

第四章 隐马尔可夫模型与贝叶斯网络

苏智勇

可视计算研究组

南京理工大学

suzhiyong@njust.edu.cn

<https://zhiyongsu.github.io>

主要内容

4.1 贝叶斯网络的基本概念

4.2 隐马尔可夫模型 (HMM)

4.3 朴素贝叶斯分类器

4.1 贝叶斯网络的基本概念

- 随机变量
- 随机变量的条件独立性
 - 两个随机变量 x 和 y 独立, 当且仅当 $p(x, y) = p(x)p(y)$
 - 随机变量间的条件独立性: 如果随机变量 x, y, z 满足 $p(x, y | z) = p(x | z)p(y | z)$, 则称 x 和 y 关于 z 条件独立,

$$p(x | y, z) = \frac{p(x, y, z)}{p(y, z)} = \frac{p(x, y | z)p(z)}{p(y | z)p(z)} = p(x | z)$$

4.1 贝叶斯网络的基本概念

- 随机过程
 - 分布（联合分布）是对一个随机变量（随机向量）的刻画，而过程是对一族随机变量的刻画！
 - 定义：随机过程是定义在 $\Omega \times T$ 上的二元函数 $X(\omega, t)$ 。对于固定的时间 t , $X(\omega, t)$ 为随机变量，简记为 $X(t)$ 或 X_t ；对于固定的 ω , $X(\omega, t)$ 为时间 t 的一般函数，称为样本函数或样本轨道，简记为 $x(t)$ 。
- 随机变量和样本函数是两个具有不同定义域和值域的单值函数。
- 随机过程既可看成是“所有随机变量的集合”，也可看成为“所有样本函数的集合”。

4.1 贝叶斯网络的基本概念

- 随机过程
 - 随机过程示意图
 - 随机变量 $X(t)$: 对于固定时间 t , 蓝色虚线圈圈表示随机变量
 - 样本函数 $X(\omega, t)$: 对于固定的 ω , $X(\omega, t)$ 为时间 t 的样本函数 (3个)
 - 股票的涨跌过程
 - 今天红-明天红-后天绿
 - 今天绿-明天绿-后天绿
 - ...

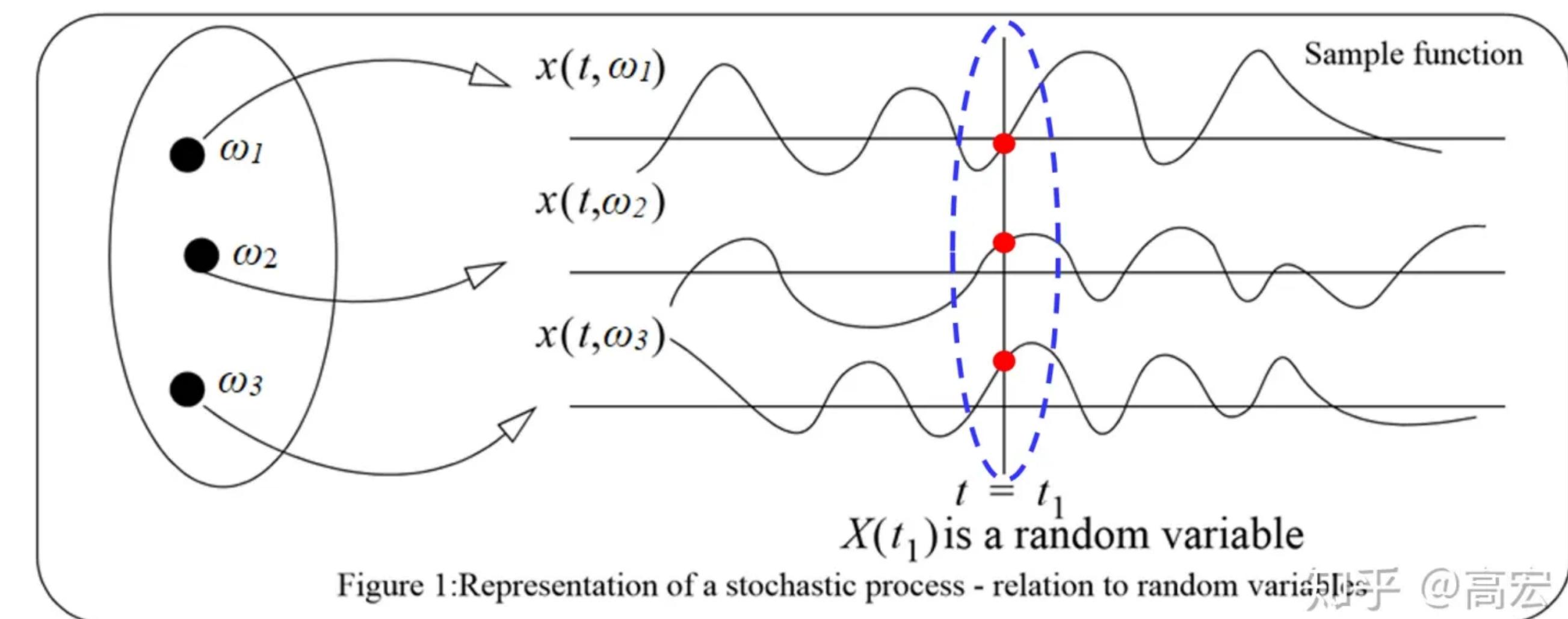


Figure 1: Representation of a stochastic process - relation to random variables
知乎 @高宏

4.1 贝叶斯网络的基本概念

- 马尔科夫性质
 - 也叫做无后效性、无记忆性，即过去只能影响现在，不能影响将来。
 - 如果 $X(t)$, $t > 0$ 为一个随机过程，则马科夫性质可以符号化成如下形式：

$$Pr[X(t + h) = y | X(s) = x(s), s \leq t] = Pr[X(t + h) = y | X(t) = x(t)], \forall h > 0$$

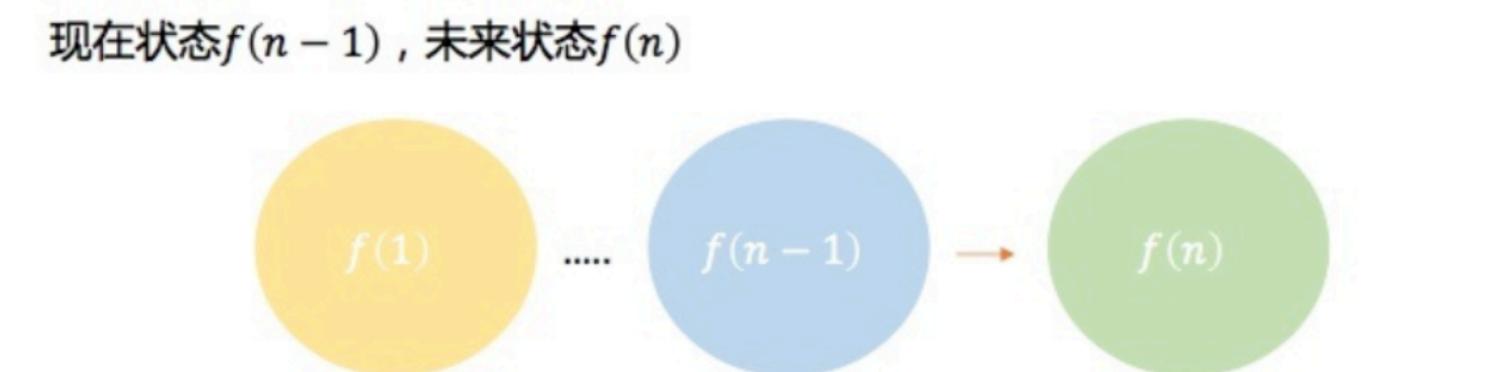
- 过去(s)并不影响将来($t + h$)的状态，但当前状态蕴含着以往所有的状态信息。

4.1 贝叶斯网络的基本概念

- 马尔科夫模型（链）

定义：具有马尔可夫性质、并以随机过程为基础模型的随机过程/随机模型被统称为马尔可夫模型（链）。

- 任性的过程：它将来的状态分布只取决于现在，跟过去无关！（一阶）
- Life is like a Markov chain, your future only depends on what you are doing now, and independent of your past.

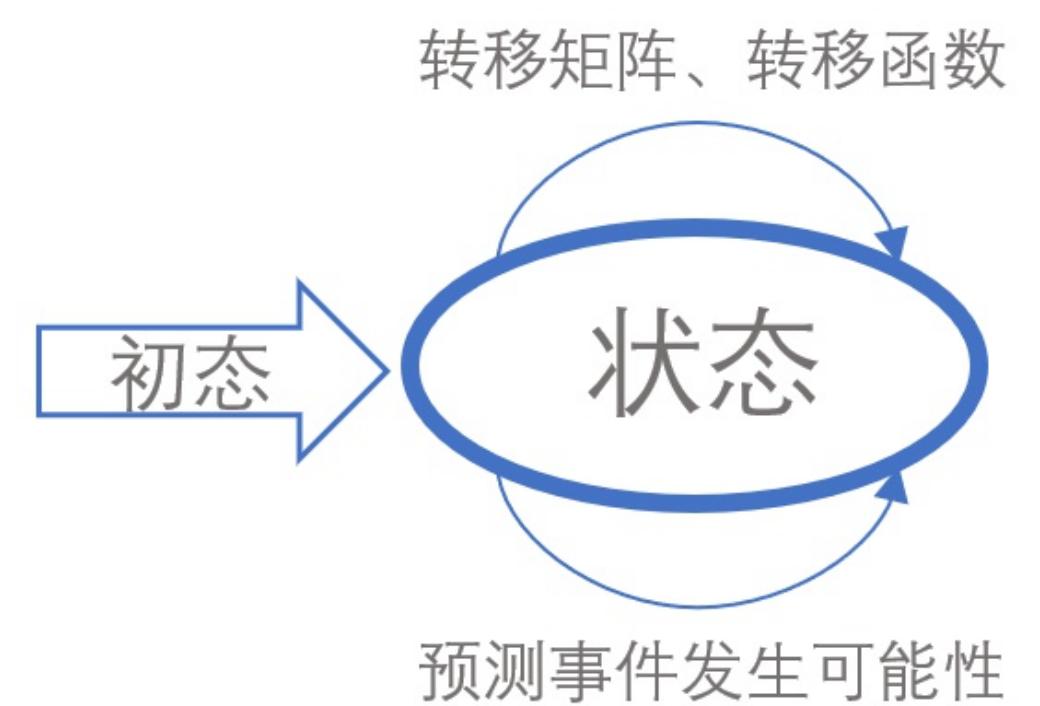


现在决定未来

4.1 贝叶斯网络的基本概念

- 马尔科夫模型（链）基本要素
 - 状态空间： $X_n = i$ 表示随机过程在 n 时刻处在 i 状态，所有状态的取值构成的集合称为“状态空间”，以符号 I 表示
 - 转移概率：把在当前时刻状态到下一时刻某状态的条件概率称作转移概率

$$P_{ij} = P(X_n = j | X_{n-1} = i)$$



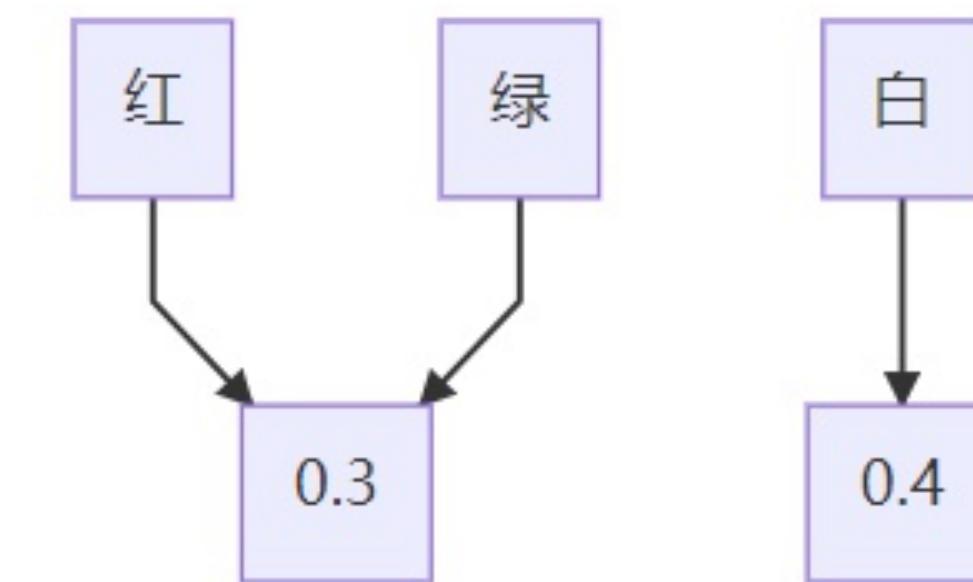
知乎 @马克波罗的鸡腿

4.1 贝叶斯网络的基本概念

- 马尔科夫模型（链）基本要素

- 转移概率矩阵：所有状态之间的转移概率组成一个矩阵。状态转移矩阵不随时间的变化。

$$P = \begin{Bmatrix} p_{11} & p_{12} & \cdots & p_{1I} \\ p_{21} & p_{22} & \cdots & p_{2I} \\ \vdots & \vdots & \ddots & \vdots \\ p_{I1} & p_{I2} & \cdots & p_{II} \end{Bmatrix}$$



知乎 @马克波罗的鸡腿

- 初始状态：即初始分布

- $\cdot p_0 = (p_0(1), p_0(2), p_0(3)) = (0.3, 0.4, 0.3)$

- $\cdot p_n(j) = P(X_n = j), j = 1, 2, 3$

4.1 贝叶斯网络的基本概念

- 马尔科夫模型（链）基本要素
 - 转移方程：对于 m 步转移，则有转移方程如下

$$p_{ij}^{(n+m)} = \sum_{k=1}^I p_{ik}^{(n)} p_{kj}^{(m)}$$

- 今天： $p_0 = (0.3, 0.4, 0.3)$
- 明天： $p_1 = p_0 \times P^1$
- 后天： $p_2 = p_1 \times P = p_0 \times P^2$



4.1 贝叶斯网络的基本概念

- 链式法则

$$P(AB) = P(A) \cdot P(B|A) = P(B) \cdot P(A|B)$$

- 高维随机变量（随机向量）的联合分布：

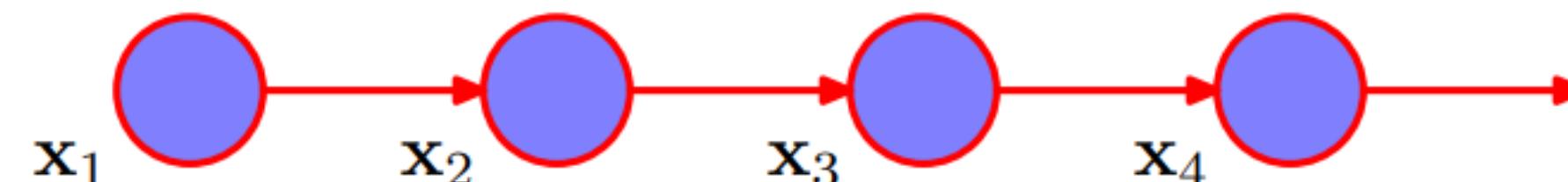
$$p(x_1, x_2, \dots, x_d) = p(x_1, x_2, \dots, x_{d-1})p(x_d | x_1, x_2, \dots, x_{d-1})$$

$$p(x_1, x_2, \dots, x_d) = p(x_1)p(x_2 | x_1)p(x_3 | x_1, x_2) \dots p(x_d | x_1, x_2, \dots, x_{d-1})$$

$$p(x_1, \dots, x_d) = p(x_1) \prod_{n=2}^d p(x_n | x_1, \dots, x_{n-1})$$

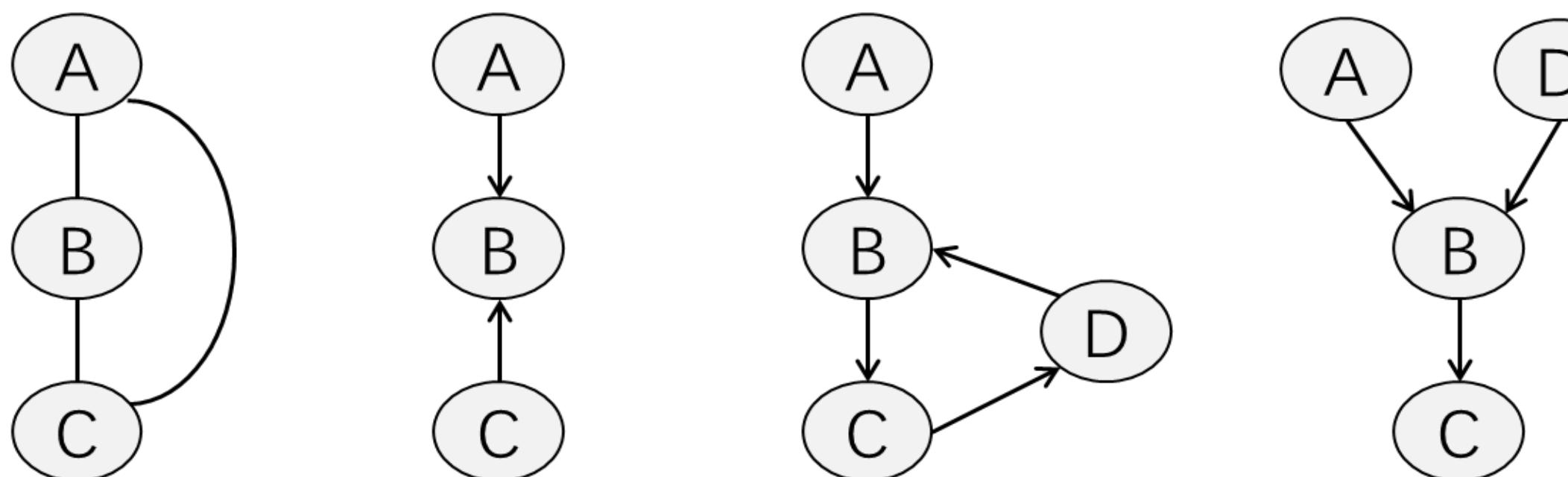
- 如果 x_{t+1} 与 x_1, x_2, \dots, x_{t-1} 关于 x_t 条件独立，则上式可以简化为

$$p(x_1, \dots, x_d) = p(x_1) \prod_{n=2}^d p(x_n | x_{n-1})$$



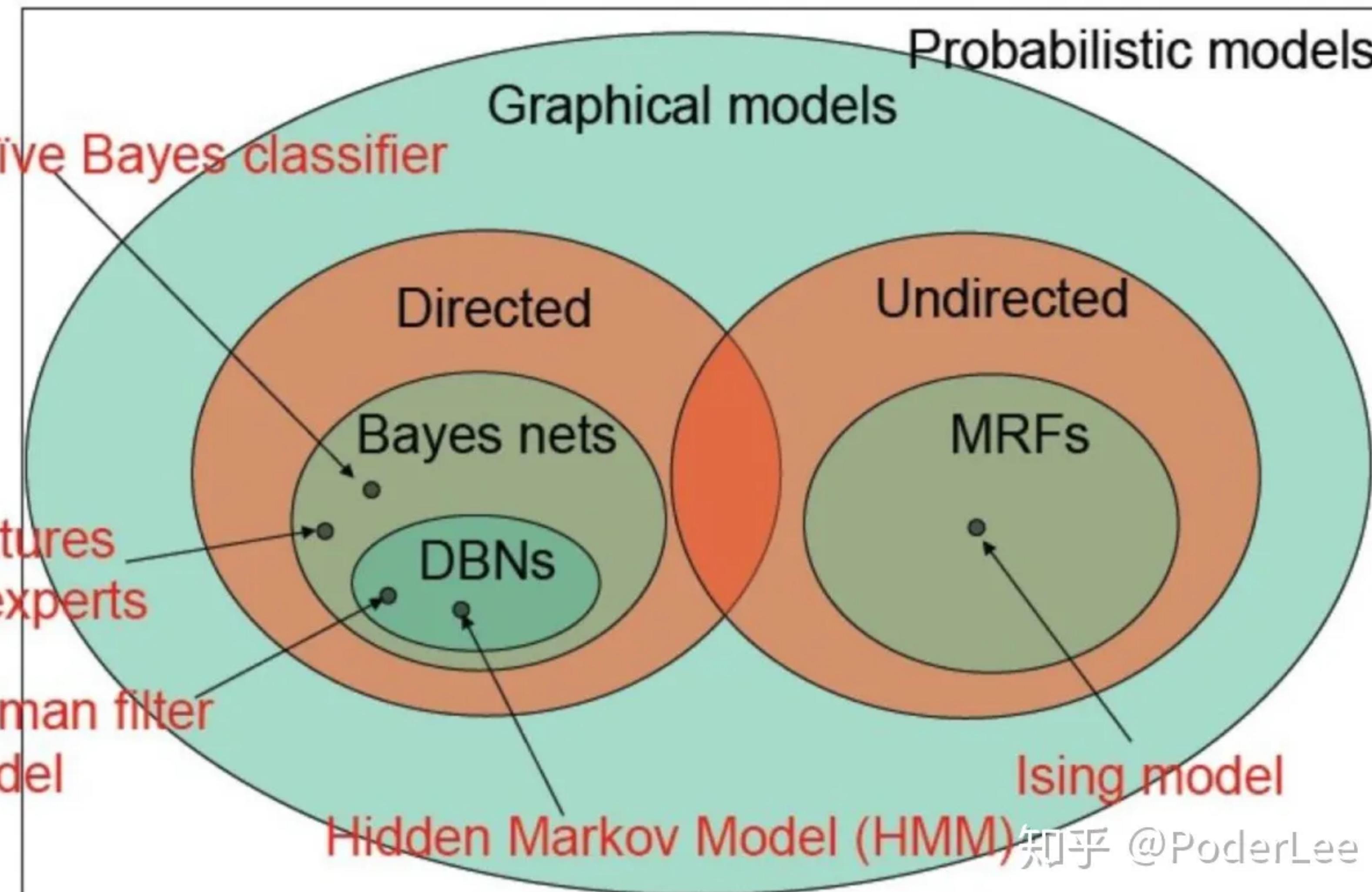
4.1 贝叶斯网络的基本概念

- 条件概率的图表示
 - 图的基本概念
 - 节点、边；有向图、无向图；有环图、无环图
 - 条件概率 $P(X = a | Y = b)$ 、联合概率 $P(X = a, Y = b)$ 、边缘概率 $P(X = a)$



