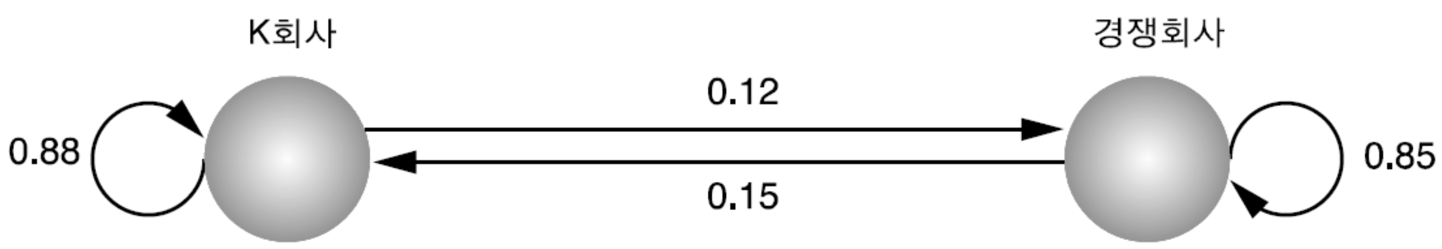
은닉 마르코프 모델(HMM)

❖ 확률 행렬과 마르코프 연쇄 (1/3)

- 확률 행렬 : 행렬의 성분이 확률로 이루어진 행렬
- 마르코프 연쇄 : 한 상태에서 다른 상태로 변할 확률이 과거의 자취보다 **현재와 바로 직전 과거**의 상태에만 의존하는 모델

K라는 분유 회사가 있다. 이 회사의 분유가 현재 분유 시장의 25%를 점유하고 있다고 한다. 전년도의 데이터에 의하면 K회사 고객의 88%가 계속 고객으로 남아 있고, 12%는 K회사의 제품에서 경쟁회사의 제품으로 바꾸었다. 경쟁회사 고객의 85%는 여전히 그 제품을 사용하고 있고, 나머지 15%는 경쟁회사의 제품에서 K회사의 제품으로 바꾼 것으로 분석되었다. 이러한 경향이 계속된다고 가정할 경우에 2년 후 그리고 더 오랜 시간이 지난 후 K회사의 시장 점유율은 어떻게 될까?



[그림 17-1] 상태-천이 다이어그램

❖ 확률 행렬과 마르코프 연쇄 (3/3)

- 상태1 : 고객이 K회사를 선택하여 K회사의 분유를 구매한다.
- 상태2 : 고객이 경쟁회사를 선택하여 그 회사의 분유를 구매한다.
- 천이 행렬 P

(행과 렬 순서로 직전과거와 현재)

$$
\begin{array}{cc}
 & \begin{array}{cc} 1 & 2 \end{array} \\
p = & \begin{bmatrix} 0.88 & 0.12 \\ 0.15 & 0.85 \end{bmatrix} \begin{array}{c} 1 \\ 2 \end{array}
\end{array}
$$

- 초기치

$$
\begin{array}{cc}
 & \begin{array}{cc} 1 & 2 \end{array} \\
p_{init} = & \begin{bmatrix} 0.25 & 0.75 \end{bmatrix}
\end{array}
$$

- 기간 t에서의 시스템의 상태

$$s_t = s_{t-1}(P) = s_{t-2}(P)(P) = \ldots = s_1(P)^{t-1}$$

- 다음해(1년후)의 시장 점유율

$$s_2 = s_1 P = [0.25, 0.75]\begin{bmatrix} 0.88 & 0.12 \\ 0.15 & 0.85 \end{bmatrix}$$

$$= [(0.25)(0.88) + (0.75)(0.15),\ (0.25)(0.12) + (0.75)(0.85)]$$

$$= [0.3325, 0.6675]$$

❖ 확률 행렬과 마르코프 연쇄 (3/3)

- 2년 후의 K회사의 시장 점유율 → 39.27%

$$s_3 = s_2 P = \begin{bmatrix} 0.3325, 0.6675 \end{bmatrix} \begin{bmatrix} 0.88 & 0.12 \\ 0.15 & 0.85 \end{bmatrix}$$

$$= \begin{bmatrix} 0.3927, 0.6073 \end{bmatrix}$$

- 이 모델에서의 결과는 **이전 해의 시장 점유율에만 영향**을 받는 것을 알 수 있다.

