

## 7장. 순차 데이터의 인식



## 들어가는 말

- 특징들의 시간성
  - 예) 지진파, 음성, 주식 거래량, 온라인 필기 문자 등
  - 이들을 순차<sup>sequential</sup> 데이터 또는 문맥 의존<sup>context-dependent</sup> 데이터라 부름
- 순차 데이터의 인식
  - 시간성의 표현과 정보 추론 방법 필요
  - 은닉 마코프 모델은<sup>HMM</sup> 가장 널리 사용되는 방법
- **HMM**
  - 19세기 마코프 모델에 토대를 둬 (7.2 절)
  - 1960년대 Baum 등이 은닉 추가하여 HMM으로 확장함 (7.3~7.5절)
  - HMM은 많은 분야에서 문제 해결 도구로 활용됨
    - 패턴인식, 컴퓨터비전, 데이터마이닝, 정보검색, 생물 정보학, 신호처리, 데이터베이스 .....
    - 응용 사례) 음성 인식, 온라인 필기 인식, DNA 열 찾기, 제스처 인식, 영어 발음 교정, 음악 인식 .....

## 7.1 순차 데이터

- 시간성이 없는 데이터
  - 특징들의 선후 관계는 무의미
- 시간성 있는 데이터 (순차 데이터)
  - 특징들의 선후 관계는 매우 중요함

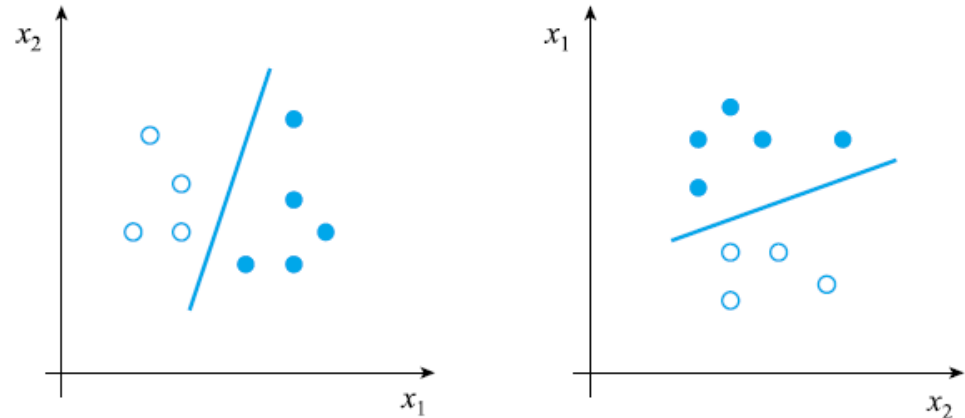


그림 7.1 시간성이 없는 데이터에서 특징의 위치를 바꿈

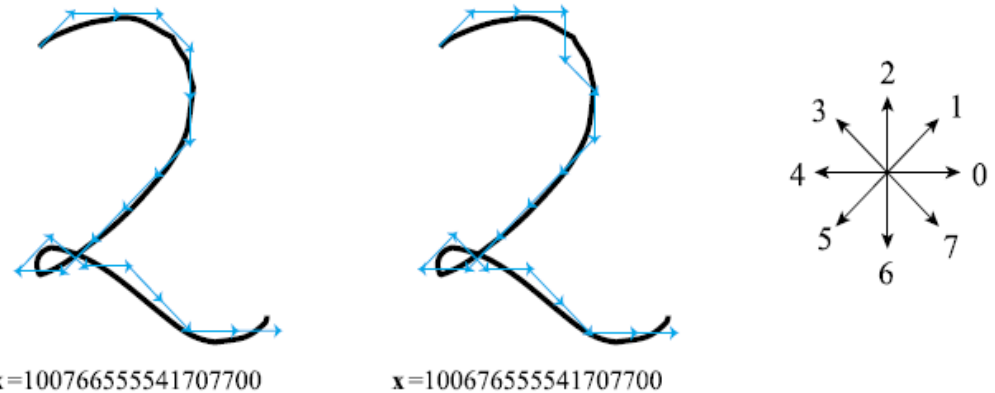


그림 7.2 시간성을 갖는 패턴에서 특징 순서를 바꾸었을 때 나타나는 물리적 왜곡 현상

## 7.1 순차 데이터

- 순차 데이터
  - 가변 길이
  - 관측 벡터로 표현

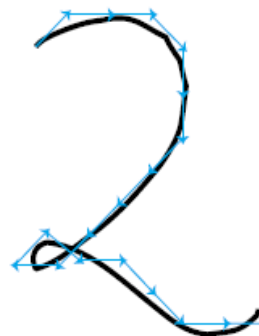
$$\mathbf{O} = (o_1, o_2, \dots, o_{t-1}, o_t, o_{t+1}, \dots, o_T)^T = o_1 o_2 \dots o_{t-1} o_t o_{t+1} \dots o_T \quad (7.1)$$

- 관측  $o_i$ 가 가질 수 있는 값의 집합을 알파벳이라 함
  - 알파벳  $V = \{v_1, v_2, \dots, v_m\}$
  - $v_i$ 를 기호라 함

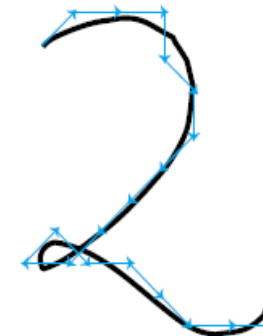
- 기호들이 시간에 따라 의존성 가짐

- 예) 그림 7.2의 온라인 숫자

- $P(o_t=2|o_{t-1}=6) \approx 0$
- $P(o_t=5|o_{t-1}=6) > P(o_t=4|o_{t-1}=6)$



x=10076655541707700



x=10067655541707700

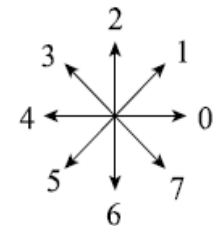


그림 7.2 시간성을 갖는 패턴에서 특징 순서를 바꾸었을 때 나타나는 물리적 왜곡 현상

## 7.2 마코프 모델

- 우선 은닉 마코프 모델에서 은닉이 빠진 **마코프 모델**을 공부하자.
  - 러시아 수학자 **Andrey Markov**가 제안
  - 시간  $t$ 에서의 관측은 가장 최근  $r$ 개 관측에만 의존한다는 가정 하의 확률 추론

$$\left. \begin{array}{l} r=0\text{이면, 0차 마코프 체인 } P(o_t | o_{t-1}o_{t-2} \cdots o_1) = P(o_t) \\ r=1\text{이면, 1차 마코프 체인 } P(o_t | o_{t-1}o_{t-2} \cdots o_1) = P(o_t | o_{t-1}) \\ r=2\text{이면, 2차 마코프 체인 } P(o_t | o_{t-1}o_{t-2} \cdots o_1) = P(o_t | o_{t-1}, o_{t-2}) \end{array} \right\} \quad (7.2)$$



(a) 0차 마코프체인



(b) 1차 마코프 체인



(c) 2차 마코프 체인

그림 7.3 마코프 체인

## 7.2 마코프 모델

### ■ 마코프 체인의 합리성은?

- 예) 최근 사흘의 날씨 정보를 가지고 오늘 날씨 추론
  - 삼일 전: 해 ( $o_{t-3}=\text{해}$ ), 이틀 전: 해 ( $o_{t-2}=\text{해}$ ), 어제: 비 ( $o_{t-1}=\text{비}$ )
  - 오늘 비올 확률은?
  - 1차 마코프 체인을 사용한다면

$$P(o_t = \text{비} \mid o_{t-1} = \text{비}, o_{t-2} = \text{비}, o_{t-3} = \text{해}) = P(o_t = \text{비} \mid o_{t-1} = \text{비})$$

- 즉 1차 마코프 체인에서는 아래 네 개 확률이 같다.
  - 정말 같을까?
  - 다르다면 오차가 무시할 정도일까?

$$P(o_t = \text{비} \mid o_{t-1} = \text{비}, o_{t-2} = \text{비}, o_{t-3} = \text{비})$$

$$P(o_t = \text{비} \mid o_{t-1} = \text{비}, o_{t-2} = \text{비}, o_{t-3} = \text{해})$$

$$P(o_t = \text{비} \mid o_{t-1} = \text{비}, o_{t-2} = \text{해}, o_{t-3} = \text{비})$$

$$P(o_t = \text{비} \mid o_{t-1} = \text{비}, o_{t-2} = \text{해}, o_{t-3} = \text{해})$$

## 7.2 마코프 모델

### ■ 마코프 모델

- 1차 마코프 체인 사용
- 2차 이상에서는 추정할 매개 변수가 많아 현실적인 문제 발생
- 알파벳을 구성하는 기호 각각을 상태로 간주

### ■ 날씨 예

- $V=\{\text{비, 구름, 해}\}$
- 기후 관측에 의해 얻은 날씨 변화 확률

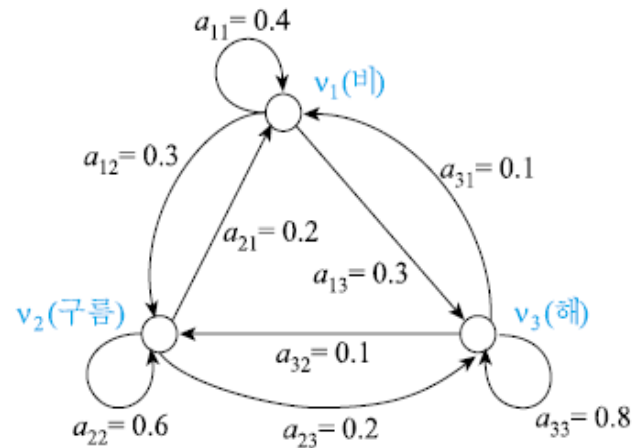
표 7.1 날씨 변화 확률

오늘 \ 내일	비	구름	해
비	0.4	0.3	0.3
구름	0.2	0.6	0.2
해	0.1	0.1	0.8

## 7.2 마코프 모델

- 상태 전이
  - 상태 전이 확률 행렬과 상태 전이도

$$\mathbf{A} = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = \begin{vmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{vmatrix}$$



(a) 상태 전이 확률 행렬

(b) 상태 전이도

그림 7.4 마코프 모델을 위한 상태 전이 확률과 상태 전이도

$$\left. \begin{array}{l} a_{ij} = P(o_t = v_j | o_{t-1} = v_i) \\ \text{여기서 } a_{ij} \geq 0 \text{ 와 } \sum_{j=1}^m a_{ij} = 1 \text{ 을 만족} \end{array} \right\}$$



## 7.2 마코프 모델

### ■ 마코프 모델로 무엇을 할 수 있나?

#### □ 관측벡터 $\mathbf{O}$ 의 확률 구하기

$$\left. \begin{aligned} P(\mathbf{O} | \text{마코프모델}) &= P(\mathbf{O} | \mathbf{A}) \\ &= P(\mathbf{O}) \\ &= P(o_1, o_2, \dots, o_T) \\ &= P(o_1)P(o_2 | o_1)P(o_3 | o_2, o_1) \cdots P(o_t | o_{t-1}, o_{t-2}, \dots, o_1) \cdots P(o_T | o_{T-1}, \dots, o_1) \\ &= P(o_1)P(o_2 | o_1)P(o_3 | o_2) \cdots P(o_t | o_{t-1}) \cdots P(o_T | o_{T-1}) \quad \text{1차 Markov model} \\ &= P(o_1) \prod_{i=1}^{T-1} P(o_{i+1} | o_i) \end{aligned} \right\} (7.4)$$