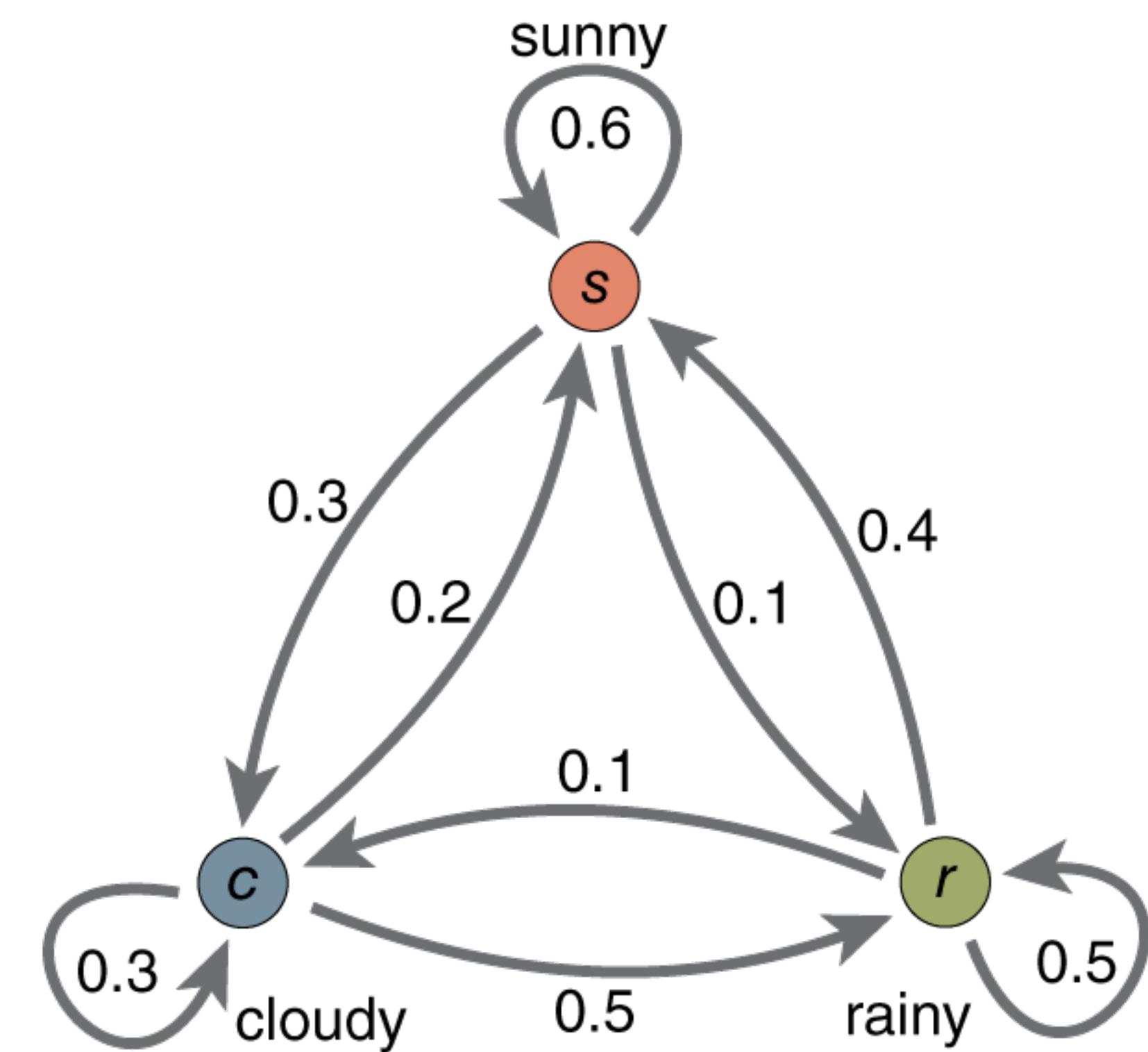
4.1 贝叶斯网络的基本概念

- 贝叶斯网
 - 一种用概率有向网表示的模型
 - 无向网的因子图
 - 通过图模型的计算



4.2 隐马尔科夫模型

- 一种特殊的动态贝叶斯网络
- HMM案例
 - 已知:
 - 天气: 状态空间 (Sunny, Rainy, Cloudy) 、天气转移概率
 - 行为: 散步、购物、内务
 - A、B跨国恋, A的对象B通过A的行为猜测可能的天气状态



4.2 隐马尔科夫模型

- HMM案例

- 条件:

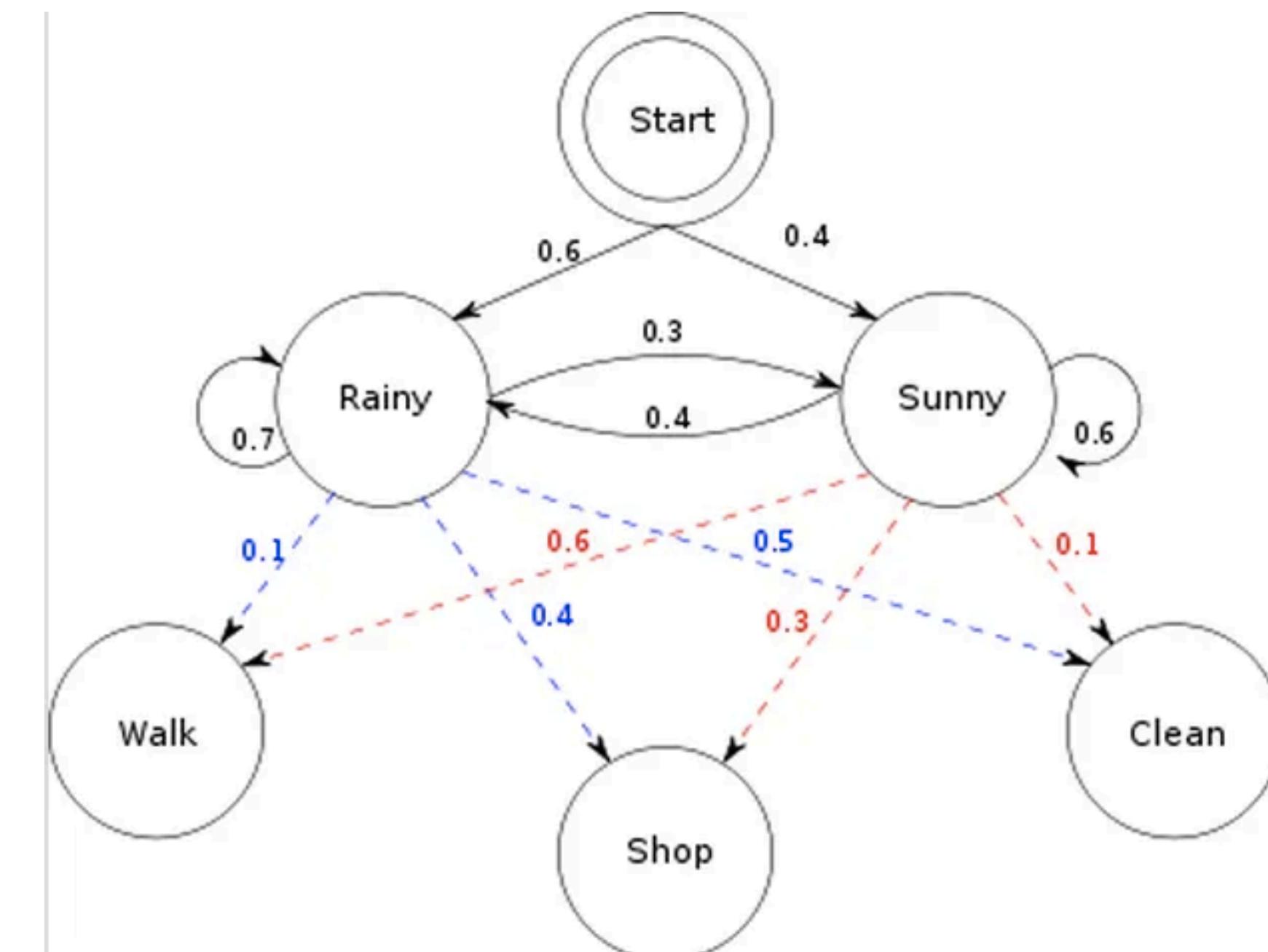
- 隐状态序列: 天气状况 (B无法直接观测到)

- 观测序列: A的行为 (A告知B)

- 行为概率: 以晴天和雨天两种情况为例

- ◆ 晴天: 散步, 购物, 内务的概率分别是0.6, 0.3, 0.1

- ◆ 雨天: 散步, 购物, 内务的概率分别是0.1, 0.4, 0.5



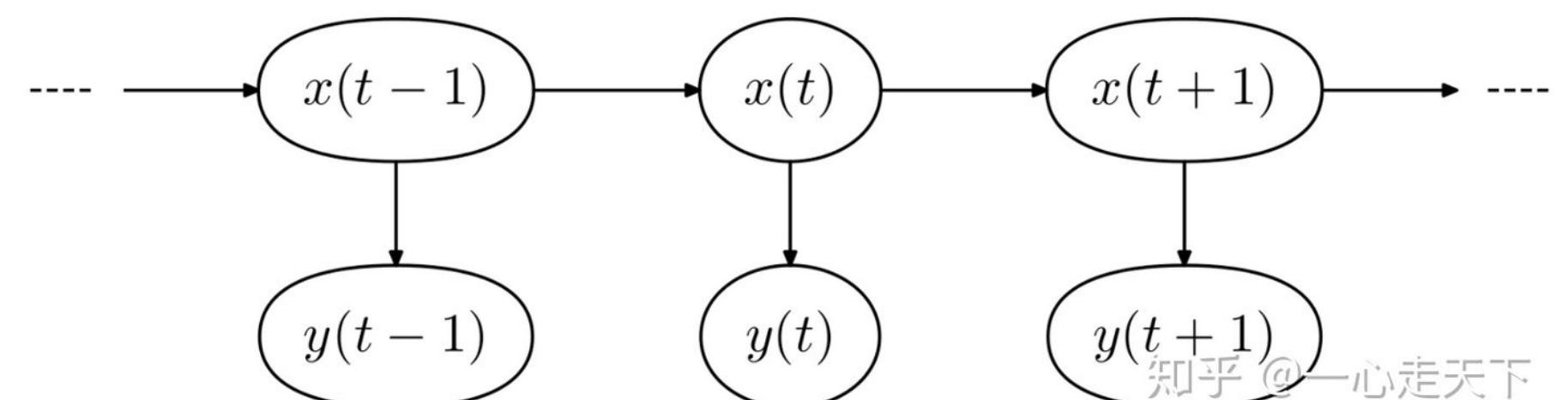
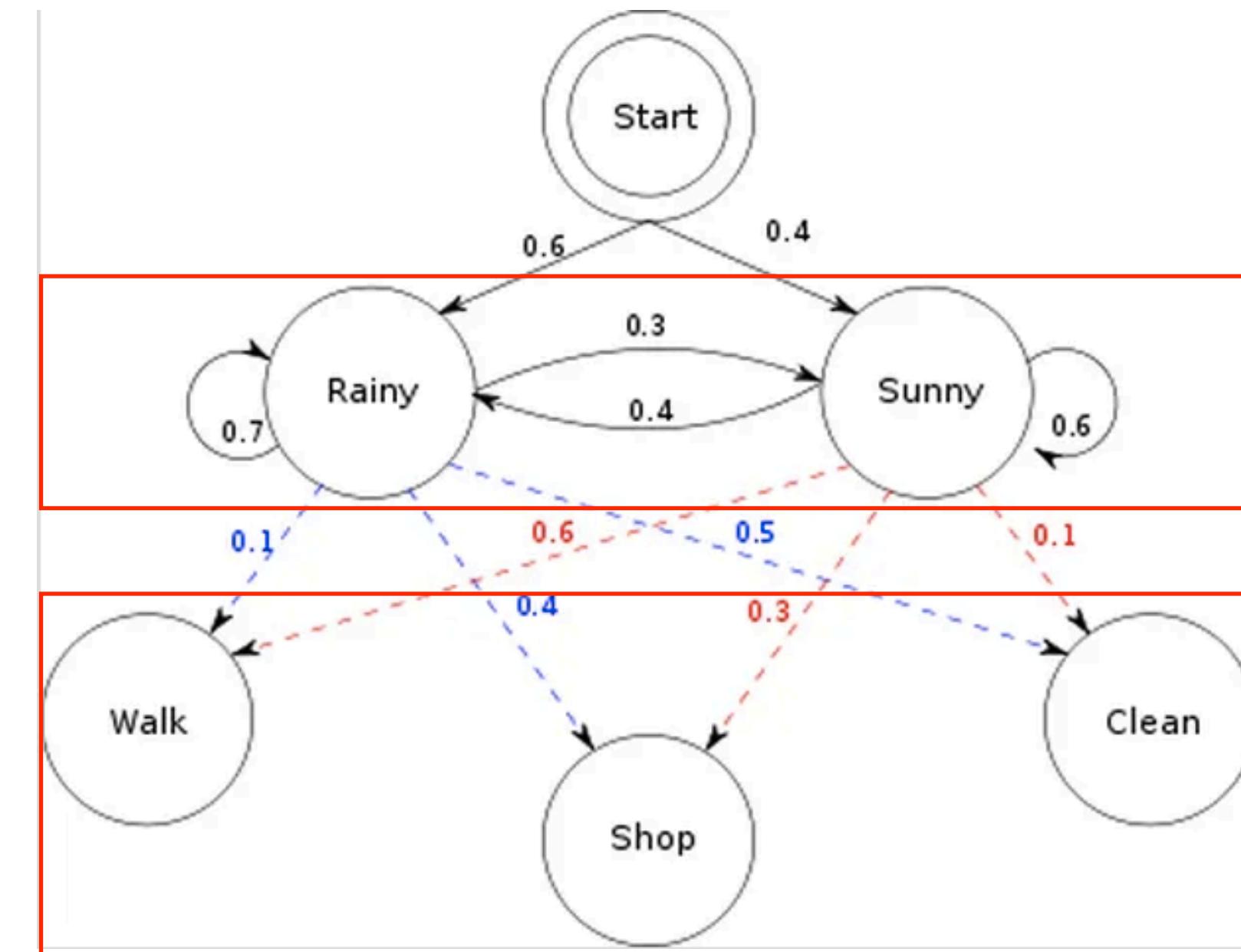
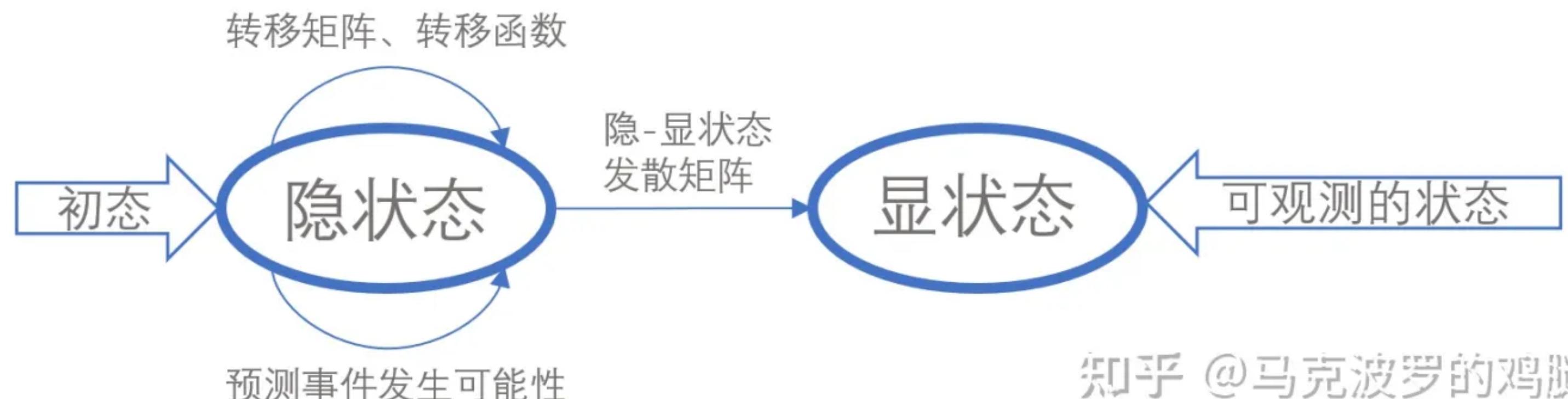
4.2 隐马尔科夫模型

- **HMM案例**
 - **问题1:** 已知整个模型，连续三天，A下班后做的事情分别是：散步，购物，内务。那么，B根据模型，计算产生这些行为的概率是多少。
 - **问题2:** 同样知晓这个模型，同样是这三件事，A要B猜，这三天她下班后北京的天气是怎么样的。这三天怎么样的天气才最有可能让她做这样事情。
 - **问题3,** 最复杂的，A只告诉B这三天她分别做了这三件事，而其他什么信息我都没有。A要B建立一个模型，晴雨转换概率，第一天天气情况的概率分布，根据天气情况她选择做某事的概率分布。（惨绝人寰）

4.2 隐马尔科夫模型

- 基本概念

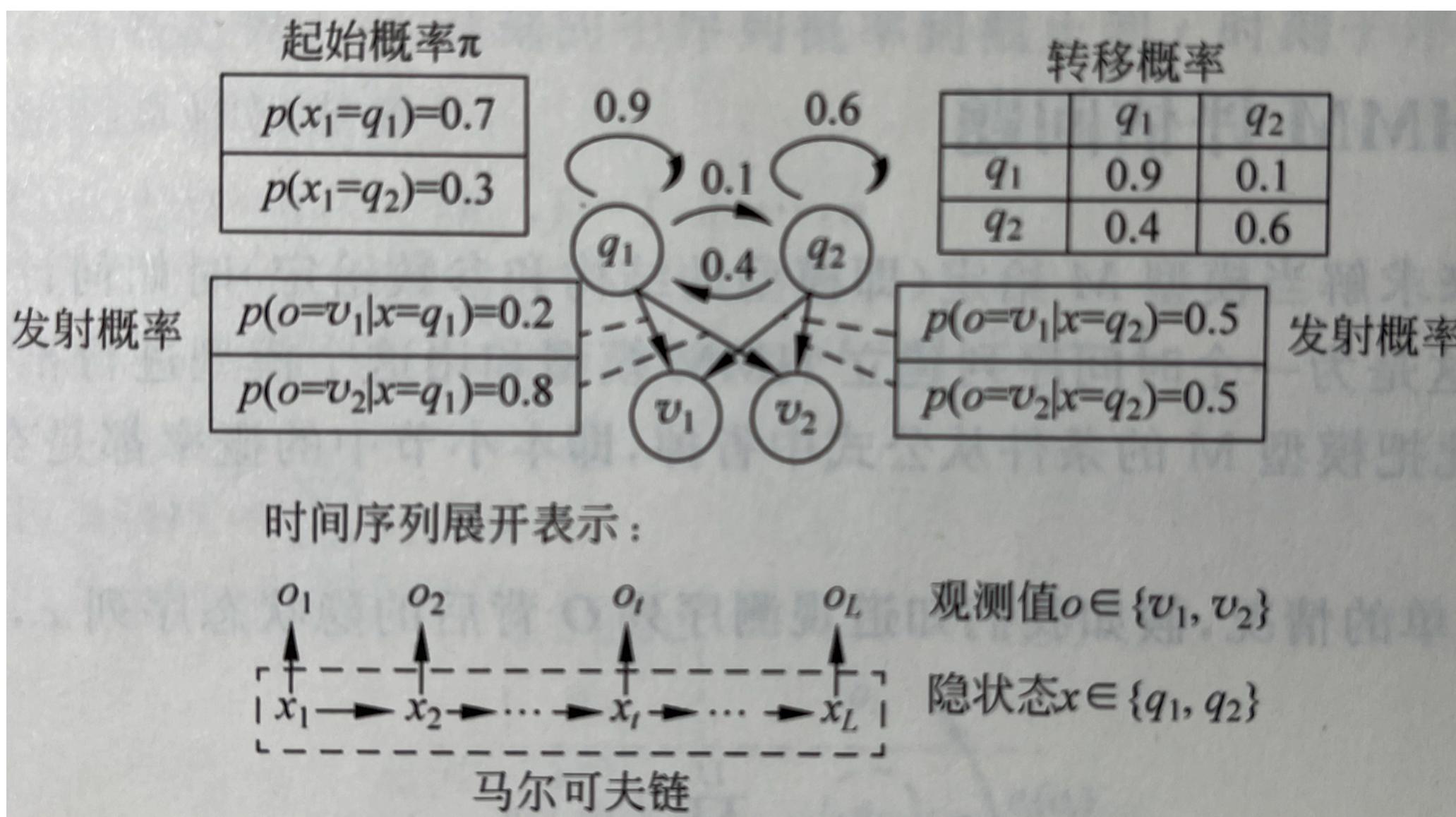
- 系统的状态取值服从马尔可夫链
- 系统状态的变化无法直接观察到
- 可能的状态和观测是贝叶斯网络的**隐节点**和**可观测节点**
- **转移概率**为隐节点之间的边
- **发射概率**为隐节点到可观测节点之间的边



知乎 @马克波罗的鸡腿

4.2 隐马尔科夫模型

- HMM三要素 $M = (A, E, \pi)$
 - 状态转移矩阵 A , 发射概率矩阵 E , 初始概率分布 π



变量表示	说明
$Q = \{q_1, q_2, \dots, q_n\}$	模型含有 n 个隐状态
$V = \{v_1, v_2, \dots, v_v\}$	观测值的取值范围
$A = [a_{ij}]_{n \times n}$	状态转移概率矩阵, a_{ij} 表示从状态 i 转到状态 j 的概率, 满足 $\sum_{j=1}^n a_{ij} = 1, \forall i$
$O = o_1 o_2 \dots o_L$	长度为 L 的观测序列, o_t 的取值为 V 中某个值
$X = x_1 x_2 \dots x_L$	长度为 L 的隐状态序列, x_t 的取值为 Q 中某个值
$E = [e_{ij}]_{n \times V}$	发射概率矩阵, $e_{ij} = p(o = v_j x = q_i)$ 表示模型隐状态取值 q_i 时观测到 v_j 的概率, 满足 $\sum_{j=1}^V e_{ij} = 1, \forall i$
$\pi = [\pi_1, \pi_2, \dots, \pi_n]$	初始概率分布, π_i 表示马氏链从该状态起始的概率, 满足 $\sum_{i=1}^n \pi_i = 1$

4.2 隐马尔科夫模型

- 隐马尔科夫模型的三种典型问题：解决时间序列的决策问题

- 模型评估问题 (Evaluation)

概率计算问题。已知HMM模型 M 和观测序列 O , 求当前观测序列出现的概率 $p(O | M)$ (似然度)。

- 隐状态推断问题 (Decoding)

预测问题。已知HMM模型 M 和观测序列 O , 求产生该序列的最有可能的隐状态序列

$$x = \arg \max_x p(x, O | M).$$

- 模型学习问题 (Learning)

已知观测序列 O 和隐层状态序列 x , 求HMM的模型参数 $\arg \max_M p(x, O | M)$ 或者

$$\arg \max_M p(O | M).$$

4.2.1 HMM模型评估问题

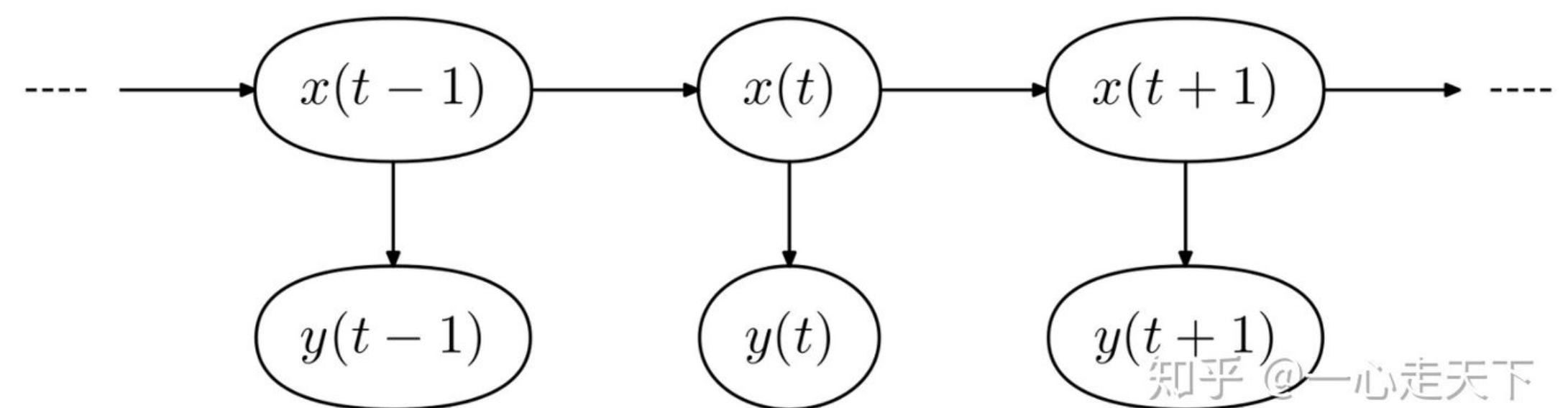
已知HMM模型 M （模型结构和参数）和观测序列 O ，求当前观测序列出现的概率 $p(O|M)$ （似然度）。

- 已知隐状态序列 x ，即 $p(x)$ 已知

似然度：
$$p(O|x) = \prod_{i=1}^L p(o_i | x_i)$$

- 观测序列 O 和隐状态序列 x 的联合概率，即 $p(O|M)$ ：

$$p(O, x) = p(O | x)p(x) = \prod_{i=1}^L p(o_i | x_i) \left(\pi_{x_1} \prod_{i=2}^L p(x_i | x_{i-1}) \right)$$



