自然语言
处理与理解

赵云蒙
华东理工大学 信息科学与工程学院
能源化工过程智能制造教育部重点实验室

2023-2024 第一学期

第6章
隐马尔可夫模型

6.1 马尔可夫模型

6.1 马尔可夫模型

★ 马尔可夫 (Andrei Andreyevich Markov)

- ✦ 1856.6.14 ~ 1922.7.20
- ✦ 俄国/苏联数学家。
- ✦ 博士导师：切比雪夫
- ✦ 在概率论、数论、函数逼近论和微分方程等方面卓有成就。他提出了用数学分析方法研究自然过程的一般图式—马尔可夫链，并开创了随机过程(马尔可夫过程)的研究。

6.1 马尔可夫模型

★ 马尔可夫模型描述

存在一类重要的随机过程：如果一个系统有 $\boldsymbol{N}$ 个状态 $S_1, S_2, \dots, S_N$，随着时间的推移，该系统从某一状态转移到另一状态。如果用 $\boldsymbol{q_t}$ 表示系统在时间 $\boldsymbol{t}$ 的状态变量，那么，$\boldsymbol{t}$ 时刻的状态取值为 $S_j(1 \le j \le N)$ 的概率取决于前 $t-1$ 个时刻$(1, 2, \dots, t-1)$的状态，该概率为：

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

6.1 马尔可夫模型

★ **假设1：**

如果在特定情况下，系统在时间 t 的状态只与其在时间 $t-1$ 的状态相关，则该系统构成一个离散的一阶马尔可夫链：

$$p\left(q_t = S_j \middle| q_{t-1} = S_i, q_{t-2} = S_k, \cdots\right) = p(q_t = S_j | q_{t-1} = S_i) \tag{6-1}$$

★ **假设2：**

如果只考虑公式(6.1)独立于时间 t 的随机过程，即所谓的不动性假设，状态与时间无关，那么：

$$p(q_t = S_j | q_{t-1} = S_i) = a_{ij}, \qquad 1 \le i, j \le N \qquad (6\text{-}2)$$

该随机过程称为**马尔可夫模型(Markov Model)**。

6.1 马尔可夫模型

★ 在马尔可夫模型中，状态转移概率 a_{ij} 必须满足下列条件：

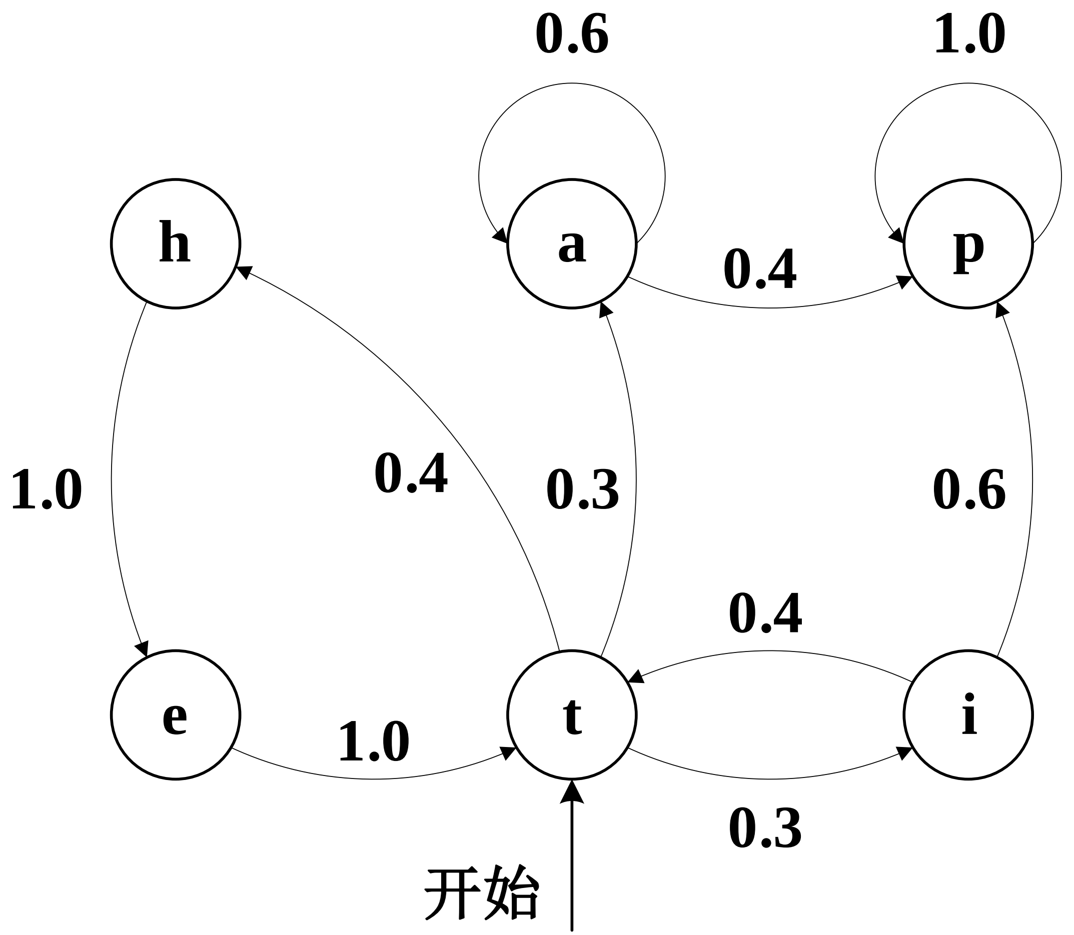
$$a_{ij} \geq 0 \tag{6-3}$$

$$\sum_{j=1}^{N} a_{ij} = 1 \tag{6-4}$$

6.1 马尔可夫模型

★ **马尔可夫链可以表示成状态图（转移弧上有概率的非确定的有限状态自动机）**

- 零概率的转移弧省略。
- 每个节点上所有发出弧的概率之和等于1。



6.1 马尔可夫模型

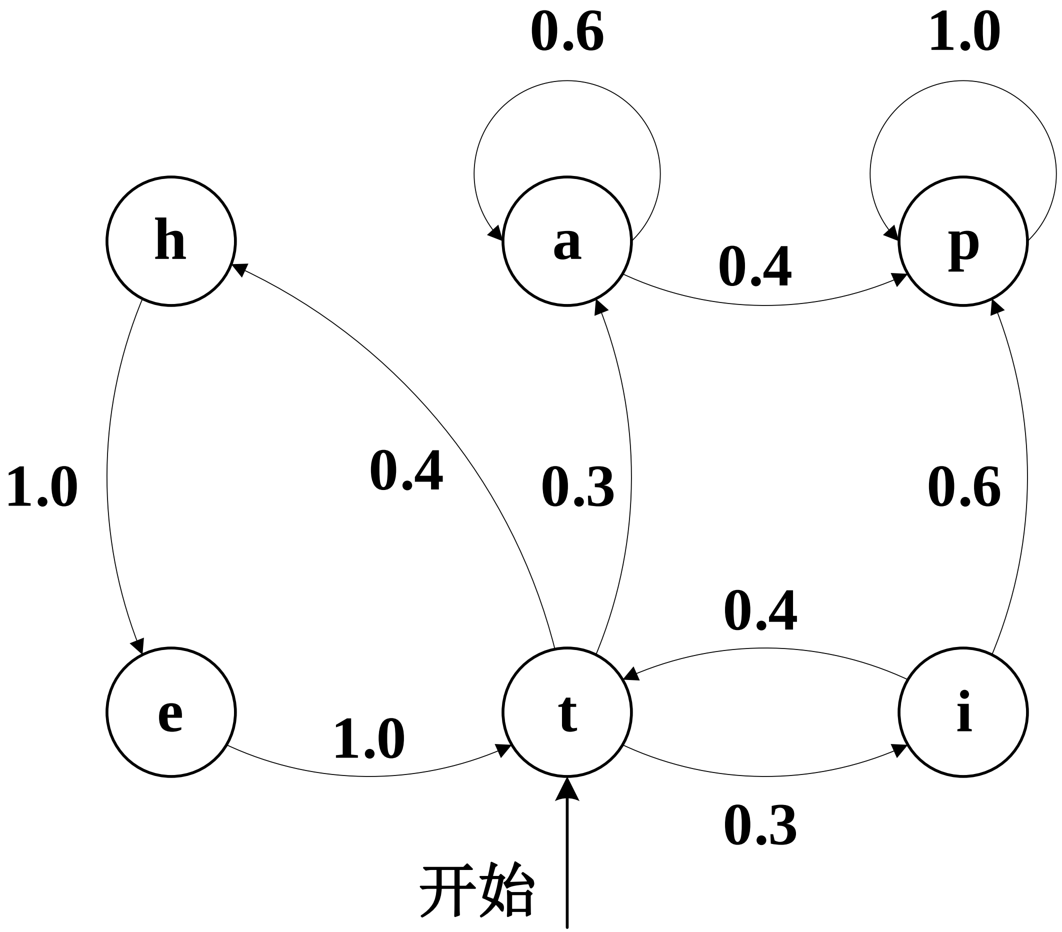
★ 状态序列$S_1, S_2, \ldots, S_T$ 的概率：

$$p(S_1, S_2, \ldots, S_T)$$
$$= p(S_1) \times p(S_2|S_1) \times p(S_3|S_1, S_2) \times \cdots \times p(S_T|S_1, \ldots, S_{T-1})$$
$$= p(S_1) \times p(S_2|S_1) \times p(S_3|S_2) \times \cdots \times p(S_T|S_{T-1})$$

$$= \pi_{S_1} \prod_{t=1}^{T-1} a_{S_T S_{T+1}} \tag{6-5}$$

其中，$\pi_i = p(q_1 = S_i)$，为初始状态概率。

6.1 马尔可夫模型



$$p(t,i,p) = p(S_1 = t) \times p(S_2 = i|S_1 = t) \times p(S_3 = p|S_2 = i)$$
$$= 1.0 \times 0.3 \times 0.6 = 0.18$$

6.2 隐马尔可夫模型

6.2 隐马尔可夫模型

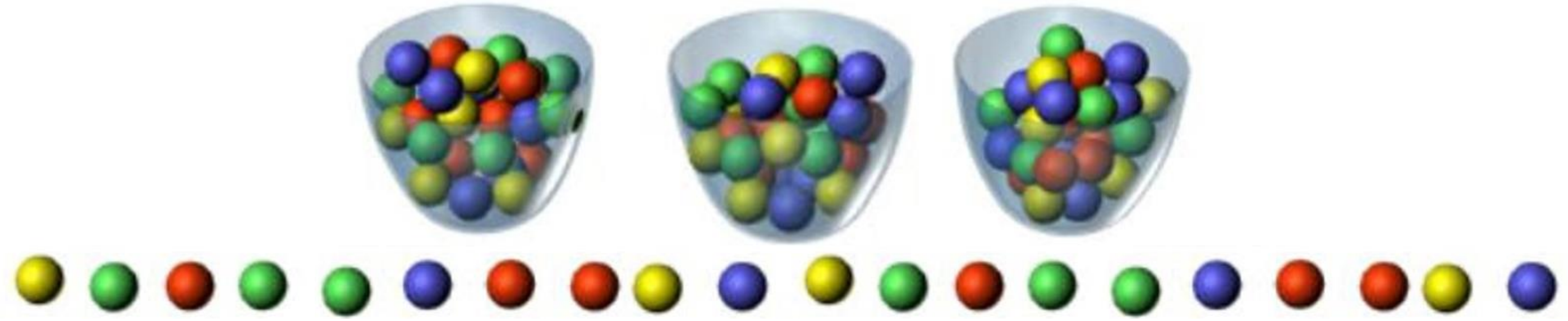
★ **隐马尔可夫模型(Hidden Markov Model, HMM)**

✦ 创建于20世纪70年代，是美国数学家鲍姆(**Leonard E. Baum**)等人提出来的。

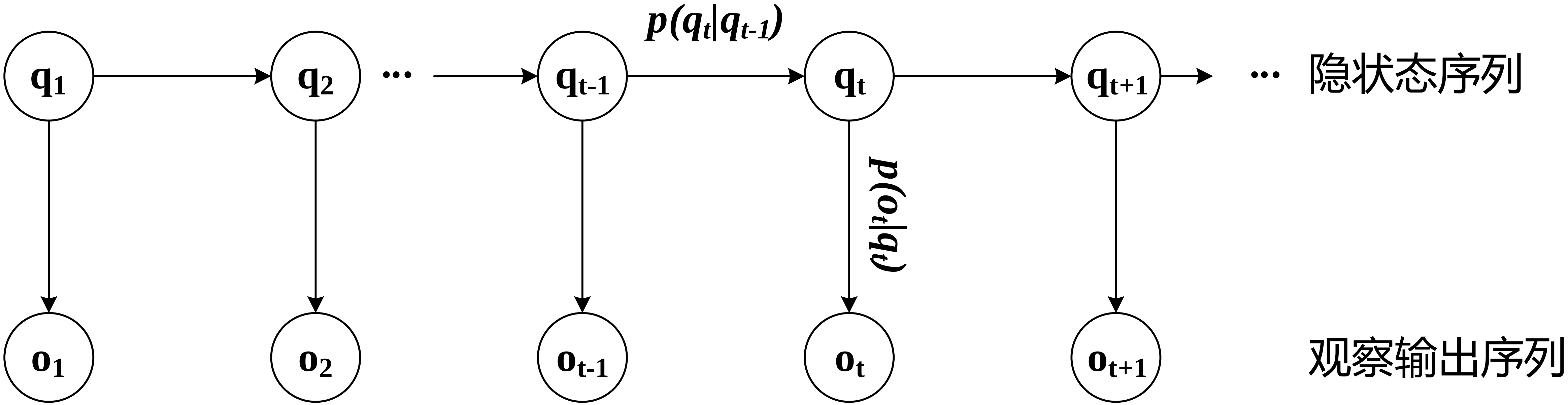
★ **描写**：该模型是一个双重随机过程，我们**不知道具体的状态序列**，**只知道状态转移的概率**，即模型的状态转换过程是不可观察的（隐蔽的），而可观察事件的随机过程是隐蔽状态转换过程的随机函数。

★ 例如：

- N 个袋子，每个袋子中有 M 种不同颜色的球。一实验员根据某一概率分布选择一个袋子，然后根据袋子中不同颜色球的概率分布随机取出一个球，并报告该球的颜色。
- 对局外人：可观察的过程是不同颜色球的序列，而袋子的序列是不可观察的。每只袋子对应 HMM 中的一个状态；球的颜色对应于 HMM 中状态的输出。



★ HMM图示



★ HMM的组成

1. 模型中的状态数为 N（袋子的数量）
2. 从每一个状态可能输出的不同的符号数 M（不同颜色球的数目）

6.2 隐马尔可夫模型

3. 状态转移概率矩阵 $\boldsymbol{A} = \boldsymbol{a}_{ij}$，$\boldsymbol{a}_{ij}$ 为实验员从一只袋子(状态 $\boldsymbol{S}_i$) 转向另一只袋子（状态 $\boldsymbol{S}_j$） 取球的概率。其中,

$$\begin{cases} a_{ij} = p\left(q_{t+1} = S_j \middle| q_t = S_i\right), \qquad 1 \le i,j \le N \\ a_{ij} \ge 0 \\ \sum_{j=1}^{N} a_{ij} = 1 \end{cases} \tag{6-6}$$

4. 从状态 $\boldsymbol{S_j}$ 观察到某一特定符号 $\boldsymbol{v_k}$ 的概率分布矩阵为：

$\boldsymbol{B = b_j(k)}$

其中，$\boldsymbol{b_j(k)}$ 为实验员从第 $\boldsymbol{j}$ 个袋子中取出第 $\boldsymbol{k}$ 种颜色的球的概率。那么，

$$\begin{cases} b_j(k) = p\left(O_t = v_k \middle| q_t = S_j\right), \qquad 1 \le j \le N, 1 \le k \le M \\ b_j(k) \ge 0 \\ \sum_{k=1}^{M} b_j(k) = 1 \end{cases} \tag{6-7}$$

5. 初始状态的概率分布为：$\boldsymbol{\pi} = \boldsymbol{\pi}_i$，其中，

$$\begin{cases} \pi_i = p(q_1 = S_i), & 1 \le i \le N \\ \pi_i \ge 0 \\ \sum_{i=1}^{N} \pi_i = 1 \end{cases} \tag{6-8}$$

★ 一般将 HMM 记为: $\boldsymbol{\mu} = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$ 或 $\boldsymbol{\mu} = (\boldsymbol{S}, \boldsymbol{O}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$用以指出模型的参数集合。

✦ $\boldsymbol{A}$，状态转移矩阵；　$\boldsymbol{B}$，输出矩阵

★ 给定HMM求观察序列

给定模型 $\mu=(A,B,\pi)$，产生观察序列 $O=O_1O_2\cdots O_T$：

(1) 令 $t=1$；

(2) 根据初始状态分布 $\pi=\pi_i$ 选择初始状态 $q_1=S_i$；

(3) 根据状态 S_i 的输出概率分布 $b_i(k)$，输出 $O_t=v_k$；

(4) 根据状态转移概率 a_{ij}，转移到新状态 $q_{t+1}=S_j$；

(5) $t=t+1$, 如果 $t<T$，重复步骤(3) (4), 否则结束。

6.2 隐马尔可夫模型

★ 三个问题：

(1) 估计问题： 在给定模型 $\mu = (A, B, \pi)$ 和观察序列$O = O_1 O_2 \cdots O_T$ 的情况下，怎样快速计算概率 $p(O|\mu)$?