#### □ 예제 7.1 “오늘 해가 떴는데 내일부터 7일 간의 날씨가 해-해-비-비-해-구름-해일 확률은 얼마인가?”

$$\begin{aligned} P(\mathbf{O} | \mathbf{A}) &= P(o_1 = \text{해}, o_2 = \text{해}, o_3 = \text{해}, o_4 = \text{비}, o_5 = \text{비}, o_6 = \text{해}, o_7 = \text{구름}, o_8 = \text{해} | \mathbf{A}) \\ &= P(\text{해}, \text{해}, \text{해}, \text{비}, \text{비}, \text{해}, \text{구름}, \text{해}) \\ &= P(\text{해}) P(\text{해} | \text{해}) P(\text{해} | \text{해}) P(\text{비} | \text{해}) P(\text{비} | \text{비}) P(\text{해} | \text{비}) P(\text{구름} | \text{해}) P(\text{해} | \text{구름}) \\ &= \pi_3 a_{33} a_{33} a_{31} a_{11} a_{13} a_{32} a_{23} \\ &= 1 * 0.8 * 0.8 * 0.1 * 0.4 * 0.3 * 0.1 * 0.2 \\ &= 1.536 * 10^{-4} \end{aligned}$$

## 7.2 마코프 모델

- 또 다른 예, 온라인 필기 숫자 인식

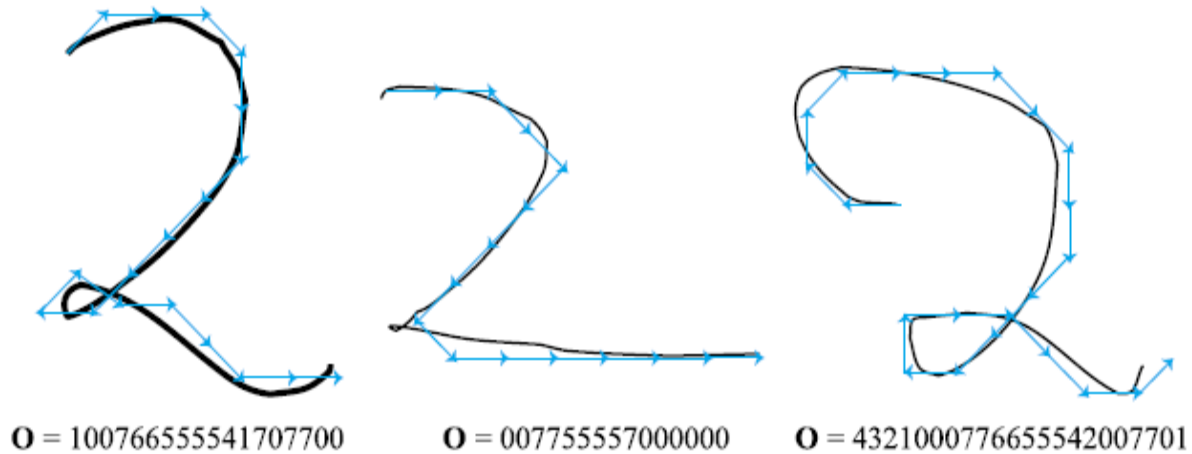


그림 7.5 온라인 필기 숫자 인식에서 부류 2의 훈련 샘플들

## 7.2 마코프 모델

- 또 다른 예, 온라인 필기 숫자 인식
  - 상태 전이 확률 행렬을 구하면,

표 7.2 상태 전이의 발생 회수

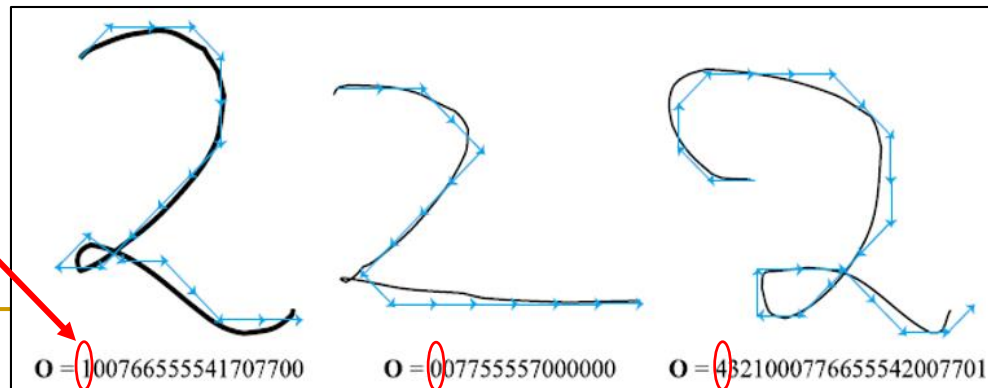
$t-1 \backslash t$	0	1	2	3	4	5	6	7
0	11	1	0	0	0	0	0	5
1	2	0	0	0	0	0	0	1
2	1	1	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0
4	0	1	1	1	0	0	0	0
5	0	0	0	0	2	8	0	1
6	0	0	0	0	0	2	2	0
7	4	0	0	0	0	1	2	4

$$A = \begin{pmatrix} 11/17 & 1/17 & 0 & 0 & 0 & 0 & 0 & 0 & 5/17 \\ 2/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/3 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2/11 & 8/11 & 0 & 0 & 1/11 \\ 0 & 0 & 0 & 0 & 0 & 1/2 & 1/2 & 0 & 0 \\ 4/11 & 0 & 0 & 0 & 0 & 1/11 & 2/11 & 2/11 & 4/11 \end{pmatrix}$$

- 초기 확률 행렬을 구하면,

$$\pi = (\pi_1, \pi_2, \pi_3, \pi_4, \pi_5, \pi_6, \pi_7, \pi_8) = \left(\frac{1}{3}, \frac{1}{3}, 0, 0, \frac{1}{3}, 0, 0, 0\right)$$

오른쪽 세개 샘플의  $t=1$ 일때의 값은 1,0,4임



## 7.2 마코프 모델

### ■ 또 다른 예, 온라인 필기 숫자 인식

- 이렇게 만든 마코프 모델로 그림 7.6의 샘플이 발생할 확률을 구하면,

“체인 코드  $1 \rightarrow 0 \rightarrow 7 \rightarrow 6 \rightarrow 5 \rightarrow 5 \rightarrow 7 \rightarrow 0 \rightarrow 0$ 으로 표현되는 샘플이 발생할 확률은 얼마인가?”

$$\begin{aligned}
 P(\mathbf{O}|\mathbf{A}) &= P(o_1=1, o_2=0, o_3=7, o_4=6, o_5=5, o_6=5, o_7=7, o_8=0, o_9=0|\mathbf{A}) \\
 &= P(1) P(0|1) P(7|0) P(6|7) P(5|6) P(5|5) P(7|5) P(0|7) P(0|0) \\
 &= \pi_2 a_{21} a_{18} a_{87} a_{76} a_{66} a_{68} a_{81} a_{11} \\
 &= (1/3) * (2/3) * (5/17) * (2/11) * (1/2) * (8/11) * (1/11) * (4/11) * (11/17) \\
 &= 0.9243 * 10^{-4}
 \end{aligned}$$

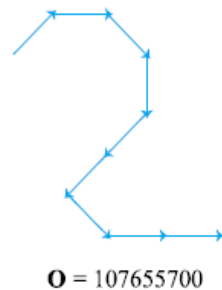


그림 7.6 이 샘플이 발생할 확률은 0.00009243

$$\mathbf{A} = \begin{matrix} & \begin{matrix} 1 & 0 & 7 & 6 & 5 & 5 & 7 & 0 & 0 \end{matrix} \\ \begin{matrix} 1 \\ 0 \\ 7 \\ 6 \\ 5 \\ 5 \\ 7 \\ 0 \\ 0 \end{matrix} & \begin{bmatrix} 11/17 & 1/17 & 0 & 0 & 0 & 0 & 0 & 0 & 5/17 \\ 2/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/3 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2/11 & 8/11 & 0 & 1/11 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/2 & 1/2 & 0 & 0 \\ 4/11 & 0 & 0 & 0 & 0 & 1/11 & 2/11 & 4/11 & 0 \end{bmatrix} \end{matrix}$$

$a_{18}$

$a_{87}$

## 7.2 마코프 모델

- 마코프 모델과 0차, 2차 마코프 체인과의 비교 (온라인 필기 숫자 예)
  - 0차 마코프 체인 : 이전 정보는 전혀 고려하지 않음

표 7.3 0차 마코프 체인을 위한 확률 표

	0	1	2	3	4	5	6	7
$t$	19 (19/55)	4 (4/55)	2 (2/55)	1 (1/55)	3 (3/55)	11 (11/55)	4 (4/55)	11 (11/55)

$$\begin{aligned} P(\mathbf{O}|\mathbf{A}_0) &= P(o_1 = 1, o_2 = 0, o_3 = 7, o_4 = 6, o_5 = 5, o_6 = 5, o_7 = 7, o_8 = 0, o_9 = 0|\mathbf{A}_1) \\ &= P(1)P(0)P(7)P(6)P(5)P(5)P(7)P(0)P(0) \\ &= (4/55)*(19/55)*(11/55)*(4/55)*(11/55)*(11/55)*(11/55)*(19/55) \\ &\quad *(19/55) \\ &= 0.3489*10^{-6} \end{aligned}$$

## 7.2 마코프 모델

- 마코프 모델과 0차, 2차 마코프 체인과의 비교 (온라인 필기 숫자 예)
  - 2차 마코프 체인 : 많은 변수값을 가짐 → 계산 복잡도 매우 높아짐
    - 추정할 매개 변수가  $64*8$ 개로 증가

표 7.4 2차 마코프 체인을 위한 전이 확률 표

$(t-2, t-1) \backslash t$	0	1	2	3	4	5	6	7
00	5 (5/9)	0 (0/9)	0 (0/9)	0 (0/9)	0 (0/9)	0 (0/9)	0 (0/9)	4 (4/9)
01	...							
...								
$ij$								
...								
77								...

- 꼭 이해하고 기억해야 할 것! (MM과 HMM의 근본적인 차이점)
  - 마코프 모델에서는 상태를 나타내는 노드에 관측 (비, 해, 구름) 그 자체가 들어 있다. 즉 상태를 볼 수 있다.
  - HMM에서는 상태를 볼 수 없다. 즉 상태가 은닉된다.

## 7.3 은닉 마코프 모델로의 발전

- 은닉 마코프 모델로의 발전
  - 마코프 모델은 한계를 가진다.
    - 보다 복잡한 현상이나 과정에 대한 모델링 능력의 한계
  - 모델의 용량을 키우기 위해,
    - 상태를 감춘다.