4.2.1 HMM模型评估问题

已知HMM模型 M （模型结构和参数）和观测序列 O ，求当前观测序列出现的概率 $p(O|M)$ （似然度）。

- 未知隐状态序列 x ，即 $p(x)$ 未知
 - 概率分解：穷举，求解当前观测序列在所有可能隐状态序列下的加权概率和

$$p(O) = \sum_x p(O, x) = \sum_x p(O, x)p(x)$$

- 计算量巨大： n^L 种隐状态序列取值组合

4.2.1 HMM模型评估问题

- 前向（向前）算法：Forward Algorithm
 - 已知当前状态，要得到结果，往后能怎么“走”？（**求果**）
 - 从前往后，迭代求解：长度为 t 的观测子序列 $o_1 o_2 \dots o_t$ 在 t 时刻隐变量取值 q_j 的概率 $\alpha_t(j)$
 - $\alpha_t(j) = p(o_1 o_2 \dots o_t, x_t = q_j)$: t 时刻，隐状态为 q_j 时，观测子序列(1-t)出现的概率
 - $\alpha_t(j) = \sum_{i=1}^n \alpha_{t-1}(i) a_{ij} e_j(o_t) = e_j(o_t) \sum_{i=1}^n \alpha_{t-1}(i) a_{ij}$, n 为隐状态数
 - $\alpha_1(j), \alpha_2(j), \dots, \alpha_T(j)$
 - 终止结果
 - $p(O) = \sum_{i=1}^n \alpha_T(i)$: 将最终结点的（不同走法）的概率求和，就是产生目前观测的可能情况

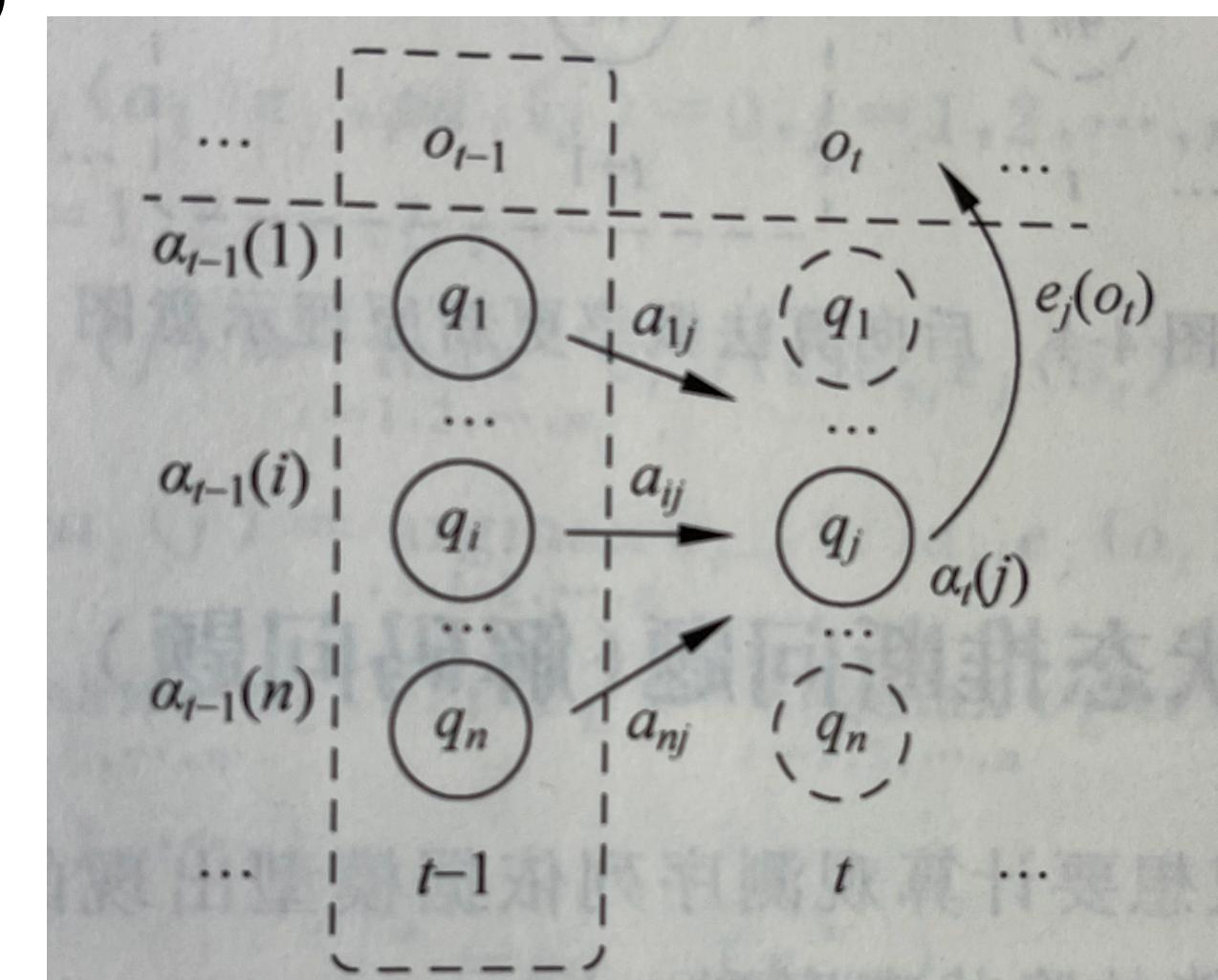


图 4-3 前向算法概率更新原理示意图

4.2.1 HMM模型评估问题

观测序列概率的前向算法

输入：模型 λ ， 观测序列 O

输出：观测序列概率 $P(O|\lambda)$

(1) 初值

$$\alpha_1(i) = \pi_i b_i(o_1), \quad i = 1, 2, \dots, N$$

(2) 递推 对 $t=1,2,\dots,T-1$,

$$\alpha_{t+1}(i) = [\sum_{j=1}^N \alpha_t(j)a_{ji}]b_i(o_{t+1}), \quad i = 1, 2, \dots, N$$

书中的解释：方括号里的值与观测概率 $b_i(o_{t+1})$ 的乘积，就是到时刻 $t+1$ ，观测序列为 $o_1, o_2, \dots, o_t, o_{t+1}$ ，且在时刻 $t+1$ 时处于状态 q_i 的前向概率

(3) 终止：

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

4.2.1 HMM模型评估问题

• 前向算法举例

例 10.2 考虑盒子和球模型 $\lambda = (A, B, \pi)$, 状态集合 $Q = \{1, 2, 3\}$, 观测集合 $V = \{\text{红, 白}\}$,

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

设 $T = 3$, $O = (\text{红, 白, 红})$, 试用前向算法计算 $P(O|\lambda)$.

知乎 @Nash

盒子对应状态集合, 用 $Q = \{1, 2, 3\}$ 表示, 球为观测集合 $V = \{\text{红球, 白球}\}$, $T = 3$ 代表抽了3次, 求在给定的模型 A, B, π 下, 抽中 $\{\text{红, 白, 红}\}$ 的概率

A , A_{11} 表示 上一次在盒子1抽奖 \rightarrow 这一次在盒子1抽奖的概率为0.5

B , 第1行表示盒子1中红白概率各为0.5, 第2行表示盒子2中红白概率四六开, 第三行表示盒子3中红白概率七三开

π , 第1行表示状态1的初始概率为0.2, 第2行表示状态2的初始概率为0.4, 第3行表示状态3的初始概率为0.4

4.2.1 HMM模型评估问题

- 前向算法举例

$$\alpha_t(j) = \sum_{i=1}^n \alpha_{t-1}(i) a_{ij} e_j(o_t) = e_j(o_t) \sum_{i=1}^n \alpha_{t-1}(i) a_{ij}$$

初值

$$\alpha_1(i) = \pi_i b_i(o_1), i = 1, 2, \dots, N$$

- $\alpha_1(i)$ 代表着第1个时刻状态 i 的前向概率；
- π_i 代表着第 i 个状态的初始值； $b_i(o_1)$ 代表着状态 i 下的第1个时刻观测概率 o_1

案例10.2 中，初值是第一次抽球为红球的概率

盒子1,2,3 都可能抽中红球，各自的概率是

状态1（盒子1）为红球的概率 $\alpha_1(1) = \pi_1 b_1(o_1) = 0.2 * 0.5 = 0.1$

状态2（盒子2）为红球的概率 $\alpha_1(2) = \pi_2 b_2(o_1) = 0.4 * 0.4 = 0.16$

状态3（盒子3）为红球的概率 $\alpha_1(3) = \pi_3 b_3(o_1) = 0.4 * 0.7 = 0.28$

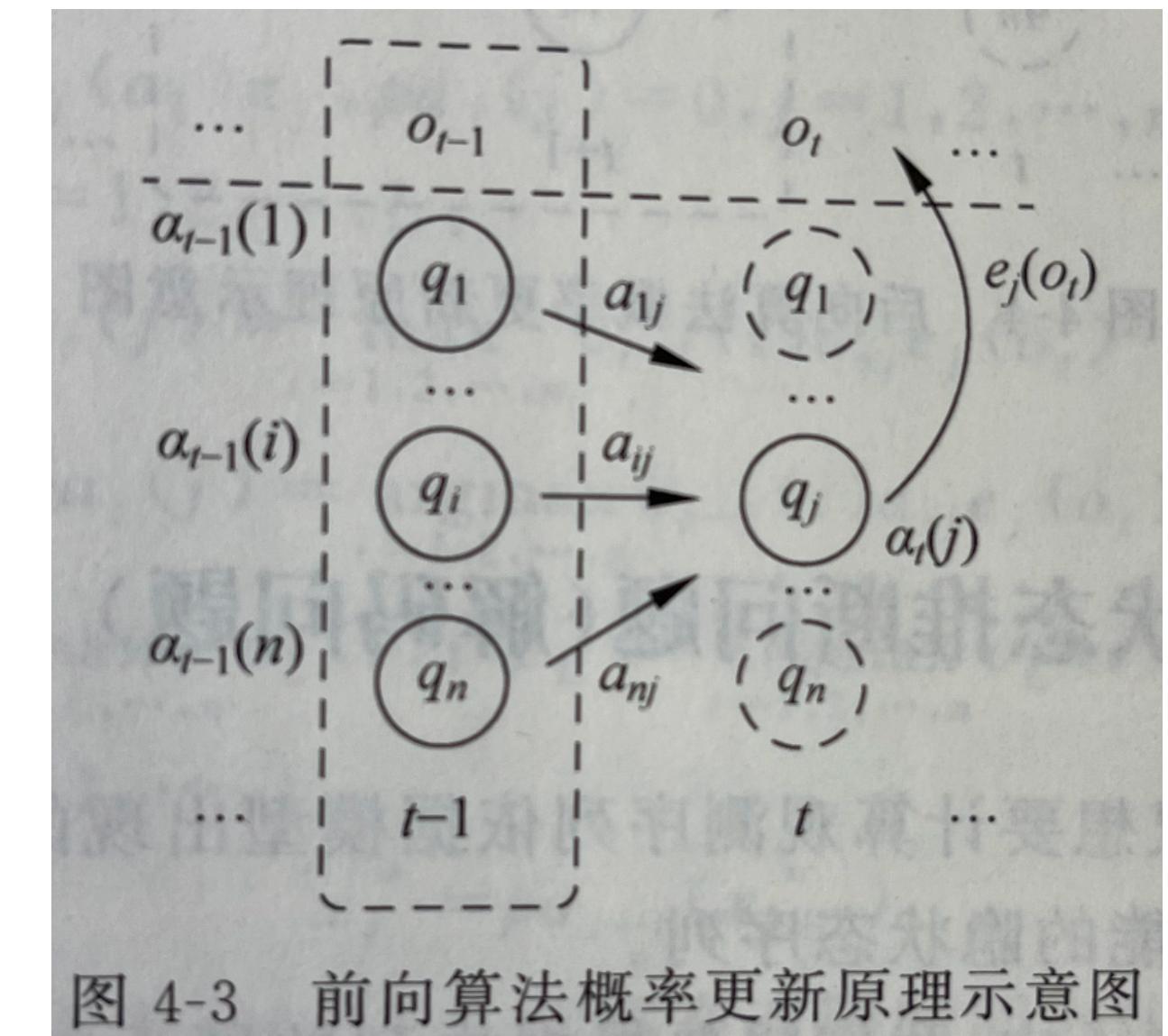


图 4-3 前向算法概率更新原理示意图

4.2.1 HMM模型评估问题

递推

$$\alpha_{t+1}(i) = \left[\sum_{j=1}^N \alpha_t(j) a_{ji} \right] b_i(o_{t+1}), i = 1, 2, \dots, N$$

1. $\alpha_{t+1}(i)$: 第 $t+1$ 时刻下, 第 i 个状态的前向概率
2. $\sum_{j=1}^N \alpha_t(j) a_{ji}$: 第 t 时刻下, 【状态 j 对应的前向概率 * t 时刻 $\rightarrow t+1$ 时刻的状态转移概率】的求和
3. $b_i(o_{t+1})$: 状态 i 下第 $t+1$ 时刻的观测概率 o_{t+1}

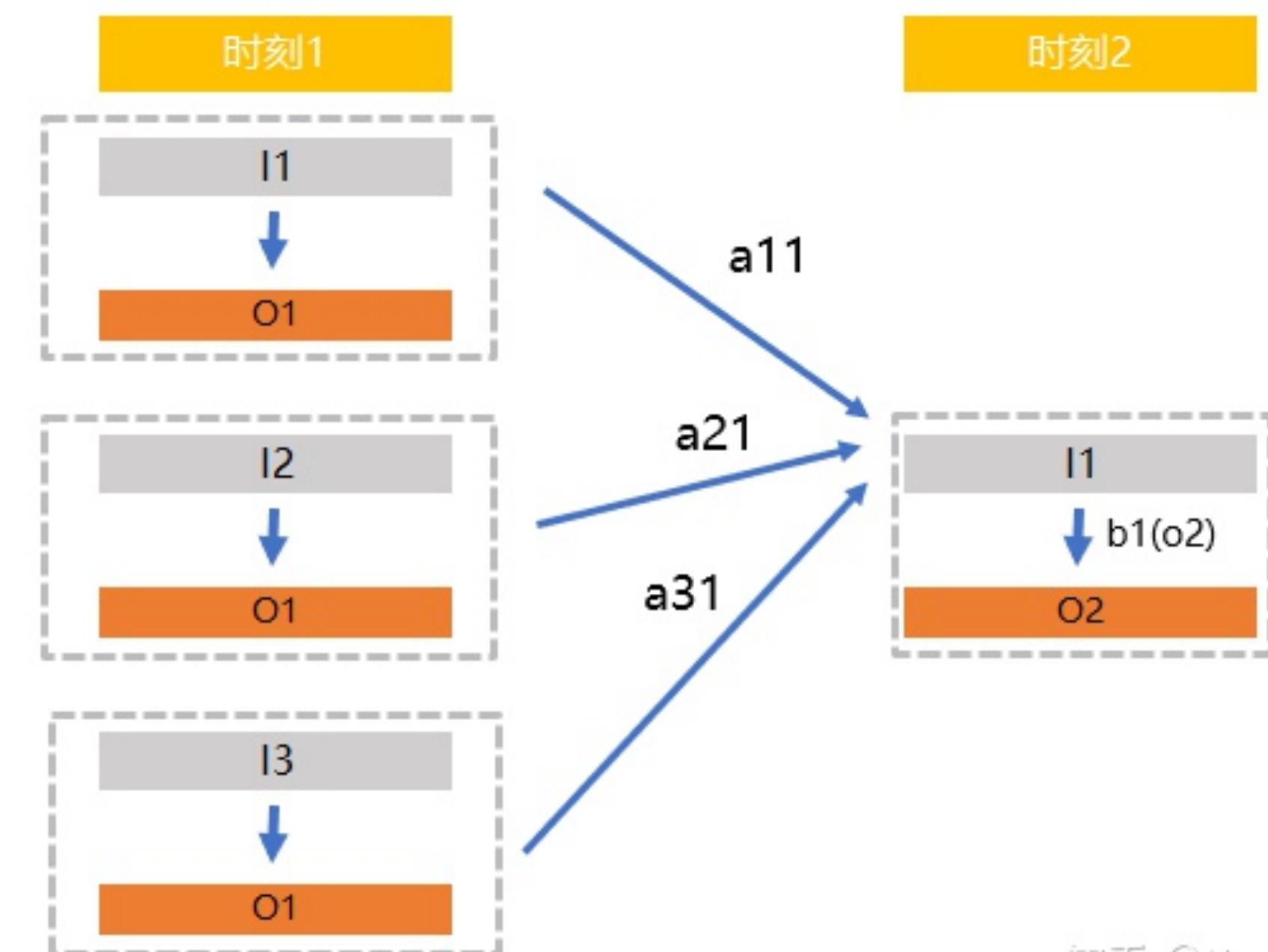
$t+1$ 时刻的状态是由 t 时刻的状态决定的。 t 时刻的状态1、状态2、状态3都可能产生 $t+1$ 时刻的状态1。

以案例10.2来说, 第2次在盒子1中抽中白球, 可能有三种前置场景: 第1次在盒子1中抽中红球、第1次在盒子2中抽中红球和第1次在盒子3中抽中红球。

三种前置场景的前向概率乘上对应的状态转移概率的求和, 就是第 $t+1$ 时刻的状态1的概率。对

应公式部分, 就是 $\sum_{j=1}^N \alpha_t(j) a_{ji}$ 。

求到了状态1的概率后, 再乘上时刻2状态1的观测概率, 就是时刻2, 观测序列 o_1, o_2 , 状态为1的前向概率。



知乎 @Nash

4.2.1 HMM模型评估问题

递推计算

时刻2

$$\alpha_2(1) = [\sum_{i=1}^3 \alpha_1(i)a_{i1}]b_1(o_2)$$

$$= [\alpha_1(1)a_{11} + \alpha_1(2)a_{21} + \alpha_1(3)a_{31}]b_1(o_2)$$

$$= [0.1 * 0.5 + 0.16 * 0.3 + 0.28 * 0.2] * 0.5 = 0.077$$

$$\alpha_2(2) = [\sum_{i=1}^3 \alpha_1(i)a_{i2}]b_2(o_2)$$

$$= [\alpha_1(1)a_{12} + \alpha_1(2)a_{22} + \alpha_1(3)a_{32}]b_2(o_2)$$

$$= [0.1 * 0.2 + 0.16 * 0.5 + 0.28 * 0.3] * 0.6 = 0.1104$$

$$\alpha_2(3) = [\sum_{i=1}^3 \alpha_1(i)a_{i3}]b_3(o_2)$$

$$= [\alpha_1(1)a_{13} + \alpha_1(2)a_{23} + \alpha_1(3)a_{33}]b_3(o_2)$$

$$= [0.1 * 0.3 + 0.16 * 0.2 + 0.28 * 0.5] * 0.3 = 0.0606$$

时刻3

$$\alpha_3(1) = [\sum_{i=1}^3 \alpha_2(i)a_{i1}]b_1(o_3)$$

$$= [\alpha_2(1)a_{11} + \alpha_2(2)a_{21} + \alpha_2(3)a_{31}]b_1(o_3)$$

$$= [0.077 * 0.5 + 0.1104 * 0.3 + 0.0606 * 0.2] * 0.5 = 0.04187$$

$$\alpha_3(2) = [\sum_{i=1}^3 \alpha_2(i)a_{i2}]b_2(o_3)$$

$$= [\alpha_2(1)a_{12} + \alpha_2(2)a_{22} + \alpha_2(3)a_{32}]b_2(o_3)$$

$$= [0.077 * 0.2 + 0.1104 * 0.5 + 0.0606 * 0.3] * 0.4 = 0.03551$$

$$\alpha_3(3) = [\sum_{i=1}^3 \alpha_2(i)a_{i3}]b_3(o_3)$$

$$= [\alpha_2(1)a_{13} + \alpha_2(2)a_{23} + \alpha_2(3)a_{33}]b_3(o_3)$$

$$= [0.077 * 0.3 + 0.1104 * 0.2 + 0.0606 * 0.5] * 0.7 = 0.05284$$

$$\alpha_t(j) = \sum_{i=1}^n \alpha_{t-1}(i)a_{ij}e_j(o_t) = e_j(o_t) \sum_{i=1}^n \alpha_{t-1}(i)a_{ij}$$

终止

$$p(O|\lambda) = \sum_{i=1}^3 \alpha_3(i)$$

$$= \alpha_3(1) + \alpha_3(2) + \alpha_3(3) = 0.13022$$

4.2.1 HMM模型评估问题

- 后向（向后）算法：Backward Algorithm
 - 已知结果，是当初怎么“走”造成的？（追因）
 - 从后往前求解： t 时刻隐状态 $x_t = q_j$ 时，观察到后续观测值 $o_{t+1} o_{t+2} \dots o_L$ 的概率 $\beta_t(j)$
 - $\beta_t(j) = p(o_{t+1} o_{t+2} \dots o_L | x_t = q_j)$: t 时刻，隐状态为 q_j 时，观测子序列(t+1 - T)出现的概率
 - $\beta_t(j) = \sum_{i=1}^n \beta_{t+1}(i) e_i(o_{t+1}) a_{ji}$
 - $\beta_L(j) = 1, \beta_{T-1}(j), \dots, \beta_1(j)$
 - 终止结果
 - $p(O) = \sum_{i=1}^n \pi_i e_i(o_1) \beta_1(i)$