(2) 序列问题： 在给定模型 $\mu = (A, B, \pi)$ 和观察序列$O = O_1 O_2 \cdots O_T$ 的情况下，如何选择在一定意义下“最优”的状态序列 $Q = q_1 q_2 \cdots q_T$，使得该状态序列“最好地解释”观察序列?

(3) 训练问题或参数估计问题： 给定一个观察序列 $O = O_1 O_2 \cdots O_T$ ，如何根据最大似然估计来求模型的参数值？即如何调节模型的参数，使得$p(O|\mu)$ 最大?

6.3 前向算法

6.3 前向算法

★ **问题1：给定模型 $\mu=(A,B,\pi)$ 和观察序列$O=O_1O_2\ldots O_T$，快速计算观察序列概率 $p(O|\mu)$**

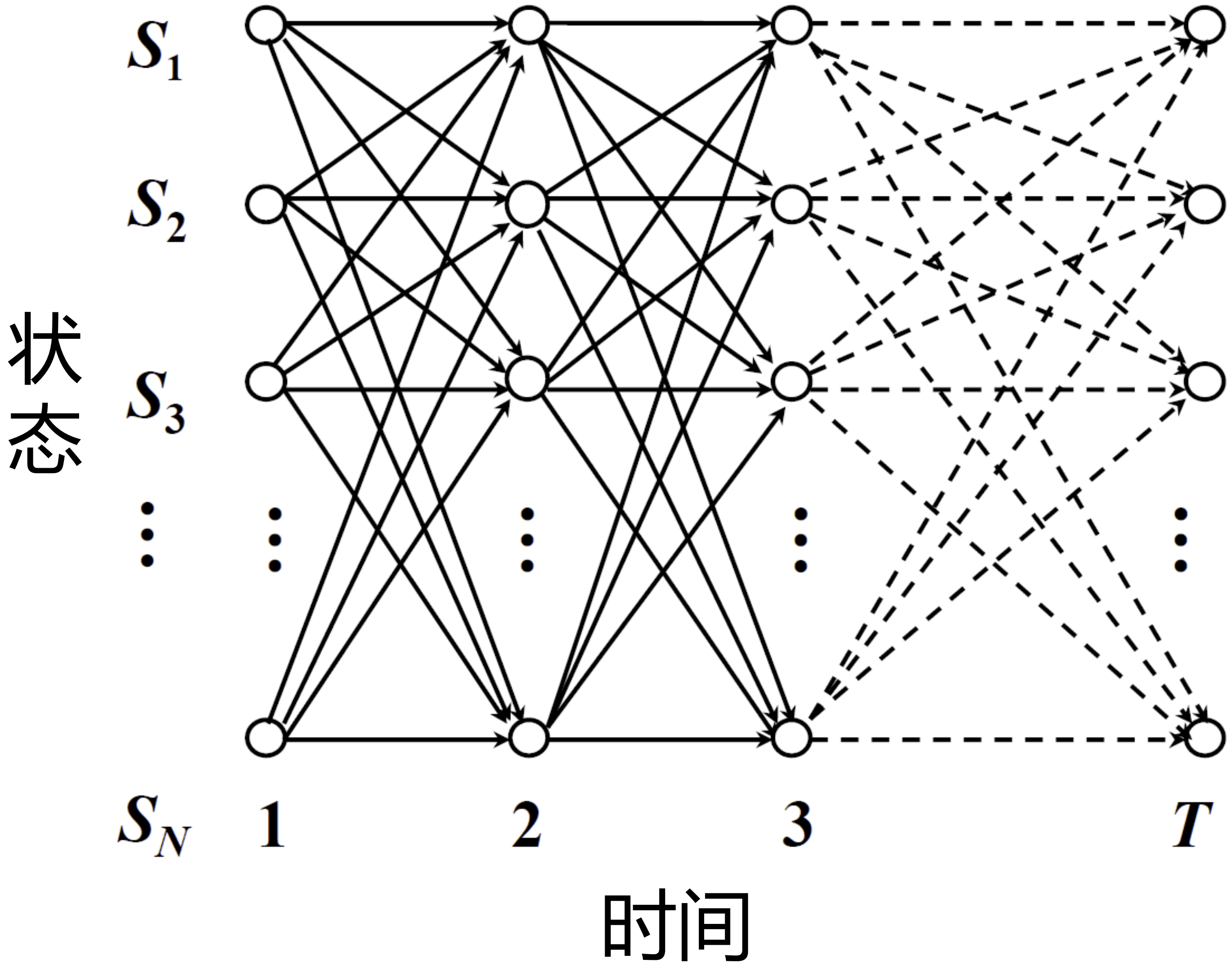
$$p(O|\mu)=\sum_{Q}p(O,Q|\mu)=\sum_{Q}p(Q|\mu)\times p(O|Q,\mu) \qquad (6\text{-}9)$$

$$p(Q|\mu)=\pi_{q_1}\times a_{q_1q_2}\times a_{q_2q_3}\times\cdots\times a_{q_{T-1}q_T} \qquad (6\text{-}10)$$

$$p(O|Q,\mu)=b_{q_1}(O_1)\times b_{q_2}(O_2)\times\cdots\times b_{q_T}(O_T) \qquad (6\text{-}11)$$

$$\boxed{p(O|\mu)=\sum_{Q}\pi_{q_i}b_{q_i}(O_1)\prod_{t=1}^{T-1}a_{q_t,q_{t+1}}b_{q_{t+1}}(O_{t+1})}$$

6.3 前向算法



困难：

如果模型 μ 有 N 个不同的状态，时间长度为 T，那么有 N^T 个可能的状态序列，搜索路径成指数级组合爆炸。

6.3 前向算法

★ **解决办法：动态规划**

★ **前向算法(The forward procedure)**

★ **基本思想：前向变量 $\boldsymbol{\alpha_t(i)}$**

★ **定义：** 前向变量 $\alpha_t(i)$ 是在时间 t，HMM输出了序列 $O_1O_2\cdots O_t$，并且位于状态 S_i 的概率

$$\alpha_t(i) = p(O_1O_2\cdots O_t, q_t = S_i|\mu) \tag{6-12}$$

如果可以高效地计算 $\alpha_t(i)$，就可以高效地求得 $p(O|\mu)$。

6.3 前向算法

✦ 因为$p(O|\mu)$ 是在到达状态 q_T 时观察到序列$O = O_1O_2 \cdots O_T$的概率（所有可能的概率之和）：

$$p(O|\mu) = \sum_{S_i} p(O_1O_2 \cdots O_T, q_T = S_i|\mu) = \sum_{i=1}^{N} \alpha_t(i) \qquad (6\text{-}13)$$

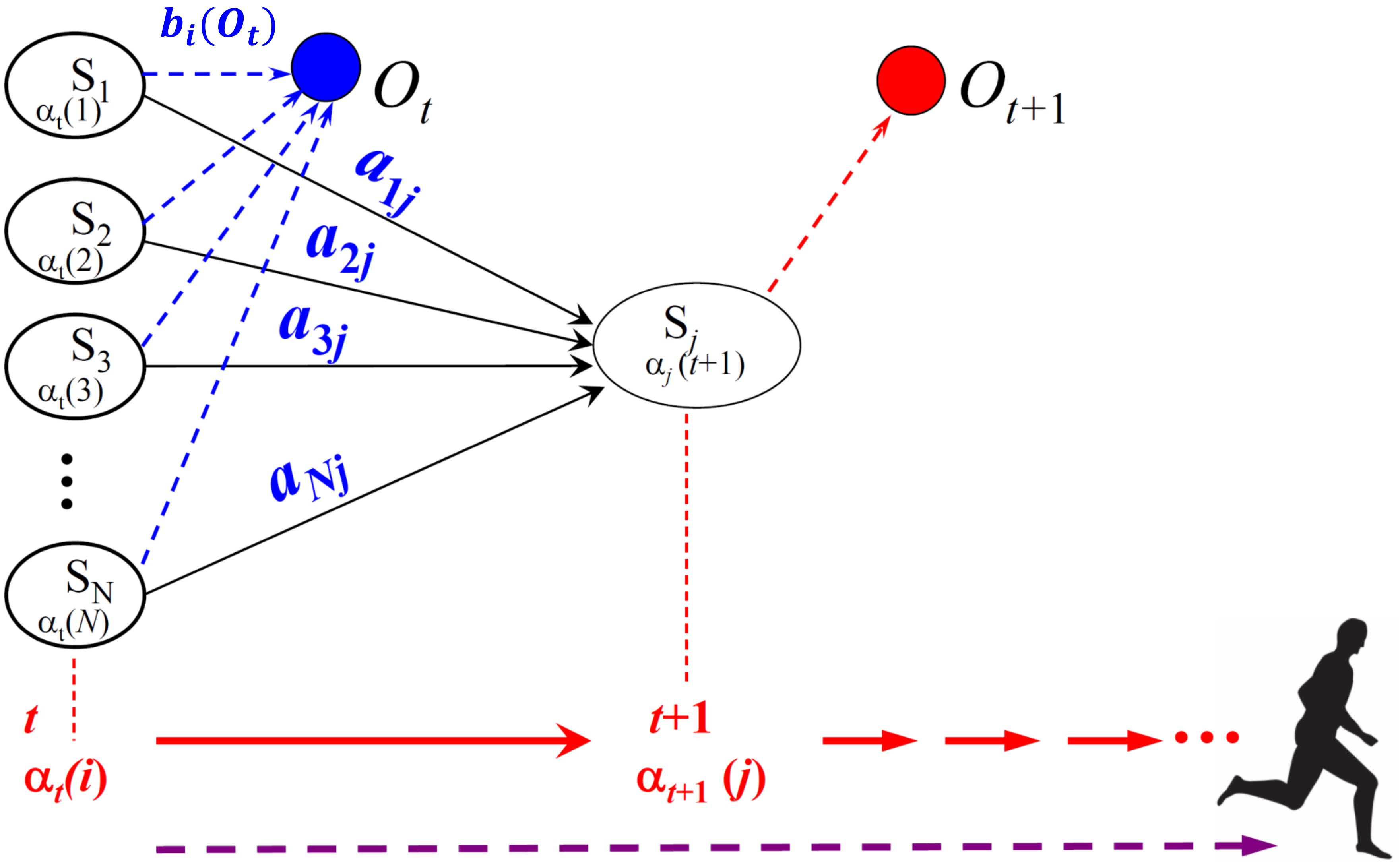
✦ **动态规划计算 $\alpha_t(i)$**： 在时间 $t+1$ 的前向变量可以根据时间 t 的前向变量的值递推计算 $\alpha_t(1), \cdots, \alpha_t(N)$的值递推计算：

$$\boxed{\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i)a_{ij}\right] \times b_j(O_{t+1})} \qquad (6\text{-}14)$$

6.3 前向算法

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] \times b_j(O_{t+1}) \tag{6-14}$$

★ 算法图解



★ 算法6.1：前向算法描述

(1) 初始化：

$$\alpha_1(i) = \pi_i b_i(O_1), \qquad 1 \le i \le N$$

(2) 循环计算：

$$a_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] \times b_j(O_{t+1}), \qquad 1 \le t \le T-1$$

(3) 结束，输出：

$$p(O|\mu) = \sum_{i=1}^{N} \alpha_T(i)$$

★ 算法的时间复杂性：

每计算一个 $\alpha_t(i)$ 必须考虑从 $t-1$ 时的所有 N 个状态转移到状态 S_i 的可能性，时间复杂性为 $O(N)$，对应每个时刻 t，要计算 N 个前向变量： $\alpha_t(1), \alpha_t(2), \cdots, \alpha_t(N)$，所以，时间复杂性为：$O(N) \times N = O(N^2)$。又因 $t = 1, 2, \cdots, T$，所以前向算法总的复杂性为：$O(N^2T)$。

★ 前向算法例题

★ 有3个盒子，每个盒子都有红色和白色两种球，分别为：

- ✦ 盒子1：5红5白
- ✦ 盒子2：4红6白
- ✦ 盒子3：7红3白

$$\pi = \begin{bmatrix} 0.2 \\ 0.4 \\ 0.4 \end{bmatrix} \quad A = \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{bmatrix} \quad B = \begin{bmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \\ 0.7 & 0.3 \end{bmatrix}$$

- ✦ 观察序列： $\{O_1, O_2, O_3\} = \{红, 白, 红\}$
- ✦ 观察符号集合： $O = \{红, 白\}$
- ✦ 状态集合： $S = \{盒子1, 盒子2, 盒子3\}$

6.3 前向算法

$$\pi = \begin{bmatrix} 0.2 \\ 0.4 \\ 0.4 \end{bmatrix} \quad A = \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{bmatrix} \quad B = \begin{bmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \\ 0.7 & 0.3 \end{bmatrix}$$

✦ 时刻1：(红色球, 盒子1) $\alpha_1(1) = \pi_1 b_1(o_1) = 0.2 \times 0.5 = 0.1$

(红色球, 盒子2) $\alpha_1(2) = \pi_2 b_2(o_1) = 0.4 \times 0.4 = 0.16$

(红色球, 盒子3) $\alpha_1(3) = \pi_3 b_3(o_1) = 0.4 \times 0.7 = 0.28$

✦ 时刻2：(白色球, 盒子1)

$$a_2(1) = \left[\sum_{i=1}^{3} a_1(i) a_{i1}\right] b_1(O_2) = [0.1 * 0.5 + 0.16 * 0.3 + 0.28 * 0.2] \times 0.5 = 0.077$$

(白色球, 盒子2)

$$a_2(2) = \left[\sum_{i=1}^{3} a_1(i) a_{i2}\right] b_2(O_2) = [0.1 * 0.2 + 0.16 * 0.5 + 0.28 * 0.3] \times 0.6 = 0.1104$$

(白色球, 盒子3)

$$a_2(3) = \left[\sum_{i=1}^{3} a_1(i) a_{i3}\right] b_3(O_2) = [0.1 * 0.3 + 0.16 * 0.2 + 0.28 * 0.5] \times 0.3 = 0.0606$$

6.3 前向算法

$$\pi = \begin{bmatrix} 0.2 \\ 0.4 \\ 0.4 \end{bmatrix} \quad A = \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{bmatrix} \quad B = \begin{bmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \\ 0.7 & 0.3 \end{bmatrix}$$

✦ 时刻3：(红色球, 盒子1)

$$a_3(1) = \left[\sum_{i=1}^{3} a_2(i)a_{i1}\right] b_1(o_3) = [0.077 * 0.5 + 0.1104 * 0.3 + 0.0606 * 0.2] \times 0.5 = 0.04187$$

(红色球, 盒子2)

$$a_3(2) = \left[\sum_{i=1}^{3} a_2(i)a_{i2}\right] b_2(o_3) = [0.077 * 0.2 + 0.1104 * 0.5 + 0.0606 * 0.3] \times 0.4 = 0.035512$$

(红色球, 盒子3)

$$a_3(3) = \left[\sum_{i=1}^{3} a_3(i)a_{i3}\right] b_3(o_3) = [0.077 * 0.3 + 0.1104 * 0.2 + 0.0606 * 0.5] \times 0.7 = 0.052836$$

✦ 观察序列概率：$P(O|\lambda) = \sum_{i=1}^{3} a_3(i) = 0.130218$

6.4 后向算法

★ **后向算法(The backward procedure)**

★ 定义后向变量 $\beta_t(i)$ 是在给定了模型 $\mu=(A,B,\pi)$ 和假定在时间 t 、状态 S_i 的条件下，模型输出观察序列 $O_{t+1}O_{t+2}\cdots O_T$的概率：

$$\beta_t(i)=p(O_{t+1}O_{t+2}\cdots O_T|q_t=S_i,\mu) \qquad (6\text{-}15)$$

6.4 后向算法

★ 与前向变量一样，运用**动态规划**计算后向量：

第一步，从时刻 t 到 $t+1$，模型由状态 S_i 转移到状态 S_j，并从 S_j 输出 O_{t+1}；

第二步，在时间 $t+1$，状态为 S_j 的条件下，模型输出观察序列 $O_{t+2}O_{t+3}\cdots O_T$。

6.4 后向算法

★ 第一步输出O_{t+1}的概率： $\sum_{j=1}^{N} a_{ij} b_j(O_{t+1})$

★ 第二步输出 $O_{t+2} \cdots O_T$ 的概率按后向变量的定义为：$\beta_{t+1}(j)$

于是，有归纳关系：

$\beta_t(i) = p(O_{t+1}O_{t+2} \cdots O_T | q_t = S_i, \mu)$ $\sum_{j=1}^{N} a_{ij} b_j = p(O_{t+1} | q_t = S_i, q_{t+1} = S_j, \mu)$ $\beta_{t+1}(j) = p(O_{t+2}O_{t+3} \cdots O_T | q_{t+1} = S_j, \mu)$

$$\boxed{\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)} \qquad (6\text{-}16)$$

