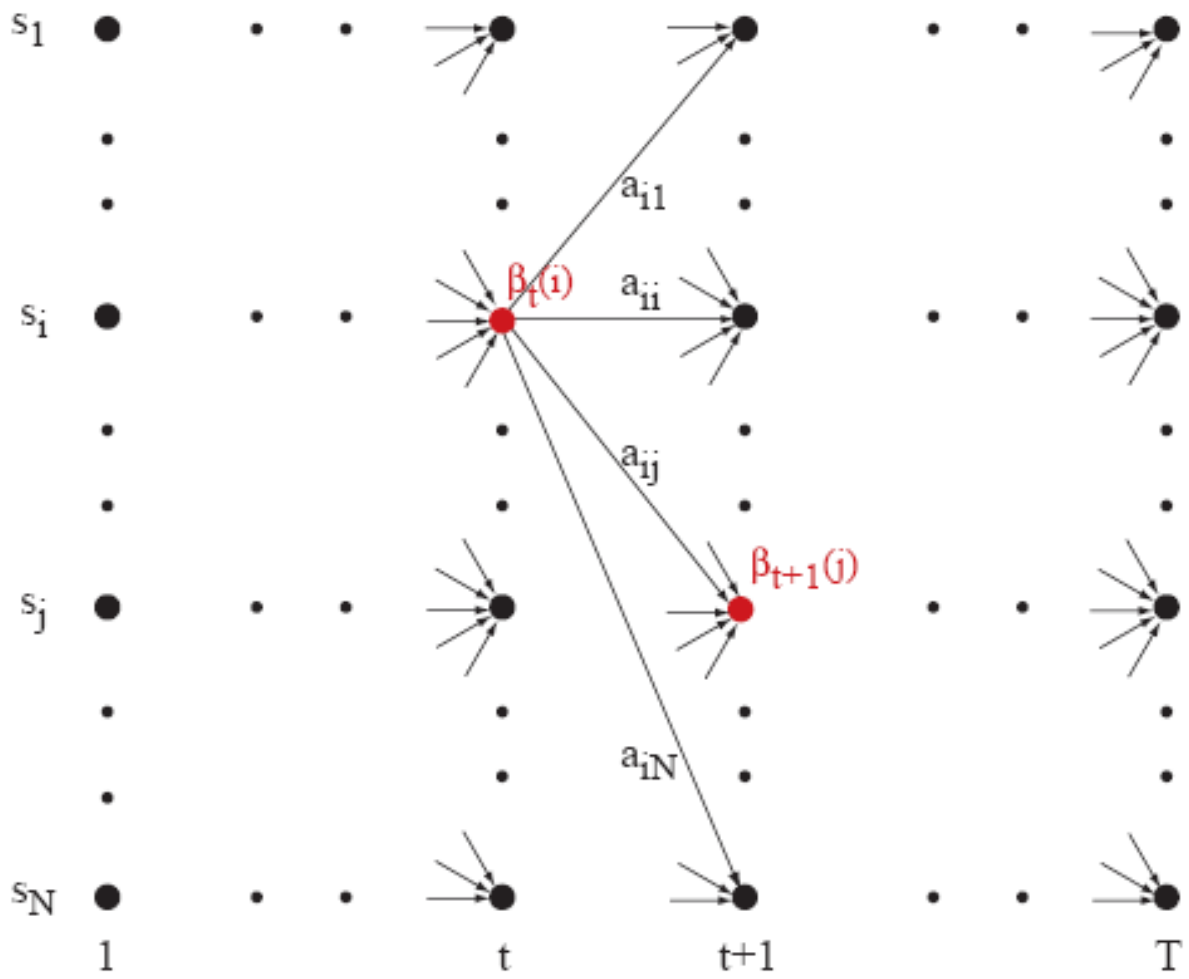
$$\beta(t,i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta(t+1,j)$$