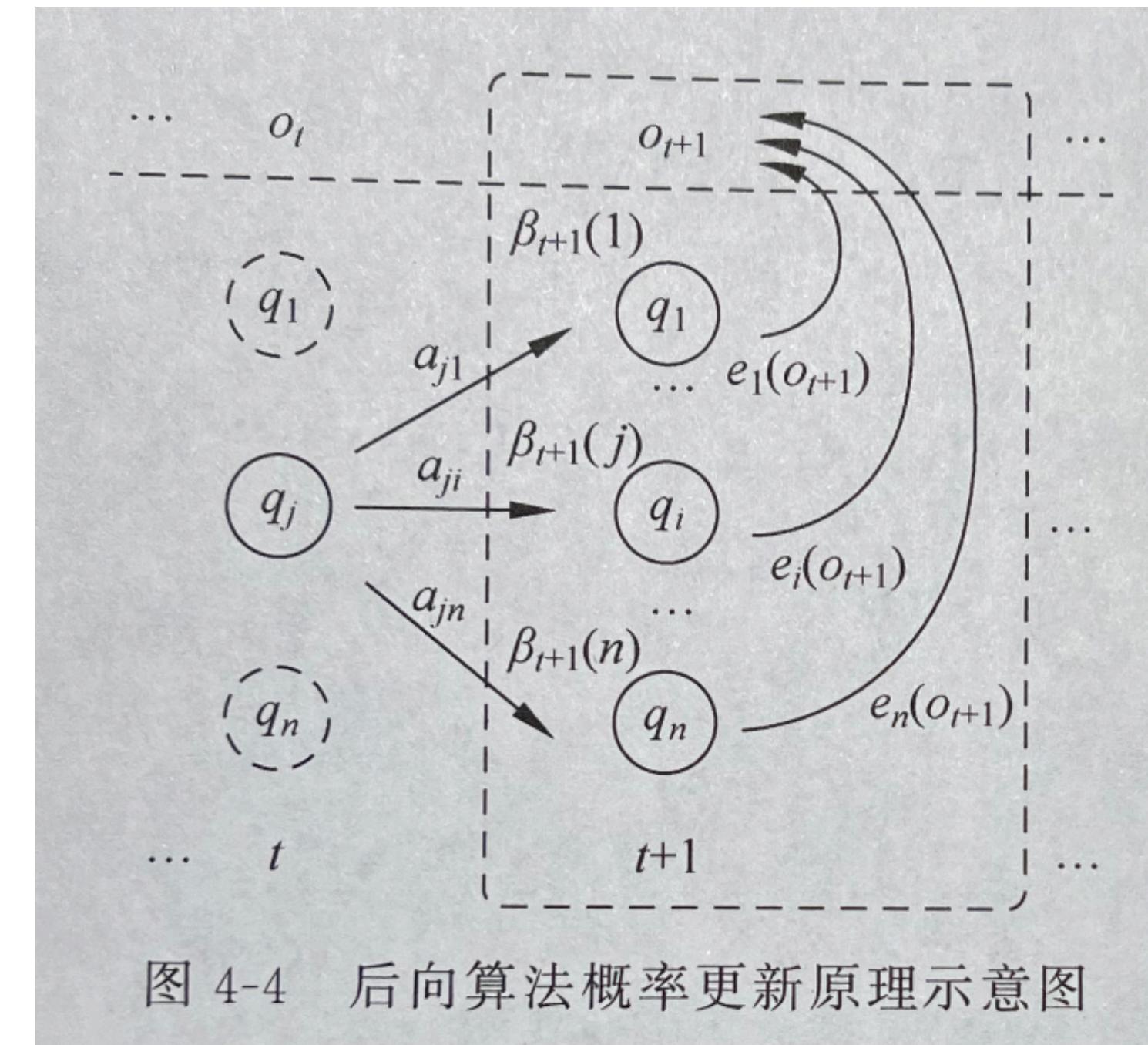


图 4-4 后向算法概率更新原理示意图

4.2.1 HMM模型评估问题

观测序列概率的后向算法

输入：隐马模型 λ ， 观测序列 O

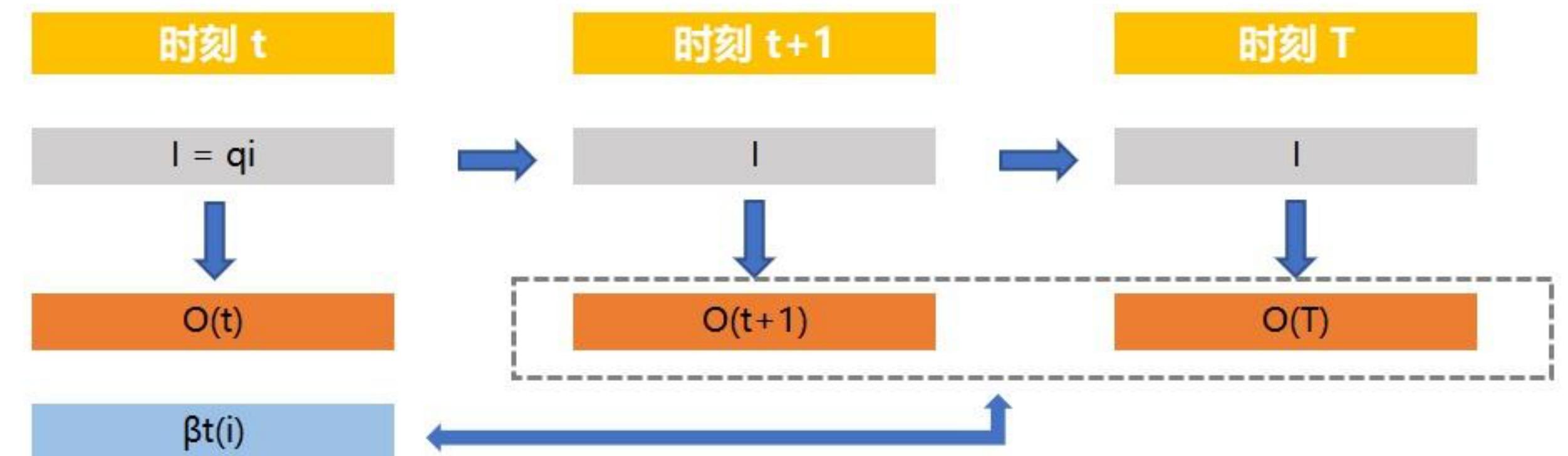
输出：观测序列概率 $P(O|\lambda)$

$$(1) \beta_T(i) = 1, i = 1, 2, \dots, N$$

$$(2) \text{ 对 } t = T - 1, T - 2, \dots, 1$$

$$\beta_t(i) = \sum_{j=1}^N a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) i = 1, 2, \dots, N$$

$$(3) P(O|\lambda) = \sum_{i=1}^N \pi_i b_i(o_1) \beta_1(i)$$



4.2.1 HMM模型评估问题

- 后向算法举例

给定模型 $\lambda = (A, B, \pi)$ 和观测序列 $O = (o_1, o_2, \dots, o_T)$ ，计算在模型 λ 下观测序列 O 出现的概率 $P(O|\lambda)$

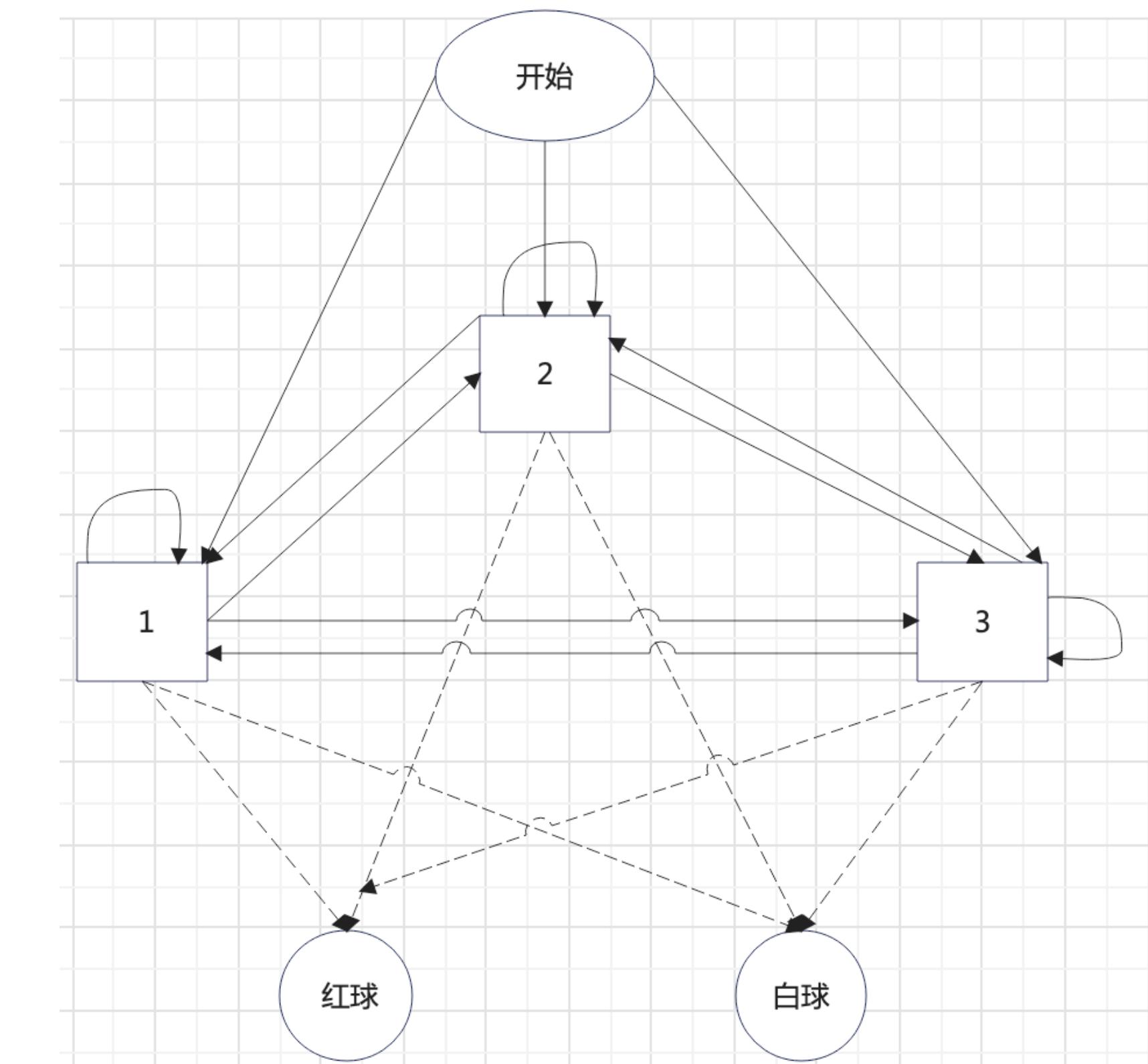
还是前面文章的那个例子，同样的条件

$$A = \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{bmatrix}$$

$$B = \begin{bmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \\ 0.7 & 0.3 \end{bmatrix}$$

$$\pi = (0.2, 0.4, 0.4)^T$$

求给定这个模型条件下，观测为 $O = (\text{红}, \text{白}, \text{红})$ 的概率



4.2.1 HMM模型评估问题

- 后向算法举例

初值

$$\beta_T(i) = 1, i = 1, 2, \dots, N$$

最终时刻 T，所有状态的后向概率都规定为1。

书中案例10.2 为例计算下

初值是第3次，三个状态（3个盒子）抽中红球的后向概率都为1（其实也没有后向序列了）

状态1（盒子1） $\beta_{T=3}(1) = 1$

状态2（盒子2） $\beta_{T=3}(2) = 1$

状态3（盒子3） $\beta_{T=3}(3) = 1$

4.2.1 HMM模型评估问题

- 后向算法举例

递推

对 $t = T - 1, T - 2, \dots, 1$

$$\beta_t(i) = \sum_{j=1}^N a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) i = 1, 2, \dots, N$$

1. $\beta_t(i)$ 是指比如 $T-1$ 时刻下第 i 个状态的后向概率；

$$2. \sum_{j=1}^N a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) :$$

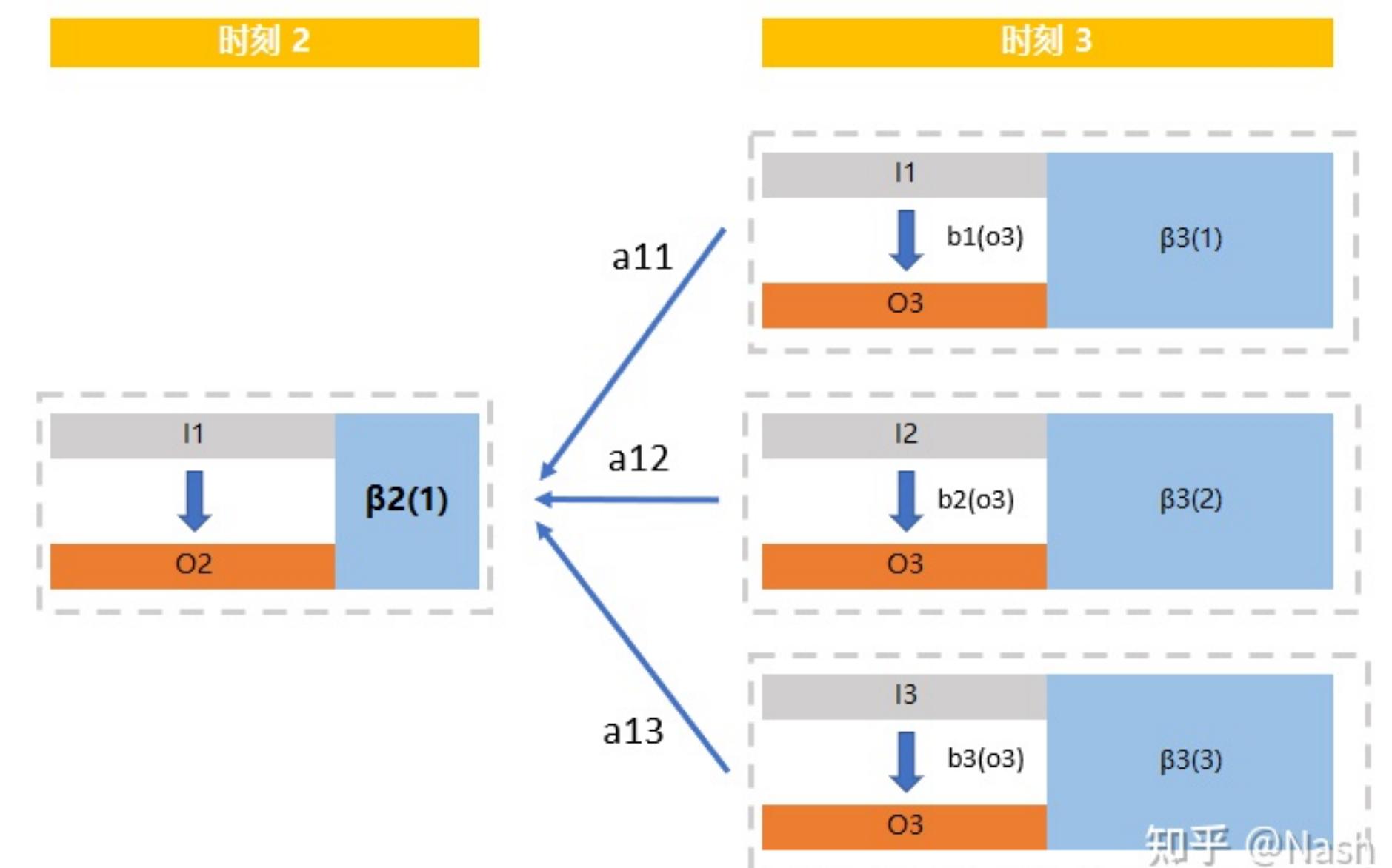
- j 对应比如 T 时刻的状态， i 对应比如 $T-1$ 时刻的状态

- a_{ij} 是时刻 $T-1 \rightarrow$ 时刻 T 的状态转移概率

- $b_j(o_{t+1})$ 是时刻 T 下状态 j 观测概率

- $\beta_{t+1}(j)$ 是时刻 T 下状态 j 的后向概率

- 总结下就是时刻 T 下【所有状态 j 的状态转移概率 * 该状态的观测概率 * 该状态的后向概率】的求和



4.2.1 HMM模型评估问题

第2次抽球 $t = T - 1 = 2$

$$\beta_{t=2}(1) = P(o_2 = \text{白}, o_3 = \text{红} | 1, \lambda) = \sum_{j=1}^3 a_{1j} b_j(o_3) \beta_3(j)$$

$$= a_{11} b_1(o_3) \beta_3(1) + a_{12} b_2(o_3) \beta_3(2) + a_{13} b_3(o_3) \beta_3(3)$$

$$= 0.5 * 0.5 * 1 + 0.2 * 0.4 * 1 + 0.3 * 0.7 * 1 = 0.54$$

$$\beta_2(2) = P(o_2 = \text{白}, o_3 = \text{红} | 2, \lambda) = \sum_{j=1}^3 a_{2j} b_j(o_3) \beta_3(j)$$

$$= a_{21} b_1(o_3) \beta_3(1) + a_{22} b_2(o_3) \beta_3(2) + a_{23} b_3(o_3) \beta_3(3)$$

$$= 0.3 * 0.5 + 0.5 * 0.4 + 0.2 * 0.7 = 0.49$$

$$\beta_2(3) = P(o_2 = \text{白}, o_3 = \text{红} | 3, \lambda) = \sum_{j=1}^3 a_{3j} b_j(o_3) \beta_3(j)$$

$$= a_{31} b_1(o_3) \beta_3(1) + a_{32} b_2(o_3) \beta_3(2) + a_{33} b_3(o_3) \beta_3(3)$$

$$= 0.2 * 0.5 + 0.3 * 0.4 + 0.5 * 0.7 = 0.57$$

第1次抽球 $t = T - 2 = 1$

$$\beta_{t=1}(1) = P(o_1 = \text{红}, o_2 = \text{白}, o_3 = \text{红} | 1, \lambda) = \sum_{j=1}^3 a_{1j} b_j(o_2) \beta_2(j)$$

$$= a_{11} b_1(o_2) \beta_2(1) + a_{12} b_2(o_2) \beta_2(2) + a_{13} b_3(o_2) \beta_2(3)$$

$$= 0.5 * 0.5 * 0.54 + 0.2 * 0.6 * 0.49 + 0.3 * 0.3 * 0.57 = 0.2451$$

$$\beta_1(2) = P(o_1 = \text{红}, o_2 = \text{白}, o_3 = \text{红} | 2, \lambda) = \sum_{j=1}^3 a_{2j} b_j(o_2) \beta_2(j)$$

$$= a_{21} b_1(o_2) \beta_2(1) + a_{22} b_2(o_2) \beta_2(2) + a_{23} b_3(o_2) \beta_2(3)$$

$$= 0.3 * 0.5 * 0.54 + 0.5 * 0.6 * 0.49 + 0.2 * 0.3 * 0.57 = 0.2622$$

$$\beta_1(3) = P(o_1 = \text{红}, o_2 = \text{白}, o_3 = \text{红} | 3, \lambda) = \sum_{j=1}^3 a_{3j} b_j(o_2) \beta_2(j)$$

$$= a_{31} b_1(o_2) \beta_2(1) + a_{32} b_2(o_2) \beta_2(2) + a_{33} b_3(o_2) \beta_2(3)$$

$$= 0.2 * 0.5 * 0.54 + 0.3 * 0.6 * 0.49 + 0.5 * 0.3 * 0.57 = 0.2277$$

$$p(o|\lambda) = \sum_{i=1}^3 \pi_i b_i(o_1) \beta_1(i) = \pi_1 b_1(o_1) \beta_1(1) + \pi_2 b_2(o_1) \beta_1(2) + \pi_3 b_3(o_1) \beta_1(3)$$

$$= 0.2 * 0.5 * 0.2451 + 0.4 * 0.4 * 0.2622 + 0.4 * 0.7 * 0.2277 = 0.130218$$

4.2.1 HMM模型评估问题

- 前向算法和后向算法统一

$$P(O|M) = \sum_{i=1}^n \sum_{j=1}^n \alpha_i(i) a_{ij} e_j(o_{t+1}) \beta_{t+1}(j), t = 1, 2, \dots, L-1$$

4.2.2 HMM隐状态推断问题

已知HMM模型 M 和观测序列 O , 求产生该序列的最有可能的隐状态序列 $x = \arg \max_x p(x, O | M)$

- 维特比 (Viterbi) 算法
 - 一种利用动态规划求解最优路径的算法, 一条路径对应一个状态序列
 - 递推计算在每一时刻各条部分路径的最大概率

$$v_t(j) = \max_{x_1 \cdots x_{t-1}} p(x_1 \cdots x_{t-1}, o_1 \cdots o_t, x_t = q_j | M)$$

$$\cdot v_t(j) = \max_{i=1:n} v_{t-1}(i)a_{ij}e_j(o_t)$$

$$\cdot pa_t(j) = \arg \max_{i=1:n} v_{t-1}(i)a_{ij}e_j(o_i)$$

- 记录最大值对应的隐状态路径

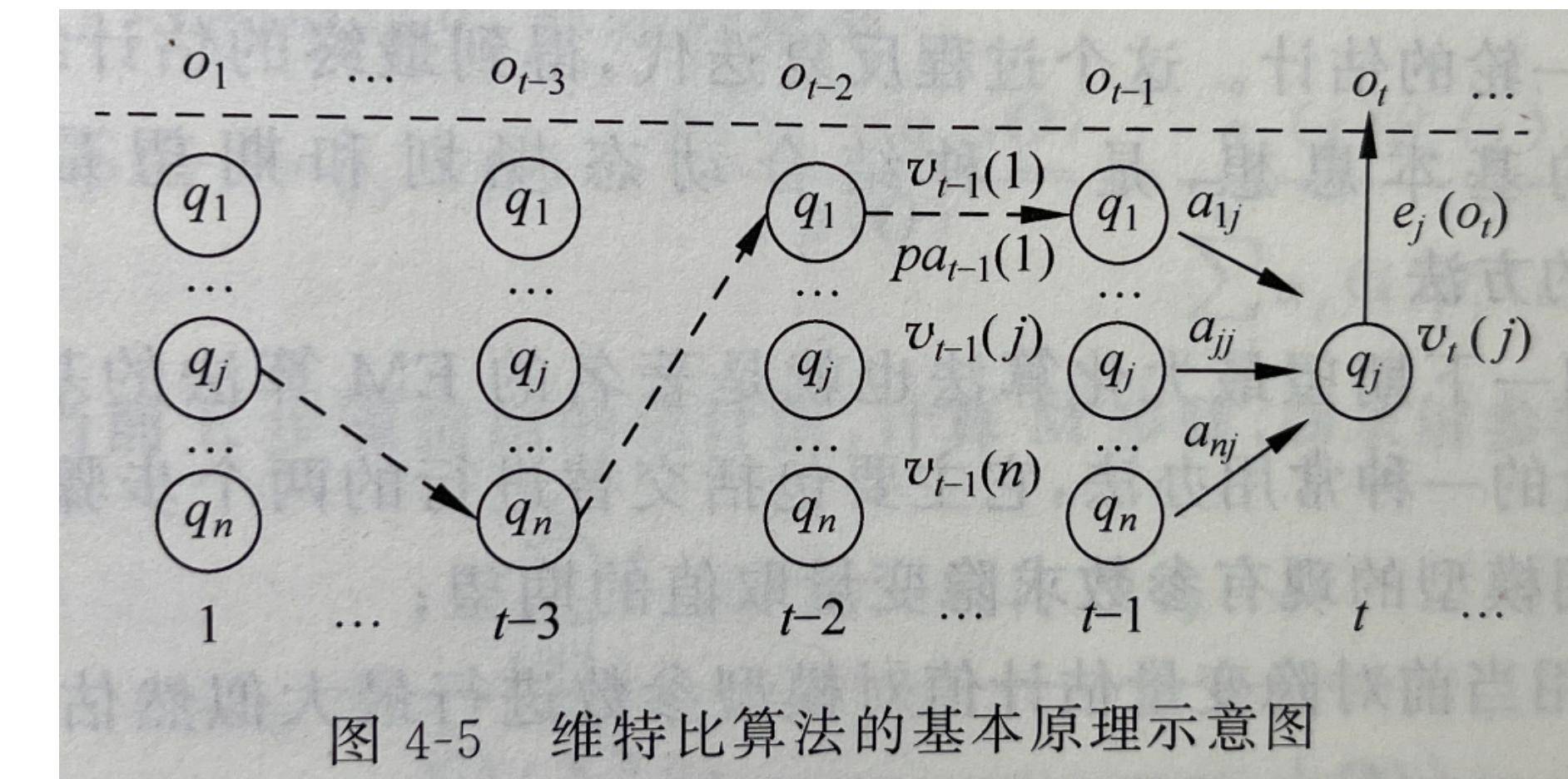
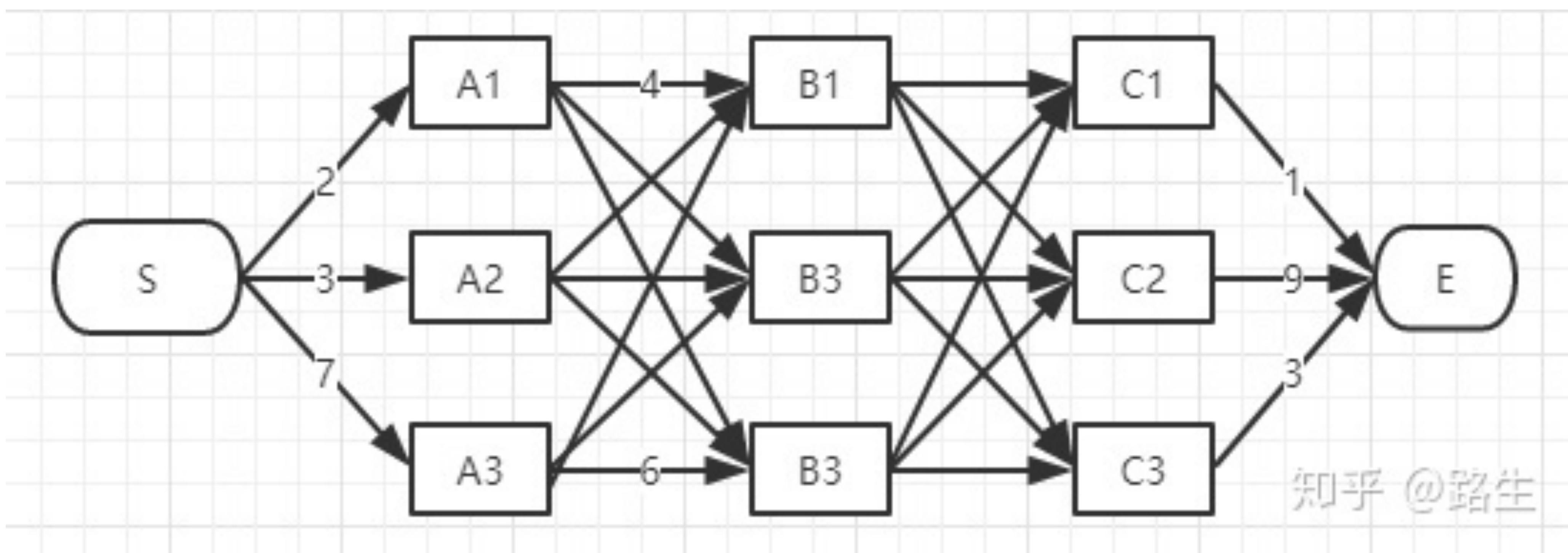
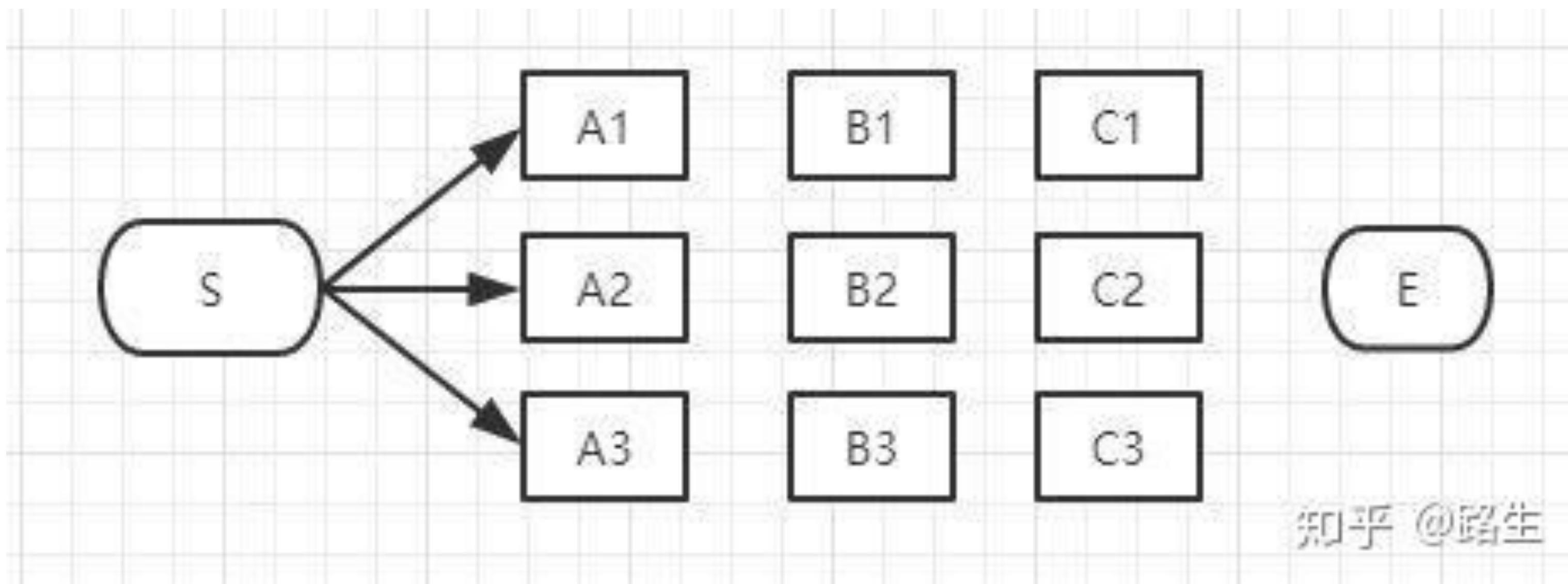