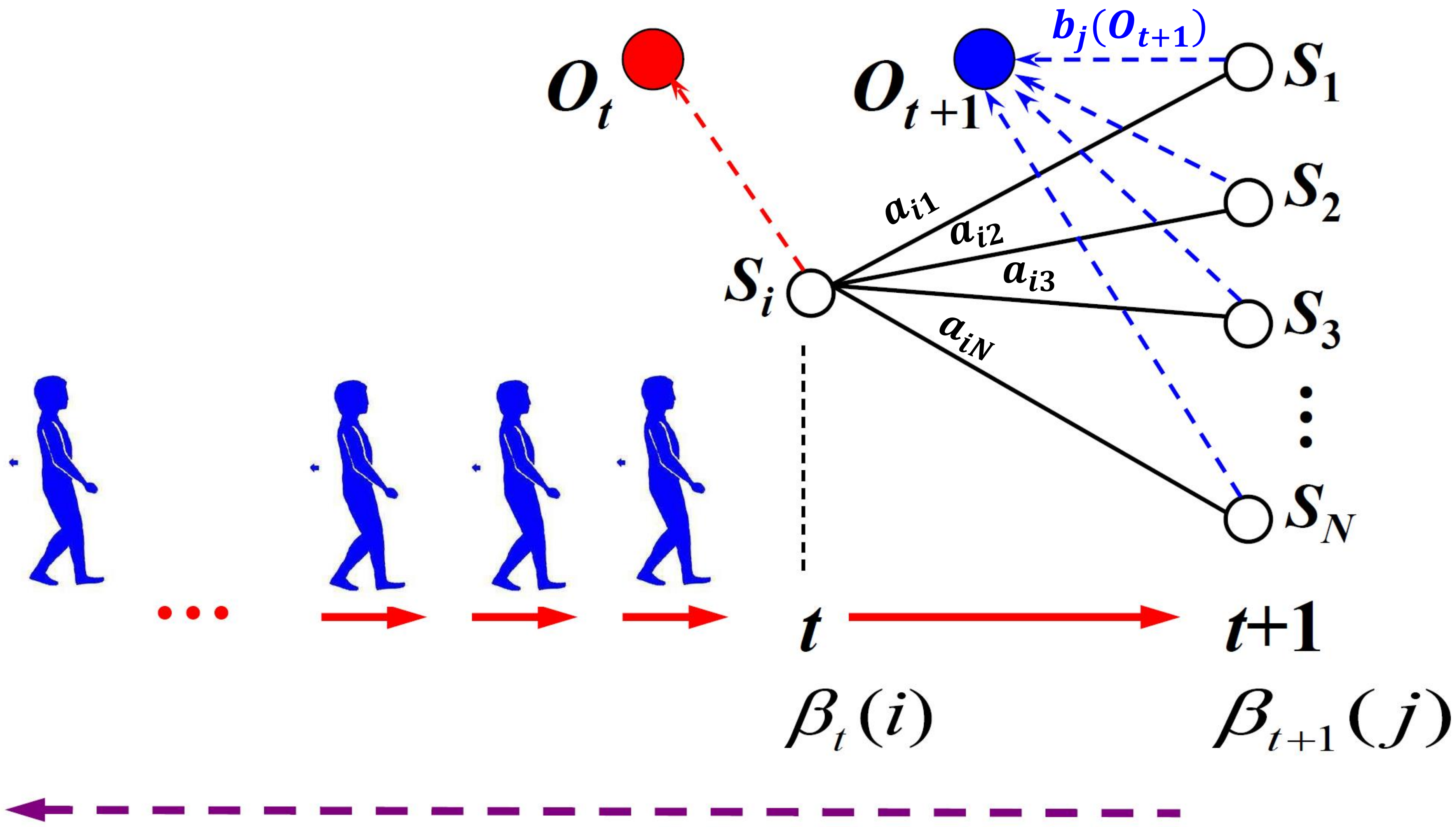
★ 归纳顺序：$\beta_T(x), \beta_{T-1}(x), \ldots, \beta_1(x)$ （x 为HMM模型的状态)

$$\boxed{\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] \times b_j(O_{t+1})} \qquad (6\text{-}14)$$

6.4 后向算法

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j) \tag{6-16}$$

★ 算法图解：



6.4 后向算法

★ **算法6.2：后向算法描述**

(1) 初始化：$\beta_T(i) = 1, \ 1 \le i \le N$

(2) 循环计算：

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j), \qquad 1 \le t \le T-1, 1 \le i \le N$$

(3) 结束，输出：$p(O|\mu) = \sum_{i=1}^{N} \pi_i \times b_i(O_1) \times \beta_1(i)$

★ 算法的时间复杂度：$O(N^2T)$

6.5 维特比(Viterbi)算法

6.5 维特比(Viterbi)算法

★ **问题2：如何发现“最优”状态序列，能够“最好地解释”观察序列**

★ 解释不是唯一的，关键在于如何理解“最优” 的状态序列?

★ **一种解释是**：状态序列中的每个状态都单独地具有概率，对于每个时刻 t $(1 \leq t \leq T)$，寻找 q_t 使该状态序列中每一个状态都单独地具有最大概率，即

$$\gamma_t(i) = p(q_t = S_i | O, \mu) \text{ 最大。}$$

6.5 维特比(Viterbi)算法

模型的输出序列 $\boldsymbol{O}$，并且
在时间 $\boldsymbol{t}$ 到达状态 $\boldsymbol{i}$ 的概率。

$$\gamma_t(i) = p(q_t = S_i | O, \mu) = \frac{p(q_t = S_i, O | \mu)}{p(O | \mu)} \tag{6-17}$$

★ 分解过程：

(1) 模型在时间 t 到达状态 S_i，并且输出 $O = O_1O_2 \cdots O_t$。根据前向变量的定义，实现这一步的概率为 $\alpha_t(i)$。

(2) 从时间 t，状态 S_i 出发，模型输出$O = O_{t+1}O_{t+2} \cdots O_T$，根据后向变量定义，实现这一步的概率为 $\beta_t(i)$ 。于是：

$$p(q_t = S_i, O|\mu) = \alpha_t(i) \times \beta_t(i) \qquad (6\text{-}18)$$

6.5 维特比(Viterbi)算法

★ 而 $p(O|\mu)$ 与时间 t 的状态无关，因此：

$$p(O|\mu) = \sum_{t=1}^{N} \alpha_t(i) \times \beta_t(i) \tag{6-19}$$

★ 将公式(6.18)和(6.19)带入(6.17)式得：

$$\gamma_t(i) = \frac{\alpha_t(i) \times \beta_t(i)}{\sum_{t=1}^{N} \alpha_t(i) \times \beta_t(i)} \tag{6-20}$$

★ ***t* 时刻的最优状态为**： $\hat{q}_t = \underset{1 \le i \le N}{\mathrm{argmax}}(\gamma_t(i))$

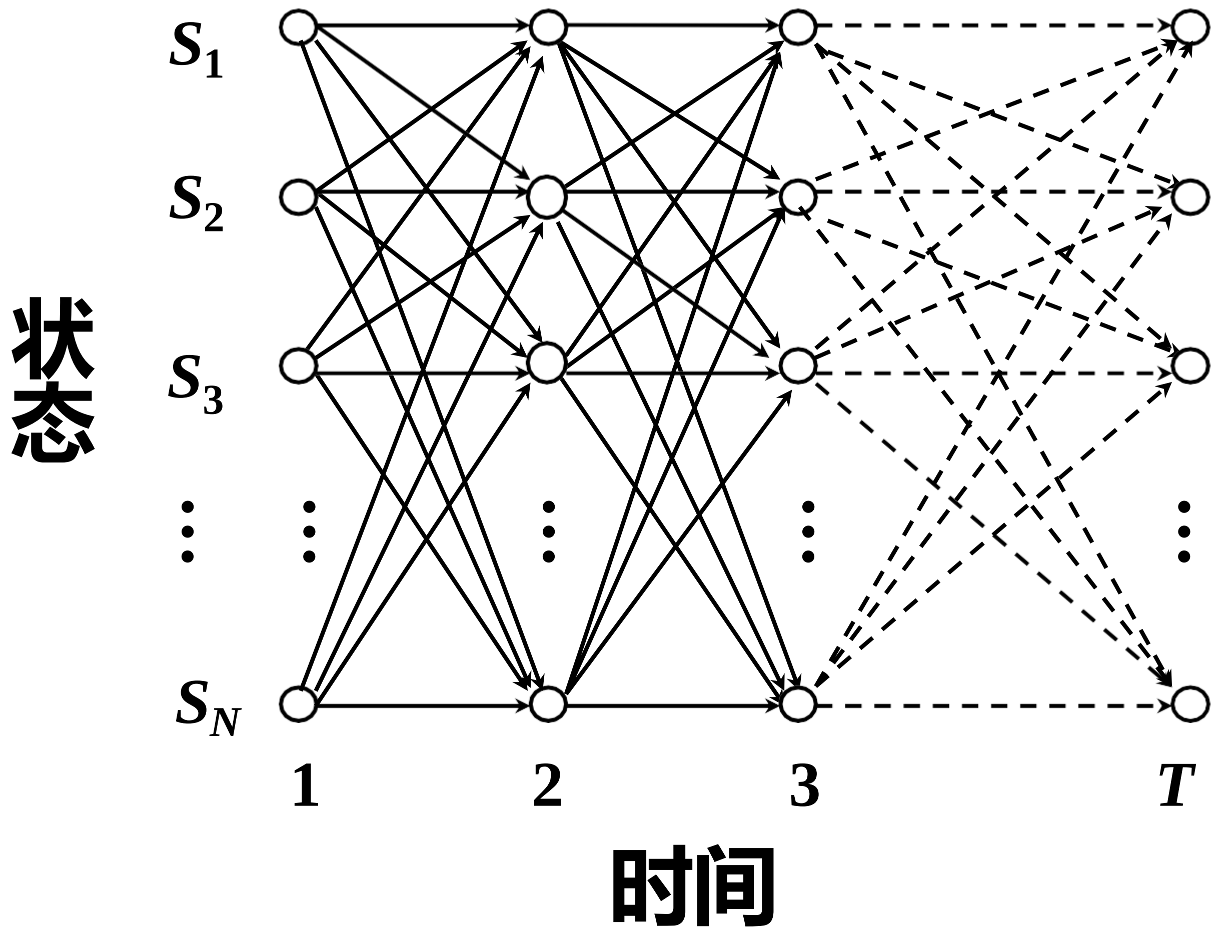
6.5 维特比(Viterbi)算法

★ **问题：**

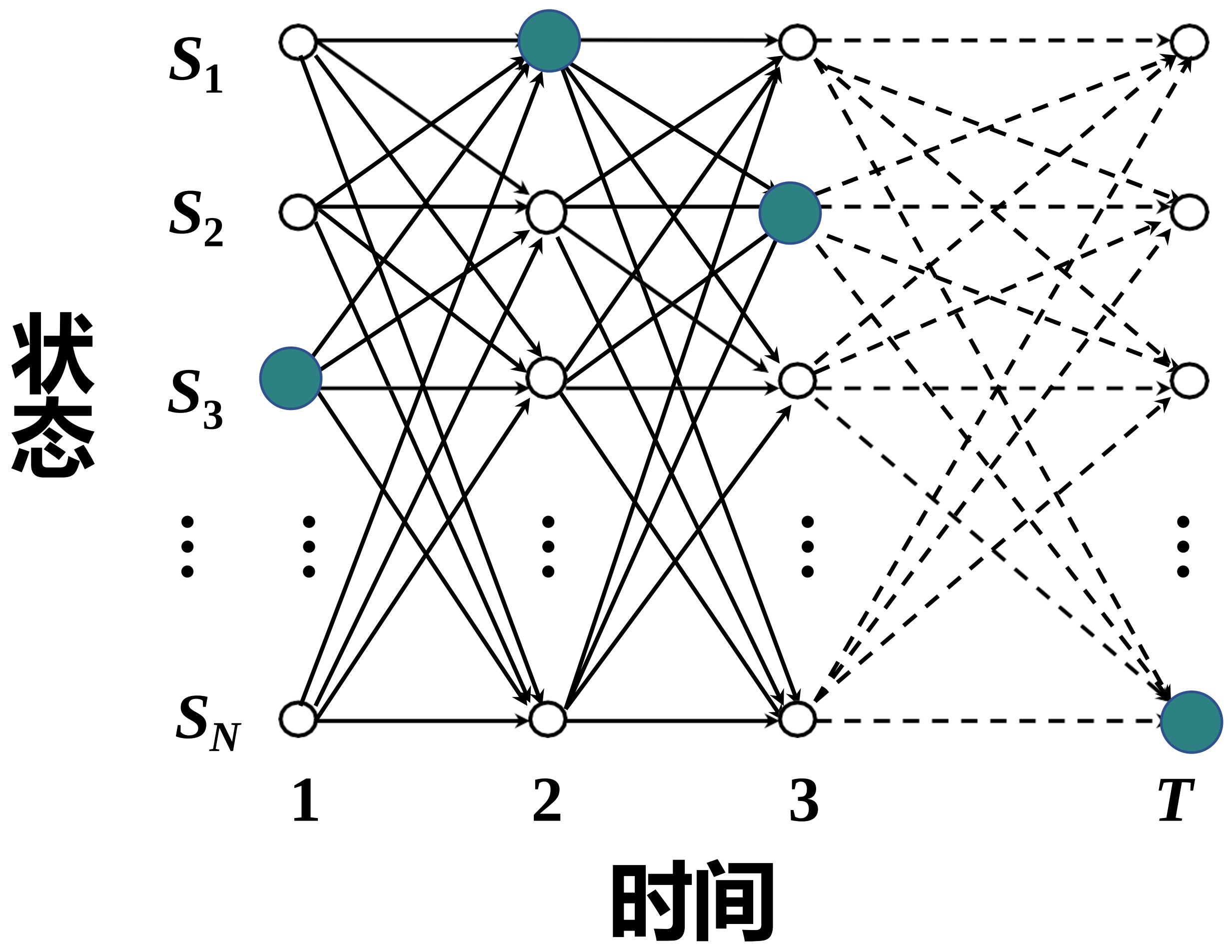
★ **每一个状态单独最优不一定使整体的状态序列最优**，可能两个最优的状态 $\hat{q}_t$ 和 $\hat{q}_{t+1}$ 之间的转移概率为0，即

$$\alpha_{\hat{q}_t \hat{q}_{t+1}} = 0$$

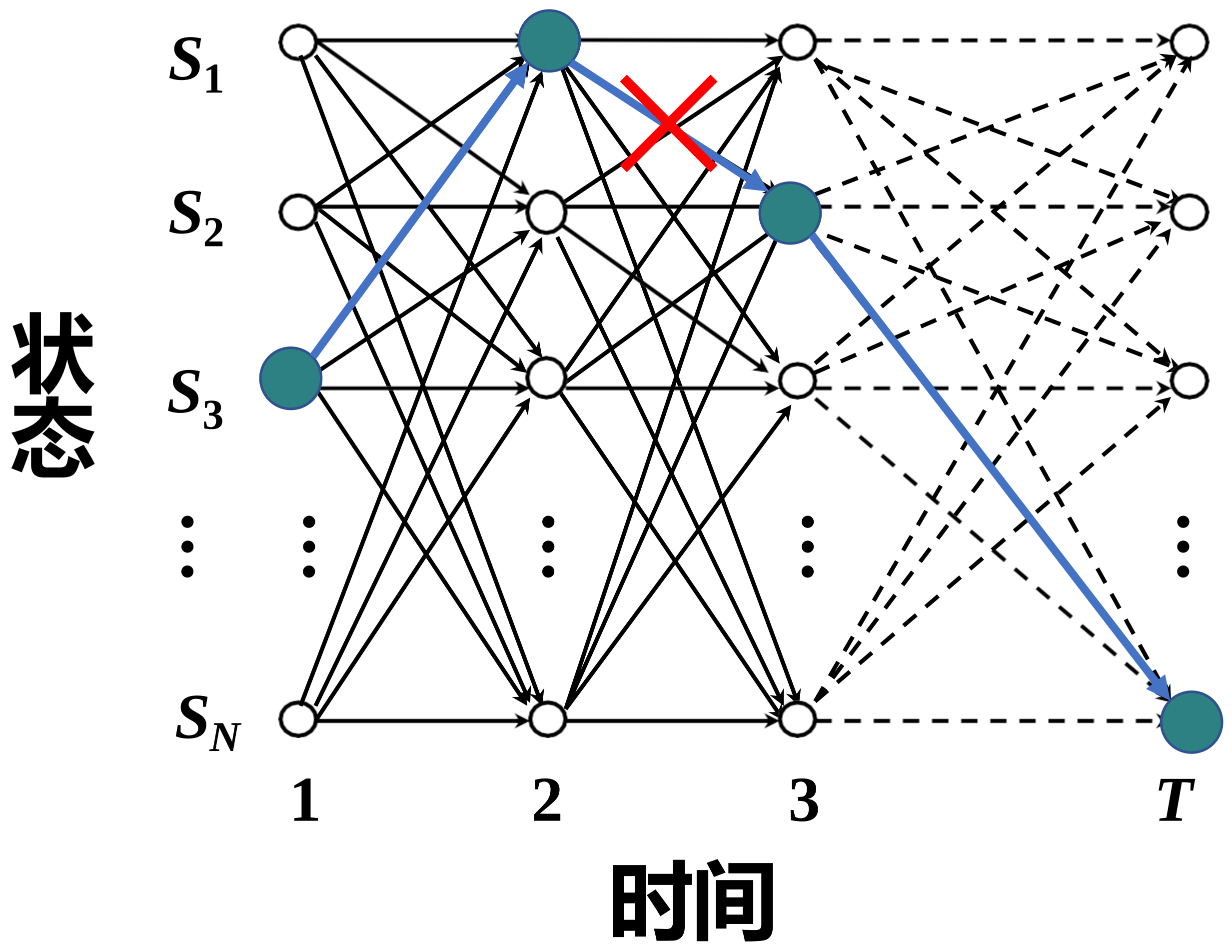
6.5 维特比(Viterbi)算法



6.5 维特比(Viterbi)算法



6.5 维特比(Viterbi)算法



这种情况下，所谓的“最优状态”序列不是合法的序列

6.5 维特比(Viterbi)算法

★ **另一种解释是**：在给定模型 μ 和观察序列 O 的条件下求概率最大的状态序列：

$$\hat{Q} = \underset{Q}{\operatorname{argmax}}\, p(Q|O,\mu) \qquad (6\text{-}21)$$