## 7.3.1 동기

### ■ 마코프 체인의 차수와 추정할 매개 변수의 수

- $r$ 차에서  $m^r(m-1)$  개의 매개 변수
- 차수에 따라 기하급수적으로 모델 크기가 증가함

- 0차 Markov model에서는  $m-1$ 개의 확률값을 추정해야함( $m$ 은 상태수). 나머지 하나는 확률 전체합이 1이라는 정보를 갖고 있으므로
- 1차에선  $m^* (m-1)$
- 2차에선  $m^*m^*(m-1)$

### ■ HMM

- 차수를 미리 고정하지 않고 모델 자체가 확률 프로세스에 따라 적응적으로 정함
- 따라서 아무리 먼 과거의 관측도 현재에 영향을 미침
- 즉  $P(\text{해}|\text{비},\text{비}), P(\text{해}|\text{비},\text{비},\text{비}), P(\text{해}|\text{비},\text{비},\dots,\text{비}), P(\text{해}|\text{비},\text{비},\text{해},\dots,\text{해})$ 가 모두 다름
- 이런 뛰어난 능력에도 불구하고 모델 크기는 감당할 정도임

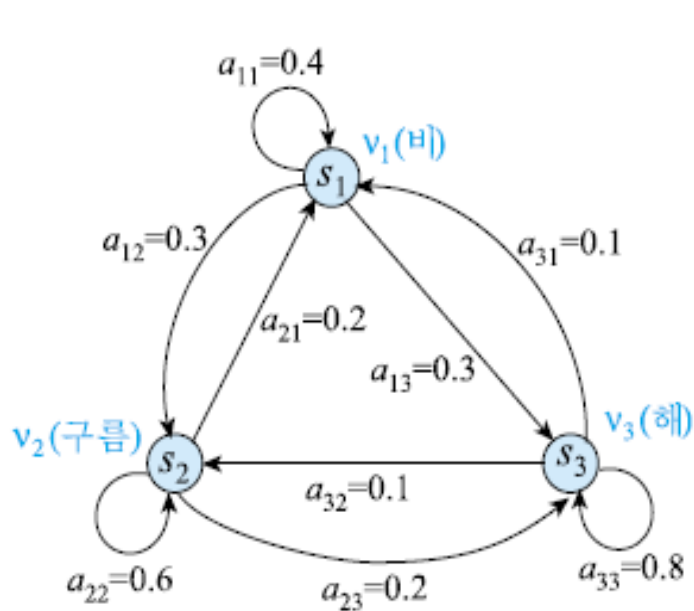
### ■ 어떤 비결에 의해 이런 뛰어난 능력이 가능한가?

- 상태를 감춘다.

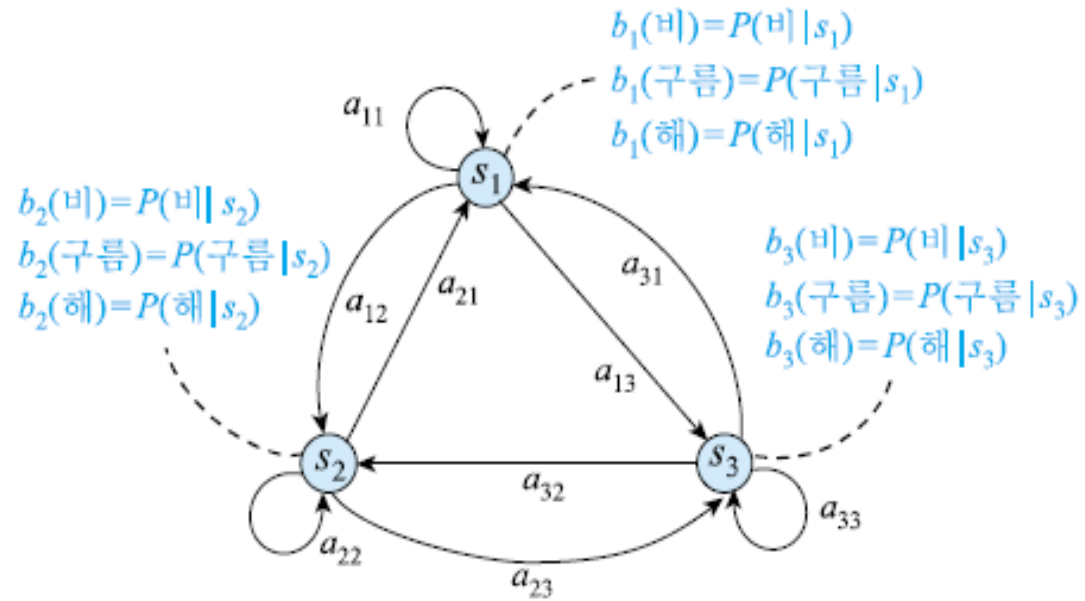


## 7.3.1 동기

- 상태를 감추면,
  - 마코프 모델이 은닉 마코프 모델이 된다.



(a) 마코프 모델



(b) 은닉 마코프 모델

그림 7.7 마코프 모델을 은닉 마코프 모델로 확장

## 7.3.1 동기

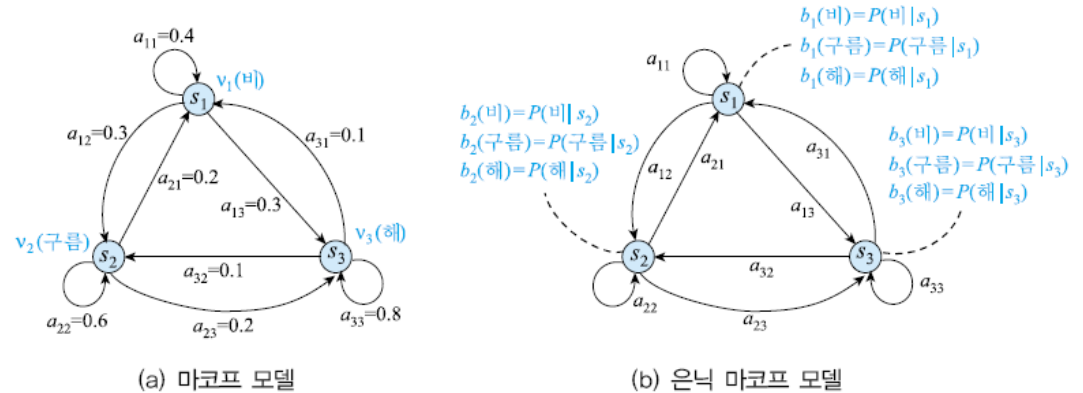


그림 7.7 마코프 모델을 은닉 마코프 모델로 확장

- 그림 7.7의 해석 (예를 들어, 해→해→비→구름이 관측되었다.)
  - 마코프 모델
    - 그것은 상태  $s_3 \rightarrow s_3 \rightarrow s_1 \rightarrow s_2$ 에서 관측한 것이다.
  - 은닉 마코프 모델
    - 모든 상태에서 {비,해,구름}이 관측 가능하므로  $s_3 \rightarrow s_3 \rightarrow s_1 \rightarrow s_2$ 일 수도 있고  $s_1 \rightarrow s_2 \rightarrow s_3 \rightarrow s_1$ 일 수도 있다. 가능한 경우는 몇 가지일까?
    - 그들의 확률은 다르다. 예를 들어  $s_3 \rightarrow s_3 \rightarrow s_1 \rightarrow s_2$ 일 확률은

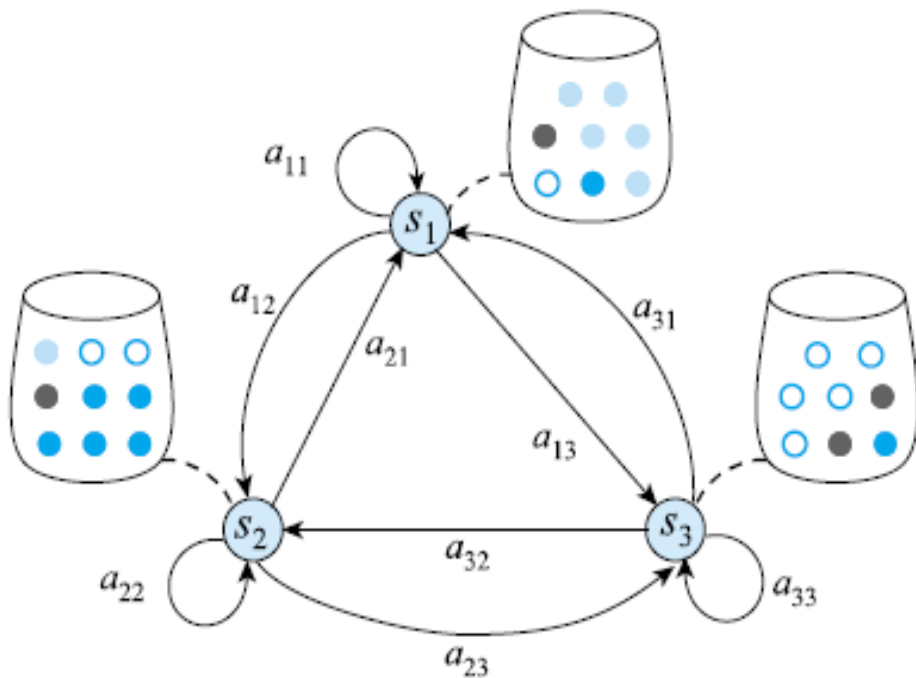
해 → 해 → 비 → 구름이  $s_3 \rightarrow s_3 \rightarrow s_1 \rightarrow s_2$ 에서 관측되었을 확률 =

$$[\pi_3 * b_3(\text{해})] * [a_{33} * b_3(\text{해})] * [a_{31} * b_1(\text{비})] * [a_{12} * b_2(\text{구름})]$$

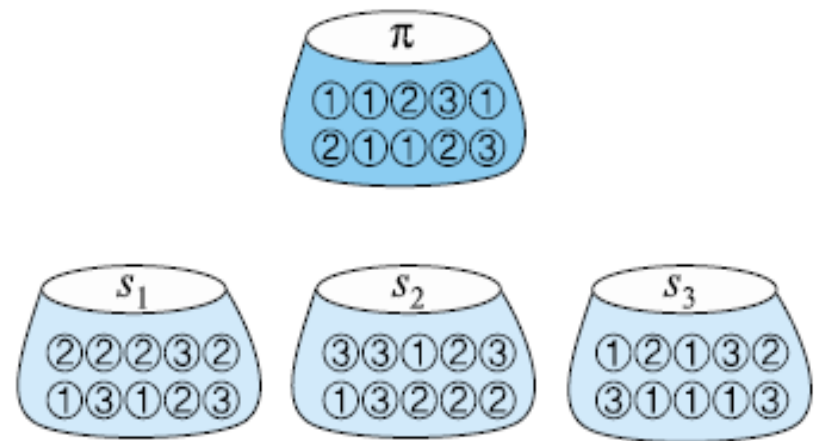
## 7.3.2 HMM의 예

### ■ 예제 7.3 공을 담은 항아리 [Rabiner89]

- $n$  개의 항아리, 공의 색깔  $m$  개
  - 항아리가 상태이고 공의 색깔을 관찰
- 그림 7.8은  $n=3, m=4$ 
  - 출력 심볼, 관측 가능한 공의 색깔:  $V=\{\text{하양(W)}, \text{검정(B)}, \text{연파랑(L)}, \text{진파랑(D)}\}$



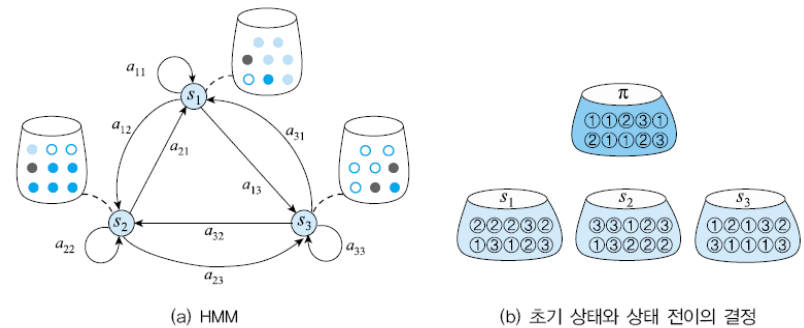
(a) HMM



(b) 초기 상태와 상태 전이의 결정

그림 7.8 공을 담은 항아리를 HMM으로 모델링

## 7.3.2 HMM의 예



(a) HMM

(b) 초기 상태와 상태 전이의 결정

그림 7.8 공을 담은 항아리를 HMM으로 모델링

### ■ 실험 시나리오

이 예에서의 실험을 정리해 보자. 실험이 시작되면  $\pi$ -항아리에서 카드를 하나 뽑아 번호를 보고 다시 집어 넣는다. 카드 번호의 상태로 들어간다. 그 상태에 해당하는 공 항아리에서 공을 하나 뽑고 색깔을 확인하고 다시 집어 넣는다. 색깔을  $o_1$ 에 기록한다. 그 상태에 해당하는 카드 항아리에서 카드를 하나 뽑고 번호를 보고 다시 집어 넣는다. 카드 번호의 상태로 이동한다. 그 상태에 해당하는 공 항아리에서 공을 하나 뽑고 색깔을  $o_2$ 에 기록한다. 이런 과정을 반복한다.