图 4-5 维特比算法的基本原理示意图

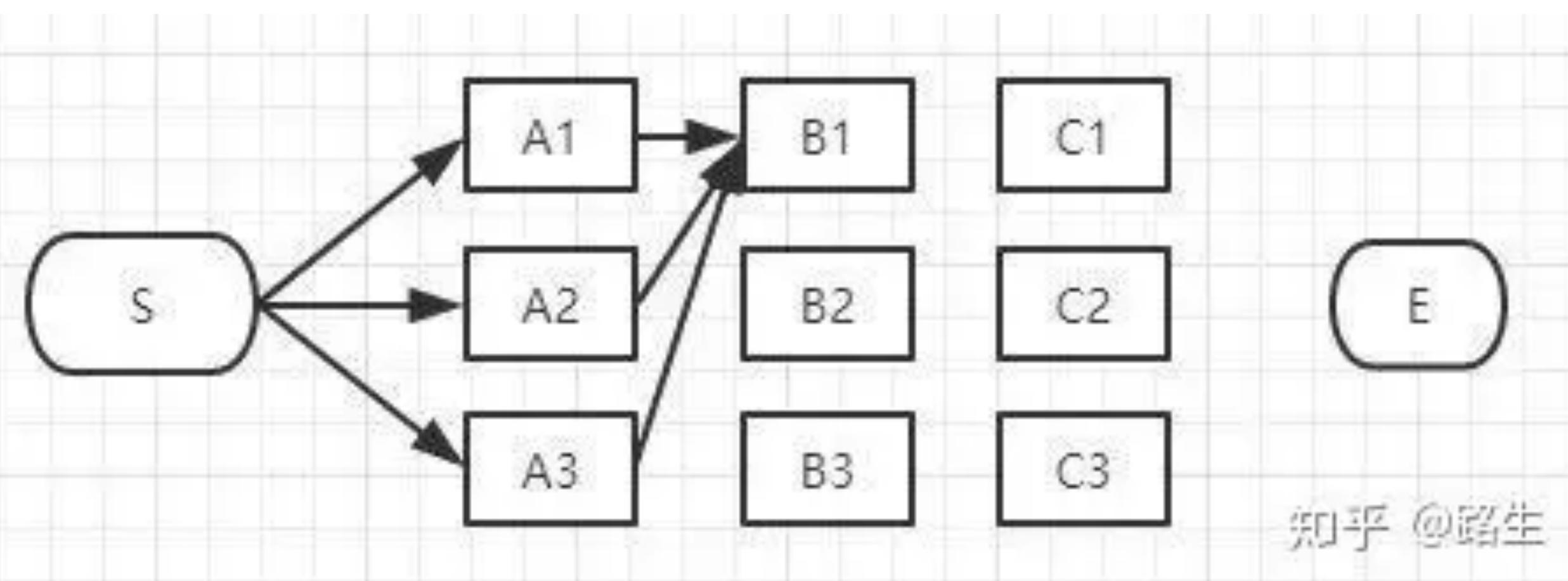
4.2.2 HMM隐状态推断问题



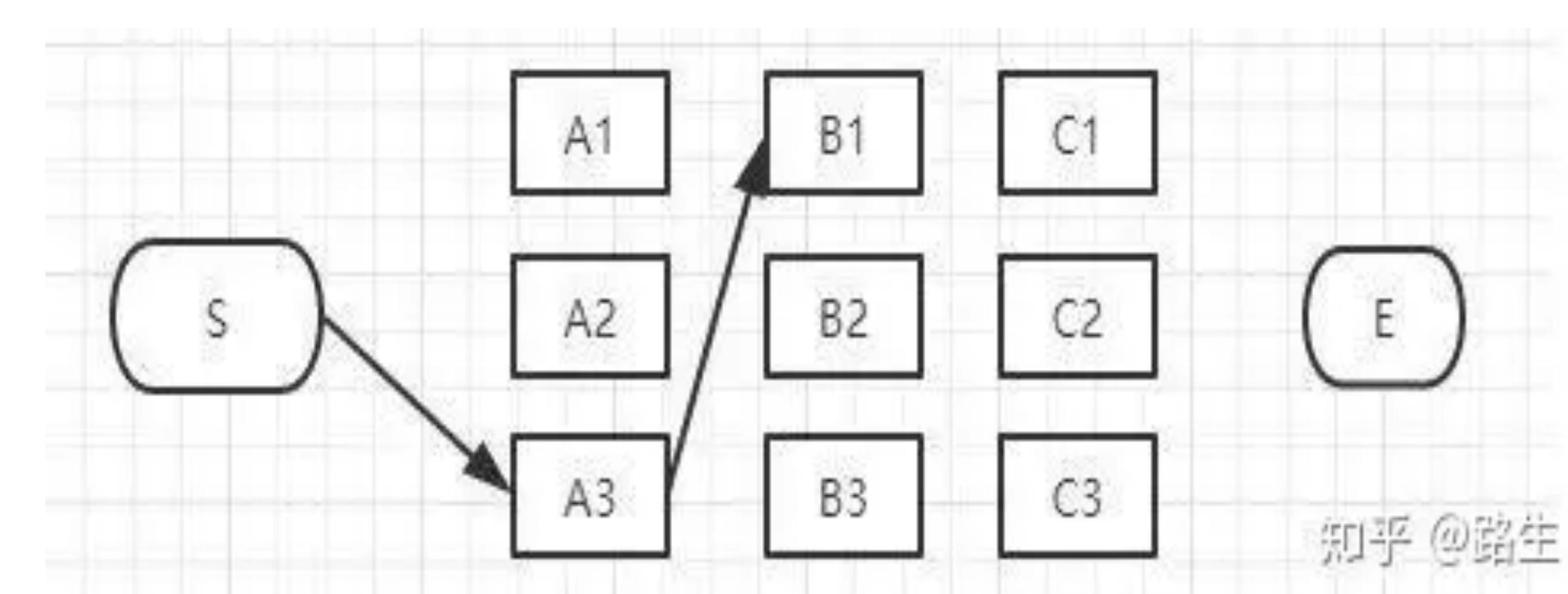
4.2.2 HMM隐状态推断问题



知乎 @路生

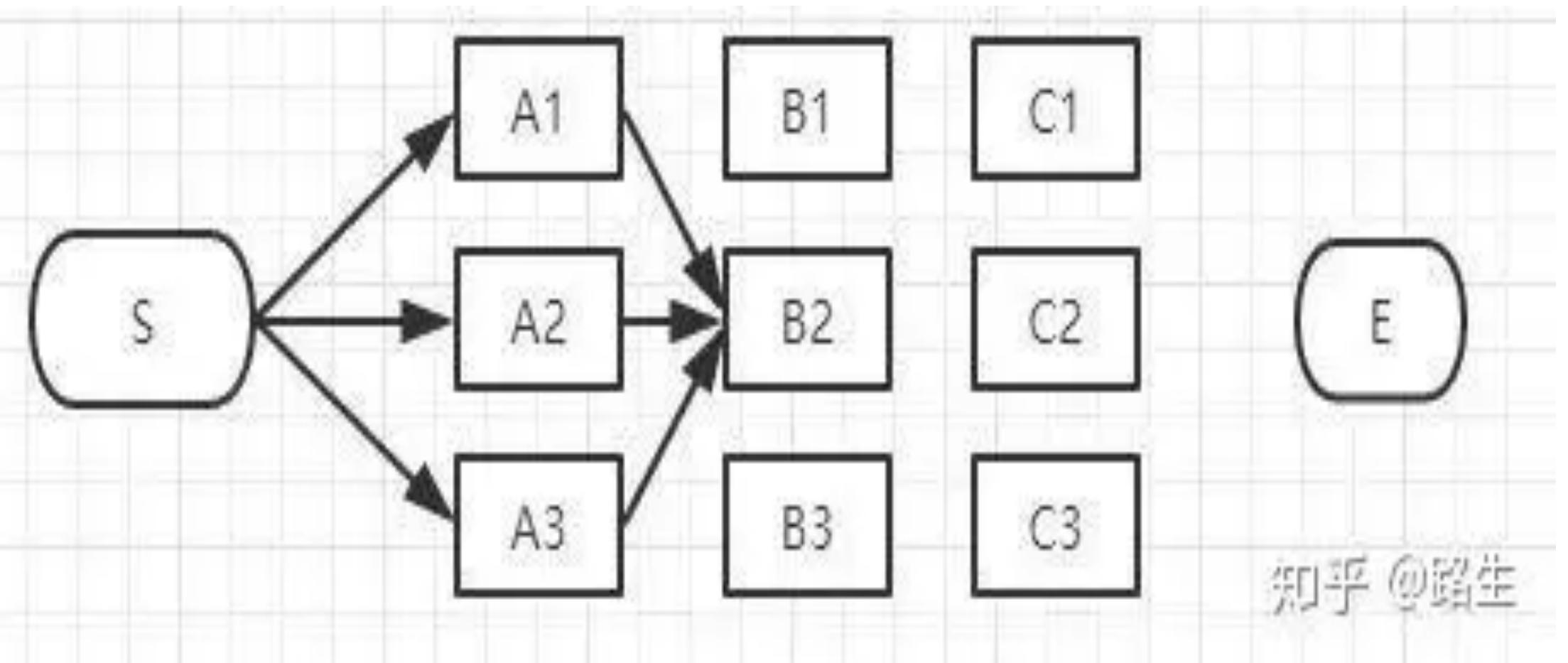


知乎 @路生

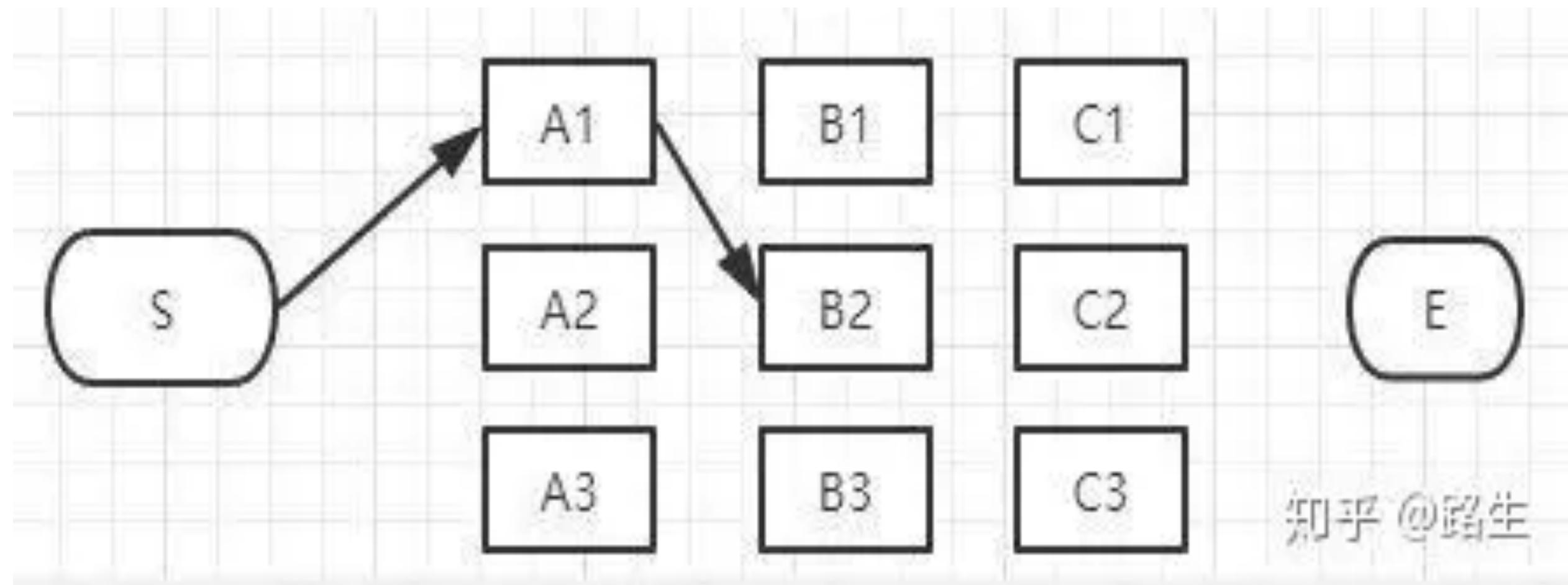


知乎 @路生

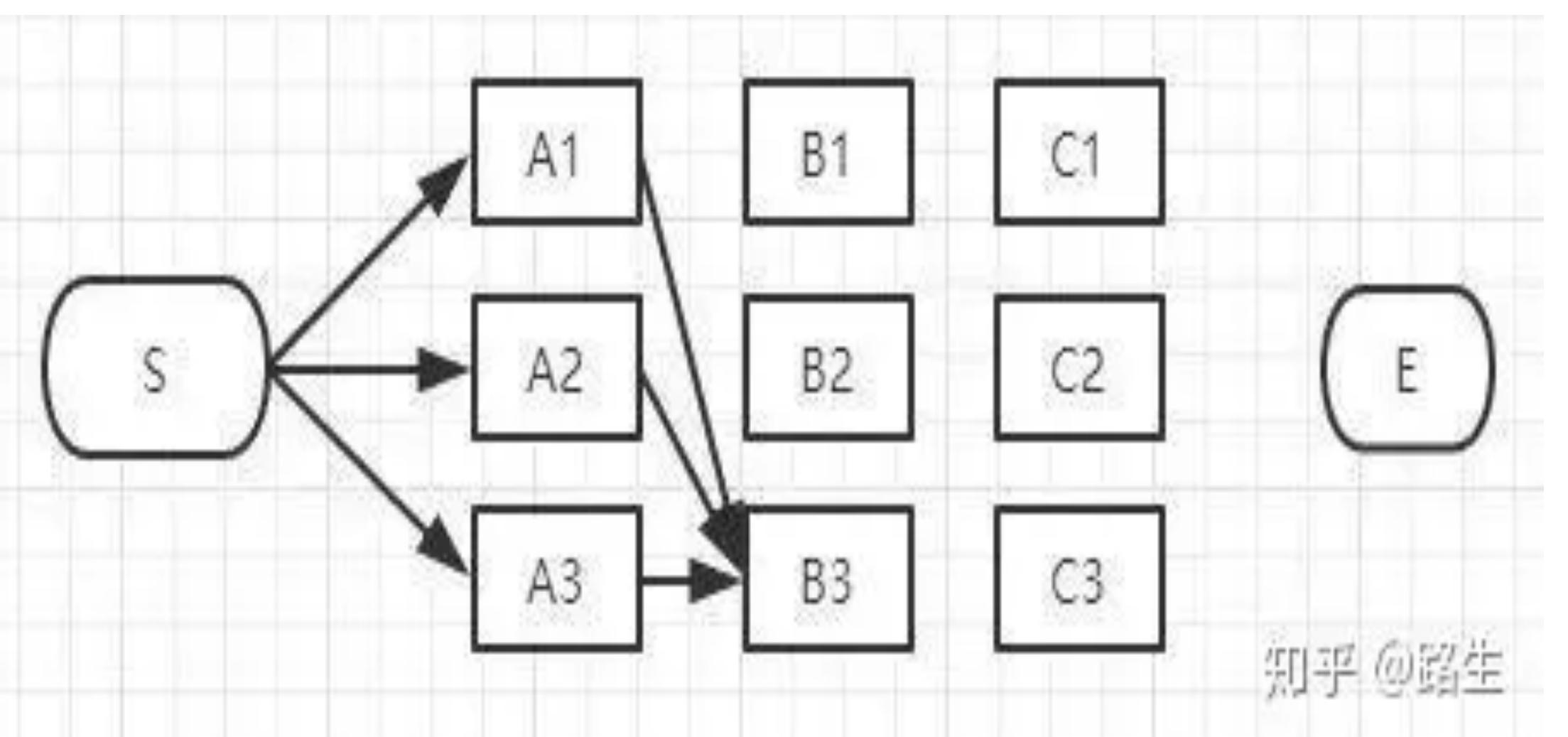
4.2.2 HMM隐状态推断问题



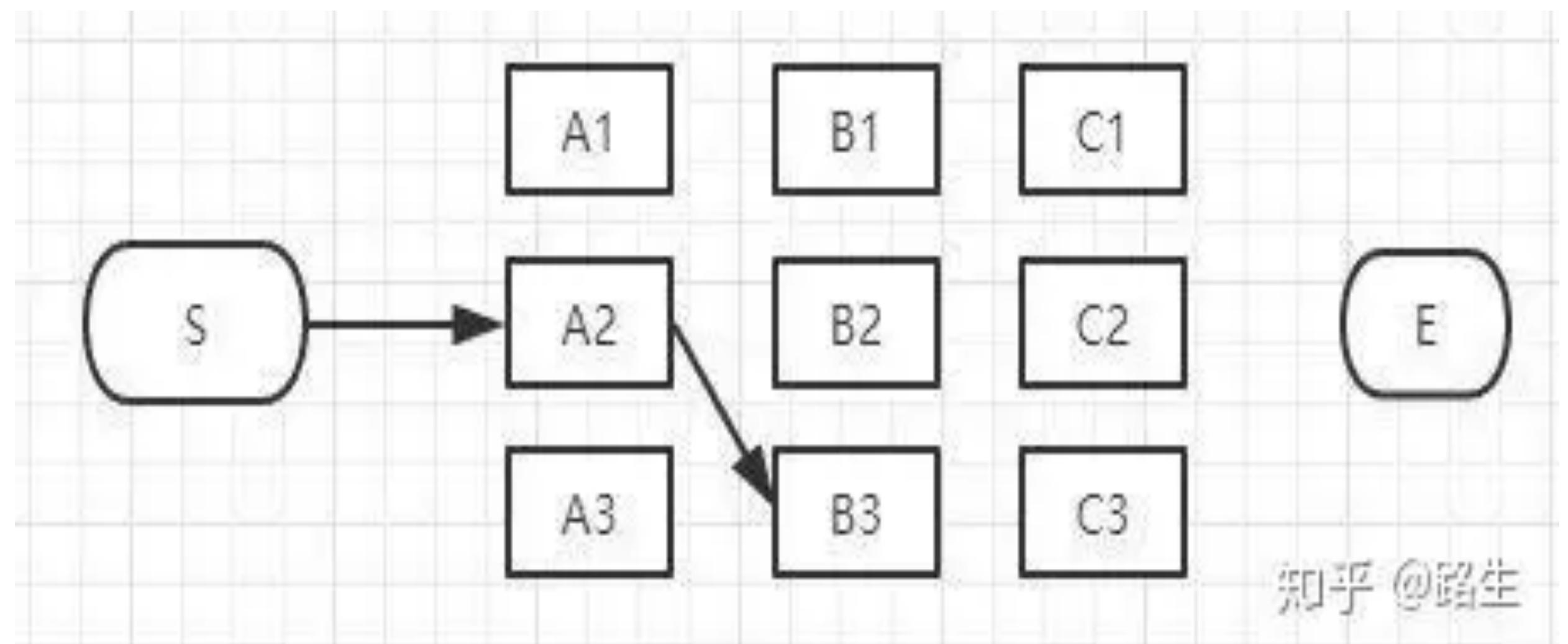
知乎 @路生



知乎 @路生

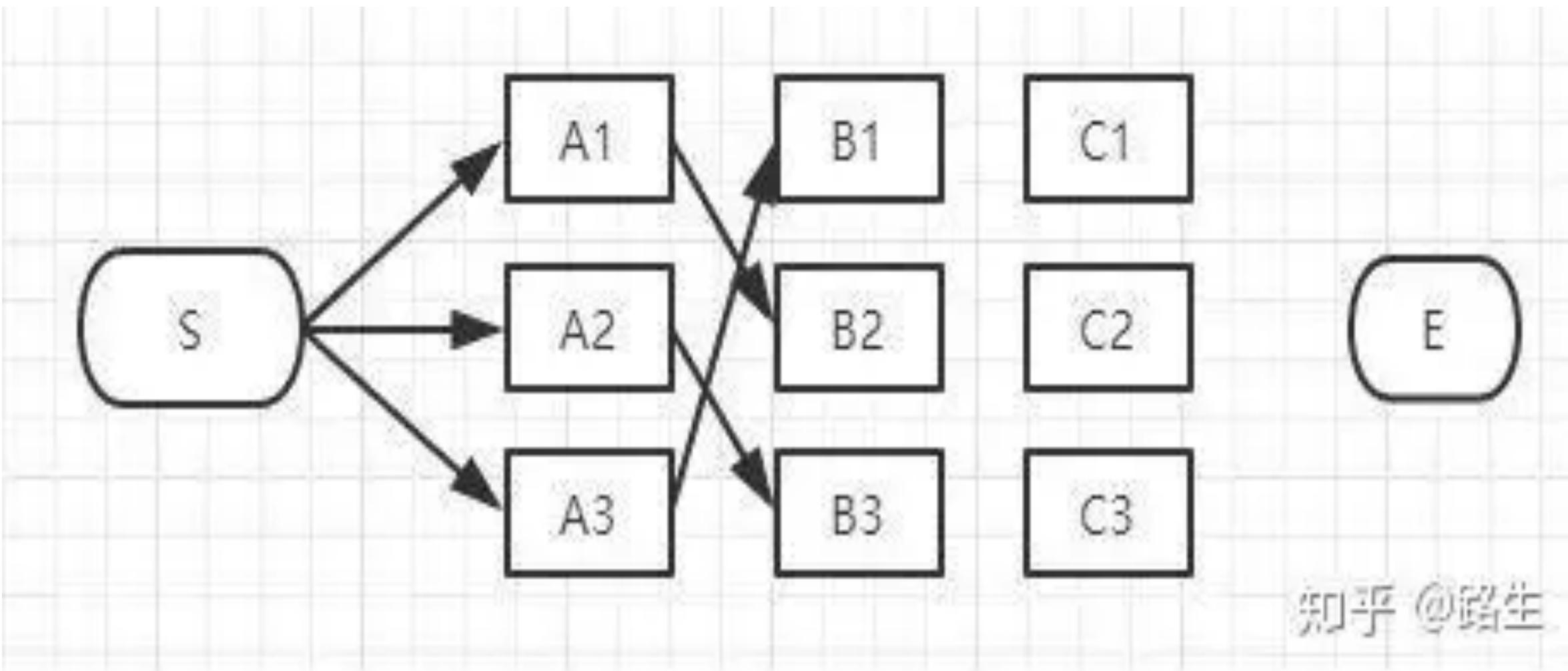


知乎 @路生

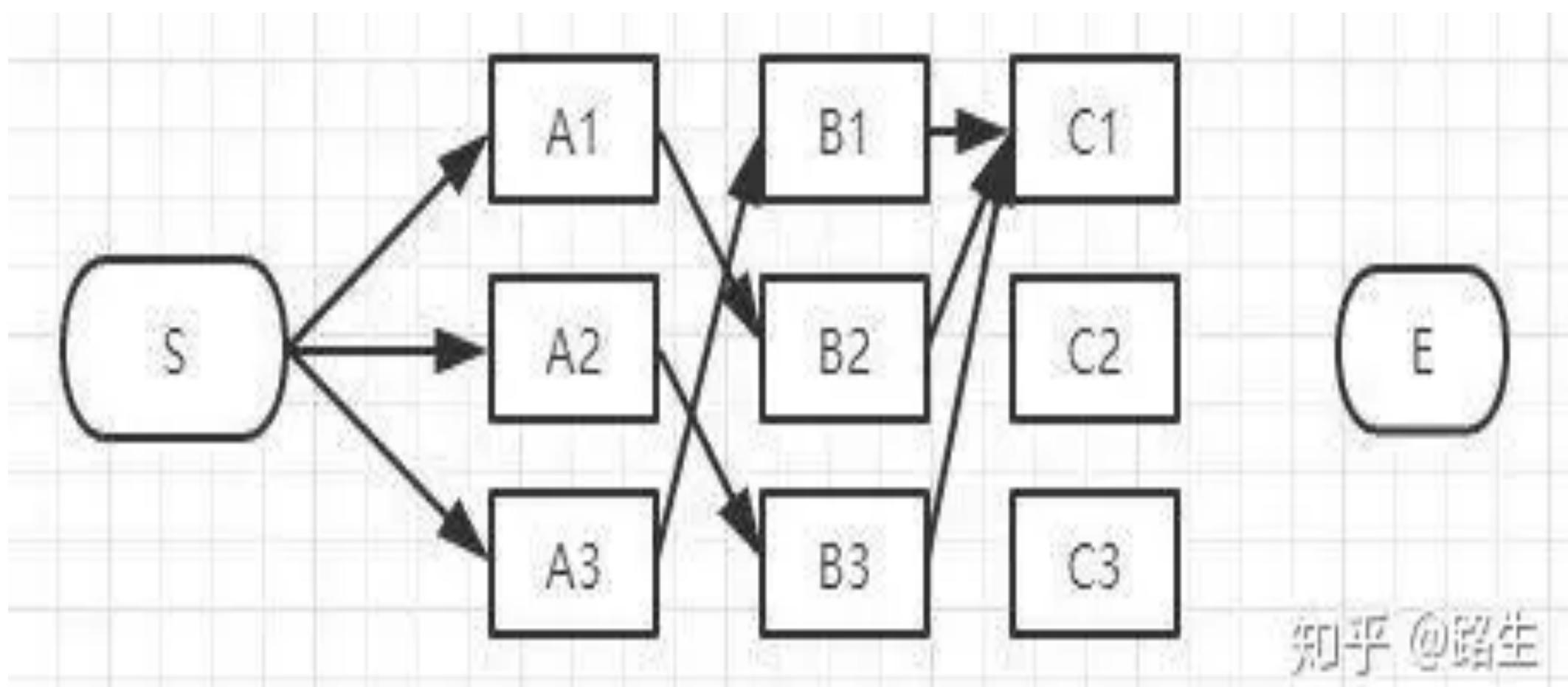


知乎 @路生

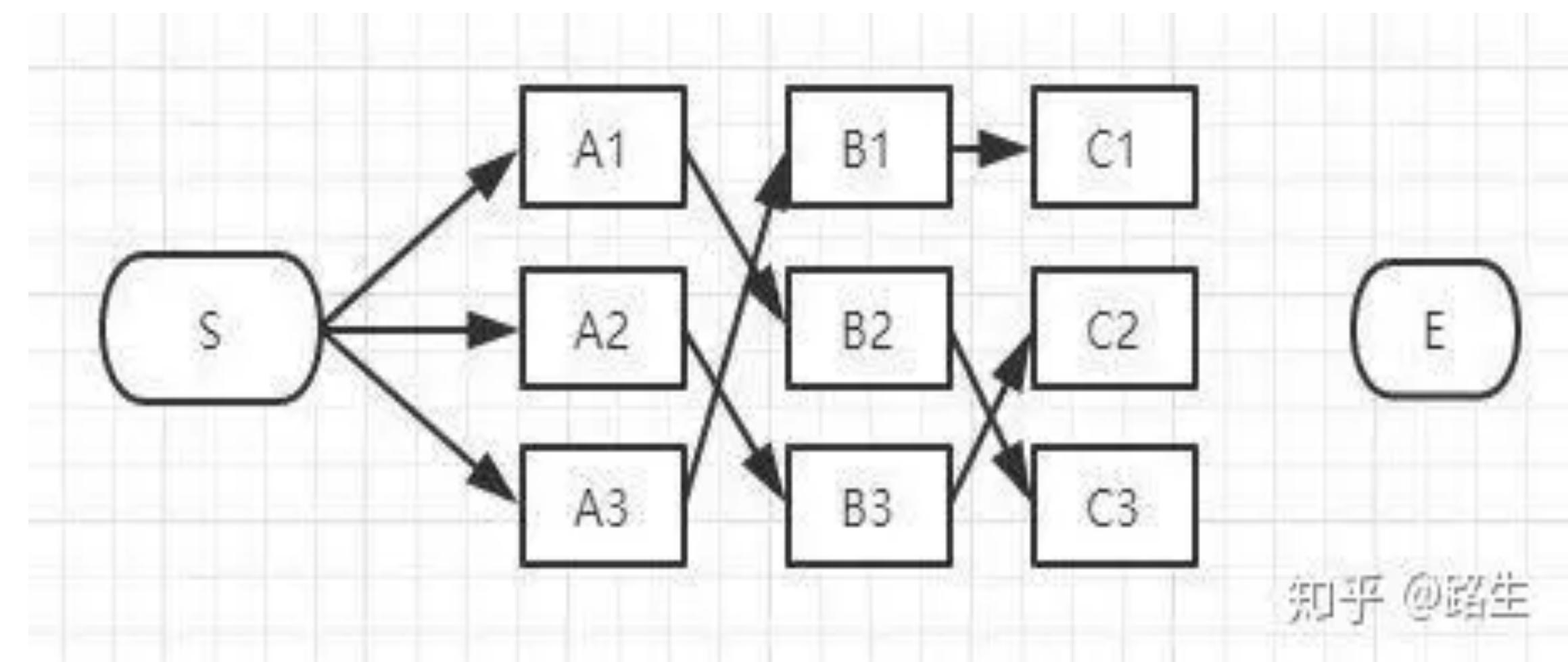
4.2.2 HMM隐状态推断问题



知乎 @路生

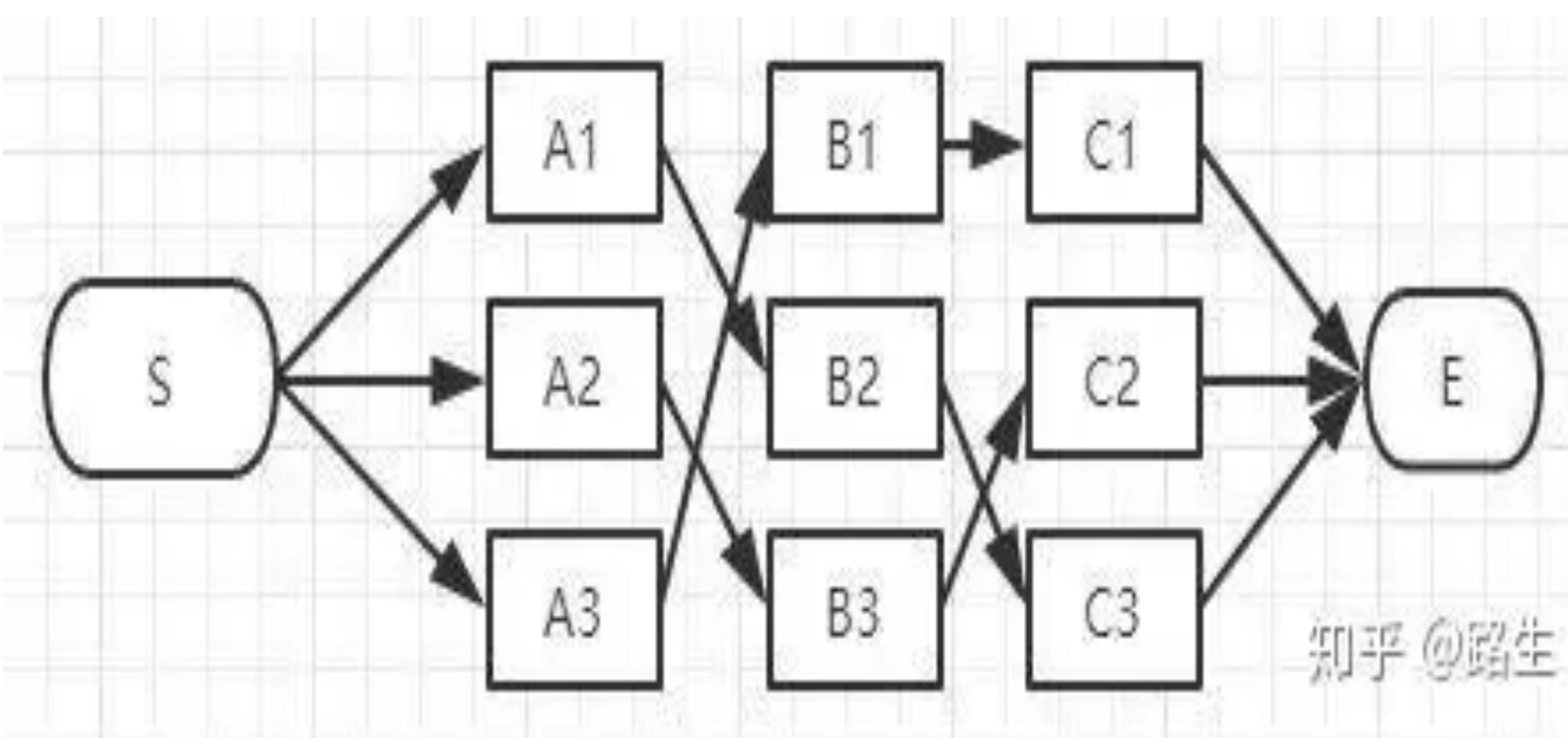


知乎 @路生

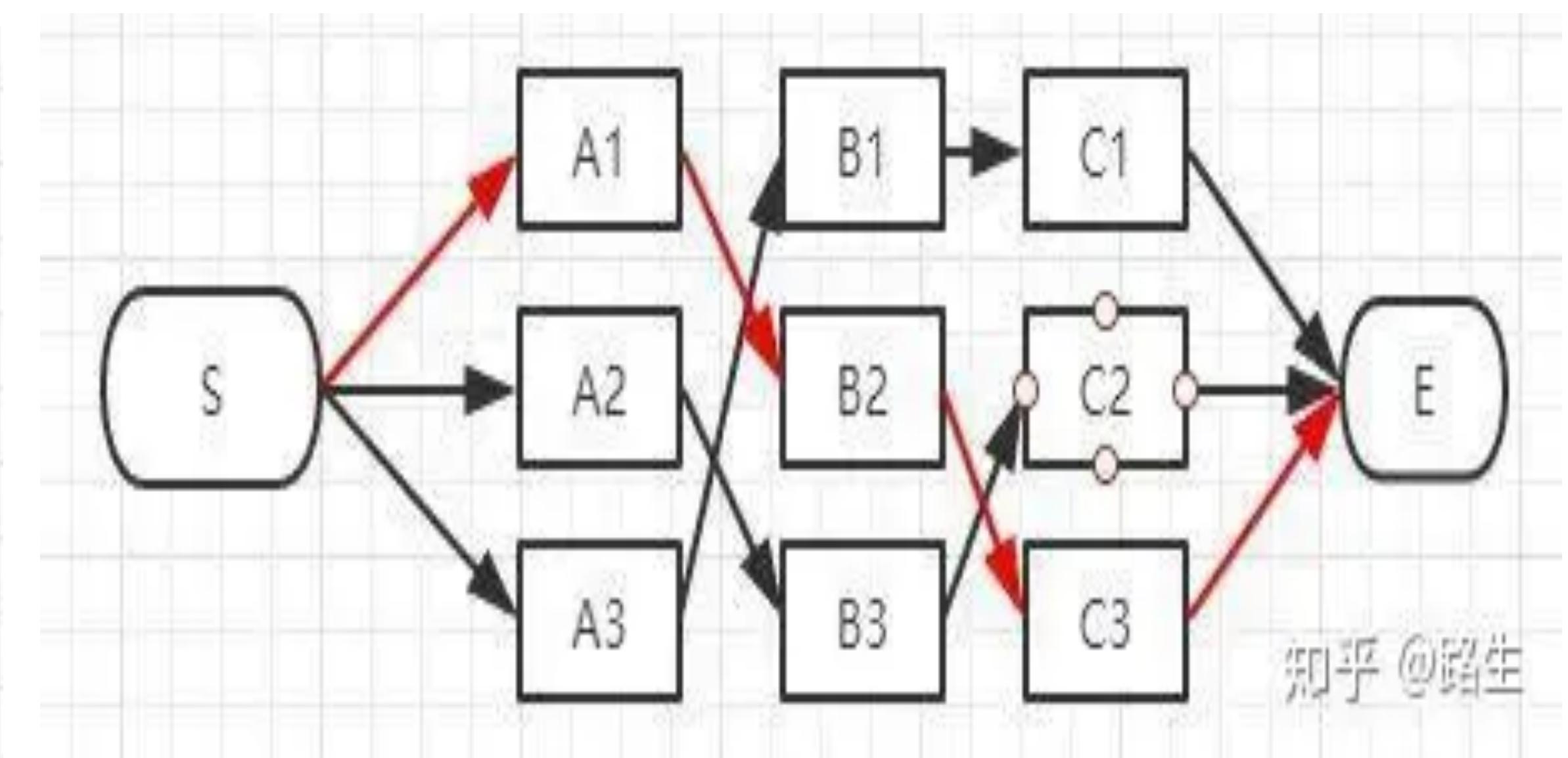


知乎 @路生

4.2.2 HMM隐状态推断问题



知乎 @路生



知乎 @路生

4.2.3 HMM模型学习问题

已知模型结构、观测值取值范围 V 、模型隐状态集 Q ，根据观测序列 O 的样本，学习模型参数 A 、 E 、 π 。

- 隐状态序列 x 已知：最大似然估计法

$$\hat{a}_{ij} = \frac{\#T_{ij}}{\sum_{j'}^n \#T_{ij'}}, \quad \hat{e}_{ik} = \frac{\#E_{ik}}{\sum_{k'}^V \#E_{ik'}}, \quad \hat{\pi}_i = \frac{\#S_i}{S}$$

- $\#T_{ij}$: 隐状态序列从状态 q_i 变为 q_j 的次数
- $\#E_{ik}$: 隐状态为 q_i 时观测值为 V_k 的次数