- 이처럼 한 상태에서 다른 상태로 변할 확률이 현재의 상태에만 의존하는 모델을 "1차 마르코프 연쇄" 라고 한다.

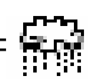
02_ 마르코프 가정, 마르코프 모델, 마르코프 과정

❖ 마르코프 가정, 마르코프 모델, 마르코프 과정 (1/3)

- 어느 지역에 매일 변하는 날씨에 대한 마르코프 모델.
- 단, 이 지역의 날씨 하루를 단위로 변하며, 관측 가능한 날씨는 맑음( ),비( ), 흐림( ) 가운데 날씨 예측은 과거 날씨의 이력을 바탕으로 내일 날씨가 어떻게 될 것인가를 결정
- 날씨 예측에 관한 통계적인 모델을 설정해 보자. 1일전 (q_{n-1}), 2일전 (q_{n-2}), 3일전 (q_{n-3}), 등의 과거의 날씨에 따라 오늘 (q_n), 날씨가 어떠했다는 통계가 있을 경우,

$$P(q_n | q_{n-1}, q_{n-2}, \cdots, q_1)$$

- 즉, n 째 날의 날씨는 $q_n \in$ {맑음( ),비( ),흐림( )} 이고, 그 날씨가 될 확률은 알려져 있는 과거(q_{n-2}, q_{n-1} ...) 의 날씨에 의존
- 확률을 사용하면 과거 날씨 이력을 바탕으로 내일, 모레의 날씨도 확률적으로 예측을 할 수 있을 것이다. 예를 들어, 만약 과거 3일간의 날씨가 {맑음( ),비( ), 흐림( )} 의 순서로 관측되었다면, 내일 날씨가 비가 올 확률은 다음과 같은 식으로 표현된다.

$P(q_4 =$  $| q_3 =$ $,\ q_2 =$  $q_1 =$ $)$

❖ 마르코프 가정, 마르코프 모델, 마르코프 과정 (2/3)

- n이 클 수록 우리가 수집해야 하는 관측 데이터도 많아져야 한다는 것

- 가변 n = 6이라고 가정하면, 우리가 수집해야 하는 과거 날씨 통계 데이터가 무려 $3^{(6-1)}$ = 243 개나 필요하게 된다. 그러므로 다음과 같은 "마르코프 가정" 이라고 하는 좀더 간단한 가정이 필요하다.

- 그래서, 어떤 열 $q_1, q_2, \ldots q_n$ 에 대하여 1차 마르코프 가정:

$$P(q_t = S_j \mid q_{t-1} = S_i, q_{t-2} = S_k, \cdots) = P(q_t = S_j \mid q_{t-1} = S_i)$$

- 마르코르 가정을 사용하여 어떤 열 $\{q_1, q_2, \ldots q_n\}$이 관측될 확률은 과거와 현재의 관측 결과의 결합 확률(joint probability)로 표현

$$\begin{aligned} &P(q_1,\ q_2, \cdots, q_{T-1}, q_T) \\ &\quad = P(q_1)P(q_2 \mid q_1)(q_3 \mid q_1,\ q_2), \cdots, P(q_{T-1} \mid q_1, q_2, \cdots, q_{T-2})P(q_T \mid q_1, q_2, \cdots, q_{T-1}) \\ &\quad = P(q_1)P(q_2 \mid q_1)(q_3 \mid q_2), \cdots, P(q_{T-1} \mid q_{T-2})P(q_T \mid q_{T-1}) \end{aligned}$$

02_ 마르코프 가정, 마르코프 모델, 마르코프 과정

❖ 마르코프 가정, 마르코프 모델, 마르코프 과정 (3/3)

- 우변의 두 상태 간의 천이 확률이 시간에 관하여 독립이라고 가정하면, 상태 천이 확률은 다음과 같이 표현할 수 있다.

$$\begin{cases} a_{ij} \ge 0, \ \forall_{i,j} \\ \sum_{j=1}^{N} a_{ij} = 1, \ \forall_{i} \end{cases}$$