★ 这种解释避免了前一种理解引起的“断点”问题，根据这种理解，优化的不是状态序列中的单个状态，而实整个状态序列，不合法的序列概率为0，因此，不可能被选为最优状态序列。

6.5 维特比(Viterbi)算法

★ **Viterbi 算法：动态搜索最优状态序列。**

★ 安德鲁·维特比（Andrew Viterbi）

✦ 1935.3.9

✦ 1967年提出维特比算法，但没有申请专利

✦ 与厄文·雅各布创立了高通公司

★ **定义：Viterbi 变量**是在时间 t 时，模型沿着某一条路径到达 S_i，并输出观察序列 $O = O_1 O_2 \cdots O_t$ 的最大概率：

$$\delta_t(i) = \max_{q_1, q_2, \cdots, q_{t-1}} p(q_1, q_2, \cdots, q_t = S_i, O_1 O_2 \cdots O_t | \mu) \quad \text{(6-22)}$$

★ 递推计算：$\delta_{t+1}(j) = \max_j \left[\delta_t(i) \times a_{ij}\right] \times b_j(O_{t+1})$ (6-23)

6.5 维特比(Viterbi)算法

★ **算法6.3：Viterbi 算法**

(1) 初始化：$\delta_1(i)=\pi_i b_i(O_1)$，$1\le i\le N$

路径记忆变量$\psi_t(i)$，记录该路径上状态s_i的前一个状态（$t-1$时刻的状态）概率最大的路径变量：$\psi_1(i)=0$

(2) 递推计算：

$$\delta_t(j)=\max_{1\le i\le N}[\delta_{t-1}(i)\cdot a_{ij}]\cdot b_j(O_t)，\quad 2\le t\le T,1\le j\le N$$

$$\psi_t(j)=\underset{1\le i\le N}{\operatorname{argmax}}[\delta_{t-1}(i)\cdot a_{ij}]\cdot b_j(O_t)，\quad 2\le t\le T,1\le i\le N$$

6.5 维特比(Viterbi)算法

(3) 结束，输出：

$$\hat{Q}_T = \underset{1\le i\le N}{\operatorname{argmax}}[\delta_T(i)], \qquad \hat{p}\left(\hat{Q}_T\right) = \max_{1\le i\le N} \delta_T(i)$$