- $\#S_i$: 初始状态为 q_i 的次数， S 为总样本数

- 隐状态序列 x 未知：最大方差估计法（Expectation-Maximum, EM）

4.2.3 HMM模型学习问题

我们有两枚硬币(coin A & coin B)，这两枚硬币是用特殊材质做的，硬币A抛出正面 (Head)和反面(Tail)的概率为 θ_A 和 $1-\theta_A$ ，硬币B抛出正面和反面的概率为 θ_B 和 $1-\theta_B$ 。我们不知道 θ_A 和 θ_B ，因此想通过不断的抛硬币来推测出 θ_A 和 θ_B ，为了方便，写成向量形式：

$$\theta = (\theta_A, \theta_B)。$$

因为有两枚硬币，我们随机地在硬币A和硬币B中挑一个(概率相等，各为50%)，然后再用选中的硬币独立地抛10次，为了使整个事件更具说服力，我们选硬币抛硬币的整个过程重复做了5次。因此，总的来说选了5次硬币，抛了 $5 \times 10 = 50$ 次。

选了5次硬币，每次记为 $z_i \in \{A, B\}$ ，5次合到一起记为 $z = (z_1, z_2, z_3, z_4, z_5)$ ；每选1次硬币(抛10次)，我们记录其中正面出现的次数 $x_i \in \{0, 1, \dots, 10\}$ ，5次合到一起记为 $x = (x_1, x_2, x_3, x_4, x_5)$ 。于是，很容易我们就可以评估出 θ_A ：

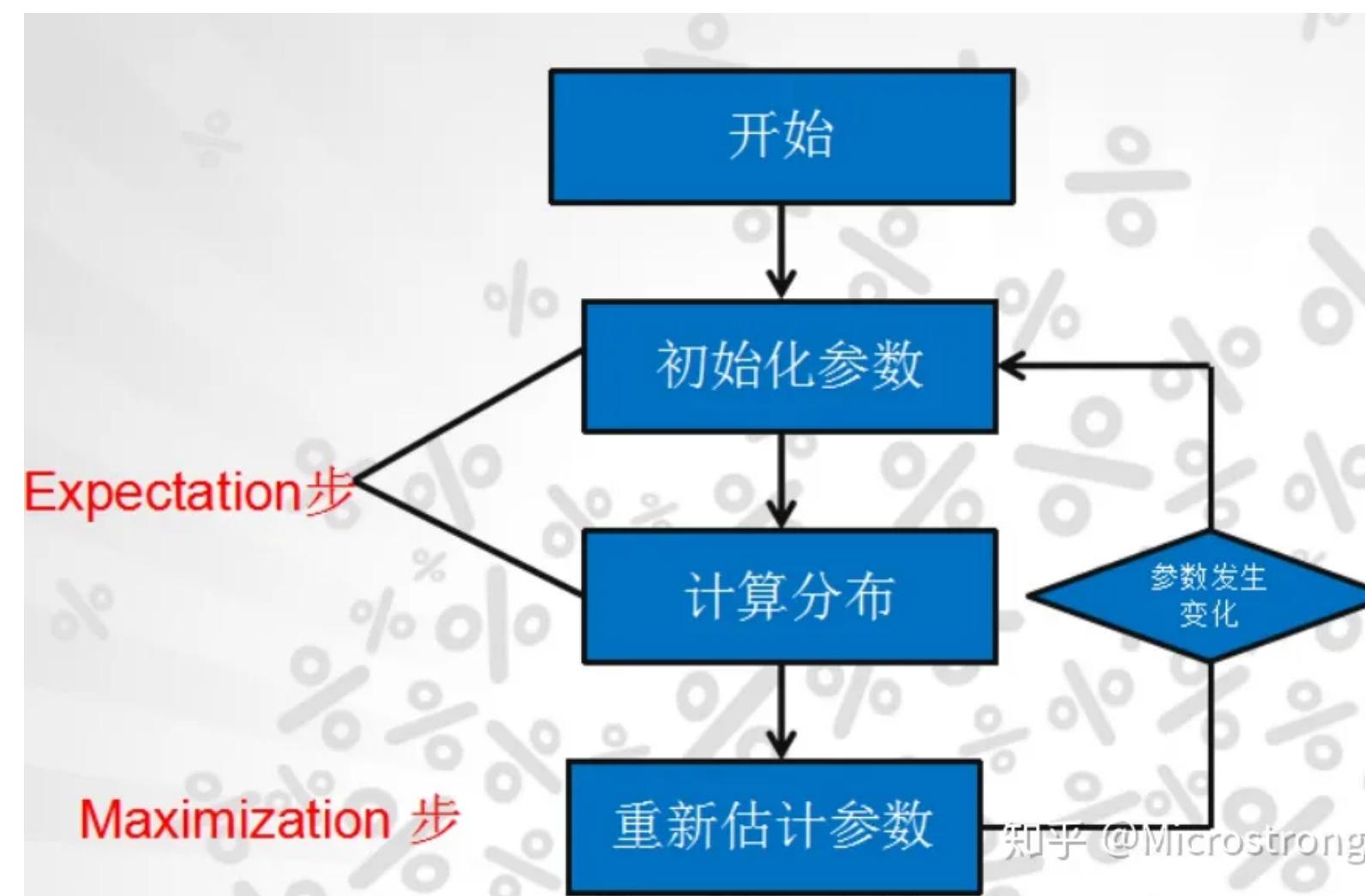
$$\theta_A = \frac{A \text{ 硬币抛出正面的次数}}{A \text{ 硬币抛出的总次数}}$$

这种通过观测来评估模型参数的方法称为极大似然评估(Maximum likelihood estimation)。

4.2.3 HMM模型学习问题

- EM算法

- 问题：已知一组独立的样本 $X = \{x_1, x_2, \dots, x_m\}$, 求模型 $p(x, z)$ 的参数 θ , 使 $p(x, z)$ 最大, 其中 z 是隐变量。
- E步骤：利用模型现有参数求隐变量取值的期望，猜测最可能的概率分布
- M步骤：利用当前对隐变量估计值对模型参数最大似然估计，更新模型