※ $a_{ij} = P(q_t = S_j \mid q_{t-1} = S_i)$

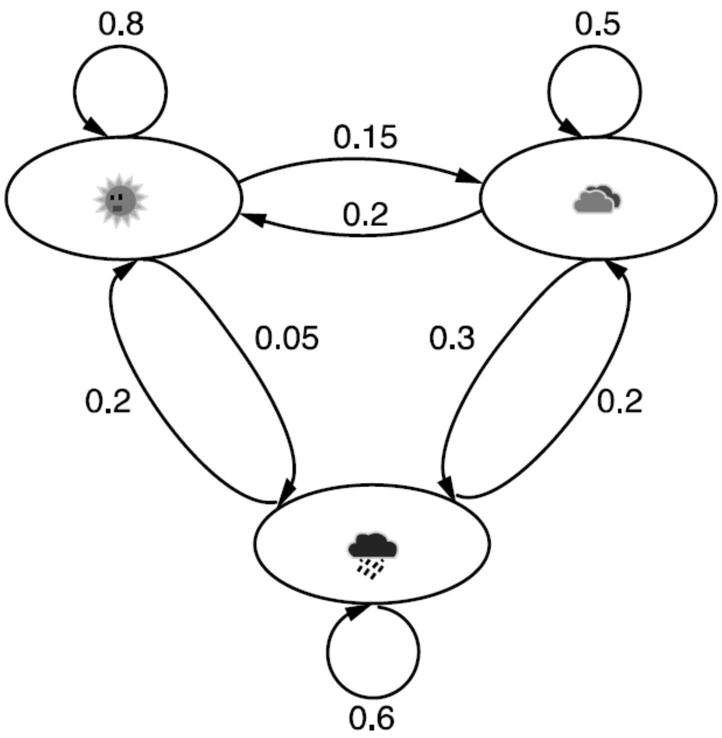
- 1차 마르코프 모델에서 이들 확률 값들을 상태 천이 다이어그램으로 표현 할 수 있다. 날씨 영역은 3개의 상태(S={ , , })를 가지고 매일 위의 천이 확률 표와 같은 확률값 ($P(q_n|q_{n-1})$)에 따라 가능한 다른 상태로의 천이가 됨.

오늘의 날씨	내일의 날씨		
	맑음()	비()	흐림()
맑음()	0.8	0.05	0.15
비()	0.2	0.6	0.2
흐림()	0.2	0.3	0.5

$$\begin{pmatrix} 0.8 & 0.05 & 0.15 \\ 0.2 & 0.6 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{pmatrix}$$

❖ 마르코프 가정, 마르코프 모델, 마르코프 과정 (3/3)

- 천이 확률값들은 1차 마르코프 모델의 상태 천이 다이어그램으로 표현할 수 있다. 날씨영역은 세 가지 상태(S = ={ , , }) 를 나타낸다. 그리고 날씨는 천이 확률표와 같이 확률값 P(qn|qn − 1)에 따라 다른 상태로의 천이가 이루어진다.



오늘의 날씨	내일의 날씨		
	맑음()	비()	흐림()
맑음()	0.8	0.05	0.15
비()	0.2	0.6	0.2
흐림()	0.2	0.3	0.5

❖ 마르코프 가정, 마르코프 모델, 마르코프 과정 (3/3)

오늘 날씨(q_1)가 (☀)일 경우에 내일 날씨(q_2)가 (☀)이 되고 모레 날씨(q_3)가 (🌧)가 될 확률은 얼마일까?

$$P(q_2=\text{☀},\ q_3=\text{🌧} \mid q_1=\text{☀})$$

$$=P(q_3=\text{🌧} \mid q_2=\text{☀},\ q_1=\text{☀})\cdot P(q_2=\text{☀} \mid q_1=\text{☀})$$

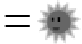
$$=P(q_3=\text{🌧} \mid q_2=\text{☀})\cdot P(q_2=\text{☀} \mid q_1=\text{☀}) \leftarrow \text{마르코프 가정} \qquad (17.12)$$

$$=0.05\times 0.8$$

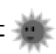
$$=0.04$$

02_ 마르코프 가정, 마르코프 모델, 마르코프 과정

❖ 마르코프 가정, 마르코프 모델, 마르코프 과정 (3/3)

그렇다면 어제, 오늘의 날씨가 각각 $q_1=$, $q_2=$일 때, 내일 날씨가 $q_3=$이 될 확률은 얼마일까?