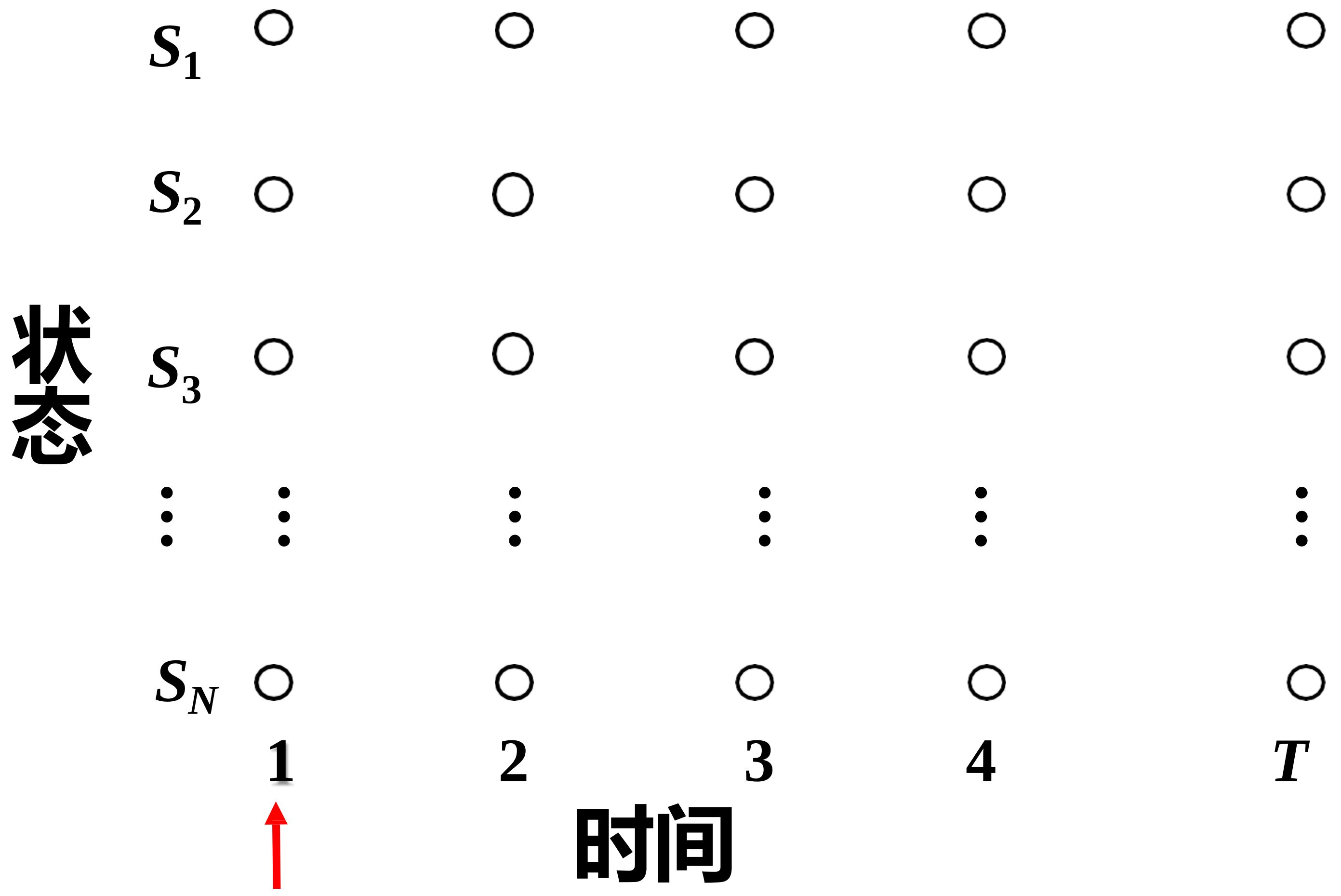
(4) 通过回溯得到路径（状态序列）：

$$\hat{q}_t = \psi_{t+1}(\hat{q}_{t+1}), \qquad t = T-1, T-2, \cdots, 1$$

★ 算法的时间复杂度：$O(N^2T)$

6.5 维特比(Viterbi)算法

★ 图解 Viterbi 搜索过程：

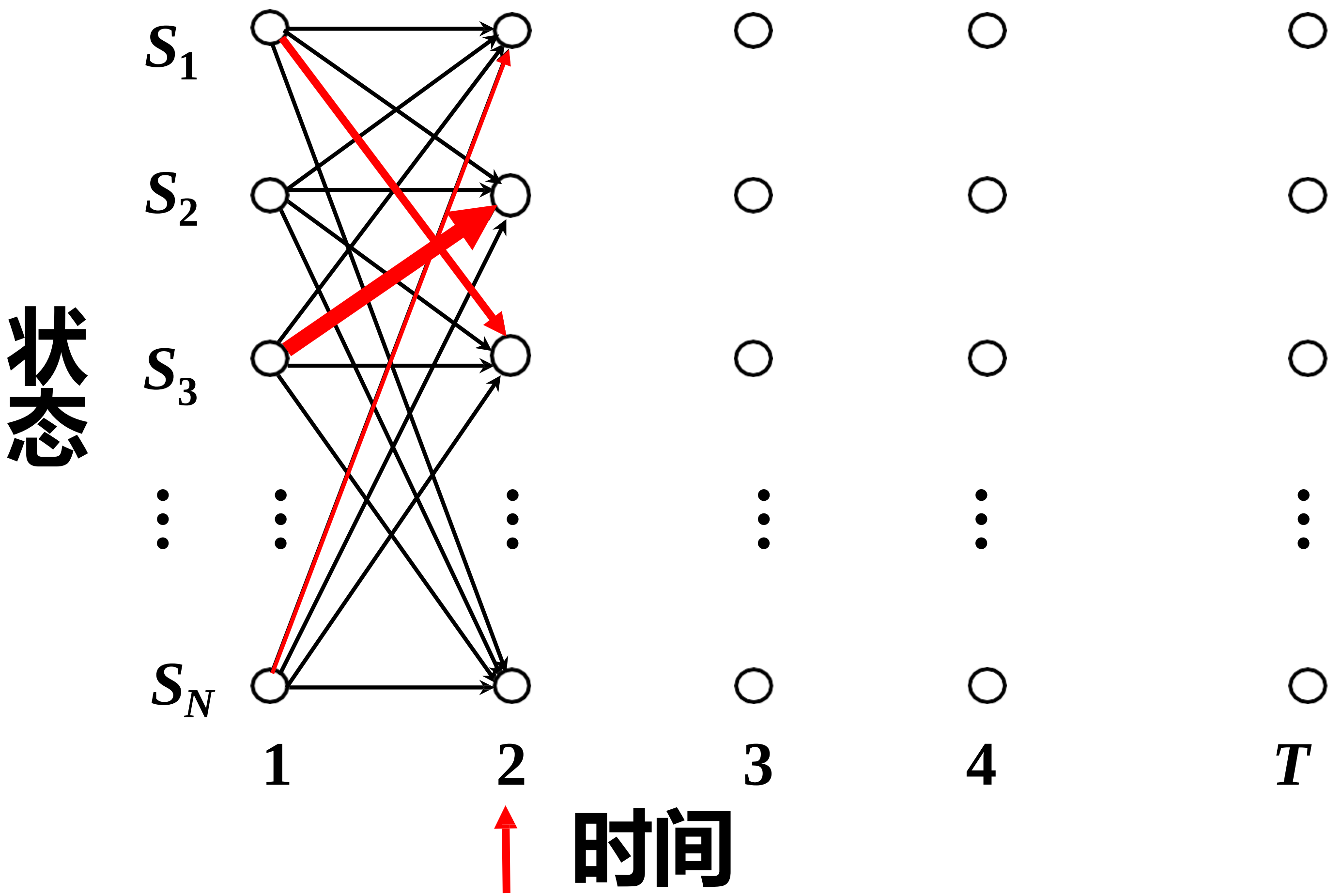


剪枝策略：

①$\delta_t(j) \geq \Delta$

②$N_{Path} \leq \sigma$

★ 图解 Viterbi 搜索过程：



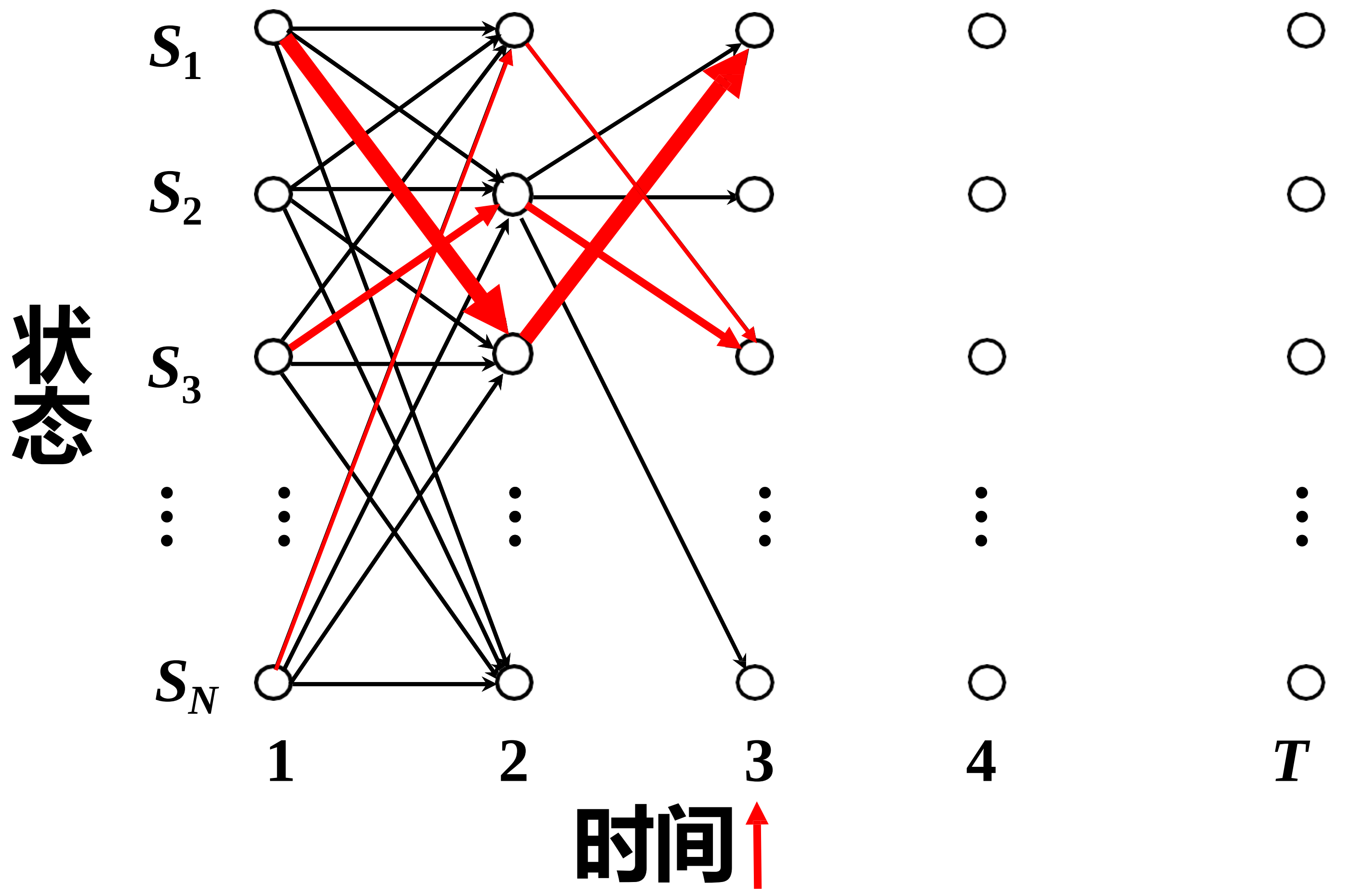
剪枝策略：

①$\delta_t(j) \geq \Delta$

②$N_{Path} \leq \sigma$

6.5 维特比(Viterbi)算法

★ **图解 Viterbi 搜索过程：**

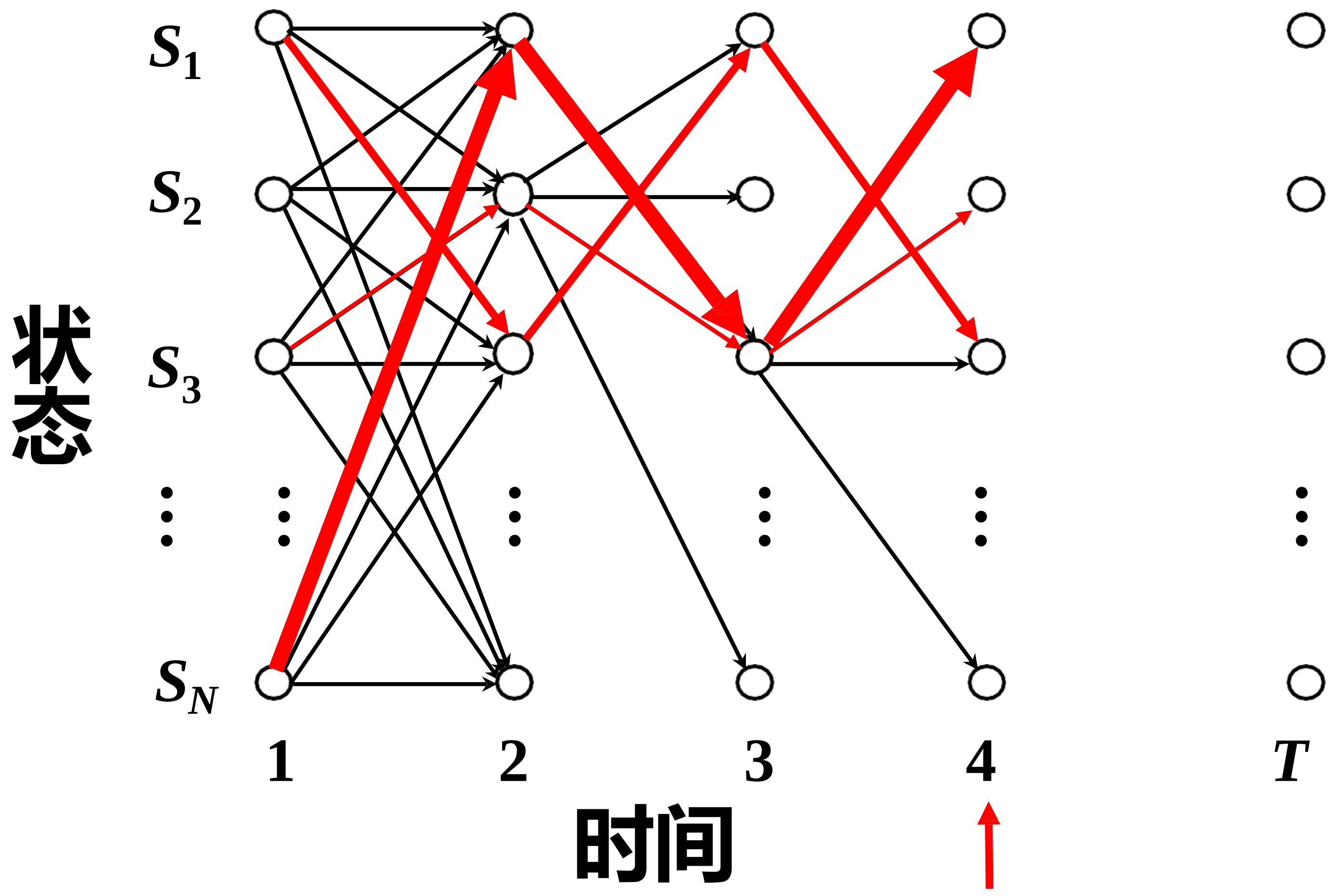


剪枝策略：

①$\delta_t(j) \geq \Delta$

②$N_{Path} \leq \sigma$

★ 图解 Viterbi 搜索过程：

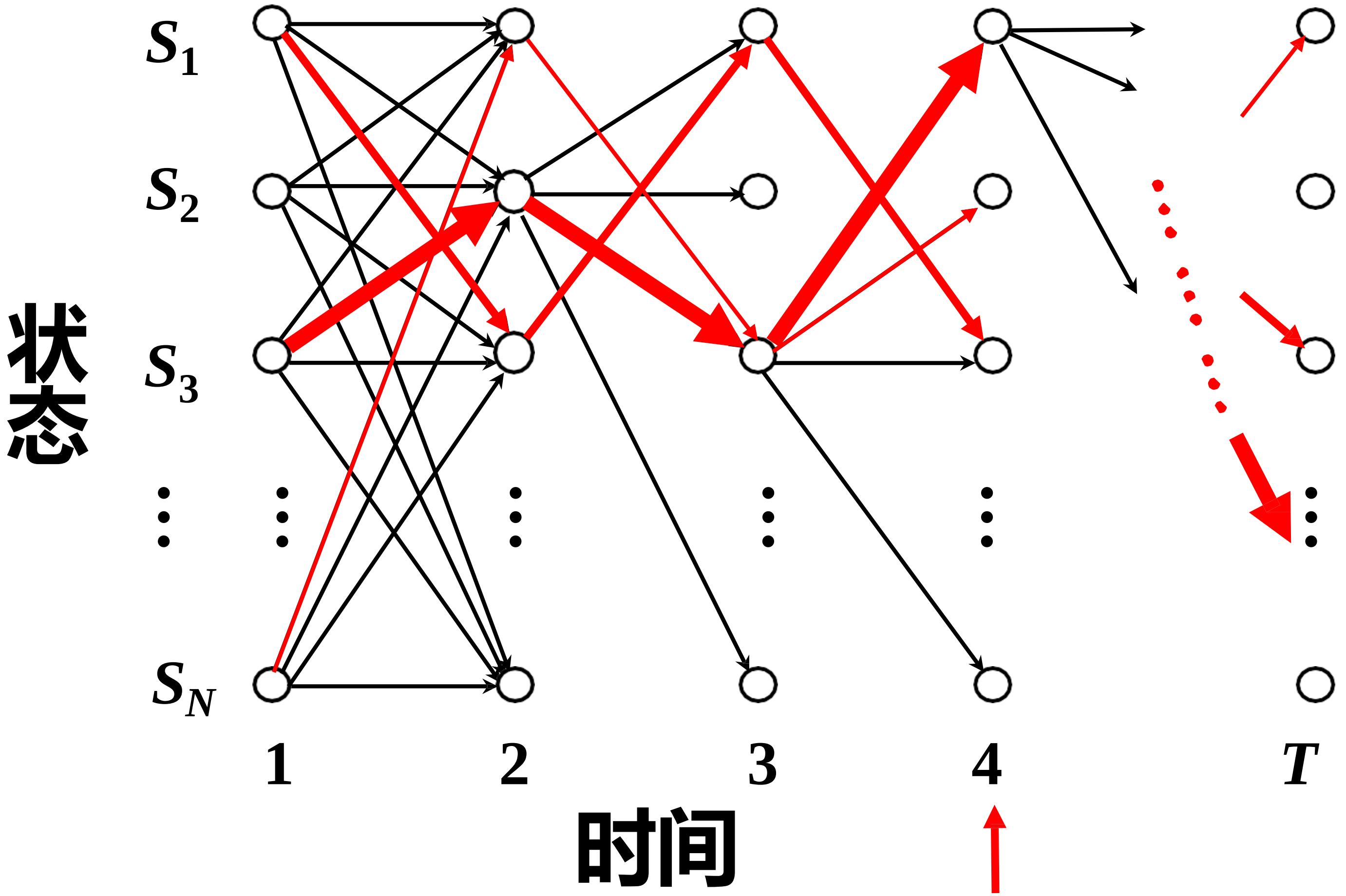


剪枝策略：

①$\delta_t(j) \geq \Delta$

②$N_{Path} \leq \sigma$

★ 图解 Viterbi 搜索过程：



剪枝策略：

①$\delta_t(j) \geq \Delta$

②$N_{Path} \leq \sigma$

★ 维特比(Viterbi)算法例题

★ 观察序列：$O = \{红, 白, 红\}$

$$\pi = \begin{bmatrix} 0.2 \\ 0.4 \\ 0.4 \end{bmatrix} \quad A = \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{bmatrix} \quad B = \begin{bmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \\ 0.7 & 0.3 \end{bmatrix}$$

6.5 维特比(Viterbi)算法

$$\pi = \begin{bmatrix} 0.2 \\ 0.4 \\ 0.4 \end{bmatrix} \quad A = \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{bmatrix} \quad B = \begin{bmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \\ 0.7 & 0.3 \end{bmatrix}$$

✦ $t = 1$时刻：

$$\delta_1(1) = \pi_1 b_1(O_1) = 0.2 \times 0.5 = 0.1$$

$$\delta_1(2) = \pi_2 b_2(O_1) = 0.4 \times 0.4 = 0.16$$

$$\delta_1(3) = \pi_3 b_3(O_1) = 0.4 \times 0.7 = 0.28$$

$$\psi_1(1) = \psi_1(2) = \psi_1(3) = 0$$

$$\delta_t(j) = \max_{1 \le i \le N}[\delta_{t-1}(i) \cdot a_{ij}] \cdot b_j(O_t), \quad 2 \le t \le T, 1 \le j \le N$$

$$\psi_t(j) = \underset{1 \le i \le N}{\operatorname{argmax}}[\delta_{t-1}(i) \cdot a_{ij}] \cdot b_j(O_t), \quad 2 \le t \le T, 1 \le i \le N$$

6.5 维特比(Viterbi)算法

$$\pi = \begin{bmatrix} 0.2 \\ 0.4 \\ 0.4 \end{bmatrix} \quad A = \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{bmatrix} \quad B = \begin{bmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \\ 0.7 & 0.3 \end{bmatrix}$$

$$\delta_1(1) = \pi_1 b_1(O_1) = 0.2\times0.5 = 0.1$$
$$\delta_1(2) = \pi_2 b_2(O_1) = 0.4\times0.4 = 0.16$$
$$\delta_1(3) = \pi_3 b_3(O_1) = 0.4\times0.7 = 0.28$$

✦ $t = 2$时刻:

$$\delta_2(1) = \max_{1\le j\le 3}[\delta_1(i)a_{i1}]b_1(O_2) = \max_{1\le j\le 3}[0.1\times0.5, 0.16\times0.3, \mathbf{0.28}\times\mathbf{0.2}]\times0.5 = 0.028$$

$$\psi_2(1) = 3$$

$$\delta_2(2) = \max_{1\le j\le 3}[\delta_1(i)a_{i2}]b_2(O_2) = \max_{1\le j\le 3}[0.1\times0.2, 0.16\times0.5, \mathbf{0.28}\times\mathbf{0.3}]\times0.6$$
$$= 0.0504$$

$$\psi_2(2) = 3$$

$$\delta_2(3) = \max_{1\le j\le 3}[\delta_1(i)a_{i3}]b_3(O_2) = \max_{1\le j\le 3}[0.1\times0.3, 0.16\times0.2, \mathbf{0.28}\times\mathbf{0.5}]\times0.3 = 0.042$$

$$\psi_2(3) = 3$$

$$\delta_t(j) = \max_{1\le i\le N}[\delta_{t-1}(i)\cdot a_{ij}]\cdot b_j(O_t), \quad 2\le t\le T, 1\le j\le N$$

$$\psi_t(j) = \operatorname*{argmax}_{1\le i\le N}[\delta_{t-1}(i)\cdot a_{ij}]\cdot b_j(O_t), \quad 2\le t\le T, 1\le i\le N$$

6.5 维特比(Viterbi)算法

$$\pi = \begin{bmatrix} 0.2 \\ 0.4 \\ 0.4 \end{bmatrix} \quad A = \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{bmatrix} \quad B = \begin{bmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \\ 0.7 & 0.3 \end{bmatrix}$$

$\delta_2(1) = 0.028$
$\delta_2(2) = 0.0504$
$\delta_2(3) = 0.042$

✦ $t = 3$时刻：

$$\delta_3(1) = \max_{1\le j\le 3}[\delta_2(i)a_{i1}]b_1(O_3) = \max_{1\le j\le 3}[0.028\times 0.5, \mathbf{0.0504\times 0.3}, 0.042\times 0.2]\times 0.5$$
$$= 0.00756$$

$$\psi_3(1) = 2$$

$$\delta_3(2) = \max_{1\le j\le 3}[\delta_2(i)a_{i2}]b_2(O_3) = \max_{1\le j\le 3}[0.028\times 0.2, \mathbf{0.0504\times 0.5}, 0.042\times 0.3]\times 0.4$$
$$= 0.01008$$

$$\psi_3(2) = 2$$

$$\delta_3(3) = \max_{1\le j\le 3}[\delta_2(i)a_{i3}]b_3(O_3) = \max_{1\le j\le 3}[0.028\times 0.3, 0.0504\times 0.2, \mathbf{0.042\times 0.5}]\times 0.7 = 0.0147$$

$$\psi_3(3) = 3$$

✦ 隐藏序列结果：(3,3,3)

$$\delta_t(j) = \max_{1\le i\le N}[\delta_{t-1}(i)\cdot a_{ij}]\cdot b_j(O_t),\quad 2\le t\le T, 1\le j\le N$$

$$\psi_t(j) = \operatorname*{argmax}_{1\le i\le N}[\delta_{t-1}(i)\cdot a_{ij}]\cdot b_j(O_t),\quad 2\le t\le T, 1\le i\le N$$

·四：隐马尔可夫模型

★ 有3个盒子，每个盒子都有红色和白色两种球，分别为：

- ✦ 盒子1：6红4白
- ✦ 盒子2：3红7白
- ✦ 盒子3：4红6白

$$\pi = \begin{bmatrix} 0.3 \\ 0.3 \\ 0.4 \end{bmatrix} \quad A = \begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.2 & 0.5 & 0.3 \\ 0.3 & 0.5 & 0.2 \end{bmatrix} \quad B = \begin{bmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \\ 0.4 & 0.6 \end{bmatrix}$$

(1) 使用前向算法计算观察序列 $\{O_1, O_2, O_3, O_4\} = \{红, 白, 红, 白\}$的概率

(2) 使用后向算法计算观察序列 $\{O_1, O_2, O_3, O_4\} = \{白, 红, 白, 红\}$的概率

(3) 使用维特比算法计算观察序列 $\{O_1, O_2, O_3, O_4\} = \{红, 白, 红, 白\}$的最优状态序列

6.6 参数学习

6.6 参数学习

★ **问题3：模型参数学习**

★ 给定一个观察序列 $O = O_1 O_2 \cdots O_T$，如何根据最大似然估计来求模型的参数值？或者说如何调节模型 μ 的参数，使得 $p(O|\mu)$ 最大？即估计模型 μ 中的 $\pi_i, a_{ij}, b_j(k)$ 使得观察序列 O 的概率 $p(O|\mu)$ 最大。

★ **前向后向算法 (Baum-Welch or forward-backward procedure)**

6.6 参数学习

★ 如果HMM的状态序列 $Q = q_1 q_2 \cdots q_T$ 和产生的观察序列 $O = O_1 O_2 \cdots O_T$ 已知，则用最大似然估计来计算 μ 的参数：

$$\bar{\pi}_i = \frac{t = 1\text{时刻状态为}S_i\text{的次数}}{t = 1\text{时刻所有状态的总数}} = \frac{\delta(q_1, S_i)}{\sum_{i=1}^{N} \delta(q_1, S_i)}$$

$$\bar{a}_{ij} = \frac{Q\text{中从状态 } q_i \text{ 转移到 } q_j \text{ 的次数}}{Q\text{中所有从状态 } q_i \text{ 转移到另一状态（包括 } q_j\text{ ）的总数}}$$

$$= \frac{\sum_{t=1}^{T-1} \delta(q_t, S_i) \times \delta(q_{t+1}, S_j)}{\sum_{t=1}^{T-1} \delta(q_t, S_i)} \qquad (6\text{-}24)$$

其中，$\delta(x, y)$ 为 克罗内克δ函数(Kronecker delta)，
当$x = y$ 时，$\delta(x, y) = 1$，否则 $\delta(x, y) = 0$。

6.6 参数学习

★ 类似地，

$$\bar{b}_j(k) = \frac{Q\text{中从状态 } q_j \text{ 输出符号 } v_k \text{ 的次数}}{Q \text{ 到达 } q_j \text{ 的总次数}}$$

$$= \frac{\sum_{t=1}^{T}\delta\left(q_t, S_j\right) \times \delta(O_t, v_k)}{\sum_{t=1}^{T}\delta\left(q_t, S_j\right)} \tag{6-25}$$

其中，v_k 是模型输出符号集中的第 k 个符号。

6.6 参数学习

★ **期望值最大化算法 (Expectation-Maximization, EM)**

——处理无法计算状态序列次数时（存在隐变量时）

★ **基本思想**：初始化时随机地给模型的参数赋值（遵循限制规则，如：从某一状态出发的转移概率总和为1），得到模型 μ_0，然后可以从 μ_0 得到从某一状态转移到另一状态的期望次数，然后以期望次数代替公式中的次数，得到模型参数的新估计，由此得到新的模型 μ_1，从 μ_1 又可得到模型中隐变量的期望值，由此重新估计模型参数。循环这一过程，参数收敛于最大似然估计值。