이 실험이 우리 눈에 다 보이는 것이 아니라는 사실을 기억해야 한다. 보이는 것은 공의 색깔, 즉  $\mathbf{O} = o_1 o_2 \dots o_T$  뿐이다. 실험이 커튼 뒤에서 이루어지고 실험하는 사람이 단지 공의 색깔만 불러 준다고 생각하면 된다. 우리가 가진 것은  $\mathbf{O}$  뿐이다. ■■■

- $\mathbf{O} = (\text{빨강}, \text{하양}, \text{하양}, \text{파랑})^T$ 이 관찰되었다면,
  - 발생 확률은? 어느 상태에서 관측되었을까?

## 7.3.2 HMM의 예

- 예제 7.4 여자 친구의 삶 [Wikipedia]
  - 여자 친구의 일상을 관찰,  $V=\{\text{산책, 쇼핑, 청소}\}$
  - 날씨는 해와 비 (상태)
  - 내가 가진 정보

### 날씨

$$P(\text{비 온 다음 날 비}) = P(q_t = \text{비} | q_{t-1} = \text{비}) = 0.7$$

$$P(\text{비 온 다음 날 해}) = P(q_t = \text{해} | q_{t-1} = \text{비}) = 0.3$$

$$P(\text{해 뜬 다음 날 비}) = P(q_t = \text{비} | q_{t-1} = \text{해}) = 0.4$$

$$P(\text{해 뜬 다음 날 해}) = P(q_t = \text{해} | q_{t-1} = \text{해}) = 0.6$$

$$P(\text{비}) = 0.6$$

초기상태 확률

$$P(\text{해}) = 0.4$$

- 답할 수 있나?

- 그녀가 일주일 연속으로 쇼핑만 할 확률은?
- 그꼬저께 산책, 그저께 산책, 어제 청소했다는데 3일간 그곳 날씨는?
- 그꼬저께 산책, 그저께 산책, 어제 청소했다는데 오늘과 내일 무얼할까?

### 날씨에 따른 그녀의 행동

$$P(\text{비 오는 날 산책}) = P(o_t = \text{산책} | q_t = \text{비}) = 0.1$$

$$P(\text{비 오는 날 쇼핑}) = P(o_t = \text{쇼핑} | q_t = \text{비}) = 0.4$$

$$P(\text{비 오는 날 청소}) = P(o_t = \text{청소} | q_t = \text{비}) = 0.5$$

$$P(\text{해 뜬 날 산책}) = P(o_t = \text{산책} | q_t = \text{해}) = 0.6$$

$$P(\text{해 뜬 날 쇼핑}) = P(o_t = \text{쇼핑} | q_t = \text{해}) = 0.3$$

$$P(\text{해 뜬 날 청소}) = P(o_t = \text{청소} | q_t = \text{해}) = 0.1$$

## 7.3.2 HMM의 예

- HMM을 사용하면 확률 추론할 수 있다. 어떻게?

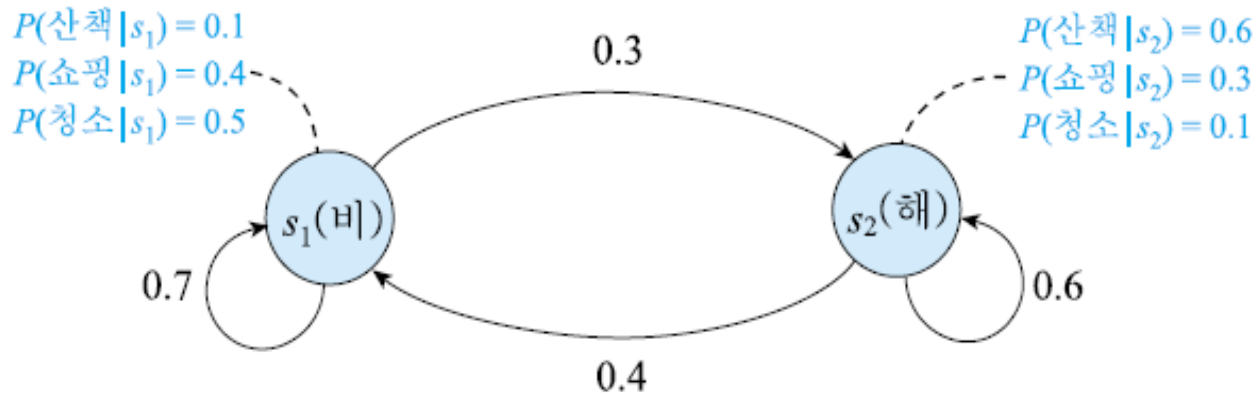


그림 7.9 그녀의 삶을 HMM으로 모델링

### 7.3.3 구성 요소와 세 가지 문제

- 앞의 예제는 한쪽 면만 조망
  - 모델을 알 때 확률 추론하라.
    - 그녀의 삶 예에서는 날씨와 그녀의 행위에 대한 확률 분포 알고 있음
  - 또 다른 측면
    - 어느 것이 상태인가? 상태는 몇 가지인가? (아키텍처 설계)
    - 확률 분포는 어떻게 구하나? (학습)
    - 예) 온라인 필기 인식
      - 무엇이 상태이며, 상태를 몇 가지로 할까?
      - 가진 것은 오로지 훈련 집합

### 7.3.3 구성 요소와 세 가지 문제

#### ■ 아키텍처

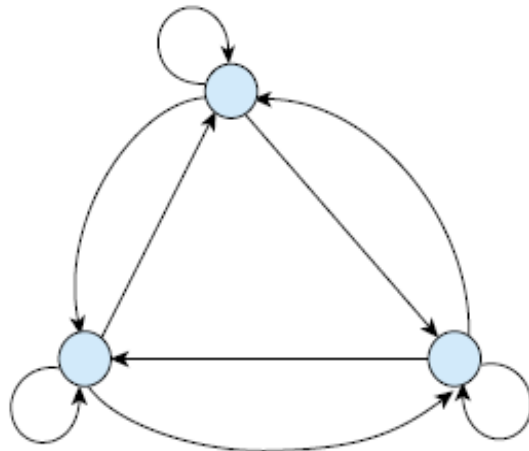
##### □ HMM은 가중치 방향 그래프로 표현

- 노드가 상태
- 상태로 사용할 것이 명확한 경우도 있지만 그렇지 않은 경우도 있다.

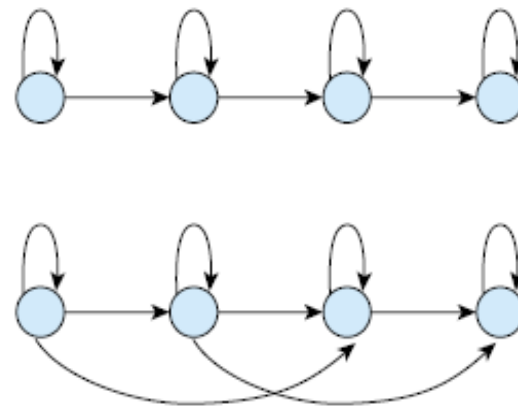
##### □ 대표적인 아키텍처

- 어고딕 모델과 좌우 모델
- 응용의 특성에 따라 적절한 아키텍처 선택 중요

- 예를 들어 음성 인식은 좌우 모델이 적당함. 왜? → 음성은 어떤 부분 패턴이 지나간 후에 그것이 다시 나타나지 않으므로 “좌우모델처럼 어떤 상태로 들어가면 뒤로 돌아갈 수 없는게 적당함”



(a) 어고딕 모델



(b) 좌우 모델



### 7.3.3 구성 요소와 세 가지 문제

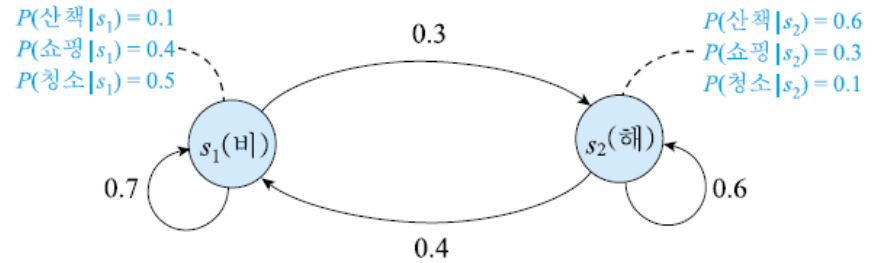


그림 7.9 그녀의 삶을 HMM으로 모델링

#### ■ 세 가지 매개변수 $\Theta = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$

##### 1. 상태 전이 확률 행렬 $\mathbf{A} = |a_{ij}|$

$$a_{ij} = P(q_{t+1} = s_j | q_t = s_i), 1 \leq i, j \leq n$$

$\circ$  이 때  $\sum_{j=1}^n a_{ij} = 1$

##### 2. 관측 행렬 $\mathbf{B} = |b_j(v_k)|$

$$b_j(v_k) = P(o_t = v_k | q_t = s_j), 1 \leq j \leq n, 1 \leq k \leq m$$

$\circ$  이 때  $\sum_{k=1}^m b_j(v_k) = 1$

##### 3. 초기 확률 벡터 $\boldsymbol{\pi} = |\pi_i|$

$$\pi_i = P(q_1 = s_i), 1 \leq i \leq n$$

$\circ$  이 때  $\sum_{i=1}^n \pi_i = 1$

### 7.3.3 구성 요소와 세 가지 문제

#### ■ 세 가지 문제

1. **평가 문제:** 모델  $\Theta$ 가 주어진 상황에서, 관측벡터  $O$ 에 대한  $P(O|\Theta)$ 는?
2. **디코딩 문제:** 모델  $\Theta$ 가 주어진 상황에서, 관측벡터  $O$ 를 알 경우, 최적의 상태열은?
3. **학습 문제:** 훈련집합  $X=\{O_1, \dots, O_N\}$ 이 주어져 있을 때 HMM의 매개 변수  $\Theta$ 는?