4.2.3 HMM模型学习问题

- EM算法推导流程

2.1.1 凸函数

设是定义在实数域上的函数，如果对于任意的实数，都有：

$$f'' \geq 0$$

那么是凸函数。若不是单个实数，而是由实数组成的向量，此时，如果函数的 Hesse 矩阵是半正定的，即

$$H'' \geq 0$$

是凸函数。特别地，如果 $f'' > 0$ 或者 $H'' > 0$ ，称为严格凸函数。

4.2.3 HMM模型学习问题

- EM算法推导流程

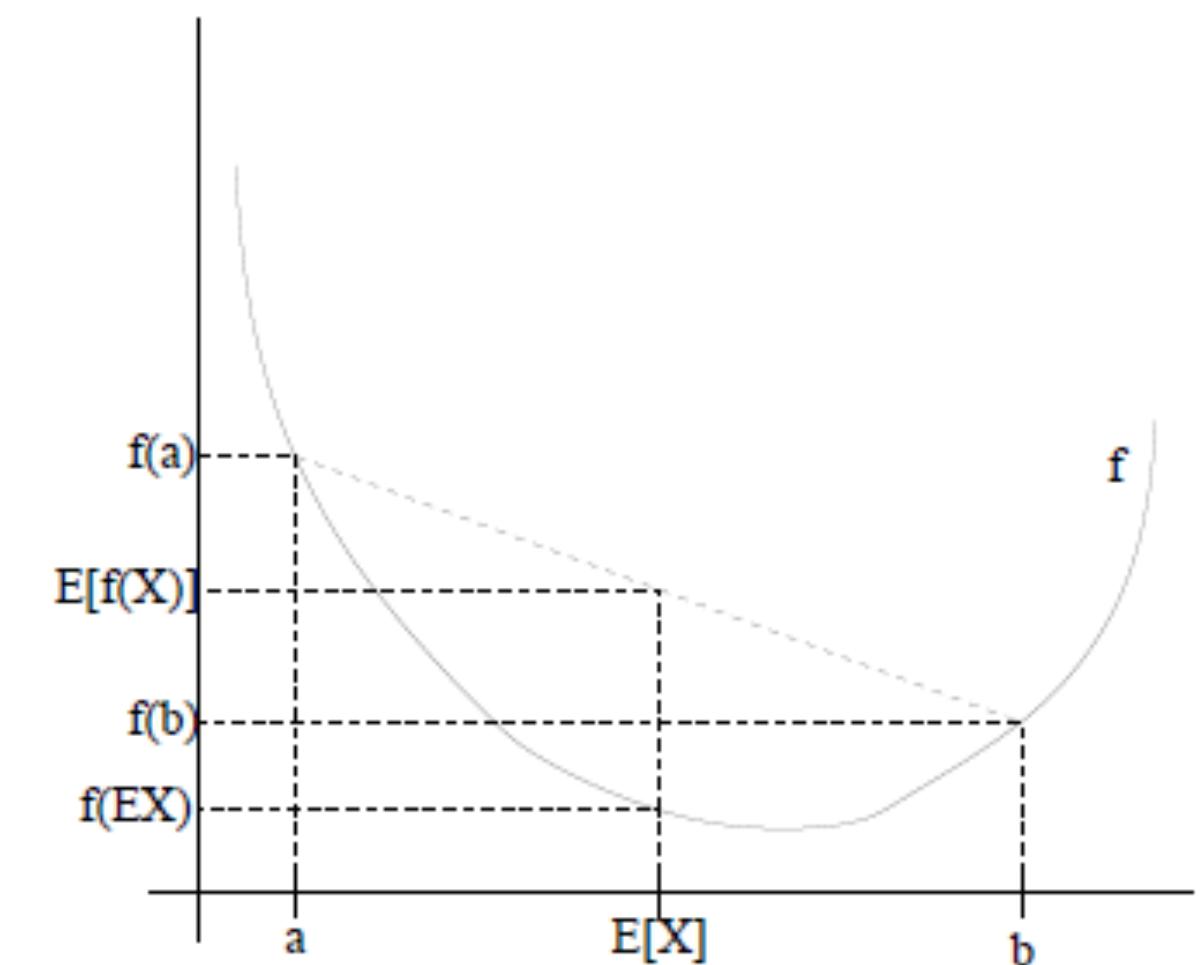
2.1.2 Jensen不等式

如下图，如果函数 f 是凸函数， x 是随机变量，有 0.5 的概率是 a ，有 0.5 的概率是 b ， x 的期望值就是 a 和 b 的中值了那么：

$$E[f(x)] \geq f(E(x))$$

其中， $E[f(x)] = 0.5f(a) + 0.5f(b)$ ， $f(E(x)) = f(0.5a + 0.5b)$ ，这里 a 和 b 的权值为 0.5， $f(a)$ 与 a 的权值相等， $f(b)$ 与 b 的权值相等。

特别地，如果函数 f 是严格凸函数，当且仅当： $p(x = E(x)) = 1$ (即随机变量是常量) 时等号成立。



4.2.3 HMM模型学习问题

2.1.3 期望

- EM算法推导流程

对于离散型随机变量 X 的概率分布为 $p_i = p\{X = x_i\}$ ，数学期望 $E(X)$ 为：

$$E(X) = \sum_i x_i p_i$$

p_i 是权值，满足两个条件 $1 \geq p_i \geq 0$ ， $\sum_i p_i = 1$ 。

若连续型随机变量 X 的概率密度函数为 $f(x)$ ，则数学期望 $E(X)$ 为：

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx$$

设 $Y = g(X)$ ，若 X 是离散型随机变量，则：

$$E(Y) = \sum_i g(x_i) p_i$$

若 X 是连续型随机变量，则：

$$E(Y) = \int_{-\infty}^{+\infty} g(x) f(x) dx$$

4.2.3 HMM模型学习问题

- EM算法推导流程

对于 m 个相互独立的样本 $x = (x^{(1)}, x^{(2)}, \dots, x^{(m)})$ ，对应的隐含数据 $z = (z^{(1)}, z^{(2)}, \dots, z^{(m)})$ ，此时 (x, z) 即为完全数据，样本的模型参数为 θ ，则观察数据 $x^{(i)}$ 的概率为 $P(x^{(i)} | \theta)$ ，完全数据 $(x^{(i)}, z^{(i)})$ 的似然函数为 $P(x^{(i)}, z^{(i)} | \theta)$ 。

假如没有隐含变量 z ，我们仅需要找到合适的 θ 极大化对数似然函数即可：

$$\theta = \arg \max_{\theta} L(\theta) = \arg \max_{\theta} \sum_{i=1}^m \log P(x^{(i)} | \theta)$$

增加隐含变量 z 之后，我们的目标变成了找到合适的 θ 和 z 让对数似然函数极大：

$$\theta, z = \arg \max_{\theta, z} L(\theta, z) = \arg \max_{\theta, z} \sum_{i=1}^m \log \sum_{z^{(i)}} P(x^{(i)}, z^{(i)} | \theta)$$

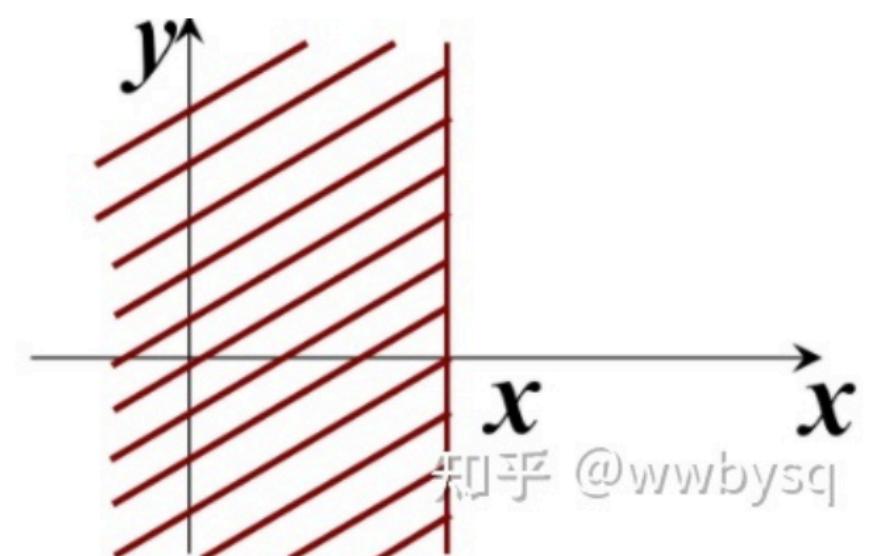
不就是多了一个隐变量 z 吗？那我们自然而然会想到分别对未知的 θ 和 z 分别求偏导，这样做可行吗？

4.2.3 HMM模型学习问题

• EM算法推导流程

理论上是可行的，然而如果对分别对未知的 θ 和 z 分别求偏导，由于 $\log P(x^{(i)} | \theta)$ 是 $P(x^{(i)}, z^{(i)} | \theta)$ 边缘概率（建议没基础的同学网上搜一下边缘概率的概念），转化为 $\log P(x^{(i)} | \theta)$ 求导后形式会非常复杂（可以想象下 $\log(f_1(x) + f_2(x) + \dots)$ 复合函数的求导），所以很难求解得到 θ 和 z 。那么我们想一下可不可以将加号从 \log 中提取出来呢？我们对这个式子进行缩放如下：

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= P(X \leq x, Y < +\infty) \\ &= F(x, +\infty) \end{aligned}$$



$$\sum_{i=1}^m \log \sum_{z^{(i)}} P(x^{(i)}, z^{(i)} | \theta) = \sum_{i=1}^m \log \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{P(x^{(i)}, z^{(i)} | \theta)}{Q_i(z^{(i)})} \quad (1)$$

$$\geq \sum_{i=1}^m \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)} | \theta)}{Q_i(z^{(i)})} \quad (2)$$

上面第(1)式引入了一个未知的新的分布 $Q_i(z^{(i)})$ ，满足：

$$P(X = x_i, Y = y_j) = p_{ij}$$

$$P(X = x_i) = \sum_{j=1}^{\infty} p_{ij}$$

$$\sum_z Q_i(z) = 1, 0 \leq Q_i(z) \leq 1$$

第(2)式用到了 Jensen 不等式（对数函数是凹函数）：

$$\log(E(y)) \geq E(\log(y))$$

4.2.3 HMM模型学习问题

其中：

- EM算法推导流程

$$E(y) = \sum_i \lambda_i y_i, \lambda_i \geq 0, \sum_i \lambda_i = 1$$

$$y_i = \frac{P(x^{(i)}, z^{(i)} | \theta)}{Q_i(z^{(i)})}$$

$$\lambda_i = Q_i(z^{(i)})$$

也就是说 $\frac{P(x^{(i)}, z^{(i)} | \theta)}{Q_i(z^{(i)})}$ 为第 i 个样本， $Q_i(z^{(i)})$ 为第 i 个样本对应的权重，那么：

$$E(\log \frac{P(x^{(i)}, z^{(i)} | \theta)}{Q_i(z^{(i)})}) = \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)} | \theta)}{Q_i(z^{(i)})}$$

上式我实际上是我们构建了 $L(\theta, z)$ 的下界，我们发现实际上就是 $\log \frac{P(x^{(i)}, z^{(i)} | \theta)}{Q_i(z^{(i)})}$ 的加权求和，由于上面讲过权值 $Q_i(z^{(i)})$ 累积和为1，因此上式是 $\log \frac{P(x^{(i)}, z^{(i)} | \theta)}{Q_i(z^{(i)})}$ 的加权平均，也是我们所说的期望，这就是Expectation的来历啦。下一步要做的就是寻找一个合适的 $Q_i(z)$ 最优化这个下界(M步)。

4.2.3 HMM模型学习问题

- EM算法推导流程

假设 θ 已经给定，那么 $\log L(\theta)$ 的值就取决于 $Q_i(z)$ 和 $p(x^{(i)}, z^{(i)})$ 了。我们可以通过调整这两个概率使下界逼近 $\log L(\theta)$ 的真实值，当不等式变成等式时，说明我们调整后的下界能够等价于 $\log L(\theta)$ 了。由 Jensen 不等式可知，等式成立的条件是随机变量是常数，则有：

$$\frac{P(x^{(i)}, z^{(i)} | \theta)}{Q_i(z^{(i)})} = c$$

其中 c 为常数，对于任意 i ，我们得到：

$$P(x^{(i)}, z^{(i)} | \theta) = c Q_i(z^{(i)})$$

方程两边同时累加和：

$$\sum_z P(x^{(i)}, z^{(i)} | \theta) = c \sum_z Q_i(z^{(i)})$$

4.2.3 HMM模型学习问题

- EM算法推导流程

由于 $\sum_z Q_i(z^{(i)}) = 1$ 。从上面两式，我们可以得到：

$$\sum_z P(x^{(i)}, z^{(i)} | \theta) = c$$

$$Q_i(z^{(i)}) = \frac{P(x^{(i)}, z^{(i)} | \theta)}{c} = \frac{P(x^{(i)}, z^{(i)} | \theta)}{\sum_z P(x^{(i)}, z^{(i)} | \theta)} = \frac{P(x^{(i)}, z^{(i)} | \theta)}{P(x^{(i)} | \theta)} = P(z^{(i)} | x^{(i)}, \theta)$$

其中：

边缘概率公式： $P(x^{(i)} | \theta) = \sum_z P(x^{(i)}, z^{(i)} | \theta)$

条件概率公式： $\frac{P(x^{(i)}, z^{(i)} | \theta)}{P(x^{(i)} | \theta)} = P(z^{(i)} | x^{(i)}, \theta)$

从上式可以发现 $Q(z)$ 是已知样本和模型参数下的隐变量分布。

4.2.3 HMM模型学习问题

- EM算法推导流程

如果 $Q_i(z^{(i)}) = P(z^{(i)} | x^{(i)}, \theta)$, 则第 (2) 式是我们的包含隐藏数据的对数似然的一个下界。

如果我们能极大化这个下界, 则也在尝试极大化我们的对数似然。即我们需要极大化下式:

$$\arg \max_{\theta} \sum_{i=1}^m \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)} | \theta)}{Q_i(z^{(i)})}$$

至此, 我们推出了在固定参数 θ 后分布 $Q_i(z^{(i)})$ 的选择问题, 从而建立了 $\log L(\theta)$ 的下界, 这是 E 步, 接下来的 M 步骤就是固定 $Q_i(z^{(i)})$ 后, 调整 θ , 去极大化 $\log L(\theta)$ 的下界。

去掉上式中常数的部分 $Q_i(z^{(i)})$, 则我们需要极大化的对数似然下界为:

$$\arg \max_{\theta} \sum_{i=1}^m \sum_{z^{(i)}} Q_i(z^{(i)}) \log P(x^{(i)}, z^{(i)} | \theta)$$

4.2.3 HMM模型学习问题

输入：观察数据 $x = (x^{(1)}, x^{(2)}, \dots, x^{(m)})$, 联合分布 $p(x, z|\theta)$, 条件分布 $p(z|x, \theta)$, 极大迭代次数 J 。

- **EM算法流程**

1) 随机初始化模型参数 θ 的初值 θ^0

2) for j from 1 to J:

- E步：计算联合分布的条件概率期望：

$$Q_i(z^{(i)}) := \underline{P(z^{(i)} | x^{(i)}, \theta)}$$

- M步：极大化 $L(\theta)$, 得到 θ :

$$\theta := \arg \max_{\theta} \sum_{i=1}^m \sum_{z^{(i)}} Q_i(z^{(i)}) \log P(x^{(i)}, z^{(i)} | \theta)$$

- 重复E、M步骤直到 θ 收敛

输出：模型参数 θ

4.2.3 HMM模型学习问题

- EM算法案例

例：假设有两枚硬币A、B，以相同的概率随机选择一个硬币，进行如下的掷硬币实验：共做5次实验，每次实验独立的掷十次(H代表正面朝上)。a是在知道每次选择的是A还是B的情况下进行，b是在不知道选择的是A还是B的情况下进行，问如何估计两个硬币正面出现的概率？

4.2.3 HMM模型学习问题

a Maximum likelihood



HTTTTHHTHTH
 HHHTTHHHHH
 HTHHHHHTHH
 HTHTTTHTHTT
 THHHHTHHHTH

5 sets, 10 tosses per set

Coin A	Coin B
	5 H, 5 T
9 H, 1 T	
8 H, 2 T	
	4 H, 6 T
7 H, 3 T	
24 H, 6 T	9 H, 11 T

$$\hat{\theta}_A = \frac{24}{24 + 6} = 0.80$$

$$\hat{\theta}_B = \frac{9}{9 + 11} = 0.45$$

CASE a

已知每个实验选择的是硬币A还是硬币B，重点是如何计算输出的概率分布，这其实也是极大似然求导所得。

$$\begin{aligned} \underset{\theta}{\operatorname{argmax}} \log P(Y|\theta) &= \log((\theta_B^5(1-\theta_B)^5)(\theta_A^9(1-\theta_A))(\theta_A^8(1-\theta_A)^2)(\theta_B^4(1-\theta_B)^6)(\theta_A^7(1-\theta_A)^3)) \\ &= \log[(\theta_A^{24}(1-\theta_A)^6)(\theta_B^9(1-\theta_B)^{11})] \end{aligned}$$

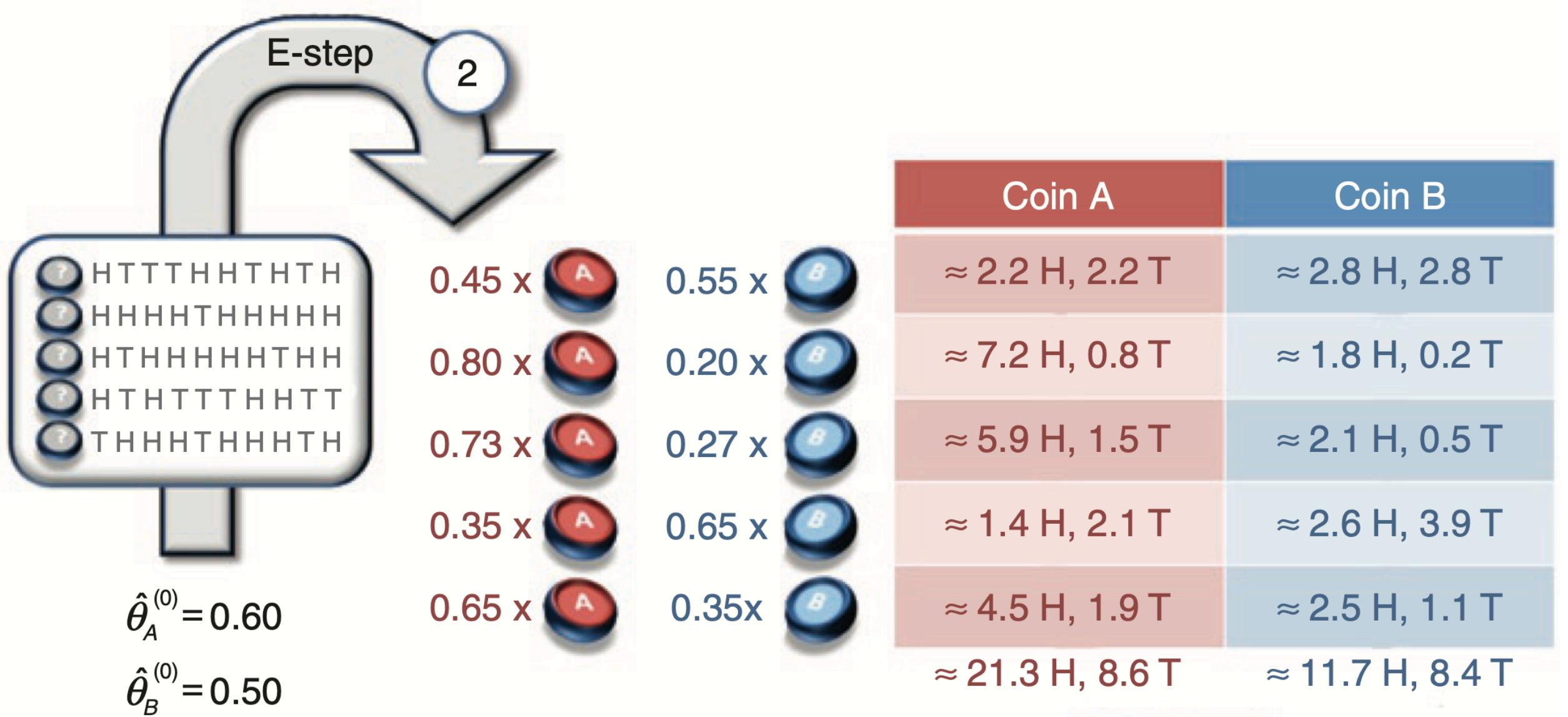
上面这个式子求导之后发现，5次实验中A正面向上的次数再除以总次数作为即为 $\hat{\theta}_A$ ，5次实验中B正面向上的次数再除以总次数作为即为 $\hat{\theta}_B$ ，即：

$$\hat{\theta}_A = \frac{24}{24 + 6} = 0.80$$

$$\hat{\theta}_B = \frac{9}{9 + 11} = 0.45$$

4.2.3 HMM模型学习问题

b Expectation maximization



$$\hat{\theta}_A^{(1)} \approx \frac{21.3}{21.3 + 8.6} \approx 0.71$$

$$\hat{\theta}_B^{(1)} \approx \frac{11.7}{11.7 + 8.4} \approx 0.58$$

$$\hat{\theta}_A^{(10)} \approx 0.80$$

$$\hat{\theta}_B^{(10)} \approx 0.52$$

$$P(B|A) = \frac{P(AB)}{P(A)}$$

E-Step: $Q_i(z^{(i)}) := P(z^{(i)} | x^{(i)}, \theta)$

(1) 计算第一次选硬币，且投出的结果为H-T-T-T-H-H-T-H-T-H时，选的是硬币A的概率

根据二项分布定义，如果第一次选的是A硬币，投10次有5次为正面的概率为：

$$P(HTTTHHTHTH|A) = C_{10}^5 (0.6)^5 (1-0.6)^5$$

同理，可计算出如果第一次选的是B硬币，投10次有5次为正面的概率：

$$P(HTTTHHTHTH|B) = C_{10}^5 (0.5)^5 (1-0.5)^5$$

再由贝叶斯定律可计算出，投币结果为H-T-T-T-H-H-T-H-T-H，且这一结果是硬币A投出的概率为：

$$\begin{aligned} P(A|HTTTHHTHTH) &= \frac{P(HTTTHHTHTH|A)P(A)}{P(HTTTHHTHTH|A)P(A) + P(HTTTHHTHTH|B)P(B)} \\ &= \frac{C_{10}^5 (0.6)^5 (1-0.6)^5 \times 0.5}{C_{10}^5 (0.6)^5 (1-0.6)^5 \times 0.5 + C_{10}^5 (0.5)^5 (1-0.5)^5 \times 0.5} \\ &\approx 0.45 \end{aligned}$$