6.6 参数学习

★ 给定模型 μ 和观察序列 $O = O_1O_2 \dots O_T$，那么，在时间 t 位于状态 S_i，时间 $t+1$ 位于状态 S_j 的概率：

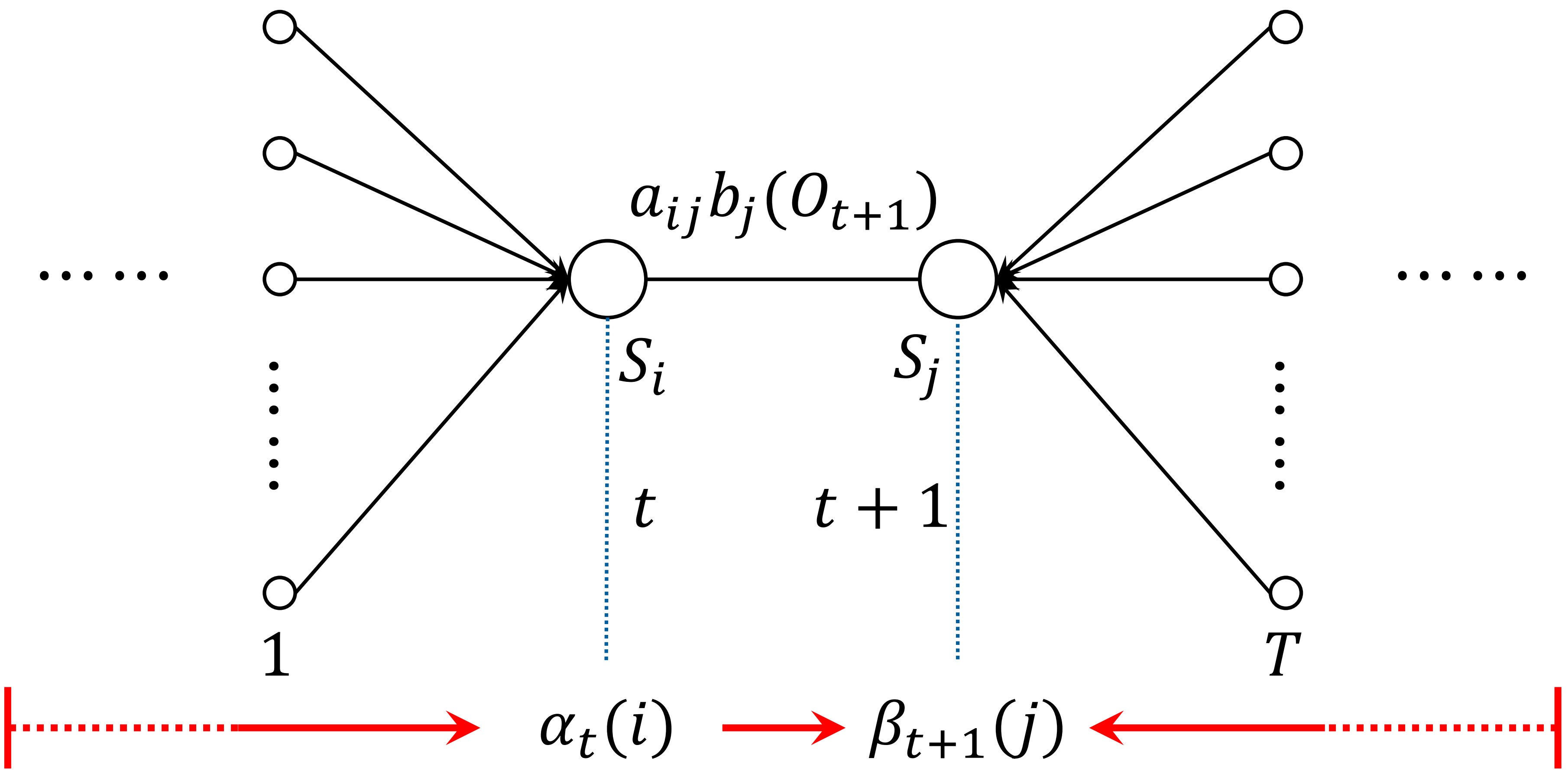
$$\xi_t(i,j) = p\left(q_t = S_i, q_{t+1} = S_j \middle| O, \mu\right)$$

$$= \frac{p(q_t = S_i, q_{t+1} = S_j, O|\mu)}{p(O|\mu)}$$

$$= \frac{\alpha_t(i) \times a_{ij}b_j(O_{t+1}) \times \beta_{t+1}(j)}{p(O|\mu)}$$

$$= \frac{\alpha_t(i) \times a_{ij}b_j(O_{t+1}) \times \beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i) \times a_{ij}b_j(O_{t+1}) \times \beta_{t+1}(j)} \qquad (6\text{-}26)$$

★ 图解搜索过程：



$$\boldsymbol{\alpha_t(i) \times a_{ij}b_j(O_{t+1}) \times \beta_{t+1}(j)}$$

★ 那么，给定模型 μ 和观察序列 $O = O_1 O_2 \dots O_T$，在时间 t 位于状态 S_i 的概率为：

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i, j) \tag{6-27}$$

由此，模型 μ 的参数可由下面的公式重新估计：

(1) q_1为 S_i 的概率：

$$\pi_i = \gamma_1(i) \tag{6-28}$$

6.6 参数学习

(2)

$$\bar{a}_{ij}=\frac{Q\text{中从状态 } q_i \text{ 转移到 } q_j \text{ 的期望次数}}{Q\text{中所有从状态 } q_i \text{ 转移到下一状态（包括 } q_j\text{ ）的期望次数}}$$

$$=\frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \tag{6-29}$$

(3)

$$\bar{b}_j(k)=\frac{Q\text{中从状态 } q_j \text{ 输出符号 } v_k \text{ 的期望次数}}{Q\text{ 到达 } q_j \text{ 的期望次数}}=\frac{\sum_{t=1}^{T}\gamma_t(j)\times\delta(O_t,v_k)}{\sum_{t=1}^{T}\gamma_t(j)}$$

(6-30)

6.6 参数学习

★ 算法6.4：Baum-Welch算法 (前向后向算法)描述：

(1) 初始化：随机地给$\pi_i, a_{ij}, b_j(k)$赋值，使得

$$\begin{cases} \sum_{i=1}^{N} \pi_i = 1 \\ \sum_{j=1}^{N} a_{ij} = 1, \qquad 1 \le 1 \le N \\ \sum_{k=1}^{M} b_i(k) = 1, \qquad 1 \le i \le N \end{cases} \tag{6-31}$$

由此得到模型 μ_0，令 $i = 0$。

(2) 执行EM算法：

E-步：由模型 μ_i 根据公式(6.26) 和(6.27) 计算期望值 $\xi_t(i,j)$ 和 $\gamma_t(j)$ 。

M-步：用E-步中所得到的期望值，根据公式(6.28- 6.30) 重新估计 $\pi_i, a_{ij}, b_j(k)$ 得到模型 μ_{i+1}。

循环：$i = i + 1$，重复执行E-步和M-步，直到 $\pi_i, a_{ij}, b_j(k)$ 的值收敛：$|\log p(O|\mu_{i+1}) - \log p(O|\mu_i)| < \varepsilon$ 。

(3) 结束算法，获得相应的参数

✦ EM 算法举例：

⅄ 估计某高校的学生被电信诈骗的比例

Q1：

$X = ?$

Q2：

$Y = ?$

新调查方法：

向5个人发放同一个问题，
不记录问题是什么，只记录回答

A1	是，是，否，否，否
A2	是，否，否，否，否
A3	是，是，是，否，否

5.4 语言模型的自适应

Q1	被电信诈骗过吗?
Q2	网恋过吗?
A1	是，是，否，否，否
A2	是，否，否，否，否
A3	是，是，是，否，否

✦ **EM 算法举例:**

⅄ **估计某高校的学生被电信诈骗的比例**

① 初始化

$X(\text{被电诈}) = 0.3 \qquad Y(\text{网恋}) = 0.6$

② 期望 Expectation

	Q1	Q2
A1		
A2		
A3		

$$P(A1|Q1) = \frac{P(A1, Q1)}{P(Q1)} = \frac{P(A1, Q1)}{P(A1, Q1) + P(A2, Q2)}$$

$$= \frac{0.3 \times 0.3 \times 0.7 \times 0.7 \times 0.7}{0.3 \times 0.3 \times 0.7 \times 0.7 \times 0.7 + 0.6 \times 0.6 \times 0.4 \times 0.4 \times 0.4}$$

$$\approx 0.57$$

④ 迭代

③ 最大 Maximization

	Q1		Q2	
	是	否	是	否
A1	1.14	1.71	0.43	1.72
A2	0.81	3.24	0.19	0.76
A3	0.81	0.54	2.19	1.46
T	2.76	5.49	3.24	3.51

$X = 0.33 \qquad Y = 0.48$

⑤ 收敛

$X = 0.35 \qquad Y = 0.35$

★ HMM使用中注意的问题

★ Viterbi 算法运算中的小数连乘，出现溢出

- ✦ 取对数

★ Baum-Welch 算法的小数溢出

- ✦ 放大系数
- ✦ 参阅[Rabiner and Juang, 1993: pp. 365-368]
- ✦ 参阅http://htk.eng.cam.ac.uk/

6.7 HMM应用举例

6.7 HMM应用举例

★ **汉语的自动分词 与 词性标注问题**

★ 举例：

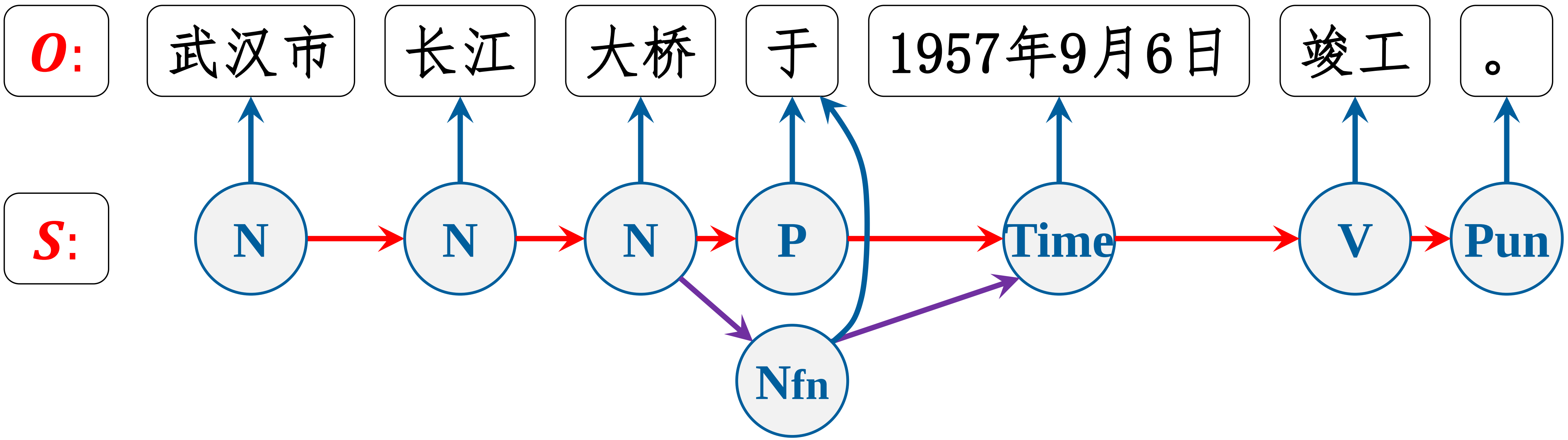
武汉市长江大桥于1957年9月6日竣工。

★ 可能的切分：

① 武汉市/N 长江/N 大桥/N 于/P 1957年9月6日/Time 竣工/V 。/Pun

② 武汉/N 市长/N 江大桥/N 于/P 1957年9月6日/Time 竣工/V 。/Pun

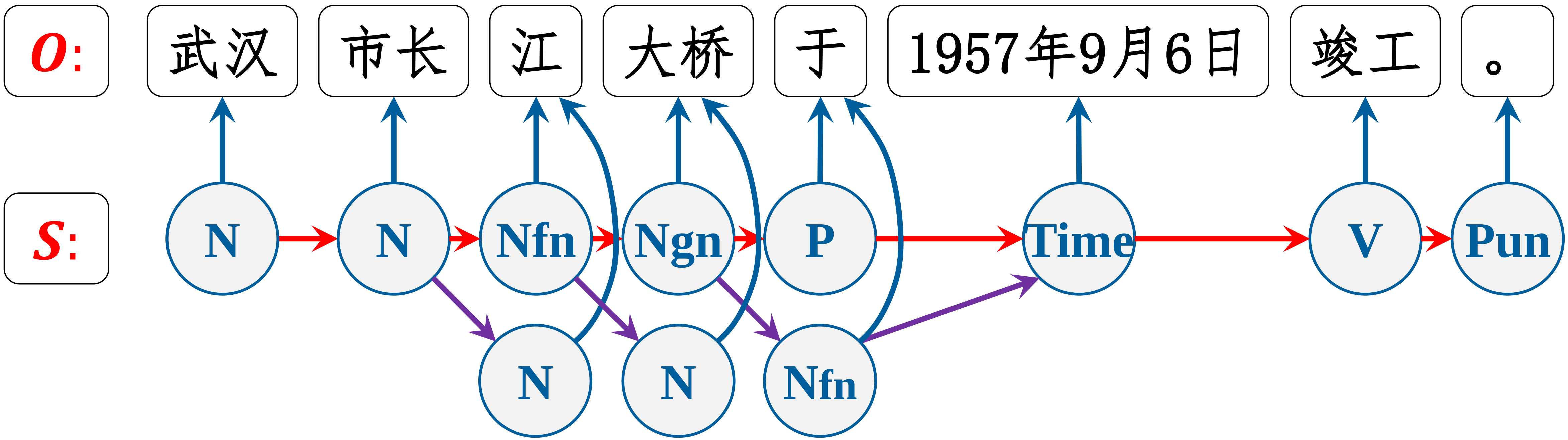
6.7 HMM应用举例



①武汉市/N 长江/N 大桥/N 于/P 1957年9月6日/Time 竣工/V 。/Pun

②武汉市/N 长江/N 大桥/N 于/Nrf 1957年9月6日/Time 竣工/V 。/Pun

6.7 HMM应用举例



①武汉/N 市长/N 江/Nfn 大桥/Ngn 于/P 1957年9月6日/Time 竣工/V 。/Pun

②武汉/N 市长/N 江/N 大桥/Ngn 于/P 1957年9月6日/Time 竣工/V 。/Pun

③武汉/N 市长/N 江/Nfn 大桥/N 于/P 1957年9月6日/Time 竣工/V 。/Pun

④武汉/N 市长/N 江/Nfn 大桥/N 于/Nfn 1957年9月6日/Time 竣工/V 。/Pun

6.7 HMM应用举例

★ 构造HMM的思路：

(1) 假设模型中状态（词性）的数目为词性符号的个数N

(2) 状态序列（词性序列）的马尔可夫性质

假设在统计意义上每个词性的概率分布只与上一个词的词性有关（即词性的二元语法），而每个单词的概率分布只与其词性相关，那么，通过对已经分词并做了词性标注的训练语料进行统计。

(3) 状态转移（词性到词性的转移）概率矩阵

(4) 从状态（词性）观察到输出符号（单词）的概率分布矩阵

(5) 求概率

对于任何一个给定的观察值序列(单词串),可以通过Viterbi算法得到一个可能性最大的状态值序列(词性串)。

6.7 HMM应用举例

★ 进一步解释：

(1) 估计HMM模型 $\mu = (A, B, \pi)$ 的参数；

(2) 对于任意给定的一个输入句子及其可能的输出序列 O，求找所有可能的 O 中使概率 $p(O|\mu)$ 最大的解；

(3) 快速地选择“最优”的状态序列（词性序列），使其最好地解释观察序列。

6.7 HMM应用举例

★ **用HMM 解决问题必须考虑的几个问题：**

(1) 如何确定状态、观察及其各自的数目？

(2) 参数估计：初始状态概率、状态转移概率、输出概率如何确定？

★ 思路：

✦ 如果把汉语自动分词结果作为观察序列 $O = O_1 O_2 \ldots O_T$ ， 那么，我们要求解的是：$\hat{O} = \underset{O}{\operatorname{argmax}}\, p(O|\mu)$ 。

✦ 对于词性标注而言，则需求解： $\hat{Q} = \underset{Q}{\operatorname{argmax}}\, p(O|\mu)$ 。

6.7 HMM应用举例

★ 问题1：模型参数

(1) 观察序列：单词序列

(2) 状态序列：词类标记序列

(3) 状态数目 N：为词类标记符号的个数，如北大语料库词类标记符号数为106个；

(4) 输出符号数 M：每个状态可输出的不同词汇个数，如汉语介词 P 约有60个，连词 C 约有110个，即状态 P 和 C 分别对应的输出符号数为60、110。

★ 参数估计

(1)如果**无任何标注语料**：需要一部有词性标注的词典，采用无指导学习方法：

a) 获取词类个数(状态数)；

b) 获取对应每种词类的词汇数(输出符号数)；

c) 利用EM迭代算法获取初始状态概率、状态转移概率和输出符号概率。

6.7 HMM应用举例

(2) 若**有大规模分词和词性标注语料**：有指导学习方法

咱们/rr 中国/ns 这么/rz 大{da4}/a 的{de5}/ud 一个/mq 多/a 民族/n 的{de5}/ud 国家/n 如果/c 不/df 团结/a ，/wd 就/d 不/df 可能/vu 发展/v 经济/n ，/wd 人民/n 生活/n 水平/n 也/d 就/d 不/df 可能/vu 得到/v 改善/vn 和{he2}/c 提高/vn 。/wj