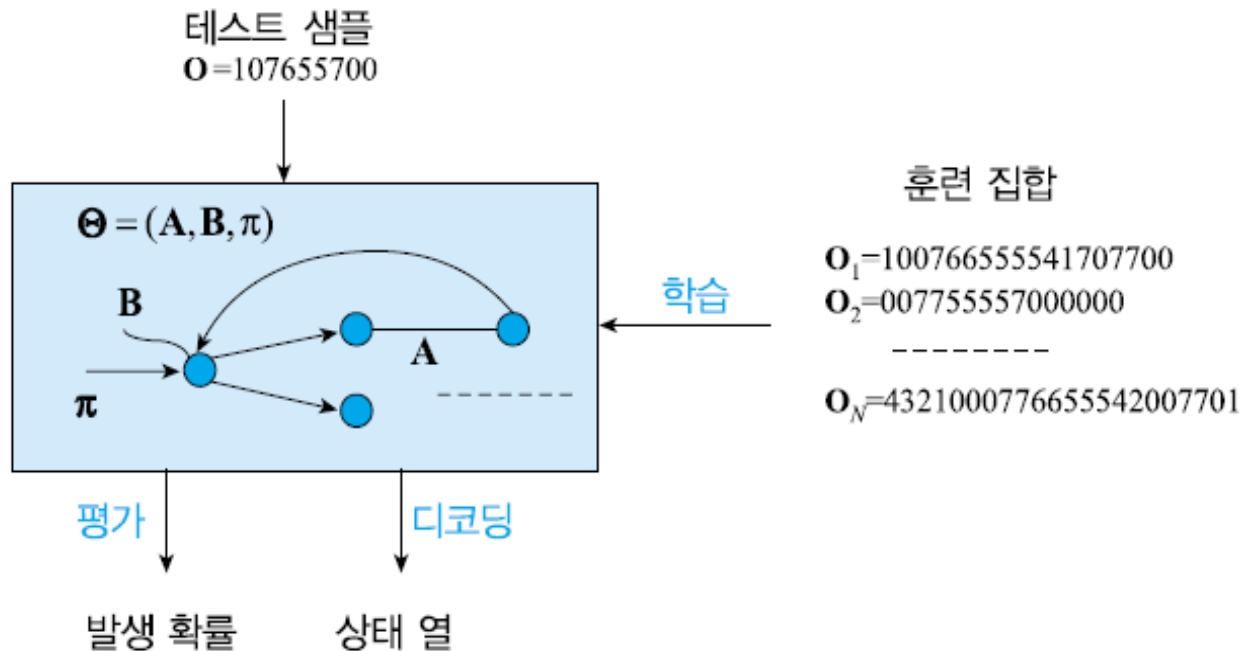


그림 7.11 HMM의 세 가지 문제

## 7.4 알고리즘

- 이제 세 가지 문제를 풀기 위한 알고리즘을 공부하자.
  - **평가 문제:** 동적 프로그래밍 이용
  - **디코딩 문제:** 동적 프로그래밍 이용 (Viterbi 알고리즘)
  - **학습 문제:** EM 알고리즘 (Baum-Welch 알고리즘)

## 7.4.1 평가

### ■ 평가 문제란?

- 모델  $\Theta$ 가 주어진 상황에서, 관측벡터  $\mathbf{O}$ 에 대한  $P(\mathbf{O} | \Theta)$ 는?
- 예) 그녀가 일주일 연속으로 쇼핑만 할 확률은?

### ■ 평가 문제를 풀어 보자.

- HMM에서는  $\mathbf{O}$ 를 관측한 상태 열을 모른다.
- 우선  $\mathbf{O}$ 를 관측한 상태 열을 안다고 가정하고 그것을  $\mathbf{Q}=(q_1, \dots, q_T)$ 라 하자.
- 그럼 아래 식을 유도할 수 있다.

$$P(\mathbf{O}, \mathbf{Q} | \Theta) = \pi_{q_1} b_{q_1}(o_1) a_{q_1 q_2} b_{q_2}(o_2) \cdots a_{q_{T-1} q_T} b_{q_T}(o_T) \quad (7.10)$$

- 문제  $P(\mathbf{O} | \Theta)$ 는 모든 상태 열에 대해 (7.10)의 식을 더하여 구할 수 있다.

$$P(\mathbf{O} | \Theta) = \sum_{\text{모든 } \mathbf{Q}} P(\mathbf{O}, \mathbf{Q} | \Theta)$$

## 7.4.1 평가

### ■ 예제 7.5 여자 친구의 삶

- “그녀가 오늘 산책, 내일 산책, 모레 청소, 그리고 글피 쇼핑할 확률은?”
- 즉  $O=(o_1=\text{산책}, o_2=\text{산책}, o_3=\text{청소}, o_4=\text{쇼핑})$ 일 때,  $P(O|\Theta)$ 는?

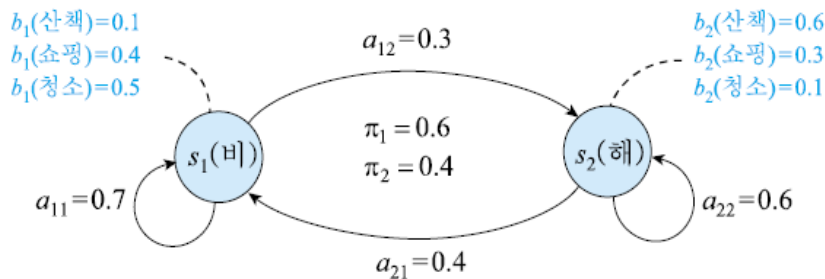


그림 7.12 그녀의 삶 시나리오의 HMM

- 모든 상태 열을 나열하면,

$Q_1 = \text{비비비비}, Q_2 = \text{비비비해}, Q_3 = \text{비비해비}, Q_4 = \text{비비해해},$   
 $Q_5 = \text{비해비비}, Q_6 = \text{비해비해}, Q_7 = \text{비해해비}, Q_8 = \text{비해해해},$   
 $Q_9 = \text{해비비비}, Q_{10} = \text{해비비해}, Q_{11} = \text{해비해비}, Q_{12} = \text{해비해해},$   
 $Q_{13} = \text{해해비비}, Q_{14} = \text{해해비해}, Q_{15} = \text{해해해비}, Q_{16} = \text{해해해해}$

## 7.4.1 평가

### ■ 예제 7.5 여자 친구의 삶

- 상태 열  $Q_1$ 에 대해 구해 보면,

$$\begin{aligned}P(\mathbf{O}, Q_1 | \Theta) &= \pi_1 b_1(\text{산책}) a_{11} b_1(\text{산책}) a_{11} b_1(\text{청소}) a_{11} b_1(\text{쇼핑}) \\ &= 0.6 * 0.1 * 0.7 * 0.1 * 0.7 * 0.5 * 0.7 * 0.4 \\ &= 4.116 * 10^{-4}\end{aligned}$$

- 모든 상태 열에 대한 값을 더하여 답을 구해 보면,

$$\begin{aligned}P(\mathbf{O} | \Theta) &= \sum_{i=1}^{16} P(\mathbf{O}, Q_i | \Theta) \\ &= 0.4116 * 10^{-3} + 0.1323 * 10^{-3} + 0.02216 * 10^{-3} + 0.02268 * 10^{-3} \\ &\quad + 0.6048 * 10^{-3} + 0.1944 * 10^{-3} + 0.10368 * 10^{-3} + 0.11664 * 10^{-3} \\ &\quad + 0.9408 * 10^{-3} + 0.3024 * 10^{-3} + 0.04608 * 10^{-3} + 0.05184 * 10^{-3} \\ &\quad + 4.8384 * 10^{-3} + 1.5552 * 10^{-3} + 0.82944 * 10^{-3} + 0.93312 * 10^{-3} \\ &= 0.0111\end{aligned}$$

## 7.4.1 평가

- 답은 구할 수 있다. 하지만,
  - 예제 7.5에서 상태 열의 개수는  $2^4=16$  가지
  - 일반적으로 상태의 수가  $n$ 이고 관측 열  $O$ 의 길이가  $T$ 라면  $n^T$  가지의 상태 열
  - 각각의 상태 열은  $2T-1$ 번의 곱셈. 따라서 시간 복잡도는  $\Theta(n^T T)$ 
    - 예)  $n=5, T=30$ 이라면  $5.4948 \times 10^{22}$  번의 곱셈 필요. 계산 폭발!
- 보다 효율적인 알고리즘이 있다.
  - 기본 아이디어는 동적 프로그래밍에 의한 중복 계산 제거
  - 예) 예제 7.5의  $Q_1$ ='비비비비',  $Q_2$ ='비비비해'의 계산
    - $P(O, Q_1 | \Theta) = \pi_1 b_1(\text{산책}) a_{11} b_1(\text{산책}) a_{11} b_1(\text{청소}) a_{11} b_1(\text{쇼핑})$
    - $P(O, Q_2 | \Theta) = \pi_1 b_1(\text{산책}) a_{11} b_1(\text{산책}) a_{11} b_1(\text{청소}) a_{12} b_2(\text{쇼핑})$
    - 빨간 부분은 중복계산하는 부분 → **dynamic programming**으로 이를 중복 계산할 필요없음

## 7.4.1 평가

### ■ 동적 프로그래밍

□ 그림 7.13의 격자에 16개의 모든 상태 열이 들어 있다.

■ 예, 진한 파랑은  $Q_1$ ='비비비비', 연파랑은  $Q_7$ ='비해해비'에 해당

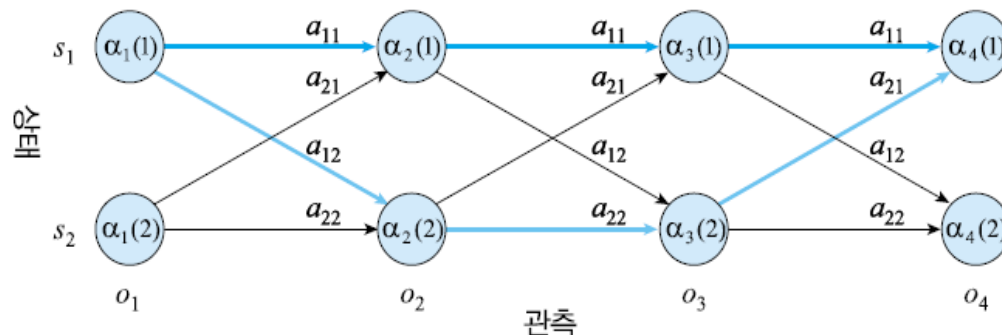


그림 7.13 전방 계산을 위해 상태를 격자 모양으로 펼침

□  $\alpha_t(i)$  는 관측 벡터의 일부  $o_1 o_2 \dots o_t$ 를 관측하고 시간  $t$ 에  $s_i$ 에 있을 확률

$$\left. \begin{aligned} \alpha_t(i) &= P(o_1, o_2, \dots, o_t, q_t = s_i \mid \Theta) \\ &= \left[ \sum_{j=1}^n \alpha_{t-1}(j) a_{ji} \right] * b_i(o_t) \end{aligned} \right\}$$