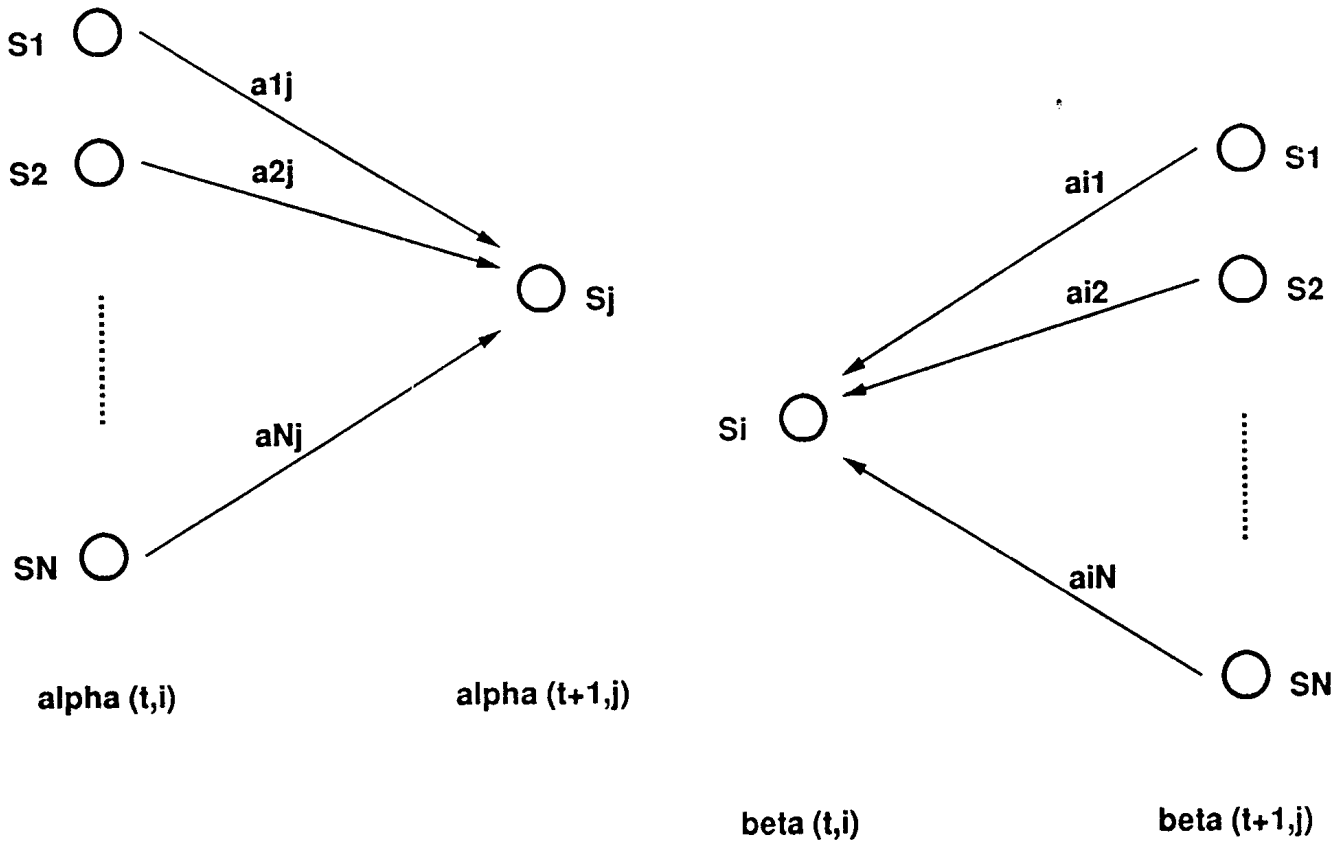
$$i = 1,2,\cdots,N \qquad t = T-1,\cdots,2,1$$

Termination :

$$P(\mathcal{O}|\lambda) = \sum_{i=1}^{N} \beta(1,i)$$

Backward Procedure - Illustration





Significance of alpha/beta terms

$$\alpha(t,i)\beta(t,i) = P(o_1, \cdots, o_t, \cdots, o_T,\ s_i \text{@time } t|\lambda)$$

Solution for problem (1) :

$$P(\mathcal{O}|\lambda) = \sum_{i=1}^{N} \alpha(t,i)\beta(t,i) \qquad \text{each time } t$$

References

- *Rabiner and Juang, Fundamentals of Speech Recognition, Prentice Hall*
- *Rabiner, A tutorial on HMM and selected applications in Speech Recognition*
- *J Glass, Speech Recognition, Spring 2003, Open Course Ware, MIT*
- *Speech Recognition, Course AM 0282*
- *Andrew Moore, Tutorial on HMM @ http://www.autonlab.org/tutorials/hmm.html*
- *C Bechetti, Speech Recognition: Theory and C++ Implementation*
- *Various Other sources*