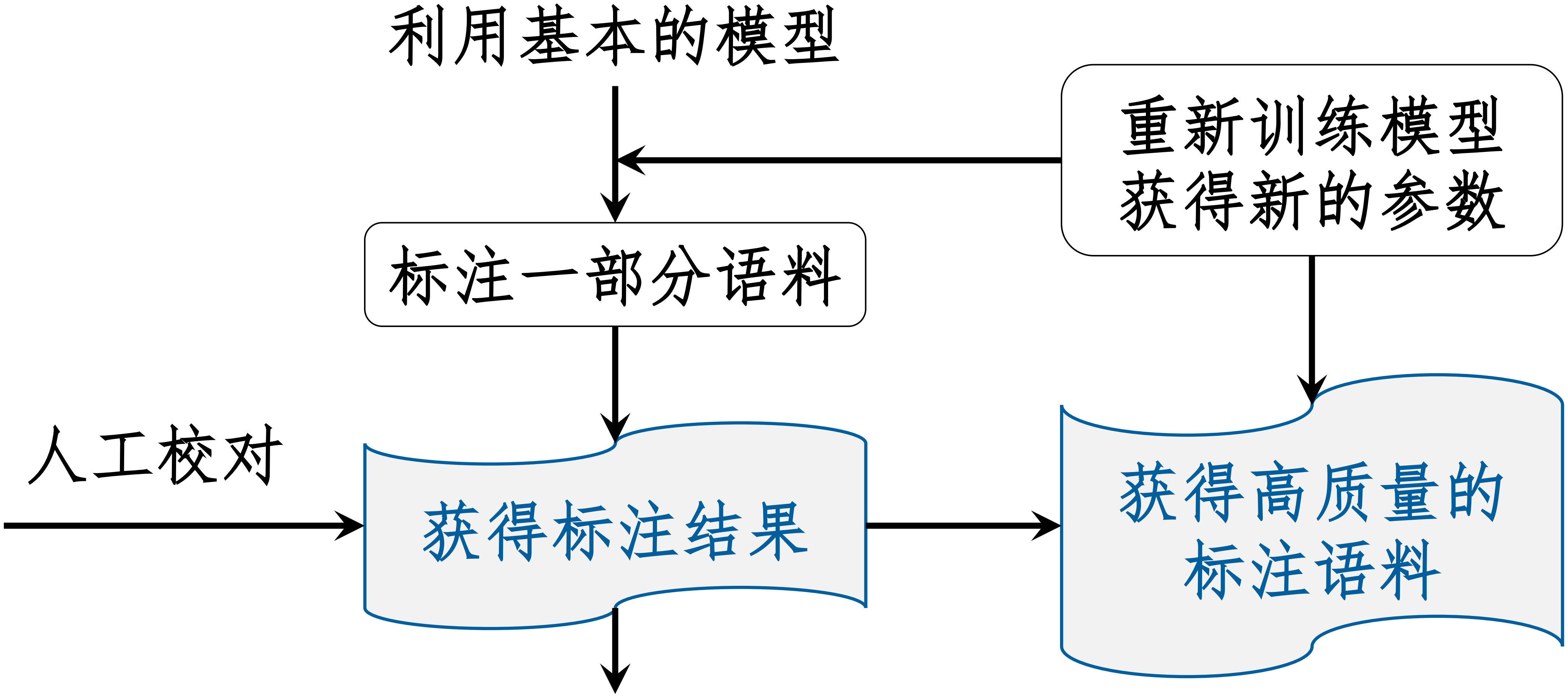
可以从这些标注语料中抽取出所有的词汇和词类标记，并用最大似然估计方法计算各种概率。

$$\bar{\pi}_{pos_i} = \frac{POS_i\text{出现在句首的次数}}{\text{所有句首的个数}}$$

$$\bar{a}_{ij} = \frac{\text{从词类}POS_i\text{转移到}POS_j\text{的次数}}{\text{所有从状态}POS_i\text{转移到另一}POS(\text{包括}POS_j)\text{的总数}}$$

$$\bar{b}_j = \frac{\text{从状态}POS_j\text{输出词汇}w_k\text{的次数}}{\text{状态}POS_j\text{出现的总次数}}$$

★ 一般来说，需要通过错误驱动的机器学习方法修正模型的参数：



6.7 HMM应用举例

★ **问题2：如何获取观察序列?**

—借助于其他工具，获得n-best的粗切分(所有可能的切分)。

★ 本地主叫通话时长1400分钟。

本地/ 主叫/ 通话/ 时长/ **1400**/ 分钟/ 。

本/ 地主/ 叫/ 通话/ 时/ 长/ **1400**/ 分钟/ 。

本/ 地主/ 叫/ 通话/ 时长/ **1400**/ 分钟/ 。

★ 负责任

负/ 责任

负责/ 任

负/ 责/ 任

6.7 HMM应用举例

★ 分词实验：以“负责任”为例

✦ 利用部分《人民日报》语料

词类 / 词	A	C	Q	NF	NG	NL	V	VN	总计
负责	4	0	0	0	0	0	177	50	231
任	0	4	11	59	2	4	98	0	178
其他	34469	25475	24232	11453	4550	25670	184488	42674	353011
总计	34473	25479	24243	11512	4552	25674	184763	42724	353420

6.7 HMM应用举例

$O_1 = w_1 w_2 =$ 负责/ 任, $\quad p(O_1|\mu) = 5.4 \times 10^{-6}$

$O_2 = w_1 w_2 =$ 负/ 责任, $\quad p(O_2|\mu) = 9.3 \times 10^{-6}$

$O_3 = w_1 w_2 w_3 =$ 负/ 责/ 任, $\quad p(O_3|\mu) = 4.3 \times 10^{-6}$

$$p(O_2|\mu) > p(O_1|\mu) > p(O_3|\mu)$$

第二种切分结果可能性较大：负/ 责任

6.7 HMM应用举例

★ **分词性能测试：** Ref. 汉语自动分词和中文人名识别技术研究[硕士学位论文]，浙江大学，2006

- 封闭测试：《人民日报》1998年1月份的部分切分和标注语料，约占训练语料的1/10，计78396个词，含中国人名1273个。准确率（人名识别前）：**90.34%**。

- 开放测试：《人民日报》1998年2月份的部分切分和标注语料，也占训练语料的1/10，共82347个词，含中国人名2316个。准确率（人名识别前）：**86.32%**。

6.7 HMM应用举例

★ **词性标注性能测试：** 应用于词性标注的隐马尔可夫模型参数估计[硕士学位论文]，大连理工大学，2006

- 采用有指导的参数估计方法；
- 训练语料：北京大学标注的《人民日报》2000年1、2、4月份的语料；
- 封闭测试：2000年2月20-29日的标注语料，词性标注的精确率为：**95.16%**；
- 开放测试：2000年3月1-7日的语料，词性标注的精确率为：**88.45%**。

6.7 HMM应用举例

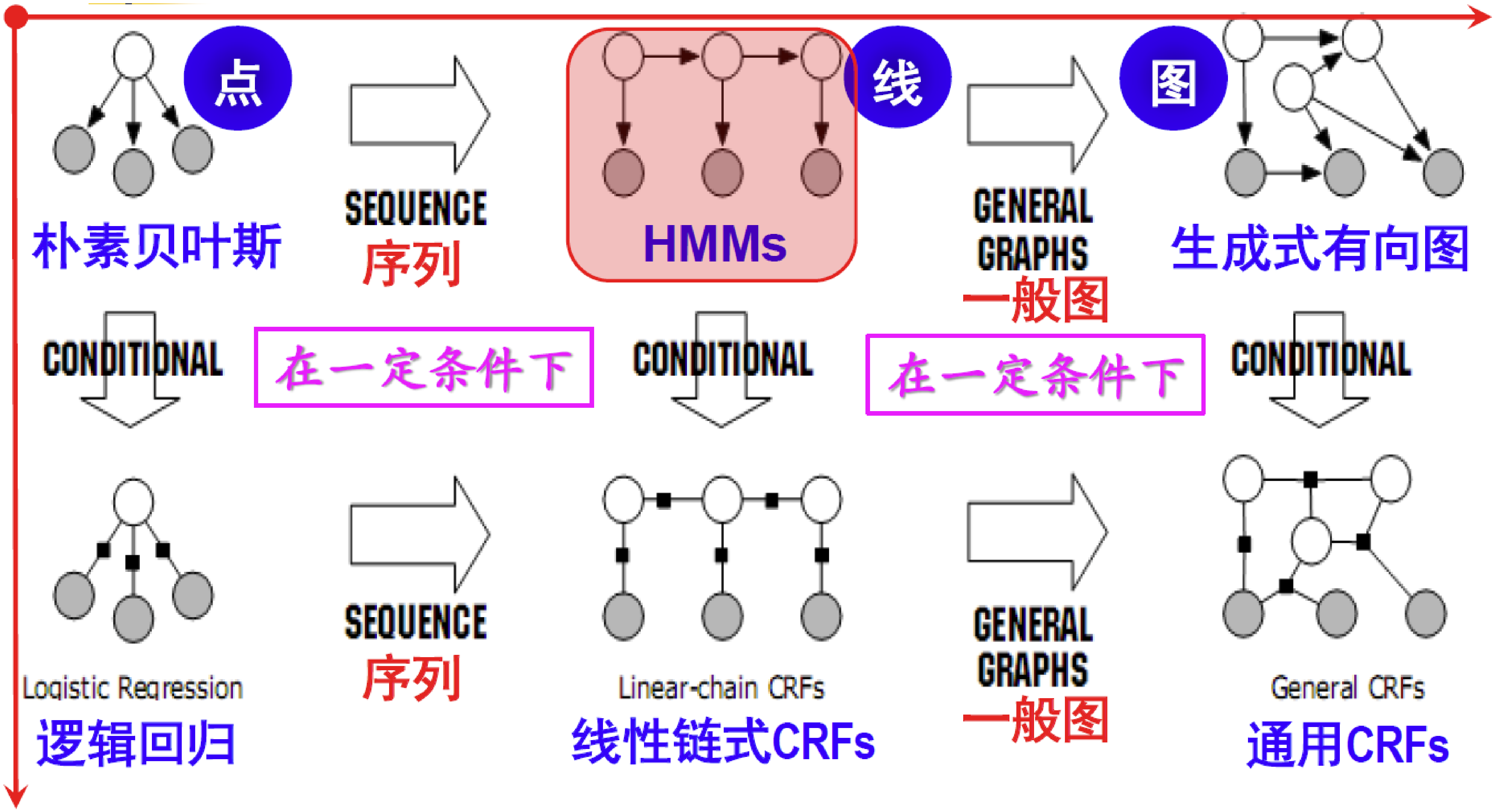
★ 训练语料规模对模型参数的影响：

★ 选用北大标注的2000年《人民日报》语料作为训练数据。5个训练语料集大小不同：C1为2月份的；C2为1月及2月份的；C3为1、2和4月份的；C4为1、2、4和9月份的；C5为1、2、4、9和10月份五个月的。采用相同的测试集（2000年3月份前7天的语料），观察词性标注的精确率变化：

语料	C1	C2	C3	C4	C5
准确率%	86.16	90.85	88.45	88.82	89.04

6.8 CRFs及其应用

NLP中概率图模型的演变



6.8 CRFs及其应用

★ 提出

- **提出动因：** 在NLP和图像处理中有一类问题是进行序列标注和结构划分，而n-gram是利用当前时刻 t 之前已经发生的时间信息。
- **条件随机场(conditional random fields, CRFs)**于2001年由 J. Lafferty 等人提出，是用于标注和划分序列结构数据的概率化结构模型，在自然语言处理和图像处理中得到了广泛应用。
- **基本思路：** 给定观察序列 $\boldsymbol{X}$，输出标注序列 $\boldsymbol{Y}$，通过计算 $\boldsymbol{P}(\boldsymbol{Y}|\boldsymbol{X})$求解最优标注序列。

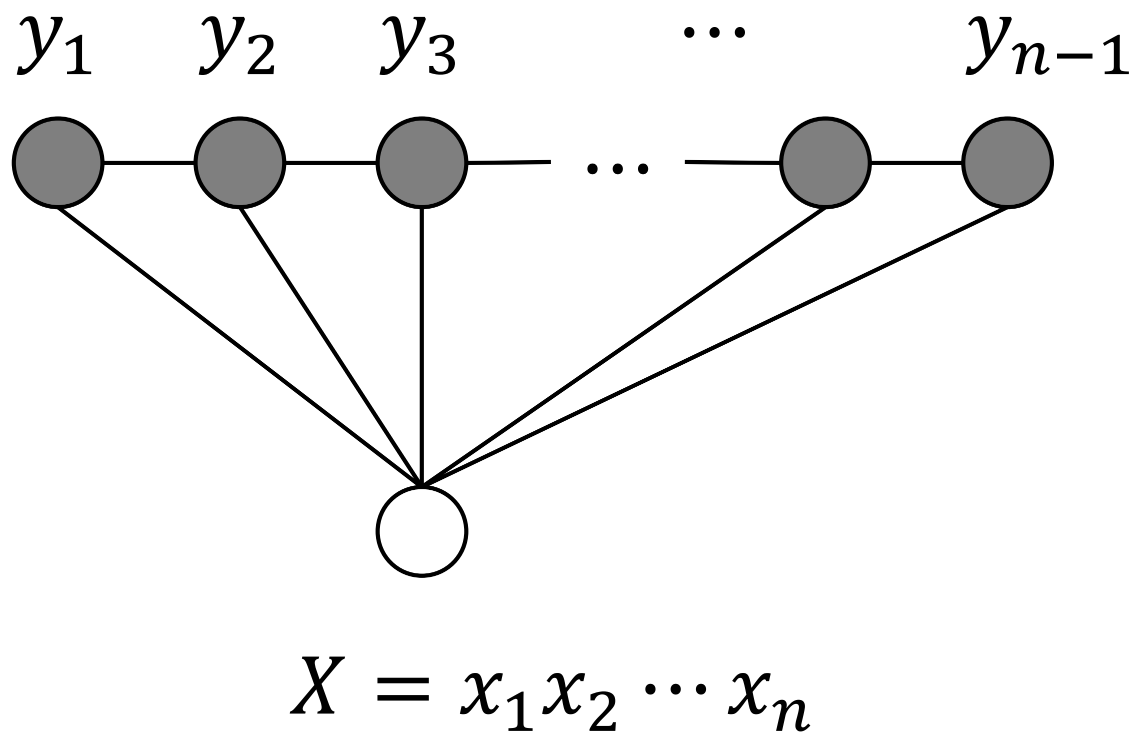
★ 定义

- ✦ 设 $\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{E})$ 为一个无向图，$\boldsymbol{V}$ 为结点集合，$\boldsymbol{E}$ 为无向边的集合
- ✦ 设$\boldsymbol{X}$与$\boldsymbol{Y}$是随机变量，$\boldsymbol{P}(\boldsymbol{Y}|\boldsymbol{X})$是在给定$\boldsymbol{X}$的条件下$\boldsymbol{Y}$的条件概率分布
- ✦ $\boldsymbol{Y}=\{\boldsymbol{Y}_{\boldsymbol{v}}|\boldsymbol{v}\in\boldsymbol{V}\}$，即 $\boldsymbol{V}$ 中每个结点对应于一个随机变量 $\boldsymbol{Y}_{\boldsymbol{v}}$
- ✦ 如果以观察序列 $\boldsymbol{X}$ 为条件，每个随机变量 $\boldsymbol{Y}_{\boldsymbol{v}}$ 都满足以下马尔可夫特性：

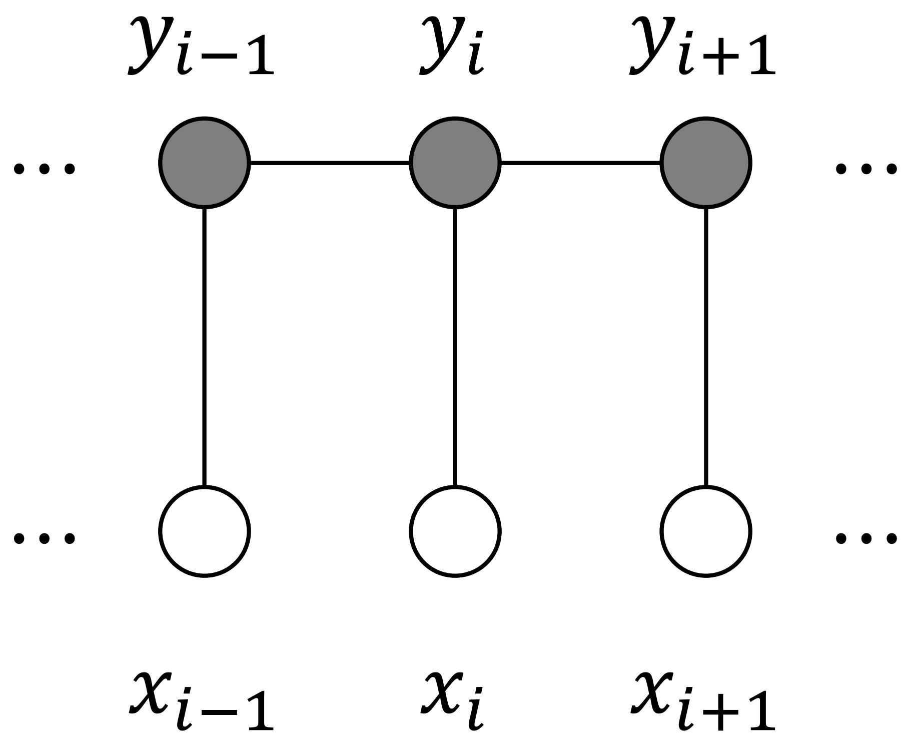
$$P(Y_v|X,Y_w,w\neq v)=p(Y_v|X,Y_w,w\sim v) \qquad (6\text{-}32)$$