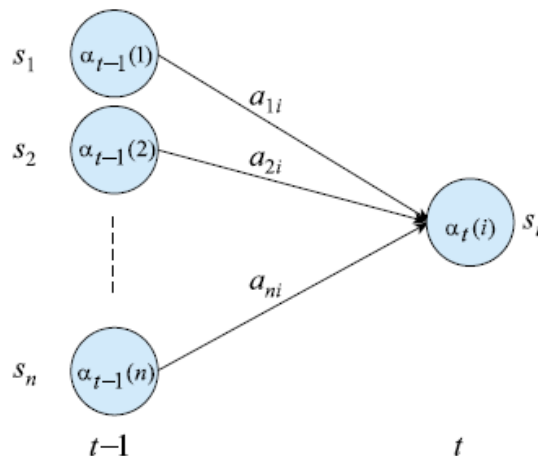


그림 7.14  $\alpha_t(i)$ 의 순환 계산



## 7.4.1 평가

### ■ 전방 알고리즘 forward algorithm

- 초기식을 포함시킨 완벽한 식으로 정리하고 그것을 가상 코드로 적으면,

알고리즘 [7.1] 전방 계산에 의한 관측 벡터의 발생 확률 계산

입력: HMM  $\Theta = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$ , 관측 벡터  $\mathbf{O} = o_1 o_2 \dots o_T$

출력:  $\mathbf{O}$ 의 발생 확률  $P(\mathbf{O} | \Theta)$

알고리즘:

1. 배열  $\alpha[1 \dots T][1 \dots n]$ 을 생성하라.
2. **for** ( $i = 1$  **to**  $n$ )  $\alpha[1][i] = \pi_i * b_i(o_1);$  // (7.13)
3. **for** ( $t = 2$  **to**  $T$ )
4.   **for** ( $i = 1$  **to**  $n$ ) {
5.        $sum = 0;$
6.       **for** ( $j = 1$  **to**  $n$ )  $sum = sum + \alpha[t-1][j] * a_{ji};$
7.        $\alpha[t][i] = sum * b_i(o_t);$        // 라인 5-7이 (7.14)
8.     }
9.    $sum = 0;$
10. **for** ( $j = 1$  **to**  $n$ )  $sum = sum + \alpha[T][j];$  // (7.15)
11.  $P(\mathbf{O} | \Theta) = sum;$

$$\alpha_1(i) = \pi_i b_i(o_1), \quad 1 \leq i \leq n$$

$$\alpha_t(i) = \left[ \sum_{j=1}^n \alpha_{t-1}(j) a_{ji} \right] * b_i(o_t), \quad 2 \leq t \leq T, 1 \leq i \leq n$$

$$P(\mathbf{O} | \Theta) = \sum_{j=1}^n \alpha_T(j)$$

시간 복잡도는  $\Theta(n^2 T)$

$n=5, T=30$ 이라면 750 번의 곱셈 이전의 낱말 계산과 비교해 보라.

## 7.4.1 평가

### 예제 7.6 전방 계산에 의한 확률 평가

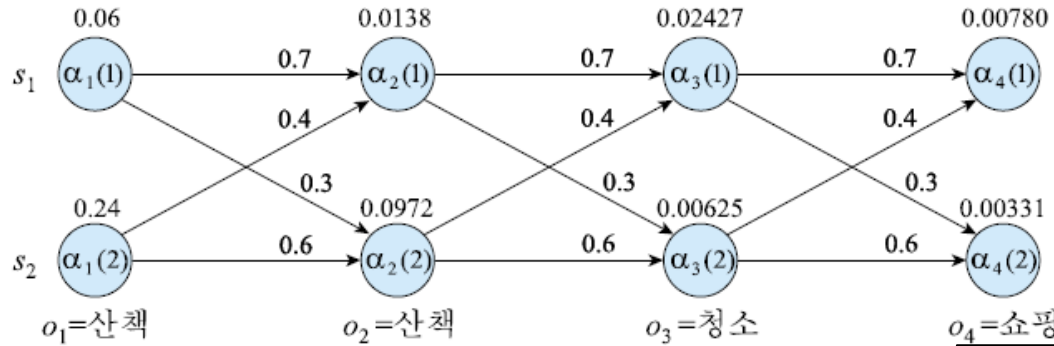


그림 7.15 전방 계산 예

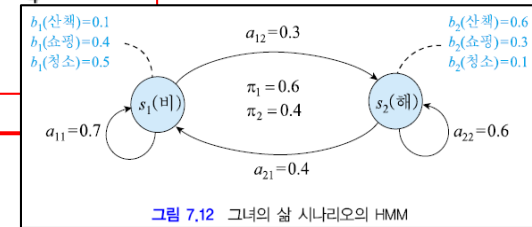


그림 7.12 그녀의 삶 시나리오의 HMM

$$\alpha_1(i) = \pi_i b_i(o_1), \quad 1 \leq i \leq n$$

$$\alpha_t(i) = \left[ \sum_{j=1}^n \alpha_{t-1}(j) a_{ji} \right] * b_i(o_t), \quad 2 \leq t \leq T, 1 \leq i \leq n$$

$$P(\mathbf{O} | \Theta) = \sum_{j=1}^n \alpha_T(j)$$

$t=1$ 일 때  $\alpha[1][1] = \pi_1 * b_1(\text{산책}) = 0.6 * 0.1 = 0.06$

$\alpha[1][2] = \pi_2 * b_2(\text{산책}) = 0.4 * 0.6 = 0.24$

$t=2$ 일 때  $\alpha[2][1] = (\alpha[1][1] * a_{11} + \alpha[1][2] * a_{21}) * b_1(\text{산책}) = (0.06 * 0.7 + 0.24 * 0.4) * 0.1 = 0.0138$

$\alpha[2][2] = (\alpha[1][1] * a_{12} + \alpha[1][2] * a_{22}) * b_2(\text{산책}) = (0.06 * 0.3 + 0.24 * 0.6) * 0.6 = 0.0972$

$t=3$ 과  $4$ 일 때  $\alpha[3][1] = (\alpha[2][1] * a_{11} + \alpha[2][2] * a_{21}) * b_1(\text{청소}) = (0.0138 * 0.7 + 0.0972 * 0.4) * 0.5 = 0.02427$

$\alpha[3][2] = (\alpha[2][1] * a_{12} + \alpha[2][2] * a_{22}) * b_2(\text{청소}) = (0.0138 * 0.3 + 0.0972 * 0.6) * 0.1 = 0.00625$

$\alpha[4][1] = (\alpha[3][1] * a_{11} + \alpha[3][2] * a_{21}) * b_1(\text{쇼핑}) = (0.02427 * 0.7 + 0.00625 * 0.4) * 0.4 = 0.00780$

$\alpha[4][2] = (\alpha[3][1] * a_{12} + \alpha[3][2] * a_{22}) * b_2(\text{쇼핑}) = (0.02427 * 0.3 + 0.00625 * 0.6) * 0.3 = 0.00331$

답은  $P(\mathbf{O} | \Theta) = \alpha[4][1] + \alpha[4][2] = 0.0111$

## 7.4.2 디코딩

### ■ 디코딩 문제란?

- 모델  $\Theta$ 가 주어진 상황에서, 관측벡터  $\mathbf{O}$ 를 얻었을 때 그것의 최적 상태열은?
- 예) 그꼬저께 산책, 그저께 산책, 어제 청소했다는데 3일간 그곳 날씨는?
- 무엇을 기준으로 최적을 판단할 것인가?

### ■ 생각 1

- 예를 들어,  $\mathbf{O}=(o_1=\text{산책}, o_2=\text{산책}, o_3=\text{청소}, o_4=\text{쇼핑})$ 이었다면, 산책은  $s_1$ 보다  $s_2$ 가 확률이 높으므로  $s_2$ , 청소는  $s_1$ 이 확률이 높으므로  $s_1$ , 쇼핑은  $s_1$ 이 확률이 높으므로  $s_1$ 을 취함. 즉  $\mathbf{O}$ 의 최적 상태열은  $\mathbf{Q}=s_2s_2s_1s_1$ ='해해비비'로 함
- 합리적인가?

## 7.4.2 디코딩

- 합리적인 생각은,

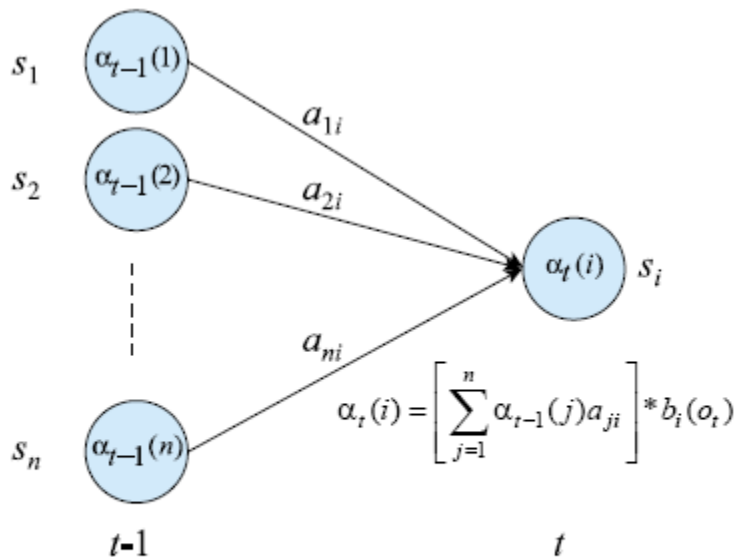
- $P(\mathbf{O}, \mathbf{Q} | \Theta)$ 를 기준 함수로 채택하고 이것을 최대화하는  $\hat{\mathbf{Q}}$ 를 찾아야 한다. 즉,

$$\hat{\mathbf{Q}} = \underset{\text{모든 } \mathbf{Q}}{\operatorname{arg\,max}} P(\mathbf{O}, \mathbf{Q} | \Theta) \quad (7.16)$$

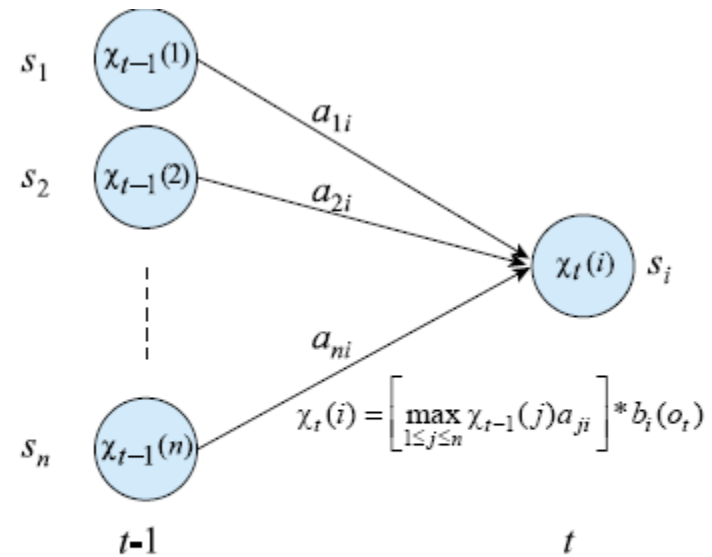
- 평가 문제와 비슷하다.

- 모든  $\mathbf{Q}$ 에 대해 계산하고 (그것을 더하는 대신) 그것 중 가장 큰 것을 취함
    - 이런 낱낱 방법은 계산 폭발
    - 동적 프로그래밍을 이용하자.