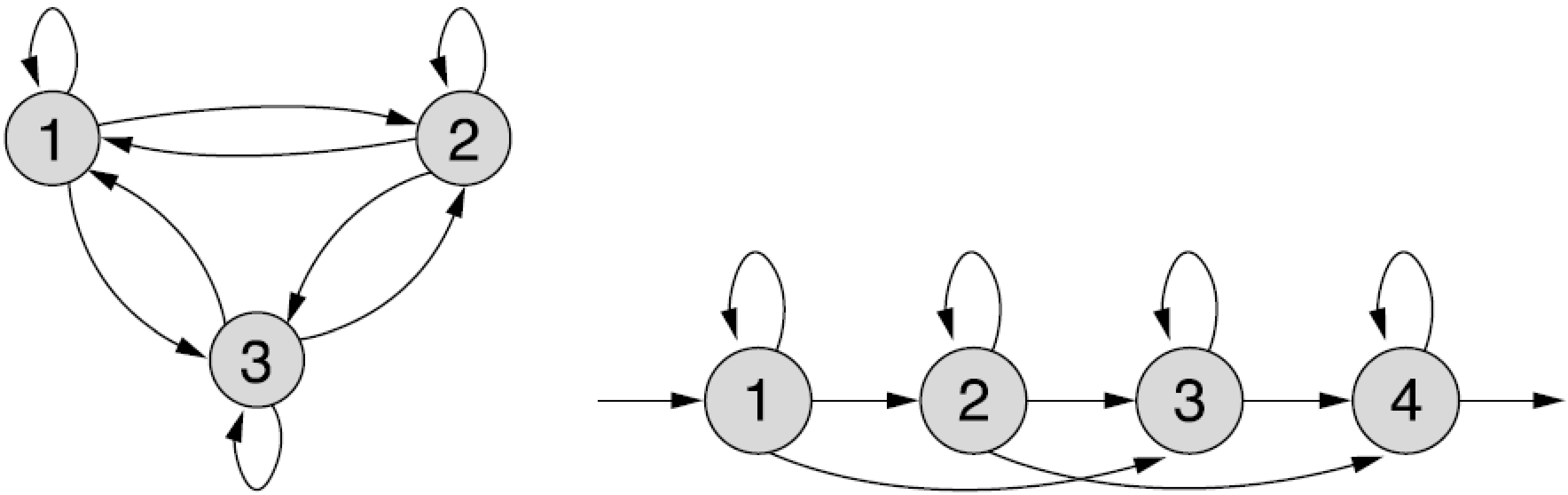
$P(q_3=$$\mid q_2=$$,q_1=$$)$

$=P(q_3=$$\mid q_2=$$)$ ← 마르코프 가정 (17.14)

$=0.2$

❖ 마르코프 모델

- 마르코프 모델은 상태 천이의 형태에 따라 날씨 HMM모델과 같은 에르고딕(ergodic)모델과 음식인식에 많은 적용되는 좌우(left-right)모델이 있다.



(a) 에르고딕(ergodic) 모델 (b) 좌우(left−right) 모델

[그림 17−3] 마르코프 모델의 종류

03_ 은닉 마르코프 모델

❖ 날씨 HMM 모델 (1/4)

- 마르코프 모델에서 설명한 날씨 모델을 은닉 마르코프 모델로 바꾸어 날씨를 숨겨 보자. 그런데 날씨를 숨긴다는 가정이 어떤 상황이 될까?

- 여러분이 어느 외딴 집에 갇혀 있다고 가정해보자. 날씨를 아는 방법은 매일 밥을 가져다 주는 감시자가 우산을 가지고 있는지( ) 안 가지고( ) 있는지를 확인하는 것이 오늘 날씨를 알 수 있는 유일한 실마리가 된다고 하자.

- 아래 표와 같은 우산 확률을 가정 할 수 있을 것이다. 날씨가 맑으면 감시자가 우산을 가지고 있을 확률은 0.1 비가 오면 그 확률이 0.8, 흐린 날씨이면 그 확률이 0.3이 된다는 의미이다.