注意：其中 $P(A)$ 和 $P(B)$ 都为0.5，因为抽取A硬币和抽取B硬币是等可能的。

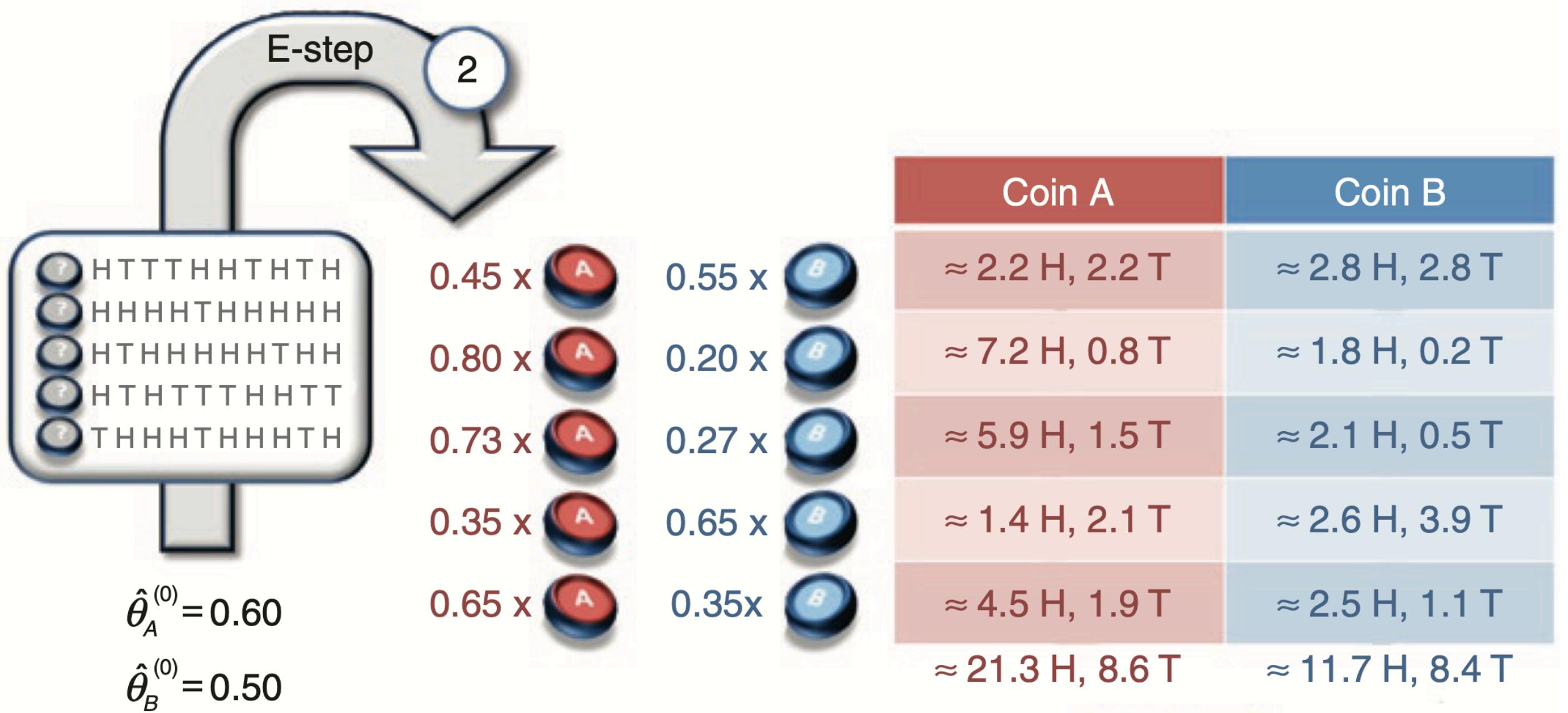
同理，投币结果为H-T-T-T-H-H-T-H-T-H，且这一结果是硬币B投出的概率为：

$$P(B|HTTTHHTHTH) \approx 0.55$$

4.2.3 HMM模型学习问题

(2) 再依次计算第二到第五次选硬币时，抽取到A/B硬币的概率。所有结果汇总如下表：

b Expectation maximization



$$P(B | A) = \frac{P(AB)}{P(A)}$$

Observation	Coin A	Coin B
HTTHHHTHTH	0.45	0.55
HHHHTHHHHH	0.80	0.20
HTHHHHHTHH	0.73	0.27
HTHTTTHHTT	0.35	0.65
THHHHTHHHTH	0.65	0.35
Total	2.98	2.02
Score	21.24	11.76

其中，Coin A列表示投出该行硬币正反面情况下，推测是用A硬币来投的概率。

Total行是计算的5次试验的总和，而Score行的计算如下：

$$\text{Score} = (0.45 \times 5) + (0.8 \times 9) + (0.73 \times 8) + (0.35 \times 4) + (0.65 \times 7) = 21.24$$

仔细回想，这里的Score其实就是5次选币（50次投币）事件的期望值(Expectation)！以上整个过程也正是在计算 $P(z^{(i)} | x^{(i)}; \theta)$ ，即

每次观测(Observation)索引 $\Rightarrow i$

当前 θ 参数条件 $\Rightarrow (\theta_A, \theta_B)$

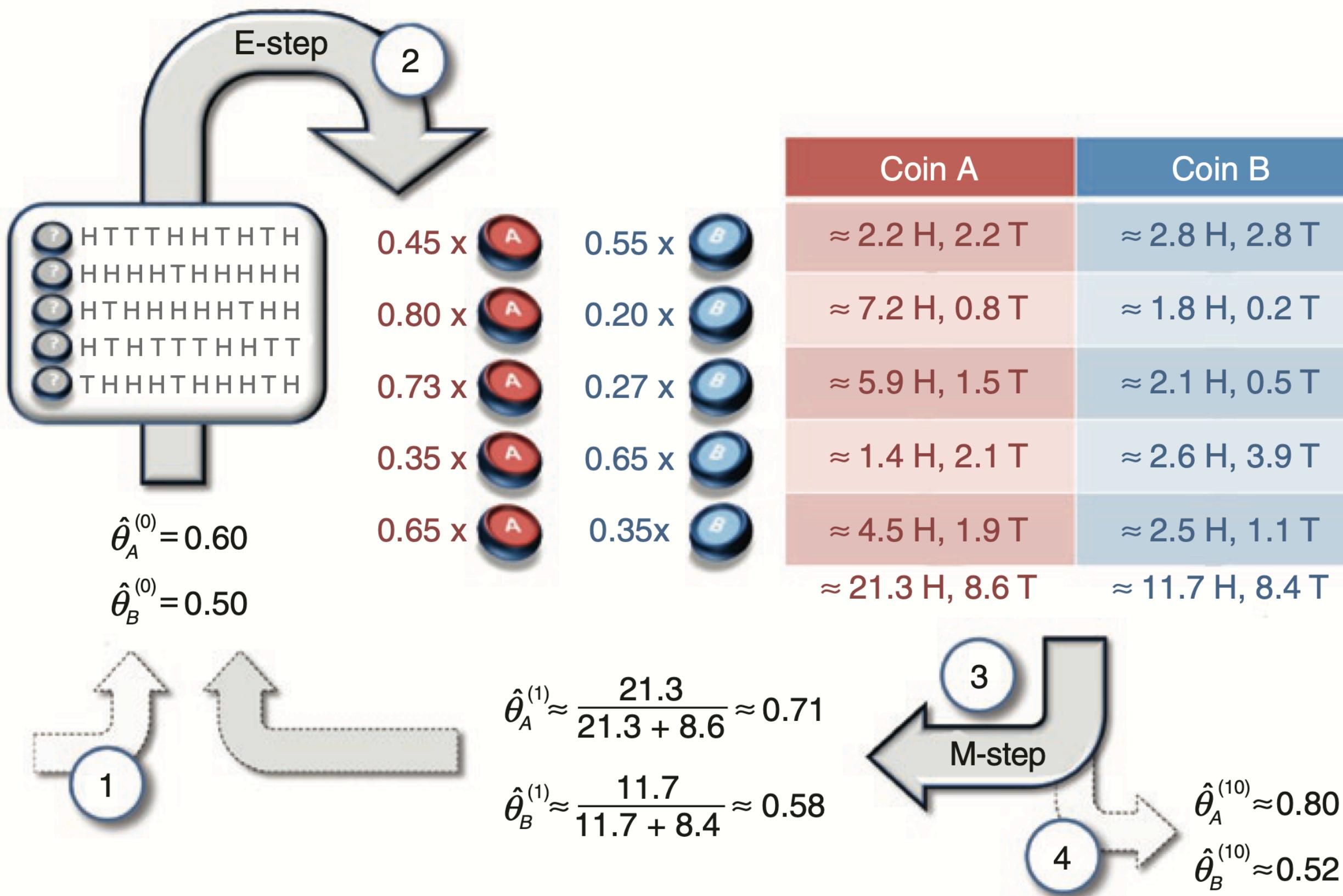
每次观测，如H-T-T-T-H-H-T-H-T-H $\Rightarrow x^{(i)}$

挑选的硬币 (Coin A, Coin B) $\Rightarrow z^{(i)}$

的概率 $P(z^{(i)} | x^{(i)}; \theta)$

4.2.3 HMM模型学习问题

b Expectation maximization



另外，我们可以在文中看到如下表格，计算的也是一样的，以第二行为例Coin A列的2.2H表示第一次选硬币，再抛10次时硬币为H的总和0.45×5； Coin B列的2.8H表示0.55×5

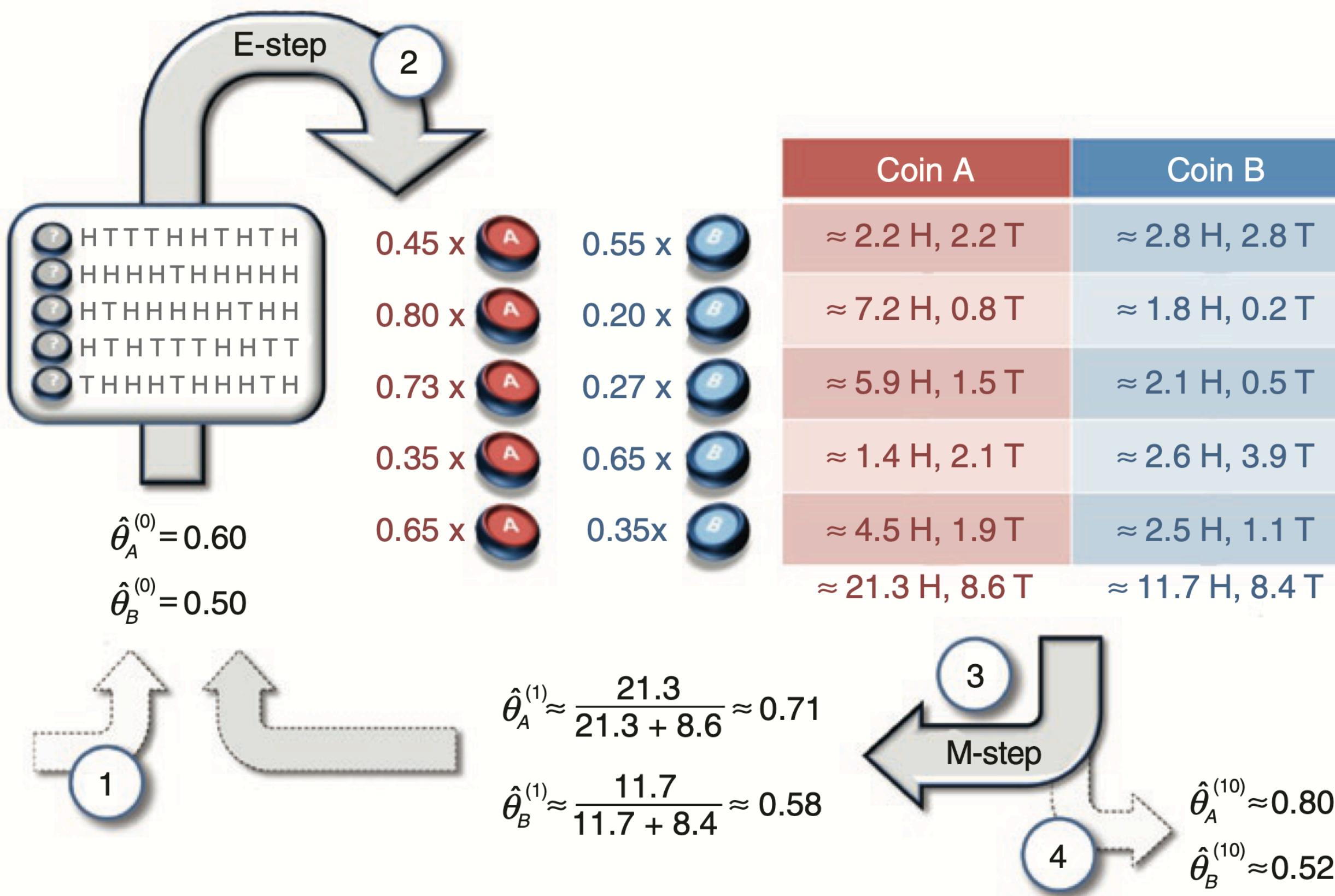
	Coin A	Coin B
	$\approx 2.2 \text{ H}, 2.2 \text{ T}$	$\approx 2.8 \text{ H}, 2.8 \text{ T}$
	$\approx 7.2 \text{ H}, 0.8 \text{ T}$	$\approx 1.8 \text{ H}, 0.2 \text{ T}$
	$\approx 5.9 \text{ H}, 1.5 \text{ T}$	$\approx 2.1 \text{ H}, 0.5 \text{ T}$
	$\approx 1.4 \text{ H}, 2.1 \text{ T}$	$\approx 2.6 \text{ H}, 3.9 \text{ T}$
	$\approx 4.5 \text{ H}, 1.9 \text{ T}$	$\approx 2.5 \text{ H}, 1.1 \text{ T}$
	$\approx 21.3 \text{ H}, 8.6 \text{ T}$	$\approx 11.7 \text{ H}, 8.4 \text{ T}$

同理，我们也可以算出观测到是HTTTHHTHTH时Coin A抛出反面的概率得分2.2T

$$P(B | A) = \frac{P(AB)}{P(A)}$$

4.2.3 HMM模型学习问题

b Expectation maximization



$$\theta := \arg \max_{\theta} \sum_{i=1}^m \sum_{z^{(i)}} Q_i(z^{(i)}) \log P(x^{(i)}, z^{(i)} | \theta)$$

M步：求出似然函数下界 $Q(\theta, \theta^i)$, y_j 代表第 j 次实验正面朝上的个数, μ_j 代表第 j 次实验选择硬币 A 的概率, $1 - \mu_j$ 代表第 j 次实验选择硬币 B 的概率。

$$\begin{aligned} Q(\theta, \theta^i) &= \sum_{j=1}^5 \sum_z P(z|y_j, \theta^i) \log P(y_j, z|\theta) \\ &= \sum_{j=1}^5 \mu_j \log(\theta_A^{y_j} (1 - \theta_A)^{10-y_j}) + (1 - \mu_j) \log(\theta_B^{y_j} (1 - \theta_B)^{10-y_j}) \end{aligned}$$

针对L函数求导来对参数求导, 例如对 θ_A 求导:

$$\begin{aligned} \frac{\partial Q}{\partial \theta_A} &= \mu_1 \left(\frac{y_1}{\theta_A} - \frac{10 - y_1}{1 - \theta_A} \right) + \cdots + \mu_5 \left(\frac{y_5}{\theta_A} - \frac{10 - y_5}{1 - \theta_A} \right) = \mu_1 \left(\frac{y_1 - 10\theta_A}{\theta_A(1 - \theta_A)} \right) + \cdots \\ &\quad + \mu_5 \left(\frac{y_5 - 10\theta_A}{\theta_A(1 - \theta_A)} \right) \\ &= \frac{\sum_{j=1}^5 \mu_j y_j - \sum_{j=1}^5 10\mu_j \theta_A}{\theta_A(1 - \theta_A)} \end{aligned}$$

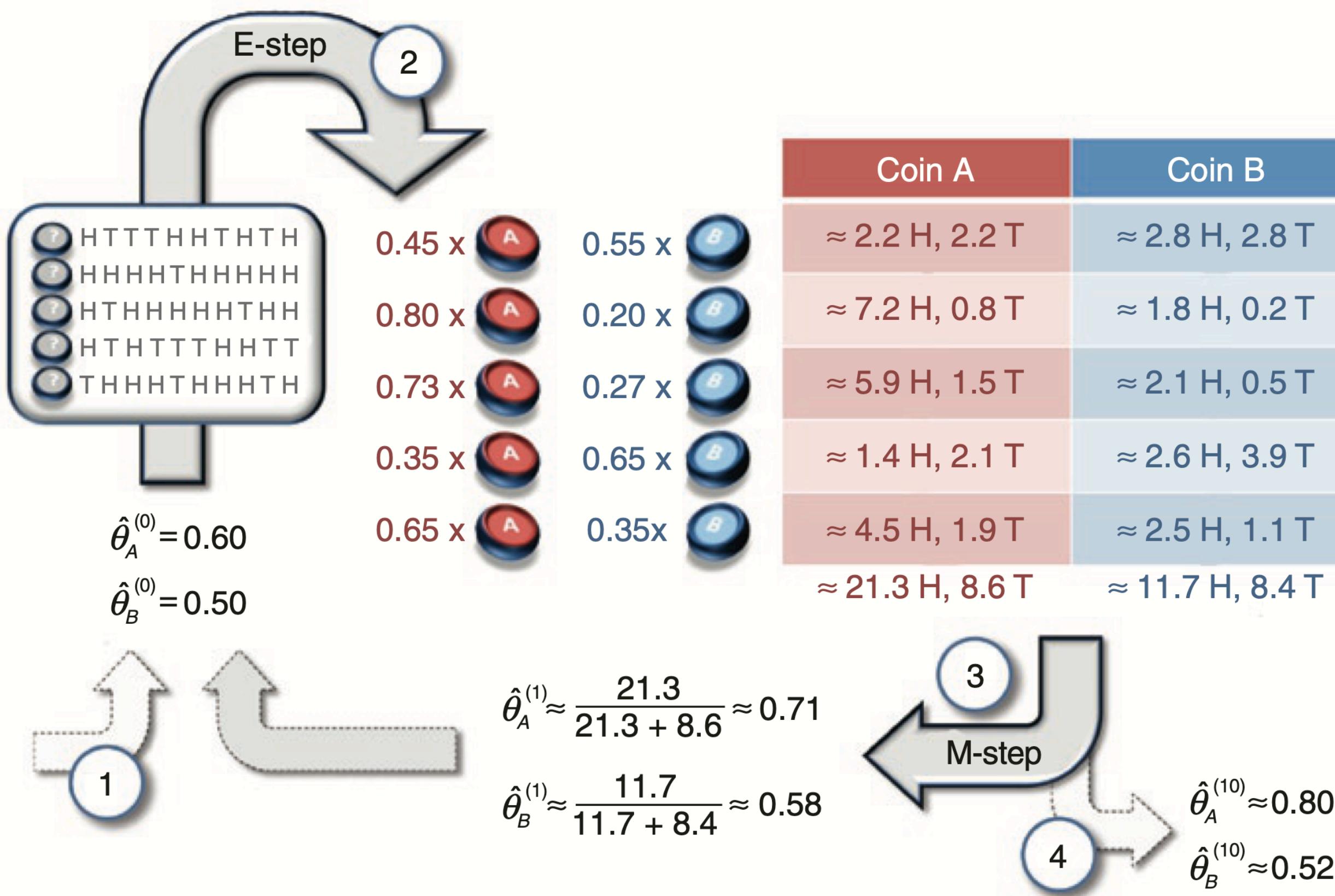
求导等于 0 之后就可得到图中的第一次迭代之后的参数值:

$$\hat{\theta}_A^{(1)} = 0.71$$

$$\hat{\theta}_B^{(1)} = 0.58$$

4.2.3 HMM模型学习问题

b Expectation maximization



第二轮迭代：基于第一轮EM计算好的 $\hat{\theta}_A^{(1)}$, $\hat{\theta}_B^{(1)}$ ，进行第二轮 EM，计算每个实验中选择的硬币是 A 和 B 的概率（E步），然后在计算M步，如此继续迭代……迭代十步之后
 $\hat{\theta}_A^{(10)} = 0.8$, $\hat{\theta}_B^{(10)} = 0.52$

4.3 朴素贝叶斯分类器

- 贝叶斯分类器

$$P(B_i | A) = \frac{P(AB_i)}{P(A)} = \frac{P(B_i)P(A | B_i)}{P(A)}$$

$$P(\text{类别} | \text{特征}) = \frac{P(\text{特征} | \text{类别})P(\text{类别})}{P(\text{特征})}$$

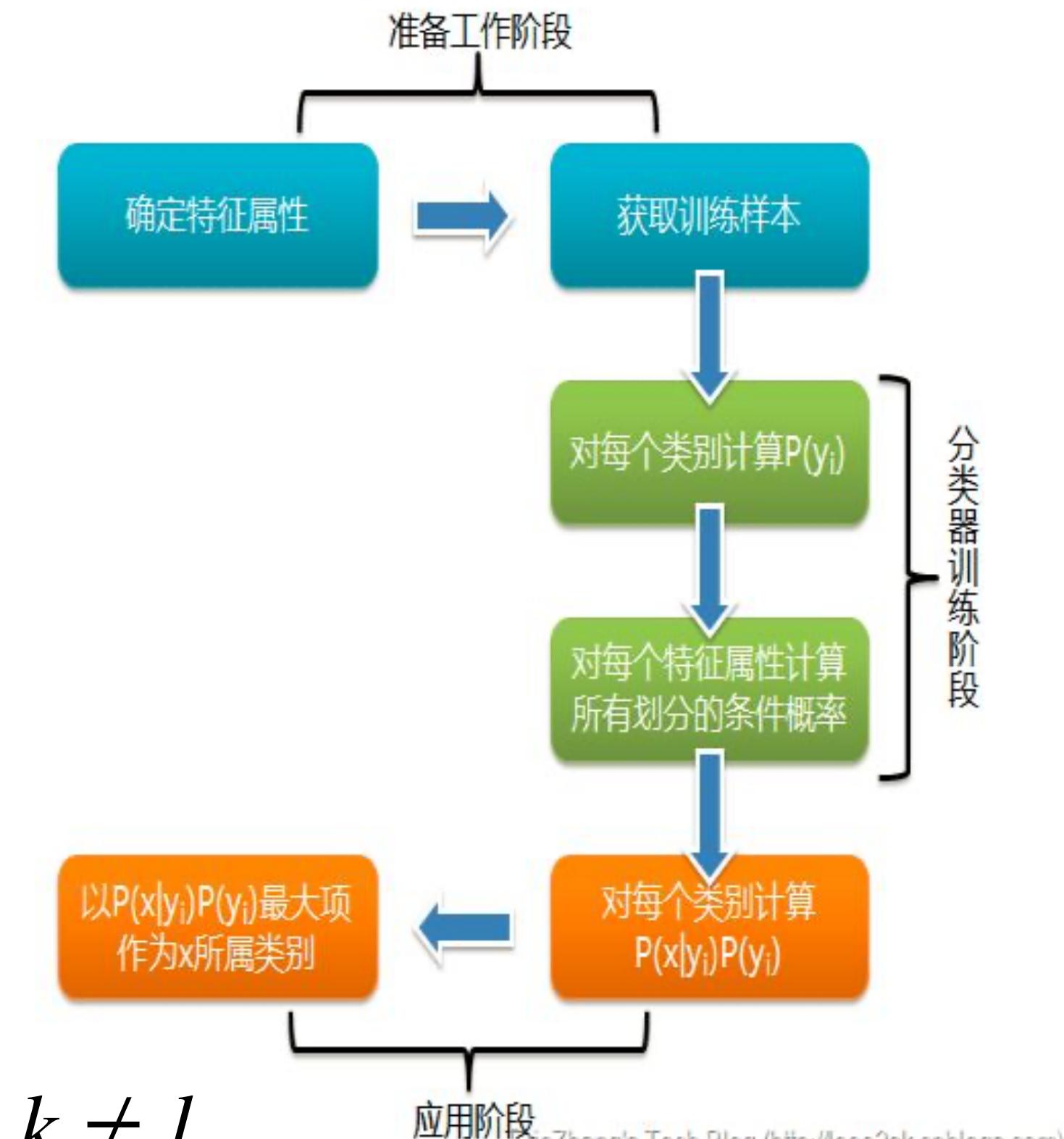
- 朴素贝叶斯分类器

- 假设各个特征取值只依赖类别标签，**特征间互相独立**：**朴素**

$$p(x_l x_k | \omega_i) = p(x_l | \omega_i)p(x_k | \omega_i), \quad l, k = 1, \dots, d, k \neq l$$

- 联合概率可以分解为

$$p(x_1, x_2, \dots, x_d, \omega_i) = p(x_1 | \omega_i)p(x_2 | \omega_i) \cdots p(x_d | \omega_i)p(\omega_i)$$



4.3 朴素贝叶斯分类器

- 朴素贝叶斯分类器

- 类别先验概率估计

$$p(Y = \omega_i) = \frac{\sum_{i=1}^N I(y_i = \omega_i)}{N}$$

- 特征的条件概率：参数的极大似然估计

$$p(x_j = v_i \mid Y = \omega_i) = \frac{\sum_{i=1}^N I(x_j^{(i)} = v_i, y_i = \omega_i)}{\sum_{i=1}^N I(y_i = \omega_i)}$$

4.3 朴素贝叶斯分类器

- 朴素贝叶斯分类器
 - 零概率问题：不能因为某事件没有观察到就武断地认为该事件的概率为 0
 - 拉普拉斯平滑：对概率值平滑矫正，解决样本量少或者特征取值概率较低的情况

$$\bullet p(Y = \omega_i) = \frac{\sum_{i=1}^N I(y_i = \omega_i) + 1}{N + C}$$

$$\bullet p(x_j = v_i | Y = \omega_i) = \frac{\sum_{i=1}^N I(x_j^{(i)} = v_i, y_i = \omega_i) + 1}{\sum_{i=1}^N I(y_i = \omega_i) + S_j}$$