$\chi_t(i)$ : 관측벡터의 일부를 관측하고 시간  $t$ 에서 상태  $s_i$ 에 있을 최대 확률



(a) 평가 문제 (합 연산)



(b) 디코딩 문제 (최대 선택 연산)

그림 7.16 평가 문제와 디코딩 문제의 연산 차이

## 7.4.2 디코딩

### ■ Viterbi 알고리즘

- 전방 계산 ( $\tau$ 는 매 단계에서 최적 경로 기록)

$$\chi_1(i) = \pi_i b_i(o_1), \quad 1 \leq i \leq n \quad (7.18)$$

$$\left. \begin{aligned} \chi_t(i) &= \left[ \max_{1 \leq j \leq n} \chi_{t-1}(j) a_{ji} \right] * b_i(o_t), \quad 2 \leq t \leq T, 1 \leq i \leq n \\ \tau_t(i) &= \arg \max_{1 \leq j \leq n} \left[ \chi_{t-1}(j) a_{ji} \right], \quad 2 \leq t \leq T, 1 \leq i \leq n \end{aligned} \right\} \quad (7.19)$$

- 최적 경로 역추적

$$\left. \begin{aligned} \hat{q}_T &= \arg \max_{1 \leq j \leq n} \chi_T(j) \\ \hat{q}_t &= \tau_{t+1}(\hat{q}_{t+1}), \quad t = T-1, T-2, \dots, 1 \end{aligned} \right\} \quad (7.20)$$

## 7.4.2 디코딩

### ■ Viterbi 알고리즘

- 가상 코드로 쓰면,

알고리즘 [7.2] 디코딩을 위한 Viterbi 알고리즘

입력: HMM  $\Theta = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$ , 관측 벡터  $\mathbf{O} = o_1 o_2 \cdots o_T$

출력: 최적 경로  $\hat{\mathbf{Q}} = (\hat{q}_1, \hat{q}_2, \dots, \hat{q}_T)$

알고리즘:

1. 배열  $\chi[1 \cdots T][1 \cdots n]$ 와  $\tau[1 \cdots T][1 \cdots n]$ 을 생성하라.
2. **for** ( $i = 1$  **to**  $n$ )  $\chi[1][i] = \pi_i * b_i(o_1)$ ; // (7.18)
3. **for** ( $t = 2$  **to**  $T$ )
4.   **for** ( $i = 1$  **to**  $n$ ) {
5.      $\chi[t-1][j] * a_{ji}$ ,  $1 \leq j \leq n$  중에 가장 큰 것의 첨자를  $k$ 라 둔다.
6.      $\chi[t][i] = \chi[t-1][k] * a_{ki} * b_i(o_t)$ ;     // (7.19)
7.      $\tau[t][i] = k$ ;     // (7.19)
8.   }
- // 지금부터 (7.20)에 의한 경로 역 추적
9.  $\chi[T][j]$ ,  $1 \leq j \leq n$  중에 가장 큰 것의 첨자를  $\hat{q}_T$ 로 한다.
10. **for** ( $t = T-1$  **to**  $1$ )  $\hat{q}_t = \tau[t+1][\hat{q}_{t+1}]$ ;     // (7.20)

시간 복잡도는  $\Theta(N^2T)$

## 7.4.2 디코딩

### ■ 예제 7.7 Viterbi 알고리즘에 의한 디코딩

$$t=1 \text{ 일 때 } \chi[1][1] = \pi_1 * b_1(\text{산책}) = 0.6 * 0.1 = 0.06$$

$$\chi[1][2] = \pi_2 * b_2(\text{산책}) = 0.4 * 0.6 = 0.24$$

$$t=2 \text{ 일 때 } \chi[2][1] = \max(\chi[1][1] * a_{11}, \chi[1][2] * a_{21}) * b_1(\text{산책}) = (0.24 * 0.4) * 0.1 = 0.0096$$

$$\tau[2][1] = 2$$

$$\chi[2][2] = \max(\chi[1][1] * a_{12}, \chi[1][2] * a_{22}) * b_2(\text{산책}) = (0.24 * 0.6) * 0.6 = 0.0864$$

$$\tau[2][2] = 2$$

$$t=3 \text{ 과 } 4 \text{ 일 때 } \chi[3][1] = \max(\chi[2][1] * a_{11}, \chi[2][2] * a_{21}) * b_1(\text{청소}) = (0.0864 * 0.4) * 0.5 = 0.01728$$

$$\tau[3][1] = 2$$

$$\chi[3][2] = \max(\chi[2][1] * a_{12}, \chi[2][2] * a_{22}) * b_2(\text{청소}) = (0.0864 * 0.6) * 0.1 = 0.00518$$

$$\tau[3][2] = 2$$

$$\chi[4][1] = \max(\chi[3][1] * a_{11}, \chi[3][2] * a_{21}) * b_1(\text{쇼핑}) = (0.01728 * 0.7) * 0.4 = 0.00484$$

$$\tau[4][1] = 1$$

$$\chi[4][2] = \max(\chi[3][1] * a_{12}, \chi[3][2] * a_{22}) * b_2(\text{쇼핑}) = (0.01728 * 0.3) * 0.3 = 0.00156$$

$$\tau[4][2] = 1$$

역추적하면  $\hat{q}_4 = 1, \hat{q}_3 = 1, \hat{q}_2 = 2, \hat{q}_1 = 2$

답은  $\hat{\mathbf{Q}} = (\hat{q}_1 = s_2, \hat{q}_2 = s_2, \hat{q}_3 = s_1, \hat{q}_4 = s_1)$

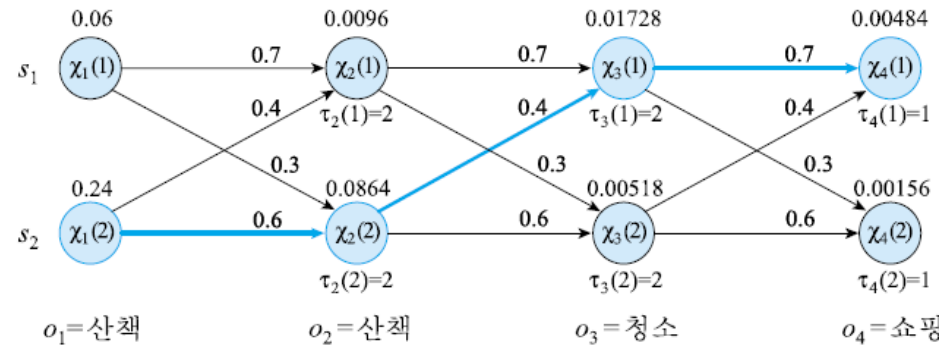


그림 7.17 Viterbi 알고리즘의 전방 계산과 최적 경로 역 추적 (파란 색이 최적 경로)

### 7.4.3 학습

- 학습 문제란?
  - 관측 벡터  $\mathbf{O}$ 가 주어져 있을 때 HMM의 매개 변수  $\Theta$ 는?
    - 즉  $\mathbf{O}$ 의 발생 확률을 최대로 하는  $\Theta$ 를 찾는 문제
  - 평가와 디코딩의 반대 작업
  - 평가와 디코딩 같은 분석적 방법 없음
    - 수치적 방법 필요
- 학습의 목적을 적어보면,

$$\hat{\Theta} = \arg \max_{\text{모든 } \Theta} P(\mathbf{O} | \Theta) \quad (7.21)$$



## 7.4.3 학습

- 학습 알고리즘 스케치
  - 반복적 개선

알고리즘 [7.3]

HMM 학습 알고리즘 스케치

입력: HMM 아키텍처 ( $n, m$ , 그래프 형태 등), 관측 벡터  $\mathbf{O} = o_1 o_2 \dots o_T$

출력: HMM  $\hat{\Theta} = (\mathbf{A}, \mathbf{B}, \pi)$

알고리즘:

1.  $\Theta$ 를 초기화 하라.
2. 적절한 방법으로  $P(\mathbf{O} | \Theta^{\text{new}}) > P(\mathbf{O} | \Theta)$ 인 개선된  $\Theta^{\text{new}}$ 를 찾아라.
3. 만족스러우면  $\hat{\Theta} = \Theta^{\text{new}}$ 로 하고 멈추고, 그렇지 않으면  $\Theta = \Theta^{\text{new}}$ 로 하고 2로 가라.

## 7.4.3 학습

- 가진 것과 찾아야 하는 것
  - 가진 것  $\mathbf{O}$ , 찾아야 하는 것  $\Theta = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$
  - $\mathbf{O}$ 로 직접  $\Theta$ 를 찾을 수 없다.
    - 은닉 변수 latent variable  $\gamma$ 와  $\kappa$  등장
    - 가우시언 혼합 추정과 유사성 있음
- Baum-Welch 알고리즘의 골격

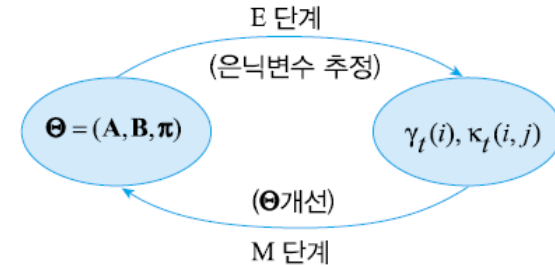


그림 7.18 Baum-Welch 학습 알고리즘은 EM 알고리즘의 일종임

### 알고리즘 [7.4] HMM 학습을 위한 Baum-Welch 알고리즘 스케치

입력: HMM 아키텍처 ( $n, m$ , 그래프 형태 등), 관측 벡터  $\mathbf{O} = o_1 o_2 \dots o_T$

출력: HMM  $\hat{\Theta} = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$

알고리즘:

1.  $\Theta$ 를 초기화 하라.
2. // 개선된  $\Theta^{\text{new}}$  찾기
  - 2.1 (E 단계)  $\Theta$ 로  $\gamma_t(i), 1 \leq t \leq T, 1 \leq i \leq n$  과  $\kappa_t(i, j), 1 \leq t \leq T-1, 1 \leq i, j \leq n$  을 추정하라.
  - 2.2 (M 단계)  $\gamma_t(i)$ 와  $\kappa_t(i, j)$ 를 가지고  $\Theta^{\text{new}}$ 를 찾아라.
3. 만족스러우면  $\hat{\Theta} = \Theta^{\text{new}}$ 로 하고 멈추고, 그렇지 않으면  $\Theta = \Theta^{\text{new}}$ 로 하고 2로 가라.