날씨	우산 확률
맑음( )	0.1
비( )	0.8
흐림( )	0.3

❖ 날씨 HMM 모델 (2/4)

날씨 마르코프 과정에서는 날씨의 관측이 가능하므로, 어떤 날씨 열 $\{q_1, q_2, \cdots, q_n\}$이 관측될 확률은 (17.15)와 같이 결합 확률(joint probability)로 바로 표현할 수 있었다.

$$P(q_1, q_2, \cdots, q_n) = \prod_{i=1}^{n} P(q_i \mid q_{i-1}) \tag{17.15}$$

그러나 이제 실제 날씨는 직접 관측이 불가능하며 은닉(hidden)되어 있다. 그러므로 어떤 날씨 $q_n \in$ {맑음(☀), 비(🌧), 흐림(☁)}이 관측될 확률은 감시자가 i번째 날에 우산을 가져 왔다면 o_i = ☂로, 우산을 가져오지 않았다면 o_i = ~~☂~~가 되어 우산 열 o_i의 관측을 통해서만 추정할 수 있을 뿐이다. 이것은 조건부 확률 $P(q_i \mid o_i)$에 해당하며 베이즈의 정리에 따라 다음과 같이 표현된다.

$$P(q_i \mid o_i) = \frac{P(o_i \mid q_i) P(q_i)}{P(o_i)} \tag{17.16}$$

❖ 날씨 HMM 모델 (3/4)

만약 확률 과정에서 날씨열($Q=\{q_1, q_2, \cdots, q_n\}$)과 우산열($O=\{o_1, o_2, \cdots, o_n\}$)이 주어진다면, 관측되는 우산열에 따른 날씨열이 일어날 조건부 확률은 (17.17)과 같다.

$$P(q_i, \cdots, q_n \mid o_i, \cdots, o_n) = \frac{P(o_1, \cdots, o_n \mid q_1, \cdots, q_n)P(q_1, \cdots, q_n)}{P(o_1, \cdots, o_n)} \qquad (17.17)$$

(17.17)에서 날씨열의 확률($P(q_1, \cdots, q_n)$)과 우산열의 확률($P(o_1, \cdots, o_n)$)이 주어진다면, $P(o_1, \cdots, o_n | q_1, \cdots, q_n)$는 모든 i에 대하여 독립이므로 $\prod_{i=1}^{n} P(o_i \mid q_i)$가 성립하게 된다.

감시자가 우산을 가지고 있는지(), 안 가지고 있는지()의 우산열에 대한 관측 확률에서 바깥 날씨가 (맑음(), 비(), 흐림()) 중의 하나가 될 날씨열에 대한 확률을 구하기 위해서

❖ 날씨 HMM 모델 (4/4)

날씨에 독립인 우산열이 발생할 사전 확률 $P(o_1,\ \cdots,\ o_n)$은 동일하므로 제거될 수 있다. 그러므로 실제 확률에 비례하는 우도(likelihood) 확률(L)은 (17.18)과 같이 표현된다.

$$P(q_1, q_2, \cdots, q_{T-1}, q_T) = P(q_1)P(q_2 \mid q_1)(q_3 \mid q_2), \cdots, P(q_{T-1} \mid q_{T-2})P(q_T \mid q_{T-1})$$