- ✦ 其中，$\boldsymbol{w}\sim\boldsymbol{v}$ 表示两个结点在图中是邻近结点
- ✦ 则称 $\boldsymbol{G}$ 为概率分布 $\boldsymbol{P}(\boldsymbol{Y}|\boldsymbol{X})$ 的条件随机场。

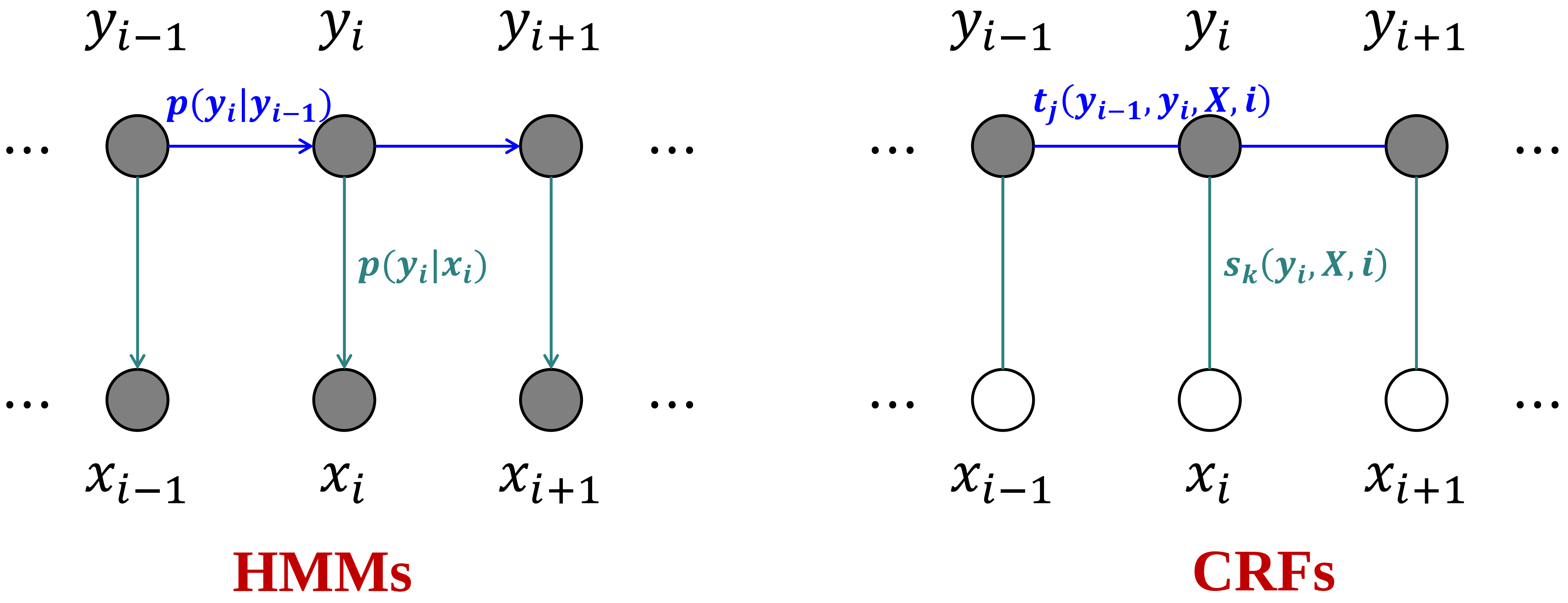
★ 理论上，只要在标注序列中描述了一定的条件独立性，$\boldsymbol{G}$ 的图结构可以任意的。序列标注问题可以建模为简单的链式结构图，结点对应标注序列$\boldsymbol{Y}$中的元素。如下图所示



或



★ HMMs 生成模型 vs. CRFs 判别模型



✦ CRFs 中的空心节点x表示该节点并不是由模型生成的。

6.8 CRFs及其应用

★ 在给定观察序列$\boldsymbol{X}$ 时，某个特定标注序列$\boldsymbol{Y}$的概率可以定义为：

$$P(Y/X)=\exp\left(\sum_j \lambda_j t_j(y_{i-1},y_i,X,i)+\sum_k \mu_k s_k(y_i,X,i)\right) \quad (6\text{-}33)$$

- ✦ 其中， $t_j(y_{i-1},y_i,X,i)$ 是转移函数，表示对于观察序列 X 的标注序列在 i 及 $i-1$ 位置上标注的转移概率；
- ✦ $s_k(y_i,X,i)$是状态函数，表示观察序列 X 在 i 位置的标注概率；
- ✦ λ_j 和 μ_k 分别是 t_j 和 s_k 的权重, 需要从训练样本中估计出。

6.8 CRFs及其应用

★ 可以定义一组关于观察序列的 $\{0,1\}$ 二值特征 $b(X,i)$，表示训练样本中某些特征的分布，如：

$$b(X,i)=\begin{cases}1 & \text{如果}X\text{的}i\text{位置为某个特定的词}\\ 0 & \text{否则}\end{cases}$$

★ 转移函数可以定义为如下形式：

$$t_j(y_{i-1},y_i,X,i)=\begin{cases}b(X,i) & \text{如果}y_{i-1}\text{和}y_i\text{满足某种搭配条件}\\ 0 & \text{否则}\end{cases}$$

★ 也可以把状态函数写成如下形式：

$$s(y_i,X,i)=s(y_{i-1},y_i,X,i)$$

6.8 CRFs及其应用

★ 由此，特征函数可以统一表示为：

$$F_j(Y,X) = \sum_{i=1}^{n} f_j(y_{i-1}, y_i, X, i) \tag{6-34}$$

其中，每个局部特征函数$f_j(y_{i-1}, y_i, X, i)$表示状态特征$s(y_{i-1}, y_i, X, i)$转移数$t(y_{i-1}, y_i, X, i)$。

★ 条件随机场定义的条件概率可以由下式给出：

$$p(Y|X,\lambda) = \frac{1}{Z(X)} \exp\left(\sum_j \lambda_j \cdot F_j(Y,X)\right) \tag{6-35}$$

其中，$Z(X)$为归一化因：$Z(X) = \sum_Y \exp(\sum_j \lambda_j \cdot F_j(Y,X))$

★ **实现CRFs 也需要解决如下三个问题：**

(1) 特征选取

(2) 参数训练

(3) 解码

6.8 CRFs及其应用

★ **应用举例：由字构词（基于字标注）的分词方法**

(Character-based tagging) Ref. Xue and Converse, 2002

- 该方法由N.Xue(薛念文) 和S. Converse 提出，发表在2002年第一届国际计算语言学学会(ACL)汉语特别兴趣小组SIGHAN组织的汉语分词评测研讨会上(https://aclanthology.org/sigs/sighan/)

★ **基本思想：** 将分词过程看作是字的分类问题：每个字在构造一个特定的词语时都占据着一个确定的构词位置(即词位)。一般情况下，每个字只有4个词位：词首(B)、词中(M)、词尾(E)和单独成词(S)。

6.8 CRFs及其应用

★ 乒乓球拍卖完了。

(1) 乒乓球/ 拍/ 卖/ 完/ 了/ 。/

(2) 乒乓球/ 拍卖/ 完/ 了/ 。/

(3) 乒/B 乓/M 球/E 拍/S 完/S 了/S 。/S

★ 在字标注过程中，对所有的字根据预定义的特征进行词位特征学习，获得一个概率模型，然后在待切分字串上，根据字与字之间的结合紧密程度，得到一个词位的分类结果，最后根据词位定义直接获得最终的分词结果。

6.8 CRFs及其应用

★ 乒/B 乓/M 球/E 拍/S 卖完了。

B，E，M，S?

★ 当前字的前后*n*个字

★ 当前字左边字的标注

★ 当前字在词中的位置

★ ……

① 特征选取

✦ 一元特征（状态函数）：当前字、当前字的前一个字、当前字的后一个字

$$s_1(y_i, X, i) = \begin{cases} 1 & \text{如果当前字是“拍”，当前字的标注} y_i \text{是} M \\ 0 & \text{否则} \end{cases}$$

$$s_2(y_i, X, i) = \begin{cases} 1 & \text{如果当前字是“拍”，当前字的标注} y_i \text{是} E \\ 0 & \text{否则} \end{cases}$$

……

乒/**B** 乓/**M** 球/**E** 拍/**S** 卖/? 完了。

① 特征选取

✦ 二元特征（转移函数）：对应各标签间转移函数的特征

$$t_1(y_{i-1}, y_i, X, i) = \begin{cases} 1 & \text{如果前一个字的标注} y_{i-1} \text{是} B\text{，当前的标注} y_i \text{是M} \\ 0 & \text{否则} \end{cases}$$

$$t_2(y_{i-1}, y_i, X, i) = \begin{cases} 1 & \text{如果前一个字的标注} y_{i-1} \text{是} M\text{，当前的标注} y_i \text{是M} \\ 0 & \text{否则} \end{cases}$$

……

乒/B 乓/M 球/E 拍/S 卖/? 完了。

② 参数训练

★ 通过训练语料估计特征权重 λ_j, 使其在给定一个观察序列 X 的条件下, 找到一个最有可能的标注序列 Y， 即条件概率$P(Y|X)$ 最大。

★ 条件概率已由上文的(6-35)式给出：

$$p(Y|X,\lambda)=\frac{1}{Z(X)}\exp\left(\sum_j \lambda_j\cdot F_j(Y,X)\right)$$

$$Z(X)=\sum_Y \exp(\sum_j \lambda_j\cdot F_j(Y,X))$$

6.8 CRFs及其应用

★ 为了训练特征权重 λ_j，需要计算模型的损失和梯度。由梯度更新 λ_j，直到 λ_j 收敛。

★ 损失函数定义为负对数似然函数:

$$L(\boldsymbol{\lambda}) = -\log p(Y|X,\boldsymbol{\lambda}) + \frac{\varepsilon}{2}\boldsymbol{\lambda}^2 \quad (\varepsilon\text{取值范围：}10^{-6}\sim10^{-3})$$

★ 损失函数的梯度为:

$$\frac{\partial L(\boldsymbol{\lambda})}{\partial \lambda_j} = \frac{\partial \log Z(X)}{\partial \lambda_j} - F_j(Y,X) + \varepsilon\boldsymbol{\lambda}$$

6.8 CRFs及其应用

③ 解码

★ 条件随机场解码的过程就是根据模型求解的过程，可以由维特比(Viterbi)算法完成。维特比算法是一个动态规划算法，动态规划要求局部路径也是最优路径的一部分。

③ 解码

★ 以中文分词为例：乒乓球拍卖完了

★ 维特比算法就是在下面由标注组成的矩阵中搜索一条最优的路径。

乒	乓	球	拍	卖	完	了
B	B	B	B	B	B	B
M	M	M	M	M	M	M
E	E	E	E	E	E	E
S	S	S	S	S	S	S

★ 分词结果：乒/B 乓/M 球/M 拍/E 卖/S 完/S 了/S

6.8 CRFs及其应用

乒	乓	球	拍	卖	完	了
B	B	B	B	B	B	B
M	M	M	M	M	M	M
E	E	E	E	E	E	E
S	S	S	S	S	S	S

★ 到达每个标注的分数由以下三部分组成：

- ✦ 标注的一元特征权重$\boldsymbol{W}$。分别用$\boldsymbol{W}_1^B$表示第一个字被标注为$\boldsymbol{B}$的权重，用$\boldsymbol{W}_1^S$表示第一个字被标注为$\boldsymbol{S}$的权重，等等。
- ✦ 标注的路径得分$\boldsymbol{R}$。分别用$\boldsymbol{R}_2^B$ 表示第二个字被标注为$\boldsymbol{B}$时的路径得分，用$\boldsymbol{R}_2^E$ 表示第二个字被标注为$\boldsymbol{E}$的路径得分，等等。
- ✦ 前一个字的标注到当前字标注转移的特征权重$\boldsymbol{T}$。用$\boldsymbol{T}_{BM}$表示由标注$\boldsymbol{B}$到$\boldsymbol{M}$的转移特征权重，类似地，其他转移特征权重分别记为：$\boldsymbol{T}_{BE}$、$\boldsymbol{T}_{MM}$、$\boldsymbol{T}_{ME}$、$\boldsymbol{T}_{EB}$、$\boldsymbol{T}_{ES}$、$\boldsymbol{T}_{SB}$和$\boldsymbol{T}_{SS}$等。

6.8 CRFs及其应用

乒	乓	球	拍	卖	完	了
B	B	B	B	B	B	B
M	M	M	M	M	M	M
E	E	E	E	E	E	E
S	S	S	S	S	S	S

★ 利用下式迭代计算每一字被标注为每一种标注的分数:

$$R_{i+1}^{B} = \max\{T_{EB} \times R_i^E, T_{SB} \times R_i^S\} \times W_{i+1}^{B}$$
$$R_{i+1}^{E} = \max\{T_{BE} \times R_i^B, T_{SE} \times R_i^E\} \times W_{i+1}^{E}$$
$$R_{i+1}^{S} = \max\{T_{ES} \times R_i^E, T_{SS} \times R_i^S\} \times W_{i+1}^{S}$$

…… ……

6.8 CRFs及其应用

1	2	3	4	5	6	7
乒	**乓**	**球**	**拍**	**卖**	**完**	**了**
B	B	B	B	B	B	B
M	M	M	M	M	M	M
E	E	E	E	E	E	E
S	S	S	S	S	S	S

★ **第一步**：计算第一个字“乒”的标注分数（以标注B为例）。由于不存在转移特征，故路径权重R_1^B为：

$$R_1^B = W_1^B = \lambda_1 \times f(null, 乒, B) + \lambda_2 \times f(乒, B) + \lambda_3 \times f(乒, B, 乓)$$

★ $f(\blacksquare)$表示特征，其中$f(null, 乒, B)$表示当前字“乒”被标注为B，前一个字为空；$f(乒, B)$表示当前字“乒”被标注为B；$f(乒, B, 乓)$表示当前字“乒”被标注为B，且后一个字为“乓”。

★ 特征的权重λ_1、λ_2和λ_3都可以从训练中得到（参数训练部分）。

6.8 CRFs及其应用

1	2	3	4	5	6	7
乒	乓	球	拍	卖	完	了
B	B	B	B	B	B	B
M	M	M	M	M	M	M
E	E	E	E	E	E	E
S	S	S	S	S	S	S

★ **第二步**：计算第二个字“乓”的标注分数（以标注B为例）。首先计算一元权重W_2^B，继而由上一个字的路径权重计算当前路径权重R_2^B为：

$$R_2^B = \max\{T_{EB} \times R_1^E, T_{SB} \times R_1^S\} \times W_2^B$$

★ 同样，对于“乓”字的标注S、M和E分别计算R_2^M、R_2^E和R_2^S。

6.8 CRFs及其应用

1	2	3	4	5	6	7
乒	乓	球	拍	卖	完	了
B	B	B	B	B	B	B
M	M	M	M	M	M	M
E	E	E	E	E	E	E
S	S	S	S	S	S	S

★ **第三步**：依据第二步迭代计算直至最后一个“了”字，得到R_7^E，R_7^S。比较这两个值，确定最优路径，然后以该值的标注点为起始点回溯，得到整个句子的最优路径。回溯过程：

★ 由：$\max\{R_7^E, R_7^S\} = R_7^S$，可推出“了”字标注为S；

★ 由：$R_7^S = \max\{T_{ES} \times R_6^E, T_{SS} \times R_6^S\} \times W_7^S = T_{SS} \times R_6^S \times W_7^S$

★ 可推出“完”字标注为S；依次回溯至第一个字，解码完毕。

6.8 CRFs及其应用

★ **关于条件随机场模型的实现工具：**

★ CRF++（C++版）

✦ http:// https://taku910.github.io/crfpp/

★ CRFSuite（C语言版）

✦ http://www.chokkan.org/software/crfsuite/

★ MALLET（Java版，通用的自然语言处理工具包，包括分类、序列标注等机器学习算法）：

✦ http://mallet.cs.umass.edu/

★ NLTK（Python版，通用的自然语言处理工具包，很多工具是从MALLET中包装转成的Python接口）：

✦ http://nltk.org/

★ **参考文献：**

★ [1]J. Lafferty, A. McCallum, and F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proc.ICML'2001,*pages 282-289

★ [2]H. M. Wallach. Conditional Random Fields: An Introduction. *CIS Technical Report MS-CIS-04-21*, Univ. of Penn., 2004

6.8 本章小结

★ **HMM的构成：**

①状态数 ②输出符号数 ③初始状态的概率分布 ④状态转移的概率 ⑤输出概率

★ **HMM 的三个基本问题：**

①快速计算给定模型的观察序列概率: 前/后向算法

②求最优状态序列：**Viterbi** 算法

③参数估计：**Baum-Welch** 算法

★ **模型实现中需要注意的问题：**小数溢出