## 7.4.3 학습

### ■ E 단계 ( $\gamma$ 와 $\kappa$ 의 추정)

- $\gamma_t(i)$ 는 시간  $t$ 에서 상태  $s_i$ 에 있을 확률
- $\kappa_t(i, j)$ 는 시간  $t$ 에  $s_i$ ,  $t+1$ 에  $s_j$ 에 있을 확률

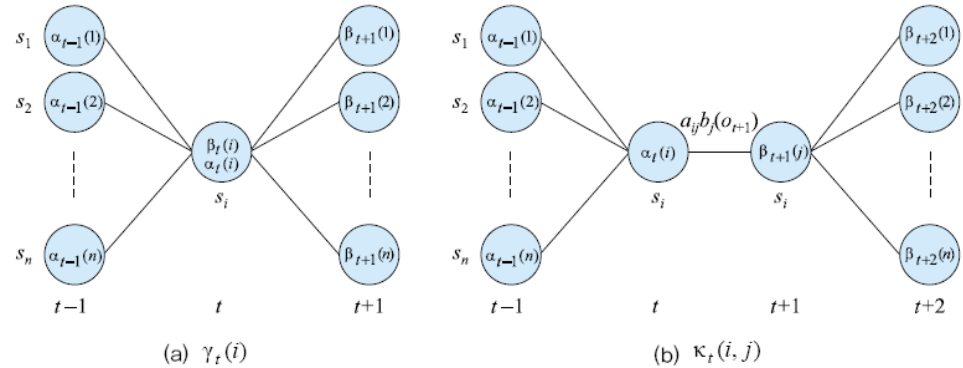


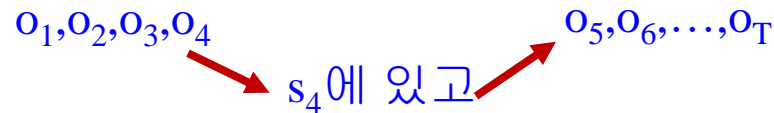
그림 7.19 은닉 변수  $\gamma_t(i)$ 와  $\kappa_t(i, j)$ 의 역할

$$\gamma_t(i) = P(q_t = s_i | \mathbf{O}, \Theta) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^n \alpha_t(j)\beta_t(j)}, \quad 1 \leq t \leq T, 1 \leq i \leq n \quad (7.22)$$

시간  $t$ 에서  $s_i$ 상태에 있을 때  $o_t$ 관측되는 것이 중요. 그 외  $T-1$ 개의 관측은 무방함.

$$\kappa_t(i, j) = \frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{k=1}^n \sum_{l=1}^n \alpha_t(k)a_{kl}b_l(o_{t+1})\beta_{t+1}(l)}, \quad 1 \leq t \leq T-1, 1 \leq i, j \leq n \quad (7.26)$$

$\gamma_t(i)$ 는 관측벡터  $o_1, o_2, \dots, o_T$ 가 주어진 조건하에서 시간  $t$ 일때  $s_i$ 에 있을 확률임. 이는  $\alpha_t(i) * \beta_t(i)$ 로 표현됨



(1)  $o_1 \sim o_4$ 가 주어진 상황에서  $s_4$ 에 있을 확률  $\alpha_t(4)$

(2)  $t=4$ 에서  $s_4$ 이후,  $t=5$ 부터 이후에 나머지  $o_5 \sim o_T$ 도 관측될 확률:  $\beta_t(4)$

## 7.4.3 학습

### ■ E 단계 ( $\gamma$ 와 $\kappa$ 의 추정)

- $\gamma_t(i)$ 는 시간  $t$ 에서 상태  $s_i$ 에 있을 확률
- $\kappa_t(i, j)$ 는 시간  $t$ 에  $s_i$ ,  $t+1$ 에  $s_j$ 에 있을 확률

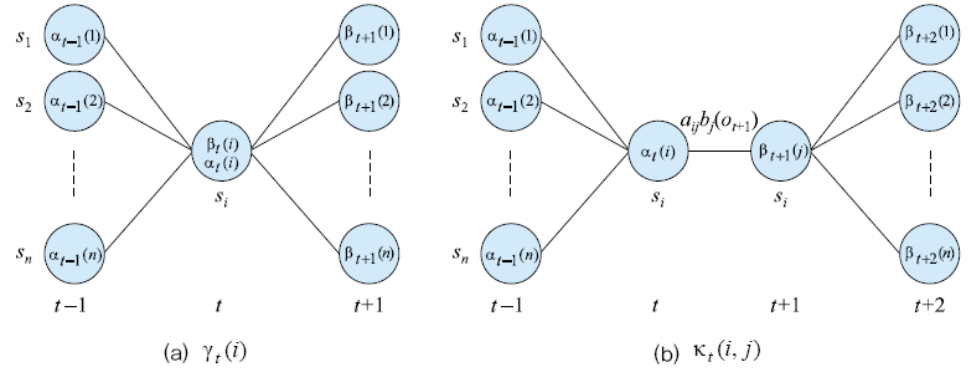


그림 7.19 은닉 변수  $\gamma_t(i)$ 와  $\kappa_t(i, j)$ 의 역할

$$\gamma_t(i) = P(q_t = s_i | \mathbf{O}, \Theta) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^n \alpha_t(j)\beta_t(j)}, \quad 1 \leq t \leq T, 1 \leq i \leq n \quad (7.22)$$

$$\kappa_t(i, j) = \frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{k=1}^n \sum_{l=1}^n \alpha_t(k)a_{kl}b_l(o_{t+1})\beta_{t+1}(l)}, \quad 1 \leq t \leq T-1, 1 \leq i, j \leq n \quad (7.26)$$

- $\gamma$ 와  $\kappa$ 를 구하는데 필요한  $\alpha$ 와  $\beta$ 는

$$\alpha_1(i) = \pi_i b_i(o_1), \quad 1 \leq i \leq n \quad (7.13)$$

$$\alpha_t(i) = \left[ \sum_{j=1}^n \alpha_{t-1}(j)a_{ji} \right] * b_i(o_t), \quad 2 \leq t \leq T, 1 \leq i \leq n \quad (7.14)$$

$$\beta_T(i) = 1, \quad 1 \leq i \leq n \quad (7.24)$$

$$\beta_t(i) = \sum_{j=1}^n a_{ij} b_j(o_{t+1}) \beta_{t+1}(j), \quad 1 \leq t \leq T-1, 1 \leq i \leq n \quad (7.25)$$

## 7.4.3 학습

### ■ M 단계 ( $\Theta$ 의 추정)

- E 단계에서 구한  $\gamma$ 와  $\kappa$ 로  $\Theta^{\text{new}}$ 의 재추정(reestimation)
- $P(O|\Theta^{\text{new}}) > P(O|\Theta)$ 이어야 함. 즉  $\Theta^{\text{new}}$ 는  $\Theta$ 보다  $O$ 를 ‘잘 설명해야’ 함

$$\begin{aligned}
 a_{ij}^{\text{new}} &= \frac{s_i \text{에서 } s_j \text{로 이전할 기대값}}{s_i \text{에서 이전할 기대값}} \\
 &= \frac{t=1 \text{일 때 } s_i \text{에서 } s_j \text{로 이전할 확률} + \dots + t=T-1 \text{일 때 } s_i \text{에서 } s_j \text{로 이전할 확률}}{t=1 \text{일 때 } s_i \text{에서 이전할 확률} + \dots + t=T-1 \text{일 때 } s_i \text{에서 이전할 확률}} \\
 &= \frac{\sum_{t=1}^{T-1} \kappa_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)}, \quad 1 \leq i, j \leq n
 \end{aligned}
 \tag{7.27}$$

$$\begin{aligned}
 b_i(v_k)^{\text{new}} &= \frac{s_i \text{에서 } v_k \text{를 관측할 기대값}}{s_i \text{에 있을 기대값}} \\
 &= \frac{o_t = v_k \text{인 모든 } t \text{에 대해 } s_i \text{에 있을 확률의 합}}{t=1 \text{일 때 } s_i \text{에 있을 확률} + \dots + t=T \text{일 때 } s_i \text{에 있을 확률}} \\
 &= \frac{\sum_{t=1}^T \gamma_t(i)}{\sum_{t=1}^T \gamma_t(i)}, \quad 1 \leq i \leq n, 1 \leq k \leq m
 \end{aligned}
 \tag{7.28}$$

$$\pi_i^{\text{new}} = t \text{가 } 1 \text{일 때 } s_i \text{에 있을 확률} = \gamma_1(i), \quad 1 \leq i \leq n
 \tag{7.29}$$

## 7.4.3 학습

### ■ Baum-Welch 알고리즘

- 초기화는 어떻게?
- 멈춤 조건은 어떻게?

### ■ 특성

- 수렴한다.
- 욕심 알고리즘이므로 전역 최적점 보장 못한다.

### ■ 어고딕과 좌우 모델

- 크기 문제 (어고딕 모델)
- 다중 관측 학습 (좌우 모델)

#### 알고리즘 [7.5]

#### HMM 학습을 위한 Baum-Welch 알고리즘

입력: HMM 아키텍처 ( $n, m$ , 그래프 형태 등), 관측 벡터  $\mathbf{O} = o_1 o_2 \dots o_T$

출력: HMM  $\hat{\Theta} = (\mathbf{A}, \mathbf{B}, \pi)$

알고리즘:

1.  $\Theta$ 를 초기화 하라.

2. **repeat** {

// E 단계

// 라인 3의  $\alpha$ 는 (7.13)~(7.14),  $\beta$ 는 (7.24)~(7.25)

3.  $\mathbf{O}$ 와  $\Theta$ 를 가지고  $\alpha_t(i), 1 \leq t \leq T, 1 \leq i \leq n$  과  $\beta_t(i), 1 \leq t \leq T, 1 \leq i \leq n$  을 구한다.

// 라인 4의  $\gamma$ 는 (7.22),  $\kappa$ 는 (7.26)

4.  $\alpha$ 와  $\beta$ 로  $\gamma_t(i), 1 \leq t \leq T, 1 \leq i \leq n$  과  $\kappa_t(i, j), 1 \leq t \leq T-1, 1 \leq i, j \leq n$  을 계산한다.

// M 단계

5.  $\gamma$ 와  $\kappa$ 를 가지고  $a_{ij}^{new}, 1 \leq i, j \leq n$ 을 추정한다. // (7.27)

6.  $\gamma$ 를 가지고  $b_i(v_k)^{new}, 1 \leq i \leq n, 1 \leq k \leq m$ 을 추정한다. // (7.28)

7.  $\gamma$ 를 가지고  $\pi_i^{new}, 1 \leq i \leq n$ 을 추정한다. // (7.29)

8.  $\Theta = (\mathbf{A}^{new}, \mathbf{B}^{new}, \pi^{new})$ ;

9. } **until** (멈춤 조건);

10.  $\hat{\Theta} = \Theta$ ;

## 7.5 부연 설명

1. HMM의 출력은 확률. 큰 장점이다.
2. 샘플 생성 능력이 있다. HMM은 생성<sup>generative</sup> 모델이다.

### 알고리즘 [7.6] HMM에 의한 관측 벡터 생성

입력: HMM  $\Theta = (\mathbf{A}, \mathbf{B}, \pi)$ , 관측 벡터의 길이  $T$

출력: 관측 벡터  $\mathbf{O} = o_1 o_2 \dots o_T$

알고리즘:

1.  $\pi$ 에 따라 초기 상태  $q_1$ 을 결정한다.
2.  $t = 1$ ;
3. **while** ( $t \neq T$ ) {
4.      $\mathbf{B}$ 에 따라 상태  $q_t$ 에서 관측 값을 결정하고 그것을  $o_t$ 에 기록한다.
5.      $\mathbf{A}$ 에 따라 다음 상태  $q_{t+1}$ 을 결정한다.
6.      $t++$ ;
7. }

## 7.5 부연 설명

3. HMM을 예측 목적으로 사용할 수 있다.
4. HMM을 분류기로 사용할 수 있다.
  - 부류 별로 독립적으로 HMM 구축
  - (7.30)으로 분류

$$\mathbf{O} \text{를 } q = \arg \max_j P(\mathbf{O} | \Theta_{\omega_j})P(\omega_j) \text{일 때 } \omega_q \text{로 분류하라.} \quad (7.30)$$

5. 적절한 아키텍처를 선택해야 한다.
  - 그녀의 삶은 어떤 모델?
  - 음성 인식은 어떤 모델?
6. 상태의 개수를 적절히 해야 한다.
7. 훌륭한 공개 소프트웨어가 있다.
  - 캠브리지 대학의 HTK (HMM Toolkit)