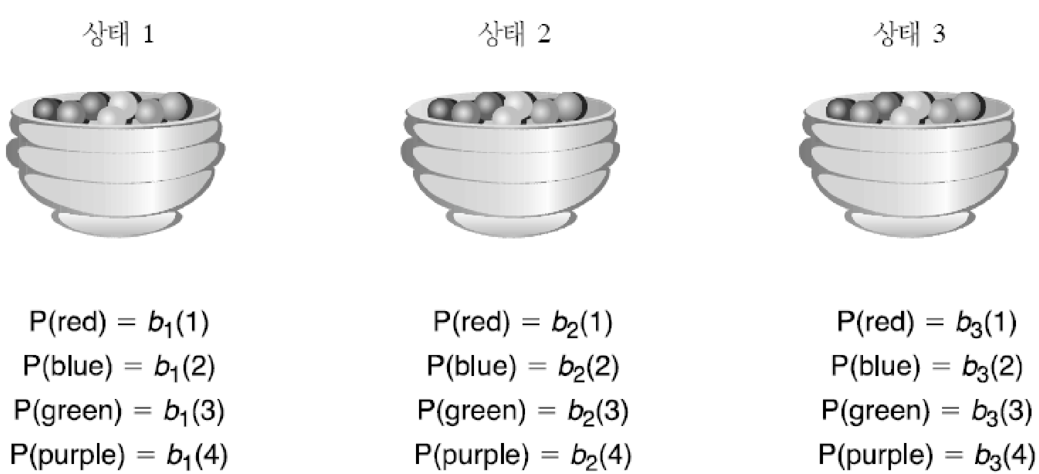
$$P(q_1, \cdots, q_n \mid o_1, \cdots, o_n) \propto L(q_1, \cdots, q_n \mid o_1, \cdots, o_n) = P(o_1, \cdots, o_n \mid q_1, \cdots, q_n)P(q_1, \cdots, q_n) \qquad (17.18)$$

만약 1차 마르코프 가정일 경우에 (17.18)은 (17.19)와 같이 간략화된다.

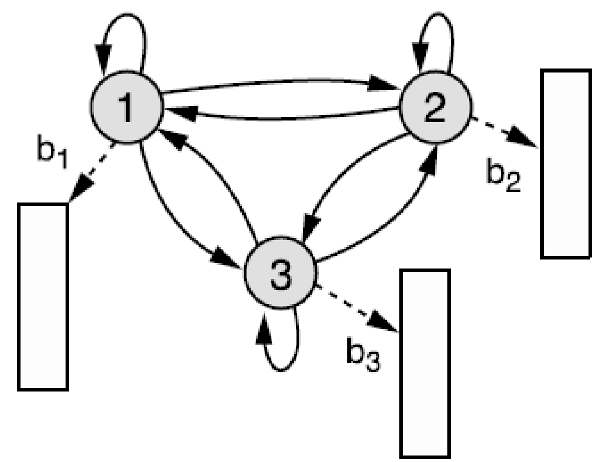
$$P(q_1, \cdots, q_n \mid o_1, \cdots, o_n) \propto L(q_1, \cdots, q_n \mid o_1, \cdots, o_n) = \prod_{i=1}^{n} P(o_i \mid q_i) \prod_{i=1}^{n} P(q_i \mid q_{i-1}) \qquad (17.19)$$

❖ 커튼 뒤에 숨어진 항아리 속의 공 HMM 모델(1/2)

어떤 방에 커튼이 드리워져 있고 커튼 뒤에는 어떤 사람과 네 가지 색깔(red, blue, green, purple)의 공이 들어 있는 3개의 항아리가 있다. 커튼 뒤의 사람이 어떤 확률 과정에 따라서 항아리를 선택하고 항아리 속에서 공을 꺼낸다. 그리고 여러분에게 공을 보여주고 항아리 속으로 그 공을 도로 집어넣는 과정을 반복한다.



(a) 항아리 속 공 HMM 모델의 출력 확률



(b) 항아리 속 공 HMM 모델의 상태 천이도

❖ 커튼 뒤에 숨어진 항아리 속의 공 HMM 모델(2/2)

- 이 모델의 경우, HMM은 다음과 같은 구성 성분들로 모델링 된다.

① N: 모델의 상태 수(항아리의 수): $S=\{S_1, S_2, \cdots, S_N\}$

② M: 이산적인 관측 심벌 개수(항아리 속에 있는 공의 색깔): $O=\{o_1, o_2, \cdots, o_M\}$

③ 상태 천이 확률 행렬 $\mathrm{A}=a_{ij}$

※ $a_{ij}=P(q_{t+1}=S_j \mid q_t=S_i)$, $1\le i, j< N$: 상태 i에서 상태 j로 천이 확률

④ 관측 심벌 확률 분포 $B=b_j(k)$,

※ $b_j(k)=P(O_t=o_k \mid q_t=S_j)$, $1\le k< M$: 상태 j에서 k가 관측될 확률

⑤ 초기 상태 분포 $\pi=\pi_i$,

※ $\pi_i=P(q_1=S_i)$, $1\le i< N$